Journal Homepage: - www.journalijar.com
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)
Article DOI: 10.21474/IJAR01/7621
DOI URL: http://dx.doi.org/10.21474/IJAR01/7621

RESEARCH ARTICLE

NEXT GENERATION SEQUENCE ANALYSIS OF SEQUENCES FROM SRA FILE OF HIV-1 ENVELOPE PROTEIN.

Sahil Sinha, Ganesh Chandra Sahoo§, Sindhu Prabha Rana, RK Topno, Md Yusuf Ansari, Manas Ranjan Dikhit, Rishikesh Kumar, Krishna Panday and Pradeep Das.

Manuscript Info

Manuscript History

Received: 23 June 2018
Final Accepted: 25 July 2018
Published: August 2018

Keywords:-
Next Generation Sequencing (NGS), Human Immunodeficiency Virus-1 (HIV-1) 454 (Roche) envelope protein, Clustal X, TreeView, NextGENe, Single-nucleotide polymorphism (SNPs), Self-made program (SMP).

Abstract

Acquired Immunodeficiency Syndrome (AIDS) caused by Human Immunodeficiency Virus is a serious public health concern globally. Therefore, the Next-Generation Sequencing (NGS) analysis is critical elucidating different parts of envelope sequences of Human Immunodeficiency Virus-1 (HIV-1). The Phylogenetic relationships with other viral sex transmitted diseases (STD), bacterial STD, fungal and protozoan sequences have revealed that different amino acid substitutions have accumulated over years and have shown clustering with other Human Immunodeficiency Virus sequences. Single-nucleotide polymorphism (SNP) analysis has shown variations occurred in different regions at various frequencies, as per the need of the HIV genome for replication. The Human Immunodeficiency Virus (HIV) reads coverage map shows peak variation positions at various frequency levels of the Human Immunodeficiency Virus-1 (HIV-1) Sequence Read Archive (SRA) file. Its nucleotide range from 84000 to 98000 shows the highest frequency of SNPs. Human Immunodeficiency Virus (HIV-1) 454 (Roche) analysis has shown different expression patterns of different Human Immunodeficiency Virus (HIV) samples and Single-Nucleotide Polymorphism at different frequencies in the HIV population showed causing drug resistance abilities. The coverage curve map shows the two most coverage ranges of nucleotide sequences between “80000-100000” and “180000-200000” against the tag number of the same range between “60-65” on the y-axis. The self-made program (SMP) taking Human Immunodeficiency Virus-1 (HIV-1) envelope protein sequence files from different countries conveyed information about their mutational changes, substitutions, transitions, trans-version, ratio of transition versus trans-version and silent mutation occurring at different sequences of nucleotide. The docking score between the protein 2NY1 and its inhibitor BMS-378806 revealed the strength of the binding ability between them.

Corresponding Author:- Sahil Sinha.

Introduction:-

Acquired Immunodeficiency Syndrome (AIDS) caused firstly by Human Immunodeficiency Virus belongs to genus *Lent virus* and family *Retroviridae*. It is responsible for the failure of the immune system in the infected body that leads to opportunistic infection with another virus and bacterial infection (Gyorkey, Melnick et al. 1987). HIV became a public health concern, demanding global health development in the year 1981 (Montagnier 2002, Girard, Osmanov et al. 2006). HIV could get infected by the transfer of pre-ejaculate which is the life-threatening infection and eventually weakens the body defences by destroying certain types of white blood cells (Hawkins, Blott et al. 2005, Checkley, Luttge et al. 2011, Gobind 2014, Dikhit, Kumar et al. 2017). Thus, every year the number of people dies from AIDS-related causes and it is getting declined from 2.3 million people to 1.6 million between the years 2005 to 2012. There has been remarkable progress in getting the biology of HIV-1 and its recognition by the human immune system, yet an efficacious HIV-1 vaccine has not been developed. The main challenges for the development of the vaccine against HIV-1 are immutability, genetic variability, and diversity. It leads to the creation of a plethora of antigens changing constantly on the surface of viral glycoprotein. Its structural features disguise conserved receptor-binding sites from the immune system, and potential epitopes are shielded from antibodies by the presence of carbohydrate moieties (Stein 1990, Dikhit, Amit et al. 2017). The rapid advance of computation work and Bioinformatics approaches has increased our ability to understand HIV-1 structure and antibody approach towards immunogenic design. Now newly acquired knowledge, the risk of HIV infection can be reduced by the possibility of using vaccine & opened up new and promising pathways towards the progress of developing an immediate effective HIV-1 vaccine (Schwartländer, Stover et al. 2011, Dikhit, Ansari et al. 2016). Although to prevent HIV infection remains the ultimate goal, Vaccines examination that could remarkably change the course of the disease and the cause of infection of people infected with HIV, which could provide positive health concern and benefactor for infected individuals and the larger community (Kelly and Kalichman 2002) (Masetshaba 2016).

To cope up with & tackle this pandemic infection, the next generation sequencing (NGS) analyses from the SRA file of HIV-1 envelope proteins have been performed. The reason behind selecting the NGS analysis as it can analyse and manage a large amount of data elucidating more specific information. More data means more information can be deduced and conveyed. Analyses of HIV 454 data elucidated information which can help researchers to develop genomic-based drugs for HIV-infected people and also find SNPs in the HIV population causing drug resistance (Gifford, Liu et al. 2009). Data analysis through NGS is also cost effective, high throughput sequencing, it's accurate and fast and also more advanced than previous sequencing methods (Sanger methods). The NGS software & servers can analyse large amount of HIV envelope protein data to get the frequency of different parts of envelope sequences of HIV.

The main aim of our research is to get the diversity & different expression patterns of HIV samples, cause drug resistance in the HIV envelope protein population (Shafer 2006). Targeting on those different frequencies and variability of sequences HIV envelope protein population which plays an important role in accumulating different patterns and expressions of the HIV envelope proteins which are worldwide problem that affects a large amount of population through SRA file sequences and SNP analysis.

Methods and methodology:-

2.1 TRIMMING OF HIV SEQUENCE (Sequence Read Archive) THROUGH PERL SCRIPT

HIV envelope protein SRA file (about 100 sequences) was trimmed through Perl Scripting considering it for further NGS analysis. It removed the unwanted nucleotides from the HIV envelope protein to focus on the important targeted region of the sequence. The Perl package was installed developed and the Perl programming script was implemented to trim the sequences. Sequences were trimmed from both beginning and end using the script. Each nucleotide sequence was processed in BLAST basic local alignment search tool (BLAST) to find its similar sequences (Dikhit et al. 2013, Chandra Sahoo, Ansari et al. 2014, Ansari, Equbal et al. 2016). This finds the portion of local identities between the sequences. The program compares nucleotides SRA sequences to sequences of all databases and calculates the statistical importance of matches. From NCBI query the unwanted nucleotide (no identical sequence) range was removed or trimmed by the Perl programming script (Larkin, Blackshields et al. 2007, Dikhit, Moharana et al. 2014). The length of a sequence is 260 and query range is from 5 to 240 then length from 0 to 4 from beginning and length 20 is removed from 240 to the last value of sequence length which is 260. The program has been written & developed in Perl programming scrip. Here the script developed was for removing a substring at given position and length. Save the BLAST output files (in text format) in the same directory where the script is to the destined folder through the command line (type Perl blast.pl) (Little, Holte et al. 2002).

2.2 Phylogenetic Analysis of Trimmed Sequences with Other Species (Viral, Bacterial, Fungal & Protozoan Sequences)

Crustal X is a windows base interface for multiple sequence alignment (MSA) programs (Sahoo, Rani et al. 2009, Kar, Suryadevara et al. 2013, Kar, Ansari et al. 2013, Dikhit, Mahantesh et al. 2018). It provides a desegregated environment for analysing the data and performing profile-profile sequence alignments well as multiple sequence alignments profile. Each of trimmed important sequence was taken after doing clustering of each 50 sequences out of around 500 sequences. There are two groups formed for making the cluster of each 50 sequences. From these two sub-group, there are 2 or 3 sequences of one sub-cluster were taken as good or important sequences based on its similarities. Further these sequences were considered for clustering along with other viral sequences (like *Herpes Simplex,* Epstein Bar virus, Hepatitis virus A, E, Human Papillomavirus, Mollusc Contagious virus), bacterial/STD sequences (like *Neustria gonorrhoea, Treponema palladium*, *Homophiles Dicey*, *Chlamydia Trachomatis*), fungal/STD (*Tineacruris Candidiasis*), protozoal (*Trichomoniasis*) and different HIV strains from different countries (includes India, USA, Malawi, South Africa, Ethiopia, Bangladesh, China). They were taken to observe the diversity and clustering of all the selected trimmed sequences. Among these sequences, all retrovirus (HIV) makes a sub-cluster with other viral, bacterial and other sequences to get phylogenetic trees in Tree View program. The Tree View is a simple and easy program for displaying phylogenies which run on the different platform (Windows PCs, Apple, and Macintosh). Here's the text file (.txt) converts into denying file (.did) in the Crustal X program. This .dnd file was taken in Tree View by derived phylogenetic trees and find out the clustering and diversity pattern of these sequences (Sahoo, Dikhit et al. 2009, Sahoo, Rani et al. 2009, Rani, Nischal et al. 2013, Sahoo, Dikhit et al. 2013, Kumar, Das et al. 2014, Sahoo, Rani et al. 2014, Ansari, Equbal et al. 2016).

2.3 Comparison of HIV-1 (SRA) Trimmed Sequences (Envelop) with the Reference HIV Drug Resistance Sequence to Find SNPs Frequencies through NGS Software (NextGENe)

NextGENe Software is the tool for analysing the data from the Next Generation DNA sequences (Larkin, Blackshields et al. 2007, Sivakumaran, Husami et al. 2013). This software package contributed significantly for accurate and easy identification of SNPs at different frequency levels and various mutational changes occurring from the genome. While operating NextGENe software, it goes three steps that are condensation, sequence assembly, and sequence alignment. In condensation, there is parsing sequencing reads into the shorter keywords to the correct bases. Whereas in Sequence Alignment, the short reads sequences match read to a reference sequence. In sequence assembly, it is used to polish and lengthen the short sequence reads into fragments of manageable size and improved accuracy. NextGENe software was running on protocol given in the software package. The sample data, taken was HIV-1 (SRA) and it was analysed and compared with the reference sample taken as HIV drug resistance fasta file. In the beginning (SRA_data.fasta) was downloaded from NCBI, open NextGENe run wizard, Select application type: for application type: select Illumina, Transcriptome, Sequence condensation, Sequence assembly and Sequence alignment, click next, click load, three files from SRA_data.fasta were loaded, click load for reference file, one file was taken from SRA_data.fasta, click format conversion. FASTQ was selected, click add, one file was loaded, click default setting, click ok. Select condensation: click inspects input files, click next, click default setting, click finish, select NextGENe, ok, NextGENe viewer opens. The NextGENe Sequence Alignment matches a short sequence reads to a reference sequence. The reference sequence can be a small genome or genomic region or it can be a whole large genome reference such as the human, mouse, or rat genome.

2.4 SNP Detection by Self-Made Program

Self-Made Program was developed on php platform and was implemented to find diversities and mutational changes occurring in the different HIV samples globally. The different HIV envelope protein samples having different strains from the different countries around the globe was processed and made it run on a self-made program to find the ratio of nucleotides available in transversion out of total nucleotides and the number of nucleotides available in transition out of the total number of nucleotides, how many silent mutations are there and how much total mutation are taking place in the whole envelope protein. Also, we got an amino acid mismatch taking place in the sample HIV envelope protein against its reference envelope protein. From the given table, a bar graph was made having the sequence number showing bar having a ratio of Transition vs Transversion bar graph showing silent mutation was made too in an excel sheet.

2.5 Molecular Docking

The docking was performed according to our previously described methodology with certain modifications (Sahoo, Dikhit et al. 2009, Kumar Jayaswal, Rani et al. 2010, Sahoo, Basu et al. 2013, Sahoo, Dikhit et al. 2013, Anwar,

Dikhit et al. 2014, Sahoo, Yousuf Ansari et al. 2014, Purkait, Singh et al. 2015). For docking, the downloaded 3D structure of the 2NY1 protein in PDB selected from the HIV envelope protein (Wang, Xiao et al. 2009). Its inhibitor BMS-378806 was also taken from PubChem and their docking was processed in Discovery Studios (Ansari, Dikhit et al. 2012). Discovery Studio (DS) is a commercial molecular modeling program for biological macromolecules (proteins, nucleic acid). In PDB, envelope protein of HIV-1 executed and then click over 26 structure hit. The X-Ray resolution and the PDB file of protein 2NY1 was downloaded (Abhishek, Sardar et al. 2017). The inhibitor BMS-378806 was taken from Pubchem and saved as SDF file [26]. The protein was prepared in DS. Further, 'Define and Edit binding site' was executed and from receptor cavities, 10 binding sites were detected. Docking was processed for each site (Dikhit, Purkait et al. 2016, Mansuri, Ansari et al. 2016). Ligand and 'Input protein' and was docked (Kar, Ansari et al. 2013, Mansuri, Ansari et al. 2016).

Result and Discussion:-

3.1 Trimming of HIV-1 (SRA) Sequences by Using a Perl Script

HIV envelope proteins had huge amount of data which was making it difficult to analyse them together at once. Therefore the trimming of the sequences were crucial, so that we could focus on the only important targeted part of the sequences and remove the unwanted part of the sequences. The HIV-1 (SRA) sequences were trimmed after doing BLAST; we got the wide range of NCBI query. After looking at the query of the nucleotide range, we were able to remove the unwanted range of nucleotide by implementing the Perl programming script which we had developed. Hence, we got the trimmed length of the targeted sequence. Suppose the length of a sequence (SRA-SRR002680.1.1) is 253 and query range is from 5 to 253 then length from 0 to 4 from beginning and length 22 is removed from 231 to the last value of sequence length which is 253.

Format having rows and columns. Through these values, we get clear ideas about the composition of each sequence that how much similar or identical they actually are to each other, and in which dimension they are different from each other.

3.2 Sequence Analysis through Clustering and Phylogenetic Analysis with Other Relevant Sequences.

Phylogenetic tree was constructed from the HIV trimmed sequences and (shown in figure 1)other viral sequences like (*Herpes Simplex, Epstein_Bar virus, Hepatitis virus A, E,* Human *Papiloma* virus, *Molluscum Contagisum* virus), bacterial/STD sequences (like *Hemophilus ducreyi, Chlamydia trachomatis, Neisseria gonorrhea, Treponema pallidum*), fungal/STD like (*Tineacruris, Candidiasis* and protozoal like *Trichomoneisis)* and also different HIV strains from countries such as Malawi, South Africa, Ethopia, Bangladesh, China, India and USA were taken to observe their evolutionary relationship and diversities with each other. We inferred from this phylogeny analysis that the retrovirus population is present in the form of clusters with other viral, bacterial, fungal and protozoan sequences at their different accumulation positions globally. In these clusters there are accumulations of different species sequences with retroviral sequences according to their closeness, similarities, and distant similarities with each other in every different cluster. In this Phylogenetic analysis of all the sequences of the HIV envelope protein clustered together in one branch of the Phylogenetic tree, the accumulated retroviral sequences of HIV strains from USA, were also closer to other viral sequences which includes *Herpes virus, Hepatitis A, E, Epstein Bar virus, Molluscum contagious virus and Human Papillomavirus*. While in other cluster there are some retroviral sequences accumulated together making a cluster in position. Mutation is likely to occur in all the strains from different countries around the globe. This cluster is similar to the sequences like is *Hepatitis, Human Papillomavirus, Trichomonas vaginitis and Hemophile* Ducreyi in a cluster. Among these sequences; (Herpes virus is similar to *Mollscum contagious virus) and closer to Hepatitis A, E, Epstein Bar virus, Mollscum contagious virus and Human Papillomavirus. Trichomonas vaginitis*is similar to *Human Papilloma virus* and are closer to *Hemophile* Ducreyi. It was noticed that *Chlamydia* is similar to *Klebsiella* but they are not similar to *Neisseria* and *Homo sapiencapase* 12. All the viral sequences making cluster together among which all the HIV sequences sub clustered with each other. Thus, there is other sub clustered group of HIV sequences from different countries like Malawai, Ethopia, China, South Africa, Iran are clustering together and are closer to each other in this group. Retroviral sequence from Malawi and Ethopia is similar to each other.

3.3 Comparison of HIV-1 (SRA) Trimmed Sequences with the Reference HIV Drug Resistance to Find SNPs Frequencies through NextGENe:

The comparison of HIV-1 (SRA) trimmed sequence is to HIV reference has suggested in figure 2a and 2b, there is the two HIV read coverage map of raw HIV-1 (SRA) respectively from NextGENe software generated the file. In this graph, the blue ticks identify locations of novel SNPs. The peaks tell us the variation of sequences. The large

peak has shown more variation in the sequences & is more low peak has shown fewer variations in the sequences of HIV envelope protein of HIV-1. The peak where there is a darker region shows a region of more SNPs. The peak position between the nucleotide ranges from 84k to 98K is showing the highest frequency of SNPs. The Coverage Curve report displays the coverage distribution of samples reads along the reference sequence without directional information. Reference sequence (HIV-1) regions that are highlighted in red indicate regions where the coverage falls below the user-set mutation filter coverage threshold (shown in the figure 2a & 2b). It shows the number of reads aligned at the SNP location. In the coverage curve map shows the maximum coverage range of nucleotides at range from 80000-100000 & 180000-200000 against the tag number between the ranges of 60-65 on the y-axis. NextGENe produces a chart with the sequence tag number on the x-axis and coverage of each tag on the y-axis. The grey and pink peaks from both coverage maps tell us about the SNPs. The more SNPs have occurred where there is the highest peak & it is wise verse of the peaks.

3.4 Detection by Self-Made Program

Self-Made program was developed on php platform and was implemented to find diversities and mutational changes occurring in the different HIV samples globally. The different HIV envelope protein samples having different strains from the different countries around the globe was processed and made it run on a self-made program to find the ratio of nucleotides available in transversion out of total nucleotides and the number of nucleotides available in transition out of the total number of nucleotides, how many silent mutations are there and how much total mutation are taking place in the whole envelope protein. Also, we got an amino acid mismatch taking place in the sample HIV envelope protein against its reference envelope protein. From the given table, a bar graph was made having the sequence number showing bar having a ratio of Transition vs Tran version bar graph showing silent mutation was made too in an excel sheet.

3.5 Molecular Docking:

Computer based molecular docking was used to estimate the most probable binding geometries for large libraries of the target molecules, if the protein structure is available (Rana, Dikhit et al. 2012, Mansuri, Kumar et al. 2017, Dikhit, Ansari et al. 2018). Through docking we want to know binding ability of the proteins, their most potential ligands, their binding affinity. In molecular docking various biological processes related with proteins are regulated or enabled by specific binding of small organic molecules (ligands) to the proteins (Rani, Dikhit et al. 2011, Chauhan, Ansari et al. 2017, Kumar, Rana et al. 2017). Large number of drugs is known to work by binding a target protein. Docking helps us to know about the molecules which could bind to the active site, and about the sites where active molecules bind. Docking studies have revealed the interaction studies between ligand proteins of total eleven conformers were formed. Docking helped us to derive a table having LibScore_Dreiding, LibScore_Dreiding, -PLP1, -PLP2 and Dock_Score. The table shows the dock score of all 11 conformers of the ligand BMS-378806 which are 3.014, 23.282, 23.099, 20.548, 20.312, 19.545, 16.908, 16.141, 14.667, 12.126, & 12.078. Amino acid residues taking part in this ASN-A:377, ASN-A:262, GLU-A:211, ARG-A:252, SER-A:447, CYS-A:445, SER-A: 446; Vander Waals taking part in this interaction is PHE-A:210, PRO-A:212, VAL-A:254, PHE-A:376, CYS-A:378, VAL-A:255, SER-A:256, LEU-A261 as shown in figure 5 and Table 2.

Discussion:-

The main objective of this is looking at the severity of HIV-1 infection globally and causing deaths to the mass of people among women, children and all age group of people made us focus on the problems caused by HIV infection. So the next generation sequencing (NGS) data was taken for analysis because there are huge data for analysis. With the help of Perl scripting, we can get the important length of sequences that have maximum similarities and unwanted sequences were removed automatically. The table formed with the help of Perl scripting from the BLAST NCBI server. It gave us the accurate & detailed information about its file name, hit found, length, score, E-score, identities and gaps. Further with those trimmed sequences phylogenetic analysis was done with the diverse HIV strains taken from different countries along with other viral, bacterial and fungal sequences. Phylogenetic analysis gave us the clear idea of its diversity and its similarities and differences with other sequences. It also gave us the clear ideas about which sequences have drug-resistant from the study of evolutionary relationships. Phylogenetic analysis gave us the information of diversity of HIV sequences with other sequences of virus, bacteria and fungus along with different HIV strains from different countries (Malawi, China, South Africa, Bangladesh, Iran, USA, India, and Ethiopia). Phylogenetic analysis, makes us known that diversities of bacterial or viral sequences. Further the phylogenetic tree tells us that the *Herpes virus* and *Molluscum* are closer to each other than viruses *(Hepatitis E* then to *Kaposi's* then to *Epstein bar* and then to *Treponema* or *Trichophyton*). It is also suggested that the *Herpes virus* is most closely related to *Hepatitis E* than *Molluscan* then to *Kaposi'* and then to *Herpes* and *Treponema*.

Herpes and *Epstein* are closely related to each other than *Molluscum* then to *Treponema* or *Trichophyton.* Ahead of this the use of some NGS software and its generated data from the raw HIV-1 envelope proteins was the critical part of the research elucidating coverage map and the coverage curve map showing the SNPs peaks at the various frequency levels of the sequences telling us about the variety of sequences. Analysis of HIV 454 data by NextGENe has elucidated the frequency of different parts of envelope sequences of HIV. The HIV reads coverage map shows us the peaks of SNPs which tell us about the variety of sequences at different regions at various frequency levels. To find out, the SNPs in the HIV population has made with different expression patterns. SNPs gave us information about the genetic variations and mutations taking place at various frequencies of the sequence to find out the drug resistant towards the HIV infection. SNPs could provide help as markers which we can predict and inform our decisions about numerous aspects of medical care, including specific diseases, adverse reactions and effectiveness of various drugs and to specific drugs. SNPs are also critical for the disease detection. The distribution of reads conveyed about the region of the HIV-1 envelope population where and at what ranges the variation of sequences have happened causing drug resistance among the population.

In the self-made program, we took envelope HIV-1 protein from different countries and got the idea about its transition, trans-version, and ratio of TS vs TV, silent mutations and total mutations. This elucidated about amino acid substitutions, mutations and variation taking place in retroviral population globally. Through docking, it revealed about the docking score between the protein 2NY1 and inhibitor BMS-378806 which was done to predict the strength of binding ability between the protein and inhibitors. All this together revealed the interaction taking place and the participants involved in the interaction so that we could deduce about the regions where its variations can lead to get information about the action which is taking place causing drug resistance among the retrovirus population causing drug resistance globally (Van Dyk 2010).

Conclusion:-

HIV infection is the serious, deadly infection causing AIDS, which is declared as a pandemic disease affecting the worldwide across the globe. To encounter it, analysis of (HIV-1) 454 race of envelope protein was very critical. By using various software and programs, analysis of HIV 454 data has elucidated the frequency of different parts of envelope sequences of HIV. The phylogenetic analysis has deduced the evolutionary origin of HIV trimmed sequences of the different strain from different countries with other viral sequences of other STDs, bacterial STD sequences, fungal sequences and protozoan sequences have revealed that different amino acid substitutions have accumulated over years and have shown clustering with other HIV sequences. SNP analyses have shown grey picks density that conveyed about the variations which have occurred in different regions at various frequency levels as per the requirement of the HIV genome for replication. Analysis of the short read sequence of HIV generated from 454 (Roche) by next generation sequencing technology is very handy to collect the information from the distribution of different expression patterns of different HIV samples globally and also find SNPs in the HIV population causing drug resistance. SMP helped to get mutational changes occurred, tran-version vs transition, silent mutations, and total mutations occurred. A highest docking score between its protein and inhibitor revealed the binding strength between them. This analysis is also useful for developing genomic-based drugs, reducing the complexity of the pandemic disease like AIDS.

Conflict of interest:

All authors found no conflict of interest

Acknowledgements:-

This study was supported by a grant for setting up biomedical informatics centre from Indian Council of Medical Research (ICMR), Govt. of India.

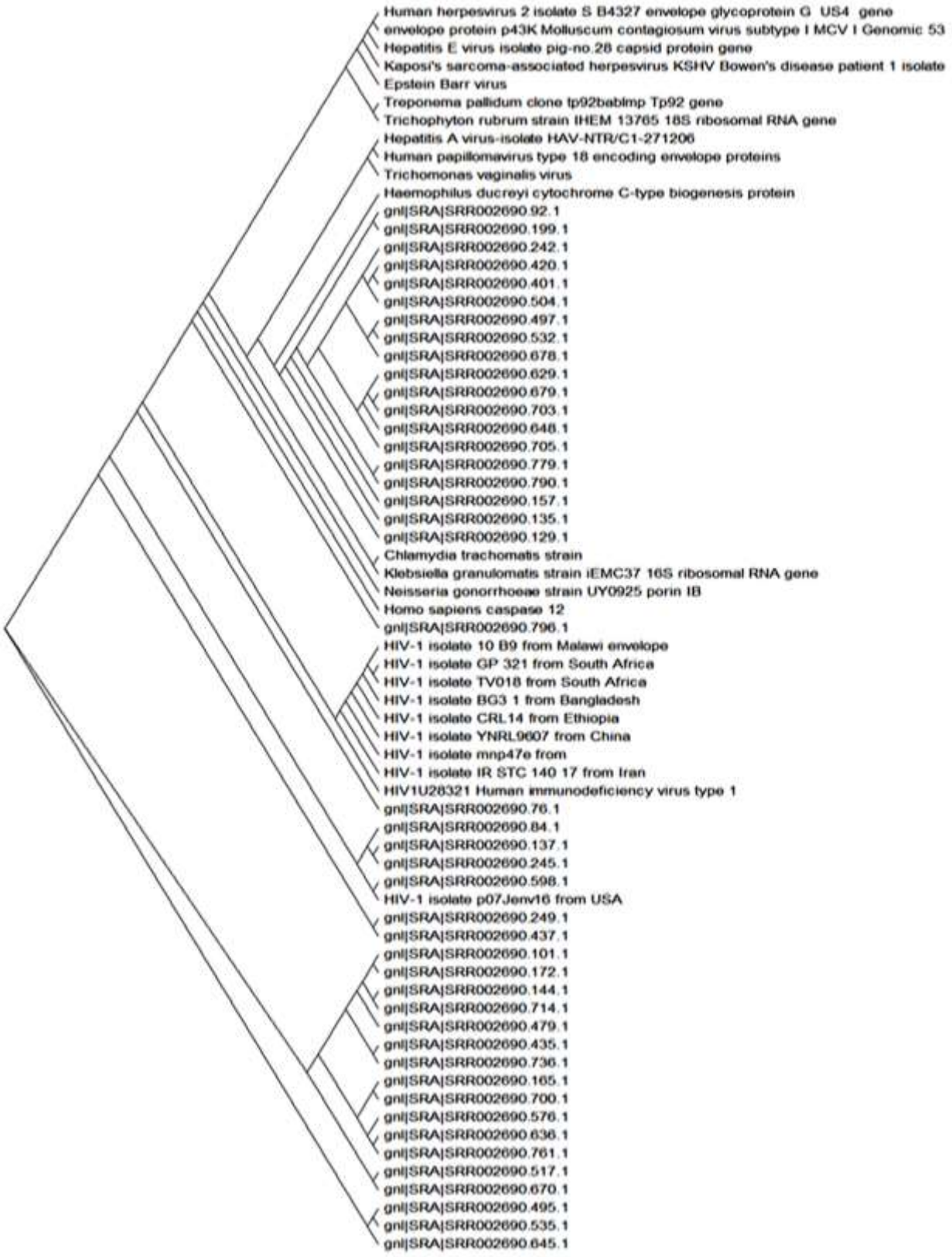

FIG. 1 Phylogenetic tree was constructed for the analysis of evolutionary relation of HIV sequences with other sequences. The trimmed sequences were selected by doing clustering of each 50 sequences out of around 1000

sequences which were trimmed. Making cluster of each 50 sequences we get two groups of cluster in clustalx. From these two groups of cluster 2 or 3 sequences were taken as good or important sequences according its similarities. Then selected trimmed sequences along with other viral sequences (like *Herpes Simplex, Epstein_Bar virus, Hepatitis virus A, E, Human Papiloma virus, Molluscum Contagisum virus)*, bacterial/STD sequences (like *Hemophilus ducreyi, Chlamydis trachomatis, Neisseria gonorrhea, Treponema pallidum*), fungal/STD (like *Tineacruris, Candidiasis* and protozoal (like Trichomoneisis) were processed in Crustal X. It gave us its .dnd file which was loaded in Tree View which constructed Phylogenetic tree. We get more better and convenient diversity of trees and also we can get information about retroviral sequences and it's closeness and similarities. According to the distances of trees and it's taxas and clades we can get good knowledge about their diversity and closeness of similarities and distance similarities. The similarities of some of the HIV (SRA) were more close to other viral sequences like *Herpes virus, Hepatitis A, E, Eptein_Bar virus, Human Papilloma virus and Molluscumcontagisum virus*, while some of the bacterial and fungal sequences were not that close to HIV trimmed sequences as other viral sequences were to trimmed sequences but they showed close relatedness in similarities with other HIV (SRA) trimmed sequences.

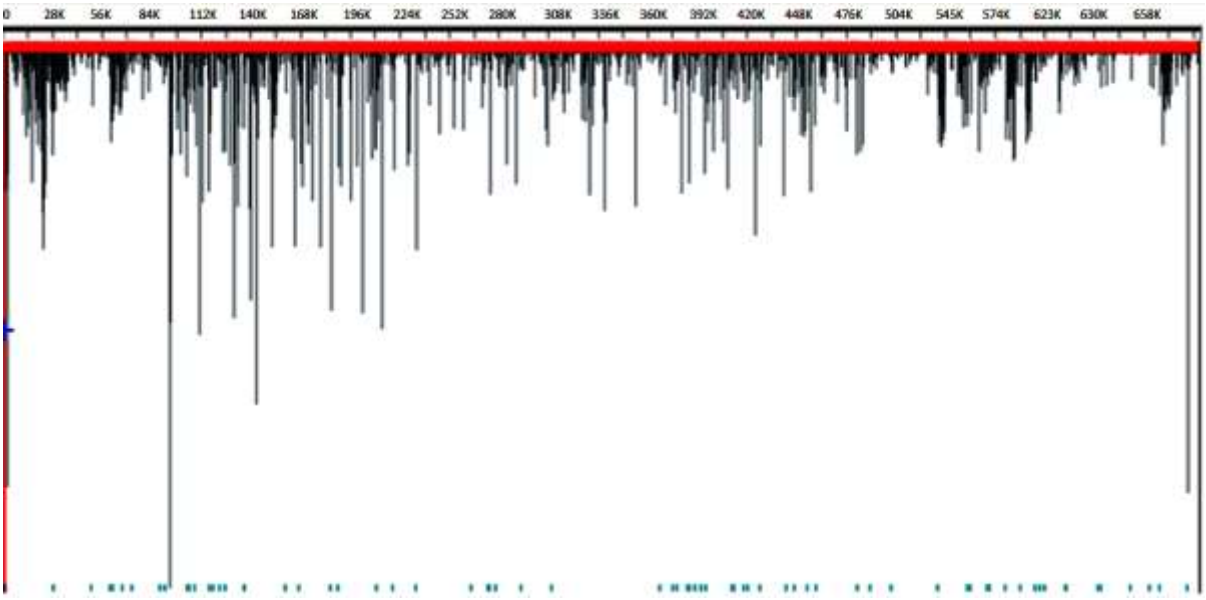

FIG 2AShows the blue ticks which identify locations of novel SNPs. The peaks tell us the variation of sequences. The peak where it is large it shows the more variations of sequences are more and peak where it is low it shows less variations of sequences. The peaks where there is darker shows region of more SNPs.

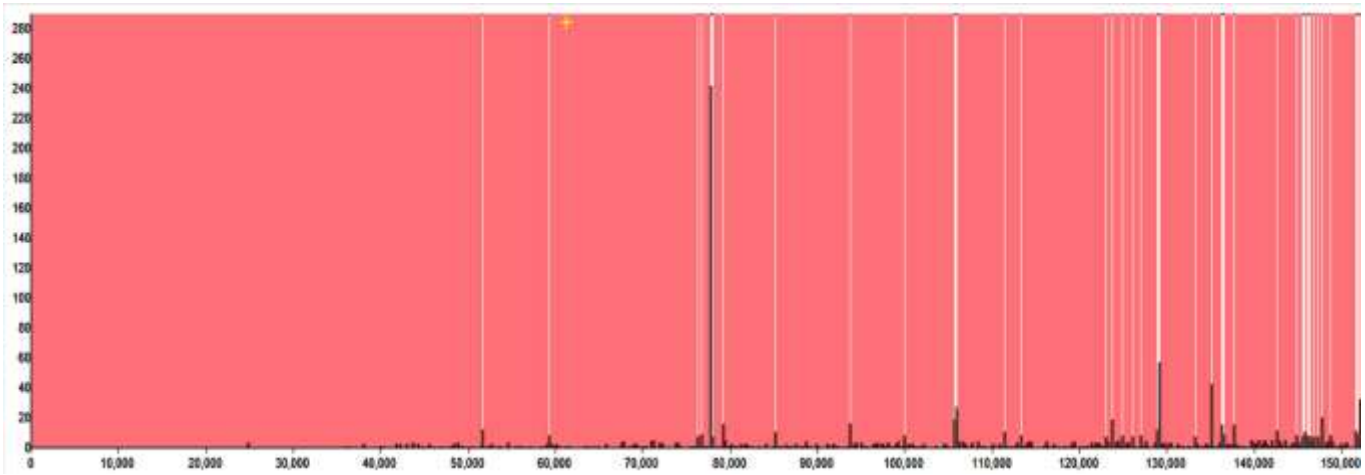

FIG. 2B In the Coverage Curve report displays the coverage distribution of samples reads along the reference sequence without directional information. It shows the number of reads aligned at the SNP location. Next Gene produces a chart with the sequence tag number on the x-axis and coverage of each tag on the y-axis. The grey and pink peaks from both coverage map tell us about the SNPs .The peak is highest then there are more SNPs detected and it is wise versa of the peaks.

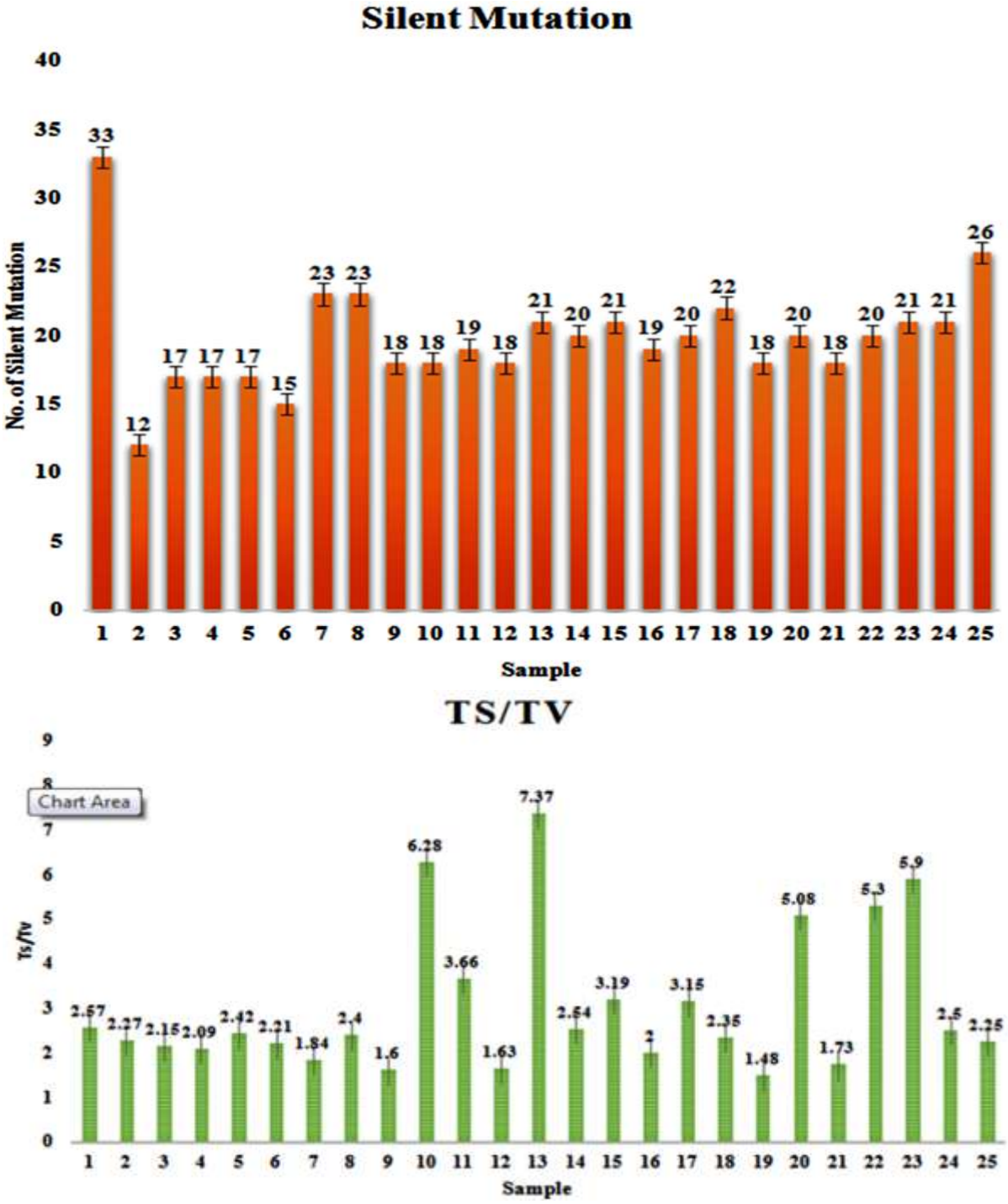

FIG 3A and 3B Showing bar graph which was made in excel sheet showing ratio of Ts vs TV, Silent mutation. In the Ts vs TV graph the sequence number 13 is showing the maximum TV ratio value of 1.48. For bar diagram of silent

mutation sequence number 25 is showing the maximum silent mutation of sequence number 1 showing maximum mutation of 33 in number.

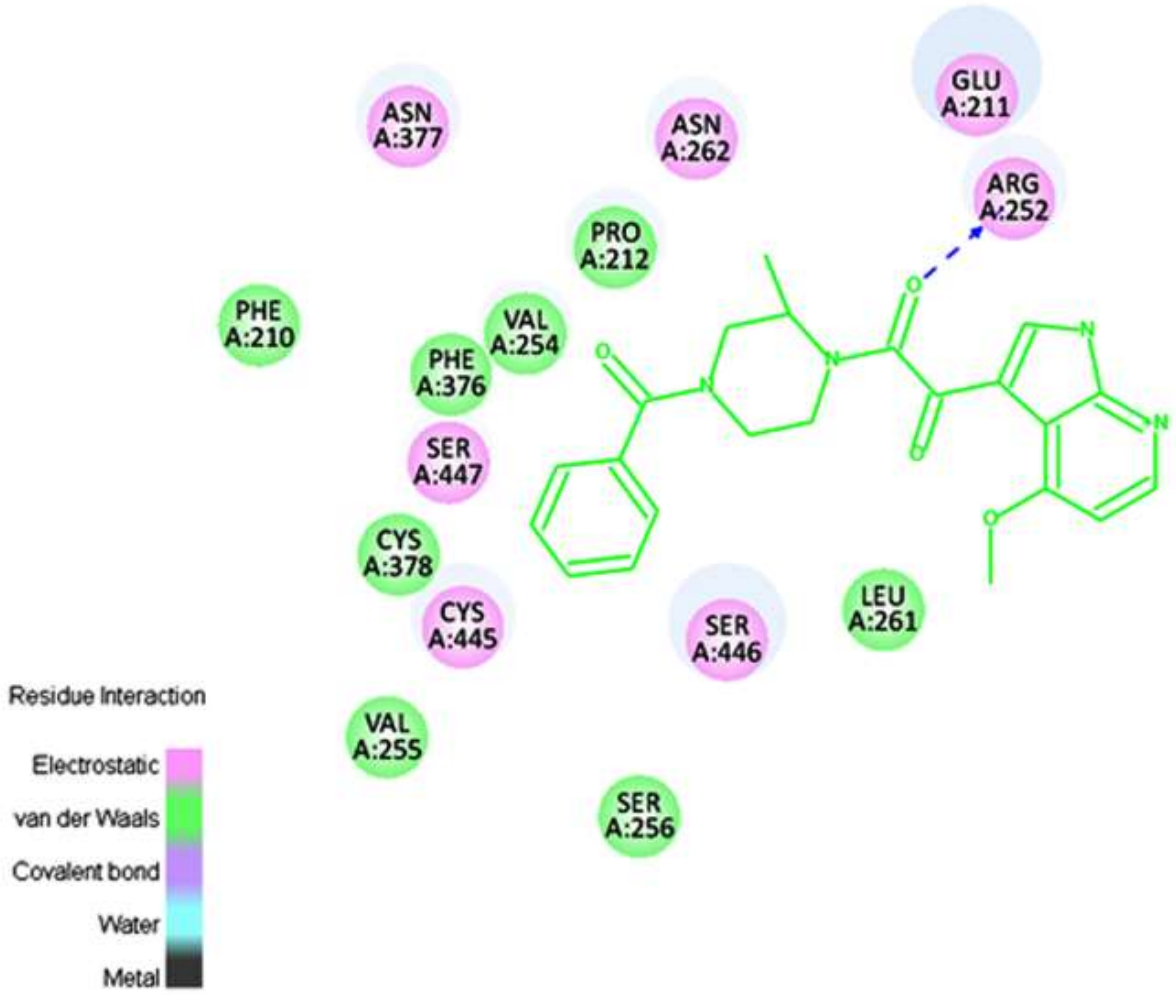

Fig 4 Amino acid residues taking part in this ASN-A: 377, ASN-A: 262, GLU-A: 211, ARG-A: 252, SER-A: 447, CYS-A: 445, SER-A: 446. Vander Waals taking part in this interaction is PHE-A: 210, PRO-A: 212, VAL-A: 254, PHE-A: 376, CYS-A: 378, VAL-A: 255, SER-A: 256, LEU-A261

Table No 1

Serial No	Trans version(TV)	Transition(Ts)	Ts/TV	Silent Mutation	Total Mutation	Total Nucleotide
1	7	18	18/7=2.57	33	58	2583
2	11	25	25/11=2.27	12	48	2583
3	19	41	41/19=2.15	17	80	2583
4	22	46	46/22=2.09	17	82	2583
5	19	46	46/19=2.42	17	82	2583
6	19	42	42/19=2.21	15	83	2583
7	26	48	48/26=1.84	23	91	2583
8	20	48	48/20=2.40	23	91	2583
9	20	33	33/20=1.60	18	58	2583
10	7	44	44/7=6.28	18	77	2583
11	15	55	55/15=3.66	19	96	2583

12	22	36	36/22=1.63	18	62	2583
13	8	59	59/8=7.37	21	104	2583
14	24	61	61/24=2.54	20	102	2583
15	21	67	67/21=3.19	21`	116	2583
16	28	56	56/26=2.00	19	95	2583
17	20	63	63/20=3.15	20	111	2583
18	28	66	66/28=2.35	22	117	2583
19	29	43	43/29=1.48	18	73	2583
20	12	61	61/12=5.08	20	102	2583
21	21	37	37/21=1.78	18	65	2583
22	10	53	53/10=5.3	20	89	2583
23	16	59	59/10=5.9	21	104	2583
24	24	60	60/24=2.50	21	105	2583
25	24	54	54/24=2.25	26	100	2583

Table no 1. showing mutational changes between HIV sample sequences from different countries in the SMP (self made program), where submission of the sequences from the different countries were submitted and all the mutational changes was noticed in comparison with each other. It gave information about Trans-version, Transition, Trans-version vs. Transition, silent mutations and total mutations.

Table No 2

LibScore_Dreiding	LibScore_Dreiding	-PLP1	-PLP2	Dock Score
0.83	0.17	75.83	78.27	3.014
3.59	2.54	89.25	88.53	23.282
2.95	1.76	86.65	88.47	23.099
2.78	1.53	88.60	90.23	20.548
3.03	2.03	86.13	90.37	20.312
3.03	1.85	85.32	89.49	19.545
3.04	1.86	89.15	89.29	16.908
2.67	1.58	87.27	86.47	16.141
2.83	1.59	85.02	83.47	14.667
3.14	2.31	89.88	91.23	12.126
3.75	2.82	92.13	92.54	12.078

Table no 2. Showing docking table having LibScore_Dreiding, LibScore_Dreiding, -PLP1, -PLP2 and Dock Score. The table is showing the dock score of all 11 confirmers of the ligands BMS-378806 which are 3.014, 23.282, 23.099, 20.548, 20.312, 19.545, 16.908, 16.141, 14.667, 12.126, & 12.078

Reference:-

1. Abhishek, K., A. H. Sardar, S. Das, A. Kumar, A. K. Ghosh, R. Singh, S. Saini, A. Mandal, S. Verma and A. Kumar (2017). "Phosphorylation of translation initiation factor 2-alpha in Leishmania donovani under stress is necessary for parasite survival." Molecular and cellular biology **37**(1): e00344-00316.
2. Ansari, M. Y., M. R. Dikhit, G. C. Sahoo and P. Das (2012). "Comparative modeling of HGPRT enzyme of L. donovani and binding affinities of different analogs of GMP." International journal of biological macromolecules **50**(3): 637-649.
3. Ansari, M. Y., A. Equbal, M. R. Dikhit, R. Mansuri, S. Rana, V. Ali, G. C. Sahoo and P. Das (2016). "Establishment of correlation between in-silico and in-vitro test analysis against Leishmania HGPRT to inhibitors." International journal of biological macromolecules **83**: 78-96.
4. Anwar, S., M. R. Dikhit, K. P. Singh, R. K. Kar, A. Zaidi, G. C. Sahoo, A. K. Roy, T. Nozaki, P. Das and V. Ali (2014). "Interaction between Nbp35 and Cfd1 proteins of cytosolic Fe-S cluster assembly reveals a stable complex formation in Entamoeba histolytica." PloS one **9**(10): e108971.
5. Chandra Sahoo, G., Y. Ansari, S. Rana, M. Ranjan Dikhit, R. Kamal Topno, K. Pandey and P. Das (2014). "Molecular Modeling and Ligand-Protein Interaction of N-Protein of Chandipura Virus." Letters in Drug Design & Discovery **11**(2): 211-221.
6. Chauhan, A. S., M. Y. Ansari, R. Mansuri, M. R. Dikhit, V. Ali, G. C. Sahoo and P. Das (2017). "Computational elucidation, mutational and hot spot-based designing of potential inhibitors against human acid-

sensing ion channels (hASIC-1a) to treat various physiological conditions." Journal of Biomolecular Structure and Dynamics: 1-18.
7. Checkley, M. A., B. G. Luttge and E. O. Freed (2011). "HIV-1 envelope glycoprotein biosynthesis, trafficking, and incorporation." Journal of molecular biology **410**(4): 582-608.
8. Dikhit, M. R., A. Amit, A. K. Singh, A. Kumar, S. Sinha, R. K. Topno, R. Mishra, V. N. R. Das, K. Pandey and G. C. Sahoo (2017). "Vaccine potential of HLA A2 epitopes from Leishmania cysteine protease type III (CPC)." Parasite immunology.
9. Dikhit, M. R., M. Y. Ansari, V. Ali, R. K. Topno, J. P. Majhee, G. C. Sahoo and P. Das (2018). "Computational elucidation of novel antagonists and binding insights by structural and functional analyses of serine hydroxymethyltransferase and interaction with inhibitors." Gene Reports **10**: 17-25.
10. Dikhit, M. R., M. Y. Ansari, R. Mansuri, B. R. Sahoo, B. Dehury, A. Amit, R. K. Topno, G. C. Sahoo, V. Ali and S. Bimal (2016). "Computational prediction and analysis of potential antigenic CTL epitopes in Zika virus: A first step towards vaccine development." Infection, Genetics and Evolution **45**: 187-197.
11. Dikhit, M. R., A. Kumar, S. Das, B. Dehury, A. K. Rout, F. Jamal, G. C. Sahoo, R. K. Topno, K. Pandey and V. Das (2017). "Identification of potential MHC Class II-restricted epitopes derived from Leishmania donovani antigens by reverse vaccinology and evaluation of their CD4+ T-Cell responsiveness against visceral leishmaniasis." Frontiers in immunology **8**: 1763.
12. Dikhit, M. R., V. Mahantesh, A. Kumar, A. Amit, B. Dehury, M. Ansari, V. Ali, V. Das, G. C. Sahoo and S. Bimal (2018). "Mining the proteome of Leishmania donovani for the development of novel MHC class I restricted epitope for the control of visceral leishmaniasis." Journal of cellular biochemistry.
13. Dikhit, M. R., K. C. Moharana, B. R. Sahoo, G. C. Sahoo and P. Das (2014). "LeishMicrosatDB: open source database of repeat sequences detected in six fully sequenced Leishmania genomes." Database **2014**.
14. Dikhit, M. R., B. Purkait, R. Singh, B. R. Sahoo, A. Kumar, R. K. Kar, M. Y. Ansari, S. Saini, K. Abhishek and G. C. Sahoo (2016). "activity of a novel sulfonamide compound 2-nitro-N-(pyridin-2-ylmethyl) benzenesulfonamide against Leishmania donovani." Drug design, development and therapy **10**: 1753.
15. Gifford, R. J., T. F. Liu, S.-Y. Rhee, M. Kiuchi, S. Hue, D. Pillay and R. W. Shafer (2009). "The calibrated population resistance tool: standardized genotypic estimation of transmitted HIV-1 drug resistance." Bioinformatics **25**(9): 1197-1198.
16. Girard, M. P., S. K. Osmanov and M. P. Kieny (2006). "A review of vaccine research and development: the human immunodeficiency virus (HIV)." Vaccine **24**(19): 4062-4081.
17. Gobind, J. (2014). Evaluation of an HIV/AIDS prevention programme at a South African university, University of Johannesburg.
18. Gyorkey, F., J. L. Melnick and P. Gyorkey (1987). "Human immunodeficiency virus in brain biopsies of patients with AIDS and progressive encephalopathy." Journal of Infectious Diseases **155**(5): 870-876.
19. Hawkins, D., M. Blott, P. Clayden, A. De Ruiter, G. Foster, C. Gilling-Smith, B. Gosrani, H. Lyall, D. Mercey and M. L. Newell (2005). "Guidelines for the management of HIV infection in pregnant women and the prevention of mother-to-child transmission of HIV." HIV medicine **6**(S2): 107-148.
20. Kar, R., P. Suryadevara, B. Sahoo, G. Sahoo, M. Dikhit and P. Das (2013). "Exploring novel KDR inhibitors based on pharmaco-informatics methodology." SAR and QSAR in Environmental Research **24**(3): 215-234.
21. Kar, R. K., M. Ansari, P. Suryadevara, B. R. Sahoo, G. C. Sahoo, M. R. Dikhit and P. Das (2013). "Computational elucidation of structural basis for ligand binding with Leishmania donovani adenosine kinase." BioMed research international **2013**.
22. Kelly, J. A. and S. C. Kalichman (2002). "Behavioral research in HIV/AIDS primary and secondary prevention: Recent advances and future directions." Journal of consulting and clinical psychology **70**(3): 626.
23. Kumar, A., S. Das, B. Purkait, A. H. Sardar, A. K. Ghosh, M. R. Dikhit, K. Abhishek and P. Das (2014). "Ascorbate peroxidase, a key molecule regulating amphotericin B resistance in clinical isolates of Leishmania donovani." Antimicrobial agents and chemotherapy **58**(10): 6172-6184.
24. Kumar Jayaswal, P., M. Rani, C. Prakash Yadav, M. Ranjan Dikhit, G. C. Sahoo and P. Das (2010). "Molecular Modeling of Cathepsin B protein in different Leishmania strains." Journal of Integrated OMICS **1**(1): 115-123.
25. Kumar, M., S. Rana, H. Kumar, P. Kumar, M. R. Dikhit, R. Mansuri, J. Kumar and G. C. Sahoo (2017). "Computational, structural and functional aspects of hypothetical protein of Aspergillus flavus Pheromone Receptor Pre-A (PRP-A)." Journal of Applied Pharmaceutical Science Vol **7**(07): 089-097.
26. Larkin, M. A., G. Blackshields, N. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm and R. Lopez (2007). "Clustal W and Clustal X version 2.0." bioinformatics **23**(21): 2947-2948.

27. Little, S. J., S. Holte, J.-P. Routy, E. S. Daar, M. Markowitz, A. C. Collier, R. A. Koup, J. W. Mellors, E. Connick and B. Conway (2002). "Antiretroviral-drug resistance among patients recently infected with HIV." New England Journal of Medicine **347**(6): 385-394.
28. Mansuri, R., M. Y. Ansari, J. Singh, S. Rana, S. Sinha, G. C Sahoo, M. R Dikhit and P. Das (2016). "Computational elucidation of structural basis for ligand binding with mycobacterium tuberculosis glucose-1-phosphate thymidylyltransferase (RmlA)." Current pharmaceutical biotechnology **17**(12): 1089-1099.
29. Mansuri, R., A. Kumar, S. Rana, B. Panthi, M. Y. Ansari, S. Das, M. R. Dikhit, G. C. Sahoo and P. Das (2017). "In vitro evaluation of antileishmanial activity of computationally screened compounds against ascorbate peroxidase to combat amphotericin B drug resistance." Antimicrobial agents and chemotherapy **61**(7): e02429-02416.
30. Masetshaba, M. (2016). Experiences of long-term highly active antiretroviral treatment by adolescents in Tembisa, Gauteng Province.
31. Montagnier, L. (2002). "A history of HIV discovery." Science **298**(5599): 1727-1728.
32. Purkait, B., R. Singh, K. Wasnik, S. Das, A. Kumar, M. Paine, M. Dikhit, D. Singh, A. H. Sardar and A. K. Ghosh (2015). "Up-regulation of silent information regulator 2 (Sir2) is associated with amphotericin B resistance in clinical isolates of Leishmania donovani." Journal of Antimicrobial Chemotherapy **70**(5): 1343-1356.
33. Rana, S., M. R. Dikhit, M. Rani, K. C. Moharana, G. C. Sahoo and P. Das (2012). "CPDB: cysteine protease annotation database in Leishmania species." Integrative Biology **4**(11): 1351-1357.
34. Rani, M., M. R. Dikhit, G. Chandra, P. Das, P. K. Thakur, V. Tyagi, F. Ahmad and M. I. Hassan (2011). "Molecular modeling of the dimensional structure of EF-1α from Leishmania donovani and its interaction with several inhibitors." Journal of Natural Science, Biology and Medicine **2**(3): 74.
35. Rani, M., A. Nischal, G. C. Sahoo and S. Khattri (2013). "Computational analysis of the 3-D structure of human GPR87 protein: implications for structure-based drug design." Asian Pacific Journal of Cancer Prevention **14**(12): 7473-7482.
36. Sahoo, B. R., M. Basu, B. Swain, M. R. Dikhit, P. Jayasankar and M. Samanta (2013). "Elucidation of novel structural scaffold in rohu TLR2 and its binding site analysis with peptidoglycan, lipoteichoic acid and zymosan ligands, and downstream MyD88 adaptor protein." BioMed research international **2013**.
37. Sahoo, G. C., M. R. Dikhit, M. Rani, M. Y. Ansari, C. Jha, S. Rana and P. Das (2013). "Analysis of sequence, structure of GAPDH of Leishmania donovani and its interactions." Journal of Biomolecular Structure and Dynamics **31**(3): 258-275.
38. Sahoo, G. C., M. R. Dikhit, M. Rani and P. Das (2009). "Homology modeling and functional analysis of LPG2 protein of Leishmania strains." J Proteomics Bioinform **2**: 32-50.
39. Sahoo, G. C., M. Rani, M. Y. Ansari, C. Jha, S. Rana, M. R. Dikhit, K. C. Moharana, R. Kumar and P. Das (2014). "Structure, evolution and virtual screening of NDM-1 strain from Kolkata." International journal of bioinformatics research and applications **10**(3): 235-263.
40. Sahoo, G. C., M. Rani, M. R. Dikhit, W. A. Ansari and P. Das (2009). "Structural modeling, evolution and ligand interaction of KMP11 protein of different leishmania strains." J Comput Sci Syst Biol **2**(2): 147-158p.
41. Sahoo, G. C., M. Yousuf Ansari, M. R. Dikhit, M. Kannan, S. Rana and P. Das (2014). "Structure prediction of gBP21 protein of L. donovani and its molecular interaction." Journal of Biomolecular Structure and Dynamics **32**(5): 709-729.
42. Schwartländer, B., J. Stover, T. Hallett, R. Atun, C. Avila, E. Gouws, M. Bartos, P. D. Ghys, M. Opuni and D. Barr (2011). "Towards an improved investment approach for an effective response to HIV/AIDS." The Lancet **377**(9782): 2031-2041.
43. Shafer, R. W. (2006). "Rationale and Uses of a Public HIV Drug-Resistance Database." The Journal of infectious diseases **194**(Supplement_1): S51-S58.
44. Sivakumaran, T. A., A. Husami, D. Kissell, W. Zhang, M. Keddache, A. P. Black, B. T. Tinkle, J. H. Greinwald Jr and K. Zhang (2013). "Performance evaluation of the next-generation sequencing approach for molecular diagnosis of hereditary hearing loss." Otolaryngology–Head and Neck Surgery **148**(6): 1007-1016.
45. Stein, Z. A. (1990). "HIV prevention: the need for methods women can use." American Journal of Public Health **80**(4): 460-462.
46. Van Dyk, A. C. (2010). HIVAIDS care and counselling: a multidisciplinary approach, Pearson South Africa.
47. Wang, Y., J. Xiao, T. O. Suzek, J. Zhang, J. Wang and S. H. Bryant (2009). "PubChem: a public information system for analyzing bioactivities of small molecules." Nucleic acids research **37**(suppl_2): W623-W633.