Extended Abstract International Journal of Bioassays 2020
Vol.9 No.2

Biochemistry 2020: Genus specific protein patterns of viruses- Sandeep Bansode- Dr D Y Patil Biotechnology and Bioinformatics Institute, India

Sandeep Bansode

Dr D Y Patil Biotechnology and Bioinformatics Institute, India

In the era of emerging and re-emerging viral infections, diagnostics and its allied fields have a major role to play in combating the diseases. The enormous amount of the molecular sequence data available in the public domain has the potential to contribute in a major way in the development of novel diagnostic tools. One of the perquisites for such a study is the identification of signature sequences example small stretches of protein/nucleotide sequences that are unique to a given family, organism or genus. There exist many resources in the public domain archiving signature sequences of proteins based on sequence similarity or identity. However, these resources do not take into account the taxonomic information which has a significant role to play in viral diagnostics. The present study is an effort to clearly take into account the taxonomic information. Thereby derive genus-specific signature sequences of viral proteins. The preliminary data for obtaining patterns viz, MSA-multiple sequence alignment is obtained from Virgen database. An in-house developed Perl script is used to derive the patterns from multiple sequence alignment. The patterns are then validated by the search against the non-redundant protein sequence database at the National Center for Biotechnology Information, thereby enabling the computation of their sensitivity and specificity. Such validation requires datasets pertaining to true-positives and true-negatives. The true-positive dataset is obtained from the taxonomy database at National Center for Biotechnology Information by formulating an Entrez query such that the total number of species belonging to a given genus is retrieved. The true-negative dataset constituted of any protein sequence that belongs to a genus other than the one in question. Of the 262 proteins belonging to (RNA viruses) 19 families in Virgen, patterns could be detected for 126 proteins, all of which clearly distinguished false-positive and true-positives sequences. These patterns when mapped onto their corresponding 3D structures [25 unique entries of Protein Data Bank] are found to be part of important functional regions like active site and dimerisation interface. The unique viral signature sequence / peptide thus obtained have applications not only in detection assays and as therapeutics but also can serve as putative targets for viral vaccines.

Viral infection involves a large number of protein-protein interactions between virus and its targeted host. These interactions range from the initial binding of viral coat proteins to host membrane receptors to hijack the host transcriptions machinery by virus proteins. Various viral diseases are caused by an infection with pathogenic viruses. For instance, Ebola virus diseases are a highly contagious and fatal disease caused by infection with Ebola virus. During the 2014 Ebola epidemic, the world witnessed over 26,000 cases and more than 10,000 deaths. There is no specific vaccine or effective treatment for Ebola virus disease Despite the increased number of known virus-host Protein-protein interactions, viral infection mechanism is not fully understood. Thus, identifying interactions between virus proteins and host proteins helps understand the mechanism of viral infection and develop treatments and vaccines.

Many computational methods have been developed to predict Protein-protein interactions. However, most of these methods predict Protein-protein interactions within a single species and cannot be used to predict Protein-protein interactions between different species because they do not distinguish interactions between proteins of the same species from those of different species. Recently, a few computational methods have been developed to predict virus-host Protein-protein interactions using machine learning methods. For instance, a homology-based method predicts Protein-protein interactions between H. sapiens and M. tuberculosis H37Rv. Support vector machine model developed by Cui et al. predicted Protein-protein interactions between human and two types of viruses (hepatitis C virus and human papillomavirus). However, these methods are intended for Protein-protein interactions between virus of a single type and host of a single type. Recent computational methods developed for predicting virus-host Protein-protein interactions are also limited to Protein-protein interactions between human and the human immunodeficiency virus 1 and cannot predict Protein-protein interactions of new viruses or new hosts which have no known Protein-protein interactions to the methods. A recent SVM model called DeNovo can exceptionally predict Protein-protein interactions of new viruses with a shared host

The emergence of high throughput technologies for genome sequencing, microarrays and proteomics transformed biology into a data-rich information science. Sequencing the complete genome of an organism is the first step in generating the 'parts list' of life. One of the first efforts involved the sequencing of Haemophilus influenzae in 1995. As of July 2006, more than 404 organisms have been sequenced completely. Furthermore, the genome sequencing projects of ~609 eukaryotic species and~933 prokaryotic have been launched. Enormous data generated by the genome sequencing projects are archived in both dedicated genomic resources and public domain databases. While the complete genome sequencing of the model organisms and microbes are taking the centre-stage, viral genome sequencing continues to be individual efforts. Viruses are a

This work is partly presented at International Conference on Biochemistry, Proteomics & Bioinformatics- February 16-17, 2020 held at Singapore

diverse group of organisms and are most abundant. The genome size of viruses varies from a few hundreds to millions of bases SV-40 was the first virus for which the complete genome sequence was obtained in the late 70s. About more than4000 viruses have been sequenced so far by virologists all over the world with an objective to study antigenic variation, geographic distribution, spread and evolution. These independent efforts enabled viruses to attain the status of 'best-represented taxa' with the highest number of whole genomes sequenced. However, due to lack of concerted efforts, viral genomic sequences only added to the entries in the public repositories until recently. The Genome OnLine Database is a tracking system for genome sequencing and provides the update of various genome-sequencing projects but does not have any mechanism to specifically monitor viral genome sequencing initiatives.

Whole-genome sequence data of viruses offer unlimited opportunities for data mining and knowledge discovery. The complete genome sequences of two large viral genomes Mimivirus and PolyDeoxyribonucleic acid virus substantiate this fact. Varying coding density and the occurrence of genes associated with metabolic pathways in these Deoxyribonucleic Acid viruses offer interesting opportunities in viral genomics in general and in understanding the evolution of viruses in particular. However, it is known that in the absence of curation and functional annotation of the genomic data and the utility of the sequence data is minimal and the sequence merely remains as an entry in the database. Bioinformatics provides a large number of databases, tools and approaches for mining huge sequence data. Although there exist numerous genome databases for the model organisms and microbes, there are a few databases, which archive viral genomic data. Many databases are the synthesis of experimental work carried out in the respective laboratories. As a result, these compilations are highly specialized.

This work is partly presented at International Conference on Biochemistry, Proteomics & Bioinformatics- February 16-17, 2020 held at Singapore