Evolutionary Conservation of the PA-X Open Reading Frame in Segment 3 of Influenza A Virus

Mang Shi, Brett W. Jagger, Helen M. Wise, Paul Digard, Edward C. Holmes, and Jeffery K. Taubenberger

Center for Infectious Disease Dynamics, Department of Biology, The Pennsylvania State University, University Park, Pennsylvania, USA; Viral Pathogenesis and Evolution Section, Laboratory of Infectious Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland, USA; The Roslin Institute, The University of Edinburgh, Midlothian, Scotland, United Kingdom; and Fogarty International Center, National Institutes of Health, Bethesda, Maryland, USA

PA-X is a fusion protein of influenza A virus encoded in part from a +1 frameshifted X open reading frame (X-ORF) in segment 3. We show that the X-ORFs of diverse influenza A viruses can be divided into two groups that differ in selection pressure and likely function, reflected in the presence of an internal stop codon and a change in synonymous diversity. Notably, truncated forms of PA-X evolved convergently in swine and dogs, suggesting a strong species-specific effect.

It was recently determined that segment 3 of influenza A virus (IAV) encodes a second protein, made in part from a +1 open reading frame (ORF) embedded within the PA gene. This ORF is accessed via ribosomal frameshifting to produce a fusion protein, denoted PA-X, whose N-terminal 191 amino acids are derived from the PA ORF, while the C-terminal sequence, most commonly 61 amino acids long, is derived from the PA ORF (3); this is indicative of selective constraint against nonsynonymous changes in PA-X and hence of functional importance. However, some strains of influenza A virus possess stop codons in the X-ORF, leading to a truncated PA-X protein. The most notable examples are the human 2009 pandemic H1N1 viruses (H1N1pdm), which possess a TGG (Trp)-to-TAG (stop) nonsense mutation at codon 42 in the X-ORF, resulting from a synonymous mutation in the PA gene (6). To determine how PA-X has evolved across IAV as a whole and particularly to instances of cross-species transmission, we performed an evolutionary analysis of PA-X in 10,164 influenza A viruses that reveal the frameshifted portion of the PA-X protein. These data indicate that segment 3 of influenza B virus.

We assembled all available nonidentical PA protein-coding sequences from GenBank and translated them in frame 1 to reveal the frameshifted portion of the PA-X protein. These data comprised avian influenza (all subtypes), human H3N2, human seasonal H1N1, human H1N1pdm, swine “classical” (CS) H1N1, swine “Eurasian” (EA) H1N1, swine “triple-reassortant” (TR) H1N1, equine H3N8, equine H7N7, canine H3N8, canine H3N2, and bat influenza virus (italics in Table 1), although the truncated proteins in the very small equine H7N7 and bat influenza virus data sets are due to stop codon mutations other than those at codon 42. The CS H1N1 viruses are particularly noteworthy, as the truncated form of PA-X occurs in a cluster of viruses sampled between 1985 and 2009, which fall within a group of generally older (1930 to 2006) swine CS viruses that possess a full-length PA-X (Fig. 2). This phylogenetic pattern suggests that swine CS viruses with the truncated form of PA-X were directly derived from those with full-length PA-X sequences. In addition, as the origins of the PA segment in human H1N1pdm lie with the TR swine influenza virus (2), the stop codon mutation at X-ORF codon 42 in this virus was clearly inherited from swine.

Overall, our phylogenetic analysis suggests that the nonsense mutation at X-ORF codon 42 was fixed at least four times independently—twice in swine and twice in dogs. Uniquely,

Received 29 June 2012 Accepted 30 August 2012 Published ahead of print 5 September 2012

Copyright © 2012, American Society for Microbiology. All Rights Reserved.
doi:10.1128/JVI.01677-12
First number is the total data set. The representative sample size used for the analysis of the non-stop codon state in the stop codon group. H1N1pdm, 2009 pandemic H1N1; swine CS, "classical" swine H1N1; swine EA, Eurasian avian H1N1-like swine H1N1; swine TR, triple-reassortant swine H1N1.

The classical swine viruses can be divided into two groups representing those with and without the stop codon mutation at codon 42. Note that there has been a single reversion to the non-stop codon state in the stop codon group. Resting X-ORF is substantially higher than that for the X-ORF encoded in frame 0 (Table 1). Although accurately estimating selection pressures in sequences with dual reading frames is notoriously difficult (7), this observation indicates that PA-X is able to tolerate a high number of amino acid changes. As synonymous mutations in frame 0 will either be nonsynonymous or nonsense mutations in frame 1, this is likely to be the case for many viral proteins encoded by +1 ORFs (5).

The conservation of the decanucleotide frameshift motif (Fig. 1) and the maintenance of the full-length +1 X-ORF in the majority of IAV genomes infecting diverse host species (6) suggest that PA-X is important for influenza A virus biology. Interestingly,

\[
\begin{align*}
\text{d}_s/\text{d}_k \text{ for:} & \\
\text{X-ORF} & \text{frame 0} & \text{Rest} & \text{frame 0} & \text{X-ORF} & \text{frame 0} & \text{Rest} & \text{frame 0} & \text{d}_s/\text{d}_k \\
\text{Avian influenza} & 4,361 (80) & 19 (2) & 0.50 & 0.52 & 4.08 & 16.44 & 0.25 & 0.12 & 3.20 \\
\text{Human H3N2} & 2,296 (49) & 8 (5) & 0.09 & 0.09 & 0.93 & 1.62 & 0.43 & 0.16 & 4.92 \\
\text{Human H1N1s} & 853 (50) & 3 (1) & 0.10 & 0.11 & 0.53 & 1.23 & 0.57 & 0.10 & 3.88 \\
\text{Human H1N1pdm} & 1,916 (56) & 1,914 (1,914) & 0.09 & 0.10 & 2.18 & 2.18 & 1.00 & 0.08 & 6.94 \\
\text{Swine CS} & 121 (52) & 6 (3) & 0.17 & 0.17 & 1.95 & 1.67 & 11.7 & 0.08 & 21.29 \\
\text{Swine CS–stop?} & 99 (53) & 98 (98) & 0.15 & 0.16 & 3.18 & 3.60 & 0.88 & 0.10 & 3.95 \\
\text{Swine EA} & 152 (60) & 0 (0) & 0.33 & 0.28 & 3.18 & 2.76 & 1.15 & 0.14 & 5.58 \\
\text{Swine TR} & 248 (73) & 225 (225) & 0.46 & 0.30 & 2.18 & 1.40 & 1.40 & 0.41 & 3.26 \\
\text{Equine H3N8} & 80 (35) & 0 (0) & 0.12 & 0.09 & 0.25 & 0.73 & 0.34 & 0.49 & 1.15 \\
\text{Equine H7N7} & 2 (2) & 2 (0) & 0.01 & 0.01 & 0.23 & 0.19 & 0.12 & 1.06 & 2.6 \\
\text{Bat influenzaa} & 2 (2) & 2 (0) & 0.01 & 0.01 & 0.14 & 0.13 & 1.08 & 0.05 & 0.59 \\
\text{Canine H3N8} & 26 (6) & 25 (24) & 0.03 & 0.01 & 0.07 & 0.05 & 1.40 & 0.41 & 1.26 \\
\text{Canine H3N2} & 8 (6) & 8 (7) & 0.03 & 0.01 & 0.16 & 0.05 & 3.20 & 0.18 & 4.13 \\
\text{Influenza B} & 190 (34) & 190 (AUb) & 0.05 & 0.06 & 1.07 & 1.11 & 0.96 & 0.05 & NAc \\
\end{align*}
\]

\footnote{a Virus groups with truncated PA-X proteins and associated data are in italics.}
\footnote{b Human H1N1s, seasonal H1N1; human H1N1pdm, 2009 pandemic H1N1; swine CS, “classical” swine H1N1; swine EA, Eurasian avian H1N1-like swine H1N1; swine TR, triple-reassortant swine H1N1.}
\footnote{c Reading frame encoding the PA protein.}
\footnote{d First number is the total data set. The representative sample size used for the analysis of d_s and d_k is shown in parentheses.}
\footnote{e First number is the number of viruses from the total data set that are truncated because of premature stop codons. The number of viruses with stop codon mutations at codon 42 is shown in parentheses.}
\footnote{f Because the very small sample size these sequences were analyzed with MEGA5 rather than PAML.}
\footnote{g AU, sequence alignment is uncertain in this region.}
\footnote{h For the full-length X-ORF the analysis of d_s/d_k was performed on all 61 amino acids present in the protein, while for sequences encoding the truncated protein this analysis was restricted to the 41 amino acids prior to the position corresponding to the stop codon.}
\footnote{i NA, contains too many stop codons for meaningful analysis.}

**TABLE 1** Frequency of stop codon mutations and the numbers of synonymous and nonsynonymous nucleotide substitutions per site for different influenza viruses
there are other possible synonymous mutations in the PA gene that would result in a PA-X truncated more severely than the commonly observed 41-amino-acid product of X-ORF. That these are not often seen suggests that stop codon mutations at X-ORF codon 42 renders the protein functionally distinct from products of 61-codon X-ORF isolates. As there is no evidence for decreased synonymous variability in the overlapping 0 frame of these truncated isolates, there are likely to be few selective constraints on X-ORF in these particular lineages. It is therefore probable that the protein domains encoded by the truncated X-ORFs have lost or altered functionality compared to the PA-Xs encoded by full-length X-ORFs. This hypothesis will need to be evaluated experimentally, especially in the context of particular host species. Indeed, as X-ORF protein truncation appears to be associated with IAV lineages circulating in particular hosts, i.e., pigs and dogs, there may be some species specificity to the evolution of a truncated PA-X protein. That a 41-amino-acid X-ORF protein evolved convergently in both IAV subtypes that infect dogs (H3N2 and H3N8) is particularly noteworthy and suggests that the truncation of this protein may be associated with the adaptation and emergence of influenza virus in this host species.

REFERENCES