Large-scale Analysis of Influenza A Virus Sequences Reveals Universally Conserved Residues of Matrix Proteins

Vivek Darapaneni¹*

¹ Department of virology and computational biochemistry, Saket Institute for Biomolecular Research, Visakhapatnam, India

Abstract
The matrix proteins of Influenza A Virus are multifunctional proteins. The matrix proteins of Influenza A Virus play imperative roles in the virus life cycle. The objective of the present study was to identify the residue conservation in the matrix proteins of Influenza A Virus. The study was based on 2836 amino acid sequences for the M1 protein and 3331 sequences for the M2 protein. Both the matrix proteins showed similar level of sequence conservation. On the whole, this study exposed residues which are universally conserved among different viral subtypes. These universally conserved residues might be involved in either structure stabilizing or protein-protein interactions. The conserved residues identified in the present study in conjunction with structural analysis of matrix proteins could form basis for designing universal anti-influenza drugs which are resistant to mutations arising in the future.

Keywords: Conservation; Universally; Drugs; Mutation; Resistance; Anti-influenza

Peer Reviewer: Tzi Bun NG, School of Biomedical Sciences, The Chinese University of Hong Kong, Hong Kong, China; Kai Zhang, Department of Oral Biology, The State University of New York at Buffalo, United States

Received: December 2, 2014; Accepted: January 15, 2015; Published: January 20, 2015

Competing Interests: The authors have declared that no competing interests exist.

Copyright: 2015 Darapaneni V et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

*Correspondence to: Vivek Darapaneni, Department of virology and computational biochemistry, Saket Institute for Biomolecular Research, Visakhapatnam, India; Email: darapanenivivek.sibr@gmail.com
Introduction

Influenza virus belongs to orthomyxoviridae family, which is enveloped with negative sense RNA genome. A, B and C are major serotypes of influenza virus [1]. Influenza A Virus (IAV) with its global dominance is a significant pathogen among the three serotypes. IAV is responsible for 250,000 to 500,000 human deaths worldwide [2] and it also results in considerable losses among domesticated birds [3]. IAV is responsible for a devastating pandemic during 1918/1919 ‘Spanish Flu’ (H1N1) which was responsible for at least 40 million deaths [4, 5] and for pandemics with reduced severity in the years: 1957 ‘Asian Influenza’ (H2N2), 1968 ‘Hong Kong Influenza’ (H3N2), 1977 ‘Russian Influenza’ (H1N1) and 2009 ‘Swine Flu’ (H1N1) [6-8]. When a pathogenic IAV of animal origin attains the potential for efficient human-human transmission, then influenza pandemics seems to arise [9, 10]. This potential can be achieved either by mutation or reassortment of RNA segments of animal and human viruses [9, 10]. In 1997, a highly pathogenic avian influenza H5N1 virus has emerged which can be a threat in the future. The emergence of H5N1 virus in the year 1997 and H1N1 virus in the year 2009 carrying drug resistant mutations have increased concerns [11, 12]. In recent past, pathogenic avian influenza viruses H6N1, H7N9, H7N7, H9N2 and H10N7 have acquired the ability to infect humans [13]. Therefore, in the future it is highly likely that a novel avian influenza A virus will acquire the potential to transmit efficiently among humans and results in a devastating pandemic. Therefore, it is of paramount importance to study the evolution of influenza viral proteins.

Viral nucleocapside of IAV consists of eight separate segments and segment seven encodes the matrix (M) proteins, namely M1 protein and M2 protein [1]. The M1 protein is a structural protein which is found abundantly in the virion. Being a multifunctional protein, the M1 protein plays an important role in the infectious lifecycle of the virus. It stabilizes the architecture of the virion by forming matrix layer between the lipid envelope and vRNPs [14]. It plays critical role: in inhibition of RNA transcription [15-17]; in regulating the trafficking of vRNPs across nucleus [18, 19] and in nuclear localization of RNA by interacting with vRNPs and RNA [20-22]. Through a network of protein–protein interactions, it plays an essential role in virion assembly [22]. It plays an important role in virus budding by bringing viral components to the site of budding [23], this is achieved through interactions with viral factors such as hemagglutinin, neuraminidase and matrix protein 2 [24-26] and through recruitment of host cell components essential for the completion of the bud [24, 27, 28]. The M1 protein (Mw=28kDa) is 252 amino acids residues long, consisting of four domains namely, N-terminal domain (NTD; residues 1-67), linker domain (LD; residues 68-87), middle domain (MD; residues 88-165) and C-terminal domain (CTD; residues 166-252) [22].

The M2 protein (Mw=11kDa) is 97 amino acids residues long transmembrane protein. The M2 protein exists as a homo-tetramer and plays an important role in the lifecycle of the virus [29, 30]. The M2 protein monomer consists of three domains namely, N-terminal domain (NTD; residues 1-23; extra-cellular), transmembrane domain (TD; residues 24-46) and C-terminal domain (CTD; residues 47-97; intracellular) [31]. The proton channel activity of M2 protein is significant for releasing the vRNPs from M1-vRNP complex into cytosol [32] during early phase of infection. The M2 protein was recently found to impede the fusion of autophagosomes to lysosomes, thereby helping in survival of infected cell [33]. M2 protein plays important role in viral budding, scission and the release of the virus [34]. Recently
M2 protein was shown to induce cell apoptosis by stabilizing Hsp40-P58IPK complex, thereby affecting the lifecycle of infected cell and influenza virus replication [35]. Specific regions of functional importance with corresponding amino acid residue positions are shown in table 1.

**Table 1** Functional regions of influenza A virus matrix proteins.

<table>
<thead>
<tr>
<th>Proteins</th>
<th>Functions</th>
<th>Residue position</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Matrix Protein 1</strong></td>
<td>M1 Protein and Cellular Receptor of Activated C Kinase Interaction Region</td>
<td>1-67 [36]</td>
</tr>
<tr>
<td></td>
<td>Death Domain Associated Protein 6 Binding Motif</td>
<td>23-48 [37]</td>
</tr>
<tr>
<td></td>
<td>Nuclear Export Signal</td>
<td>59-68 [38]</td>
</tr>
<tr>
<td></td>
<td>Transcription Inhibition Domains</td>
<td>90-108 and 129-164 [19]</td>
</tr>
<tr>
<td></td>
<td>M1 Protein and vRNP Interaction Region</td>
<td>91-105 [39, 40]</td>
</tr>
<tr>
<td></td>
<td>Nuclear Localization Signal</td>
<td>101-105 [39, 41, 42]</td>
</tr>
<tr>
<td></td>
<td>Nuclear Export Protein Binding Motif</td>
<td>101-105 [39, 41, 42]</td>
</tr>
<tr>
<td></td>
<td>RNA Binding Domain</td>
<td>76-105 [40]</td>
</tr>
<tr>
<td></td>
<td>Zinc Binding Domain</td>
<td>148-162 [15, 16]</td>
</tr>
<tr>
<td></td>
<td>Heat Shock Protein 70 Binding Region</td>
<td>128-165 [43]</td>
</tr>
<tr>
<td></td>
<td>Heat Shock Cognate Protein 70 Binding Site</td>
<td>102-201 [40]</td>
</tr>
<tr>
<td></td>
<td>M1 Protein and Cyclophilin A Interaction Region</td>
<td>88-165 [19]</td>
</tr>
<tr>
<td></td>
<td>Phosphorylation Sites</td>
<td>53, 70, 161, 185 [37]</td>
</tr>
<tr>
<td></td>
<td>Sumoylation Site</td>
<td>242 [44]</td>
</tr>
<tr>
<td><strong>Matrix protein 2</strong></td>
<td>Channel Activity Of M2</td>
<td>37-41 [45]</td>
</tr>
<tr>
<td></td>
<td>Cholesterol Recognition/Interaction Amino Acid Consensus (CRAC) Motifs</td>
<td>48-62 [46]</td>
</tr>
<tr>
<td></td>
<td>Inhibition Of Epithelial Sodium Channels</td>
<td>53-62 [47]</td>
</tr>
<tr>
<td></td>
<td>Acylation Site</td>
<td>50 [48]</td>
</tr>
<tr>
<td></td>
<td>M1-M2 Interaction Region</td>
<td>71-73 [49]</td>
</tr>
<tr>
<td></td>
<td>Phosphorylation Site</td>
<td>64 [50]</td>
</tr>
<tr>
<td></td>
<td>LC3 Interacting Region Motif</td>
<td>91-94 [51]</td>
</tr>
<tr>
<td></td>
<td>M2 Protein and Human Annexin A6 Interaction Region</td>
<td>44-97 [52]</td>
</tr>
</tbody>
</table>

Currently there are no licensed drugs available targeting the M1 protein and there are two licensed drugs targeting the M2 protein which are channel blockers namely, rimantadine and amantadine [33]. Due to the emergence of high rate of drug resistant mutation, these drugs are rendered ineffective [12]. Therefore it is extremely important to study these mutations in order to design new drugs, which are resistant to mutations. Analyses of sequences at protein or nucleotide level were carried out previously [53-56]. These studies were carried on both small number of sequences and large number of sequences, but no apparent definition of residue conservation was given. The aims of the present study were to detect the degree of conservation of the matrix proteins among all IAV subtypes from all the hosts, to facilitate the identification of universally conserved sites. The mapping of the conservation scores onto the structures of the matrix proteins suggests prospective antiviral binding sites which are resistant to mutations. Additionally, our results propose highly conserved sites with unidentified functions, which might be experimentally validated in future.
Methods

Sequence analysis and protein structure

The sequences of the IAV matrix proteins were obtained from National Centre for Biotechnology Information (NCBI) influenza virus resource [57]. Full-length sequences from all hosts and subtypes were chosen, whilst duplicate protein sequences were removed. For alignment of the collected proteins sequences, MUSCLE version 3.8 [58] was used with default parameters. Multiple refinements of the obtained alignment were carried out resulting in 26-30 iterations, until no further improvement was attained. For sequence conservation plots, plotcon program (EMBOSS package) (http://emboss.bioinformatics.nl/cgi-bin/emboss/plotcon) was used with comparison matrix EBLOSUM62 (default) and window size of 10 [59]. The experimental structure of the M1 protein (residues 1-158) was obtained from Protein Data Bank (PDB) [60] with the identifier 1AA7 [41]. The experimental structure of the M2 protein (residues 23-60) was obtained from Protein Data Bank (PDB) [60] with the identifier 2LOJ [61].

Conservation analysis

By providing multiple sequence alignment and protein structure files as an input, conserved regions were identified and mapped onto the experimental protein structures using ConSurf server (http://consurf.tau.ac.il/) [62-65]. By taking evolutionary relationships among protein sequences into account, ConSurf algorithm produces consequential conservation scores. ConSurf algorithm gives more emphasis to those protein sequences which are evolutionarily distant, thus producing conservation scores which are significant [62-65]. The ensuing conservation scores are criterion scores with an average of 0 and a standard deviation of 1. The residues with score < 0 denote higher conservation and those with score > 0 are variable residues [62-65]. The Bayesian algorithm is used to evaluate the confidence intervals of calculated conservation scores [62-65]. The conservation score given by ConSurf server is divided into scale of nine grades which are given for the purpose of visualization [62-65]. Most variable positions are placed in grade one (turquoise), intermediately conserved positions are placed in grade five (white), and most conserved positions are placed in grade nine (maroon) [62-65].

Results

Multiple alignment of protein sequences

For the M1 protein 2836 sequences were obtained from the database, which were mainly from avian (40%), human (25%) and swine (22%) viruses. For the M2 protein 3331 sequences were obtained from the database, which were mainly from avian (37%), human (26%) and swine (22%) viruses. The figure 1 shows the overview of the sequence alignment of the matrix proteins, revealing the fact that albeit the sequences belong to different lineages of IAV, they had a high degree of similarity. In general, both the proteins showed similar level of sequence conservation, but the M2 protein displayed larger variation in
conservation pattern when compared to the M1 protein. In particular, at residue position 28 of the M2 protein sequence there was a minimal conservation as specified by the minimal similarity score in the figure 1.

**Figure 1** Overview of the sequence alignments for the M1 protein (a) and the M2 protein (b) obtained using plotcon program [59] with window size of ten residues.
Conserved and variable residues

The variable and conserved residues in the matrix proteins were identified using ConSurf server [62-65] and are illustrated in figure 2 and figure 3. The highly variable residues of grades 1-2, conserved residues of grade 7 and highly conserved residues of grades 8-9 are clustered together in Table 2. The sequence conservation of matrix proteins is shown in figure 2. The conservation scores were projected onto the backbone of the protein structures and are shown in figure 3.

Table 2 Highly Variable residues (grades 1-2); Conserved residues (grades 7) and highly conserved residues (grades 8-9) of the matrix proteins identified using the ConSurf server [62-65]. Residues shown in bold did not show any variation among the sequences analysed.

<table>
<thead>
<tr>
<th>Residue Classification</th>
<th>M1 protein</th>
<th>M2 protein</th>
</tr>
</thead>
</table>
Figure 2 The amino acid sequences of the M1 protein (a) and the M2 protein (b) color coded according to the conservation scores obtained from ConSurf server [62-65].

Conservation scores for the M1 protein were obtained between -1.161 (maximum conservation) and 3.700.
(maximum variability) by the ConSurf server [62-65]. In general the M1 protein is conserved with 46% of the residues belong to grades 7-9 (conserved), while 25% of the residues belong to grades 1-3 (variable). Altogether, twenty five residue positions (10% of total residues) were found to be highly conserved (grade 9) and are shown in the table 2. Out of these highly conserved residues, only 36% of the residues were found to be exposed at the surface of the protein whereas the remaining 64% are buried residues. Residue Met1 and Gly122 showed no variation at all among 2836 sequences analyzed. In total, thirty eight residues (15% of total residues) were found to be highly variable (grade 1). Out of these highly variable residues, 53% of the residues were found to be exposed at the surface of the protein whereas the remaining 47% are buried residues. The residue positions 15, 95, 101, 167, 207 and 227 showed highest variations among all the sequences analyzed.

For the M2 protein, conservation scores were obtained between -1.119 (maximum conservation) and 2.817 (maximum variability) by the ConSurf server [62-65]. In general the M2 protein is conserved with 48.5% of the residues belong to grades 7-9 (conserved), while 26.9% of the residues belong to grades 1-3 (variable). Altogether, thirteen residue positions (10% of total residues) were found to be highly conserved (grade 9) and are shown in the table 2. Out of these highly conserved residues, only 39% of the residues were found to be exposed at the surface of the protein whereas the remaining 61% are buried residues. Residue Gly34 showed no variation at all among 3331 sequences analyzed with exception of three sequences which contained Glu34. In total, nineteen residues (7.5% of total residues) were found to be highly variable (grade 1). Out of these highly variable residues, 63% of the residues were found to be exposed at the surface of the protein whereas the remaining 37% are buried residues. The residue positions 10, 27, 28, 55 and 89 showed highest variations among all the sequences analyzed.

Discussion

The objective of the present study was to determine the degree of conservation of the matrix proteins among all the isolates of influenza A viruses. The matrix proteins from all hosts were analyzed together to facilitate the identification of universally conserved residues of potential pandemic viruses that might arise in future due to either a event of mutation or reassortment. The conserved residues detected on the matrix proteins may have either functional importance or structural importance [66]. On the contrary, variable sites arise as a result of either adaptation or evolutionary pressure to evade the host immune system.

In the established functionally important regions of the M1 protein high conservation was found, for example in the transcription inhibition domains (residues 90-108 and residues 129-164), Lys102, Ile131, Met135, Gly136, Ile154 and Arg163 are conserved at grade 9. A lone highly conserved residue Lys102 belonging to grade 9 was found in the nuclear localization signal (residues 101-105), nuclear export protein binding motif (residues 101-105), M1 protein and vRNP interaction region (residues 91-105) and RNA binding domain (residues 76-105). Residues Ser53 and Thr65-Thr67, were found to be highly conserved (grade 9) in the nuclear export signal (residues 59-68). Ala25 and Ile154 are only highly conserved (grade 9) residues found in death domain associated protein 6 binding motif (residues 23-48) and zinc binding domain (residues 148-162) respectively. Significant conservation was found in the heat shock protein 70 binding region (residues 128-165); residues Ile131, Met135, Gly136, Ile154 and Arg163...
were found to be highly conserved in this region. Residues Lys102, Gly122, Leu124, Ile131, Met135, Gly136, Ile154, Arg163, Pro171, Ile173, Glu176, Arg178, Met179 and Met189 were found to be highly conserved (grade 9), in the heat shock cognate protein 70 binding site (residues 102-201). In the M1 protein and cellular receptor of activated C kinase 1 interaction region (residues 1-67), residues Met1, Ala25, Ser53, Gly61 and Thr65-Thr67 are conserved at grade 9. Another region of conservation can be found in the M1 protein and cyclophilin A interaction region (residues 88-165), where residues Lys102, Gly122, Leu124, Ile131, Met135, Gly136, Ile154, and Arg163 were found to be highly conserved.

In the recognized functionally important regions of the M2 protein high conservation was found, for example in the region which is important for channel activity that is, $^{37}HXXXW^{41}$ motif residues His37 and Leu40 were found to be highly conserved (grade 9). The M1-M2 proteins interaction region (residues 71-73) was found to be highly conserved, with residues Ser71-Arg73 are conserved at grade 9. In the region which is associated with Inhibition of epithelial sodium channels (residues 53-62), residues Arg53 (grade 8) and Gly62 (grade 7) were found to be conserved. Significant conservation was found in the LC3 Interacting Region Motif (residues 91-94) with residues Phe91-Asn93 conserved at grade 7. In cholesterol recognition/interaction amino acid consensus motif (residues 48-62), only three residues were found to be conserved namely, Lys49 (grade 9), Arg53 (grade 8) and Gly62 (grade 7). The identified M2 protein and human Annexin A6 interaction region (residues 44-97) was found to contain conserved residues Asp44-Leu46, Lys49, Arg53, Gly62, Pro63, Ser64, Gly67, Pro69, Ser71-Tyr76, Gln80, Gln81, Ala83, Val84, Val86, Asp87 and His90-Asn93 which were either conserved at grade 8 or grade 9.

In culmination, this study has identified that matrix proteins showed a pattern of conserved and variable residues among all IAV hosts and subtypes. By identifying drug binding sites near the conserved residues in the matrix proteins, it will help in developing anti-influenza drugs which are unlikely to get ineffective in case of mutation of IAV into drug resistant form. Moreover the anti-influenza drugs targeting these binding sites are universally efficient against IAV of human, avian and swine origin. Further work arising from this study should characterize the function of the previously unknown highly conserved residues.

Acknowledgement

The author received no funding for this work.

References

2. WHO. Influenza (seasonal). Fact sheet No. 211, 2009


55. Le L, Jacek L. Study on phylogenetic relationships, variability, and correlated mutations in M2 proteins of influenza virus A. *PloS one*. 2011, 6:e22970


