International Research Journal of Biological Sciences ISSN 2278-3202
Vol. **1(7),** 52-54, November **(2012)** *I. Res. J. Biological Sci.*

Short Communication

In-Silico Identification of New Genes in HIV-1 by ORF Prediction Method

Dwivedi Vivek Dhar[1], Pandey Amit[2] and Mishra Sarad Kumar[3]
[1]Department of Bioinformatics, UCST, Dehradun, INDIA
[2]Forest Pathology Division, FRI, Dehradun, INDIA
[3]Department of Biotechnology, DDU Gorakhpur University, Gorakhpur, INDIA

Available online at: www.isca.in
Received 5th September 2012, revised 21st October 2012, accepted 25th October 2012

Abstract

In the present study the complete genome sequence of Human Immunodeficiency Virus-1 was searched and retrieved from Genbank database with accession number NC_001802.1. The sequence of HIV-1 applied for gene Identification by ORFs prediction method. ORF Finder revealed that total 39 ORFs were identified in all six possible reading frames (RF). Three ORFs were identified in +1 RF, four in +2RF, six in +3RF, six in -1RF, thirteen in -2RF, and seven in -3RF. The translated sequences of all identified ORFs were searched in NCBI nonredundant database using Protein BLAST. All ORFs found in +1, +2, +3RF, and ORF_4 in -1RF, ORF_{10} in -2RF, ORF_2 and ORF_3 in -3RF were already identified by previous workers but the remaining ORFs in -1,-2, and -3 RFs, eleven different ORFs translations showed similarity with different protein sequences of different organisms in nonredundant protein sequence database.

Keywords: HIV-1, genome analysis, proteins, ORFs, reading frames.

Introduction

HIV is a member of the genus *Lentivirus*, part of the family of Retroviridae[1]. Lentiviruses have many morphological and biological properties in common. Many species are infected by lentiviruses, which are characteristically responsible for long-duration illnesses with a long incubation period[2]. Lentiviruses are transmitted as single-stranded, positive-sense, enveloped RNA viruses. Upon entry into the target cell, the viral RNA genome is reverse transcribed into double-stranded DNA by a virally encoded reverse transcriptase that is transported along with the viral genome in the virus particle. The resulting viral DNA is then imported into the cell nucleus and integrated into the cellular DNA by a virally encoded integrase and host co-factors [3]. Once integrated, the virus may become latent, allowing the virus and its host cell to avoid detection by the immune system. Alternatively, the virus may be transcribed, producing new RNA genomes and viral proteins subsequently packaged and released from the cell as new virus particles again restarting the replication cycle. Two types of HIV have been characterized: HIV-1 and HIV-2. HIV-1 is the virus that was initially discovered and termed both LAV and HTLV-III [4, 5, 6]. HIV-1 is the most common and pathogenic strain of the virus[7]. Scientists divide HIV-1 into a major group (Group M) and two or more minor groups. Each group is believed to represent an independent transmission of SIV into humans. There are nine genes encoding 19 proteins already identified in HIV-1[8]. But the investigation of new genes in this virus indicates probable presence of other products hitherto unreported. The key to its evolution, identification and lineage related information may also lie in these regions including development of its clinical detection and management.

In the present study, we performed the *in silico* studies of complete genome sequence of HIV-1 to identify new genes and their relations with other sequences in the NCBI nonredundant database.

Material and Methods

The complete genome sequence of Human Immunodeficiency Virus-1 was searched and retrieved from Genbank database available at NCBI[9]. The complete genome sequence of this virus was subjected to gene prediction using ORF Finder tool[10]. All identified ORFs translations were searched in NCBI nonredundant database to find out the similarity with other sequences using protein BLAST program[11].

Results and Discussion

The complete genome sequence of HIV-1 was searched and retrieved (accession number **NC_001802.1**). The ORF Finder showed 39 ORFs in all six possible reading frames (RFs).

In the +1RF three ORFs were predicted: ORF_1 identified between position 3955 to 4122 was 168 bp long, ORF_2 between 5377 to 5595 was 219 bp long, and ORF_3 between 5608 to 5856 was 249 bp long.

In the +2RF four ORFs were predicted: ORF_1 identified between 797 to 955 was 159 bp long, ORF_2 identified between 5105 to 5341 was 237 bp long, ORF_3 identified between 5771 to 8341 was 2571 bp long, and ORF_4 identified between 5771 to 8341 was 2571 bp long.

In the +3RF six ORFs were predicted: ORF$_1$ identified between 336 to 1838 was 1503 bp long, ORF$_2$ identified between 4125 to 4226 was 102 bp long, ORF$_3$ identified between 4587 to 5165 was 579 bp long, ORF$_4$ identified between 5889 to 6023 was 135 bp long, ORF$_5$ identified between 8343 to 8714 was 372 bp long, and ORF$_6$ identified between 8859 to 8963 was 105 bp long.

In the -1RF six ORFs were predicted: ORF$_1$ identified between 767 to 895 was 129 bp long, ORF$_2$ identified between 917 to 1042 was 126 bp long, ORF$_3$ identified between 1895 to 2005 was 111 bp long, ORF4 identified between 2672 to 2800 was 129 bp long, ORF$_5$ identified between 5936 to 6085 was 150 bp long, and ORF$_6$ identified between 6563 to 6760 was 198 bp long.

In the -2RF thirteen ORFs were predicted: ORF$_1$ identified between 2071 to 2283 was 213 bp long, ORF$_2$ identified between 2365 to 2490 was 126 bp long, ORF$_3$ identified between 2527 to 2628 was 102 bp long, ORF$_4$ identified between 3121 to 3396 was 276 bp long, ORF$_5$ identified between 4276 to 4398 was 123 bp long, ORF$_6$ identified between 4453 to 4578 was 126 bp long, ORF$_7$ identified between 5887 to 6069 was 183 bp long, ORF$_8$ identified between 6109 to 6264 was 156 bp long, ORF$_9$ identified between 6541 to 6768 was 228 bp long, ORF$_{10}$ identified between 6919 to 7488 was 570 bp long, ORF$_{11}$ identified between 7681 to 7827 was 147 bp long, ORF$_{12}$ identified between 8182 to 8298 was 117 bp long, and ORF$_{13}$ identified between 8740 to 8913 was 174 bp long.

In the -3RF seven ORFs were predicted: ORF$_1$ identified between 636 to 767 was 132 bp long, ORF$_2$ identified between 2157 to 2261 was 105 bp long, ORF$_3$ identified between 2334 to 2438 was 105 bp long, ORF4 identified between 2443 to 3131 was 189 bp long, ORF$_5$ identified between 3495 to 3602 was 108 bp long, ORF$_6$ identified between 5571 to 5690 was 120 bp long, and ORF$_7$ identified between 8358 to 8462 was 105 bp long.

The translated sequences of all identified ORFs were searched in NCBI nonredundant database using Protein BLAST. As a result all ORFs found in +1, +2, +3RF, and *ORF$_4$ in -1RF*, ORF$_{10}$ in -2RF, ORF$_2$ and ORF$_3$ in -3RF were already identified and showed similarity with already defined protein of HIV-1, but in the remaining ORFs identified in -1,-2, and -3 RFs, eleven different ORFs showed similarity with different protein sequences in nonredundant database, and eleven ORFs did not show any significant similarity in the database. The all new identified ORFs along with their positions, length, reading frame numbers, accession numbers of similar proteins, and percentage similarity is listed in table-1.

Table -1
New ORFs identified in HIV-1 genome

RF Number	ORF Position	ORF Length		Similar Protein Accession no.	Maximum Identity
		Nucleotide	Amino Acid		
-1	767-895	129	42	XP_002774057.1	56%
-1	917-1042	126	41	No result found	-
-1	1895-2005	111	36	ABC46990.1	62%
-1	5936-6085	150	49	XP_002702453.1	54%
-1	6563-6760	198	65	EHK17473.1	41%
-2	2071-2283	213	70	ZP_08206794.1	42%
-2	2365-2490	126	41	No result found	-
-2	2527-2628	102	33	No result found	-
-2	3121-3396	276	91	YP_001236654.1	44%
-2	4276-4398	123	40	No result found	-
-2	4453-4578	126	41	No result found	-
-2	5887-6069	183	60	XP_001602651.1	46%
-2	6109-6264	156	51	No result found	-
-2	6541-6768	228	75	ZP_10281217.1	52%
-2	7681-7827	147	48	ZP_09669895.1	43%
-2	8182-8298	117	38	No result found	-
-2	8740-8913	174	57	XP_003476767.1	55%
-3	636-767	132	43	XP_002313615.1	61%
-3	2943-3131	189	62	No result found	-
-3	3495-3602	108	35	No result found	-
-3	5571-5690	120	39	No result found	-
-3	8358-8462	105	34	No result found	-

Conclusion

Sequence analysis of the genome of an organism leads to characterization and identification of the concerned organism. The identified ORFs may have bearing on the genotypic or phenotypic traits. All the genes identified in a genome do not participate in the formation of a protein. Some genes need a suitable environment for activation and expression. In HIV-1, 39 ORFs were identified by ORF prediction method. In 39 ORFs, 17 ORFs were already identified by previous workers. The remaining 22 ORFs may be suggested as genes because in 22 ORFs, 11 ORFs showed the similarity with other sequences of different organisms. It may be suggested that the sequences and functions of these ORFs have evolutionary relation with matching organisms in the databases. The remaining ORFs did not show evolutionary record in the databases, thus probably indicating their functions unique to this virus. The functional attributes of these ORFs should be verified in wet lab experimentation and analysis. This identification can significantly contribute in the understanding of the evolutionary and functional analysis of ORFs at molecular level. However, owing to the considerable importance of these ORFs, more contributions are warranted for its detailed investigation.

Acknowledgements

We are thankful to department of biotechnology DDU Gorakhpur University, Gorakhpur for providing laboratory facilities & encouragement and to director of UCST, Dehradun for their cooperation during the study. We are grateful to the Head, Forest Pathology Division, FRI, Dehradun for his kind support and necessary suggestions whenever needed.

References

1. Fauquet C.M. and Fargette D., International Committee on Taxonomy of Viruses and the 3,142 unassigned species, *Virol. J.*, **2**, 64 **(2005)**

2. Levy J.A., HIV pathogenesis and long-term survival, *AIDS* **7 (11)**, 1401–10 **(1993)**

3. Smith J.A., Daniel R., Following the path of the virus: the exploitation of host DNA repair mechanisms by retroviruse, *ACS Chem Biol,* **1 (4)**, 217–26 **(2006)**

4. Gilbert P.B., McKeague I.W., Eisen G., Mullins C., Gueye N.A., Mboup S., Kanki P.J., Comparison of HIV-1 and HIV-2 infectivity from a prospective cohort study in Senegal, *Statistics in Medicine,* **22 (4),** 573–593, **(2003)**

5. Barre
S. F., Chermann J.C., Rey F., Nugeyre M.T., Chamaret S. and Gruest J., Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome, *Science,* **220**, 868–70, **(1983)**

6. Clavel F., Mansinho K., Chamaret S., Guetard D., Favier V., Nina J, Human immunodeficiency virus type 2 infection associated with AIDS in West Africa, *N Engl J Med.*, **316**, 1180–5 **(1987)**

7. Weiss R.A., How does HIV cause AIDS?, *Science*, **260(5112)** 1273–9 **(1993)**

8. Zhao J., Tang S., Ragupathy V., Carr J.K., Wolfe N.D., Awazi B. and Hewlett I., Identification and genetic characterization of a novel CRF22_01A1 recombinant form of HIV type 1 in Cameroon, *AIDS Res Hum Retroviruses*, **26(9)**, 1033-45, **(2010)**

9. Dennis A., Benson I., Karsch M., David J., Lipman J.O., and David L.W., GenBank, *Nucleic Acids Res*, **36**(Database issue), D25–D30, **(2008)**

10. David L.W., Deanna M.C., Scott F., Alex E.L., Thomas L.M., Joan U.P., Gregory D., Schuler L. Schriml M., Edwin S., Tatiana A.T., and Lukas W., Database resources of the National Center for Biotechnology, *Nucleic Acids Res*, **1**, **31(1)**, 28–33 **(2003)**

11. Altschul S.F., Gish W., Miller W., Myers E.W. and Lipman D.J., Basic local alignment search tool, *J Mol Biol,* **215,** 403-10 **(1990)**