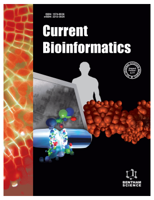
Current Bioinformatics, 2017, *12*, 303-311

RESEARCH ARTICLE

HIV-1 Nucleotide Sequence Comprehensive Analysis: A Computational Approach



José Irahe Kasprzykowski[1,2,3,*], Kiyoshi Ferreira Fukutani[2], Helton Fábio[2], Aldina Maria Prado Barral[2] and Artur Trancoso Lopo de Queiroz[1,2]

[1]Applied Computing Post Graduate Program, Feira de Santana State University, Feira de Santana-BA, Brasil; [2]Laboratory of Immunoparasitology, Gonçalo Moniz Institute, Oswaldo Cruz Foundation, Salvador-BA, Brasil; and [3]Biotechnology, Health and Investigative Medicine Post Graduate Program Gonçalo Moniz Institute, Oswaldo Cruz Foundation, Salvador-BA, Brasil

ARTICLE HISTORY

Received: March 11, 2016
Revised: September 29, 2016
Accepted: October 18, 2016

DOI: 10.2174/1574893611666161027142611

Abstract: ***Background:*** Acquired Immunodeficiency Syndrome (AIDS) is a large-scale pandemic caused by the infection of Human Immunodeficiency Virus (HIV). This virus infects over 40 million people worldwide. In the search for pandemic control, many drug resistance tests have been performed, resulting in the generation of large genomic data amount. These data are stored in biological databases, increasing on a daily basis. However, the majority of genomic data lacks important information, regarding virus subtype distribution, in the primary databases, e.g. GenBank.

Objective: A novel software tool to obtain, index and analyze highly mutational virus data, such as all HIV-1 sequence data from GenBank.

Method: The software aligns all sequences containing a complete genome (HXB2) for mapping purposes. In addition, all sequences with subtype references are locally aligned to classify all data into genotypic niches.

Results: Our results detail the prevalence of every subtype from a global HIV-1 sequence perspective, highlighting increases in the number of sequences related to recombinant subtypes. We were also able to identify country-based distribution of sequences according to geographical data distribution. All data were analyzed on a reasonable timescale, particularly in comparison to classic methods.

Conclusion: Our software represents an important contribution to HIV molecular epidemiology and offers a technique to rapidly classify new sequences, in addition to providing insight about sequence coverage density, subtype and country distribution. This data, together with cross-referencing, will aid in the generation of a novel, comprehensive and updated HIV-1 database.

Keywords: Comparative genomics, sequence subtyping, HIV, sequence mapping, bioinformatics, genomics, highly mutant virus.

1. INTRODUCTION

Viruses-associated diseases are a serious public health problem worldwide. Acquired Immunodeficiency Syndrome (AIDS) has no cure and affects people of almost all the countries [1]. Human Immunodeficiency Virus (HIV) is the associated etiological agent of AIDS and the World Health Organization (WHO) estimates that HIV infects over than 40 million people worldwide with 2.3 million new cases every year [2]. There are two types of HIV, however, the type 1 (HIV-1) is the major associated with AIDS pandemic [3-5]. The HIV-1 has four main groups; M, N, O and P [6]. Related to over 33 million infections around the world, strains from group M are the most prevalent in the pandemic [6]. Due the virus reproduction speed, recombination and the high mutation rate show high genetic variability with 9 subtypes and 61 recombinant, corresponding to 70 variants [3, 6]. This genetic variability shows notorious influence in the disease progression, because different genotypes could present different cellular tropism, viral replication and antiretroviral drug susceptibility [7, 8].

During the viral life cycle, several viral proteins interact with the host proteins [9]. These proteins are encoded by two major gene sets; structural genes, such as *gag*, *pol* and *env*

[*] Address correspondence to this author at the Laboratory of Immunoparasitology, Gonçalo Moniz Institute, Salvador-BA, Brasil; Tel/Fax: +55-71 331762251; E-mail: irahe22@gmail.com

and regulatory genes, such as *Vif*, *Vpr*, *Vpu*, *Nef* [5]. In the last thirty years, these proteins have been used as targets to reduce the viral replication and to develop new drugs [8]. However, the viral resistance is still observed in clinical trials, being the key problem of drug development [10]. The virus adaptation to selective pressure imposed by drugs can reduce viral titer. Thus the resistant strain arises, able to avoid the pressure, prevails as the resistant strain starts a new infection scenario [11].

Over the past two decades, the following two strategies have often been used to find drugs against AIDS. One target is the HIV reverse transcriptase [12-17]; the other is the protease inhibitors [18-20] based on the distorted key theory [19, 21, 22]. Thus, developments in the new approach offering such treatments and vaccine acknowledgment about viral distribution, dynamics, infection, prevalence and genomics are required [23]. The mutagenic potential and diverse antigen/epitope assortment are serious challenges on the vaccine development, control and cure [24, 25]. To improve the knowledge, new information regarding host and virus is needed [26].

Therefore, studies with primary databases such as GenBank (http://www.ncbi.nlm.nih.gov/genbank) remain important. The GenBank sequence database is one of the biggest nucleotide databases [27]. This database stores over 154 billion nucleotide bases, corresponding to 167 million sequences, which is also a part of an international collaboration between the National Center for Biotechnology Information (NCBI), the European Molecular Biology Laboratory (EMBL) Data Library from the European Bioinformatics Institute (EBI) and the DNA Data Bank of Japan (DDBJ) [27]. There are over 580,000 HIV-1 sequences available on this integrated database. The HIV-1 sequences deposited in the GenBank represent only a small genome fragment usually as a part of a single genotype [23, 28-33]. Thus, to get more tangible information regarding HIV-1, more extensive analysis is necessary. The primary information stored in the databases could be used to understand and predict events or variables relevant to epidemic control and cure development [5].

However, some of the available variables, such as the viral subtype and genomic location are not present on the GenBank database. To obtain these information, it is necessary reanalyze the primary data. The golden standard for subtyping classification is the phylogeny analysis [34, 35]. Although broadly used, this technique is optimized to analyze only small and local databases. Moreover, phylogeny requires more computational resources to perform subtype classification analysis.

Thus, many scientists use heuristic methods to expedite complex analysis. However, heuristic approaches lack precision and are directly dependent on the database confirmation [36]. Herein, we developed an integrated system to obtain index and classified all nucleotide sequence from the GenBank. The software used optimal alignment algorithm to map and subtype sequences by comparing them to the subtype of reference sequences. Nonetheless, to reduce the amount of alignments, a pre analysis was performed. At this point, all recombinant subtypes were grouped for recombination derivation.

2. MATERIALS AND METHOD

2.1. Sequence Acquisition and Storage

The National Center for Biotechnology Information (NCBI) offers many tools for sequence management. One of these is the Global Query Cross-Database Search System (Entrez). Entrez offers a series of e-tools over Simple Object Access Protocol (SOAP), which employs a data transfer protocol based on eXtensive Markup Language (XML) and Socket connections. These tools are used to query all 38 Entrez databases and search for abstracts, gene information, taxonomy, 3D structures and nucleotide sequences [37]. Entrez employs the ESearch e-tool to select the database to be accessed, as well as the list of nucleotide sequence IDs, obtained by EFetch. The data retrieval system was developed to facilitate access to several databases within NCBI, including GenBank [38].

Regarding the Nucleotide database, the number of downloadable sequences is limited to protect the server when multiple requests are sent simultaneously. Limits are applied to the EFetch tool and up to 200 sequences can be downloaded at once [39]. If the query contains more than 200 sequences, the POST method must be used. This method employs an http (Hypertext Transfer Protocol) or https (Secure Hypertext Transfer Protocol) connection, transferring data directly as a stream. However, the POST method limits the connection request frequency and blocks the user IP if the maximum number of connections is reached. Thus, retrieving large amount of sequence data from the GenBank represents a complex task.

The development of the NSeek acquiring-tool was based on the limits imposed by each NCBI service. The first step in acquiring a nucleotide sequence is to acquire the individual identifiers (ID) utilizing the ESearh tool. However, if the local database contains some of the target sequences, it is unnecessary to download all of the IDs. Thus, the local database ID sequence list is compared to those retrieved from GenBank. Hence, a list with non-downloaded IDs is generated and the sequences are downloaded in the subsets containing 200 IDs each. Each sublist is then processed within a different thread, enabling simultaneous downloads. If one sequence download via SOAP method fails, NSeek automatically starts the download via the POST method. All the repeated download requests are delayed by 30 minutes, thereby avoiding NCBI service overload. All threads are synchronized with each other and the task manager reschedules the failed download, effectively avoiding the downloading process interruption.

After downloading all sequences the modulation to local database process starts. A scheduled task is created to demodulate each sequence and its features simultaneously, avoiding the download process impairing. When this task is finished, the sequence is modulated and then stored, with its features and cross-reference information, in the local database. Nevertheless, if errors occur in downloading or modulation process that cannot be quickly solved, a failsafe process is needed. Thus, we developed a manual import module. This module allows the user to import previously-downloaded GenBank (gb) files from the local database. However, some multi-sequence gb files are large size (*e.g.*

30 GB). These large files could disrupt the sequence of the modulation process. Addressing this issue, a gradual reading algorithm was developed to reduce memory allocation in the importation process. The algorithm reads line-by-line and applies regular expression identification. Both single and multiple sequence files can be modulated by NSeek. In summary, these novel approaches allow the download of GenBank stored-sequence and its indexing and insertion in the local databases, manually or automatically.

2.2. Sequence Mapping

Thus, we developed a mapping module to map all the sequences regarding HIV with its reference genome. The mapping process is based on a modified Needleman-Wusch algorithm, with the Gotoh affine gap penalty implementation for the alignment. This mathematical based algorithm performs a sequence similarity maximization, and uses a non-heuristic dynamic programming approach to infer the best alignment possible between the sequences [40]. However, utilization of non-heuristic algorithm is considered as the time and resource consuming process.

Therefore, several new technologies were used to better address this problem. The algorithm was modified so that the resource consumption could be reduced. For this purpose, we replaced the well-known Needleman matrix by two dynamic arrays. This way, the matrix was assembled and traced back simultaneously, reducing the alignment time and the use of resource. With the limited use of resource, more alignments could be made at the same time thus, overall decreasing the time exponentially. Thereat, a density map was created regarding the global distribution of HIV sequences. Within this map, all the fragments were represented within their limits. This map can also help to identify epitope density in the database, increasing researchers' datasets with more variable sequences and more concrete information in a short period of time.

2.3. Sequence Subtyping

A module of the genotype of all HIV-1 sequences available in the GenBank was developed to address the lack of information about sequence distribution. Thus, the module performed a non-heuristic alignment based on the modified Smith & Waterman algorithm with Gotoh affine gap penalty implementation. All modifications were performed to reduce time and resource consumption, optimizing the genotyping process.

The HIV genotyping process requires the comparison of all query sequences to a reference set. This corresponds to over 40 million alignments. To avoid such extensive comparison, we performed a sequence grouping based on recombination derivation. Thus, all the sub-subtypes (the recombinants) were grouped and linked to the "pure" subtype (the group M subtypes) showing its genomic characteristics. The sequence was then aligned to pure subtypes. Then, the alignment with the chained sub-subtypes, that showed high similarity in the previous pure alignment, was performed. Therefore, over 60% needed alignments were reduced, with no further precision losses. The subtype process is illustrated in Fig. **1**.

Furthermore, only similar measurement was performed by the algorithm to reduce the resource usage. The trace back and matrix calculation were removed, reducing 70% time and memory allocation. Moreover, after the comparison, the similar matrix was removed from the algorithm, reducing the memory consumption. After that, two dynamic arrays were

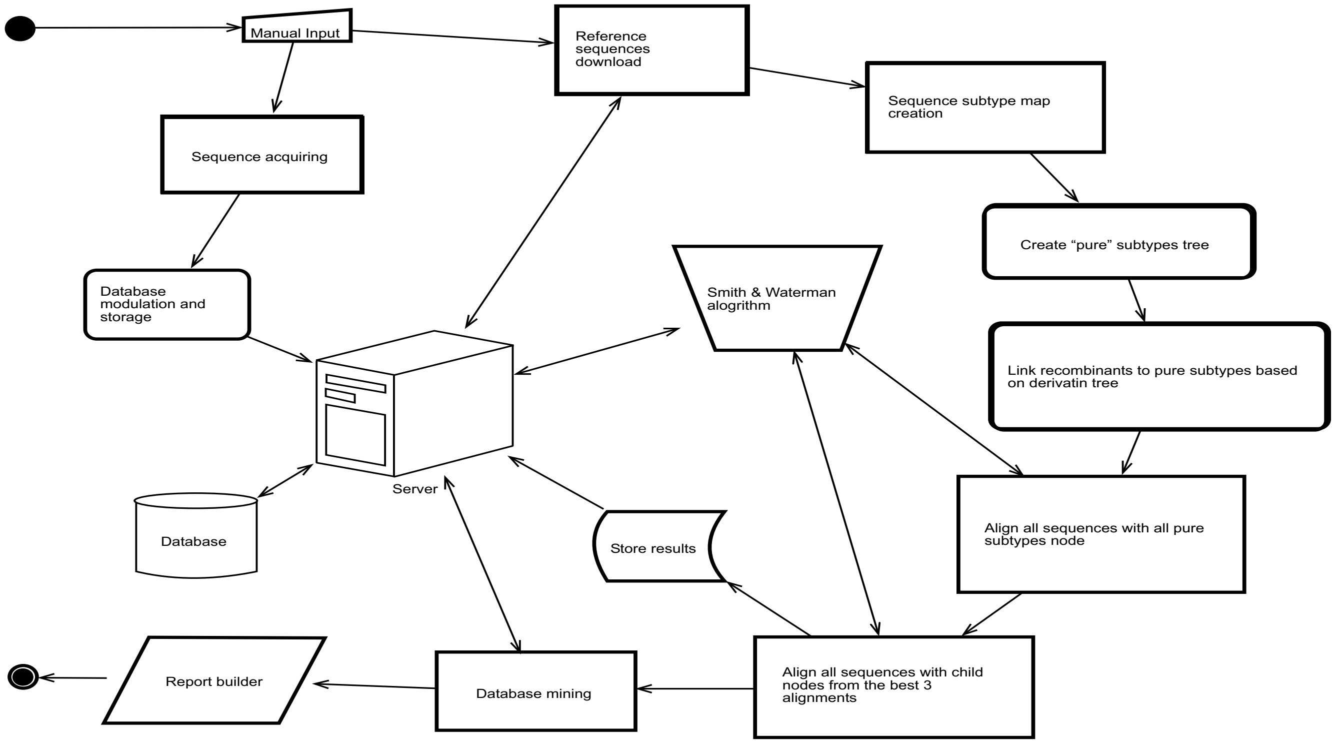

Manual Input
Reference sequences download
Sequence subtype map creation
Sequence acquiring
Create "pure" subtypes tree
Database modulation and storage
Smith & Waterman alogrithm
Link recombinants to pure subtypes based on derivatin tree
Server
Database
Store results
Align all sequences with all pure subtypes node
Report builder
Database mining
Align all sequences with child nodes from the best 3 alignments

Fig. (1). Sequence subtyping flow chart.

used in the measurement of similar score. The last line and the actual line were needed to perform this admeasurement. The performance enhancement allowed several simultaneous alignments execution. The number of simultaneous processes depends on the availability hardware resource.

A task manager was developed for supervising task creation, task pool management and resource consumption.In this way, the alignment process was monitored and enhanced. The manager read up to 200 sequences and created tasks guided by the resource analysis. Each task held the query and all chained reference sequences. First, the query was aligned to all "pure" reference sequences and the top-3 alignments with the best similarity scores were moved to the next step. After this, the query was aligned to the reference chained sequences associated to the top-3 alignment. The manager scheduled the writing process right after the termination of the reading process. In this way, the processing time was reduced.

2.3.1. Subtyping Validation

The validation was performed for comparing the classical approach to the grouping algorithm. The classical approach performed all reference set comparisons while the grouping algorithm followed the reference chain. In both the approaches, the alignment algorithm was non-heuristic. Thus, the results of subtyping process were compared using a confusion matrix.

3. RESULTS AND DISCUSSIONS

The tool was used to analyze the massive data in a reliable time with a high precision level. Currently, there are no subtyping tools available to analyze whole HIV sequences present in the GenBank with good performance and accuracy. Moreover, the accuracy did not differ between the classical and the optimized subtyping approach (data not shown). However, the optimized tool reduced 50% of the processing time as it took 1 day and 7 hours with the same accuracy, while the classical approach took 2 days and 10 hours for subtyping. Despite the phylogenetic process (Gold standard) is best applied in small datasets, its application is a time consuming process [34, 35], because its use in large datasets is a complex and almost unfeasible task. Thus, the reliable genotyping tool with standard genotyping procedures is essential to organism surveillance and further vaccine development [41].

Despite massive sequences related to HIV are deposited in the GenBank, the majority does not represent significant information individually, because these sequences are mainly partial gene sequences (Fig. **2**). While the whole genome has 9,500 base pairs and the average sequence length in the database is 1,005 base pairs. The ordered data according to the whole genome allows the identification of specific genomic areas without intense evaluation. Thus, this identification could drive new genomic and molecular epidemiology studies on the HIV**.**

The mapping process identified the beginning and end of the sequence according to the reference. Thereby, we were able to identify partial and total coverage for each HIV-1 gene (Fig. **3**). The results showed a high frequency of partial coverage of structural genes (*gag*, *pol* and *env*) compared to the total coverage, with a corresponding density of 66.41%. This suggests a low representation bias of the nonstructural genes in the dataset. The *pol* gene was the most prevalent in GenBank, reaching 47.8% of all HIV-1 sequences (Fig. **3**). However, no unrepresented region was observed. The lowest coverage region was 5' LTR with 1.44% density.

Regarding subtypes prevalence, the B was the most observed, with 45.96% prevalence, followed by C (2.93%) and A1 (2.46%) subtypes, as shown in Fig. **4**. Furthermore, two HIV-1 strains can hybridize, generating a new circulating virus with a mosaic genome [42, 43]. It occurs when different subtypes infect the same cell. The recombination process gives rise a double subtype strain (resulting from 2 subtypes coinfection) or even more complex strains (three or more subtype coinfection). These processes make the disease control and surveillance difficult worldwide.

Each recombinant presents particular characteristics and dynamics. Thus, with the analysis, we were able to identify the HIV-1 subtype prevalence as shown in Fig. **5**. Moreover, there were total 47.21%recombinant sequences in the database (Fig. **4**). The most prevalent recombinant was the CRF_03AB with 13.61%, followed by CRF01_AE with 12.67% (Fig. **5**).

The graphic representation offers a novel intuitive perspective and useful insights regarding biological systems. Several graphic approaches have been applied successfully in other biological issues, such as enzyme-catalyzed reactions [44-47] slow conformational change [48], protein folding kinetics and folding rates [49], the inhibition of HIV-

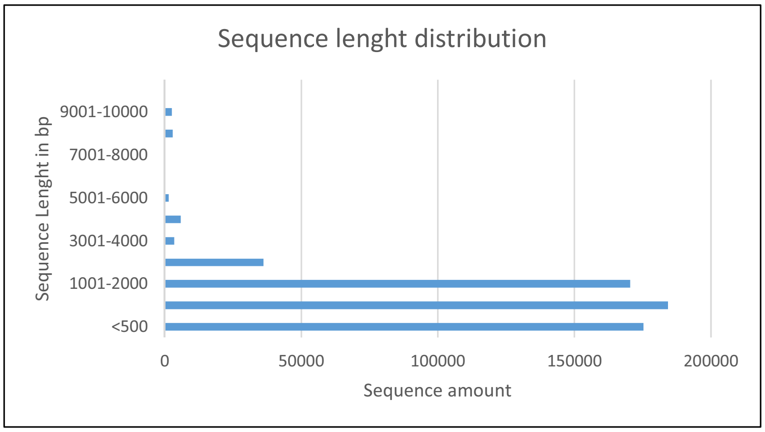


Fig. (2). HIV-1 sequence length distribution on GenBank database considering all sequences available.

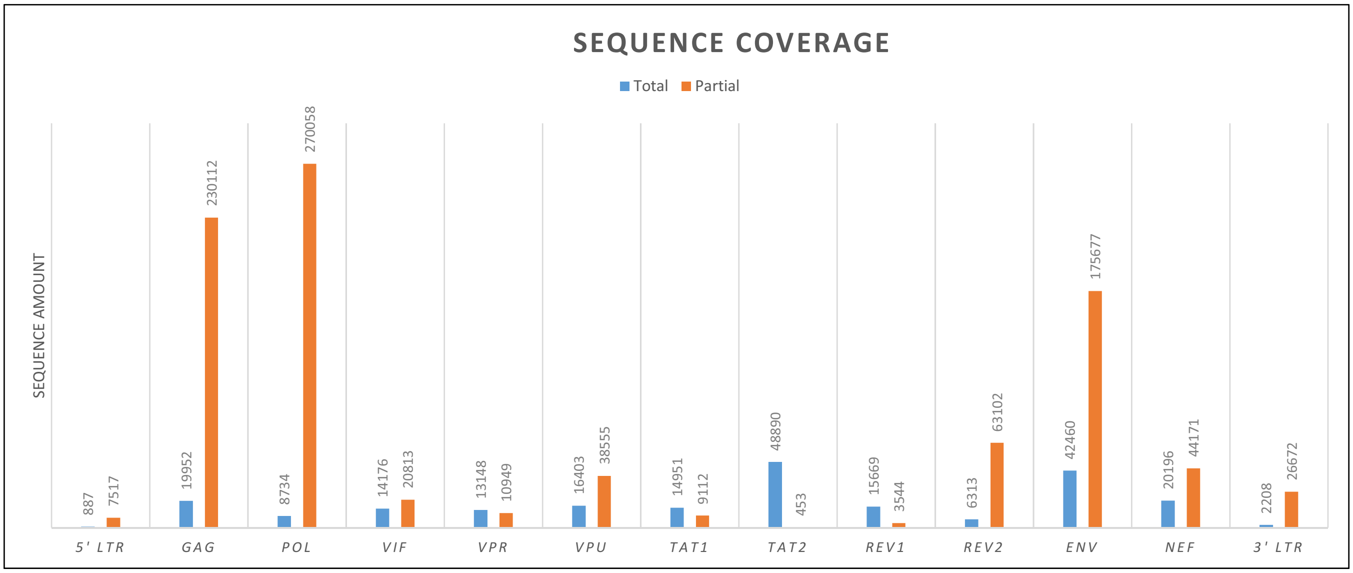


Fig. (3). Sequence coverage of structural and regulatory genes.

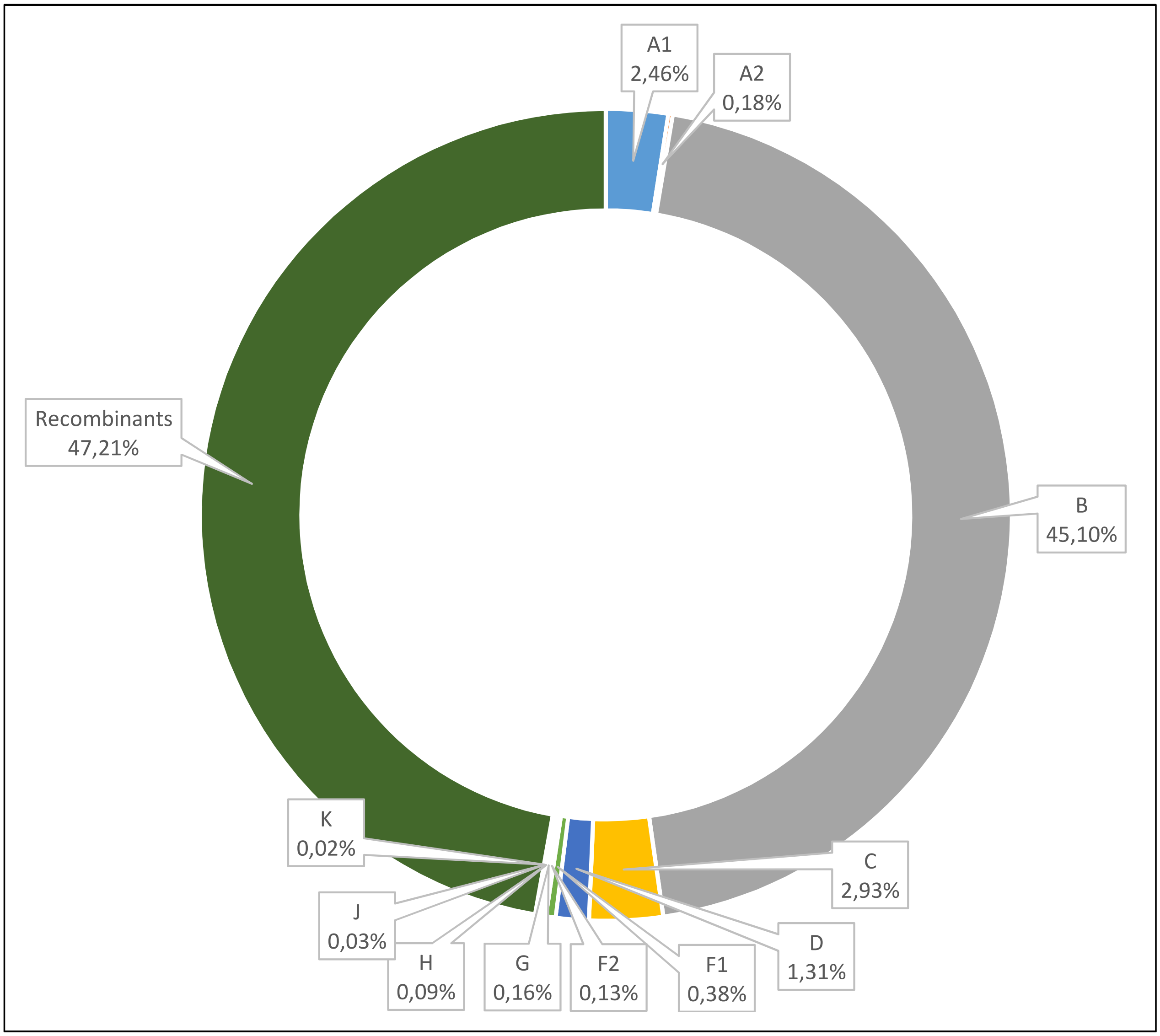


Fig. (4). Pure and Recombinant Subtype Distribution.

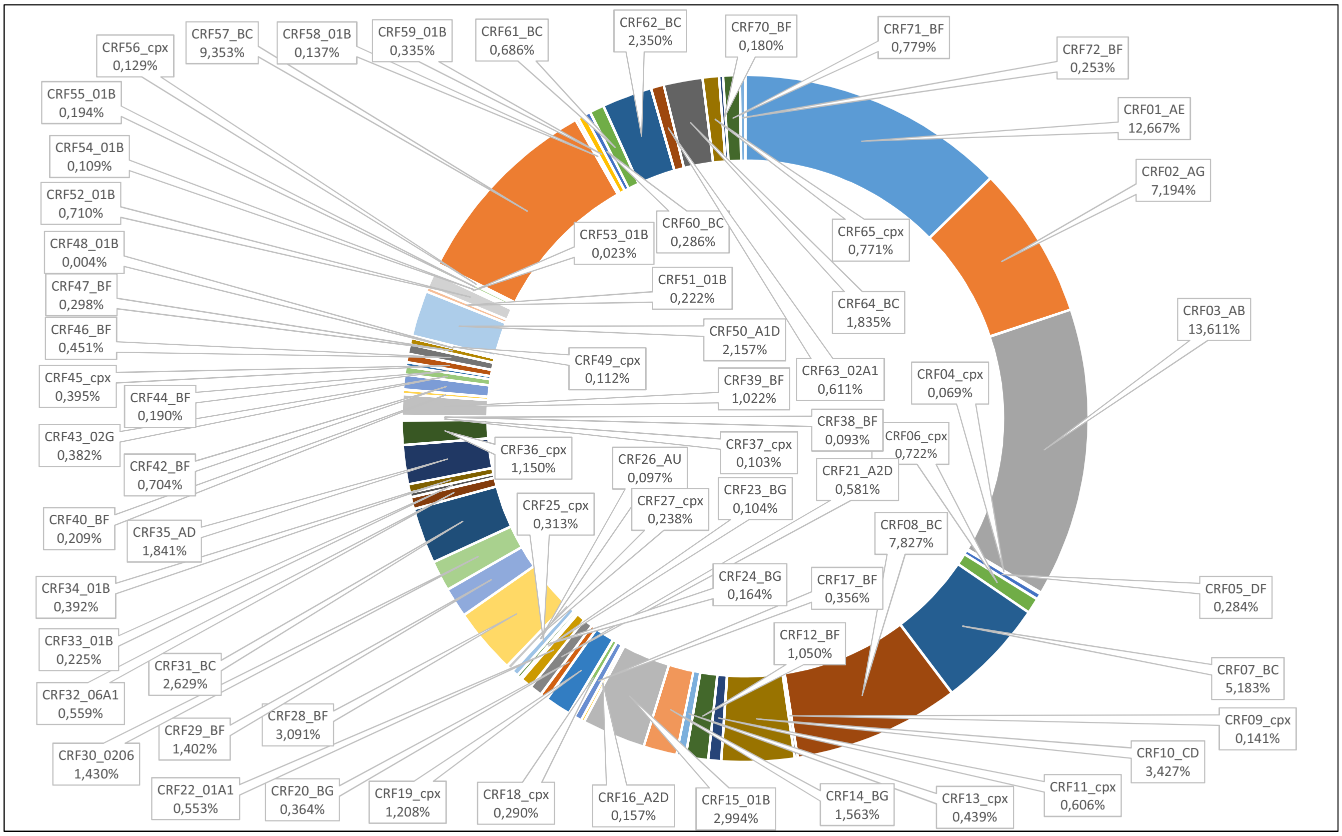


Fig. (5). Recombinant sequence distribution.

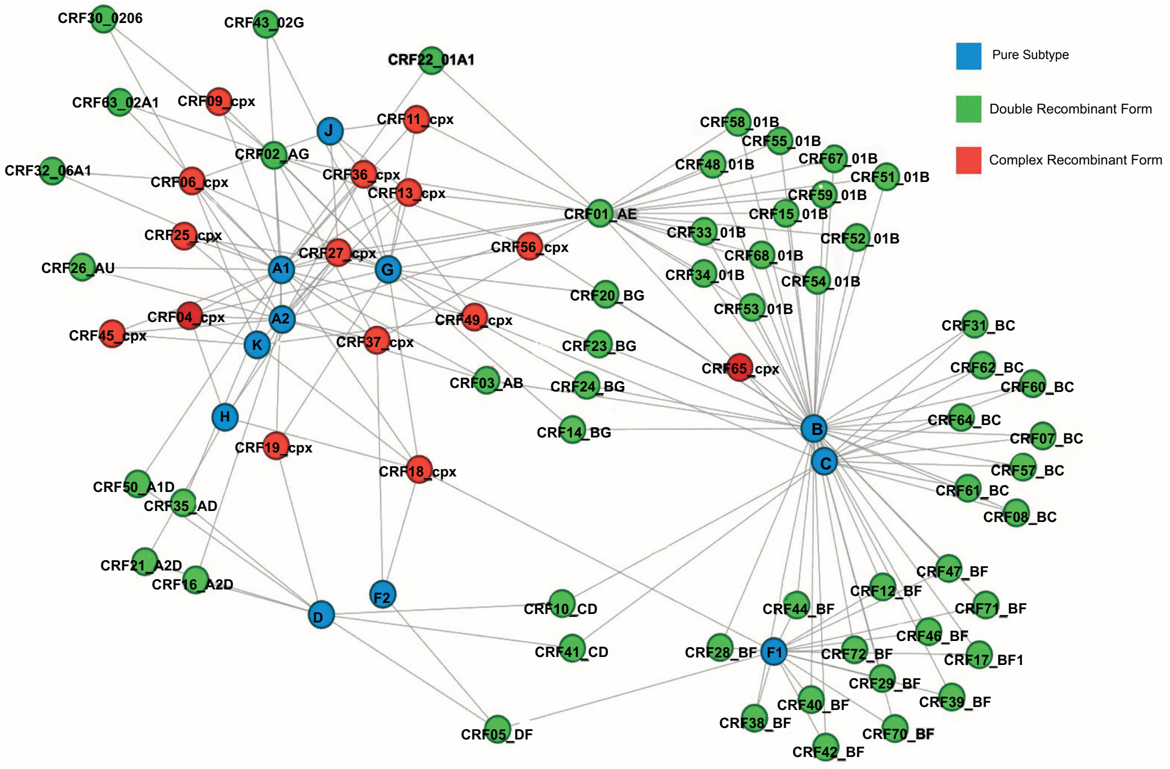


Fig. (6). HIV-1 Subtype and Recombinant derivation new subspecies heat spot. Complex recombinants also known as "CPX" are result of coinfection by several different virus strain creating a genomic mosaic. The relation between double and complex subtype represented a double subtype's majority. However, complex subtypes still represent 6.68% of all recombinant sequences classified. With all sequences subtypes, we were able to create an expansion based on recombination derivation of all sequences on database and its subtypes. This distribution expansion can be addressed in Fig. **5**.

1 reverse transcriptase [12-14], non-steady drug metabolism systems [50], hepatitis B viral infections [51], HBV virus gene missense mutation [52], the evolution of biological sequences [53] and the use of *wenxiang* graphs [54] applied to protein-protein interactions [55, 56].In this study, we applied complex network modeling to offer an enhanced perspective regarding the derivation of variability in HIV-1 (Fig. **6**). This node network was created using the observed subtype distribution and the Yifan Hu method of mathematical modeling. This graphic approach highlighted close relationship between some pure recombinants. These insights allowed the development of a subtyping derivation strategy, dramatically reducing the overall alignment time required.

After classification, each sequence was indexed in the database, allowing filtered searches and easy individual dataset creation. We classified the recombinants as double" sub-subtypes and "complex" sub-subtypes. The double represents two circular virus strain originated sequences and 3 complex or others. This provided an insight into the comprehensive genomic evaluation and a virus mutational dynamics, assisting in the prevalence studies. Regarding virus adaptation, the primary challenge in the vaccine and drug development [57], the databases assist in escape mutation surveillance. The virus interaction with the host immune system occurs when CD8+ T-cells recognize HIV-1 antigens. These epitopes have 8 to 11 amino acid length and bounds to MHC class I molecules [58].

However, when the escape mutations takes place within an epitope coding region, the binding process does not occur. This impairs the immune response leading virus persistence. Our system allows the user to evaluate the fragmented sequences in a complete genome context. This facilitates the studies on epitope and mutation, because all the sequences are related to HXB2 reference genome. This standardization is crucial to sequence-feature comparison

In summary, the alignment process for a large amount of sequences is a complex computational task. This is associated to viral adaptation [24]. The HIV-1 evades the immune system by using diverse ways such as, external glycoprotein, heavy glycosylation, escape mutations and recombination [59]. These viral features dramatically increase the complexity of alignment and comparison process making the analysis of all sequences deposited in database an intricate task. Thus, the analysis was applied to all 582,678 sequences available in the GenBank using a single server. The subtyping process took 1 day and 7 hours, and the original serial approach was suspended (after a 6-day runtime) and was estimated to last 97 years. Both the analyses were performed using a dual hex core Intel Xeon® Processor X5650 processor containing 16 individual cores, 32GB RAM running Ubuntu 14.04 O.S.

4. CONCLUSION

The recent development of various genome analysis tools has brought about significant progress [60-65]. These tools have been successfully used in the genome analysis [61, 66-78]. Thus, interesting outcomes will be obtained by using these state-of-the-art tools on HIV genome. The recent progress regarding the role of HIV-1 proteins, specifically env genes [79] and the role during the infection, immunogenicity and host interaction is taken into consideration [80, 81]. However, we presented a new user-friendly software for massive data analysis. The software was optimized to high mutational viruses sequence analysis, such HCV and HIV, and this analysis was performed in a reliable time.

Moreover, the results showed a new tendency regarding recombinant distribution and highlighted the genome fragmentation in the database (gene coverage bias), with the structural genes being more represented. Thus, this analysis of all the sequences from HIV provides new basis for epidemics, subtypes and fragmentation distribution.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work was partly supported by Oswaldo Cruz Foundation and Feira de Santana State University.

REFERENCES

[1] Hemelaar J, Gouws E, Ghys PD, Osmanov S. Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. AIDS 2006; 20(16): W13-W23.

[2] UNAIDS, W.H.O. Report on the Global AIDS Epidemic. 2013.

[3] UK Colaborative Group on HIV Droug Resistance. The increasing genetic diversity of HIV-1 in the UK, 2002-2010. AIDS 2014; 28: 773-80.

[4] Neher RA, Leitner T. Recombination rate and selection strength in HIVintrapatient evolution. PLoS Comput Biol 2010; 6(1): e1000660.

[5] Li G, Piampongsant S, Faria NR, *et al.* An integrated map of HIV genome-wide variation from a population perspective. Retrovirology 2015; 12: 18.

[6] Hemelaar J. The origin and diversity of the HIV-1 pandemic. Trends Mol Med 2012; 18: 182-92.

[7] Geretti AM, Harrison L, Green H, *et al.* Effect of HIV-1 subtype on virologic and immunologic response to starting highly active antiretroviral therapy. Clin Infect Dis 2009; 48: 1296-305.

[8] Kantor R. Impact of HIV-1 pol diversity on drug resistance and its clinical implications. Curr Opin Infect Dis 2006; 19: 594-606.

[9] Engelman A, Cherepanov P. The structural biology of HIV-1: mechanistic and therapeutic insights. Nat Rev Microbiol 2012; 10: 279-90.

[10] Perry CM. Elvitegravir/cobicistat/ emtricitabine /tenofovir disoproxil fumarate single-tablet regimen: A review of its use in the management of HIV-1 infection in adults. Drugs 2014; 74: 75-97.

[11] Smyth RP, Davenport MP, Mak J. The origin of genetic diversity in HIV-1. Virus Res 2012; 169: 415-29.

[12] Althaus IW, Chou JJ, Gonzales AJ, *et al.* Steady-state kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-87201E. J Biol Chem 1993; 268: 6119-124.

[13] Althaus IW, Gonzales AJ, Chou JJ, *et al.* The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase J Biol Chem 1993; 268: 14875-80.

[14] Althaus IW, Chou JJ, Gonzales AJ, *et al.* Kinetic studies with the nonnucleoside HIV-1 reverse transcriptase inhibitor U-88204E. Biochemistry 1993; 32: 6548-54.

[15] Althaus IW, Chou JJ, Gonzales AJ, *et al.* Kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-90152E. Biochem Pharmacol 1994; 47: 2017-28.

[16] Althaus IW, Chou KC, Lemay RJ, *et al.* The benzylthio-pyrididine U-31,355 is a potent inhibitor of HIV-1 reverse transcriptase. Biochem Pharmacol 1996; 51: 743-50.

[17] Chou KC, Kézdy FJ, Reusser F. Review: Steady-state inhibition kinetics of processive nucleic acid polymerases and nucleases. Anal Biochem 1994; 221: 217-30.

[18] Chou KC. A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. J Biol Chem 1993; 268: 16938-48.

[19] Chou KC. Review: prediction of human immunodeficiency virus proteasecleavage sites in proteins. Anal Biochem 1996; 233: 1-14.

[20] Shen HB, Chou KC. HIVcleave: a web-server for predicting HIV protease cleavage sites in proteins. Anal Biochem 2008; 375: 388-90.

[21] Gan YR, Huang H, Huang YD, *et al.* Synthesis and activity of an octapeptide inhibitor designed for SARS coronavirus main proteinase. Peptides 2006; 27: 622-5.

[22] Du QS, Sun H, Chou KC. Inhibitor design for SARS coronavirus main protease based on "distorted key theory". Med Chem 2007; 3: 1-6.

[23] Sanabani SS, Pessôa R, Soares de Oliveira AC, *et al.* Variability of HIV-1 genomes among children and adolescents from são paulo, Brazil. PLoS One 2013; 8(5): e62552.

[24] McBurney SP, Ross TM. Viral sequence diversity: challenges for AIDS vaccine designs. Expert Rev Vaccines 2008; 7: 1405-17.

[25] Butler IF, Pandrea I, Marx PA, Apetrei C. HIV genetic diversity: biological and public health consequences. Curr HIV Res 2007; 5: 23-45.

[26] Schroeder SA. Much accomplished, much to do. J Gen Intern Med 2013; 9: 1104-7.

[27] National Center for Biotechnology Information. GenBank:Scope. http://www.ncbi.nlm.nih.gov/books/NBK153518/ (Accessed Aug 27, 2015).

[28] Rolland M, Edlefsen PT, Larsen BB, *et al.* Increased HIV-1 vaccine efficacy against viruses with genetic signatures in Env V2. Nature 2012; 490: 417-20.

[29] Herbeck JT, Rolland M, Liu Y, *et al.* Demographic processes affect HIV-1 evolution in primary infection before the onset of selective processes. J Virol 2011; 85: 7523-34.

[30] Brown BK, Darden JM, Tovanabutra S, *et al.* Jackson Foundation, Rockville, Maryland 1 ; The Division of AIDS, National Institutes of Health, Bethesda, Maryland 2 ; Armed Forces Research Institute of Medical Sciences, Bangkok, Thailand 3 ; and Walter Reed Army Institute of Research, Rockvil. Society 2005; 79(10): 6089-101.

[31] Fernández-García A, Revilla A, Vázquez-de Parga E, *et al.* The analysis of near full-length genome sequences of HIV type 1 subtype a viruses from russia supports the monophyly of major intrasubtype clusters. AIDS Res Hum Retroviruses 2012; 28: 1340-3.

[32] van der Kuyl AC, Berkhout B. The biased nucleotide composition of the HIV genome: a constant factor in a highly variable virus. Retrovirology 2012; 9: 92.

[33] Mayrose I, Stern A, Burdelova EO, *et al.* Synonymous site conservation in the HIV-1 genome. BMC Evol Biol 2013; 13: 164.

[34] Leitner T, Escanilla D, Franzén C, Uhlén M, Albert J. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. Proc Natl Acad Sci U S A 1996; 93: 10864-9.

[35] Valsamakis A. Molecular testing in the diagnosis and management of chronic hepatitis B. Clin Microbiol Rev 2007; 20: 426-39.

[36] Chakraborty A, Bandyopadhyay S. FOGSAA: Fast Optimal Global Sequence Alignment Algorithm. Sci Rep 2013; 3: 1746.

[37] National Center for Biotechnology Information. Entrez Programming Utilities Help. http://www.ncbi.nlm.nih.gov/books/NBK25501/ (Accessed Aug 27, 2015).

[38] McEntyre J. Linking up with entrez, Trends Genet 1998; 14(1): 39-40.

[39] National Center for Biotechnology Information. E-Utils:EFetch. http://www.ncbi.nlm.nih.gov/books/NBK25499/#chapter4.EFetch (Accessed Aug 27, 2015).

[40] Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970; 48(3): 443-53.

[41] Gifford R, de Oliveira T, Rambaut A, *et al.* Assessment of automated genotyping protocols as tools for surveillance of HIV-1 genetic diversity. AIDS 2006; 20(11): 1521-9.

[42] Peeters M. Recombinant HIV sequences: their role in the global epidemic. In: theoretical biology and biophysics group, editors HIV sequence compendium los Alamos: los Alamos national laboratory; 2000: pp 39-54.

[43] Baryshev PB, Bogachev VV, Gashnikova NM. HIV-1 genetic diversity in russia: CRF63_02A1, a new HIV type 1 genetic variant spreading in siberia. AIDS Res Hum Retroviruses 2014; 30(6): 592-7.

[44] Jiang SP, Liu WM, Fee CH. Graph theory of enzyme kinetics: 1. Steady-state reaction system. Scientia Sinica 1979; 22: 341-58.

[45] Forsen S. Graphical rules for enzyme-catalyzed rate laws. Biochem J 1980; 187: 829-35.

[46] Chou KC. Graphic rules in steady and non-steady enzyme kinetics. J Biol Chem 1989; 264: 12074-9.

[47] Zhou GP, Deng MH. An extension of Chou's graphic rules for deriving enzyme kinetic equations to systems involving parallel reaction pathways. Biochem J 1984; 222: 169-76.

[48] Lin SX, Neet KE. Demonstration of a slow conformational change in liver glucokinase by fluorescence spectroscopy. J Biol Chem 1990; 265: 9670-5.

[49] Chou KC. Review: Applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady state systems. Biophys Chem 1990; 35: 1-24.

[50] Chou KC. Graphic rule for drug metabolism systems. Curr Drug Metab 2010; 11: 369-78.

[51] Xiao X, Shao SH, Chou KC. A probability cellular automaton model for hepatitis B viral infections. Biochem Biophys Res Commun 2006; 342: 605-10.

[52] Xiao X, Shao S, Ding Y, Huang Z, Chen X, Chou KC. An Application of Gene Comparative Image for Predicting the Effect on Replication Ratio by HBV Virus Gene Missense Mutation. J Theor Biol 2005; 235: 555-65.

[53] Wu ZC, Xiao X, Chou KC. 2D-MH: A web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids. J Theor Biol 2010; 267: 29-34.

[54] Chou K-C, Lin WZ, Xiao X. Wenxiang: a web-server for drawing wenxiang diagrams. Nat Sci 2011; 3: 862-5.

[55] Zhou GP. The disposition of the LZCC protein residues in wenxiang diagram provides new insights into the protein-protein interaction mechanism J Theor Biol 2011; 284: 142-8.

[56] Zhou GP, Huang RB.The pH-Triggered Conversion of the PrP(c) to PrP(sc.). Curr Top Med Chem 2013; 13(10): 1152-63.

[57] Snoeck J, Fellay J, Bartha I, Douek DC, Telenti A. Mapping of positive selection sites in the HIV-1 genome in the context of RNA and protein structural constraints. Retrovirology 2011; 8: 87.

[58] Roider J, Meissner T, Kraut F, *et al.* Comparison of experimental fine-mapping to in-silico prediction results of HIV-1 epitopes reveals ongoing need for mapping experiments. Immunology 2014;143(2): 193-201.

[59] Taylor BS, Hammer SM. The challenge of HIV-1 subtype diversity. N Engl J Med 2008; 359: 1965-6.

[60] Chen W, Lei TY, Jin DC, Lin H, Chou KC. PseKNC: a flexible web-server for generating pseudo K-tuple nucleotide composition. Anal Biochem 2014; 456: 53-60.

[61] Chen W, Zhang X, Brooker J, Lin H, Zhang L, Chou KC. PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. Bioinformatics 2015; 31: 119-20.

[62] Guo SH, Deng EZ, Xu LQ, *et al.* iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. Bioinformatics 2014; 30: 1522-9.

[63] Liu B, Liu F, Fang L, Wang X. repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. Bioinformatics 2015; 31: 1307-9.

[64] Liu B, Liu F, Fang L, Wang X, Chou KC. repRNA: a web server for generating various feature vectors of RNA sequences. Mol Genet Genomics 2016; 291: 473-81.

[65] Liu B, Liu F, Wang X, Chen J, Fang L, Chou KC. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res 2015; 43: W65-W71.

[66] Chen W, Feng PM, Lin H, Chou KC. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res 2013; 41: e68.

[67] Qiu WR, Xiao X, Chou KC. iRSpot-TNCPseAAC: Identify recombination spots with trinucleotide composition and pseudo amino acid components. Int J Mol Sci 2014; 15: 1746-66.

[68] Lin H, Deng EZ, Ding H, Chen W, Chou KC. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in

prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Res 2014; 42: 12961-72.

[69] Chen W, Feng PM, Lin H, Chou KC. iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. Biomed Res Int 2014; 623-149.

[70] Chen W, Feng PM, Deng EZ, Lin H, Chou KC. iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. Anal Biochem 2014; 462: 76-83.

[71] Chen W, Lin H, Chou KC. Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. Mol Biosyst 2015; 11: 2620-34.

[72] Chou KC. Impacts of bioinformatics to medicinal chemistry. Med Chem 2015; 11: 218-34.

[73] Liu B, Long R. iDHS-EL: Identifying DNase I hypersensi-tivesites by fusing three different modes of pseudo nucleotide composition into an en-semble learning framework. Bioinformatics 2016; 32: 2411-8.

[74] Liu B, Fang L, Long R, Lan X. iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. Bioinformatics 2016; 32: 362-89.

[75] Liu B, Wang S, Long R, Chou KC. iRSpot-EL: identify recombination spots with an ensemble learning approach. Bioinformatics 2017; 33(1): 35-41.

[76] Chen W, Feng P, Ding H. iRNA-Methyl: Identifying N6-methyladenosine sites using pseudo nucleotide composition. Anal Biochem 2015; 490: 26-33.

[77] Chen W, Tang H, Ye J, Lin H. iRNA-PseU: Identifying RNA pseudouridine sites. Mol Ther Nucleic Acids 2016; 5: e332.

[78] Chen W, Feng P, Ding H, Lin H, Chou KC. Using deformation energy to analyze nucleosome positioning in genomes. Genomics 2016; 107: 69-75.

[79] Dev J, Park D, Fu Q, *et al.* Structural basis for membrane anchoring of HIV-1 envelope spike. Science 2016; 353: 172-5.

[80] Sirois S, Sing T. Review: HIV-1 gp120 V3 loop for structure-based drug design. Curr Protein Pept Sci 2005; 6: 413-22.

[81] Sirois S, Touaibia M, Roy R. Review: Glycosylation of HIV-1 gp120 V3 loop: towards the rational design of a synthetic carbohydrate vaccine. Curr Med Chem 2007; 14: 3232-42.