Phylogenetic Reconstruction of Influenza B From 1940 to 1998. Are Means Important For The End?

L. M. Boykin[1,2] **H. Lu** [1] **C. A. Macken** [1]
lboykin@lanl.gov hlu@lanl.gov cam@lanl.gov
P. H. Mukundan [3] **A. K. Naik** [4]
phm@lanl.gov ashwin@lanl.gov

[1] Influenza Sequence Database, Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, New Mexico - 87544, USA
[2] Department of Biology, University of New Mexico, Albuquerque, New Mexico - 87131, USA
[3] Health Sciences Center, University of New Mexico, Albuquerque, New Mexico - 87131, USA
[4] Department of Biochemical and Biological Sciences, University of Houston, 3801 Cullen Blvd., Houston Science Center, Houston, Texas - 77023, USA

Keywords: influenza, phylogeny, DNArates, PAML, evolution

1 Introduction

Influenza has been a major cause of mortality and morbidity this century. Influenza B, according to recent reports is currently at a peculiar stage having two co-circulating lineages, which are evolving independently, with multiple sub lineages. Continuation of this trend is likely to lead to two antigenically distinct subtypes. Vaccines at present contain only one subtype and hence vaccines developed against one subtype may not be effective against the other.

To understand the evolutionary trend of Influenza B, it is very important to reconstruct the evolution of the virus as accurately as possible. Earlier attempts by others have shortcomings in reconstruction due to smaller dataset, unreliable methods or models [3, 4]. The goal of this project is to include as many sequences as possible and use the most accurate models and methods [1]. The availability of super-computing facility at the Los Alamos National Laboratory has made it possible for us to use the computationally intensive maximum likelihood methods along with models of molecular evolution.

We show in this study that the choice of evolutionary model significantly affects the branchlengths of the reconstructed tree whereas the topology is not significantly affected. We also report descrepencies in results due to the choice of the method used to determine the variable rates of substitution among sites.

2 Method and Results

2.1 Methods

Data: Sequences were obtained from Influenza Sequence Database(http://www.flu.lanl.gov) at Los Alamos National Laboratory, Los Alamos. 213 sequences of Influenza B were used for the final analysis.

Multiple Sequence alignment: Multiple sequence alignment was done using clustalW 1.7 and refined by hand using multiple sequence alignment editor, MASE and Wisconsin Genetic Group's GCG package.

Table 1: Comparison of DNArates and PAML.

Categories	1	2	3	4	5
PAML	711	0	0	176	142
DNArates	711	176	115	23	4

Phylogenetic Analysis: PAUP* was used for distance, parsimony and maximum likelihood analyses. Concurrently, fastDNAml, DNA rates and PAML were used to validate phylogenetic analysis.

Rates of substitution calculation: DNArates, PAML and Gamma Distribution by PAUP were used to determine rate of substitutions.

2.2 Results

The final tree was constructed using Hasegawa Kishino Yano Model of molecular evolution [2] along with variable rates of substitutions along the sequences. The tree obtained has similar topology compared to previously published trees, but with one major difference. The branchlengths are considerably shorter, possible because most of the previous trees were distance and parsimony based which do not consider branch lengths when searching for the trees. The Second major difference was the estimated time of divergence of the Yamagata-88 and Victoria-87 lineages. Previous reports have estimated the divergence time to be in the late 1970's. Our estimation seems to indicate the divergence time to be in the late 1960's. More work has to be done to conclusively estimate the time of divergence.

Two of the most widely used softwares to estimate the rates of DNA substitutions, DNA rates and PAML, were compared. A smaller representative data set of 71 taxa was chosen amoungst the 213 sequences. All 1031 sites were divided into 5 categories by both DNA rates and PAML. Table 1 shows the distribution of sites into categories by both PAML and DNA rates. PAML assigns a large number of sites to the highly variable categories while DNA rates has a more uniform distribution of sites in to categories.

References

[1] Lio, P. and Goldman, N., Models of molecular evolution and phylogeny, *Genome Res.*, 8(12):1233–1244, 1998.

[2] Hasegawa, M., Kishino, H. and Yano, T., Dating of the human-ape splitting by a molecular clock of mitochondrial DNA, *J. Mol. Evol.*, 22(2):160–174, 1985.

[3] Lindstrom, S.E., Hiromoto, Y., Nishimura, H., Saito, T., Nerome, R. and Nerome, K., Comparative analysis of evolutionary mechanisms of the hemagglutinin and three internal protein genes of influenza B virus, *J. Virol.*, 73(5):4413–4426, 1999.

[4] Luo, C., Morishita, T., Satou, K., Tateno, Y., Nakajima, K., and Nobusawa, E., Evolutionary pattern of Influenza B virus based on the HA and NS genes during 1940 to 1999: Origin of NS genes after 1997. *Archives of Virology*, 144:1881–1891, 1999.