Virus Genotyping : Standard Operating Procedure

1. Purpose

This document describes the ViPR computing process, whereby virus subtypes are determined. The method is a command line back-end data processing tool. Some components were derived from the work of Los Alamos HCV database team, or the VBRC (Virus Bioinformatics Resource Center) project. Currently, version 1.2 operates on Hepatitis C (HCV), Dengue, West Nile, Japanese encephalitis, St Louis encephalitis, Tick-borne encephalitis, Yellow Fever, Murray Valley encephalitis, and Bovine viral diarrhea virus. It is theoretically portable to other virus families, provided that reference sequences for the relevant subtypes are available.

2. Method Description

The method uses phylogenetic trees to find, for an 'unknown' input, the closest reference genome from all subtypes. It is a relativist method, rather than employing an absolute homology standard such as the *e-value* from *Blastn*. Operationally it builds a tree from input DNA sequence in GBK or FASTA file, and determines the best standard genotype neighbor, relative to others. Sensitivity of the method is set by defining Branching Indices (BI) cutoffs, derived from phylogeny tree characteristics. Output files contain the genotype of the virus, within tab-delimited text.

The vipr_genotype package includes the Perl script vipr_genotype.pl, Perl modules Genotype.pm and Draw_graph.pm, and an MSA for each of the viral species. Dependencies of the method are on these packages in the executable PATH, for alignment and phylogeny compute: MUSCLE, dnadist of the Felsenstein "Phylip" package, and FastME, a "minimum evolution" tree inference engine. The dnadist/FastME approach is used because of its speed, while other tree-generating programs such as Phylip and PhyML can be used with minimal code change if deemed necessary.

3. Input Data Preparation

The input for the method is one in GenBank format (GBK) or FASTA format, or more genomes in separate files in a directory. During operation, the DNA sequence is aligned by MUSCLE with the corresponding reference sequences. This MSA is then used to generate a distance matrix by dnadist, which is then used by FastME to generate the phylogeny tree. The BI value (0-1) for the input genome is calculated as the ratio of distance between grandparent and parent of the target and distance between sibling (a reference) and grandparent. If the BI is above certain threshold, it is assumed that the target belonged to the subtype indicated by the sibling node. For HCV, the threshold is 0.711, according to Hraber *et. al* (1).

Command options of the vipr_genotype.pl script:

-t	Taxon of the input genome. Valid options are: HCV, DENGUE, STLOUIS, WESTNILE, JAPENCEPH, TKBENCEPH, YELLOWFEVER, BOVDIARRHEA1, MURRAY. If absent, a separate blastn run will be performed to determine the species. However this process isn't as accurate as it should be, so the user supplied species is preferred.
-d	Directory of the input file(s)
-i	Name of input genome file in GenBank or FASTA format. If omitted, process all GenBank files in directory.
-recomb	Also run the recombination calculation, where a sliding window of 400 nucleotides is moved by 100 nucleotides each step, to produce a genotype profile along the sequence.

Sample commands:

```
$ ./vipr_genotype.pl -t HCV -d ./ -i AB119282.gb
$ ./vipr_genotype.pl -t HCV -d ./ -i AB119282.faa
$ ./vipr_genotype.pl -t HCV -d ./
$ ./vipr_genotype.pl -t HCV -recomb -d ./ -i AB119282.gb
$ ./vipr_genotype.pl -t HCV -recomb -d ./
```

4. Output Data, Processing and Display

Results of the genotype analysis are output in tsv file named *_genotype_summary.tsv with rows of data, one for each input sequence. The genotype analysis generates a single output line, listing the job_name, job_type, start and end of the sequence, BI value, genotype, time stamp, status, and any comment.

```
Job_name  Job_type      start..end Branch_Index    genotype        Date(ymdhms)    Status  Comment
AB119282  genotype       1..9658   BI=0.765676     genotype=1b     20111109104143  Success comment=
```

Results of the recombination analysis are output in .tsv file named *_recombination_summary.tsv with rows of data, one for each input sequence, delimited in format:

```
AB119282  recomb_summ    1..10262  BI=0.000000     genotype=1b     20111109104326  Success comment=
AB119282  recomb_window  1..400    BI=0.286235     genotype=1a     20111109104144  Success comment=
......
```

For a recombination analysis, a summary line is first generated to give the summary of all steps. Any subtype with >=5 steps with BI>threshold is listed in the genotype field and separated by comma. This is followed by output lines labeled by job_type=recomb_window for each step.

If the method does not produce an answer, such as when there is no current similar strain, a ‘Failed’ error notice is supplied to the data processor for analysis. This helps distinguish between technical and scientific bugs. The method exits on its own.

Other files include the following:
1. Input DNA sequence in FASTA format,
2. MSA in both FASTA and Phylip formats,
3. Phylogeny tree in Newick format, and
4. A bar graph of the BI profile in .gif format.

5. References

1. Hraber, P., Kuiken, C., Waugh, M., Geer, S., Bruno, W.J., Leitner, T. (2008) Classification of hepatitis C virus and human immunodeficiency virus-1 sequences with the branching index. *J Gen Virol. Sep;89 (Pt 9):2098-107*. PMID: 18753218.
2. Hraber, P.T., Leach, R.W., Reilly, L.P., Thurmond, J., Yusim, K., Kuiken, C; Los Alamos HIV database team. (2007) Los Alamos hepatitis C virus sequence and human immunology databases: an expanding resource for antiviral research. *Antivir Chem Chemother*. 18(3):113-23. Erratum in: *Antivir Chem Chemother. 18(4):243*. PMID: 17626595.
3. Kuiken, C., Simmonds, P. (2009) Nomenclature and numbering of the hepatitis C virus. *Methods Mol Biol. 510:33-53*. PMID: 19009252.
4. Desper, R. & Gascuel, O. (2002). Fast and Accurate Phylogeny Reconstruction Algorithms Based on the Minimum-Evolution Principle. *Journal of Computational Biology 19(5), pp. 687-705*.

