

# VirusFinder: A Web-Based Virus Identification System Using Viral Nucleotide Signatures

Yan Ding<sup>1</sup>, Hong Cao<sup>1</sup> and Shu Wen<sup>2</sup>

<sup>1</sup> Department of Bioinformatics  
American Type Culture Collection  
10801 University Blvd.  
Manassas, Virginia 20110, USA  
Email: yding@atcc.org

<sup>2</sup> Department of Microecology  
Dalian Medical University, 465  
Zhongshan Road, Dalian 116027, China  
Email: wsn2008cn@hotmail.com

## Abstract

*VirusFinder was developed as a web-based virus identification system using viral signatures. Nucleotide signature-based detection has been proven to be a rapid and accurate pathogen identification approach in natural outbreaks or acts of bioterrorism. Species-specific signatures of selected families of HFVs including arenaviruses, bunyaviruses, filoviruses, and flaviviruses have been identified by comparative genomics. The unique signatures for each individual species of HFVs were derived from the multiple sequence alignments and were then saved into the signature database after validation. VirusFinder was implemented using J2EE technology with an intelligent signature recognition algorithm. As such, species-level differentiation and identification of HFVs was automatically achieved given an input of unknown nucleotide sequences. Results show that VirusFinder is able to identify HFVs at species level with 100% accuracy.*

## 1. Introduction

It is clear that nucleotide signatures may be exploited to quickly and accurately identify pathogens involved in natural outbreaks or acts of bioterrorism [1]. It has been proven that comparative genomic analysis is a powerful approach for identification of microbial/viral signatures [2]. However, all wet lab approaches need some clues about the pathogen(s) involved and are usually time-consuming. Moreover, there are chances of getting false-positive or false-negative interpretations due to experimental and

or systematic errors. While a computer-aided virus classification was proposed [3], the usability of the system is limited because the virus classification largely depends on the availability of physical-chemical properties of the unknown agent. The ideal solution for accurate and rapid identification of unknown agents is to match the unknown sequence with the pathogen nucleotide signature database.

One of the main ideas in this work is to identify pathogen signatures using Ebola-like hemorrhagic fever virus as an example biothreat agent. Another important idea is implementation of a web-based virus identification and information system with an intelligent signature recognition algorithm.

## 2. Signature Identification

Four families of HFVs (*arenaviruses*, *bunyaviruses*, *filoviruses*, and *flaviviruses*) that cause viral hemorrhagic fever have been identified by comparative genomics. The results shown in Figure 1 illustrate the Ebola-like virus signatures at species level. Ebola-like viruses are



Figure 1. Multiple alignment of GP gene in Ebola viruses

members of a family of negative-stranded RNA viruses, the *Filoviridae*. The Ebola glycoprotein (GP) gene was selected as a target for analysis. The GP protein is a virulence factor in Ebola pathogenesis necessary for virus replication. Its strong species-specific conservation offers a convenient way to accurately identify various Ebola species. Multiple alignment was generated using the Clustal X program [4]. The consensus sequence was derived based on the principle of minimum evolution. Unique signature sites within the GP gene of each species were identified by comparing the aligned GP genes from four Ebola virus species with the consensus nucleotide sequence (data not shown). Candidate signatures were screened against the GenBank database for potential interference from similar nucleotide sequences. Probes were tested by RT-PCR to ensure that each one amplifies pathogen DNA of the predicted size. Candidate signatures that pass these rigorous selection processes were then populated into the viral signature database.

### 3. Implementation

VirusFinder was developed as a multi-tiered J2EE enterprise application as shown in Figure 2. The outermost presentation tier provides a

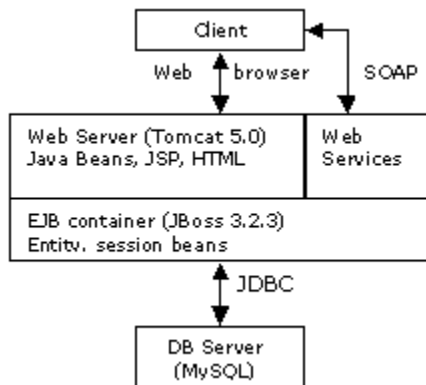


Figure 2. Architecture design of VirusFinder

browser-based graphical user interface based on HTML, JSP, and a set of web services. Initial client requests are handled by Tomcat 5.0 as the Web server. The browser and web service interfaces communicate with the application tier through a developed Application Programmer Interface (API). The middle tier implemented with Enterprise Java Beans is loaded by the JBoss 3.2.3 application server. The middle tier sits above the MySQL server as the database tier. VirusFinder was also implemented as Web

Services for enhancement of data exchange and interoperability through SOAP (Simple Object Access Protocol). The web server takes a

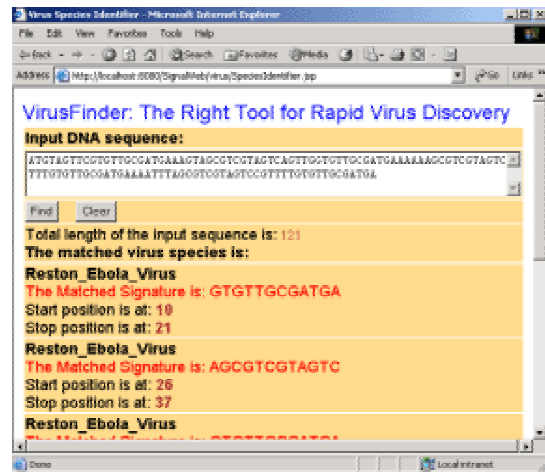


Figure 3. Snapshot of VirusFinder output

nucleotide sequence as input and displays the total length of input sequence, the matched virus species name, the matched nucleotide signatures, and the start and stop positions of the matched signatures as the outputs as shown in Figure 3.

### 4. Conclusions

VirusFinder provides the first line of biodefense for rapid and accurate virus detection and identification. It is reliable, easy to use, robust, and expandable to identify any pathogens once their nucleotide signatures are obtained. The Web Services feature enhances the data exchange and interoperability of the system.

### References

- [1] R.T. Nelson, J. Hua, and J.K. Lodge. Identification of virulence mutants of the fungal pathogen *Cryptococcus neoformans* using signature-tagged mutagenesis. *Genetics*, 2001, 157:935-947.
- [2] S. Lapa, M. Mikheev, et al. Species-level identification of orthopoxviruses with an oligonucleotide microchip. *J Clin Microbiol*, 2002, 40: 753-757.
- [3] A.S. Kolaskar and P.S. Nail. Computerization of virus data and its usefulness in virus classification. *Intervirology*, 1992, 34:133-141.
- [4] J.D. Thompson, T.J. Gibson, and D.G. Higgins. The CLUSTAL\_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.*, 1985, 25:4876-4882.