Journal of Microbiology, Immunology and Infection (2012) **45**, 404–410

Available online at www.sciencedirect.com

SciVerse ScienceDirect

journal homepage: www.e-jmii.com

ORIGINAL ARTICLE

Characterization of subtypes of the influenza A hemagglutinin (HA) gene using profile hidden Markov models

Yu-Nong Gong [a,†], **Guang-Wu Chen** [b,c,†,*], **Shin-Ru Shih** [c,d]

[a] *Graduate Institute of Electrical Engineering, Chang Gung University, Taoyuan, Taiwan*
[b] *Department of Computer Science and Information Engineering, Chang Gung University, Taoyuan, Taiwan*
[c] *Research Center for Emerging Viral Infections, Chang Gung University, Taoyuan, Taiwan*
[d] *Department of Medical Biotechnology and Laboratory Science, Chang Gung University, Taoyuan, Taiwan*

Received 29 July 2011; received in revised form 27 September 2011; accepted 16 October 2011

KEYWORDS
HA subtype;
Influenza A virus;
Profile hidden Markov model;
Sequence analysis

Background: The influenza A virus has evolved into 16 hemagglutinin (HA) subtypes with different antigenic properties. Thus far typing has been primarily assay based, but the many sequences available from the US National Center for Biotechnology Information (NCBI) offer alternative ways of characterizing the HA gene.
Methods: All available HA sequences from the NCBI were analyzed. The software package HMMER was used to score how a training sequence fitted a profile hidden Markov model (profile HMM) constructed from the consensus sequence of one particular HA subtype, Hx, where $x = 1$ to 16. Scores from sequences of the same subtype and from other subtypes were then compared to see if they were separable. This approach was implemented in a stepwise manner, utilizing a sliding window of 100 amino acids with 10-amino-acid increments to build many subtype-specific models, and then assessing which 100-amino acid segments yielded the desired differentiability.
Results: Segment-based analysis revealed domains that correlate to HA sequence heterogeneity from one subtype to the others. For example, we showed that H1 segments covering only the second half of HA are not statistically separable from H2, H5 and H6 within the same region, suggesting evolutionary relatedness for these subtypes. The HA1 domain was found to be mostly differentiable between subtypes, which is in line with wet-lab findings that the domain is antigenicity-rich. We also reported a couple of regions that can be conveniently used to characterize all HA subtypes.

* Corresponding author. Department of Computer Science and Information Engineering, Chang Gung University, 259 Wen-Hua 1st Road, Kwei-Shan, Taoyuan 333, Taiwan.
E-mail address: gwchen@mail.cgu.edu.tw (G.-W. Chen).

[†] These authors contributed equally to this work.

doi:10.1016/j.jmii.2011.12.018

Conclusion: We established an analysis framework for assessing sequence-subtype association to provide insights into HA subtypes with close evolutionary relationships.

Introduction

The influenza A virus is a negative-stranded RNA molecule consisting of eight genomic segments. The virus is divided into subtypes on the basis of major differences in surface protein hemagglutinin (HA) and neuraminidase (NA), including 16 HA and nine NA antigenic subtypes.[1] Through continuous mutation, the HA gene has evolved into the observed multiple subtypes, which help the virus escape tracking by the human immune system, and have resulted in several pandemics.[2,3]

Conventional investigation of influenza A virus HA subtypes has been primarily based on biological assays such as the hemagglutination inhibition (HI) test.[1,4] These assay-based diagnoses are labor and time consuming, and often give rise to a negative result if antigenic changes have occurred as a result of alterations in the HA gene.[5] In recent years HA sequences have become increasingly available, so that pair-wise comparison[6,7] and a database search[8,9] are now alternative ways of characterizing the HA gene. While Rost[10] indicated that sequence alignment may produce unambiguous results when the pair-wise sequence identity of long alignments is greater than 40%, Park et al[11] also suggested that if sequence identities of related proteins are less than 30%, a database search will often fail to detect the relationship between the query and target strains. Nevertheless, those methods require that reference strains of known subtypes are chosen for comparison. The task of choosing appropriate representative strains itself often requires extensive biological knowledge, and the choices made can inevitably affect the outcome.

To make use of all available sequences, and thus avoid the need to choose subtype-specific HA representatives during sequence comparison, we multiply aligned all the HA sequences from a given HA subtype and built on their consensus a probabilistic model called a profile hidden Markov model (profile HMM), using the software package HMMER.[12] Previously we reported the use of profile HMMs to molecularly detect enterovirus type 71 (EV71) strains.[13] A profile HMM[12,14–16] is both statistically and probabilistically intrinsic, which makes it ideal for a quantitative assessment of how an individual sequence belongs to a profile. Profile HMMs have been extensively used in the Pfam database of protein families,[17] which allows protein families to be searched for, classified and characterized.

In the present study we expanded the use of profile HMMs in detecting EV71, and found it useful in revealing HA fragments that show specific association to a given subtype. All nonredundant influenza A HA protein sequences in GenBank, the US National Institutes of Health genetic sequence database, were downloaded for investigation. Instead of constructing a single model based on the entire HA gene, we used a sliding window of 100 amino acids, with 10-amino-acid increments, to build many segment-based, type-specific profiles, and then evaluated how these 100-amino-acid segments might serve as molecular targets in separating sequences of one specific HA subtype from other subtypes. Such segment-based analysis helps to establish a framework for assessing sequence-subtype association to identify major antigenicity domains, and provides insights into viral subtypes with close evolutionary relationships.

Methods

Sequence retrieval and pre-processing

A total of 22,503 HA protein sequences were collected from the Influenza Virus Resource[18] of the US National Center for Biotechnology Information (NCBI) for all viruses isolated up to February 2009 (inclusive) for model development and training. They were manually curated by removing short sequences (less than 100 amino acids; 370 sequences), redundant sequences (approximately 30% of the total sequence count), and sequence records that contained an unknown amino acid, represented in the single-letter alphabet as 'X' (742 sequences). The resulting 16,198 sequences were further grouped into 8958 full-length sequences (at least 500 amino acids long) and 7240 partial sequences (Table 1). Other than the described training sequences, we gathered an additional 3520 HA sequences from March 2009 to January 2011 to validate our method.

Building profile HMMs and computing for HMMER raw scores

Profile HMMs for each of the HA subtypes were constructed using the software package HMMER version 3.0.[12] Subtype-specific full-length training sequences for each subtype from Table 1 were firstly aligned using the computer program MUSCLE (multiple sequence comparison by log-expectation), version 3.8.31,[19] and the resulting multiple sequence alignments (MSAs) were each read by the program *hmmbuild* in HMMER, which in turn generated one model for each subtype. Together we built 14 subtype-specific models, except for subtypes H14 and H15, which contained, respectively, only three and six full-length sequences. They were excluded from the HMM model building proposed in this study because of limited sample sizes for yielding statistically significant results.

The HMMER program *hmmsearch* was used to compute for an HMMER raw score for any given training sequence per established HMM profile of subtype Hx ($x = 1$ to 16 excluding 14 and 15). An HMMER raw score represents the log-odds the training sequence belongs to the given model created by *hmmbuild*. To maximize the effectiveness of our

Table 1 Statistics for influenza hemagglutinin (HA) sequences used (for the purpose of model building, training, and testing)

Subtype	Training sequence			Testing sequence
	Total	Full-length[a]	Partial-length	Total
H1	3465	1521	1944	275 (seasonal) 2712 (H1N1pdm)
H2	253	176	77	0
H3	6229	2108	4121	265
H4	403	352	51	11
H5	2842	2340	502	171
H6	598	511	87	5
H7	726	614	112	3
H8	64	55	9	0
H9	1150	851	299	63
H10	171	153	18	6
H11	144	135	9	2
H12	69	63	6	2
H13	58	54	4	4
H14[b]	3	3	0	0
H15[b]	6	6	0	0
H16	17	16	1	1
Total	16,198	8958	7240	3520

[a] Full-length sequences are 500 amino acids or longer.

[b] Used only in computing non-Hx scores for other subtypes. Seasonal = seasonal influenza virus; H1N1pdm = pandemic H1N1 influenza virus.

molecular diagnosis, we also utilized full-length sequences to train our models in addition to the listed partial ones. Scores produced from full-length sequences, however, would be unrealistically higher as these sequences were already used in the construction of their profile HMM earlier. To correct such over-fitted scores, we employed a cross-validation method described in our previous report.[13] As the size of HA varies among subtypes, we have provided 16 reference strains[1] for describing the coordinates of HA alignments in Additional file 1 (http://vbrc.cgu.edu.tw/phmms/Additional_file_1.fas). To reveal sequence similarity among those subtypes, we additionally used the program *hmmemit* to generate one subtype-specific consensus sequence from each of the 16 HA MSAs (the same ones used to generate profile HMMs), and showed a grand MSA that is presented in Additional file 2 (http://vbrc.cgu.edu.tw/phmms/Additional_file_2.fas).

HMMER raw score normalization and conversion to Z-score

HMMER raw scores produced above were found in general to be proportional to the length of the locally aligned segment between a query sequence and a profile. As a result we normalized each of these scores by this length, and further transformed the normalized raw scores into a Z-score representation, a quantity mathematically represented by

$$Z = \frac{X - \mu}{\sigma} \tag{3}$$

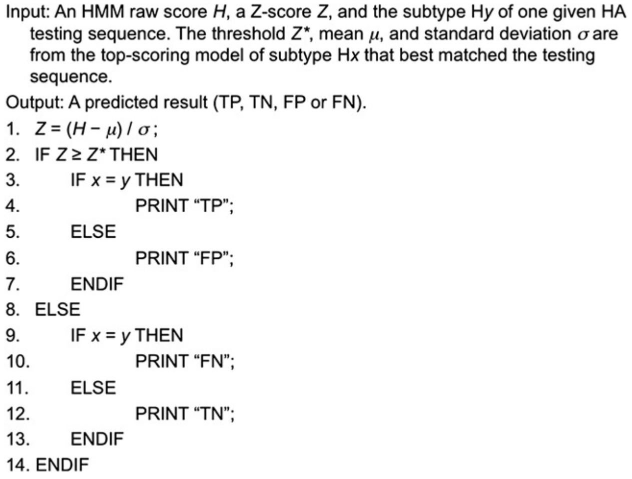
where X is a normalized HMMER raw score to be converted, σ the standard deviation, and μ the mean of the normalized HMMER raw scores.

Segment-based analysis

To better reveal the differentiability of training samples on different parts of the HA gene, we implemented a segment-based analysis, in which we defined multiple 100-amino-acid segments incremented by a 10-amino acid sliding window from the N- to the C-terminus of the entire HA gene. We used these 100-amino-acid segments to build 658 subtype- and position-specific models (47 for each of the 14 subtypes). In quantitatively assessing the separation of the two types of scores, we defined a separation distance S as the subtraction of the largest non-H*x* Z-score from the smallest H*x* Z-score, where *x* represents any one of the 14 HA subtypes, and graphed a subtype-specific S distribution for these 100-amino-acid segments. A positive S value suggests that a cut-off Z-score threshold exists in separating non-H*x* sequences from H*x* ones, based on the profile HMM of H*x*. This allows us to assess which 100-amino-acid segment(s) may best differentiate between the scores of a given subtype and the remaining subtypes.

Validation of the proposed molecular diagnosis

Only segment-based models having positive S values are incorporated in our molecular diagnosis. For each of these models, the arithmetic mean of the minimal H*x* Z-score and the maximal non-H*x* Z score was precomputed as a cut-off value Z^* for a binary classification test[20] described in Fig. 1. Every testing sample of subtype H*y* was firstly aligned to each of these recruited segment-based models. The top-scoring model H*x* emerged and its threshold Z* retrieved. A result was deemed to be positive if the computed Z-score for the testing sample was $\geq$ Z* and the testing sequence displayed the same subtype as the



```
Input: An HMM raw score H, a Z-score Z, and the subtype Hy of one given HA
  testing sequence. The threshold Z*, mean μ, and standard deviation σ are
  from the top-scoring model of subtype Hx that best matched the testing
  sequence.
Output: A predicted result (TP, TN, FP or FN).
$Z = (H - \mu) / \sigma$;
IF $Z \geq Z^*$ THEN
    IF $x = y$ THEN
        PRINT "TP";
    ELSE
        PRINT "FP";
    ENDIF
ELSE
    IF $x = y$ THEN
        PRINT "FN";
    ELSE
        PRINT "TN";
    ENDIF
ENDIF
```

Figure 1. A binary test used to validate the proposed segment-based hidden Markov model (HMM) analysis.

matched model, and further became a true positive (TP) if $x = y$, or a false positive (FP) if $x \neq y$. Otherwise the result was classed as negative for failing to determine the subtype, either a false negative (FN) if $x = y$, or a true negative (TN) if $x \neq y$.

Results

The HMMER raw score distribution of all training sequences evaluated with H1 profile (based on full-length H1 sequences) is shown in Fig. 2 as an example, in which the separation of H1 versus non-H1 scores is readily assessed. There exists a clear separation of H1 and non-H1 sequences for a length of approximately 400 amino acids or longer. An inset of Fig. 2 stretching for 200—300 amino acids and scores 400—600 shows the mixing of H1, H2, H5, and H6 sequences. Graphs similar to Fig. 2 for other subtypes were also recorded (data not shown). In particular, we found the mixing of HMMER raw scores in various groupings, including {H3, H4, H14}, {H7, H10, H15}, {H8, H9, H12}, and {H11, H13, H16}, in addition to the mixture of H1, H2, H5 and H6 sequences shown in Fig. 2. These groupings of subtypes were also found to conform to the reported phylogenetic tree topology.[1,21,22]

To better define a quantitative measurement of the intended separation for scores, we further transformed the HMMER raw scores into Z-scores. Taking the same data shown in Fig. 2, for example, all HMMER raw scores were firstly scaled by the length of alignment (the horizontal axis) that produced the raw scores in *hmmsearch*, and then converted into Z-scores. Fig 3A shows the distribution of Z-scores, in which the separation between H1 (circles) and non-H1 (triangles) sequences can be easily assessed. The crossover of H1 and non-H1 scores in Fig. 3A suggests that either some H1 sequences (with low Z-scores) did not score high enough to fit this H1 profile, or that some of the non-H1 sequences were probabilistically similar to the H1 subtype in yielding relatively high Z-scores.

Upon observing that the scores from different subtypes might cross over and thus result in a negative S value as exemplified in Fig. 3A, we performed an alternative analysis based on many smaller and overlapped segments, and rebuilt a subtype- and position-specific profile HMM on each of these segments. Full-length HA alignments were divided into multiple 100-amino-acid segments, each representing

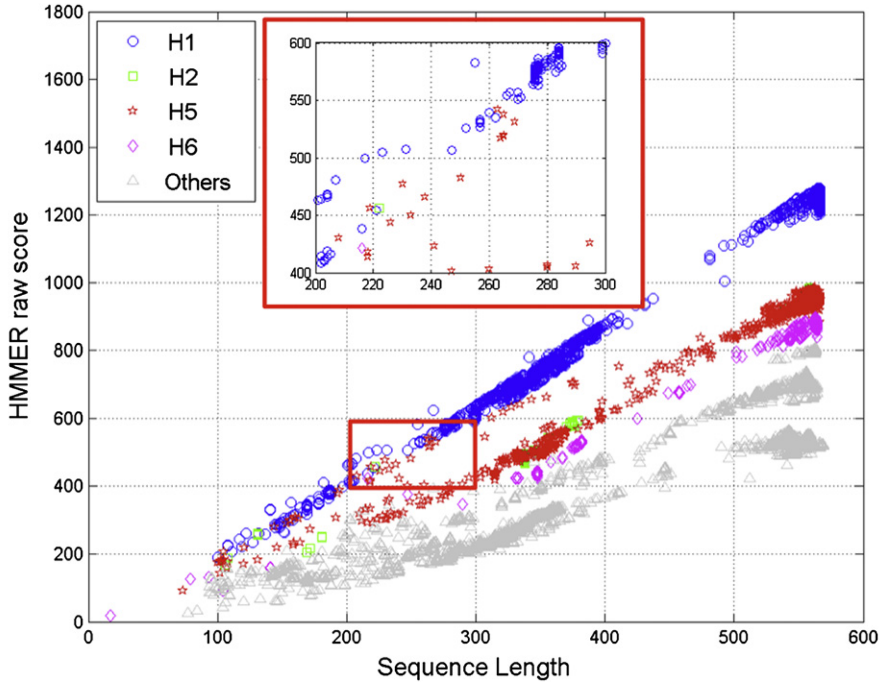


Figure 2. Length-specific HMMER raw scores generated by training samples with respect to a profile HMM generated from the multiple sequence alignment of H1 full-length sequences. The horizontal axis displays the matched hemagglutinin (HA) sequence length in amino acids revealed by *hmmsearch*. The vertical axis represents the HMMER raw scores.

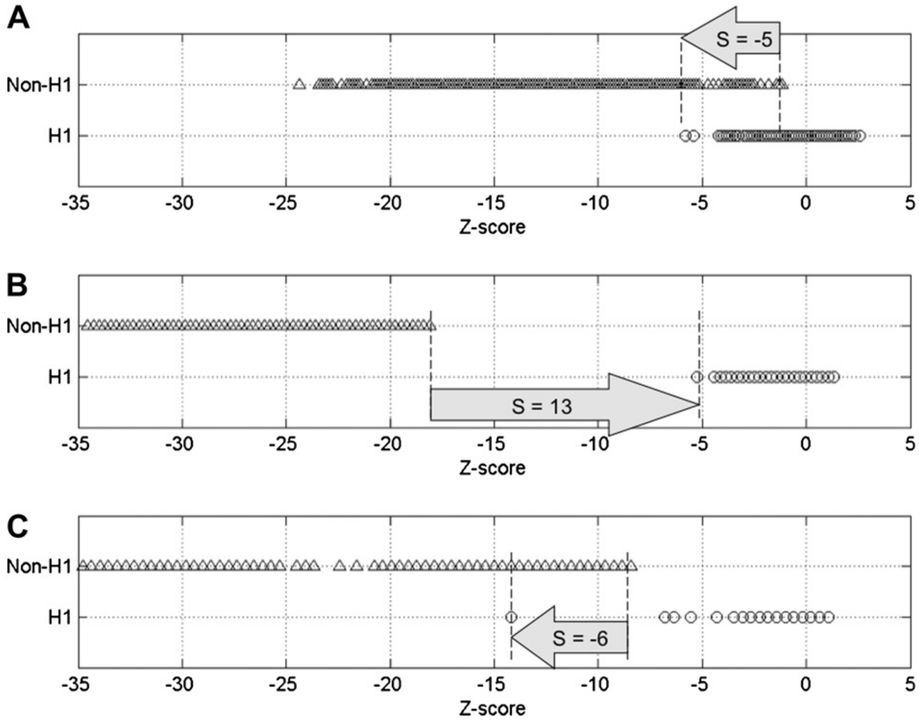


Figure 3. (A) Analysis of the entire HA gene gave a Z-score distribution for H1 and non-H1 strains that showed a negative S value of −5. Segment-based analysis of H1 versus non-H1 strains gave (B) a Z-score distribution for the 18th segment of H1 (or positions 171—270) showing a positive S value of 13, and (C) the 35th segment of H1 (or positions 341—441) showed a negative S value of −6.

a segment of alignment containing positions 1—100, 11—110, 21—120, and so on, in which the coordinate labels are subtype-specific, according to the prototype strains shown in Additional file 1. As illustrated in Fig. 3B, the 18th segment is apparently capable of yielding separable Z-scores for H1 and non-H1 sequences with an S value of 13. Fig. 3C, on the contrary, shows that the 35th segment gives a minimal S value of −6. This 100-amino-acid segment is not able to differentiate between H1 and non-H1 sequences, because of one single low-scoring H1 sequence with a Z-score of approximately −14.

Following the construction and training of 658 subtype-specific, segment-based models, we used an additional 3520 sequences to validate our method. Five test cases (Table 2) were performed according to all segments of positive S values. Over 99% of the testing sequences were TP for various cases. The count for TN ranges from 2 to 22, as fewer models were available for the purpose of classification as a result of applying a more stringent (bigger) S value. The number of negative cases increased significantly from a count of four in Case 3 to 22 in Cases 4 and 5, which allowed us to conveniently choose S values no less than two as a criterion to recruit useful 100-amino-acid segment-based models with which to perform the proposed molecular investigation.

Table 2 Binary classification tests for recruited segments with various S cut-off values

Case	Recruited segments		Testing sequences	
	S value	Count	TP/TN	FP/FN
1	S > 0	585	3518/2	0/0
2	S > 1	564	3517/3	0/0
3	S > 2	545	3516/4	0/0
4	S > 3	529	3498/22	0/0
5	S > 4	513	3498/22	0/0

FN = false negative; FP = false positive; TN = true negative; TP = true positive.

The aforementioned segment-based analysis reveals what segments are useful in differentiating one influenza A HA subtype from others. Graphs of segments with variable S values are presented in Fig. 4, where the 14 investigated subtypes are separately shown in different sub-figures based on their groupings of HMMER raw scores. Segments showing more positive S values are considered better molecular targets associating to one particular subtype. Taking +2 as a significant S value for separating one specific subtype from the others (dotted line in Fig. 4), we see that most of those 100-amino-acid segments have well exceeded this threshold for H8 to H13 and H16 in Figs. 4C to 4E, except for H13 segments from 251_350 to 341_440 in Fig. 4E. Further checking revealed that all 17 H16 sequences included in the training set hit unusually high HMMER raw scores to H13 model and caused this failure in differentiation. In other words, the H13 and H16 sequences are more similar in these mentioned segments than in any other part of HA.

Discussion

For a given subtype-specific profile, training data were used to reveal how homotypic (of the same H*x* type) and

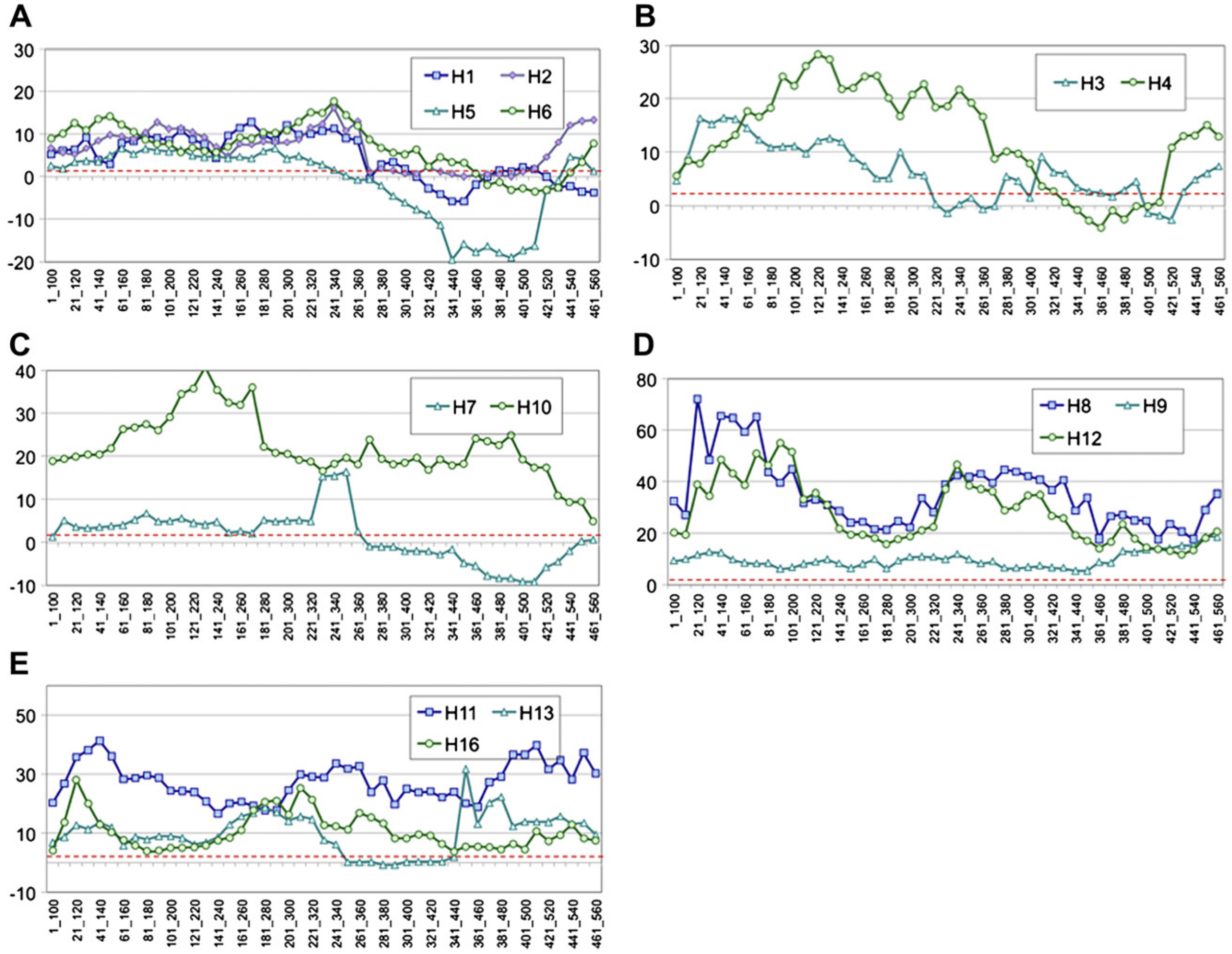


Figure 4. Graphs of S values for the gene segments (A) H1, H2, H5, and H6; (B) H3 and H4; (C) H7 and H10; (D) H8, H9, and H12; and (E) H11, H13, and H16. The grouping was based on the evolutionary relatedness revealed in this study. Graphs for H14 and H15 were excluded for their lack of training sequences from the National Center for Biotechnology Information (NCBI). The dashed line shows a threshold of S = 2.

heterotypic (of non-H*x* type) sequences are separated on the basis of their HMMER raw scores with respect to this model. It was found, however, that the homo- and hetero-typic scores did not completely separate the sequences, as exemplified in Fig. 2. Among the possible reasons for this incomplete separation is that some of the short sequences matched to a model position containing insufficient "subtype-specific information". This would have lead to either a homotypic sequence not scoring high enough with respect to the model, or a heterotypic sequence that did not hit a score low enough to separate it from the homotypic ones.

To justify this hypothesis, we performed the profile HMM analysis in a segment-based manner. The reason for choosing a 100-amino-acid sliding window is that many DNA sequencers can easily produce a nucleotide sequence that is at least 300 base pairs long. This makes an HA sequence to be molecularly characterized likely to comprise 100 or more amino acids. Using segment-based analysis, we see in Fig. 4A that H1 segments containing those non-H1 sequences with elevated HMMER raw scores in Fig. 2 mostly exhibited S values of two or less, with many of them further plunged into negative territory in the second half of HA. This example clearly demonstrates, based on our findings, that certain H1 domains are evolutionarily closer to some subtypes than others. The locations of these nondifferentiable segments are also subtype-dependent, and can be easily visualized in Fig. 4.

It is noted that some segments with negative S values in Fig. 4 are largely the result of a number of non-H*x* sequences that produced high HMMER raw scores to the H*x* model of interest. For example, we observed all H16 sequences included in the training set hit relatively high HMMER raw scores to the H13 model, and resulted in 10 consecutive H13 segments with $S < 2$ (Fig. 4E). If we removed all H16 sequences from our training set, the S values of all these H13 segments completely recovered into positive territory. We can further illustrate this by using the grand MSA of the consensus sequences for all 16 subtypes, shown in Additional file 2, in which the genomic span of these H13 segments with $S < 2$ (a 190-amino-acid stretch from positions 251 to 440) shows 90.5% identity to H16, compared to only 78.1% for the remaining HA. Similar observations were also made on some segments from other subtypes with low S values. For example, a stretch of H5 segments with S < 2 (320 amino acids long, from positions 241 to 560), shown in Fig. 4A, is 71.2% identical to H1, 81.2% to H2, and 70.6% to H6, whereas for the remaining HA stretches the percentages are lower at 53.7%, 67.9%, and 49%, respectively. Another example is a 290-amino-acid H7 stretch (20 segments with S < 2, from positions 271 to 560 in Fig. 4C), with 76.1% identity to H10 and 84.1% to H15, whereas for the remaining HA the degrees of identity are only 55.7% and 76.2%, respectively.

We noted that segments with negative S values (Figs. 4A to 4C) are mostly located in the mid- to after-part of HA gene. Those 100-amino-acid segments within approximately the first two-thirds of this surface glycoprotein, also known as the HA1 domain, are mostly differentiable between subtypes. This phenomenon is in line with other reported findings,[23,24] that the HA1 domain has much richer antigenicity than the remaining HA2 domain. Our method further pinpointed segments with large S values that may contain important molecular information for the evolution of influenza A virus into various subtypes. Segments with large S values that are common among subtypes, such as segments 21_120 to 211_310, represent genomic locations that can be used in distinctly characterizing all 16 HA subtypes, and can be useful for designing diagnostic probes or polymerase chain reaction (PCR) primers for molecular detection.

Depending on what population is to be detected or characterized, the "subtype" column of Table 1 can be easily re-assigned to a phenotype of interest, for example, "Human H1" or even "Recent Avian H1". The analysis framework proposed here is equally applicable to other influenza gene segments or genes from other organisms, provided that a homo- and hetero-typic grouping can be made and respective training samples properly assembled.

Acknowledgments

We thank Chang Gung Memorial Hospital (grants CMRPD250033, CMRPD260011, CMRPD260012) and the National Science Council of Taiwan (grant NSC96-2221-E-182-059) for financially supporting this research.

References

1. Fouchier RAM, Munster V, Wallensten A, Bestebroer TM, Herfst S, Smith D, et al. Characterization of a novel influenza A virus hemagglutinin subtype (H16) obtained from black-headed gulls. *J Virol* 2005;**79**:2814—22.
2. Palese P. Influenza: old and new threats. *Nat Med* 2004;**10**(12 Suppl):S82—7.
3. Pandemic (H1N1) 2009 Update 112. World Health Organization (WHO) http://www.who.int/csr/don/2010_08_06/en/index.html [accessed 13.07.11].
4. Röhm C, Zhou N, Süss J, Mackenzie J, Webster RG. Characterization of a novel influenza hemagglutinin, H15: criteria for determination of influenza A subtypes. *Virology* 1996;**217**:508—16.
5. Mackay WG, van Loon AM, Niedrig M, Meijer A, Lina B, Niesters HG. Molecular detection and typing of influenza viruses: are we ready for an influenza pandemic? *J Clin Virol* 2008;**42**:194—7.
6. Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci U S A* 1990;**87**:2264—8.
7. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 1970;**48**:44353.
8. Pearson WR. Comparison of methods for searching protein sequence databases. *Protein Sci* 1995;**4**:1145—60.
9. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol* 1981;**147**:195—7.
10. Rost B. Twilight zone of protein sequence alignments. *Protein Eng* 1999;**12**:85—94.
11. Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, et al. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J Mol Biol* 1998;**284**:1201—10.
12. Eddy SR. Profile hidden Markov models. *Bioinformatics* 1998;**14**:755—63.
13. Chen G-W, Hsiung CA, Chyn J-L, Shih SR, Wen CC, Chang IS. Revealing molecular targets for enterovirus type 71 detection by profile hidden Markov models. *Virus Genes* 2005;**31**:337—47.
14. Eddy SR. What is a hidden Markov model? *Nat Biotechnol* 2004;**22**:1315—6.
15. Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A* 1987;**84**:4355—8.
16. Gribskov M, Veretnik S. Identification of sequence pattern with profile analysis. *Meth Enzymol* 1996;**266**:198—212.
17. Sonnhammer EL, Eddy SR, Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* 1997;**28**:405—20.
18. Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, et al. The influenza virus resource at the national center for biotechnology information. *J Virol* 2008;**82**:596—601.
19. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004;**32**:1792—7.
20. Fawcett T. *ROC graphs: notes and practical considerations for researchers*. Amsterdam: Kluwer; 2004.
21. Air GM. Sequence relationships among the hemagglutinin genes of 12 subtypes of influenza A virus. *Proc Natl Acad Sci U S A* 1981;**78**:7639—43.
22. Dugan VG, Chen R, Spiro DJ, Sengamalay N, Zaborsky J, Ghedin E, et al. The evolutionary genetics and emergence of avian influenza viruses in wild birds. *PLoS Pathog* 2008;**4**:e1000076.
23. Stevens J, Corper AL, Basler CF, Taubenberger JK, Palese P, Wilson IA. Structure of the uncleaved human H1 hemagglutinin from the extinct 1918 influenza virus. *Science* 2004;**303**:1866—70.
24. Shih AC, Hsiao TC, Ho MS, Li WH. Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution. *Proc Natl Acad Sci U S A* 2007;**104**:6283—8.