Phylogenetic Clustering of Protein Sequences Using Recurrence Quantification Analysis

Akshay Yadav1, V. K. Jayaraman2,*, Mohan Kale3, and Urmila Kulkarni-Kale1

1Bioinformatics Centre, University of Pune, India
2Centre for Development of Advanced Computing (C-DAC), Pune, India
3Department of Statistics, University of Pune, India

Molecular phylogeny analysis (MPA) investigates differences in molecular sequences to analyse the evolutionary-ary relationships of organisms or their bio-macromolecules. Currently there are two categories of methods for MPA viz. distance based and character based. All the methods require multiple sequence alignment (MSA) as a prerequisite to MPA and are usually followed by bootstrap analysis. MSA is efficient in terms of time, computations and memory requirement only when size and number of sequences are small. MSA of whole proteomes is computationally intensive and time consuming. In this paper, an attempt is made to propose a new alignment free approach of MPA using Recurrence Quantification Analysis (RQA). The protein sequence is converted into a numeric sequence by assigning a unique score in the form of real numbers to each amino acid and various RQA features are extracted from the numeric sequence. These features are used as co-ordinates for calculating the distance between the sequences using an appropriate distance function. The distance matrix, thus obtained is used for clustering using Neighbour-Joining method. Requirement of time for clustering and tree construction is observed to be significantly reduced in comparison with the alignment-based algorithms. As an example, a test case involving an application of this method for clustering of 59 polyprotein sequences of family Flaviviridae is demonstrated and phylogenetic tree thus obtained was found to be fairly accurate with 95% accuracy. Only 3 misclassifications out of 59 were observed. Thus, the proposed method has potential to be a reasonable alternative for existing MPA methods.

Keywords: Recurrence Quantification Analysis, Molecular Phylogeny Analysis, Flaviviridae.

1. INTRODUCTION

Molecular phylogeny analysis (MPA) of protein sequences involves determination of how sequences evolve and might have been derived from ancestral sequences during the process of evolution. The evolutionary relationships among the sequences are depicted by placement of the sequences as the outer branches on a tree. The branching relationships on the inner part of the tree then reflect the degree to which the different sequences are related.1 There are currently two types of methods for MPA viz. character based and distance based. Distance based methods calculate all-to-all distances between sequences under comparison and the clustering is carried out on the basis of the distance matrix.2 The clustering algorithms compute a tree based on a distance matrix starting from the most similar sequence pairs. In character-based methods, clustering is directly carried out on the basis of character differences at various positions in the multiple sequence alignment (MSA).3 These traditional methods however require MSA of the sequences and bootstrap analysis of the sequence data. MSA becomes computationally intensive and time consuming when number and size of the sequences increases.4 Bootstrap analysis is used for assessment of reliability of phylogenetic tree by assigning a certain degree of confidence to the relationships between the sequences using statistical measures. Therefore there is a need for alternate alignment-free methods for phylogenetic analysis. To overcome these limitations, many alignment free methods have been proposed for phylogenetic analysis of whole genome and proteomes. Ding et al. used a simple 4k-dimension feature representation vector, where k is the length of a word for phylogenetic analysis of DNA sequences.5 Qi et al. gave systematic methodology for inferring phylogeny of microbial organisms from the oligopeptide content, i.e., frequency of amino acid k-strings in their complete proteomes.6 Kolekar et al. used the Inter arrival time distribution (Return Time Distribution) approach of k-mers for phylogenetic analysis and genotyping of Mumps and Dengue viruses.7,9 Recurrence plot enables the visual analysis of m dimensional phase space trajectory through a 2D plot of recurrences. Both axes of the recurrence plot are ordered sequences. Recurrence plot is the visualization of a square recurrence matrix (RM) of
distance elements within a cut-off limit. An example of the recurrence plot is illustrated in the Figure 1. The quantification analysis of recurrence plots involves a set of descriptors that describe small-scale structures in the plots, such as single dots, and diagonal, vertical, or horizontal lines.

Recurrence Quantification is a modern technique for quantifying time-series data in phase space. Eckmann et al. introduced the concept of “recurrence” in 1987 as a tool for visualizing the recurrences of a state \( x_i \) in phase space. Since recurrence plots were limited to only visual analysis, Zhilut and Webber quantified RQA in terms of mathematical equations and extracted a set of important features from RQA plots. The quantification of nonlinear signals by the means of these RQA features captures useful information that would normally not be evident in the raw signal. Webber et al. encoded the proteins according to the hydrophathy indices of the 20 amino acids, and analyzed the resulting time series data using recurrence quantification. RQA has also been used to understand the structure function relationships from the hydrophobicity profiles of the primary structure. RQA has also been used to understand the problem of protein solubility. We have also demonstrated use of RQA features for discrimination between aggregating and non-aggregating proteins.

We propose a new alignment-free method for phylogenetic analysis of protein sequences using RQA. The protein sequences are first converted to numerical sequences by assigning a score to each amino acid. These numerical sequences are subsequently treated as a time series data for Recurrence analysis. Finally, the recurrence plot features were used as coordinates for calculating the distance between the sequences using an appropriate distance function.

### 2. MATERIALS AND METHODS

#### 2.1. Dataset

Fifty-nine polyprotein sequences of the viruses that belong to family Flaviviridae were compiled from the VirGen database. The viruses of the family Flaviviridae (ssRNA enveloped viruses) are important arthropod-borne viruses and cause of many diseases in humans. They produce a broad spectrum of clinical responses in humans ranging from asymptomatic infection to fulminant encephalitis or haemorrhagic fever. The polyprotein sequences were taken from all the 4 genera of family Flaviviridae viz. *Flavivirus, Hepacivirus, Pestivirus* and *Pegivirus*.

#### 2.2. Scoring

The protein sequences were converted to numerical sequences by assigning a score to each amino acid using any one of the five factor values (Table I) given by Atchely et al. These scores are the numeric patterns of amino acid variability that are produced by using multivariate statistical analyses on almost 500 amino acid attributes. Each set of five factor values correspond to particular amino acid features. Factor I is a polarity index. Factor II is a secondary structure factor. Factor III relates to molecular size or volume. Factor IV reflects relative amino acid composition in various proteins, number of codons coding for an amino acid, and amino acid composition. Factor V refers to electrostatic charge with high coefficients on isoelectric point and net charge.

#### 2.3. Recurrence Quantification Analysis

Each of the 59 polyprotein sequences of family Flaviviridae were converted to numerical sequences using all the 5 factor amino acid scores. These numerical sequences were treated as a time series for Recurrence analysis. A recurrence of a state at position \( i \) at a later position \( j \) can be visualized in 2D recurrence plots as a square matrix of black and white dots (Fig. 1). Both axes are the ordered sequences and each black dot indicates a recurrence. This recurrence plot may be mathematically expressed as:

\[
R_{ij} = \Theta(e - \| x_i - x_j \|), \quad x_i \in X \times \cdots, \quad i, j = 1, \ldots, N
\]

where \( N \) is the number of states considered for analysis, \( e \) is a threshold distance, \( \| \cdot \| \) is the norm, and \( \Theta \) the Heaviside function. To decide whether a given point is recurrent, it must fall within the threshold distance, \( e \).

The distances within the matrix, their rescaling, and recurrence plot texture the recurrence parameters viz. Embedding dimension (\( M \) or EMBED), Delay (\( \tau \) or DELAY), Radius (RADIUS)²⁰

### Table I. The five factor scores of all the amino acids.

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Factor I</th>
<th>Factor II</th>
<th>Factor III</th>
<th>Factor IV</th>
<th>Factor V</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>−0.591</td>
<td>−1.302</td>
<td>−0.733</td>
<td>1.57</td>
<td>−0.146</td>
</tr>
<tr>
<td>C</td>
<td>−1.343</td>
<td>0.462</td>
<td>−0.862</td>
<td>−1.02</td>
<td>−0.255</td>
</tr>
<tr>
<td>D</td>
<td>1.05</td>
<td>0.302</td>
<td>−3.656</td>
<td>−0.259</td>
<td>−3.242</td>
</tr>
<tr>
<td>E</td>
<td>1.367</td>
<td>−1.453</td>
<td>1.477</td>
<td>0.113</td>
<td>−0.637</td>
</tr>
<tr>
<td>F</td>
<td>−1.006</td>
<td>−0.690</td>
<td>1.891</td>
<td>−0.397</td>
<td>0.412</td>
</tr>
<tr>
<td>G</td>
<td>−0.384</td>
<td>1.652</td>
<td>1.33</td>
<td>1.045</td>
<td>2.064</td>
</tr>
<tr>
<td>H</td>
<td>0.336</td>
<td>−0.417</td>
<td>−1.673</td>
<td>−1.474</td>
<td>−0.078</td>
</tr>
<tr>
<td>I</td>
<td>−1.239</td>
<td>−0.547</td>
<td>2.131</td>
<td>0.393</td>
<td>0.816</td>
</tr>
<tr>
<td>K</td>
<td>1.831</td>
<td>−0.561</td>
<td>0.533</td>
<td>−0.277</td>
<td>1.648</td>
</tr>
<tr>
<td>L</td>
<td>−1.019</td>
<td>0.987</td>
<td>−1.505</td>
<td>1.266</td>
<td>−0.912</td>
</tr>
<tr>
<td>M</td>
<td>−0.663</td>
<td>−1.524</td>
<td>2.219</td>
<td>−1.005</td>
<td>1.212</td>
</tr>
<tr>
<td>N</td>
<td>0.945</td>
<td>0.828</td>
<td>1.299</td>
<td>−0.169</td>
<td>0.933</td>
</tr>
<tr>
<td>P</td>
<td>0.189</td>
<td>2.081</td>
<td>−1.628</td>
<td>0.421</td>
<td>−0.832</td>
</tr>
<tr>
<td>Q</td>
<td>0.931</td>
<td>−0.179</td>
<td>−3.005</td>
<td>−0.503</td>
<td>−1.853</td>
</tr>
<tr>
<td>R</td>
<td>1.538</td>
<td>−0.055</td>
<td>1.502</td>
<td>0.44</td>
<td>2.897</td>
</tr>
<tr>
<td>S</td>
<td>−0.228</td>
<td>1.399</td>
<td>−4.76</td>
<td>0.67</td>
<td>−2.647</td>
</tr>
<tr>
<td>T</td>
<td>−0.032</td>
<td>0.326</td>
<td>2.213</td>
<td>0.908</td>
<td>1.313</td>
</tr>
<tr>
<td>V</td>
<td>−1.337</td>
<td>−0.279</td>
<td>−0.544</td>
<td>1.242</td>
<td>−1.262</td>
</tr>
<tr>
<td>W</td>
<td>−0.595</td>
<td>0.009</td>
<td>0.672</td>
<td>−2.128</td>
<td>−0.184</td>
</tr>
<tr>
<td>Y</td>
<td>0.26</td>
<td>0.830</td>
<td>3.097</td>
<td>−0.838</td>
<td>1.512</td>
</tr>
</tbody>
</table>
The Embedding dimension \((M)\) or EMBED parameter is used for higher dimensional reconstruction of the time series by the method of time delays, as introduced by Takens (1981). This theorem states that the topological features of any higher-dimensional system consisting of multiple coupled variables can be reconstructed from a single measured variable of that system. The reconstruction is performed by defining time-delayed vectors \(Y_i\) of \(M\) points \(P_i\) that are delayed or offset in time \((\tau\) or Delay).

The RADIUS parameter implements a cut-off limit (Heavy-side function) that transforms the distance matrix (DM) into the recurrence matrix (RM).

### 2.4. Recurrence Features

The quantification analysis of recurrence plots involves a set of descriptors that describe small-scale structures in the plots, such as single dots, and diagonal, vertical, or horizontal lines. Fourteen recurrence features (Table II) were extracted from the recurrence plots of the protein sequences for calculating the distance between the sequences.

### 2.5. Distance Function

Normalized distance function was used to calculate the distances between the sequences using the above-mentioned features as coordinates.

\[
d(X, Y) = \frac{\sum |x_i - y_i|}{(1 + \sum |x_i - y_i|)}
\]

The phylogenetic trees were generated using Neighbor program in PHYLIP package (http://evolution.genetics.washington.edu/phylip.html) using Neighbor-Joining (NJ) method for clustering. The entire process of optimization of RQA parameters was automated using perl codes developed in-house.

### 3. RESULTS

The three recurrence parameters were optimized for obtaining the phylogenetic tree with best possible genus specific clades. The best tree where only 3 out of 59 taxa misclassified was found for EMBED = 2, DELAY = 0.1 and RADIUS = 2. The highest accuracy for clustering was obtained by using the factor II amino acid scores. Factor II amino acid scores reflects the relative propensity for various amino acids in various secondary structural states. The tree obtained is shown in Figure 2 using genus specific color-codes. As can be seen, the four genera viz., Hepacivirus (cyan), Pegivirus (pink), Pestivirus (green), mosquito-borne Flavivirus (red) and tick-borne Flavivirus (blue) are found to form four clades. It can be seen that polyproteins of 56 Flaviviruses are clustered in accordance with their taxonomical classification leading to an accuracy of 95%. However, the taxa HGBV-C and HGV-1 from the genus Pegivirus are misclassified in the Pestivirus and the taxa BVDV-1 from the genus Pestivirus is misclassified into Pegivirus. The Flavivirus clade is separated into two sub-clades according to tick-borne (blue) and mosquito-borne (red), as expected. The four DENV serotypes (DENV-1, DENV-2, DENV-3, and DENV-4) are also clustered in the same clade.

The average sequence length of the 59-polyprotein sequences is 3 kb. The generation of alignment based phylogenetic tree of this dataset takes around 72 hours while as the proposed alignment-free method based on RQA requires 3 minutes only.

### 4. CONCLUSIONS

Although the traditional alignment based algorithms for inferring molecular phylogenies are widely accepted, these algorithms became computationally intensive when the number and size of the sequences increases. The exponential rise in whole genomic and proteomic data necessitates the development of quicker/faster and optimal methods for inferring whole genome and proteome phylogenies. Protein sequences have characteristic patterns of recurrence of states which can be used to explore the evolutionary relationships between them. Closely related sequences have similar patterns which are reflected in the features extracted.
from these patterns. The proposed method for inferring the phylogenetic relationship using Recurrence quantification analysis, as evident for the phylogenetic analysis of whole polyprotein sequences of *Flaviviridae* family, can be used as a fast and reliable alternative for whole proteome phylogenies. It needs to be applied to various other types of the datasets and work in this direction is in progress.

**Acknowledgments:** V. K. Jayaraman gratefully acknowledges CSIR, New Delhi for financial assistance in the form of emeritus scientist grant. A. Akshay Yadav acknowledges the BioInformatics National Certification (BINC) fellowship awarded by Department of Biotechnology (DBT), Government of India. Urmila Kulkarni-Kale acknowledges financial support under the Centre of Excellence (CoE) grant of Department of Biotechnology (DBT), Government of India. All authors acknowledge the infrastructural facilities instituted by the DBT at Bioinformatics Centre, University of Pune.

**References and Notes**


Received: 5 May 2012. Accepted: 10 June 2012.