FLU, an amino acid substitution model for influenza proteins

Dang C.C., Le Q.S., Gascuel O., Le V.S.
College of Technology, Vietnam National University Hanoi, 144 Xuan Thuy, Cau Giay, Hanoi, Viet Nam; Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom; Methodes et Algorithmes Pour la Bioinformatique, LIRMM, Universite Montpellier II, Montpellier, France

Abstract: Background. The amino acid substitution model is the core component of many protein analysis systems such as sequence similarity search, sequence alignment, and phylogenetic inference. Although several general amino acid substitution models have been estimated from large and diverse protein databases, they remain inappropriate for analyzing specific species, e.g., viruses. Emerging epidemics of influenza viruses raise the need for comprehensive studies of these dangerous viruses. We propose an influenza-specific amino acid substitution model to enhance the understanding of the evolution of influenza viruses. Results. A maximum likelihood approach was applied to estimate an amino acid substitution model (FLU) from 113, 000 influenza protein sequences, consisting of 20 million residues. FLU outperforms 14 widely used models in constructing maximum likelihood phylogenetic trees for the majority of influenza protein alignments. On average, FLU gains 42 log likelihood points with an alignment of 300 sites. Moreover, topologies of trees constructed using FLU and other models are frequently different. FLU does indeed have an impact on likelihood improvement as well as tree topologies. It was implemented in PhyML and can be downloaded from ftp://ftp.sanger.ac.uk/pub/1000genomes/lsq/FLU or included in PhyML 3.0 server at http://www.atgc-montpellier.fr/phyml/. Conclusions. FLU should be useful for any influenza protein analysis system which requires an accurate description of amino acid substitutions.

Index Keywords: amino acid; cohort analysis; database; evolutionary biology; phylogenetics; protein; topology; virus; Orthomyxoviridae; virus protein; amino acid substitution; article; biological model; chemistry; genetics; human; Orthomyxovirus; statistical model; Amino Acid Substitution; Humans; Likelihood Functions; Models, Genetic; Orthomyxoviridae; Viral Proteins

Year: 2010
Source title: BMC Evolutionary Biology
Volume: 10
Issue: 1
Art. No.: 99
Link: Scorpus Link
Chemicals/CAS: Viral Proteins
Correspondence Address: Le, Q. S.; Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom; email: lsq@sanger.ac.uk
ISSN: 14712148
DOI: 10.1186/1471-2148-10-99

PubMed ID: 20384985
Language of Original Document: English
Abbreviated Source Title: BMC Evolutionary Biology
Document Type: Article
Source: Scopus
Authors with affiliations:
- Dang, C.C., College of Technology, Vietnam National University Hanoi, 144 Xuan Thuy, Cau Giay, Hanoi, Viet Nam
- Le, Q.S., Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom
- Gascuel, O., Methodes et Algorithmes Pour la Bioinformatique, LIRMM, Universite Montpellier II, Montpellier, France
- Le, V.S., College of Technology, Vietnam National University Hanoi, 144 Xuan Thuy, Cau Giay, Hanoi, Viet Nam
References:
- Felsenstein, J., (2004) Infering Phylogenies, , Sunderland, Massachusetts, US: Sinauer Associates
- Ziheng, Y., (2006) Computational Molecular Evolution, , Oxford, UK: Oxford University Press 1
- Opperdoes, F.R., Phylogenetic analysis using protein sequences (2003) The Phylogenetics Handbook A Practical Approach to DNA and Protein Phylogeny, pp. 207-235. , Cambridge: Cambridge University Press Salemi M, Vandamme AM
- Setubal, C., Meidanis, J., (1997) Introduction to Computational Molecular Biology, , Boston, Massachusetts, US: PWS Publishing 1
- Thorne, J., Models of protein sequence evolution and their applications (2000) Currrent Opinion in Genetics and Development, 10, pp. 602-605. , 10.1016/S0959-437X(00)00142-8
- Le, S., Gascuel, O., An improved general amino acid replacement matrix (2008) Mol Biol Evol, 25, pp. 1307-1320. , 10.1093/molbev/msn067. 18367465
- 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. , Washington DC: National Biomedical Research Foundation Dayhoff MO
- 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. , 1633570
- Adachi, J., Hasegawa, M., Model of amino acid substitution in proteins encoded by mitochondrial DNA (1996) J Mol Evol, 42, pp. 459-468. , 10.1007/BF02498640. 8642615
- Whelan, S., Goldman, N., A general empirical model of protein evolution derived from multiple protein families using a maximum likelihood approach (2001) Mol Biol Evol, 18, pp. 691-699. , 11319253
- Dimmic, M.W., Rest, J.S., Mindell, D.P., Goldstein, R.A., RtREV: An amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny (2002) J Mol Evol, 55, pp. 65-73. , 10.1007/s00239-001-2304-y. 12165843
- Nickle, D.C., Heath, L., Jensen, M.A., Gilbert, P.B., Mullins, J.I., Pond, S.K., HIV-specific probabilistic models of protein evolution (2007) PLoS ONE, 2, p. 5503. , 10.1371/journal.pone.0000503. 17551583
- Fauci, A., Race against time (2009) Nature, 435, pp. 423-424. , 10.1038/435423a
- Ghedin, E., Sengamalay, N., Shumway, M., Zaborsky, J., Feldblyum, T., Subbu, V., Spiro, D., Salzberg, S., Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution (2005) Nature, 437, pp. 1162-1166. , 10.1038/nature04239. 16208317
- Janies, D.A., Hill, A., Guralnick, R., Habib, F., Waltari, E., Wheeler, W.C., Genomic analysis and geographic visualization of the spread of avian influenza (H5N1) (2007) Systematic Biology, 56, pp. 321-329. , 10.1080/10635150701266848. 17464886
- Bao, Y., Bolotov, P., Dernovoy, D., Kiryutin, B., Zaslavsky, L., Tatusova, T., Ostell, J., Lipman, D., The influenza virus

resource at the national center for biotechnology information (2008) J Virol, 82, pp. 596-601. , 10.1128/JVI.02005-07. 17942553

- Nguyen, T., Nguyen, T., Vijaykrishna, D., Webster, R., Guan, Y., Malik Peiris, J., Smith, G., Multiple sublineages of influenza A virus (H5N1), Vietnam, 2005-2007 (2008) Emerging Infectious Diseases, 14, pp. 632-636. , 10.3201/eid1404.071343. 18394281
- Guindon, S., Gascuel, O., A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood (2003) Syst Biol, 52, pp. 696-704. , 10.1080/10635150390235520. 14530136
- Akaike, H., A new look at the statistical model identification (1974) IEEE Trans Automat Contr, 19, pp. 716-722. , 10.1109/TAC.1974.1100705
- Kishino, H., Hasegawa, M., Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea (1989) J Mol Evol, 29, pp. 170-179. , 10.1007/BF02100115. 2509717
- Goldman, N., Anderson, J., Rodrigo, A., Likelihood-based tests of topologies in phylogenetics (2000) Syst Biol, 49, pp. 652-670. , 10.1080/106351500750049752. 12116432
- Pagel, M., Meade, A., Mixture models in phylogenetic inference (2005) Mathematics of Evolution and Phylogeny, pp. 121-142. , Oxford, UK: Oxford University Press Gascuel O
- Edgar, R.C., MUSCLE: Multiple sequence alignment with high accuracy and high throughput (2004) Nucl Acids Res, 32, pp. 1792-1797. , 10.1093/nar/gkh340. 15034147
- Kevin, L., Sindhu, R., Serita, N., Randal, L., Tandy, W., Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees (2009) Science, 324, pp. 1561-1564. , 10.1126/science.1171243. 19541996
- Boni, M.F., Zhou, Y., Taubenberger, J.K., Holmes, E.C., Homologous recombination is very rare or absent in human influenza A virus (2008) Journal of Virology, 82 (10), pp. 4807-4811. , DOI 10.1128/JVI.02683-07
- He, C.Q., Xie, Z.X., Han, G.Z., Dong, J.B., Wang, D., Liu, J.B., Ma, L.Y., Li, G.R., Homologous recombination as an evolutionary force in the avian influenza a virus (2009) Mol Bio Evol, 26, pp. 177-187. , 10.1093/molbev/msn238
- Castresana, J., Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis (2000) Molecular Biology and Evolution, 17, pp. 540-552. , 10742046
- Strimmer, K., Haeseler, A.V., Nucleotide substitution models (2003) The Phylogenetics Handbook A Practical Approach to DNA and Protein Phylogeny, pp. 72-100. , Cambridge: Cambridge University Press Salemi M, Vandamme AM
- Felsenstein, J., Evolutionary trees from DNA sequences: A maximum likelihood approach (1981) Journal of Molecular Evolution, 17, pp. 368-376. , 10.1007/BF01734359. 7288891
- Fitch, W.M., 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. , 10.1007/BF00487738
- Churchill, G.A., Haeseler, A.V., Navidi, W.C., Sample size for phylogenetic inference (1992) Mol Biol Evol, 9, pp. 753-769. , 1630311
- 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. , 7659011
- Klosterman, P., Uzilov, A., Bendana, Y., Bradley, R., Chao, S., Kosiol, C., Goldman, N., Holmes, I., XRate: A fast prototyping, training and annotation tool for phylo-grammars (2006) BMC Bioinformatics, 7, p. 428. , 10.1186/1471-2105-7-428. 17018148