Non-Molecular-Clock-Like Evolution following Viral Origins in *Homo sapiens*

Wendy Mok¹, Kelly Seto² and Jon Stone³

¹Department of Biochemistry and Biomedical Sciences, McMaster University, 1280 Main Street West, Hamilton ON L8S 4K1, Canada. ²Department of Molecular and Medical Genetics, University of Toronto, 1 King’s College Circle, Toronto ON M5S 1A8, Canada. ³Department of Biology and Origins Institute, McMaster University, 1280 Main Street West, Hamilton ON L8S 4K1, Canada.

Abstract: Researchers routinely adopt molecular clock assumptions in conducting sequence analyses to estimate dates for viral origins in humans. We used computational methods to examine the extent to which this practice can result in inaccurate ‘retrodiction.’ Failing to account for dynamic molecular evolution can affect greatly estimating index case dates, resulting in an overestimated age for the SARS-CoV-human infection, for instance.

Keywords: computational biology, epidemic, mutation, virus, SARS-CoV

Introduction

Dating when viruses acquired the ability to infect human genomes is paramount to managing public health. Infections could occur directly and repeatedly via other animal hosts (e.g. Human Immunodeficiency Virus, West Nile Virus). But, if infections were manifested secondarily, as a consequence from viral sequence substitutions that allowed sustained transmission among humans—as recent reports suggest for Avian Influenza Virus—a pandemic could ensue, with casualty numbers greatly exceeding those for influenza pandemics from the past century (Nicholls, 2006; Thomas and Noppenberger, 2007).

Pinpointing viral origins in humans enables researchers to extrapolate backward to estimate index case dates, calculate mutation and substitution rates, and document genetic events that permit efficient interspecies transmission and enhanced virulence (Chen et al. 2004). These data enable researchers to extrapolate forward to predict variability and develop vaccination or management programs to prevent or respond to potential global outbreaks. Herein, we show that adopting molecular clock assumptions can yield inaccurate estimated origin times, considering as an example data from the Severe Acute Respiratory Syndrome coronavirus (SARS-CoV) infection in humans.

We compared the S-gene in SARS-CoV sequences isolated from patients included in a recently published phylogenetic tree (He et al. 2004). This gene synthesizes the Spike-protein that is involved in virus-to-host-cell-epitope interactions, so sequence changes will affect evolutionary dynamics. We observed that stepwise genetic distance was greatest immediately following initial infection and diminished to a plateau thereafter (Fig. 1). Recognizing that this change in substitution rate would violate a molecular clock assumption and could cause pairwise genetic distances to yield inaccurate evolutionary divergence estimates (especially if genetic distance calculations were performed with respect to a reference sequence representing an hypothetical common ancestor), we quantified the extent to which failing to account for dynamic SARS-CoV evolution might affect estimating an origin time.

We developed a computer simulation program to emulate sequence evolution and used it to evolve *in silico* and according to the recently published phylogenetic tree (He et al. 2004) the SARS-CoV sequences. The computer simulation program determined substitution rates on the basis of a gamma distribution function (Fig. 1). It implemented prescriptions for designating time-points for events, such as strains becoming ‘extinct’ in patients, to accord as closely as possible to definite time-points, such as dates on which patients died. And it determined origin times according to a molecular clock assumption. We ran 1000 replicates to obtain an origin time distribution, from which we could obtain a representative, estimated (e.g. median) index case date (Fig. 2).
Materials and Methods

We obtained from Genbank 51 SARS-CoV sequences isolated from infected patients at different times throughout the epidemic and included in the aforementioned published phylogenetic tree (He et al. 2004). These constituted the available, unique sequences from the 61 that were included in that phylogenetic tree. We extracted from each among the 51 patient sequences a 3767-nucleotide sequence encoding the Spike (S) protein. We aligned these sequences using ClustalX (Thompson et al. 1997) and determined stepwise genetic distances between sequentially emerging strains (i.e. according to the phylogenetic tree) to calculate substitution rates (nucleotides per site per day).

We used the observed substitution rate variation (Fig. 1) and independently published average mutation rates (He et al. 2004; Lu et al. 2004) to define a gamma distribution function (e.g. Golding, 1983) relating time (days) to expected genetic distance (substitutions per site). We used this gamma function (α = 1, θ = 4, implemented using the function GammaDistribution[1, 4] + 0.18 in Mathematica (Wolfram Research Inc, 1988)) in the computer simulation program to evolve sequences according to the phylogenetic tree, which was rooted by using as outgroups sequences obtained from strains found in civet cats (He et al. 2004). Comparable results were obtained using one- and two-parameter molecular substitution models (Jukes and Cantor, 1969; Kimura, 1980). Reassortment was unconsidered. The computer simulation program calculated Hamming distances between an hypothetical ancestor and the sequences; paired time points with these distances according to the phylogenetic tree; performed linear regressions; and extrapolated backward to 0 distance, to estimate origin times. We ran 1000 replicates and determined the median and 95% confidence interval for the resulting distribution (Fig. 2), which allowed us to estimate a representative index case date.

Results

In our computer simulations, the origin time for SARS-CoV in humans was estimated to have transpired approximately 45 days (median t = −45,
Fig. 2) prior to the actual index case date (95% confidence interval: −85, −15 days). This would correspond to mid September–mid October, 2002; December 31, 2002; or January 26, 2003 for previously published real-world estimates (August–September 2002 (Lu et al. 2003); November 16, 2002 (He et al. 2004); and December 12, 2002 (Zeng et al. 2003)).

Discussion
Adopting a molecular clock assumption generated inaccurate estimated origin times for virtual SARS-CoV infections in humans, yielding estimated initial infection dates that differed in comparison to previously published estimates for actual index case dates. We note that previously published estimates ‘redate’ inaccurately, to months prior to or weeks following the real-world index case date (He et al. 2004; Lu et al. 2004) or months following the initial outbreak (Zeng et al. 2003). Those estimates were generated using sequences that were obtained between February and April 2003, whereupon molecular modification had stabilized (Fig. 1).

Adopting a molecular clock assumption might have been valid in those proficient analyses; however, extrapolating backward on the basis of that constant modification rate—especially using pairwise genetic distances—might have been misleading.

We propose that the nonlinear, rapid divergence exhibited by the SARS-CoV immediately after initially infecting humans (t = 0–25, Fig. 1) might represent a time period during which the virus ‘settled’ before becoming ‘comfortable’ (t >25, Fig. 1) in its new environment. Whether settling constitutes a general phenomenon could be tested with data from entire genomes (i.e. including all genes) and other viruses (e.g. the Avian Influenza Virus). In the meanwhile, we recommend subjecting virus sequences to computational, non-molecular clock assumption analyses (e.g. “relaxed phylogenetics” in Drummond et al. 2006) to estimate time-points for critical epidemiological phenomena, like the viral reassortment events associated with the first human SARS-CoV case in 1997 (He et al. 2004).

Acknowledgments
Funding was transferred from the Natural Sciences and Engineering Council of Canada (USRA Scholarship to WM and Discovery Grant 261590 to JRS). Critical and illuminating comments were transmitted by reviewers 1 and 2, B. Golding, and R. Morton, and L. de Lourdes Vazquez Paz.
References

Chen, H., Deng, G., Li, Z. et al. 2004. The evolution of H5N1 influenza viruses in ducks in Southern China. *PNAS*, 101:10452–57.

Drummond, A.J., Ho, SYW., Phillips, M.J. et al. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biology*, 4:e88.

Golding, B. 1983. Estimation of DNA and protein sequence divergence: An examination of some assumptions. *Mol. Biol. Evol.*, 1:125–42.

He, J.F., Peng, G.W., Min, J. et al. 2004. Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. *Science*, 303:1666–69.

Jukes, T.H. and Cantor, C.R. 1969. Evolution of protein molecules. In Munro HN, ed. Mammalian protein metabolism. *New York: Academic Press*, pp. 21–132.

Kimura, M. 1980. A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–20.

Nicholls, H. 2006. Pandemic influenza: the inside story. *PLoS Biol.*, 4:e50.

Lu, H., Zhao, Y., Zhang, J. et al. 2004. Date of origin of the SARS coronavirus strains. *BMC Infect. Dis.*, 4:3. (Accessed 10 July 2004. URL: http://www.biomedcentral.com/1471–2334/4/3).

Thomas, J.K. and Noppenberger, J. 2007. Avian influenza: a review. *Am. J. Health Syst. Pharm.*, 64:149–65.

Thompson, J.D., Gibson, T.J., Plewniak, F. et al. 1997. The ClustalX windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucl. Acid. Res.*, 24:4876–82.

Wolfram Research, Inc., 1988. Mathematica: A System for Doing Mathematics by Computer Version 2.1. Champaign: Wolfram Research, Inc.

Zeng, F., Chow, KYC. and Leung, F.C. 2003. Estimated timing of the last common ancestor of the SARS coronavirus. *N. Engl. J. Med.*, 349:2469–70.