BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Software

Volume 10(8)

GSIT: An integrated web-tool for identification of genomic signatures in highly similar DNA sequences

Amit Tuteja[1,2†*], Kandarp Joshi[1,3†], Swati Subodh[1,4] & Navkiran Kaur[2]

[1]Institute of Molecular Medicine, New Delhi; [2]Amity Institute of Biotechnology, Amity University, Noida, India; [3]Faculty of Technology and Engineering, The Maharaja Sayajirao University of Baroda, Vadodara; [4]Council of Scientific and Industrial Research-Open Source Drug Discovery Unit, New Delhi; Amit Tuteja - Email: amittuteja1981@gmail.com; Phone: +91-9899748528; *Corresponding author
† Authors contributed equally

Received August 06, 2014; Revised August 16, 2014; Accepted August 16, 2014; Published August 30, 2014

Abstract:
Accurate identification and characterization of infectious agent and its subtype is essential for efficient treatment of infectious diseases on a target population of patients. Comparative biology of microbial populations *in vitro* and *in vivo* can identify signatures that may be used to develop and improve diagnostic procedures. Here we report Genomic Signature Identification Tool (GSIT) a web based tool for identification and validation of genomic signatures in a group of similar DNA sequences of micro-organisms. GSIT uses multiple sequence alignment to identify the unique base sites and scores them for inclusion as genomic signature for the particular strain. GSIT is a web based tool where the front-end in designed using HTML/CSS and Javascript, while back-end is run using CGI-Perl.

Availability: The server is freely available at the http://genome-sign.net/gsit.

Keywords: Infectious Diseases, Subtype Identification, Diagnostics, Genomic signatures, Comparative Genomics.

Background:
The diversity among newly emerging pathogens calls for a fast and accurate identification technique for the characterization of the pathogens. Genomic and molecular biology techniques such as plasmid profiling, restriction fragment length polymorphism (RFLP), polymerase chain reaction (PCR) and whole genome sequencing (WGS) are regularly applied in clinical settings for detection of pathogens **[1].** However, a major bottleneck in discovery of novel variants is the availability of datasets and the approach used in their discovery. Advent of fast and accurate WGS technology bridges the gap of genomic dataset availability **[2].** Comparative genomics among datasets of closely related species of same organisms is extensively used to identify genomic variation signatures. The term genomic signatures are used in various contexts such as specific expression pattern in disease or organism **[3]** or a stretch of nucleic acid/amino acid sequence specific to a particular trait **[4].** Here we define a genomic signature as a point mutation specific only to one subtype/strain of an organism. Identification of genomic signatures would not only help in

differentiating the different subtypes infecting the host but would also help in understanding the relationship between subtypes and the patients' response to treatment and may help in providing the optimal duration and type of therapy, which may ultimately improve patient management. Such approaches have been previously extensively used to develop and improve diagnostic procedures, as well as, explore the relationship between infection and drug response **[5, 6].** Here we present genomic signature identification tool (GSIT), which employees comparative genomics technique for identification of genomic signatures among subtypes/strain of same species of organism and provides a list of significant genomic signatures for many applications including pathogen characterization and epidemiological applications.

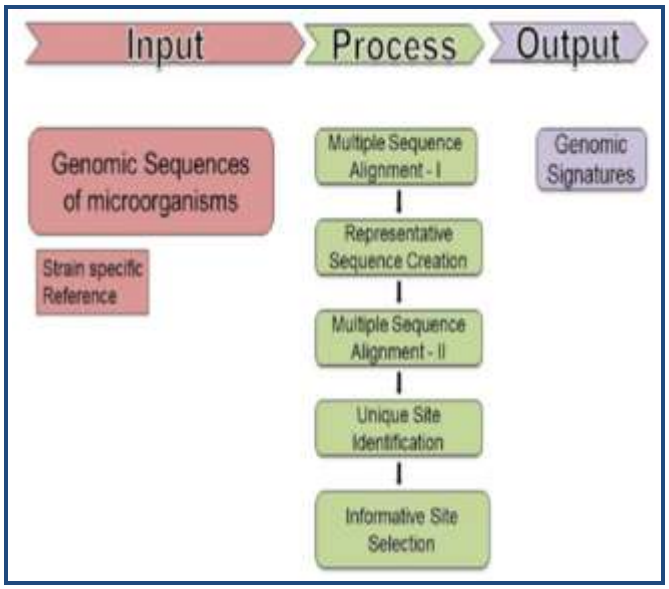

Figure 1: Organization of different modules in GSIT web application and steps employed for functioning of GSIT

Methodology:

Input:

The input to server consists of whole genome sequence of three or more subtype/strain of same species along with a reference sequence of each subtype/strain respectively. Length difference among input sequences should be less than 10% in order to confirm quality sequence alignment. Users also need to provide their functional email id where the results will be sent after the job is completed.

Algorithm:

The web server is implemented in Perl/CGI. Once the user files are uploaded to the server, multiple sequence alignment (MSA) of all the individual subtypes and representative sequences is constructed using a MSA tool ClustalW- MPI **[7]** and the alignments are used to determine the consensus sequence for each subtype. The consensus sequences of each subtype are then aligned with each other in another round of MSA. The MSA generated is parsed through custom scripts, which analyze each position of alignment and determines sites with a unique base as compared to other subtypes. Next, a probabilistic method is implemented to determine the statistical significance and subtype prediction probability of each of the unique site shortlisted in previous module. The probability of misclassification within a subtype (Pn(x | z)) and between different subtypes (Pn(x | z!)) for each site is calculated as **(Please see supplementary material for equation and explanation).**

The sites showing highest probability of classification within a subtype and lowest or null probability of misclassification both within and between subtype sequences are shortlisted and termed as genomic signatures of that particular subtype. A concise workflow showing the working of GSIT is shown in **Figure 1.**

Output:

Once the results are available, link to the result table is sent to the user via email. The output consists of a table showing the top 5 genomic signatures for each subtype along with Probabilities of misclassification. The table also shows the informative signature along with 50 bp sequences flanking the signature. Users can also download the flanking sequences to the genomic signature in fasta format as an archive file.

Proof of Concept:

We validated the pipeline using Hepatitis B as a representative model. Hepatitis B Virus (HBV) whole genome sequences (N = 174) obtained from Hepatitis B Virus database [s2as02.genes.nig.ac.jp] and reference sequences for the eight representative genotypes (A-H) obtained from NCBI- Genbank [http://www.ncbi.nlm.nih.gov/projects/genotyping/view.cgi ?db=2] were used to correlate genomic signature based identity of subtypes. A systematic analysis of HBV genotypes was performed using the GSIT tool. In all 10 Genomic Signature sites were identified for Hepatitis B Virus genotypes A-H.

Applications:

Genomic signatures find applications in a diverse set of clinical and non-clinical scenarios. Different subtypes of same viral species are known to show very different responses to therapy. An understanding of the relationship between genotypes and the patients' response to treatment may help in providing the optimal duration and type of therapy, which may ultimately improve patient management. The advantages of our technique include its ability to simultaneously compare strains at the whole-genome level with high sensitivity to detect subtle differences. These genomic signatures can be highly informative and can also be used for detection and monitoring of infectious diseases and their causative agents. Translating genomic signatures into a diagnostic and prognostic aid remains to be demonstrated.

Acknowledgement:

The authors thank Dr. Siddhartha Kundu and Dr. Vinod Scaria for useful inputs and discussions regarding algorithm and manuscript.

References:

[1] Raoult D *et al. Nat Rev Microbiol.* 2004 **2:** 151 [PMID: 15040262]

[2] Ghedin E *et al. Nature* 2005 **437**: 1162 [PMID: 16208317]

[3] Zaas AK *et al. Sci Transl Med.* 2013 **5:** 203ra126 [PMID: 24048524]
[4] Pan C *et al. PLoS One* 2010 **5**: e9549 [PMID: 20221396]
[5] Barken KB *et al. Clin Chim Acta.* 2007 **384**: 1 [PMID: 17689512]
[6] Rappuoli R, *Nat Med* 2004 **10**: 1177 [PMID: 15516917]
[7] Li KB, *Bioinformatics* 2003 **19:** 1585 [PMID: 12912844]

Edited by P Kangueane
Citation: **Tuteja *et al.*** Bioinformation 10(8): 551-554 (2014)

BIOINFORMATION

open access

Supplementary material:

Methodology:

Algorithm

The probability of misclassification within a subtype ($Pn_{(x \mid z)}$) and between different subtypes ($Pn_{(x \mid z!)}$) for each site is calculated as,

$$\boldsymbol{Pn}\ (\boldsymbol{X} \mid \boldsymbol{Z}) = \frac{\sum \boldsymbol{n}(\boldsymbol{X} \mid \boldsymbol{Z})}{\sum \boldsymbol{N}(\boldsymbol{X} \mid \boldsymbol{Z})}$$

Where, $Pn_{(X \mid Z)}$ is the probability of misclassification of n base at X position in Z type,
$\sum n_{(X \mid Z)}$ is sum of number of all the bases except base n for X position in Z type,
$\sum N_{(X \mid Z)}$ is the sum of total number of nucleotides for X position in Z type

$$\boldsymbol{Pn}\ (\boldsymbol{X} \mid \boldsymbol{Z}!) = \frac{\sum \boldsymbol{n}(\boldsymbol{X} \mid \boldsymbol{Z}!)}{\sum \boldsymbol{N}(\boldsymbol{X} \mid \boldsymbol{Z}!)}$$

Where, $Pn_{(X \mid Z!)}$ is the probability of misclassification of n base at X position in non Z type,
$\sum n_{(X \mid Z!)}$ is sum of number of all the bases except base n for X position in non Z type,
$\sum N_{(X \mid Z!)}$ is the sum of total number of nucleotides for X position in non Z type