Molecular Biology Research Communications 2016;5(2):114- 121 MBRC

Original Article Open Access

In silico comparison of Iranian HIV -1 envelop glycoprotein with five nearby countries

Maryam Ghafari, Mandana Behbahani, Hassan Mohabatkar*

Department of Biotechnology, Faculty of Advanced Sciences and Technologies, University of Isfahan, Isfahan, Iran

ABSTRACT

HIV-1 envelope (env) glycoprotein mediates an important role in entry of the virus into the susceptible target cells. As env glycoprotein of HIV-1 is highly variable in the different geographical regions, in the present study, different properties of this protein in Iran are compared with five nearby countries. The sequences of HIV-1 env glycoproteins of Iran, Afghanistan, Russia, Turkey, Pakistan and Saudi Arabia databases were collected from databases. Amino acid composition and physical and chemical properties of the proteins from these countries were studied using Protparam and COPid tools. Receiver-operating characteristic (ROC) curve analysis and Support Vector Machine (SVM) were used to evaluate association between the properties of HIV-1 env glycoprotein of Iran with five nearby countries. The results verify that amino acid composition and four physical and chemical properties (molecular weight, isoelectric point, Aliphatic Index, and grand average of hydropathicity) of HIV-1 env protein in Iran and Russia were not significantly different. In conclusion, the results indicate that in silico techniques provide valuable information for comparing HIV-1 envelop glycoprotein in different geographical locations.

Keywords: Amino acid composition; Envelope glycoprotein; HIV-1; Physical and chemical properties

INTRODUCTION

HIV-1 is one the most important pathogens and causes the majority of HIV infections globally [1, 2]. HIV-1 is a member of Retroviridae and is sorted into three

*Corresponding Author: Department of Biotechnology, Faculty of Advanced Sciences and Technologies, University of Isfahan, Isfahan, Iran
Tel: +98-3137934391
FAX: +98-3137932342
E.mail: h.mohabatkar@ast.ui.ac.ir

pISSN 2322-181X eISSN 2345-2005

typical groups; group M (main), group O (outliner), and group N (non-M/non-O) [3, 4]. The majority of the infection is caused by group M which is divided into 9 different subtypes symbolized A, B, C, D, F, G, H, J and K [5, 6]. Genome of HIV is composed of two positive strands RNA which are packaged in a protein capsid and surrounded by a lipid env [7]. HIV-1 env glycoprotein is essential for entry of the virus into the cell by binding to the specific receptors on the surface of the target cells [8]. HIV-1 env glycoprotein has proved to be useful to study of variation in HIV strains by a number of approaches [4]. Some of these approaches are based on amino acid sequences, amino acid composition and physical and chemical properties of proteins. Physical and chemical properties of viral proteins are widely used to predict various aspects of proteins such as molecular weight, isoelectric point, instability index, aliphatic index and grand average of hydropathicity (GRAVY) [9]. Several researchers have worked on evolutionary patterns of HIV-1 envelop group M in African, Asia, Europe and America [10]. In addition, phylogenetic analysis of HIV-1 env glycoprotein of different Asian countries such as India, Bangladesh, Cambodia, China and Japan were studied [11]. But the diversity of HIV-1 envelops in Iran and some nearby courtiers have not been investigated yet. In the present study, the Amino acid composition and some physical and chemical properties of HIV-1 env proteins of different subtypes in Iran and five nearby countries are studied.

MATERIALS AND METHODS

Data collection: Amino acid sequences of HIV-1 env protein from six different countries (Iran, Pakistan, Russia, Saudi Arabia, Afghanistan and Turkey) were collected from NCBI (http://www.ncbi.nlm.gov/protein). Two hundred sixty two amino acid sequences from Iran and 752 sequences from five nearby countries (349 Pakistan, 280 Russia, 62 Saudi Arabia, 34 Afghanistan and 27 Turkey) were studied. Redundant sequences (more than 95% similarity) were removed from our dataset by CD-HIT. After running CD-HIT, the number of sequences existing in the dataset in Iran, Pakistan, Russia, Saudi Arabia, Afghanistan and Turkey were 123, 123, 50, 26, 12 and 30 respectively. In the next step, all nucleotide sequences of these 6 groups were also collected from NCBI.

Server and tools:

Context-based Modeling for Expeditious Typing (COMET): All nucleotide sequences of these HIV-1 env genes of the 6 countries were analyzed using the Calibrated Population Resistance (CRP) subtyping tool COMET to predict different subtypes. COMET (v. 0.2) is a reliable tool to predict HIV-1/2 subtypes [12]. This tool is available at (http://comet.retrovirology.lu) and uses Prediction Partial Matching (PPM) compression algorithm.

ProtParam tool: ProtParam is a server which computes different physical and chemical parameters of a protein [13]. The web-server is available at

http://web.expasy.org/protparam. Four characteristics (molecular weight, isoelectric point, GRAVY and aliphatic index) of HIV-1 env glycoprotein from Iran and five nearby countries were evaluated by ProtParam.

Composition protein identification: The amino acid sequences of env glycoprotein from these six countries were analyzed using Composition Based Protein Identification (COPid). COPid is a server which analyzes composition of various types of amino acids [14]. The COPid web-server is available at http://www.imtech.res .in/raghava/ COPid. This server helps the researchers to elucidate the function of a protein and generate a phylogenetic tree from its composition.

Statistical Analysis: The data were evaluated by Receiver Operating Characteristic (ROC) curve and Molegro Data Modeller (MDM). ROC curve is a tool for organizing classifiers and visualizing their performance. ROC curves are usually used in machine learning and data mining research [15]. ROC server can calculate accuracy (ACC) and compute association between the properties of HIV-1 env glycoprotein in Iran and five nearby countries near Iran. ACC is a factor to differentiate positive and negative classes of data. When ACC value is more than 0.8, it means that the difference between two classes is significant. ACC is calculated using the following formula:

ACC = Σ True positive + Σ True negative/ Σ Total population

MDM is a cross-platform application for data mining and data visualization. MDM generates regression and classification models by partial least squares, neural networks and support vector machines [16]. SVMs are supervised learning algorithms which are broadly used in classification or regression problems [17].

RESULTS

The result of HIV-1 subtyping in Iran, Russia, Turkey, Saudi Arabia, Pakistan and Afghanistan are summarized in Table 1. The results showed that subtype A was the dominant subtype in Iran, Russia and Pakistan. But dominant subtypes in Saudi Arabia, Turkey and Afghanistan were C, B and AD, respectively.

Table 1: COMET subtyping tool Results I: Iran A: Afghanistan P: Pakistan T: Turkey R: Russia SA: Saudi Arabia

Countries	A	AD	AE	AG	B	C	D
I	46%	20%	-	-	26%	8%	-
A	-	93%	7%	-	-	-	-
P	95%	3%	2%	-	-	-	-
T	12%	-	-	-	72%	-	3%
R	70%	-	-	15%	15%	-	-
SA	3.1%	-	-	4%	20%	37%	5%

Analysis of ProtParam results using ROC curve and SVM are presented in Tables 2 and 3. Table 2 shows that ACC values of GRAVY, aliphatic index, isoelectric point and molecular weight between Iran and Russia were 0.56, 0.65, 0.62 and 0.65 respectively. But ACC values between Iran and four other countries were more (Table 2). As the results of ROC analysis between Iran and Russia were less than 0.8, physicochemical properties of viral env protein between these two countries were not significantly different. The results of MDM analysis approved our previous results (Table 3), because values of four parameters such as molecular weight, isoelectric point, aliphatic Index and GRAVY between Iran and Russia were less than 6.4 and the values between Iran and four other countries were more than 0.7. The results of ROC curve and MDM analyses showed that physicochemical properties of HIV-1 env protein in Iran and Russia were not significantly different.

Table 2: ROC Analysis of ProtParam Results I: Iran A: Afghanistan P: Pakistan T: Turkey R: Russia SA: Saudi Arabia

Classifier	I-T	I-R	I-A	I-SA	I-P
GRAVY	0.68	0.62	0.91	0.89	0.83
Aliphatic Index	0.98	0.65	0.94	0.97	0.92
PI	1.00	0.56	0.89	0.87	0.97
MW	0.73	0.65	0.93	0.80	0.84

Table 3: Performance of MDM I: Iran A: Afghanistan P: Pakistan T: Turkey R: Russia SA: Saudi Arabia

Classifier	I-T	I-R	I-A	I-SA	I-P
GRAVY	0.77	0.50	0.91	0.89	0.81
Aliphatic Index	0.97	0.64	0.93	0.98	0.89
PI	1.00	0.50	0.91	0.87	0.92
MW	0.80	0.61	0.90	0.80	0.84

The results of ROC curve analysis for amino acid composition of six databases are been shown in Table 4. The ACC values of 17 amino acids out of 20 amino acids between Iran and Russia were less than 70%. These amino acids were Asp, Ala, Glu, Phe, Ile, His, Gly, Leu, Asn, Lys, Arg, Pro, Gln, Thr, Val, Trp and Tyr. However ACC values between Iran and four other countries were significantly different.

The results of Molegro analysis for amino acid composition of six databases are shown in Table 5. The ACC values of 17 amino acids out of 20 amino acids between Iran and Russia databases were less than 0.7. The ACC of Cys, Met, and Ser between Iran and Russia were more than 0.7.

Table 4: ROC analysis results of 20 amino acids of six databases I: Iran A: Afghanistan P: Pakistan T: Turkey R: Russia SA: Saudi Arabia

Classifier	I-T	I-R	I-A	I-SA	I-P
Asp	0.85	0.62	0.90	0.71	0.92
Cys	0.98	0.74	0.90	0.95	0.96
Ala	0.81	0.66	0.90	0.84	0.83
Glu	0.91	0.56	0.82	0.71	0.83
Phe	0.99	0.67	0.90	0.91	0.99
Ile	0.84	0.56	0.86	0.70	0.84
His	0.93	0.64	0.82	0.87	0.90
Gly	1.00	0.65	0.85	0.77	1.00
Leu	1.00	0.64	0.96	0.96	1.00
Met	0.85	0.71	0.90	0.86	0.81
Asn	0.89	0.60	0.96	0.89	0.84
Lys	0.81	0.56	0.75	0.76	0.80
Arg	0.97	0.59	0.82	0.86	0.97
Pro	0.99	0.63	0.83	0.96	1.00
Ser	0.90	0.75	0.78	0.76	0.83
Gln	1.00	0.60	0.82	0.95	0.98
Thr	0.96	0.62	0.91	0.98	0.99
Val	0.95	0.56	0.80	0.69	0.86
Trp	1.00	0.66	0.88	0.96	1.00
Tyr	0.93	0.60	0.88	0.80	0.91

Table 5: Molegro analysis results of 20 amino acids of six databases I: Iran A: Afghanistan P: Pakistan T: Turkey R: Russia SA: Saudi Arabia

Classifier	I-T	I-R	I-A	I-SA	I-P
Asp	0.88	0.54	0.91	0.71	0.88
Cys	0.98	0.74	0.91	0.94	0.94
Ala	0.82	0.61	0.91	0.82	0.87
Glu	0.89	0.53	0.91	0.71	0.85
Phe	0.97	0.60	0.91	0.89	0.98
Ile	0.82	0.56	0.91	0.71	0.81
His	0.85	0.50	0.91	0.83	0.89
Gly	1.00	0.62	0.91	0.76	1.00
Leu	0.99	0.65	0.90	0.95	1.00
Met	0.81	0.66	0.91	0.83	0.85
Asn	0.87	0.69	0.90	0.87	0.83
Lys	0.82	0.52	0.91	0.70	0.80
Arg	0.95	0.54	0.91	0.88	0.95
Pro	0.97	0.50	0.91	0.96	1.00
Ser	0.86	0.72	0.91	0.72	0.82
Gln	0.99	0.50	0.91	0.95	0.97
Thr	0.93	0.65	0.90	0.97	0.98
Val	0.93	0.50	0.91	0.75	0.85
Trp	1.00	0.65	0.89	0.96	1.00
Tyr	0.93	0.58	0.91	0.73	0.89

DISCUSSION

In the present study, association between the properties of HIV-1 env glycoprotein in Iran and five nearby countries are studied. According to literature, the most variations in HIV-1 are related to the env glycoproteins gp41 and gp120 sequences [18, 19]. In the large number of cases, phylogenetic clustering of HIV-1 isolates is based on the differences in env genes nucleotide sequences [20, 21]. HIV-1 env proteins of different subtypes and sub-subtypes can vary in more than 30% of their amino acids [22, 23]. In recent decades, several phylogenetic classifications are proposed on the HIV-1 env glycoprotein in Asian, African, European and American countries. In 2006, Ahn and Son reported that codon usage patterns among the HIV-1 env proteins of different subtypes may be a useful method to predict the evolutionary patterns of pandemic viruses [10]. In 2007, Singh and Seth, analyzed amino acid sequences of HIV env Protein by means of Clustal X software and found the association between the sequences from different Asian countries [11]. Our results also demonstrate that amino acid composition and four physical and chemical properties of HIV-1 env protein in Iran and Russia were less than 0.6. According to literature when the ACC values are less than 0.6, it means that differences between positive and negative classes of data are not significant [24]. Therefore, viral env proteins in Iran and Russia were not significantly different. Physicochemical properties between Iran and four different countries Turkey, Afghanistan, Pakistan and were significantly different. The result also showed that subtype A is dominant in Iran and Russia. Some researchers have reported that subtype A can circulate at high rate among intravenous drug users and the most probable way for HIV-1 subtype A introduction into Iran is through Former Soviet Union countries[25,26,27]. These observations indicated that in silico properties of HIV-1 env protein in Iran and Russia are similar.

Acknowledgments: This study was supported by the University of Isfahan

Conflict of Interest: The authors declare that they have no competing interest.

REFERENCES

1. Kandathil A, Ramalingam S, Kannangai R, David S, Sridharan G. Molecular epidemiology of HIV. Indian J Med Res 2005;121:333-344.
2. Potter SJ, Chew CB, Steain M, Dwyer DE, Saksena NK. Obstacles to successful antiretroviral treatment of HIV-1 infection: problems & perspectives. Indian J Med Res 2004;119:217-237.
3. Sakoda T, Kasahara N, Hamamori Y, Kedes L. A high-titer lentiviral production system mediates efficient transduction of differentiated cells including beating cardiac myocytes. J Mol Cell Cardiol 1999;31:2037-2047.

4. Spira S, Wainberg MA, Loemba H, Turner D, Brenner BG. Impact of clade diversity on HIV-1 virulence, antiretroviral drug sensitivity and drug resistance. J Antimicrob Chemother 2003;51:229-240.
5. Roy CN, Khandaker I, Oshitani H. Intersubtype genetic variation of HIV-1 Tat exon 1. AIDS Res Hum Retroviruses 2015;31:641-648.
6. Khan S, Rai MA, Khanani MR, Khan MN, Ali SH. HIV-1 subtype A infection in a community of intravenous drug users in Pakistan. BMC Infect Dis 2006;6:164.
7. Berg RK, Melchjorsen J, Rintahaka J, Diget E, Søby S, Horan KA. Genomic HIV RNA induces innate immune responses through RIG-I-dependent sensing of secondary-structured RNA. PloS one 2012;7:e29291.
8. Korkut A, Hendrickson WA. Structural plasticity and conformational transitions of HIV envelope glycoprotein gp120. PloS one 2012;7:e52170.
9. Hoffer LJ. How much protein do parenteral amino acid mixtures provide? Am J Clin Nutr 2011;94:1396-1398.
10. Ahn I, Son HS. Epidemiological comparisons of codon usage patterns among HIV-1 isolates from Asia, Europe, Africa and the Americas. Exp Mol Med 2006;38: 643-651.
11. Singh S, Gupta SK, Gupta MK, Seth PP. Phylogenetic analysis of HIV-1 envelope glycoprotein in Asian countries. Proceeding of International Conference on Applied Bioenginering 2007;200-204.
12. Pineda-Pena AC, Faria NR, Imbrechts S, Libin P, Abecasis AB, Deforche K. Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: Performance evaluation of the new REGA version 3 and seven other tools. Infect Genet Evol 2013;19:337-348.
13. Gasteiger E, Hoogland C, Gattiker A, Wilkins MR, Appel RD, Bairoch A. Protein identification and analysis tools on the ExPASy server. The proteomics protocols handbook: Springer; 2005; pp:571-607.
14. Kumar M, Thakur V, Raghava GP. COPid: composition based protein identification. In Silico Biol 2008;8:121-128.
15. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett 2006;27:861-874.
16. Mikut R, Reischl M. Data mining tools. Data Min Knowl Discov 2011;1:431-443.
17. Avupati VR, Kurre PN, Bagadi SR, Muthyala MK, Yejella RP. De novo based ligand generation and docking studies of PPARδ agonists: Correlations between predicted biological activity vs. biopharmaceutical descriptors. Chem-Bio Informatics J 2010;10:74-86.
18. Gaschen B, Taylor J, Yusim K, Foley B, Gao F, Lang D. Diversity considerations in HIV-1 vaccine selection. Science 2002;296:2354-2360.
19. Yuan T, Li J, Zhang MY. HIV-1 envelope glycoprotein variable loops are indispensable for envelope structural integrity and virus entry. PLoS One 2013 1;8:e69789.
20. Robertson D, Anderson J, Bradac J, Carr J, Foley B, Funkhouser R. HIV-1 nomenclature proposal. Science 2000;288:55-56.

21. Suslov KV. AID-mediated somatic hypermutation for generation of viral envelope protein diversity in patient-specific therapeutic HIV vaccines based on induction of neutralizing antibodies. Immunol Lett 2010;128:86-87.
22. Amogne W, Bontell I, Grossmann S, Aderaye G, Lindquist L, Sönnerborg A, Neogi U. Phylogenetic analysis of Ethiopian HIV-1 subtype C near full-length genomes reveals high intrasubtype diversity and a strong geographical cluster. AIDS Res Hum Retroviruses 2016;32:471-474.
23. Becker ML, DE Jager G, Becker WB. Analysis of partial gag and env gene sequences of HIV type 1 strains from southern Africa. AIDS Res Hum Retroviruses 1995;11:1265-1267.
24. Zhu W, Zeng N, Wang N. Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations. NESUG proceedings: health care and life sciences, Baltimore, Maryland. 2010.
25. Bobkov A, Kazennova E, Selimoval L, Bobkova M, Khanina T, Ladnaya N, Karvchenko A, Pokrovsky V, Cheingsong-Popov R, Weber J. A sudden epidemic of HIV type 1 among injecting drug users in the former Soviet Union: Identification of subtype A, subtype B, and novel gag A/env B recombinants. AIDS Res Hum Retroviruses 1998;14:669-676.
26. Nabatov AA, Kravchenko ON, Lyulchuk MG, Shcherbinskaya AM, Lukashov VV. Simultaneous introduction of HIV type 1 subtype A and B viruses into injecting drug users in southern Ukraine at the beginning of the epidemic in the former Soviet Union. AIDS Res Hum Retroviruses 2002;18:891-889.
27. Sarrami-Forooshani R, Das SR, Sabahi F, Adeli A, Esmaeili R, Wahren B, Molecular analysis and phylogenetic characterization of HIV in Iran. J Med Virol 2006;78:853-863.