

Influenza-specific amino acid substitution model

Cuong D.C., Vinh L.S., Quang L.S.

College of Technology, Vietnam National University Hanoi, Hanoi, Viet Nam; Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, United Kingdom

Abstract: Amino acid substitution model is a crucial component in protein sequence comparative systems such as protein sequence similarity searching, protein sequence alignment, and protein phylogenetic analysis. Although several general amino acid substitution models have been estimated from large protein databases, they might not be appropriate for analyzing specific species. In this paper, we apply the maximum likelihood approach to all influenza protein sequences to estimate an amino acid substitution model of so-called I09 for influenza viruses. Comparing I09 with fourteen other widely used models, we achieve remarkable results: (1) a likelihood improvement of phylogenetic trees based on I09 compared with other models. Precisely, I09 results in the best likelihood in 436 out of 489 cases tested; (2) tree topologies constructed with I09 and other models are frequently different indicating that the impact of I09 is not only on the likelihood improvement but also in tree topologies; (3) marked differences between I09 and other models revealing that existing models are not be able to capture the amino acid substitution process of influenza viruses. © 2009 IEEE.

Author Keywords: Amino acid substitution model; Influenza; Protein evolution

Index Keywords: Amino acid substitution; Influenza virus; Maximum likelihood approaches; Phylogenetic analysis; Phylogenetic trees; Protein database; Protein evolution; Protein sequence alignments; Protein sequences; Tree topology; Amino acids; Knowledge engineering; Maximum likelihood estimation; Organic acids; Systems engineering; Topology; Viruses; Proteins

Year: 2009

Source title: KSE 2009 - The 1st International Conference on Knowledge and Systems Engineering

Art. No.: 5361735

Page : 19-25

Link: [Scopus Link](#)

Correspondence Address: Cuong, D. C.; College of Technology, Vietnam National University Hanoi, Hanoi, Viet Nam; email: cuongdc@vnu.edu.vn

Sponsors: College of Technology; Vietnam National University

Conference name: 1st International Conference on Knowledge and Systems Engineering, KSE 2009

Conference date: 13 October 2009 through 17 October 2009

Conference location: Hanoi

Conference code: 79895

ISBN: 9.78E+12

DOI: 10.1109/KSE.2009.27

Language of Original Document: English

Abbreviated Source Title: KSE 2009 - The 1st International Conference on Knowledge and Systems

Engineering

Document Type: Conference Paper

Source: Scopus

Authors with affiliations:

- Cuong, D.C., College of Technology, Vietnam National University Hanoi, Hanoi, Viet Nam
- Vinh, L.S., College of Technology, Vietnam National University Hanoi, Hanoi, Viet Nam
- Quang, L.S., Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, United Kingdom

References:

- Felsenstein, J., (2004) *Infering Phylogenies*, Sunderland, Massachusetts: Sinauer Associates
- Yang, Z., (2006) *Computational Molecular Evolution*, 1st ed. Oxford University Press
- Setubal, C., Meidanis, J., (1997) *Introduction to Computational Molecular Biology*, 1st ed. PWS Publishing
- Thorne, J., *Models of protein sequence evolution and their applications* (2000) *Current Opinion in Genetics and Development*, 10, pp. 602-605
- Le, S.Q., Gascuel, O., *An improved general amino acid replacement matrix* (2008) *Molecular Biology and Evolution*, 25 (7), pp. 1307-1320. , DOI 10.1093/molbev/msn067
- Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C., *A model of evolutionary change in proteins* (1978) *Atlas of Protein Sequence Structure*, 5, pp. 345-352. , M. O. Dayhoff, Ed. Washington DC: National Biomedical Research Foundation
- Jones, D.T., Taylor, W.R., Thornton, J.M., *The rapid generation of mutation data matrices from protein sequences* (1992) *Comput. Appl. Biosci.*, 8, pp. 275-282
- Adachi, J., Hasegawa, M., *Model of amino acid substitution in proteins encoded by mitochondrial DNA* (1996) *Journal of Molecular Evolution*, 42 (4), pp. 459-468
- Whelan, S., Goldman, N., *A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach* (2001) *Molecular Biology and Evolution*, 18 (5), pp. 691-699
- J., M.A., G., P.B., M., J.I., Nickle, D.C., Heath, L., Pond, S.L.K., *Hiv-specific probabilistic models of protein evolution* (2007) *PLoS ONE*, 2, p. 503
- Le, Q.S., Vinh, L.S., *Limits of maximum likelihood models for protein phylogenies* (2008) *The Workshop on Knowledge, Language, and Learning in Bioinformatics*, Hanoi
- Nachmana, M.W., Crowella, S.L., *Estimate of the mutation rate per nucleotide in humans* (2000) *Genetics*, 156, pp. 297-304
- Zhao, Z., Li, H., Wu, X., Zhong, Y., Zhang, K., Zhang, Y.-P., Boerwinkle, E., Fu, Y.-X., *Moderate mutation rate in the SARS coronavirus genome and its implications* (2004) *BMC Evolutionary Biology*, 4, p. 21. , <http://www.biomedcentral.com/1471-2148/4/21>, DOI 10.1186/1471-2148-4-21
- D., M., Dimmic, M.W., Rest, J.S., G., R.A., *Rtrev: An amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny* (2002) *J Mol. Evol.*, 55, pp. 65-73
- Fauci, A.S., *Race against time* (2009) *Nature*, 435, pp. 423-424
- G., R., H., F., E., W.-W., W., C., Janies, D.A., Hill, A., *Genomic analysis and geographic visualization of the spread of avian influenza (h5n1)* (2007) *Systematic Biology*, 56, pp. 321-329
- Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., Yaschenko, E., *Database resources of the National Center for Biotechnology Information* (2008) *Nucleic Acids Research*, 36 (SUPPL. 1), pp. D13-D21. , DOI 10.1093/nar/gkm1000
- V., D., G., R., Y., W., J., G., M., P.-A., S., G.J., *The Vinh Nguyen "Multiple sublineages of influenza a virus (h5n1), vietnam, 2005-2007"* (2008) *Emerging Infectious Diseases*, 14, pp. 632-636

- G., O.T., C., T.M., Webster, R.G., Bean, W.J., Kawaoka, Y., Evolution and ecology of influenza A viruses (1992) *Microbiological Reviews*, 56, pp. 152-179
- Strimmer, K., Von Haeseler, A., Nucleotide substitution models (2003) *The Phylogenetics Handbook A Practical Approach to DNA and Protein Phylogeny*, pp. 72-100. , M Salemi and A.-M. Vandamme, Eds. Cambridge: Cambridge University Press
- Felsenstein, J., The number of evolutionary trees (1978) *Syst. Zool.*, 27, pp. 27-33
- Nei, M., Kumar, S., (2000) *Molecular Evolution and Phylogenetics*, , Oxford University Press
- Fitch, W., Margoliash, E., A method for estimating the number of invariant amino acid position in a gene using cytochrome c as a model case (1967) *Biochem. Gene*, 1, pp. 65-71
- Churchill, G.A., Von Haeseler, A., Navidi, W.C., Sample size for phylogenetic inference (1992) *Mol. Biol. Evol.*, 9, pp. 753-769
- Yang, Z., Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites (1993) *Molecular Biology and Evolution*, 10 (6), pp. 1396-1401
- Gu, X., Fu, Y.-X., Li, W.-H., Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites (1995) *Mol. Biol. Evol.*, 12, pp. 546-557
- Edgar, R.C., MUSCLE: Multiple sequence alignment with high accuracy and high throughput (2004) *Nucleic Acids Research*, 32 (5), pp. 1792-1797. , DOI 10.1093/nar/gkh340
- Castresana, J., Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis (2000) *Molecular Biology and Evolution*, 17 (4), pp. 540-552
- Guindon, S., Gascuel, O., A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood (2003) *Systematic Biology*, 52 (5), pp. 696-704. , DOI 10.1080/10635150390235520