Poster H-7

Classification of HCV Using Whole Genome Method

Authors:
Raymond Wan (*Gilead Sciences Inc.*)
Eugenia Bastos (*SAS Institute*)

Short Abstract: Q-gram is used in clustering procedures to generate a hierarchical tree for HCV genomes. A classifier is also built using the measure to classify sequences into subtypes, showing that q-gram can be very good measure of sequence similarity and it can be used effectively in clustering viral genomes.

Long Abstract:

The idea of using fixed length words (q-gram) has not seen much use in DNA sequence analysis. In recent years, it has been used in sequence query engine like BLAT, where it is used to filter out large portion of the sequence database before more traditional and resource intensive alignment-based techniques are applied to produce the final search string. In this paper, q-gram is used to generate a distance matrix for a set of Hepatitis C (HCV) DNA sequences. The distance matrix is then used as input to standard clustering procedures from SAS language to generate a hierarchical tree showing how different subtypes of HCV are related to each other. We also used the distance measure to create a simple classifier and test it against shorter sequences from the LANL database.

Traditionally, classification of virus is based on a number of clinical criteria that includes characteristics related to culture system and disease manifestation. For HCV, the classification relies entirely on genetic information. There are now more than 80 recognized subtypes within the six families. In addition, there are recombinant subtypes made of a mixture of different subtypes. There is no easy way to genotype a viral sample precisely. The current classification method of HCV is mostly based on comparison of a small part of the viral genome, usually the 5' UTR and NS5B region. As more full-length sequence data is accumulated, it becomes possible to genotype a virus sample by whole genome comparison. Genetic proximity is much easier to measure when one looks at the whole virus rather than small pieces of the genome. This kind of analysis will also allow us to map the evolution of the virus and its acquisition of traits like drug resistance in geographical location and time. This is one reason why large scale sequencing of virus is important in the fight against viral diseases like HIV and Avian Flu. In this work, the viral sequence data came from the Los Alamos HIV database. The annotation of each sequence contains information about the viral sample, including country of origin, gene segments contained in the sequence, accession number and patient number. For this study, we will focus on the genotyping information.

Q-gram (or n-gram) methods refer to a family of algorithms that involves the breaking up of text strings into words of fixed length q (the q in q-gram). There are several flavors of q-gram. In this paper, we use overlapping q-grams to perform pair-wise comparisons of the entire collection of HVC genomes to calculate a full distance matrix. Cluster analyses were performed using the distance as input to the CLUSTER procedure of SAS. In the resulting

hierarchical tree, all 223 samples are clustered according to their subtypes, in exact agreement with subtype information contained in the annotations. This confirms the validity of the clustering methodology using q-gram distance and that q-gram distance is a good measure of sequence similarity.

In stage 2, we build a simple classifier using the full genomes. The classifier is based on the idea of average distance to a subtype cluster, which is the basis of the previous clustering algorithm. The average distance to a subtype cluster is the average q-gram distance between a test sample and every sample of a given subtype. We use this classifier to identify the subtype of the shorter sequence segments randomly selected from the remaining samples of the database. To test the classifier, we randomly selected 425 annotated sequence segments from the database. The test showed that, even with limited data, a simple classifier can work quite well. More sophisticated machine learning algorithms was not used here because of the limitation in available data.

We showed that q-gram distance is a very economical and simple tool in measuring sequence and genetic relationship. It can be use effectively as a measure of sequence similarity in clustering and classifying sequence samples of viral genomes. This method has the advantage of being computationally simple and quick. No bootstrapping, complicated clustering algorithms or multiple-alignment were needed. Q-gram method is free of arbitrary parameters and it does not need an amino acid substitution scoring table like BLOSUM. Recombination and genomic shuffling are not so easy to deal with in traditional alignment-based sequence comparison methods but this current method has no such deficiencies. It makes full use of all available information contained within the sequences. This methodology can also be easily adapted to analyze HIV, AIV or other viral genomes. If a sufficiently large library of annotated full-length sequences is built up, classification of even small segments (500 to 2,000 bases) can be performed using standard machine learning procedures. This would provide a very useful tool for understanding the spread and evolution of many viral diseases.