Abstract

Previously we have used this approach in influenza HA subtyping, and found it useful in revealing HA fragments that are abundant with "antigenic character" to differentiate one subtype from the others. In this paper our goal is to identify NA fragment(s) that are more differentiable for one NA subtype to the others. Profile Hidden Markov Models (PHMMs) were used in quantitatively evaluating how a sequence belongs to the consensus of a collection of the same subtype. A total of 6,868 NA coding sequences from GenBank were analyzed. Although strains from all subtypes were used, we are particularly interested in HA subtypes that once infected humans, including N1 and N2. Instead of constructing a single profile model for the entire NA, we used a sliding window of 100-aa with 10-aa increment in building many profiles with the index PHMM or PHMM2, followed by evaluating what 100-aa segment(s) yield the best overall differentiation power. Eventually, this analysis may provide locations for molecular targets that are characteristic to a specific subtype.

Introduction

Classification of influenza A viruses are based on their antigenic differences of surface proteins - hemagglutinin (HA) and neuraminidase (NA). Recently there are reports that such classification can be facilitated through biological sequence comparison. Those dry-land methods utilize pairwise sequence comparison or phylogenetic inference to determine if a test strain is having its HA or NA fragment resemble most to a pre-selected group of reference strains of known HA or NA subtypes. Those approaches, however, cannot turn into different result if reference strains are not chosen appropriately. Previously we have reported using profile hidden Markov models in molecularly detecting enterovirus type 71 (EV71) strains and influenza A virus subtypes. Protein sequences were retrieved from GenBank/NCBI (Influenza Virus Resource) as of March 2007. All redundant sequences of the same family, which we call a profile. Then we are able to easily recognize if, instead of which a target sequence can be classified into a specific profile among a collection of profile library. We have explored this approach in influenza NA subtyping, and found it useful in revealing the HA fragments that are abundant with features or "antigenic character" that allows us to differentiate one subtype from the others. These observations are in agreement with published results that the 1st 2/3 of the HA gene displays more "antigenic" profiles that can be used to differentiate one subtype from the others. This observation is consistent with lab conclusion that the HA domain is more antigenic-rich than NA in H3 strains. In this work we further expanded this approach in revealing "signature-rich" NA segments. A total of 5,488 full (or nearly full) NA acid sequences were downloaded from GenBank and investigated. Those locations, or so-called signature-rich segments, are believed to contain site information with binding activities that are selectively expressed for influenza A replication.

Materials and Methods

Profile Hidden Markov Models

Profile Hidden Markov Model (PHMM) is both statistically and probabilistically intrinsic, which makes it a quantitative comparison tool that has been extensively used in PAM database [3], which offers the search and classification for protein families. We presented profile HMM method for NA subtyping using the package PHMMER version 2.2 from Sean Eddy’s lab, which was originally affiliated with Washington University at St. Louis, and now is located at the Lava Lab, at the Salk Institute’s Research Campus of the Howard Hughes Medical Institute. Before building profile PHMMs, gene templates of the same subtype should be aligned. We used Clustal V version 1.8.1 [4] to align all the proteins in the profile library to the top of the multiple alignment using the program ArachneTools. Training sequences consisting of 100 N-terminal segments of the same subtype and those from different subtypes were evaluated against the profile using the program ArachneTools. The result was a training score of the profile PHMM, which is a bit score representing the log odds ratio of how one sequence belongs to the profile, which is Score = Score(q) − Score(f) (HMM) (PHMM)

Protein (HMM) and Protein (Null)

Profile (HMM) and Profile (Null) are the probabilities of target sequence according to a profile HMM and a null model, respectively. A null model in null hypothesis model of the statistics of random sequence. Thus, a positive score means the profile HMM is a better one than the null model.

Data Pre-processing

All of the influenza aminio acid sequences in GenBank were downloaded for investigation as of March 2007. There are total of 6,868 NA coding sequences in Table 1. For each subtype, we are interested in NA subtypes that once infected humans. Thus we worked on building the profiles for N1, N2, N3, N5, N7, and N8. Following the same rationale as in the PHMM profiles, we constructed a PHMM for each subtype. A total of 5,488 NA coding sequences from GenBank were analyzed. Although strains from all subtypes were used, we are particularly interested in HA subtypes that once infected humans, including N1 and N2. Instead of constructing a single profile model for the entire NA, we used a sliding window of 100-aa with 10-aa increment in building many profiles with the index PHMM or PHMM2, followed by evaluating what 100-aa segment(s) yield the best overall differentiation power. Eventually, this analysis may provide locations for molecular targets that are characteristic to a specific subtype.

Table 1 - Statistics of Influenza A Virus segments

<table>
<thead>
<tr>
<th>Subtype</th>
<th>Total Length</th>
<th>Non-Redundant</th>
<th>Full Length</th>
<th>Partial Length</th>
<th>Testing Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>N1</td>
<td>10 744</td>
<td>1,049</td>
<td>965</td>
<td>74</td>
<td>100</td>
</tr>
<tr>
<td>N2</td>
<td>10 096</td>
<td>1,466</td>
<td>1,027</td>
<td>17</td>
<td>100</td>
</tr>
<tr>
<td>N3</td>
<td>10 012</td>
<td>1,377</td>
<td>1,027</td>
<td>12</td>
<td>100</td>
</tr>
<tr>
<td>N4</td>
<td>50 124</td>
<td>1 720</td>
<td>1 470</td>
<td>37</td>
<td>100</td>
</tr>
<tr>
<td>N5</td>
<td>10 011</td>
<td>1 299</td>
<td>1 027</td>
<td>12</td>
<td>100</td>
</tr>
<tr>
<td>N7</td>
<td>10 027</td>
<td>3 786</td>
<td>3 470</td>
<td>37</td>
<td>100</td>
</tr>
<tr>
<td>N8</td>
<td>10 027</td>
<td>3 786</td>
<td>3 470</td>
<td>37</td>
<td>100</td>
</tr>
</tbody>
</table>

Results

Relationship Between Subtypes of Influenza Viruses

Length-specific HMMER bit score distribution for the entire-length NA sequences of various subtypes with respect to N1 and N2 profile PHMM in N1 and N2 subtypes. In general it was found that the raw scores are proportionally to the sequence length. There are clear separations between N1 and N2 to NA5 and N7 to N1 and N2. In full-length profiles, the N1 and N2 subtypes were utilized. The graph shows that the separation between the two types (target subtype and red circles (Non-target subtype) can be easily determined. The null model shows at least a 1.5-fold higher bit score than the N1 and N2 subtypes, which is a clear difference.

Conclusion and Discussion

1. Segment-based analysis allows us to screen out the signature-rich segments for classification.
2. Biologists can make diagnoses of influenza subtype by sequencing the NA fragments of the signature-rich segment that are most suitable for making differentiation among subtypes.

References

[1] Revealing Molecular Targets for Enterovirus Type 71 Detection by Profile Hidden Markov Models.
[5] A general method applicable to the search for similarities in the amino acid sequence of two protein.