WPI CS525-Bioinformatics Project

“Predicting subtype of HCV type 1 from NS5A and rate of mutation in NS5A in HCV subtype 1b”

Alyssa Tsiros & Dmitry Korkin, PhD

Department of Bioinformatics & Computational Biology, Worcester Polytechnic Institute, 100 Institute Road Worcester MA 01609.

Received on May 7, 2015;

ABSTRACT

Motivation: I've developed a profile hidden Markov model for predicting the subtype of HCV type 1 from the nonstructural protein 5a gene that performs at 95.26% accuracy. Furthermore, I've computed the Jukes-Cantor pairwise genetic distances between representative sequences of NS5A from HCV subtype 1b, drawn from years 1999, 2000, 2001, 2003, 2006, 2007, 2008, and 2010 and used them to construct a phylogenetic tree and to determine rate of substitutions per nucleotide per year for this gene. This method performs at an average accuracy of 54.25%.

1. INTRODUCTION

The Hepatitis C Virus (HCV) infection is the leading cause of liver disease worldwide, infecting over 170 million individuals, with more than one million new cases each year (Zein, 2000). In the United States (U.S.) alone, approximately four million people are infected by the virus, which is responsible for the deaths of over eight thousand individuals each year (Zein, 2000). The current therapy for the virus combines interferon-α with ribavirin. Interferon-α is a cytokine protein capable of stimulating an innate-immune response as well as transitioning from an innate to an adaptive-immune response and it is a commonly used immunotherapy (Brassard, 2002). Ribavirin is a nucleoside analogue that acts as a nucleoside in DNA synthesis and works to prevent viral replication in infected cells (National Library of Medicine, 2012). Treatment with interferon-α and ribavirin clears HCV in approximately 50% of patients National Library of Medicine, 2012).

The HCV virus is an enveloped RNA virus with six distinct HCV genotypes have been confirmed, each with several subtypes (Division of Viral Hepatitis, 2015). HCV genotype 1 is the most common genotype found in the U.S. as it accounts for over 70% of all HCV cases. Moreover, genotype 1 is documented as the least responsive to interferon-ribavirin treatment (Division of Viral Hepatitis, 2015). In one study by Enmoto et al., authors found a correlation between interferon response in patients with HCV type 1b and the number of amino acid substitutions in the nonstructural protein 5a (NS5A) gene (**Fig. 1**).

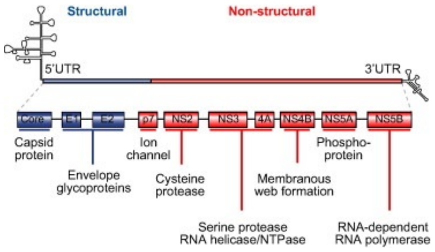


Figure 1. Structural organization of HCV RNA and proteins (Pawlotsky, 2013)

The NS5A protein has been implicated in the regulation of replication; however, its exact role is still unclear (Pawlotsky, 2013). Although inhibitors in combination with regular interferon-ribavirin therapy have been used in treatment, several documented amino acid substitutions in the NS5A protein have caused resistance in such inhibitors (Pawlotsky, 2013). Moreover, various studies indicate differing responses to interferon treatment of HCV types 1a and 1b. One study in particular reported better virological response to antiviral therapy in HCV type 1a than in type 1b (Andreoli et al., 2012).

A machine learning method for predicting subtypes of HCV type 1 from the NS5A protein could support the hypothesis that the effectiveness of interferon treatment of patients with HCV type 1 might be correlated to NS5A. Furthermore, an analysis of nucleotide substitution rate for NS5A in HCV subtypes 1a and 1b must be conducted to aid researchers and scientists in the preparation of target therapies for new and potentially resistant strains. The prediction of such nucleotide substitutions can ultimately help to determine the effectiveness of current and future treatments.

2. METHODS

The methods are presented in two subsections. First, the process of developing a profile hidden Markov model for the prediction of HCV type 1 subtypes (a or b) is described. Next, the process by which we determine the rate of substitutions per nucleotide per year for HCV subtype 1b is discussed.

1. PHMM to predict HCV type 1 subtypes from NS5A

A total of 573 HCV type 1 NS5A protein sequences was collected from the European HCV database (euHCVdb) (https://euhcvdb.ibcp.fr/euHCVdb/jsp/index.jsp) after duplicates and sequences containing non amino acid residue symbols were removed. The sequences were split into training and testing sets comprising 409 and 164 sequences, respectively. A multiple sequence alignment (MSA) of the training set was generated and a hidden Markov model (HMM) structure with the length of the alignment was constructed. A profile hidden Markov model (PHMM, **Fig. 2**) was estimated from the multiple sequence alignment and amino acid residue distributions.

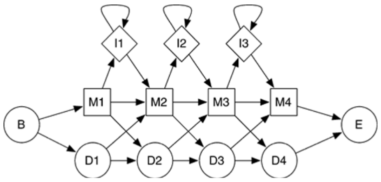


Figure 2. Profile hidden Markov model of length four with match, insertion, and delete states (Sczyrba, 2005)

A PHMM of a MSA of length L contains L match states, M_i. Match state M_i produces residue j with probability M_{ij}. Insertion states I_i emit background residues according to the π distribution. Deletion states D_i is non-emitting, thus it permits the i-th sequence residue to be passed over. At each position i of the sequence, the seven possible transitions are as followed: $M_i \longrightarrow M_{i+1}$, $M_i \longrightarrow I_i$, $M_i \longrightarrow D_{i+1}$, $I_i \longrightarrow I_i$, $I_i \longrightarrow M_{i+1}$, $D_i \longrightarrow D_{i+1}$, and $D_i \longrightarrow M_{i+1}$. A PHMM can take any state path from beginning, B, to end, E, following the transitions described above and pictured in **Figure 2**. Sequences passed through the PHMM are scored using log-odd ratios for emission probabilities and log probabilities for state transitions. $v^{M_j}(i)$ is defined as the logarithmic likelihood score of the best path for matching $x_1..x_i$ considering the current symbol a_i. $v^I_j(i)$ and $v^D_j(i)$ are defined similarly. The formula is as followed:

The training sequences were scored by the PHMM (**Fig. 3a**) and a 95% confidence interval was calculated around the mean scores for subtypes 1a and 1b. The testing sequences were assigned to the appropriate subtype category if the PHMM scores fell within the designated interval (**Fig. 3b**).

$$v_j^M(i) = \log\left[\frac{e_{M_j}(x_i)}{p(x_i)}\right] + \max\begin{cases} v^M_{j-1}(i-1) + \log\left(a_{M_{j-1},M_j}\right) \\ v^I_{j-1}(i-1) + \log\left(a_{I_{j-1},M_j}\right) \\ v^D_{j-1}(i-1) + \log\left(a_{D_{j-1},M_j}\right) \end{cases}$$

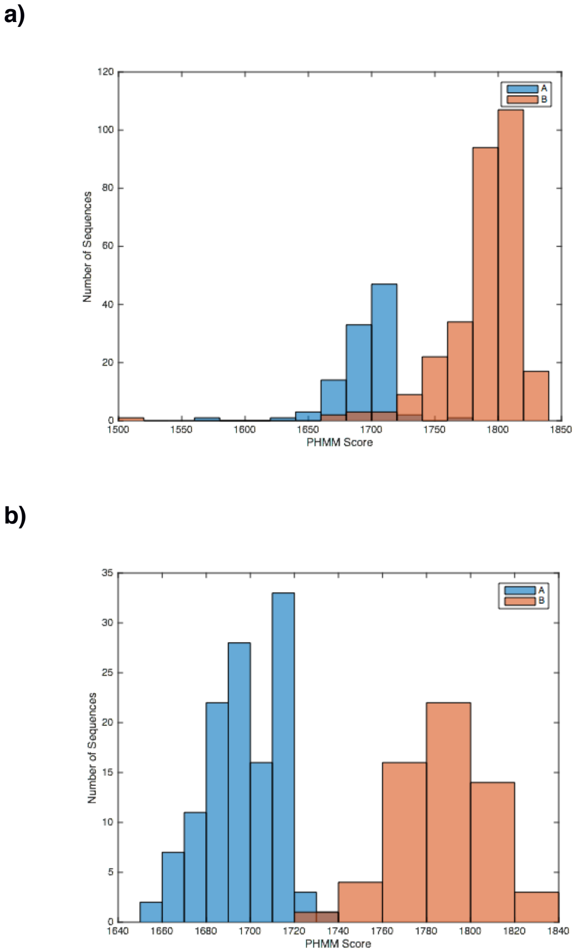


Figure 3. PHMM scores. **a)** PHMM scores of training sequences. **b)** PHMM scores of testing sequences.

i.2. Determining rate of substitutions per nucleotide per year in the NS5A protein of HCV subtype 1b

A total of 317 HCV subtype 1b NS5A DNA sequences were collected from the euHCVdb after duplicates and sequences containing non nucleotide base symbols were removed. The sequences were sorted into year of original publication, which resulted in 8 groups (1999, 2000, 2001, 2003, 2006, 2007, 2008, 20010). Any year containing three sequences or fewer was not considered for this analysis. A MSA and a PHMM (**Fig. 2**) were generated for each year to determine which sequence was the most representative of its year. Next, the pairwise Jukes-Cantor genetic distance of the representative sequences from one another was determined and a phylogenetic tree indicating these distances was produced (**Fig. 4a**). These distances were calculated as followed, where d is distance and β is the proportion of positions at which two sequences are different (β is close one for poorly related sequences and close to zero for similar sequences):

$$d = -\frac{3}{4}\ln\left(1 - \frac{4}{3}\beta\right)$$

Pairwise Jukes-Cantor distances were calculated with respect to the earliest representative sequence (1999) and a new phylogenetic tree was produced from these results (**Fig. 4b**). A least squares linear fit of genetic distances with respect to the 1999 sequence was then performed to determine the rate of substitutions per nucleotide per year. This analysis was reproduced without the genetic distance between the 1991 sequence and itself (a distance of zero) to generate a more accurate fit (**Fig. 5**). Finally, a 95% confidence interval of the linear fit was calculated and accuracy of the model was determined from a random test sequence from each year.

3. RESULTS & DISCUSSION

We grouped our results according to the two analyses conducted and described by the methods.

1. PHMM to predict HCV type 1 subtypes from NS5A

Training sequences of NS5A from HCV type 1 were scored according to the PHMM described in the methods (2.1) and a histogram was generated summarizing these scores (**Fig. 3a**). After we determined the 95% confidence interval of the mean scores for subtypes 1a and 1b (1695.1 ± 63.07 and 1789 ± 43.04, respectively), we scored the testing sequences according to the PHMM (**Fig. 3b**). Overall accuracy of the PHMM for predicting the correct HCV type 1 subtype from the NS5A sequence was 95.26%. Furthermore, the precision and recall of this method were 96.77% and 90.91%, respectively.

2. Determining rate of substitutions per nucleotide per year in the NS5A protein of HCV subtype 1b

After the most representative sequences for HCV subtype 1b for each year in question were determined from individual PHMM scores and multiple sequence alignments, we computed the Jukes-Cantor pairwise genetic distances of the sequences from one another. With these distances, we constructed a phylogenetic tree and then reconstructed the tree with respect to the 1999 sequence (**Fig. 4**).

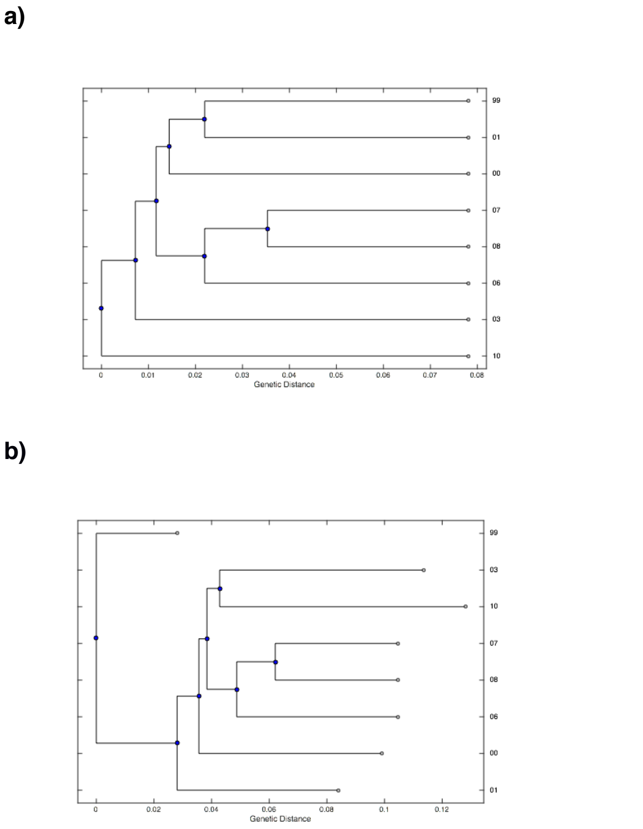
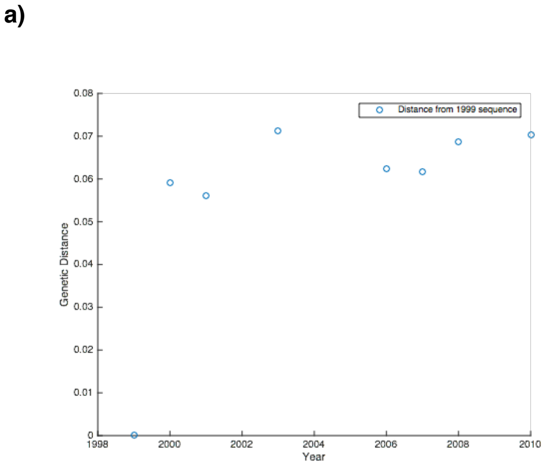


Figure 4. Phylogenetic trees. **a)** Phylogenetic tree representing genetic distances of each sequence from another. **b)** Phylogenetic tree reconstructed with respect to the earliest year, 1999.

According to **Figure 4b**, the most closely related sequences to the 1999 sequence are from years 2001 and 2000, respectively, and the most distantly related sequences to the 1999 sequence are from years 2007 and 2008, equally.

Using the genetic distances calculated from the sequence from year 1999, we generated a scatter plot to determine the rate of substitutions per nucleotide per year (**Fig. 5a**). Moreover, we reproduced this plot after removing the distance of zero of the sequence from 1999 from itself to obtain a more accurate rate (**Fig. 5b**). We also included a 95% confidence interval of the fit determined from this rate.



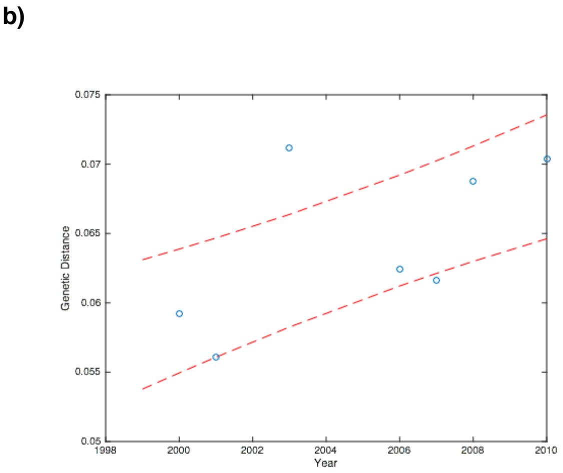


Figure 5. Genetic distance from 1999 sequence vs. year. **a)** Genetic distance with respect to the 1999 sequence, including the distance of the 1999 sequence from itself (zero). **b)** Genetic distance with respect to the 1999 sequence with 95% confidence interval.

The calculation of rate of substitutions per nucleotide per year from **Figure 5a** was 0.0036. The corrected rate of substitutions per nucleotide per year from **Figure 5b** was 0.000967. Finally, we determined the accuracy of this predicted rate by testing a random sequence from each year and determining whether or not its distance from the 1999 sequence was within the 95% confidence interval. We tested a random sequence from each year 100 times to obtain an average accuracy of 54.25%

CONCLUSION

In conclusion, we've developed a method for predicting HCV type 1 subtype from the NS5A protein and determined a rate of substitutions per nucleotide per year in NS5A of HCV subtype 1b. We show the success of our subtype prediction PHMM and the potential of the mutation rate prediction with accuracy results of 95.26% and 54.25%, respectively. Both methods could be improved with more sequenced HCV strains or from drawing sequences from additional databases. Moreover, the rate of mutation for HCV subtype 1a can be determined by the method described; however, for the scope of this project, subtype 1b was chosen as a result of more available sequences and because subtype 1b has been reported to have weaker virological response to interferon treatment. Strong results from the PHMM designed to predict HCV type 1 subtypes indicate a potential relationship between interferon sensitivity and HCV type 1 subtype, as the NS5A protein contains the interferon sensitivity region and has been implicated in resistance to typical interferon-ribaviron treatment. This study contains preliminary support of such a hypothesis and provides questions for further research.

REFERENCES

i. Division of Viral Hepatitis. (2012). Hepatitis c information for health professionals. Centers for Disease and Control. Retrieved from http://www.cdc.gov/hepatitis/HCV/index.htm

ii. Brassard, D. L., et al. (2002). Interferon-α as an immunotherapeutic protein. Journal of Leukocyte Biology. Retrieved from http://www.jleukbio.org/content/71/4/565.long#cited-by

iii. Enomoto N., Sakuma I., Asahina Y., Kurosaki M., Murakami T., Yamamoto C., Ogura Y., Izumi N., Marumo F., Sato C. (1996) Mutations in the nonstructural protein 5A gene and response to interferon in patients with chronic hepatitis C virus 1b infection.

iv. N. Engl. J. Med. 334:77–81.National Library of Medicine (2012). Ribavirin. PubMed Health. Retrieved from http://www.ncbi.nlm.nih.gov/pubmedhealth/PMHT0011979/?report=details

v. Pawlotsky, J-M. (2013). NS5A inhibitors in the treatment of hepatitis C. Journal of Hepatology, Volume 59, Issue 2, Pages 375-382, ISSN 0168-8278. Retrieved from http://www.sciencedirect.com/science/article/pii/S0168827813002092

vi. Sczyrba, A. (2005). Sequence analysis with distributed resources. Bielefeld University, Institute of Bioinformatics. Retrieved from http://bibiserv.techfak.uni-bielefeld.de/sadr2/databasesearch/hmmer/index.html

vii. Zein, N. (2000). Clinical significance of hepatitis c virus genotypes. Clin Microbiol Rev. 2000 Apr; 13(2): 223–235. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC100152/