SOFTWARE TOOL ARTICLE

Pathogen Sequence Signature Analysis (PSSA): A software tool for analyzing sequences to identify microorganism genotypes [version 1; peer review: 2 approved with reservations]

Karina Salvatierra ¹, Hector Florez ²

¹Faculty of Exact, Chemical and Natural Sciences, Universidad Nacional de Misiones, Posadas, Argentina

²Faculty of Technology, Universidad Distrital Francisco José de Caldas, Bogotá, Colombia

V1 First published: 09 Jan 2017, 6:21
<https://doi.org/10.12688/f1000research.10393.1>

Latest published: 09 Jan 2017, 6:21
<https://doi.org/10.12688/f1000research.10393.1>

Abstract

Introduction

The chikungunya virus (CHIKV) is an arbovirus vectored by *Aedes* mosquitoes that infects humans in tropical and sub-tropical areas of Asia and Africa. Recently, outbreaks have been reported in tropical and sub-tropical areas of countries that were previously unaffected (e.g., Brazil, Colombia). Currently, the following geographical genotypes have been identified through phylogenetic analysis of CHIKV E1 gene sequences: the West African (Waf), East/Central/South African (ECSA), and Asian genotypes. Outbreaks in a geographical area can happen with the same or different genotypes. Determining which genotypes are circulating in an outbreak is important for public health management.

Objectives

To create a computer-based system available online that is suitable for detecting changes in CHIKV nucleotide and amino acid sequences and identifying their corresponding geographical genotype.

Methods

We used several computer frameworks, tools, programming languages, algorithms, and infrastructure systems to build a software tool that analyzes changes in nucleotide and amino acid sequences and identifies different geographical genotypes through phylogenetic analysis.

Results

We have built an online software tool called Pathogen Sequence Signature Analysis (PSSA) that allows researchers to analyze nucleotide and amino acid sequence variations between sample CHIKV sequences taken from infected patients and obtained through conventional Sanger sequencing, to identify their corresponding geographical genotype.

Conclusion

PSSA is able to analyze sequences in a simple and effective manner,

Open Peer Review

Approval Status ? ?

	1	2
version 1		
09 Jan 2017	view	view

1. **Easwaran Sreekumar**, Rajiv Gandhi Centre for Biotechnology, Thiruvananthapuram, India

2. **Massimo Ciccozzi**, Istituto Superiore di Sanità, Rome, Italy

Any reports and responses or comments on the article can be found at the end of the article.

and includes proper documentation (i.e., UML diagrams) and also basic examples that serve to test the algorithm. Furthermore, PSSA provides various ways to visualize the data in order to aid understanding and interpretation of results.

Results provided by PSSA will be useful for the identification of circulating CHIKV genotypes and public health surveillance. PSSA is available at: <http://pssa.itiud.org>.

Keywords

Chikungunya virus, Public health, sequences, information system, phylogenetic analysis.

This article is included in the **Neglected Tropical Diseases** collection.

Corresponding author: Hector Florez (haflorezf@udistrital.edu.co)

Competing interests: No competing interests were disclosed.

Grant information: The work presented in this paper has been supported by the Information Technologies Innovation (ITI) Research Group.

Copyright: © 2017 Salvatierra K and Florez H. This is an open access article distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the **Creative Commons Zero "No rights reserved" data waiver** (CC0 1.0 Public domain dedication).

How to cite this article: Salvatierra K and Florez H. **Pathogen Sequence Signature Analysis (PSSA): A software tool for analyzing sequences to identify microorganism genotypes [version 1; peer review: 2 approved with reservations]** F1000Research 2017, 6:21 <https://doi.org/10.12688/f1000research.10393.1>

First published: 09 Jan 2017, 6:21 <https://doi.org/10.12688/f1000research.10393.1>

Introduction

Chikungunya virus (CHIKV) is an arbovirus (arthropod-borne virus), which is part of the Alphavirus genus and belongs to the Togaviridae family. It is vectored by *Aedes* mosquitoes and infects humans in tropical and sub-tropical areas of Asia and Africa¹. CHIKV has a positive-sense, single-stranded RNA genome of 12 kb, which can persist for years in humans. Symptoms include rash and febrile illness associated with severe arthralgia².

Currently, the following geographical genotypes have been identified through phylogenetic analysis of CHIKV E1 gene sequences: the West African (Waf), East/Central/South African (ECSA), and Asian genotypes³⁻⁵. However, most CHIKV phylogenetic studies have just used fragmented sequences from the glycoprotein envelope of the E1 gene, which avoids accurate assessments regarding the relations between strains and their evolutionary dynamics.

Recently, some complete sequences from the CHIKV genome have been made available, so we have used the available data for a complete E1 gene to develop an automated computational algorithm that can be used for accurate and rapid identification of this pathogen.

The online software tool that we have created, called Pathogen Sequence Signature Analysis (PSSA), will allow researchers to analyze nucleotide and amino acid sequence variations between sample CHIKV sequences taken from infected patients, and determine the corresponding genotype from phylogenetic analysis of the results. PSSA also provides various ways to visualize the data in order to aid understanding and interpretation of results.

Methods

Implementation

To build PSSA, we used standard computer-based tools, programming languages, and infrastructure systems. PSSA is based on the Object Oriented Paradigm; thus, for its design, we used the Unified Modelling Language (UML)⁶. For its development, we used version 7.1.0 of the PHP language (<https://www.php.net/>) supported by the application server Apache version 2.4.23 (<http://www.apache.org/>). PSSA's front end was developed based on version 3.3.7 of Bootstrap (<http://getbootstrap.com/>). Using Bootstrap is very convenient for this project because it is a framework that properly integrates JavaScript, CSS, and HTML for creating responsive web applications. PSSA uses the version 3.1.1 of the library JQuery (<https://www.jquery.com/>) for facilitating the use of JavaScript functionalities. After PSSA performs an analysis, it provides results in several formats. One result corresponds to an automatically generated report in pdf format, created based on version 0.0.8 of the PHP library ezpdf (<https://github.com/rebuy-de/ezpdf>). The other results are a force-directed graph, a radial tree, and a cartesian tree, which were developed supported by version 3.5.17 of the online JavaScript library called Data-Driven Documents, known as D3 (<https://www.d3js.org/>). D3 provides services for deploying data via interactive visualizations.

The geographical genotype of a sample sequence was determined based on well-defined phylogenetic clusters whose origins have been linked to a given geographic region. We analyzed all available whole genomes in GenBank database (www.ncbi.nlm.nih.gov/)

(www.ncbi.nlm.nih.gov/genbank/). However, since the E1 gene has been previously used in several studies, including Nunes *et al.*⁷, Laiton-Donat *et al.*⁸, and Volk *et al.*⁹, to determine the genotype of sample sequences, we decided to use the E1 gene. In addition, we performed extensive testing to be sure that our reference strains can accurately classify other sequences.

PSSA stores all information related to nucleotide and amino acid analysis in one relational database. In this project we created said database using version 5.7.16 of the MySQL community server (<http://dev.mysql.com/downloads/>). The database was designed to handle the required information of the proposed nucleotide and amino acid analysis, as well as phylogenetic analyses. The database is managed using version 4.6.5.2 of phpMyAdmin (<https://www.phpmyadmin.net/>), which is a project that serves to administrate databases that use MySQL. MySQL community also offers MySQL-Workbench (<https://www.mysql.com/products/workbench/>), of which version 6.3 was used to design PSSA's database.

In addition, version 4.1 of the Integrated Development Environment (IDE) EclipsePHP (<https://eclipse.org/pdt/>) was used to develop PSSA. EclipsePHP allows for creation of PHP-based projects supporting PHP, CSS, and JavaScript languages. In addition, EclipsePHP provides git services for storing projects in desired repositories. Thus, we started the development of PSSA by creating a PHP project in EclipsePHP; next, we created all required files and wrote the source code for the algorithms that performed the desired analyses; and finally, we created a git configuration to store the project in the GitHub repository system. To host PSSA, the SUSE Linux Enterprise Server 12 SP2 (<https://www.suse.com>) was used. This server includes all applications mentioned above that are necessary for PSSA to operate correctly.

The various libraries, frameworks, and software we used to develop PSSA are all under the GNU General Public License (<http://www.gnu.org/licenses/licenses.en.html>). This means that only free software was used to develop our software tool.

As reference sequences we used ECSA genotype HM045811-Ross, Asian genotype HM045810, and West African genotype HM045807, searching through all nucleotide sequences of the CHIKV E1 gene that are available on the GenBank database (www.ncbi.nlm.nih.gov/genbank/). (the first isolated identified of three genotypes).

The accession numbers for the representative or alternative CHIKV E1 sequences used in the phylogenetic analysis are as follows:

ECSA genotype: HM045823, AM258993, EF012359, AM258991, AB455494, GU199352, FJ445426, FN295485, GU301781, HM045784, HM045822, KP164568, KP164570, KP164569.

Asian genotype: HM045813, HM045800, HM045790, HM045789, EF027140, EF027141, FN295483, L37661, EF452493, FJ807897, HE806461, KF318729, KJ451622, AB860301, KP164567, KP164572, KP164571, KJ451624, KP851709, KT211035, KT211049.

West African genotype: HM045816, HM045785, HM045815, HM045818, AY726732, HM045820, HM045817.

Operation

PSSA has been developed to be run in with Google Chrome, Mozilla Firefox, Internet Explorer, and Safari; nevertheless, PSSA might run in other browsers such as Opera. To run PSSA the URL <http://pssa.itiud.org> must be typed in the web browser.

However, if the user wants to run PSSA locally, the following steps need to be followed:

1. Download PSSA from the github repository: <https://github.com/florezfernandez/pssa>.
2. Install the local server the software: Apache, PHP, MySql, and PHPMyAdmin (optional).
3. Run the database script, which is available when the project is restored from the repository.
4. The project contains a file called "connection.php" in which the information regarding the connection to the

database is configured. Update the information of the server, database, database user, and database password with the information of the local server. The default values provided with PSSA are: server = localhost, database = pssa, database user = "root", and database password = "".

Results

The most important feature of PSSA is that the algorithm has been developed to analyze sequences taken from multiple patients. Several sequences can also be submitted per patient. The analysis process is carried out via the following steps:

1. Once the user has accessed PSSA by using the corresponding web address, they must access the "Sequence Analysis" menu, in which the user can select the menu item "Chikungunya Virus". PSSA then presents the name, description, reference and alternative sequences of available gene(s) for CHIKV (Figure 1). For each gene (e.g., E1), two icons appear on the right side. The first icon deploys reference sequences while the second icon the alternative sequences.

Sequence Analysis ▾ About PSSA

Pathogen Sequence Signature Analysis

Chikungunya

Chikungunya virus (CHIKV) is an arbovirus (arthropod-borne virus) which is a member of the Alphavirus genus belonging to the Togaviridae family, vectored by Aedes mosquitoes to humans in tropical and subtropical regions of Africa and Asia. CHIKV has a single-stranded, positive-sense RNA genome of 12 kb and causes chikungunya fever (CHIK), a febrile illness associated with severe arthralgia and rash; the name is derived from a Bantu language word describing the severe arthritissigns, which can persist for years. CHIKV is transmitted to humans mainly by Aedes (Ae.) aegypti and Ae. albopictus mosquitoes. The phylogenetic analysis of CHIKV based on E1 gene sequences identified three geographically associated genotypes: the West African (Waf), East/Central/South African (ECSA), and Asian genotypes. More recent analyses indicate that the recent Indian Ocean and Indian strains form a monophyletic group within the ECSA lineage. However, most CHIKV phylogenetic studies have utilized only partial sequences from the envelope glycoprotein E1 gene, preventing a robust assessment of some of the relationships among strains and of their evolutionary dynamics.

Name	Description
E1	The viral spike proteins facilitate attachment to cell surfaces and viral entry into the cells. The E1 envelope protein is a class II fusion protein that mediates low pH-triggered membrane fusion during virus infection.

1 registries

ITI Research Group
© 2016 All rights reserved

Figure 1. Screenshot of the PSSA Sequence Analysis menu presenting the available name, description and two icons for the CHIKV E1 gene. The first icon represents the reference sequence, while the second icon represents the alternative sequences of the CHIKV E1 gene. By selecting the reference sequences, users proceed with the *mutation analysis*, and by selecting alternative sequences users proceed with the *phylogenetic analysis*.

- There are two different types of analysis in PSSA. By selecting the reference sequences, users proceed with the *mutation analysis*, which analyzes nucleotide and amino acid changes in patient sequences, and by selecting alternative sequences users proceed with the *phylogenetic analysis*, which establishes the phylogenetic relationship between the submitted sequences and determines which genotype they belong to.
- Once the type of analysis has been selected, the user can provide the patient sequences through *FASTA* files. PSSA includes an example dataset that can be used to test the system. The symbol '-' can be included in desired sequences in order to specify possible missing data; but tabs, blank spaces, and any other symbols than the ones used in these kind of sequences (i.e., A, C, T, G) are not accepted.

After patient sequences have been provided, the analysis algorithm is run and the system presents the corresponding results.

Mutation analysis

For the mutation analysis the system provides an online report that includes the nucleotide and corresponding amino acid changes in each patient's sequence (Figure 2). The report can be sent to the user via e-mail in pdf format, and contains both a summary of the results and complete details of the analysis. It also provides a force-directed graph which presents each sequence as a node and where the set of nodes deployed using the same color represents one patient (Figure 3). In addition, nodes that belong to each patient are clustered based on the number of nucleotide and amino acid changes. When a sequence contains a substantial part of the E1 gene, these results are reliable and can be used for further purposes. Users might confirm that results are reliable by reading the pdf report and comparing it to the force-directed graph.

Phylogenetic analysis

For the phylogenetic analysis, the system presents results as a radial tree (Figure 4) and a cartesian tree (Figure 5) to establish the phylogenetic relationship between sequences and determine which genotype they belong to, based on an array of alternative sequences corresponding to the E1 gene.

Figure 2. Online textual visualization of the online report for the mutation analysis that about of presents theing nucleotide changes that produce an amino acid change.

Figure 3. Force-directed graph visualization. Force-directed graph presents each sequence as a node and each set of nodes of the same color as one patient. In addition, nodes that belong to each patient are clustered based on the number of nucleotide and amino acid changes.

Figure 4. Radial tree visualization. The three genotypes are separated into the different branches, where each branch corresponds to one sequence obtained from the GenBank database.

Pathogen Sequence Signature Analysis

Phylogenetic Tree

Figure 5. Cartesian tree visualization. It presents the same information as the radial tree, but it shows the three genotypes separated into the different branches more clearly.

The algorithm is an iterative process. Thus, for each patient file, all sequences are collected by the algorithm; then, for each sequence, some instructions of the algorithm are used to compare the iterated sequence to the selected reference sequence as well as to the alternative sequences. All information regarding the analysis is stored and used to generate the reports through the visualizations described above. It is important to mention that there are two different types of analysis in PSSA, even though they are both closely related. The mutation analysis presents results as a report of the nucleotide and amino acid variations in each patient's sequence and a force-directed graph, whilst the phylogenetic analysis is based on the nucleotide substitution model and results are presented as a radial and cartesian tree to establish the phylogenetic

relationship between sequences and determine which genotype they belong to.

Conclusions

PSSA is an online software tool that provides an automated computational algorithm that guarantees accurate and reliable detection of nucleotide and amino acid sequence variations and provides various ways to visualize the data in order to aid understanding and interpretation of results.

PSSA is different to BMA¹⁰, which is another analysis tool developed in our research group, because it not only provides information regarding nucleotides and amino acid changes, but it

also compares the sequences with multiple alternative sequences to identify the genotype in a phylogenetic tree.

PSSA will be useful for the identification of circulating CHIKV genotypes in an outbreak and public health surveillance. It is a flexible tool, which implies that it could be used for evaluating other microorganisms, such as bacteria (e.g., *Mycobacterium tuberculosis*), parasites (e.g., *Leishmania*) or other viruses (e.g., Dengue, Zika).

Software availability

Software available from: <http://pssa.itiud.org>

Latest source code: <https://github.com/florezfernandez/pssa>

Archived source code as at the time of publication:

<http://dx.doi.org/10.5281/zenodo.179922>¹¹

License: GNU General Public License (GPL)

Author contributions

KS performed the literature review and drafted the manuscript. HF designed and developed the PSSA software and helped draft the manuscript.

Competing interests

No competing interests were disclosed.

Grant information

The work presented in this paper has been supported by the Information Technologies Innovation (ITI) Research Group.

Acknowledgments

The authors would like to thank Professor Jorge E. Osorio, Department of Pathobiological Sciences, University of Wisconsin-Madison (USA), for his collaboration in the project.

References

- Robinson MC: **An epidemic of virus disease in Southern Province, Tanganyika Territory, in 1952–53. I. Clinical features.** *Trans R Soc Trop Med Hyg.* 1955; **49**(1): 28–32.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Johnston RE, Peters CJ: **Alpha viruses associated primarily with fever and polyarthrititis.** In: Fields BN, Knipe DM, Howley PM (Eds.), *Field Virology*. Lippincott-Raven Publishers, Philadelphia, 1996; 843–898.
- Powers AM, Brault AC, Tesh RB, *et al.*: **Re-emergence of Chikungunya and O'nyong-nyong viruses: evidence for distinct geographical lineages and distant evolutionary relationships.** *J Gen Virol.* 2000; **81**(Pt 2): 471–479.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Schuffenecker I, Itman I, Michault A, *et al.*: **Genome microevolution of Chikungunya viruses causing the Indian Ocean outbreak.** *PLoS Med.* 2006; **3**(7): e263.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Powers AM, Logue CH: **Changing patterns of Chikungunya virus: re-emergence of a zoonotic arbovirus.** *J Gen Virol.* 2007; **88**(Pt 9): 2363–2377.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Rumbaugh J, Jacobson I, Booch G: **The Unified Modeling Language Reference Manual.** Pearson Higher Education. 2004.
[Reference Source](#)
- Nunes MR, Faria NR, de Vasconcelos JM, *et al.*: **Emergence and potential for spread of Chikungunya virus in Brazil.** *BMC Med.* 2015; **13**: 102.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Laiton-Donat K, Usme-Ciro JA, Rico A, *et al.*: **Análisis filogenético del virus del chikungunya en Colombia: evidencia de selección purificadora en el gen E1.** *Biomedica.* 2016; **36**(Supl.2): 25–34.
[Publisher Full Text](#)
- Volk SM, Chen R, Tsetsarkin KA, *et al.*: **Genome-scale phylogenetic analyses of chikungunya virus reveal independent emergences of recent epidemics and various evolutionary rates.** *J Virol.* 2010; **84**(13): 6497–650.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Salvatierra K, Florez H: **Revised Biomedical Mutation Analysis (BMA): A software tool for analyzing mutations associated with antiviral resistance [version 2; referees: 2 approved].** *F1000Res.* 2016; **5**: 1141.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Salvatierra K, Florez H: **PSSA.** *Zenodo.* 2016.
[Data Source](#)

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 15 May 2017

<https://doi.org/10.5256/f1000research.11199.r22414>

© 2017 Ciccozzi M. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Massimo Ciccozzi

Department of Infectious, Parasitic and Immunomediated Diseases, Istituto Superiore di Sanità, Rome, Italy

Salvatierra & Florez describe the development of software tool and web interface for analyzing only Chikungunya sequences.

The idea is very interesting but I have some different concerns about the utility of its application.

1. It is not well documented that Chikungunya virus can persist in human for many time
2. In the title it must be underline that the system is to identify Chikungunya virus only and maybe in the text (discussion section) the eventual possibility to expand it
3. The authors have to better describe the requirements about the sequences used in this tool
4. It is important in phylogenetic analysis to identify the algorithm used but no mention has been made in the article, no model choose in a case of maximum likelihood algorithm and so on

I think that in this form without detailed informations for users it is not possible to accept. After major revision, this can be a useful and detailed guide.

Is the rationale for developing the new software tool clearly explained?

Partly

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

No

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Evolutionary analysis and molecular epidemiology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 18 April 2017

<https://doi.org/10.5256/f1000research.11199.r21444>

© 2017 Sreekumar E. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Easwaran Sreekumar

Molecular Virology Laboratory, Rajiv Gandhi Centre for Biotechnology, Thiruvananthapuram, Kerala, India

The article by Salvatierra & Florez describes development of software tool and web interface for analyzing sequences of microorganisms. Essentially, the software is currently configured for analyzing only chikungunya sequences.

The reviewer has a number of comments/ clarifications to make:

1. Introduction: There are no confirmed evidence that the virus can persist in humans for years, as claimed by the authors "CHIKV has a positive-sense, single-stranded RNA genome of 12 kb, which can persist for years in humans". It might be possible that the reviewer has missed such reports in the literature; so a reference citation is required to support this point.
2. Reviewer feels that since the reference data included in the software is only suitable for Chikungunya virus, the authors should refrain from making a broad claim in the title of the manuscript that it can be used for other microorganisms
3. The reviewer tried to use the software to analyze two input sequences of differing length.

The user interface simply prompts 'incorrect sequence length', without giving any clue that the input sequence should be of equal length (which I 'guessed' from the test data set).

4. There are no readily accessible user guidelines (help) or link given in the web interface so that the user can do trouble shooting easily. What happens for the mutation analysis if a user gives out of frame sequences? Will it again simply prompt that 'incorrect sequence length'?
5. The manuscript does not describe the requirements for the input sequence- minimum length, maximum length, reading frame etc. except that it should be in FASTA format.
6. What is the exact algorithm used in the phylogenetic tree. Does it use Neighbor joining method or Maximum Likelihood analysis, or any other methods? What are the default settings of the 'some instructions' of the algorithms and nucleotide substitution model to compare the iterated sequences?
7. Is there any provision to do on screen editing of the sequence (I think, no) or is it that each time one need to input an edited sequence file?
8. It provides on screen outputs. Is there any way to save these outputs, and in which format?

The reviewer feels that the software does not give added advantage over many of the stand alone, free programs such as BioEdit or MEGA, unless it has more user friendly features. It provides a ready reference for a small set of known CHIKV Genotypes which would be useful for a newcomer in the field to identify the genotypes. But even for a little more advanced user, the interface has no features for a customizable analysis.

Is the rationale for developing the new software tool clearly explained?

Partly

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

No

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Host-pathogen interaction, virus evolution, Chikungunya, Dengue

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research