Applications note

REEvolution: A tool for epitope mapping and selective pressure analysis for HIV and HCV infections.

A.T.L. de Queiroz[1,3*], L.T. Rangel[2], S.R. Matioli[1] and I.M.V.G. de Carvalho-Mello[3].

[1] Departamento de Genética e Biologia Evolutiva, IB, Universidade de São Paulo, São Paulo, Brasil.

[2] Departamento de Parasitologia,ICB, Universidade de São Paulo, São Paulo, Brasil.

[3] Laboratório de Imunologia Viral, Instituto Butantan, São Paulo, Brasil.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: One of the problems in vaccine development is the identification of evolutionary stable regions of the microorganisms because variable regions on their protein coding regions may escape immunization by Darwinian selection. MHC (major histocompatibility complex) class I restricted $CD8^+$ T cells plays an important role in the control of viraemia in HIV (human immunodeficiency virus) and HCV (hepatitis C virus) infections and, therefore, they are the selective agents for the virus sequences. A web-based application tool has been designed to perform epitope mapping and to test for evidence of positive selection specifically for HIV and HCV protein coding sequences. Both data allow user to identify regions recognized by immune system under neutral evolution or positive selection. This application is an open source project that can be extended and modified for use with other targets.

1 INTRODUCTION

Cytotoxic T-Lymphocytes (CTL) eliminate virus infected cells by recognizing virus-derived peptides, called epitopes, presented by Human Leukocyte Antigen (HLA) class I molecules on the infected cell surface. The HLA-restricted CTL response is believed to play a major role in the immune control of HIV-1 and HCV infection. Nevertheless, the immune system does not eliminate the HIV infection and only 30% of HCV infected patients have viral clearance. Viral persistence is associated with viral evolution in response to immune selective pressure, rising escape mutations that allows specific virus lineages to evade immune recognition (Goulder and Watikins, 2008; Dustin and Rice, 2007).

Miyata and Yasunaga (1980) proposed a measure of selective pressure at the protein level called nonsynonymous-synonymous substitution rate ratio (dN/dS or ω). Positive selection is defined as dN/dS >1, which means that nonsynonymous mutations offers fitness advantage to the protein and have higher fixation probabilities than synonymous mutations, as well, dN/dS < or equal than 1 indicate negative purifying selection and neutral evolution.

The identification of epitopes in conserved regions or regions under neutral selection, as well early recognition of escape mutation can reveal more information about virus-host evolution and offers insights into the nature of HLA-epitope targeting, which could be applied in vaccine design (Brumme *et al*., 2009). Recently, we also demonstrated the utility of epitope mapping with evolution models in association with increase of immune response and the therapy outcome in HCV infection (de Queiróz *et al.,* 2010).

Here, we present a web based tool to perform HLA-I epitope mapping that also tests for evidence of positive selection in viral protein coding regions (HIV and HCV) to help the identification of neutral evolving epitopes, potential escape mutations or compensatory mutation. There are two required inputs from the user: The contact information (a valid e-mail address) and the DNA aligned sequence data, which must be provided in the FASTA format.

2 METHODS AND RESULTS

The tool was developed to analyze protein-coding sequence alignment from HIV-1 and HCV, and the analysis schedule is split into 3 steps:

1- *Epitope mapping*. After the server validates and retrieves the input, the multiple alignments are translated in first open reading frame and the consensus of all sequences is generated. After that, coordinates of consensus is acquired according the reference sequence (accession number: K03455 for HIV; M67463 for HCV). Only epitopes related to the coordinates are selected. All the epitopes were selected from *Los Alamos HIV Immunology Database and Compendia* (Kober *et al*., 2007) and *Immune Epitope Database* (Vita *et al*., 2010). These datasets were checked to eliminate duplicate entries. Epitope mapping is then performed, the frequencies are measured and mutations of one and two mismatches are also estimated.

2- *Test for evidence of positive selection*. To detect positive selection without averaging the dN/dS ratio throughout the phylogenetic tree five site models are used on query nucleotide alignment. The one ratio M0 codon-based ML substitution model

is used to measure nonsynonymous/synonymous substitution rate ratio(dN/dS) of entire alignment. Two likelihood models of neutral evolution (M1a and M7) and of positive Darwinian selection (M2a and M8) are used to verify evidence of positive selection by comparing the likelihood values of the different nested models using the Likelihood Ratio Test (LRT). The likelihood's models are measured, and LRT is performed. The p-value is obtained comparing to χ^2. All models used are from the PAML4 package (Yang 2007).

3- *Bayes Empirical Bayes(BEB) analysis*. If LRT suggests evidence of positive selection, the BEB analysis (Yang *et al*., 2005) is performed to infer which category the site most likely belongs to. This approach allows the identification of specific codons under positive selection by measuring the dN/dS ratio and its posterior probability.

REEvolution results will be emailed to the address given in the contact information, in a single file. The first part of file contains the dN/dS value from entire dataset, the results of model comparison with LRT. If the test suggests evidence of positive selection, the tool lists the amino acids under selection, with its positions relative to query, posterior probabilities and the dN/dS ratio with their standard errors. The second part presents the epitope mapping results. The tool measures the frequency of epitopes through the sequences from data and searches for epitopes with one or two amino acid changes. All information from database such the epitope amino acid sequence, the protein, reference location, host species, and HLA restriction elements are provided for posterior comparison. Moreover, the location of epitope in query is also provided. The file is in HTML format.

3 DISCUSSION

Several computation approaches have been used for identification of T-cell CD8 epitopes as an alternative to experimental assays. However, those prediction programs analyze only one sequence per run just for protein sequences (Lafuente and Reche, 2009), making the evolutionary information inaccessible. Our approach has the advantage of analyze multiple sequences from aligned input data and the employment of real described epitopes from virus specific databases, not by predicting them. Besides, the tool also provides the mutation frequencies of one and two amino acids in epitope region, allowing identification of possible escape mutation. However, the tool's epitope coverage is limited by the both database size. The multiple DNA sequence analysis enables the use of evolutionary models. Those models provide information about selective pressure of the entire query region and of specific amino acids under evolutionary constraints.

REEvolution is a user-friendly interface that enable researchers to identify real epitope regions under immune system selective pressure. The comprehensive result facilitates the identification of conserved epitopes between query sequences, mutation in epitope regions under positive selection or neutral evolution and compensatory fitness mutation elsewhere. This information is useful to identify CTL escape mutations in both viruses and helps the development of novel vaccines and therapeutic strategies for AIDS and Hepatitis C.

4 IMPLEMENTATION

REEvolution is written in Python language and uses a web interface developed with CherryPy to upload data and process the output. The software main page can be found online at http://bioinfo.ib.usp.br/revolver/. Source code, examples, and an comprehensive guide can also be found on this website.

ACKNOWLEDGEMENTS

We thank to Vinicius Maracajá-Coutinho, André Yoshiaki Kashiwabara and Alexandre Rossi Paschoal for comments and useful suggestions.

Funding: Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq.

Conflict of Interest: none declared.

REFERENCES

Brumme Z.L. e*t al.*(2009) HLA-associated immune escape pathways in HIV-1 Subtype B Gag, Pol and Nef proteins. *Plos One*. 4, 8, e6687.

de Queiroz *et al*. (2010) Relation of Pretreatment Sequence Diversity in NS5A Region of HCV genotype 1 With Immune Response Between pegylated-INF/ribavirin Therapy Outcomes. *J Viral Hepat.* Epub ahead of print.

Dustin, L.B., Rice C.M.(2007) Flying under the radar: the immunobiology of hepatitis C. *Annu Rev Immunol.,* 25, 71–99.

Goulder, P.J., Watkins, D.I. (2008) Impact of MHC class I diversity on immune control of immunodeficiency virus replication. Nat Rev Immunol., 8, 619–630.

Korber, B.T.M. *et al*. (2007) HIV Molecular Immunology 2006/2007. *Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, New Mexico.* LA-UR 07-4752.

Lafuente E.M., Reche P.A. (2009) Prediction of MHC-peptide binding: a systematic and comprehensive overview. *Curr Pharm Des.*15,28, 3209-20.

Miyata T., Yasunaga T. (1980) Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *J Mol Evol.* 16, 1, 23-36.

Vita R., *et al* (2010) The immune epitope database 2.0. *Nucleic Acids Res*. 38, D854-62.

Yang Z. et al.(2005) Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol.* 22,1107–1118.

Yang Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol,* 24(8),1586-91.