Tel-Aviv University
George S. Wise Faculty of Life Sciences
Graduate school

In Silico Identification of Amino Acids Comprising the Avian-to-Human Barrier in Influenza A H5N1

Thesis submitted towards the M.Sc degree in
Biochemistry at Tel-Aviv University

by: Daphna Meroz

The research was performed in
the biochemistry department
under the supervision of:
Prof. Nir Ben-Tal

Tel-Aviv University
George S. Wise Faculty of Life Sciences
Graduate school

In Silico Identification of Amino Acids Comprising the Avian-to-Human Barrier in Influenza A H5N1

Thesis submitted towards the M.Sc degree in
Biochemistry at Tel-Aviv University

by: Daphna Meroz

The research was performed in
the biochemistry department
under the supervision of:
Prof. Nir Ben-Tal

Supervisor's signature:

Date:

Table of Contents

Acknowledgments

I am greatly indebted to my advisor, Prof. Nir Ben-Tal, who always believed in me, guided and helped me.

I am most grateful to Maya Shcushan for the huge help, support and encouragement in every step of the way. I would also like to thank Yana Gofman for being such a good friend, a good listener and always being there for me. I wouldn't have done it without them,

I would also like to thank the rest of my lab – Gilad, Guy, Ofir, Matan, Inbar and Noam, - who made me smile every day, infallibly.

Last, but certainly not least, I thank my parents, sisters and grandmother for their constant love and support. And of course, Iddo who always put my research first, and encouraged me continuously.

Abstract

Highly pathogenic in humans, although yet to become widespread in the population, the H5N1 strain constitutes a major threat owing to significant similarity between avian and human infecting viruses. The hemagglutinin (HA) protein of influenza A is the main antigen on the viral surface, mediating binding to the host receptors and virus entry into the cell. An alteration from avian-to-human like recognition via HA is thought to be one of the changes that must take place before avian influenza viruses can replicate efficiently in human cells, thus acquiring the capability to cause a human pandemic. Through a computational approach, using a supervised learning algorithm and the complete H5N1 NCBI sequence database, I successfully identified essentially all known specificity determinants for avian to human transmissibility described in the literature. Interestingly, I also detected residues that form the known H5 antigenic sites as host-distinguishing positions, offering a possible immune-related mechanism for virus specificity. My analysis also identified novel specificity determinant candidates that may help decipher the basis for human vs. avian host selectivity. These new findings may provide a better understanding of the species barrier of H5N1 and assist in designing antiviral agents. The computational analysis presented here is generic and can also be applied to gain insight into the molecular basis of host discrimination in other virus strains.

1. Introduction

Influenza, often called the 'flu', is one of the main diseases of the respiratory tract in humans and is the cause of hundreds of thousands of deaths annually. The influenza virus has three types; A, B and C. Influenza A is the most virulent of the three and is responsible for seasonal epidemics and at times major human pandemics throughout the world [1]. Influenza A is part of the orthomyxovirus family and is thought to originate mainly from wild birds. The virus may be transmitted to domestic poultry that then could possibly give rise to a human virus pandemic. The influenza A virus is comprised of eight RNA segments that encode for 11 proteins. The virion envelope contains two main surface glycoproteins: hemagglutinin (HA) and neuraminidase (NA). The different combinations of HA and NA subtypes are used to classify the virus into strains. There are nine known NA serotypes (NA1-9) and 16 known HA serotypes (HA1-16) [2], of which only three have become adapted to humans (H1, H2, H3).

The current outbreak of the swine-origin influenza virus H1N1 pandemic [3,4] is the last of several major influenza pandemics that occurred throughout history by different influenza A viruses, including the highly-pathogenic 1918 pandemic, estimated to have killed over 50 million people worldwide [5], the 1957 (H2N2) and 1968 (H3N2) pandemics [6]. While attempts are underway to estimate the transmissibility and pathogenicity of this novel strain [7,8,9], it has become clear that influenza pandemics will continue to occur with the introduction of novel influenza strains into the human population from other species. One such potential and disturbing threat is the H5N1 avian influenza.

H5N1 has become an endemic strain in wild waterfowl and domestic poultry in

many parts of Southeast Asia, and has been spreading across Asia into Europe and Africa [10]. H5N1 can be highly pathogenic in birds and in the highly pathogenic state, its spread through poultry is very fast and causes disease in multiple internal organs with a mortality rate up to 100% [11]. Although tens of millions of poultry have been infected throughout the last few years, only 442 human cases were officially reported to the World Health Organization (WHO) [12], with 262 death cases as of September 24, 2009. Additionally the current H5N1 viruses are not human-to-human transmissible. However, due to its observed high virulence in the avian host, the emergence of a human-adapted H5N1 virus, either by reassortment or mutation, is a threat to public health worldwide. It is therefore crucial to monitor viral mutations of the H5N1 that may enable the efficient transmission of the virus into the human population.

The principal protein on the viral surface is hemagglutinin (HA). HA is a homo-trimer in vivo, responsible for viral binding to host receptors. The receptor-binding domain is located at the membrane-distal globular section of the trimer [13] (Figure 1). Its binding to the cell receptor enables entry of the virus into the host cell via endocytosis and fusion of the membranes of the virus and the endosome. The cellular receptors of HA are terminal sialic acids of glycoproteins and glycolipids, which can be linked in an α2,3 or α2,6 bond to galactose (SAα2,3Gal and SAα2,6Gal, respectively). The type of linkage has been shown to influence the ability of a given virus to infect different species, due to the specific binding receptors that are more common in cells of each specie: humans are more readily infected by viruses that bind to the α2,6 linkage, birds and horses are more susceptible to viruses that bind to α2,3 linkages, and pigs can be infected by viruses with either of these binding preferences [14,15]. Therefore, considerable research has focused on identifying specificity

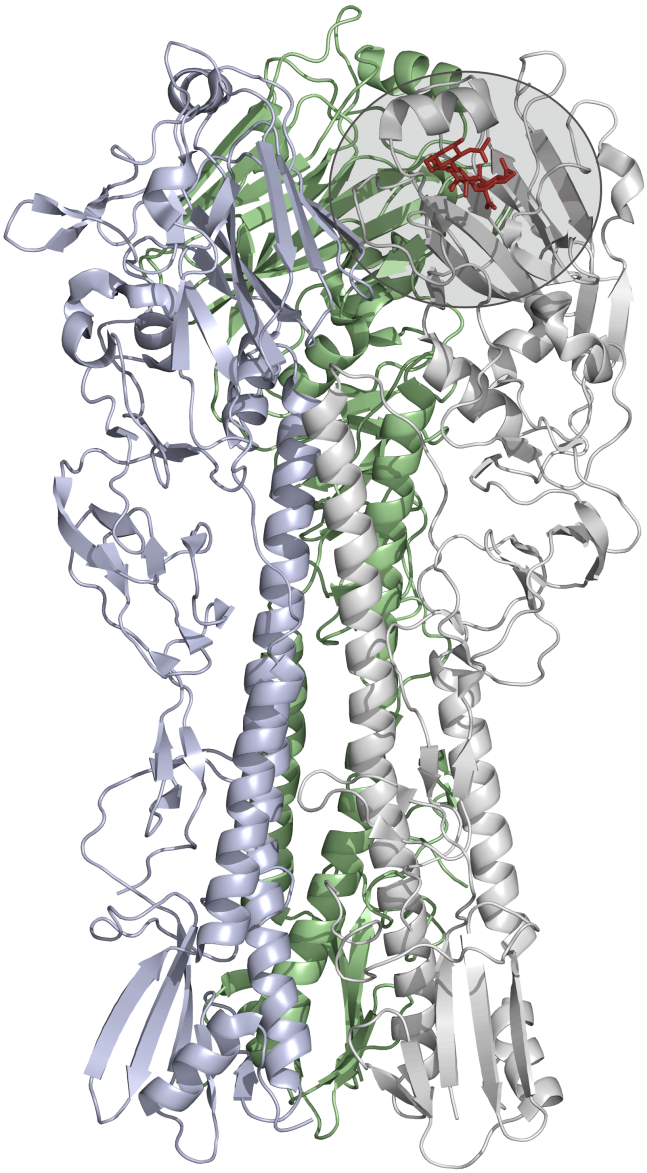

Figure 1. Structure of the hemagglutinin trimer from A/Duck/Singapore/3/97 (pdb 1jsn). The monomers are in a cartoon representation. Each monomer has been colored differently. The receptor binding site is encircled only on one monomer. The human receptor analogue (LSTc) is shown in sticks representation and coloured in red. The LSTc was modeled into the receptor-binding domain by superimposing the structures of H5 A/Duck/Singapore/3/97 and H9 A/Swine/Hong Kong/9/98 HAs (PDB 1JSN and 1JSI) with bound α2-3 and α2-6 analogs.

determinants: positions on the HA binding domain that influence its binding affinity to different receptors of various HA subtypes. To date, several positions have been identified as specificity determinants, affecting host preference of the H5N1 [16,17,18,19,20,21] (Table 1, Figure 2A).

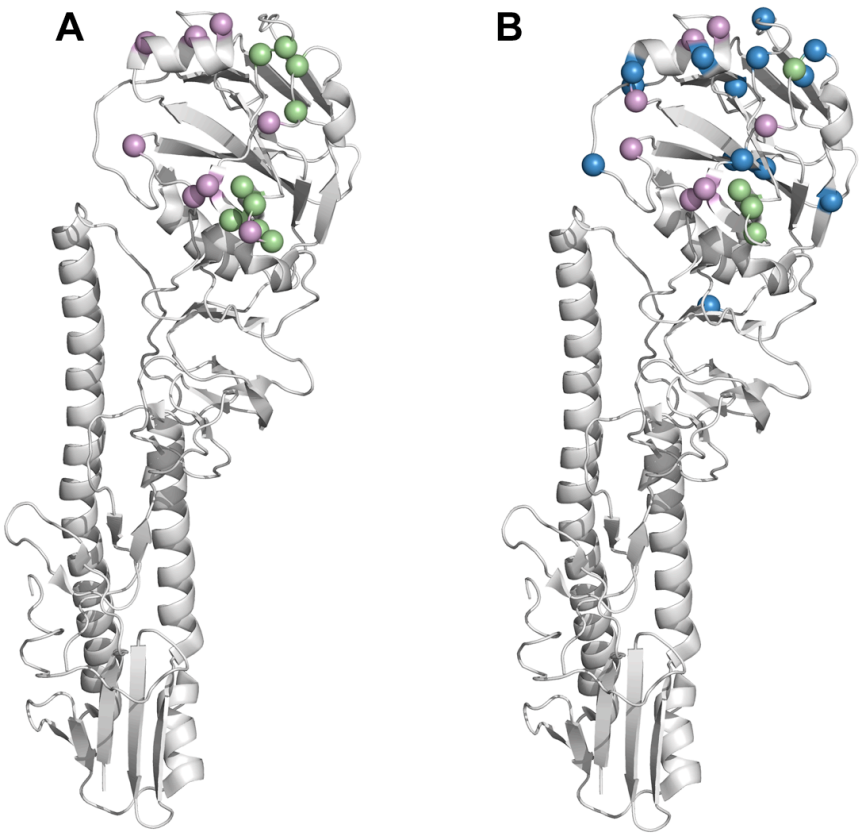

Figure 2. The known specificity determining positions and antigenic sites of HA (A) and predictions (B). The Cα atoms of key positions are mapped as spheres on a monomeric structure of the HA from A/Duck/Singapore/3/97 (pdb 1jsn), shown in a grey cartoon representation. (A) The known specificity determinants (purple) and positions known to form the antigenic sites (green) are presented on the globular receptor-binding domain. (B) The 25 most highly ranked amino acid positions that were predicted to comprise the avian-to-human species barrier of the RBD analysis. Of these, the known specificity determinants and antigenic sites are marked in purple and green (as in A) and new positions are blue. Residues 133 and 143, known as both specificity determinants and antigenic sites, are in purple. Almost all the known positions, and 14 additional new residues, were identified in the highly ranked set.

Overall there are eight known mutations that affect binding preference for the H5 hemagglutinin subtype, thus important for host recognition. Being the main surface protein, HA also features the main antigenic sites of the host immune system. As a means to evade recognition by the immune system, these sites evolve rapidly and accommodate ample mutations, a phenomenon called antigenic drift [22]. This continuous evolution is very significant in choosing which strain will be used for the next vaccine. This may also lead to 'antigenic sin' [23], referring to the 'immune impression' left by a first exposure to an influenza virus, determining all the subsequent responses for different influenza exposures. Kaverin et al. [24] defined three H5 antigenic sites; the first includes amino acid positions 140-to-145 of HA1 that forms an exposed loop near the RBD. The second site includes positions 156-157 of HA1 and the third consists of residues 129-to-133 of HA1 (Figure 2A).

The availability of growing amounts of influenza sequence data that is curated and made publically available through the NCBI influenza database [25], allows the development of computational approaches to study the evolution of influenza strains and their transmission into novel hosts. Indeed, two recent studies employed machine-learning techniques to analyze the molecular basis for host specificity in H5N1. Allen et al. examined all the H5N1 proteins but did not find any of HA's specificity determinants [26], and Wu et al. found only five positions suspected as specificity determinants, four of which were previously ascribed specific roles; two are known host-shifting positions from experiments, another is a glycosylation site and the remaining is part of an antigenic site [27].

Here, I used a similar computational approach for identifying host specificity determinants in HA. My approach uses H5N1 HA protein sequences from both avian and human origin to train a discriminative classifier that attempts to correctly predict

whether a given HA variant can infect human or avian hosts. In order to obtain biologically interpretable results, I used a classification method called Alternating Decision Trees [28] that classifies sequences using a decision tree over different positions along the HA protein. The positions selected by my method consist of almost all known annotated positions on H5N1 HA, including both specificity determinants and residues that form the antigenic sites, but also includes additional novel unknown sites that discriminate between avian and human infecting strains. I then conducted a structural analysis of the novel sites and identified a limited set of novel positions that are predicted to be of functional relevance for determining host specificity.

2. Methods

2.1. Datasets

Hemagglutinin sequences of both avian and human hosts were collected from the NCBI Influenza Database[25]. Duplicate sequences and partial sequences (less than 80% of full length) were removed from the data. Sequences were aligned using the MUSCLE program [29], and alignments were visually inspected to verify their quality. This dataset consisted of 935 avian HA sequences and 136 human HA sequences. I then used this data to create an additional dataset that contained only the receptor-binding (RBD) site (positions 114-268, H3 numbering) [30]. After removing additional duplicates, this dataset contained 544 avian isolates and 91 human isolates. Moreover, I created an additional dataset with dependencies between positions on the RBD. I created an augmented representation of each hemagglutinin RBD sequence which included features of single positions and also features for all pairs of positions. Thus each RBD sequence was represented using $155 + \binom{155}{2} = 12050$ features.

2.2.Computational Analysis

2.2.1. Adaboost Algorithm

Boosting is a machine learning algorithm for performing supervised learning. This is a method for producing a very accurate prediction rule by merging reasonable inaccurate rules-of-thumb [31]. The Adaboost (adaptive boosting) algorithm was first introduced by Freund and Schapire in 1995 [32]. The input for the algorithm is a training set $(x_1, y_1), \ldots, (x_m, y_m)$, where $x_i \subset X$ and each label $y_i \subset Y$, $Y = \{-1, +1\}$ (in our case the avian is -1 and human is +1). For a series of rounds *t*

= 1,..., T, the Adaboost algorithm [32] iteratively trains a *weak learner* with a distribution of weights for the training set. $D_t(i)$ is the weight of the given distribution on training example x_i in round t. In the first step, all weights are set equally for all training examples, and in each round the weights of incorrectly classified examples are increased (and correctly classified examples decreased), thus the weak learner is obliged to concentrate on the 'difficult' examples [32]. Ultimately, the weak learner obtains a weak hypothesis $h_t : X \to \{-1,+1\}$ that minimizes the error with respect to the distribution D_t in round t. For each weak hypothesis, the weighted error rate is $\varepsilon_t = \Pr_{i\sim Dt}[h_t(x_i) \neq y_i] = \sum_{i:h_t(x_i)\neq y_i} D_t(i)$, if $\varepsilon_t \leq 0.5$ and $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$. In the next step, we update $D_{t+1}(i) = \frac{D_t(i)\exp(-\alpha_t \bullet y_i \bullet h_t(x_i))}{Z_t}$, where Z_t is a normalization factor (so that D_{t+1} will be a distribution). The final strong hypothesis, combining the weak hypotheses is $H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$.

2.2.2. Alternating Decision Trees

As described above, Adaboost generates a strong hypotheses combined by simple rules, called weak hypotheses [32]. Theses results are obtained as an unstructured set of T hypotheses, making it hard to infer the correlations between attributes. Alternating decision trees (ADTree) are a generalization of decision trees [28]. Decision stumps are the simplest special case of decision trees which consist of a single decision node and two prediction leaves [28]. The ADTree is built by adding hypotheses according to the iteration they have been produced, introducing a structure to the set of hypotheses. The resulting structure of the set of hypotheses

can be visualized by a tree, that exhibits the connections between a hypothesis and its "parent" [28]. An additional feature of Adaboost is that it also provides a measure of confidence for each prediction, which is called the classification margin. An example of decision tree created by the ADT method is presented in Figure 16. The rectangles in the decision tree are the decision (or splitter) nodes and the ovals are the prediction nodes, the values in each oval correspond to the contribution of that node to the prediction score. The number in each decision node represents the iteration number in which that feature was selected. An instance (in my case a hemagglutinin sequence), defines a path in the alternating tree. When a path reaches a decision node, it continues with the child with the corresponding outcome for the instance in the decision node. The sign of the sum of the scores in the prediction nodes along the selected path, is the classification which the tree associates with the instance [28].

2.2.3. JBoost

Using the described dataset I employed JBoost (http://jboost.sourceforge.net/), an open source, Java implementation of the Adaboost [32] machine-learning algorithm, to identify positions in HA that separate human and avian isolates.

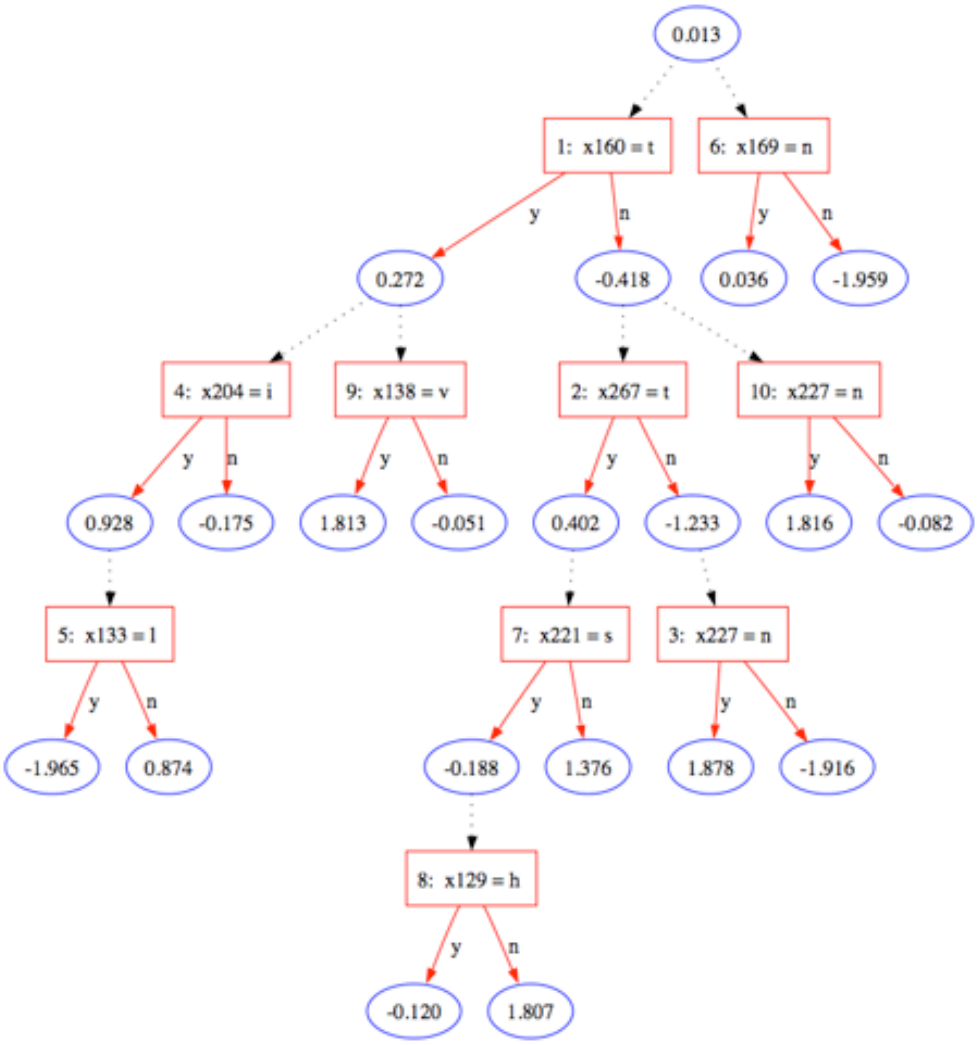

Figure 16. Representative ADT obtained after ten iterations. The ovals in the decision tree are the prediction nodes, and the rectangles represent the splitter nodes. Starting from the score of the top prediction node, and summing the scores of the relevant prediction nodes that meet the conditions of the splitter nodes, provides the final prediction score.

Ultimately, classifiers in a form of Alternating Decision Trees (ADT) [28] are generated. It has been recently used in several successful biological applications [33,34]. In my setting each data instance is an influenza HA sequence, so the dimensionality of each data point is N=530, for the entire HA data and 155 for the RBD dataset. The data labels are the species from which each isolates was obtained – human or avian. The algorithm uses the data and labels to learn an

ADT that can then be used to predict whether a given isolate was collected from a human or avian host. Ultimately the set of positions that best discriminate between human and avian isolates is selected.

2.2.4. K - fold Cross Validation

Cross-validation is an approaches for estimating how well the a model that learned from some training data is going to perform on future unseen unlabeled data (test data) [35]. In K-fold cross validation, the original training data is randomly separated into K subsets. One of the K subsets is kept as validation data for testing the model, and the remaining K – 1 subsets are used as training data. The cross-validation procedure is repeated K times, with each of the K subsets used exactly once as the validation data. Finally, the results from the K different runs are averaged. In order to measure the predictive power of my proposed method over test data I performed 10 runs of 5-fold cross-validation experiments over 100 iterations of the Adaboost algorithm, producing 50 different runs altogether.

2.2.5. Stopping criteria

While boosting algorithms have been shown to be empirically robust to over fitting, some simple criteria for choosing the number of iterations have been suggested. Here I used a stopping criteria based on the convergence of the distribution of margins over all training points. Specifically, let us denote by m_t^i the margin of the *i*-th data point in iteration *t*, and by S_t the average margin over all data point in iteration *t* - $S_t = \frac{1}{N}\sum_{i=1}^{N} m_t^i$. My stopping criterion is defined

by $(S_{t+1} - S_t)^2 < \varepsilon$, where $\varepsilon = 10^{-5}$. Moreover, it is important to examine the test error and stop when it asymptotes or begins to increase.

2.2.6. Adjusting for biases in training set size

As H5N1 is currently not human to human transmissible, there are significantly less H5N1 isolates from humans (136) then avian isolates (935) for the full sequence. A standard technique in boosting to account for biases in the label distribution is to reweight the data such that each label has equal weight. This is easily done in boosting algorithms, where each point *i* is associated with a weight w_t^i in each iteration, by tweaking $W_1 = (w_1^1, w_1^2, \cdots, w_1^N)$ to be such that $\sum_{l(i)=Human} w_1^i = \sum_{l(i)=Avian} w_1^i$. This forces the algorithm to equally focus on human isolates and on avian isolates in the first initial rounds of training.

2.3.Measuring the informativeness of selected features

In order to assess the importance of the selected features over the different decision trees created I developed a novel scoring function that is used to rank positions selected by the algorithm. My scoring function is an extension of the one suggested by Creamer et al. [36] Intuitively, given a set of decision trees generated using many different partitions of the data into train and test data, a feature is more important if it appears in many of the trees, is selected in earlier boosting iterations. Moreover, since my main concern is predicting mutations that may enable a shift of binding specificity towards a human receptor, my scoring function also takes into account the relative contribution of a given feature in assigning a sequence to the human class. More

formally, the score of a given feature i is given by $S(i) = n_i * m_{iter} * \max_{d(i)}(p_{human})$, where n_i is the number of appearances of feature i in the set of trees, m_{iter} is the mean iteration in which feature i appears, and $\max_{d(i)}(p_{human})$, is the maximal value of the human label prediction nodes taken over all of the decision nodes that contain feature i.A larger contribution score implies a greater importance of the feature for the human prediction.

2.4.Decision of the cutoff for top ranked-position

In order to choose a cutoff of a smaller subset from the list of ranked positions, I looked for a set that would cover 70% of the cumulative distribution of the computed ranking scores (Figure 4 and Figure 11).

3. Results

3.1. Analysis of the Receptor-Binding Domain of the HA protein

In order to identify a set of positions that best discriminate between human and avian isolates, I trained the Alternating Decision Tree (ADT) algorithm on the receptor-binding domain, residues 114-268 [30], of the set of human and avian H5N1 HA protein sequences. The overall mean accuracy of the model with ten runs of 5-fold cross validation was 85% (Figure 3) (86.3% and 70.5% for avian and human isolates respectively).

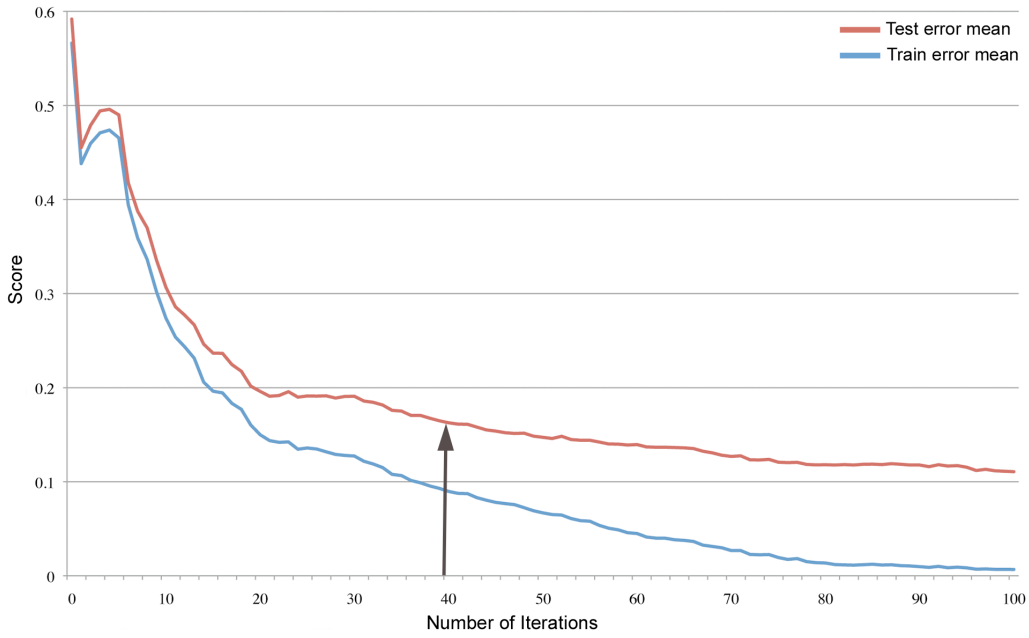

Figure 3. Mean train and test error. A plot of the mean train and test error was calculated over ten runs of 5-fold cross validation, each with 100 iterations. The blue and red curves represent the mean train and test errors, respectively. The arrow indicates the iteration number that was chosen as last. This iteration was computed by the stopping criteria described in Methods.

The full set of positions selected by the algorithm, included all the known specificity determinants (Table 2). The algorithm also detected all eleven residues known to form the three H5 antigenic sites [24] (Table 2). In order to grade the selected positions by importance, I developed a ranking function, based on a score suggested by Creamer et al. [36]. Using this function, I chose to further analyze the 25 most highly ranked positions (Figure 4, Table 2).

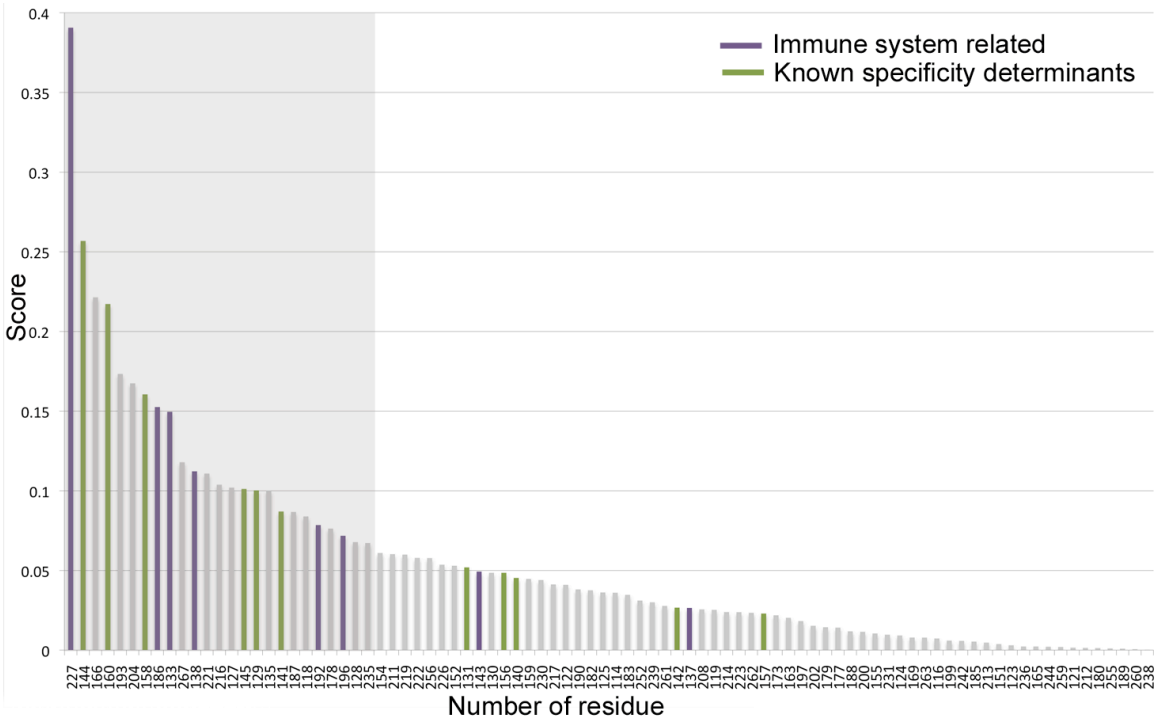

Figure 4. Ranking of the amino acid positions that emerged from the receptor binding domain analysis. Known specificity determinants are in purple, and immune system related are in green. The most highly ranked region of the distribution is shaded in gray. Residues 143 and 133 are known as both specificity determinants and antigenic positions. They are coloured in purple.

Six out of eight known specificity determinants and five out of eleven positions from the antigenic sites appeared amongst the set of the 25 most highly ranked

positions (Figure 5), and all but two antigenic sites were included in the 40 highest ranked positions. In addition to these, the algorithm also detected 14 new positions that have not been previously annotated (Figure 2B, Figure 6, Table 2). Indeed the new positions may be good candidates for specificity determinants, but some of them are particularly favorable and intriguing structure-wise. A few examples are presented below.

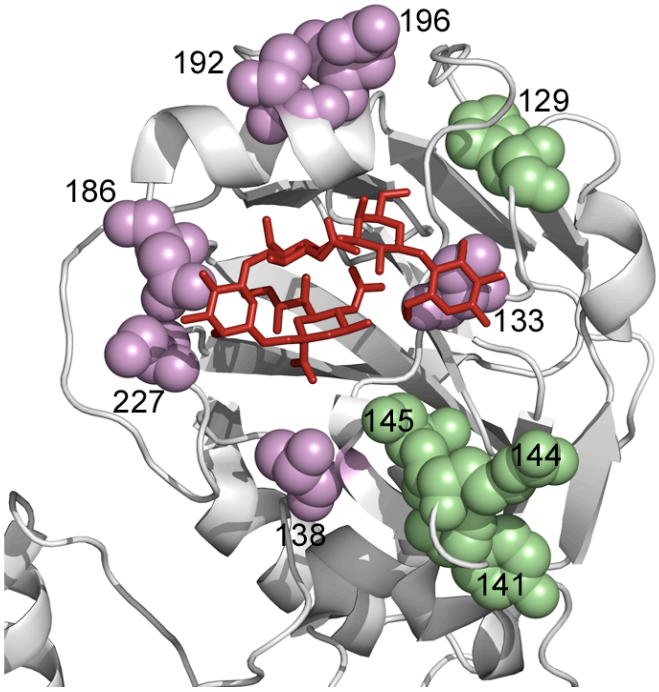

Figure 5. Known positions that were amongst the 25 most highly ranked positions in the receptor binding domain analysis. The known positions were mapped on the receptor binding domain of the H5N1 hemagglutinin structure in complex with a human receptor analogue (LSTc) in red (pdb 1jsn). Known specificity determinants are coloured in purple, residues forming antigenic sites are coloured in green.

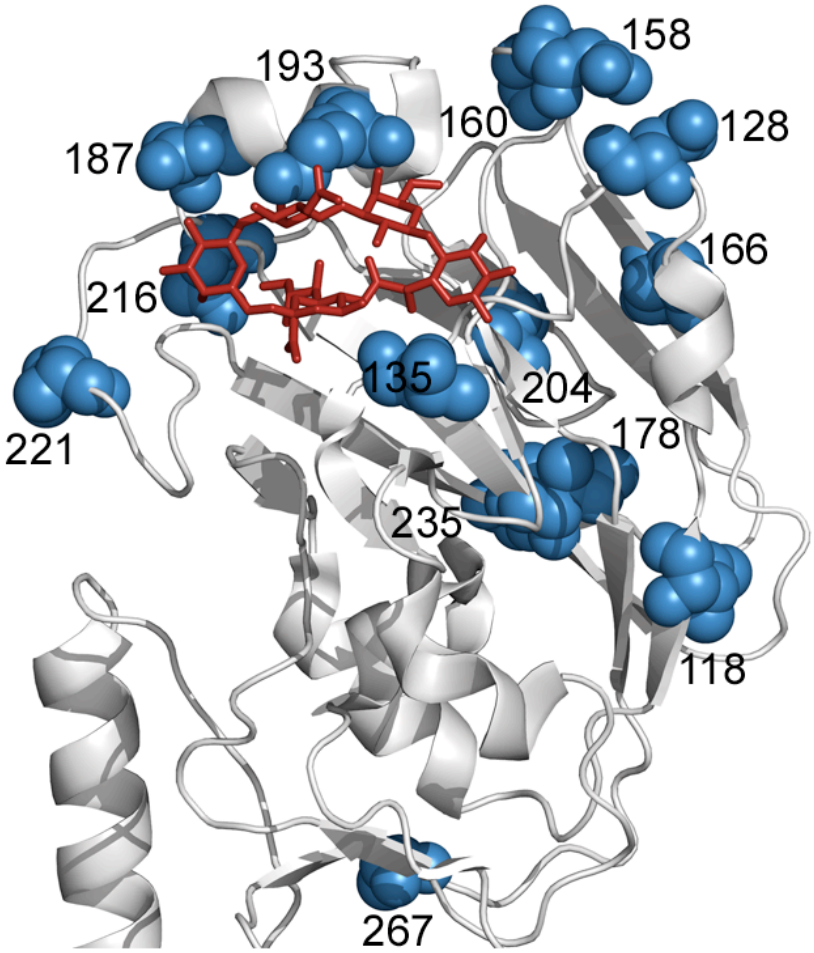

Figure 6. The new putative positions that comprise the avian-to-human species barrier. Mapping of the highly ranked amino acid positions predicted here to differentiate between avian and human on the HA receptor binding domain in complex with a human receptor analogue (LSTc) in red. (Same data as in Fig. 2A but excluding the known functional positions.) The highly ranked residues are highlighted using an all atom representation with blue van der Waals spheres, and the receptor molecule is presented using sticks representation. The LSTc was modeled into the RBD by superimposing the structures of H5 A/Duck/Singapore/3/97 and H9 A/Swine/Hong Kong/9/98 HAs (PDB 1JSN and 1JSI) with α2-3 and α2-6 analogs bound. The highly ranked positions are in close proximity to the receptor and in functionally significant locations.

Residues 128 and 135

Substitution V135G was suggested to enable a shift in specificity from avian to human. V135 is in the receptor-binding pocket where it binds the sialic acid via its backbone atoms (Figure 7). Hence, alteration to glycine may increase the backbone flexibility owing to its much smaller side-chain, affecting in turn the tight contacts with the ligand.

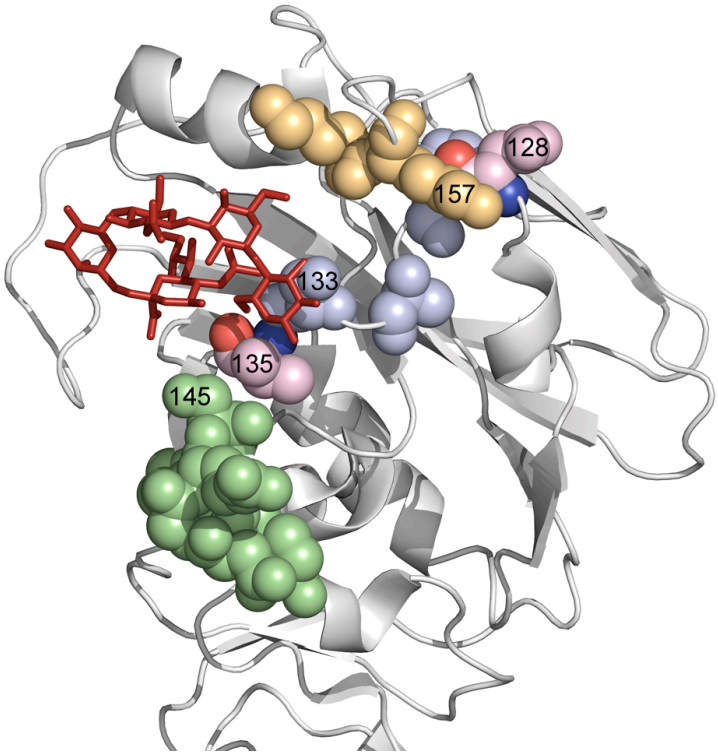

Figure 7. Residues 128 and V135. The HA receptor binding domain of the A/Duck/Singapore/3/97 (pdb 1jsn) is shown in grey cartoon representation. Residues 128 and V135 are shown as atoms spheres models, with the oxygen atom in red, the nitrogen atom in blue, and the sidechain in pink. The positions of the first known antigenic site (140-145) are in green atoms spheres model, the second antigenic site (129,131,133) are in light blue and the third antigenic site (156,157) are in light yellow. Residue 128 and V135 are in direct contact with the antigenic sites. V135 is also in contact with the receptor.

As aforementioned, Kaverin et al. [24] identified three antigenic sites: the first stretches between positions 140-145, the second includes positions 129,131,133 and the third 156-157. Interestingly, V135 is in contact with positions 133 and 145; insinuating that it might even be a part of these antigenic sites (Figure 7). Similarly, residue 128, also detected as a host specificity determinant, is in direct contact with positions 129 and 157 and could be a part of these antigenic sites (Figure 7).

A160T

The mutation A160T is suggested to alter the virus receptor binding specificity to human. Moreover, a substitution to asparagine or threonine in 158 is also proposed as a human binding characteristic (Figure 6). Interestingly, the alteration from alanine to threonine in position 160 introduces a glycosylation site in aspargine 158 [37,38], which is known to promote immune escape of the virus by masking antigenic sites.

Residues K193R and 187

Residue 193 is located in the receptor-binding domain and very close to the ligand, therefore, a mutation in this position may affect binding directly (Figure 6). Furthermore, the introduction of the mutation K193R into the H5N1 human infecting strain A/Vietnam/1203/2004 with the known H3 mutations Q226L and G228S, resulted in a considerable increase in binding to α2-6 glycans, but only a minor reduction for avian α2-3 sialoside [39]. Interestingly, Kaverin et al. 2007 [40] and Philpott et al. [41], showed that this residue was antigenically significant. Similarly, residue 187 is located in the vicinity of the receptor as well, and a mutation in this position might affect binding directly (Figure 6).

Residues V204I, 216 and 221

Positions 204 and 216 are in the interface between the receptor binding pockets of adjacent monomers (Figure 8, Figure 9). A mutation in these residues may cause an alteration in the structure of the binding sites and affect binding affinity and specificity. Also at the interface with a neighboring subunit, substitution S221P is predicted to shift the receptor-binding domain to be human-like. Besides the potential effect on inter-chain contacts, this residue is in a loop that is in close vicinity to the receptor-binding pocket (Figure 8).

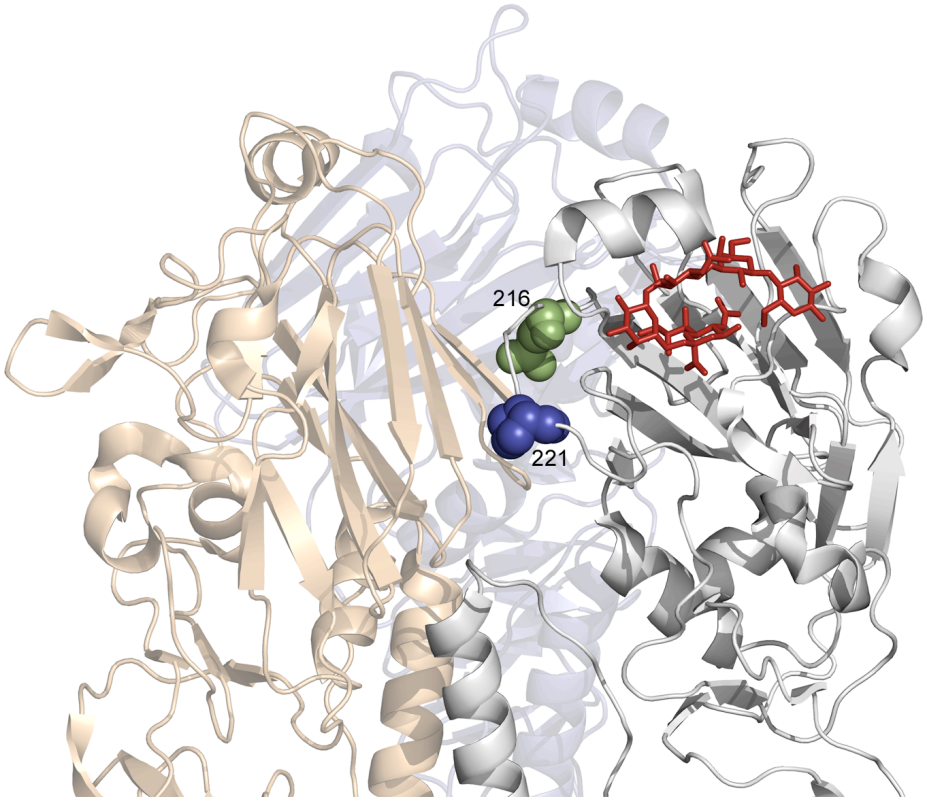

Figure 8. Residues 216 and 221. The trimer is shown in cartoon representation, with each monomer in different color. Residues 216 and 221 of the grey monomer are represented using an all atom van der Waals spheres models in green and blue respectively. The human receptor analogue (LSTc) is shown in sticks representation in red. The residues are at the interface between the receptor binding domains of the adjacent monomers. Position 221 is also in close proximity to the ligand.

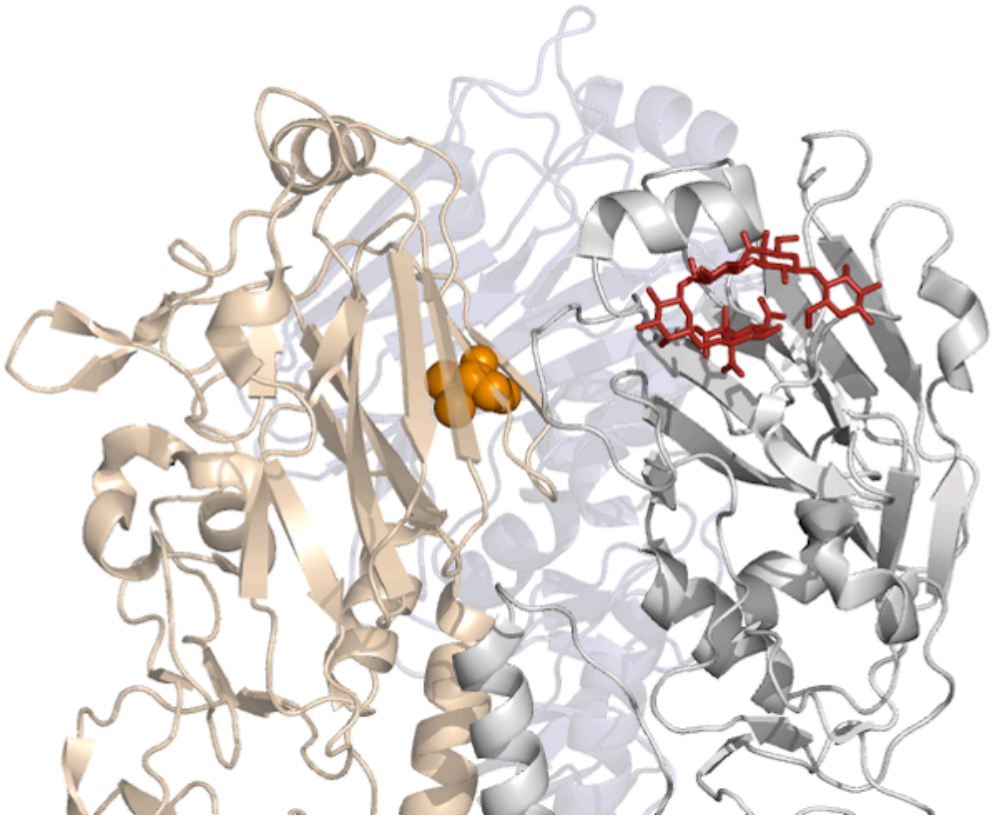

Figure 9. Amino acid 204. The trimer is shown in cartoon representation. Each monomer is coloured differently. Residue 204 of the grey monomer is represented by an all atom spheres model in orange. It is at the interface between the receptor binding domains of adjacent monomers. The human receptor analogue (LSTc) is shown in sticks representation and coloured in red in the adjacent monomer.

Proline determines directionality and stability of loops [42], and the mutation may alter the loop conformation and the structure of the binding site. Interestingly, the H5N1 viral isolate A/Vietnam/1203/2004, which is known to be among the most pathogenic in mammalian models [43], consisted of an arginine in residue 216 and serine in 221 in contrast to the avian H5N1 virus A/Duck/Singapore/1997 [Stevens et al. 2006].

Residue 178

Highly buried in the protein core, therefore a mutation may affect the packing and cause an alteration in the protein and receptor binding site structure (Figure 6).

A267N/T

Located in between chains HA1 and HA2, in close proximity to positions 70 and 71 of chain HA2 (Figure 6, Figure 10). It is predicted that a mutation from alanine to asparagine or threonine, may endorse a change in binding specificity towards a human receptor. A substitution like this may alter the current arrangement among the chains and consequently change the orientation of the receptor-binding domain.

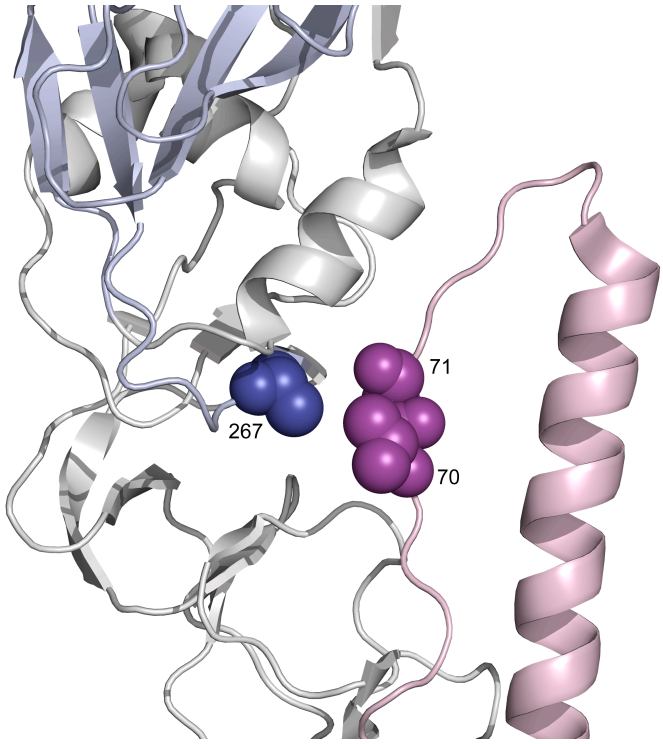

Figure 10. Amino acid 267. Chain HA1 of the hemagglutinin is in gray; HA2 in light pink and the receptor-binding site of HA1 is in light blue. Important residues are represented in all-atom spheres models: amino acid 267 is in blue and 70 and 71 in purple. Substitution from alanine to aspargine in position 267 was predicted to shift the binding specificity to a human host. Based on the structural location of this residue, I suggest that it might affect the inter-chain contact between HA1 and HA2, resulting in structural alteration in the orientation of the receptor-binding domain relative to HA2.

3.2.Analysis of simple cases: H1N1 and H3N2

The influenza strains H1N1 (excluding the 2009 sequences associated with the swine flu pandemic) and H3N2 are well established in the human population [11]. Hence, I hypothesized that the analysis for these strains would be less complex with a higher ability to distinguish between the sequences from the two hosts.

The hemagglutinin avian and human sequences were collected from the NCBI influenza database [25] and the data set was composed following the same method as described for the H5N1 strain. The resulting data sets consisted of 56 avian and 533 human sequences for the H1N1 strain and 53 avian and 1309 human sequences for the H3N2. The H1N1 sequences of new swine origin pandemic were excluded. The method was applied to analyze these strains. As expected, known specificity determinants of these strains (Table 1) have been detected for both subtypes, and the overall mean test accuracy of the model (with ten runs of 5-fold cross validation) for the H1N1 strain was 99.1% (94.4% and 99.5% for avian and human isolates respectively) and for the H3N2 strain it was 99.4% (94.7% and 99.7% for avian and human isolates respectively). These results demonstrated that the ability to differentiate between the sequences of the two hosts for the two strains was much higher than the H5N1, thus highlighting the convoluted barrier between the hosts of this strain, which is yet to be crossed.

3.3.Analysis of the whole HA protein

In order to further verify that my approach is capable of identifying functionally important sites, I conducted a second set of experiments in which the algorithm was provided with full HA sequences (rather than the receptor binding domain alone),

following the same method as described for the receptor binding domain analysis.

I hypothesized that a significant number of the detected sites would overlap with the sites selected when analyzing the receptor-binding domain and that in general, most discriminative sites would be in the receptor-binding domain. I looked at the 50 most highly ranked positions (Figure 11, Table 3). Indeed, 26 of the 50 most highly ranked positions (i.e., over 50 percent) were in the receptor-binding domain (Figure 12). Moreover, seven of the eight known specificity determinants and seven of the eleven known positions from the antigenic sites were amongst these highly ranked residues (Table 3).

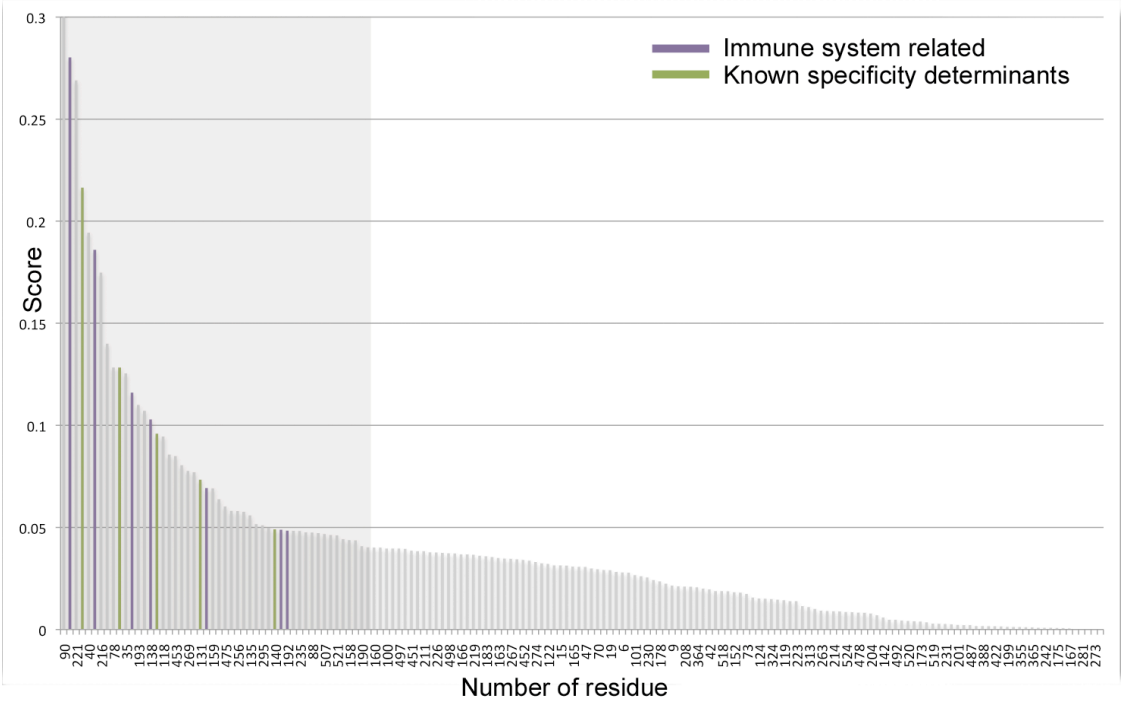

Figure 11. Ranking of the amino acid positions that emerged from the whole HA analysis. Known specificity determinants are in purple, and immune system related are in green. The most highly ranked region of the distribution is shaded in gray. For clarity, the maximum value of the score was restricted to 0.3. While position 90 received the score 0.61, all other positions were below the 0.3 threshold.

The results demonstrate the power of the approach and its ability to identify the known functional regions and residues, even when provided with a very large set of features (530 positions). Moreover, taken together with the detection of known positions in the whole HA and receptor binding domain analysis reinforces the importance of the highly ranked residues selected.

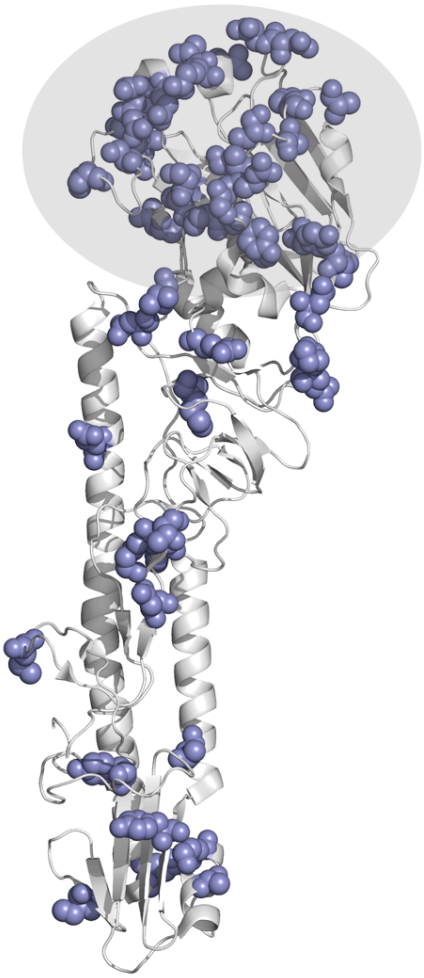

Figure 12. Highly ranked positions for the host classification of the whole HA. Mapping of the 25 most highly ranked positions on the H5N1 HA structure (pdb 1jsn). The residues are highlighted in blue using an all atom spheres representation. The receptor-binding site is encircled.

3.4.Analysis of amino-acid pairs in the receptor binding domain

While I found that many antigenic sites and specificity determinants were selected by the algorithm as positions that discriminate between human and avian isolates, I then turned to ask if there might be dependencies between positions on the receptor binding domain, using the same method employed on the previous datasets. I trained the ADT algorithm on this dataset and ranked the features using the scoring function described above. The trees obtained in these runs mostly contained pairwise features.

The overall mean accuracy of this model averaged over ten runs of 5-fold cross validation was 85% (Figure 13) (86.7% and 70.2% for avian and human isolates respectively), which was comparable to the results obtained with the single features, providing no indication of overfitting due to the large amount of features.

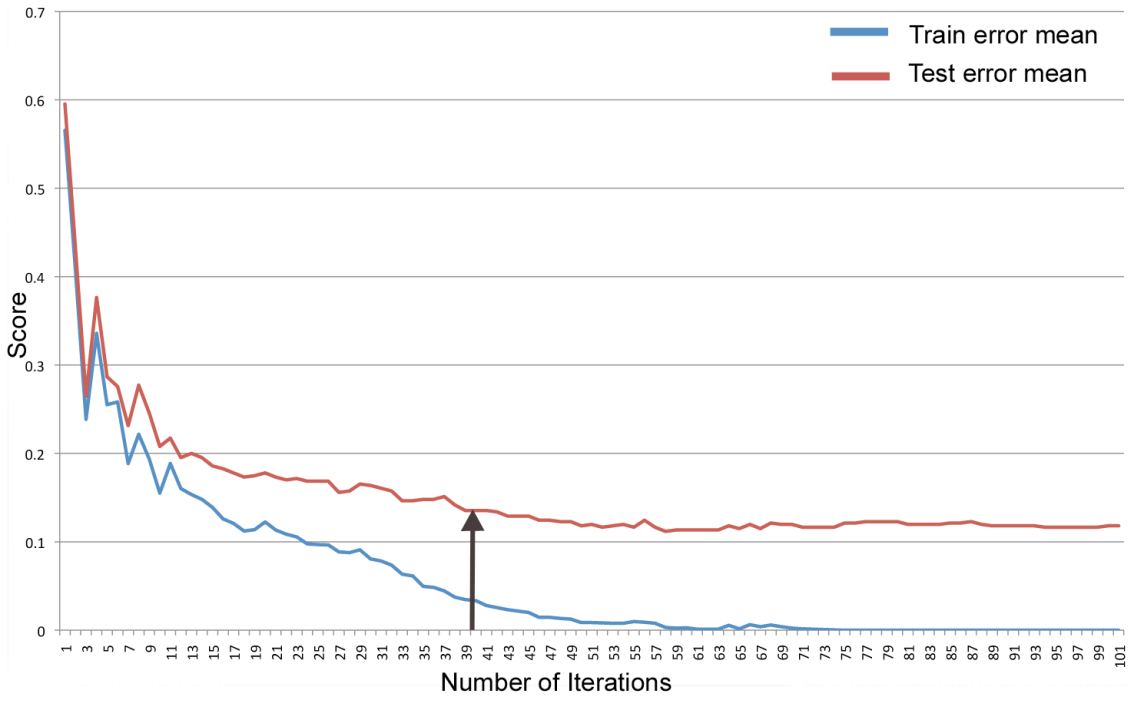

Figure 13. Mean train and test error of couple analysis of the RBD. A plot of the mean train and test error was calculated over ten runs of 5-fold cross validation, each with 100 iterations. The blue and red curves represent the mean train and test errors, respectively. The arrow indicates the iteration number that was chosen as last. This iteration was computed by the stopping criteria described in the Methods.

The full set of positions over all runs consisted of 1420 pairs. Using the ranking function I described before (see methods), I graded the pairs by importance. The distribution of the scores, showed an exponential decay-like behavior (Figure 14). Therefore, in order to determine the cutoff of the highly ranked data set, I looked for the rate of decay (exponential time constant) of second order. I chose to further analyze the 256 most highly-ranked pairs, composed of 74 unique positions (Table 4), including all the 8 known specificity determinants and the 11 known residues that form the antigenic sites.

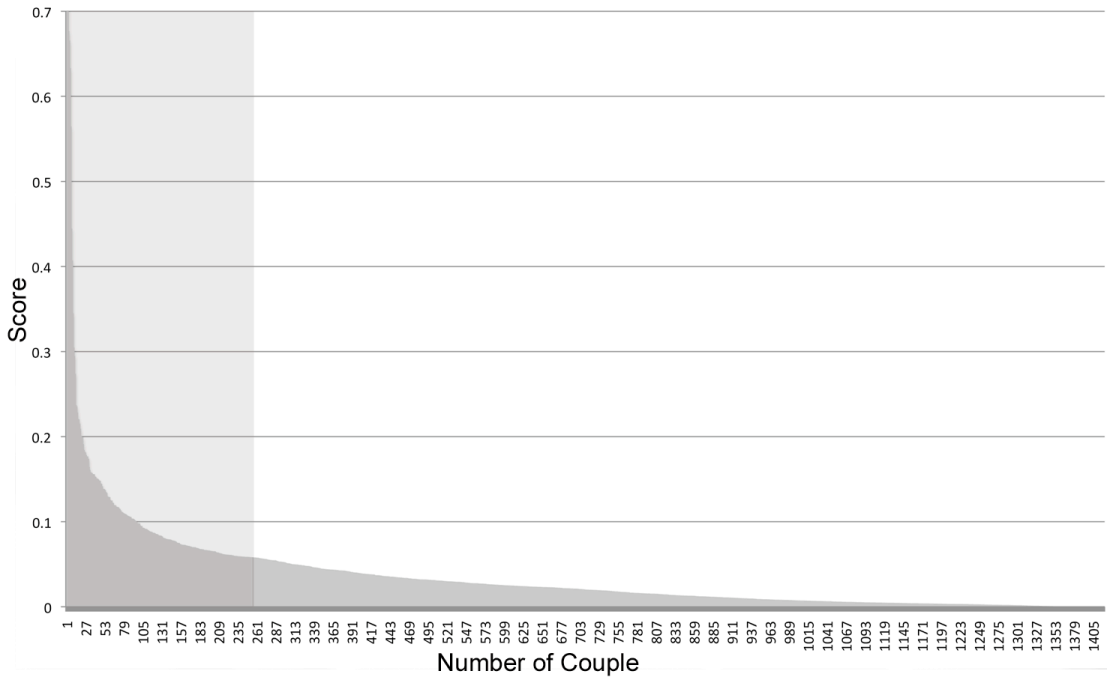

Figure 14. Ranking of the amino acid couples received from the pairwise receptor binding domain analysis. The most highly ranked region of the distribution is shaded in gray. An exponential decay behavior is seen here, and the exponential decay formulas were used to choose the marked cutoff- 256.

Next, I graded each single position using a *cumulative rank* (Table 4) that is produced by summing the scores of all the couples the position appeared in, indicating the importance of each residue in the couple analysis. Most reassuringly, nearly all the known residues and novel detected residues from the single-residue RBD analysis (see Table 2) appeared with the highest scores (Figure 15), indicating these residues have a significant role within the context of couples as well.

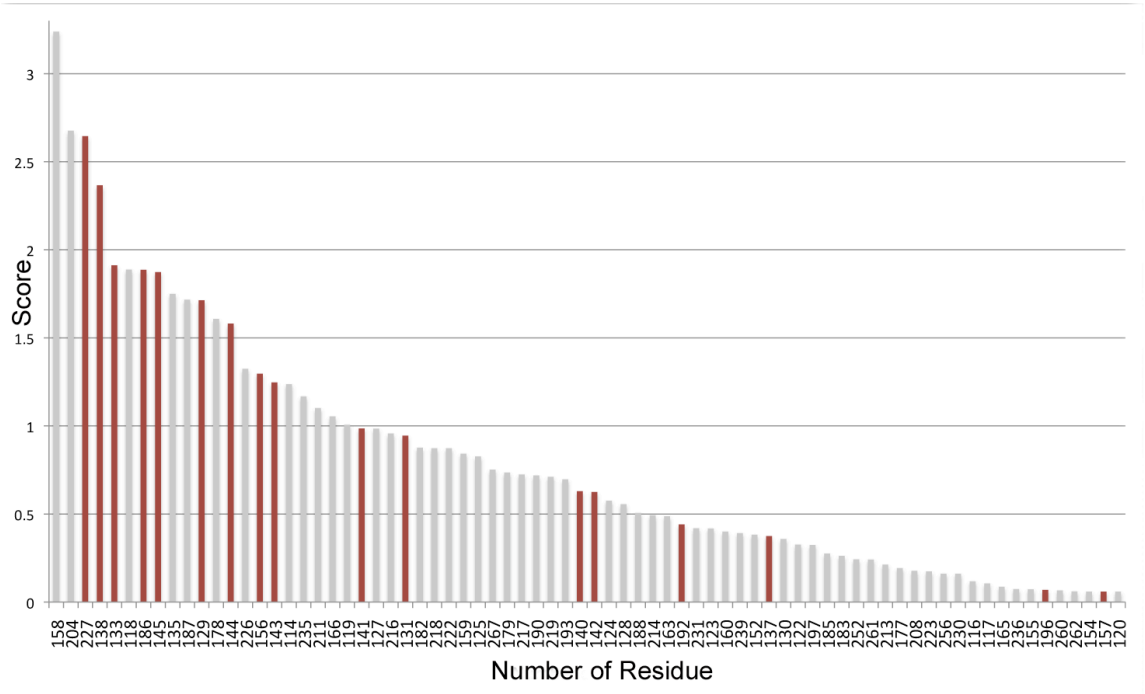

Figure 15. Ranking of the singular amino acid positions by their importance in the pairwise analysis of the receptor binding domain. Known specificity determinants and immune system related are in red. It is clear that most of the known residues are very highly ranked.

Discussion

I used a computational method to identify residues that best discriminate human and avian isolates. While a variety of classification algorithms could have been used, I used sequence-based decision trees, which have recently been applied to several interesting biological applications [32,33]. Decision trees have the advantage of identifying positions along the sequence that are discriminative but also provide a biologically interpretable output.

To evaluate the biological relevance of the computational analysis, one would turn to well known, experimentally validated, data to examine the correspondence between the computational and empiric findings. In my case, however, the available experimental evidence is mostly limited to a few specific human and avian virus variants rather than a thorough evaluation on each of the HAs in my sequence collection. Hence, there is no gold standard to which I can globally compare my results. Keeping in mind this impediment, I concentrated on the restricted available data, consisting of several positions identified as specificity determinants as well as residues implicated as antigenic sites. In spite of the limitations described above, when I applied my approach to the whole HA protein, I found that, as expected, the vast majority of highly ranked amino acid positions are in the receptor binding domain and proximate regions on the 3D structure (Figure 12), as they should.

The H5N1 virus has not yet crossed the species barrier into humans, and current infections of humans are sporadic and mostly independent of one another. Moreover, the identity between sequences from the two hosts is very high: in some cases only one amino acid distinguishes between the groups. Therefore, it is somewhat surprising that the algorithm managed to identify positions in HA that can be used to classify the infecting host of a given strain. I hypothesized that for other influenza subtypes, such

as H1N1 (excluding the sequences of the new swine-origin pandemic) or H3N2, which have crossed the species barrier between the avian and human hosts a long time ago [11], the ability to identify host specific mutations would be less challenging. Indeed, the accuracy of the method with these strains was significantly higher than with the H5N1 strain, emphasizing the intricate nature of the H5N1 problem and how fragile the species barrier is in this particular case.

Encouragingly, the highly ranked positions detected by my analysis included, in essence, almost all known specificity determinants and antigenic sites, indicating that indeed the method can identify functionally important sites on the HA protein. Most intriguingly, my method also identified a large number of antigenic sites that also discriminate between human to avian isolates. It was a bit of a surprise for me that the antigenic sites came out as specificity determining, given the highly variable nature of these positions, which the virus uses to escape immune detection via antigenic drift. I note however, that the recognition of the virus by the immune system also has a direct effect on the efficiency of infection. Recent work by Recker et al. [44] suggested that the evolution of antigenic sites may be constrained due to functional constraints. My results are in agreement with such a model, as the discriminative power of antigenic sites emerges from the presence of specific amino acids at these sites in human vs. avian isolates.

I also explored the possibility that the avian-to-human species barrier is determined by correlated mutations rather than single substitutions. To this end, I ranked all amino acid pairs of the receptor-binding domain according to their contribution to the classification. I was encouraged to see that all known positions, i.e., the specificity determinants and antigenic sites, were amongst the highly ranked pairs. Moreover, the cumulative rank (Table 4) showed that the vast majority of the

known residues and the novel positions detected by the single-residue receptor-binding domain analysis described above (Table 2) ranked high (Table 4, Figure 15). This analysis emphasized the importance of the pairwise amino-acid context of known specificity determinants, antigenic residues and novel positions detected by my analysis. Thus, illustrating the need to further understand the network of interactions of these positions which are responsible for characterizing, at least in part, the host specificity barrier.

The successful reproduction of all the known functional annotations in HA, coupled with the structural analysis of novel positions identified by my approach, suggests that other positions that were identified by the approach could be promising candidates for further experimental research on identifying host specificity determinants and antigenic sites. Comparing my analysis to that of Wu et al. [27], I successfully covered the five suggested positions by the previous study, whereas suggesting a significant number of novel candidates.

Interestingly, all the highly ranked novel positions detected in this analysis surrounded the sialic acid binding pocket, antigenic sites or were in the protein core. The strategic location of these residues reinforced their possible significance for binding to the host receptor, antigenicity or for contribution to structural stability. I thus speculate that mutations in these residues may cause an alteration in the binding specificity and recognition of antigenic sites for specific hosts. This hypothesis awaits experimental examination.

Herein I attempted to disclose molecular features involved in host binding specificity of HA, which mediates the first step in the virus infection. I restricted my analysis to HA since it is responsible for the binding to the host cell. Identifying the set of positions in HA that alter binding specificity may help improve my

understanding of the molecular details of the underlying mechanisms. It may also assist in developing surveillance tools that can help monitor viral populations in bird markets and other areas of intense contact between humans and birds. The identification of novel specificity determinant and antigenic sites may also prove useful for designing antiviral drugs which could inhibit viral entry, or novel neutralizing antibodies.

Nevertheless, as the binding process is composed of a combination of various events, which are yet to be fully understood, the same analysis could shed light on other Influenza proteins that may participate in determination of host specificity. Moreover, with an appropriate dataset, the method can be used to detect factors involved in the formation of the species barrier in other influenza strains and hosts, in particular the new intimidating swine-origin H1N1 human pandemic. Furthermore, the analysis is readily applicable for other viruses in which the same phenomenon exists.

Tables

Table 1. Known specificity determinants of the various HA subtypes. Known specificity determinants in H5N1 are marked in bold and shaded in gray.

Residue Number	Subtype	Reference
E190D	**H1**	**[16]**
G225D	**H1**	**[16]**
Q226 L	**H2, H3**	**[17]**
G228S	**H2, H3**	**[17]**
L133V	H5	[19]
S137A	H5	[20]
A138V	H5	[19]
G143R	H5	[18]
N186K	H5	[18]
T192I	H5	[20]
Q196R	H5	[18]
S227N	H5	[21]

Table 2. All positions detected by the algorithm for the analysis of the receptor-binding domain sequence of the hemagglutinin protein, over the 10 runs of 5-fold cross-validation. Overall there are 88 positions identified. Chosen highly ranked positions are marked in a black box.

Contribution Score	Residue in H3 numbering	Known Position?	Comment
0.390623	227	Specificity determinant	
0.256753	144	Antigenic position	
0.221384	166		
0.217168	160		If T in 160, glycosylation in N158
0.173285	193		Close to receptor
0.167396	204		Interface with adjacent monomer
0.160472	158		If T in 160, glycosylation in N158
0.152564	186	Specificity determinant	
0.149563	133	Specificity determinant + Antigenic position	
0.117856	267		In interface between chain HA1 and HA2
0.112168	138	Specificity Determinant	
0.110753	221		Interface with adjacent monomer
0.103787	216		Interface with adjacent monomer
0.101954	127		
0.101093	145	Antigenic position	
0.100221	129	Antigenic position	
0.10013	135		Close to receptor and antigenic site 140-145
0.086962	141	Antigenic position	
0.086692	187		Close to receptor
0.083868	118		
0.078505	192	Specificity determinant	
0.076229	178		Buried
0.071746	196	Specificity determinant	
0.06776	128		
0.067193	235		
0.060958	154		Buried

0.060216	211		
0.059922	219		
0.057841	222		
0.057684	256		
0.05359	226		
0.052894	152		
0.051897	131	Antigenic position	
0.049271	143	Specificity determinant + Antigenic position	
0.048489	130		
0.048483	156	Antigenic position	
0.04524	140	Antigenic position	
0.044683	159		
0.043946	230		
0.041173	217		
0.040936	122		
0.038021	190		
0.037497	182		
0.036177	125		
0.035952	114		
0.034675	183		
0.031036	252		
0.029965	239		
0.027732	261		
0.026606	142	Antigenic site	
0.026457	137	Specificity Determinant	
0.025536	208		
0.025221	119		
0.023851	214		
0.023795	223		
0.023369	262		
0.022926	157	Antigenic position	
0.021827	173		
0.020264	163		
0.01821	197		
0.015242	202		
0.014321	179		
0.014076	177		
0.011776	188		
0.011429	200		
0.010419	155		
0.009628	231		
0.009165	124		
0.007875	169		
0.007833	263		
0.00728	116		

0.005929	199		
0.005739	242		
0.005321	185		
0.004671	213		
0.003738	151		
0.002992	123		
0.002244	236		
0.002221	165		
0.002067	244		
0.001924	259		
0.00151	121		
0.00139	212		
0.001282	180		
0.000982	255		
0.000899	189		
0.000532	260		
0.000214	238		

Table 3. All positions detected by the algorithm for the whole sequence analysis of the hemagglutinin protein, over the 10 runs of 5-fold cross-validation. Chosen highly ranked positions are marked in a black box.

Contribution score	Number of residue	Known Residue?	Comment
0.607006	90		
0.280143	186	Specificity determinant	
0.268896	221		In interface with adjacent monomer, close to receptor
0.216394	144	Antigenic Position	
0.194362	40		
0.185909	227	Specificity determinant	
0.174777	216		In interface with adjacent monomer
0.139967	98		
0.128294	78		
0.128294	129	Antigenic Position	
0.125435	35		
0.11597	137	Specificity Determinant	Close to receptor
0.109985	193		Close to receptor
0.107093	127		
0.102908	138	Specificity determinant	
0.095879	141	Antigenic Position	
0.094466	118		
0.085681	217		
0.084863	453		
0.080476	447		
0.077656	269		
0.076984	11		
0.073367	131	Antigenic Position	
0.069219	133	Specificity determinant + Antigenic Position	Close to receptor
0.069074	159		
0.063746	408		
0.060273	475		
0.058084	356		
0.057967	256		
0.057619	187		Close to receptor
0.055808	135		Close to receptor
0.051593	363		
0.051007	295		
0.049808	57		
0.049012	140	Antigenic Position	
0.048825	143	Specificity determinant + Antigenic Position	

0.048447	192	Specificity determinant	
0.048268	76		
0.048249	235		
0.047525	472		
0.047457	88		
0.047168	533		
0.046728	507		
0.04625	197		
0.04607	521		
0.044242	79		
0.04378	158		If T in 160, glycosylation in N158
0.043684	39		
0.040761	190		Close to receptor
0.040165	24		
0.040117	160		Mutation A to T introduces glycosylation site in 158
0.040075	480		
0.039687	100		
0.039664	239		
0.039619	497		
0.039484	222		Close to receptor
0.038604	451		
0.038387	276		
0.038335	211		
0.037809	7		
0.037713	226		Close to receptor
0.03753	515		
0.037319	498		
0.037238	95		
0.036763	166		
0.036731	319		
0.036536	219		Close to receptor
0.036068	20		
0.035835	183		Close to receptor
0.035424	500		
0.035009	163		
0.034697	509		
0.03465	267		In interface between chain HA1 and HA2
0.034375	5		
0.034057	452		
0.033702	290		
0.03304	274		
0.032411	465		
0.032063	122		
0.031397	238		

0.031381	15		
0.03125	448		
0.030845	165		
0.030669	87		
0.030658	47		
0.029834	252		
0.02947	70		
0.029092	49		
0.029036	19		
0.0282	529		
0.027932	6		
0.027777	341		
0.026593	101		
0.026049	14		
0.025497	230		
0.024199	112		
0.023528	178		Buried
0.022423	479		
0.021462	9		
0.021164	326		
0.020993	208		
0.020901	359		
0.020649	364		
0.01999	182		
0.019586	42		
0.018809	406		
0.018754	518		
0.018727	128		
0.018247	152		
0.01809	113		
0.017412	73		
0.015625	116		
0.0152	124		
0.015105	188		
0.014874	324		
0.014608	75		
0.014348	119		
0.013858	350		
0.01382	223		
0.011437	272		
0.011003	313		
0.010034	314		
0.009218	263		
0.009031	286		
0.00896	214		
0.008862	179		
0.008558	524		
0.008437	367		

0.00823	478		
0.00814	123		
0.007804	204		
0.006927	155		Close to receptor
0.005875	142	Antigenic Position	
0.004758	383		
0.004696	492		
0.004338	169		
0.00421	520		
0.003992	236		
0.003924	173		
0.003477	372		
0.002856	519		
0.00279	459		
0.002647	231		
0.002466	493		
0.002233	201		
0.002168	125		
0.002086	487		
0.001656	145	Antigenic Position	
0.001637	388		
0.001597	244		
0.00157	422		
0.001446	92		
0.001253	199		
0.0012	490		
0.001164	355		
0.001087	374		
0.000993	365		
0.000791	51		
0.000788	242		
0.000774	506		
0.000764	175		
0.000594	202		
0.000557	167		
-0.00046	345		
-0.000563	281		
-0.000848	398		
-0.000994	273		
-0.001608	534		

Table 4. Ranking of amino-acid pairs of the receptor-binding domain by the cumulative ranking. The last column on the right indicates whether the residue appears in the highly ranked set of the single amino acid analysis of the RBD of the HA sequence (Table 2). Most of the known residues (specificity determinants and antigenic positions) and novel detected from the single-residue RBD analysis are highly ranked.

Rank	Residue Number	Known Position?	Included in highly ranked set of the single-residue analysis of RBD
1	158		+
2	204		+
3	227	Specificity Determinant	+
4	138	Specificity Determinant	+
5	133	Specificity Determinant + Antigenic residue	+
6	118		+
7	186	Specificity Determinant	+
8	145	Antigenic Residue	+
9	135		+
10	187		+
11	129	Antigenic Residue	+
12	178		+
13	144	Antigenic Residue	+
14	226		-
15	156	Antigenic Residue	-
16	143	Specificity Determinant + Antigenic residue	-
17	114		-
18	235		+
19	211		-
20	166		+
21	119		-
22	141	Antigenic Residue	+
23	127		+
24	216		+
25	131	Antigenic Residue	-
26	182		-
27	218		-
28	222		-
29	159		-
30	125		-
31	267		+
32	179		-
33	217		-
34	190		-

35	219		-
36	193		+
37	140	Antigenic Residue	-
38	142	Antigenic Residue	-
39	124		-
40	128		+
41	188		-
42	214		-
43	163		-
44	192	Specificity Determinant	+
45	231		-
46	123		-
47	160		+
48	239		-
49	152		-
50	137	Specificity Determinant	-
51	130		-
52	122		-
53	197		-
54	185		-
55	183		-
56	252		-
57	261		-
58	213		-
59	177		-
60	208		-
61	223		-
62	256		-
63	230		-
64	116		-
65	117		-
66	165		-
67	236		-
68	155		-
69	196	Specificity Determinant	+
70	260		-
71	262		-
72	154		-
73	157	Antigenic Residue	-
74	120		-

References

1. Zambon MC (2001) The pathogenesis of influenza in humans. Rev Med Virol 11: 227-241.
2. Fouchier RA, Munster V, Wallensten A, Bestebroer TM, Herfst S, et al. (2005) Characterization of a novel influenza A virus hemagglutinin subtype (H16) obtained from black-headed gulls. J Virol 79: 2814-2822.
3. Neumann G, Noda T, Kawaoka Y (2009) Emergence and pandemic potential of swine-origin H1N1 influenza virus. Nature 459: 931-939.
4. Dawood FS, Jain S, Finelli L, Shaw MW, Lindstrom S, et al. (2009) Emergence of a Novel Swine-Origin Influenza A (H1N1) Virus in Humans Novel Swine-Origin Influenza A (H1N1) Virus Investigation Team. New England Journal of Medicine 360: 2605-2615.
5. Johnson NPAS, Mueller J (2002) Updating the accounts: global mortality of the 1918-1920 "Spanish" influenza pandemic. Bulletin of the History of Medicine 76: 105-115.
6. Taubenberger JK, Morens DM (2009) Pandemic influenza--including a risk assessment of H5N1. Rev Sci Tech 28: 187-202.
7. Yang Y, Sugimoto JD, Halloran ME, Basta NE, Chao DL, et al. (2009) The Transmissibility and Control of Pandemic Influenza A (H1N1) Virus. Science.
8. Maines TR, Jayaraman A, Belser JA, Wadford DA, Pappas C, et al. (2009) Transmission and pathogenesis of swine-origin 2009 A(H1N1) influenza viruses in ferrets and mice. Science 325: 484-487.
9. Childs RA, Palma AS, Wharton S, Matrosovich T, Liu Y, et al. (2009) Receptor-binding specificity of pandemic influenza A (H1N1) 2009 virus determined by carbohydrate microarray. Nat Biotechnol 27: 797-799.
10. Olsen B, Munster VJ, Wallensten A, Waldenstrom J, Osterhaus AD, et al. (2006) Global patterns of influenza a virus in wild birds. Science 312: 384-388.
11. CDC Key Facts About Avian Influenza (Bird Flu) and Avian Influenza A (H5N1) Virus.
12. WHO (2009) Cumulative number of confirmed human cases of avian influenza A/(H5N1) reported to WHO. http://wwwwhoint/csr/disease/avian_influenza/country/cases_table_2009_09_24/en/indexhtml.
13. Stevens J, Blixt O, Tumpey TM, Taubenberger JK, Paulson JC, et al. (2006) Structure and receptor specificity of the hemagglutinin from an H5N1 influenza virus. Science 312: 404-410.
14. Ha Y, Stevens DJ, Skehel JJ, Wiley DC (2001) X-ray structures of H5 avian and H9 swine influenza virus hemagglutinins bound to avian and human receptor analogs. Proc Natl Acad Sci U S A 98: 11181-11186.
15. Baigent SJ, McCauley JW (2003) Influenza type A in humans, mammals and birds: determinants of virus virulence, host-range and interspecies transmission. Bioessays 25: 657-671.
16. Matrosovich M, Tuzikov A, Bovin N, Gambaryan A, Klimov A, et al. (2000) Early alterations of the receptor-binding properties of H1, H2, and H3 avian influenza virus hemagglutinins after their introduction into mammals. J Virol 74: 8502-8512.

17. Vines A, Wells K, Matrosovich M, Castrucci MR, Ito T, et al. (1998) The role of influenza a virus hemagglutinin residues 226 and 228 in receptor specificity and host range restriction. Journal of Virology 72: 7626-7631.
18. Yamada S, Suzuki Y, Suzuki T, Le MQ, Nidom CA, et al. (2006) Haemagglutinin mutations responsible for the binding of H5N1 influenza A viruses to human-type receptors. Nature 444: 378-382.
19. Auewarakul P, Suptawiwat O, Kongchanagul A, Sangma C, Suzuki Y, et al. (2007) An avian influenza H5N1 virus that binds to a human-type receptor. J Virol 81: 9950-9955.
20. Yang ZY, Wei CJ, Kong WP, Wu L, Xu L, et al. (2007) Immunization by avian H5 influenza hemagglutinin mutants with altered receptor binding specificity. Science 317: 825-828.
21. Gambaryan A, Tuzikov A, Pazynina G, Bovin N, Balish A, et al. (2006) Evolution of the receptor binding phenotype of influenza A (H5) viruses. Virology 344: 432-438.
22. Treanor J (2004) Influenza vaccine--outmaneuvering antigenic shift and drift. N Engl J Med 350: 218-220.
23. Klenerman P, Zinkernagel RM (1998) Original antigenic sin impairs cytotoxic T lymphocyte responses to viruses bearing variant epitopes. Nature 394: 482-485.
24. Kaverin NV, Rudneva IA, Ilyushina NA, Varich NL, Lipatov AS, et al. (2002) Structure of antigenic sites on the haemagglutinin molecule of H5 avian influenza virus and phenotypic variation of escape mutants. J Gen Virol 83: 2497-2505.
25. Bao YM, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, et al. (2008) The influenza virus resource at the national center for biotechnology information. Journal of Virology 82: 596-601.
26. Allen JE, Gardner SN, Vitalis EA, Slezak TR (2009) Conserved amino acid markers from past influenza pandemic strains. BMC Microbiol 9: 77.
27. Wu LC, Horng JT, Huang HD, Chen WL (2008) Identifying discriminative amino acids within the hemagglutinin of human influenza A H5N1 virus using a decision tree. IEEE Trans Inf Technol Biomed 12: 689-695.
28. Freund Y, Mason L (1999) The Alternating Decision Tree Algorithm. Proceedings of the 16th International Conference on Machine Learning: 124-133.
29. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32: 1792-1797.
30. Ha Y, Stevens DJ, Skehel JJ, Wiley DC (2002) H5 avian and H9 swine influenza virus haemagglutinin structures: possible origin of influenza subtypes. EMBO J 21: 865-875.
31. Schapire RE (2002) The Boosting Approach to Machine Learning An Overview. MSRI Workshop on NonLinear Estimation and Classification.
32. Freund Y, Schapire RE (1999) A Short Introduction to Boosting. Journal of Japanese Society for Artificial Intelligence 14: 771–780.
33. Liu K, Lin J, Zhou X, Wong S (2005) Boosting alternating decision trees modeling of disease trait information. BMC Genet 6 Suppl 1: S132.
34. Alterovitz R, Arvey A, Sankararaman S, Dallett C, Freund Y, et al. (2009) ResBoost: characterizing and predicting catalytic residues in enzymes. BMC Bioinformatics 10: 197.

35. Kohavi R (1995) A Study of Cross-Validation and Bootsrap for Accuracy Estimation and Model Selection. International Joint Conference on Artificial Intelligence (IJCAI).
36. Creamer G, Freund Y, Moore M (2005) Using Adaboost for Equity Investment Scorecards. papersssrncom.
37. Suarez DL, Perdue ML, Cox N, Rowe T, Bender C, et al. (1998) Comparisons of highly virulent H5N1 influenza A viruses isolated from humans and chickens from Hong Kong. Journal of Virology 72: 6678-6688.
38. Baigent SJ, McCauley JW (2003) Influenza type A in humans, mammals and birds: determinants of virus virulence, host-range and interspecies transmission. Bioessays 25: 657-671.
39. Stevens J, Blixt O, Chen LM, Donis RO, Paulson JC, et al. (2008) Recent avian H5N1 viruses exhibit increased propensity for acquiring human receptor specificity. J Mol Biol 381: 1382-1394.
40. Kaverin NV, Rudneva IA, Govorkova EA, Timofeeva TA, Shilov AA, et al. (2007) Epitope mapping of the hemagglutinin molecule of a highly pathogenic H5N1 influenza virus by using monoclonal antibodies. J Virol 81: 12911-12917.
41. Philpott M, Hioe C, Sheerar M, Hinshaw VS (1990) Hemagglutinin mutations related to attenuation and altered cell tropism of a virulent avian influenza A virus. J Virol 64: 2941-2947.
42. Branden C, Tooze J (1999) Introduction to protein structure. New York: Garland Publishing.
43. Govorkova EA, Rehg JE, Krauss S, Yen HL, Guan Y, et al. (2005) Lethality to ferrets of H5N1 influenza viruses isolated from humans and poultry in 2004. Journal of Virology 79: 2191-2198.
44. Recker M, Pybus OG, Nee S, Gupta S (2007) The generation of influenza outbreaks by a network of host immune responses against a limited set of antigenic types. Proc Natl Acad Sci USA 104: 7711-7716.

תקציר

למרות שהתפשטותו בקרב אוכלוסיית העולם עודנה מוגבלת, זן שפעת העופות הפתוגני בבני-אדם H5N1, מהווה איום ממשי על המין האנושי. חלבון ההמגלוטינין (hemagglutinin) של השפעת מסוג A הוא האנטיגן הראשי על המעטפת הנגיפית, המתווך את היקשרות הרצפטורים על פני התא המארח ומאפשר למעשה את הכניסה של הנגיף אל תוך התא. שינוי בהכרה של הרצפטור המאפיין תאים של עופות להכרה המאפיינת תאים הומניים על ידי ההמגלוטינין, הוא ככל הנראה אחד התנאים הראשוניים בכדי שהנגיף יוכל להשתכפל ביעילות בתאים הומניים ולגרום למגיפה. באמצעות גישה חישובית, העושה שימוש באלגוריתמים של למידה מונחית ורצפים של ה- H5N1 ממאגר המידע של NCBI, הצלחנו לזהות את כל העמדות הידועות בספרות המשפיעות על הכרה ספציפית של הרצפטור. בנוסף, זיהינו עמדות המתייגות אתרים אנטיגניים של ה- H5N1 כאתרים המבדילים בין הרצפטורים השונים, ממצא שעשוי להצביע על מכניזם הקשור למערכת החיסונית עבור הספציפיות של הנגיף. האנליזה שלנו זיהתה גם עמדות שתפקידן טרם נודע, אשר עשויות להיות אלו הקובעות ספציפיות לרצפטור. ייתכן שגילויים אלו יסייעו להבנה טובה יותר של המחסום בין המינים (עופות ובני אדם) ואף ישמשו בתכנון מעכבי נגיפים. האנליזה החישובית המוצגת כאן הינה גנרית וניתנת ליישום על חלבונים נוספים של וירוס השפעת כמו גם על וירוסים אחרים.

אוניברסיטת תל-אביב

הפקולטה למדעי החיים ע"ש ג'ורג' ס. וייז

המדרשה לתארים מתקדמים

זיהוי חישובי של חומצות אמינו המהוות את המחסום למעבר מעופות לבני אדם של וירוס השפעת מזן H5N1

חיבור זה הוגש כעבודת גמר לקראת התואר "מוסמך אוניברסיטה"

במסלול ביוכימיה באוניברסיטת תל-אביב

על-ידי

דפנה מרוז

העבודה הוכנה במחלקה לביוכימיה של אוניברסיטת תל-אביב

בהנחיית

פרופ' ניר בן-טל

חתימת המנחה:

תאריך: דצמבר 2009

אוניברסיטת תל-אביב

הפקולטה למדעי החיים ע"ש ג'ורג' ס. וייז

המדרשה לתארים מתקדמים

זיהוי חישובי של חומצות אמינו המהוות את המחסום למעבר מעופות לבני אדם של וירוס השפעת מזן H5N1

חיבור זה הוגש כעבודת גמר לקראת התואר "מוסמך אוניברסיטה"

במסלול ביוכימיה באוניברסיטת תל-אביב

על-ידי

דפנה מרוז

העבודה הוכנה במחלקה לביוכימיה של אוניברסיטת תל-אביב

בהנחיית

פרופ' ניר בן-טל

דצמבר 2009