Spring 5-1-2014

Numerical assessment of sequence conservation in flu-virus hemagglutinin

Scott S. Norton
University of Connecticut - Storrs, Scott.s.norton@hotmail.com

UNIVERSITY OF CONNECTICUT HEALTH CENTER

Numerical assessment of sequence conservation in flu-virus hemagglutinin

Honors Thesis

Scott Norton

5/1/2014

In collaboration with Dr. Gregory Huber, Ph.D, Center for Cell Analysis and Modeling (CCAM), UCHC, and Drs. Nikolay Dokholyan, Ph.D, and David Shirvanyants, Ph.D, Department of Computational and Experimental Biophysics, University of North Carolina. Funded in part by the CCAM Summer Undergraduate Research Program, UCHC.

Contents

Abstract

The flu virus was investigated to find a common recognition domain to which an antibody against human-infected viruses can bind. If such a target site is structurally and electrostaticly conserved or invariant, only a single antibody would be required to attack the virus in all cases. The sequence of one of the viral surface proteins contains 24 amino acids that do not vary through mutation. However, these amino acids are neither contiguous in sequence or in space, and the ones that are associated with each other are not readily accessible to an antibody. They do provide a first impression of which regions of the surface have the potential to serve as a common recognition site, and a broader search requiring more computational power may reveal a region of the protein surface that is structurally stable across all strains and is available for binding to a soluble antibody.

Background information

Definitions of biological terms

One of nature's most incredible feats is its ability to encode every piece of biological information, which can be immeasureably complex, as a linear sequence of a small number of molecular units. DNA, for example, are simply sequences of four nucleotides. For proteins, it's the twenty amino acids. In bioinformatics, a protein is defined at its most fundamental by its **amino acid sequence** or **protein sequence**. An amino acid monomer in a sequence is also called a **residue**. Every amino acid has four main components: an amino group, a carboxyl group, a side chain, and the α carbon which holds the monomer together. The side chain determines the identity of an amino acid. Amino acids are typically joined by a peptide bond between the carboxyl group of one and the amino group of the next. Multiple amino acids are joined in this fashion to form a polymer known as a polypeptide.

This paper will represent amino acids in one of three ways at various points: by name, as a three letter abbreviation, and as a one letter symbol. Table A-1 in the appendix can serve as a guide to navigate between the three conventions. Like a single strand of DNA, whose sequence is represented from "5' to 3'", amino acid sequences are conventionally read and represented in a single direction, from the free amino end (amino- or N-terminal) to the free carboxyl end (carboxyl-or C-terminal) along the peptide backbone. The parallels between proteins and DNA don't end there – DNA is well-accepted as the "code" for protein synthesis. Tri-nucleotide words known as **codons** translate to specific amino acids via the genetic code, which is reproduced in Table A-2.

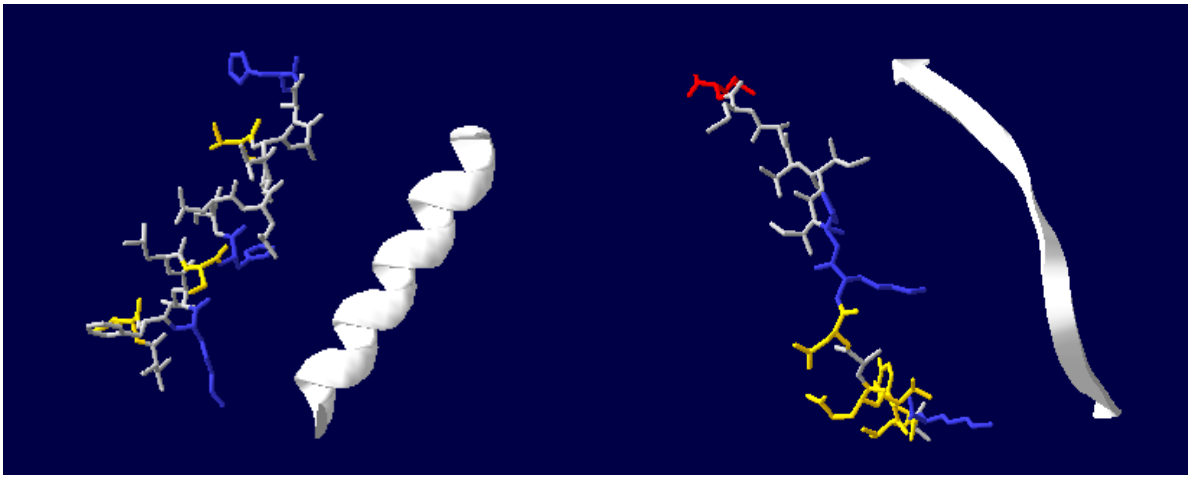

Figure 1. Examples of an α helix and a β strand, in both bond-lines and ribbon displays.

Beyond the **primary structure** level conferred by sequence alone, proteins adopt higher orders of structure denoted **secondary**, **tertiary**, and **quaternary**. Secondary structure involves non-covalent interactions between the backbone carbons, nitrogens, and oxygens to form local structures such as α helices and β strands. Figure 1 shows examples of these structures as they would appear in a Protein Databank (PDB) file viewer. The tertiary structure yields the fully-folded protein through distal interactions between residues' side chains. Finally, the quaternary structure involves interactions between the protein and other types of molecules and atoms, including sugars, metal ions, lipids, and

other proteins. Hemoglobin, for example, consists of two pairs each of α and β subunits, with each subunit being a separate polypeptide chain.

Amino-acid sequence is believed to be highly determinative of protein structure and function, and several experiments have shown that alterations in this sequence can have serious consequences on both structure and function. A famous example is evident in sickle-cell anemia, where the glutamate in the 6th position of the β-globin subunit of the blood oxygen-carrier hemoglobin, is replaced with a valine (represented E6V in the literature) [1] in a **substitution** event. This mutation causes the rest of the protein to shift by up to 12 Å beyond the 6th residue, as shown in PDB structures 4N7O [2] and 2HBS [3]. This dramatic change causes hemoglobin to aggregate into fibers which distort the shape of the red blood cell. However, not all mutations are bad; many substitutions are benign and some beneficial. In viral membrane proteins, changes to the sequence of the extracellular regions can protect the virus from detection by the host's immune system.

The fact that proteins can be represented as linear sequences facilitates their computational study. Two or more sequences can, for example, be aligned to map similarities to each other and highlight the differences between them. The alignment process is aided by a scoring or substitution matrix, a representation of the frequency of each pairwise substitution as expected in nature, scoring less frequent substitutions lower and more frequent substitutions higher. Two sequences with a high degree of similarity (a good alignment under the scoring matrix of choice) are believed to have diverged more recently than two sequences with a lower degree of similarity (more mismatches and gaps in the alignment), and from this we can reconstruct the relative evolutionary history of a large number of proteins. Phylogenetics, the science of reconstructing evolutionary relationships between biological entitites, uses sequence alignment as a tool for mapping molecular evolution and has made significant contributions to our understanding of evolutionary patterns. Molecular phylogenetics has shown, for

example, that birds are more closely related to reptiles than to mammals, suggesting that birds evolved warm-bloodedness in a separate yet convergent manner to mammals. Sequence alignment can also be used to measure the degree of conservation of a particular region in the sequence of a single protein by comparing multiple instances of that protein or those similar to it. This paper will discuss one technique for measuring sequence conservation, using flu-virus hemagglutinin as the subject.

Flu-virus surface proteins and their function in the infective process

Pandemic flu is a leading cause of death worldwide. The 1918 Spanish Flu outbreak is estimated to have killed 3-6% of the global human population [4], and the 2009 Swine Flu pandemic is confirmed to have taken over 18,000 lives [5]. In addition, seasonal epidemics of the flu virus infect 3-5 million people and kill 250-500 thousand each year [6]. Annual vaccines target the dominant strain of the flu virus for that season and grant immunity against that specific strain. However, the flu virus is able to evade the human immune system by mutating the structure of its surface antigens to decrease their affinity to blood-borne antibodies. Flu strains for which there are no host antibodies can infect the organism virtually unopposed. The virus itself has two distinct kinds of surface protein, hemagglutinin and neuraminidase, which interact with host cell membrane factors and to which antibodies can bind.

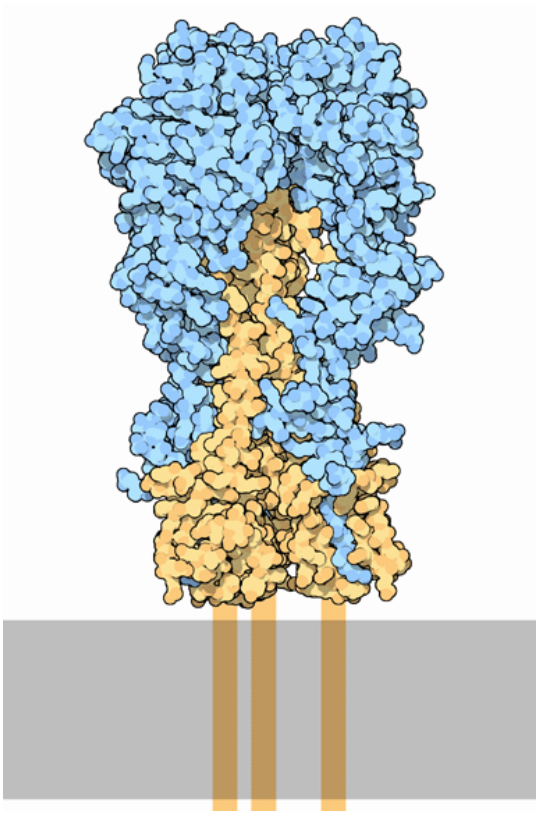

Figure 2. Flu-virus hemagglutinin, rendered from model 1RUZ. Globular head is blue, fusion peptide is yellow. (http://www.rcsb.org/pdb/education_discussion/molecule_of_the_month/images/76_1ruz.gif)

Though both proteins are highly variable, each has several subtypes with distinct sequence and structural moieties. These subtypes are used to classify individual strains of influenza. For example, H5N2 contains hemagglutinin subtype 5 (H5) and

neuraminidase subtype 2 (N2). The structural differences between hemagglutinin subtypes are fairly subtle, but even a slight change can affect specificity of interactions with receptors. Figure 3 highlights a

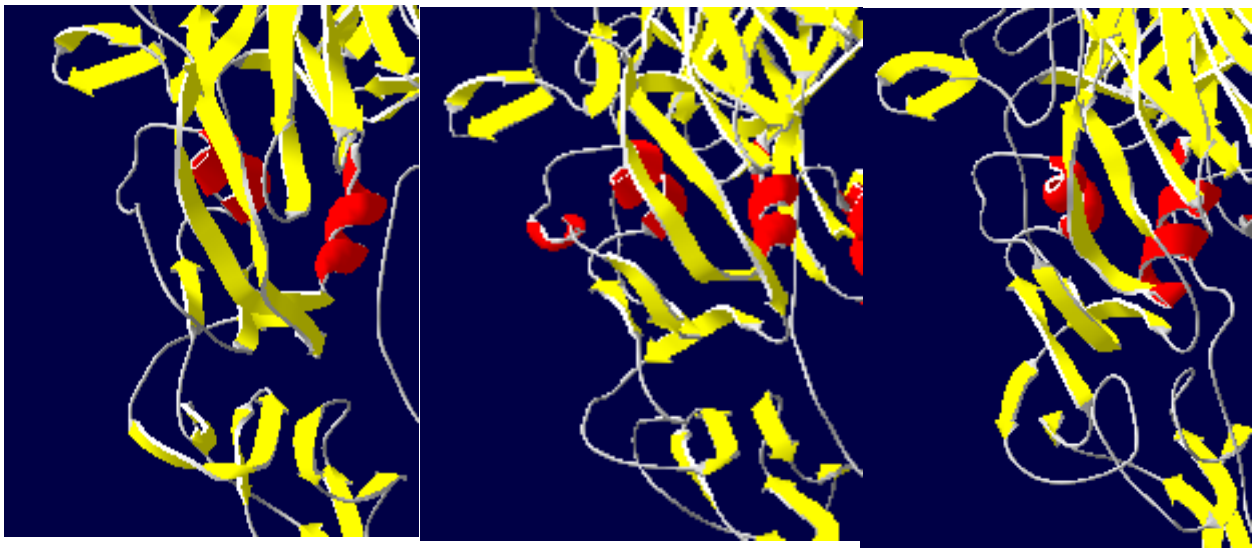

Figure 3. The same region on three structures. Left: 3S11, an H5 structure; middle: 1RUZ, an H1 structure; right: 1HGD, an H3 structure

region where the structure varies quite visibly between subtypes. Notice how the morphology of the left-hand loop varies between subtypes, with only the H1 structure containing a helix in this region. A lot of variation also exists in the loop and strands near the bottom left of the depicted region. These differences might play a role in substrate specificity, since each subtype recognizes different glycoprotein moieties (biochemical groups) on the host cell.

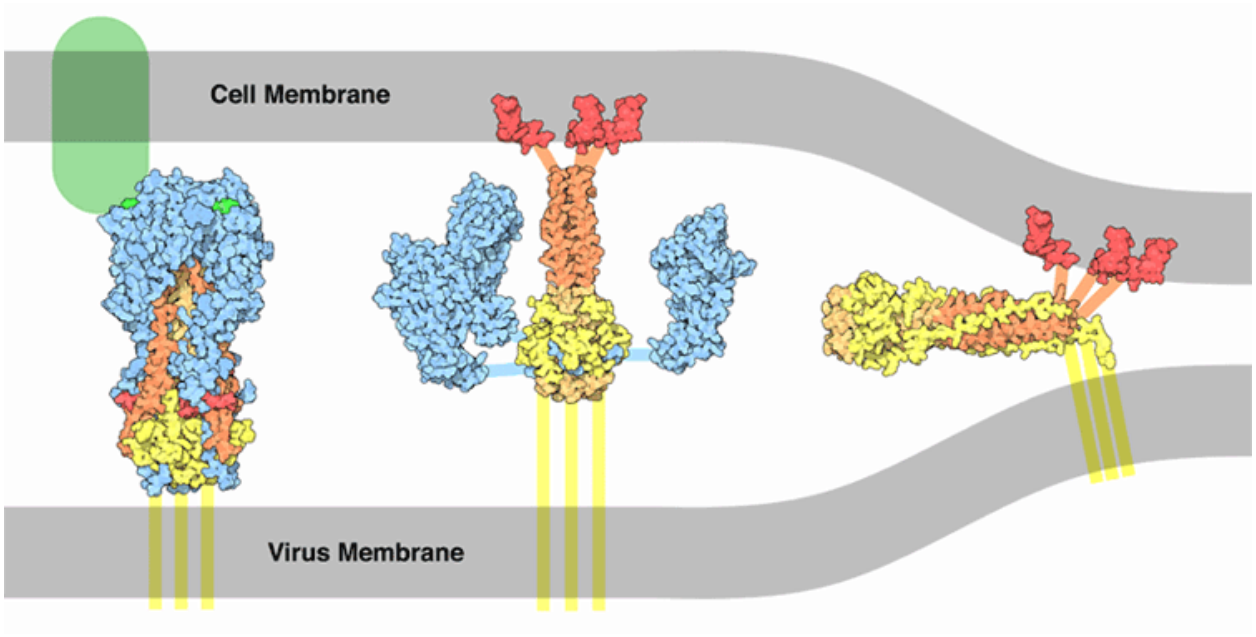

Figure 5. Schematic of hemagglutinin-mediated membrane fusion. PDB entries 1HGE, 1HTM, 1IBN, 2VIR, and 1QU1. (http://www.rcsb.org/pdb/education_discussion/molecule_of_the_month/images/76_HA-action.gif)

Hemagglutinin and neuraminidase are vital for the viral replication cycle in mammalian influenza. Flu-virus hemagglutinin is composed of three identical copies each of two subunits which make up the globular head and fusion peptide domains. Lipid anchors secure the protein to the exterior surface of the viral coat, with the globular head region oriented outwards so that it can interact with sialic-acid residues on cell-membrane receptors, triggering phagocytosis of the virus particle. Since the pattern of glycosylation on these receptors depends on the organism, the hemagglutinin subtype confers host specificity. Human-host hemagglutinin interacts with α2,6-linked sialic acid

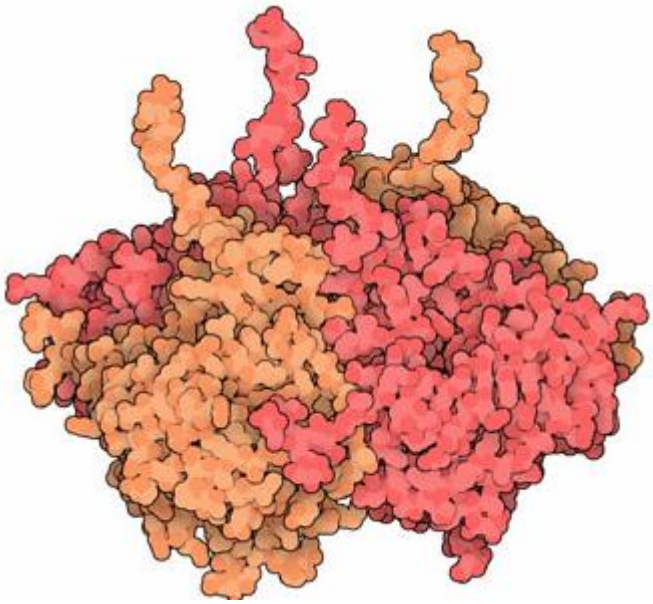

Figure 4. Neuraminidase, structure 1NN2. (http://www.rcsb.org/pdb/education_discussion/molecule_of_the_month/images/neuraminidases.jpg)

residues, whereas avian-flu hemagglutinin prefers α2,3-linked sialic acid residues. Once inside the phagosome, the cell begins constructing the mature lysosome, a process which lowers the pH of the virion's environment. This pH drop activates the fusion peptide, which fuses the viral envelope with the lysosomal membrane, allowing the viral proteins and RNA to enter the cytoplasm and produce new virus particles [7]. Newly-synthesized viral proteins pass through the cell's normal processing pathways, with surface proteins traversing the endoplasmic reticulum (ER) and Golgi before being transported in vesicles to the membrane. The surface proteins concentrate in one region of the cell membrane and recruit the capsid proteins, causing the new virus to begin budding. The bud is freed by the action of neuraminidase, which again cleaves sialic acid residues on the host cell surface to promote membrane cleavage [8].

Hemagglutinin has several domains that could be antibody targets

As the protein most directly involved in initiating the infection, hemagglutinin has been a prime target for immunization. Current vaccination techniques involve inoculating the host with an inactivated virus to raise antibodies against the virus without expressing symptoms. However, a possible future technique may be to synthesize an antibody *in vitro*, in a process that does not require an intermediate host such as a chicken. A study by Sahini *et al* identified nine regions on the hemagglutinin virus that are potential sites for antibody recognition. They performed a multiple alignment of all complete amino-acid sequences of each well-represented subtype, then performed a seeded alignment on the consensus sequence profiles using H2 as their seed. They then scored the degree of conservation at each position using a correlation-based approach, resulting in a value between 0 and 1. Regions of the globular head region that had 6 or more residues with conservation scores of 0.9 or greater were considered highly conserved. These regions were subsequently tested for solvent-accessible surface area and hydrophobicity, and ranked according to all three parameters. One of the conserved sequence regions, site 7 in Table 2 of Sahini *et al*, contains part of a known conserved epitope in H5, which is replicated in

the ELISA peptide used to detect H5 antibodies [9]. However, these structural analyses did not demonstrate this region to be a suitable target for an immune response (ranked #7 in Table 3) compared with, for instance, site 5 (ranked #1) [10].

Because hemagglutinin and neuraminidase both pass through the normal cell-processing pathways, they are both glycosylated shortly after synthesis. Sugar residues are added onto the side chains of particular amino acids, and their linkages are classified by which type of atom they bind to, usually N-linked or O-linked. Hemagglutinin and neuraminidase are the only flu-virus proteins known to be glycosylated, but to date no O-linked glycosylation has been observed [11]. This means no serine, threonine, or tyrosine residue (side chains with a terminal –OH) on either protein has ever been observed to be glycosylated. Therefore, any interactions at sites containing these residues will be with the exposed hydroxyl group, which is a potent site for hydrogen bonding.

Computational algorithms for the analysis of protein sequences

In this investigation, a large set of amino-acid sequences of hemagglutinin were aligned and then scanned for invariant positions. The algorithm used to align three or more sequences relies heavily on the global pairwise alignments of the sequences. The Needleman-Wunsch global pairwise-alignment algorithm inserts gaps into the two sequences such that the two sequences become the same length, and the alignment score is maximum. This score is determined for each position in the alignment using a predefined substitution matrix such as BLOSUM50 for amino acids [12], or NUC44 for nucleotides. If the two sequences match at a given position, the score is increased; a mismatch or a gap is penalized according to the appropriate matrix entry, with less common and nonsynonymous substitutions facing a harsher penalty. The multiple alignment is performed by first constructing a neighbor-joining guide tree using pairwise alignment scores as the distance metric. The neighbor-joining algorithm proceeds by uniformly shrinking the pairwise distances between the taxa and joining groups as they collapse [13]. Once the tree has been computed, the sequences are aligned progressively following the tree, with each

internal node representing the alignment of all its children, so that the root of the tree represents the final alignment of all the sequences.

To illustrate sequence alignment, consider the set of sequences {MKTTRSVGE, MLTPHSVGHE, MIFTQSAGD, MKTRSGG}. As you can see, these sequences vary greatly in length and composition. The Needleman-Wunsch algorithm would align each pair of these sequences with an emphasis on keeping similar groups together. The alignment between the second and fourth sequences in the given set, which happen to be the longest and shortest, respectively, produces this alignment, with a score of 5. Repeating the process for the remaining sequence pairs produces a set of ${}_4C_2 = 6$ alignments and scores. The pairwise alignment score is a one-dimensional measure of similarity between two sequences. The first sequence (MKTTRSVGE) is the most similar to each of the three others, with alignment scores of 27, 23, and 25, respectively. The second and fourth sequences are the most dissimilar at 5, followed by the second and third sequences at 12, and the third and fourth sequences at 13.

```
MLTPHSVGHE
|  |  :|  |
MKT-RS-GG-
```

To multiply align these sequences, we first build a guide tree using a distance-based algorithm. To do this, we need a distance matrix which can be computed from the alignment scores: For sequences i and j,

$$D_{ij} = \left(1 - \frac{score_{ij}}{score_{ii}}\right)\left(1 - \frac{score_{ij}}{score_{jj}}\right).$$

The distance matrix for the sample sequence set is

$$D = \begin{bmatrix} 0 & 0.310 & 0.346 & 0.245 \\ 0.310 & 0 & 0.659 & 0.829 \\ 0.346 & 0.659 & 0 & 0.557 \\ 0.245 & 0.829 & 0.557 & 0 \end{bmatrix}.$$

Note how smaller values correspond to more similar sequences and vice versa. This means the neighbor joining algorithm will place similar sequences closer together in the tree. New nodes are iteratively inserted into the tree by first joining the least distant pair of nodes, denoted i and j (in this case, sequences 1 and 4 for the first iteration) and recomputing the distances between this new node (n) and each other node (k) as

$$D_{nk} = \frac{1}{2}(D_{ik} + D_{jk} - D_{in} - D_{jn}).$$

In the first iteration on this sequence set, joining sequences 1 and 4 reduces the distance matrix to

$$D^1 = \begin{bmatrix} 0 & 0.659 & 0.447 \\ 0.659 & 0 & 0.329 \\ 0.447 & 0.329 & 0 \end{bmatrix},$$

where the rows and columns are ordered 2, 3, (1,4). The second iteration inserts a node between sequence 3 and the node joining (1,4), which reduces the distance matrix to

$$D^2 = \begin{bmatrix} 0 & 0.388 \\ 0.388 & 0 \end{bmatrix}.$$

The remaining distances in the graph can be computed using the fact that the total path length between two nodes is equal to their pairwise distance in the original distance matrix. There may be occasions where a subpath may end up being longer than the pairwise distance, and these cases typically correspond to degeneracy in the graph generated from the original distance matrix. This phenomenon arises when two sequences are distant from each other but

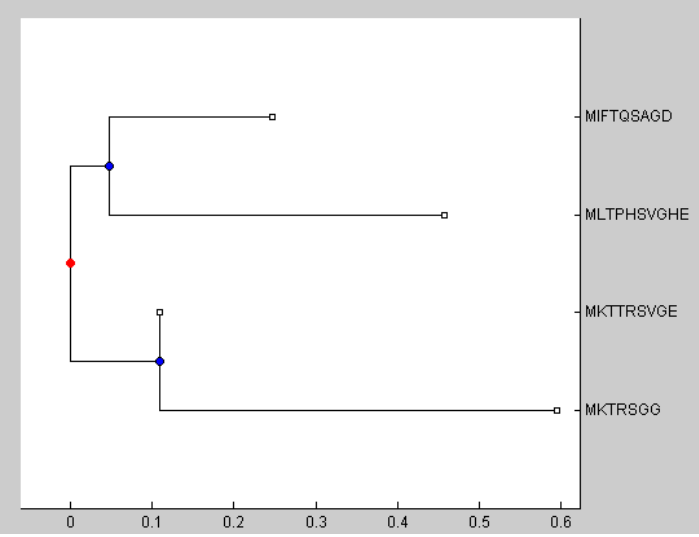

Figure 6. Neighbor-joining tree from sample sequences, generated in MATLAB.

each is close to a third sequence. The solution is to compute tree distances using only non-degenerate groups. The resulting tree, displayed in Figure 6 with a root between the two internal nodes, has topology ((2,3),(1,4)).

Finally, a multiple sequence alignment can be computed iteratively from the pairwise alignments and the rooted tree, starting with the most closely joined pair and moving along the graph until all sequences have been aligned. Gaps are inserted into sequences at each iterative step in order to preserve the alignment of each subtree; gap placement is guided by the sequence in the subtree that is closest to the internal node. The final alignment of the sample sequence set is displayed to the right.

```
M-KTTRSVG-E
ML-TPHSVGHE
MIFT-QSAG-D
M-KT-RSGG--
```

The consensus sequence is determined at each position of the alignment by measuring the character frequency profile P to construct a vector called the "consensus value" $X = MP$, where M is the substitution matrix used in the alignment. By definition, M is a symmetric matrix – that is, $M_{ij} = M_{ji}$ for all i, j – such that the diagonal elements are strictly maximal for their respective row and column ($M_{ii} > M_{ij}$ when $j \neq i$). The frequency profile P must also satisfy two constraints: $0 \leq P_i \leq 1$ for all i, and $\sum P_i = 1$. The character in the alphabet (in this case, amino acids with a gap character) corresponding to the greatest entry in the consensus value, is selected for the consensus sequence. However, X is also useful for measuring how well conserved the position is. Let Y be a vector such that $Y_i = |X - M_i|$, where M_i is the ith column of M, and define the consensus score as $S = Y \cdot P$. We shall analyze some of the properties of S as it relates to P.

Suppose that $P_k = 1$ and $P_l = 0$ for all $l \neq k$. This corresponds to every sequence having the kth character of the alphabet at that position in the alignment; the position is fully conserved. Since $X = MP$, it follows that $X = M_k$, and so $Y_k = |M_k - M_k| = 0$ while $Y_l = |M_k - M_l| > 0$ for all $l \neq k$. Thus

$S = Y \cdot P = 0$ because $Y_k = 0$ and $P_l = 0$ for all $l \neq k$. Therefore, if a position in an alignment is fully conserved, $S = 0$ at that position. However, a more powerful conclusion can be drawn about S:

> ***Theorem 1:*** *A position in a multiple-sequence alignment is fully conserved if and only if the consensus score S is exactly 0.*

Proof. Suppose the position was not fully conserved; that is, there are at least two entries in P that are nonzero. Assume $S = 0$. From here on out, we reduce all objects of interest to only those entries of vectors and columns/rows of matrices corresponding to those nonzero elements of P, since these are the only elements that will contribute to the final dot product (zero elements in either vector in a dot product contribute zero value to the dot product). Denote any object so reduced by a $'$. Then for each i, $0 \leq Y_i' \leq Y_i$ with equality only at 0. Thus if each $Y_k' = 0$, it follows that $S' = 0$, and $S = 0$. In order for $Y_k' = 0$, we must have $M'P' - M_k' = 0$ for each k. Thus all the columns of M' are the same, all the rows of M' are the same by the inherited symmetry, and we infer that all the elements of M' are equal. M', by the way, is a substitution matrix since it inherits the definitive properties from its parent M. Therefore, M' cannot have all of its elements equal by definition. Thus either $S \neq 0$, or the position is fully conserved. ■

To illustrate this consensus score in action, consider our original sequence alignment. The first column contains only an M in it, so P would be 1 at the entry corresponding to methionine, and 0 elsewhere. Thus the consensus value X would be exactly the column of the BLOSUM50 matrix, M, corresponding to methionine, and the maximum value there corresponds to the alignment of methionine with itself: 7. Thus the character chosen for the first position in the consensus sequence is M. Furthermore, the theorem states that $|X - M_i| = 0$ when M_i corresponds to methionine, so $S_1 = 0$. In the second position, we have an L, an I, and two gaps, so P would be 0.25 at leucine and isoleucine, 0.5 at the gap,

and 0 elsewhere. The contribution of M to X would then be one-fourth the leucine column, one-fourth the isoleucine column, and one-half the gap column, which gives a total of

$$\left[-\frac{13}{4}, -\frac{17}{4}, -\frac{17}{4}, -\frac{9}{2}, -\frac{7}{2}, -\frac{15}{4}, -\frac{17}{4}, -\frac{9}{2}, -\frac{17}{4}, -\frac{3}{4}, -\frac{3}{4}, -4, -\frac{5}{4}, -\frac{9}{4}, -\frac{17}{4}, -4, -3, -\frac{15}{4}, -3, -\frac{5}{4}, -\frac{9}{2}, -\frac{3}{2}\right].$$

We end up with a tie at the leucine and isoleucine positions, at $-\frac{3}{4}$ each. A choice must be made for the consensus character at that position, and at this point the choice can be arbitrary between the two winners. The computation of S can easily take up a page when dealing with a 21-by-21 matrix such as BLOSUM50, so the calculation steps shall be omitted. The result of this computation is $S_2 = 7.6389$. Continuing in this fashion produces the full consensus sequence as `M-KT-RSVG-E` and the full set of consensus scores

$$S = [0.000, 7.639, 9.629, 0.000, 7.751, 6.527, 0.000, 8.321, 0.000, 5.876, 6.606].$$

Note that the zeros in S correspond exactly to the invariant methionine, threonine, and glycine residues in the original sequence set.

So the consensus score indicates precisely which residues in the hemagglutinin alignment are invariant. These residues were mapped onto each of three crystal structures of hemagglutinin obtained from the PDB [14]. From there, a search was conducted on the space of protein structures for a complementary surface to a single region on three hemagglutinin subtypes. This search was performed using a fingerprint comparison algorithm developed by the Dokholyan group at the University of North Carolina School of Medicine. This algorithm uses Zernike invariants, also called Zernike moments or Zernike

$$\text{Aberration} = 3 - 8\rho^2 + 6\rho^4 + \rho\left(5 - 60\rho^2 + 210\rho^4 - 280\rho^6 + 126\rho^8\right)\text{Cos}[\theta] + \rho\left(5 - 60\rho^2 + 210\rho^4 - 280\rho^6 + 126\rho^8\right)\text{Sin}[\theta] + \rho^2\left(-10 + 60\rho^2 - 105\rho^4 + 56\rho^6\right)\text{Sin}[2\theta]$$

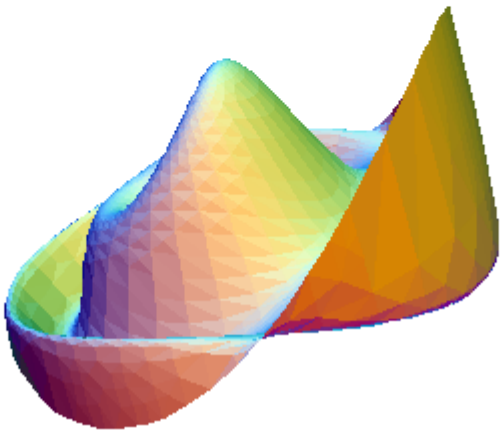

Figure 7. Zernike object with $c_{0,0}$=1, $c_{2,0}$=-1, $c_{3,0}$=1, $c_{8,-2}$=1, $c_{9,-1}$=1, and $c_{9,1}$=1, where $c_{i,j}$ is the coefficient of the (i,j) Zernike component [22].

polynomials, to encode the surface of a molecule's electron density map. Zernike moments are akin to Fourier transforms in that they allow an approximate reconstruction of an image in three dimensions using waveforms. They also have the advantage of invariance under rotation, hence the synonym "Zernike invariants". The definition of a Zernike polynomial depends on two indices n and m such that

$n - m$ is even and $|m| \leq n$. A Zernike polynomial is defined on two variables: the distance ρ of the point from the origin, and the azimuthal angle θ measured counterclockwise from the x axis. Each polynomial $z_{n,m}(\rho, \theta)$ is the product of a degree n polynomial in ρ which is even when n is even and odd when n is odd, with either $\sin(|m|\theta)$ if $m < 0$ (odd invariant), $\cos(m\theta)$ if $m > 0$ (even invariant), or 1 if $m = 0$. Since even and odd invariants differ only by the choice of trigonometric function, invariants are identified only by their even subscripts ($m \geq 0$). The first (and simplest) Zernike polynomial is $z_{0,0}(\rho, \theta) = 1$, whereas the function $z_{8,2}(\rho, \theta) = \rho^2(-10 + 60\rho^2 - 105\rho^4 + 56\rho^6)\cos(2\theta)$ is a lot more complex. Letting n vary from 0 to 20 and m vary from 0 to n creates a 121-dimensional vector-valued function, which generates the fingerprint. Using $n \leq 20$ provides a compromise between precision and speed when comparing fingerprints [15].

A bank of fingerprints can be generated for a protein by performing this algorithm at each point on the surface map. The distance between two fingerprints is measured using the cosine of the angle between them, with values close to -1 indicating strong complementarity. These fingerprints can be used to search a database for structural homologs, similar to a BLAST (Basic Local Alignment Search Tool) search. When conducting the search, each fingerprint from the query bank is compared against every protein in the database, and the search returns the fingerprint from its bank that is most strongly complementary to the query fingerprint [15].

Computational methods

Sequence alignment and conservation scoring

The first stage of this investigation was a multiple-sequence alignment of the flu-virus hemagglutinin database. The sequences were downloaded from the NCBI Influenza Virus Resource [16]. We considered only those sequences that were complete and that originate from a human-host virus. The exclusion of

sequence fragment entries improves the alignment near the N- and C-termini. Since the objective of the investigation was to identify a surface that can be incorporated into a human vaccine, the host species restriction was necessary. In all, 5440 sequences match these criteria.

The sequences were then aligned using MATLAB's bioinformatics toolbox. Due to the size of the dataset, the alignment was done in two phases. In phase one, the dataset was subdivided into 17 sets of 320 sequences, and each set was aligned separately. The output from each alignment was a consensus profile and a vector of alignment scores at each position based on the BLOSUM50 scoring matrix. A lower score indicates a higher degree of conservation at that position, and a value of 0 indicates an invariant residue. In phase two, the profiles were aligned against each other, and gaps introduced into each profile were represented as a value of 100 in the corresponding score vector. The seventeen vectors from phase one and the score vector from phase two were averaged to obtain a score for each position in the total alignment.

fpCompare as a tool for aligning molecular surfaces

Three hemagglutinin crystal structures were obtained from the PDB for analysis: an H1 from the 1918 pandemic (PDB:1RUZ [17]), an H3 from a 1968 infection (PDB:1HGD [18]), and an H5 from a pathogenic avian influenza virus (PDB:3S11 [19]). The selection of these proteins was random within their subtypes, with the intention of choosing a representative of each of several subtypes. These proteins were viewed in Jmol [20] and Pymol [21], and the invariant residues were selected on each to visually determine their location in the protein structure and their suitability as recognition sites. Suitability criteria included proximity to the protein's surface, proximity to other invariant or highly conserved residues, and the presence of hydrophilic residues.

For each hemagglutinin structure, a fingerprint was computed at each significant point, and fpCompare [15] was run on each fingerprint set against a database of fingerprints of the entire non-redundant PDB.

fpCompare transforms the database fingerprint through translation and rotation to attain a global minimum distance, which is an unbounded positive number. A threshold distance of 150 was used to determine significance of a hit. The results for each hemagglutinin structure were visually compared against each other on the basis of agreement of fit. In order to classify a surface as a common binding partner, it must be complementary to the same region of the protein surface regardless of mutation. Orientations in which the surfaces intersected each other, or in which the recognition site on hemagglutinin was not accessible to a large molecule, were discarded.

```
001   MKTII ALSYI LCLVF AQKYP FNDNN ADNNS TATLC IGYHA NNNTD TVDTV
051   LEKNV TVTHS VNLLE DSHNG KLCDL RGVAP LHLGN CNIAG WILGN PECDG
101   FQNKK WWSYI VETKS YDNGT CYPGD FPDYE ELREQ LSSVG SFERF EIFPK
151   ESSWP NHDTN KGVSA ACPHR GNKSF YKNLI WLLTH KGNYK YPKLN KSYPN
201   NKGKD VLYLW GVHHP STDAD QQSLY QNADG YVFVG TSRYQ QKFIP EIGSR
251   PKVRD QEGRM NYYWT LVEPG DKILF EATGN LIAPR YAFAM ERNGG SGIMR
301   SDAPI HDCNS ECQTP NGAIN NDLTS LPFQN VHPIT YGKCP KYVKS NKLRL
351   ATGMR NVPSI QSRAY QRRKS RGLFG AIAGF IEGGW TGMVD GWYGY HHQNE
401   QGSGY AADLK STQNA IDQIT NKVNS VIEKM NTQFH AVGKE FNHLE KRIEN
451   LNKKV DDGFI DIWTY NAELL VLLEN ERTLD FHDSN MKNLY EKVRS QLRNN
501   AKDMG NGCFE FYHKC DNACM ESVKN GTYDY PKYSD EAKLN REQID GVKLE
551   SGYIY WILWI YYSTV ACSLV CVVSL GAISF WMCSN GNLLL QCRIC ILDQN
601   FRNFR I
```

Figure 8. The consensus sequence, with invariant residues highlighted in red.

Results

Sequence alignment revealed sites of zero variation

Scores for each subset alignment averaged at around 3 with the maximum scores reaching 14, not including gaps. There were a few positions that had low scores in some subsets and in the consensus, but were treated as gaps in the remaining subsets. The average score vector had 24 positions that were 0, indicating those residues that are invariant in the database. The final consensus sequence is given in Figure 8, and includes both the globular head (001-371) and the fusion peptide (372-606). The twenty-four invariant residues (red) are sparsely distributed throughout the protein, and many of them are

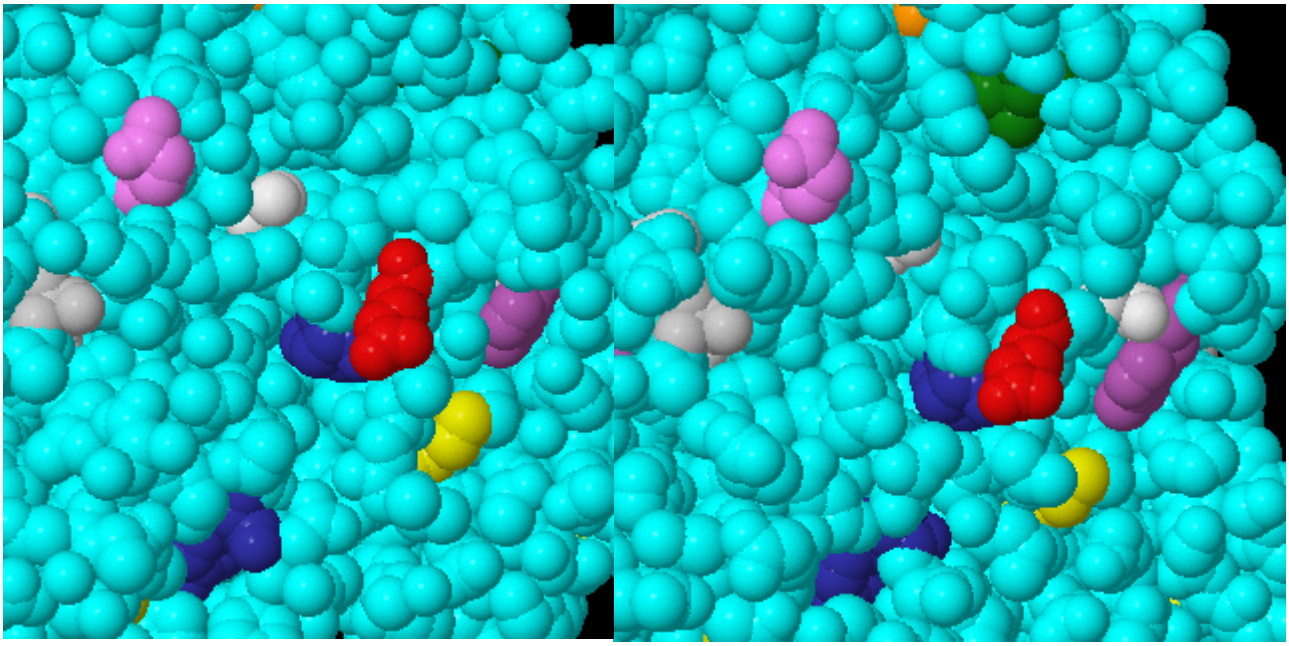

Figure 9. E382, W385, and C508 on structures 3S11 (left) and 1RUZ (right)

nonpolar hydrocarbons or aromatic. Of particular note is E382, which is associated with two bulky aromatic invariants (F380 and W385) and a nearby cysteine, C508. E382 is also associated with S411 and Q413, and the six residues form a small pocket in the molecular surface. This pocket appears to be an ideal epitope for binding, and it has the advantage of being located near a short palindromic sequence between E382 and the N-terminus of the fusion peptide that is also highly conserved, though this is hydrophobic and often buried on the interior of the protein. The conservation scores for each position can be found in Table A-3 in the appendix.

fpCompare returned false positives

Using a distance threshold of 30, the fingerprint comparisons returned several high-scoring hits for each of the hemagglutinin structures, of which 16 are shared by all three (Table 1). These hits include a bacteriophage T4 gene product (PDB:1EL6), an anthranilate phosphoribosyl transferase (PDB:1O17), and a nicotinate phosphoribosyl transferase (PDB:1YBE). In visually comparing the associations of each of these 16 potential binding partners with hemagglutinin, we found that each of the hits had a different

or, in some cases, impossible interaction with at least one of the hemagglutinin molecules, disqualifying it as a common binding partner.

3S11			1HGD			1RUZ		
Point	PDB	Distance	Point	PDB	Distance	Point	PDB	Distance
58810	1el6-C	288.537	60380	1el6-C	297.532	65450	1el6-C	289.325
80030	1el6-C	296.939						
6860	1erz-B	294.326	36030	1erz-B	287.756	20730	1erz-B	269.75
36500	1erz-B	294.326						
39500	1erz-B	294.326						
34740	1ktz-B	276.258	15310	1ktz-B	269.343	45810	1ktz-B	297.471
			49510	1ktz-B	269.343			
			84100	1ktz-B	293.405			
39330	1kzq-B	293.229	79670	1kzq-B	247.331	4800	1kzq-B	299.898
40430	1kzq-B	228.161						
50660	1o17-D	277.815	12110	1o17-D	252.159	7460	1o17-D	295.136
			47850	1o17-D	252.159			
51790	1oqe-J	277.82	9310	1oqe-J	272.967	54520	1oqe-J	233.805
43790	1sqj-B	294.625	51510	1sqj-B	271.898	10210	1sqj-B	260.055
			77430	1sqj-B	271.898	28100	1sqj-B	266.801
						45800	1sqj-B	260.055
17980	1tjv-D	294.461	1730	1tjv-D	281.687	20680	1tjv-D	295.306
54010	1tjv-D	294.461						
39430	1vh0-F	299.504	77440	1vh0-F	292.283	85630	1vh0-F	292.833
20620	1wxr-A	275.315	45370	1wxr-A	294.768	49480	1wxr-A	285.558
						65640	1wxr-A	268.458
3210	1xi0-B	258.391	86000	1xi0-B	299.865	13260	1xi0-B	297.718
780	1ybe-B	278.858	3960	1ybe-B	298.841	10220	1ybe-B	261.181
			66010	1ybe-B	295.588	22610	1ybe-B	298.335
			88540	1ybe-B	295.588	44900	1ybe-B	261.181
						83880	1ybe-B	261.181
21140	1yt4-A	296.311	5820	1yt4-A	246.548	18500	1yt4-A	282.145
			77440	1yt4-A	291.026	76840	1yt4-A	280.887
14890	2d13-D	249.498	15880	2d13-D	269.465	990	2d13-D	296.975
24680	2d13-D	255.411				30390	2d13-D	298.223
						33360	2d13-D	296.975
						33370	2d13-D	296.975
						80230	2d13-D	296.975
80030	2pz1-A	291.768	87140	2pz1-A	286.092	92060	2pz1-A	251.086
58830	2z07-B	271.203	80830	2z07-B	238.642	20730	2z07-B	292.016
						30390	2z07-B	296.771

Table 1. Data on the top scoring hits from fpComapre

Discussion

The hydrophilic pocket containing E382, S411, and Q413 is structurally conserved across hemagglutinin subtypes, residing on the same part of the structure in all three molecules analyzed. However, so is the hydrophobic N-terminal domain of the fusion peptide (GLFGAIAGFI). Since this peptide may be an acid-

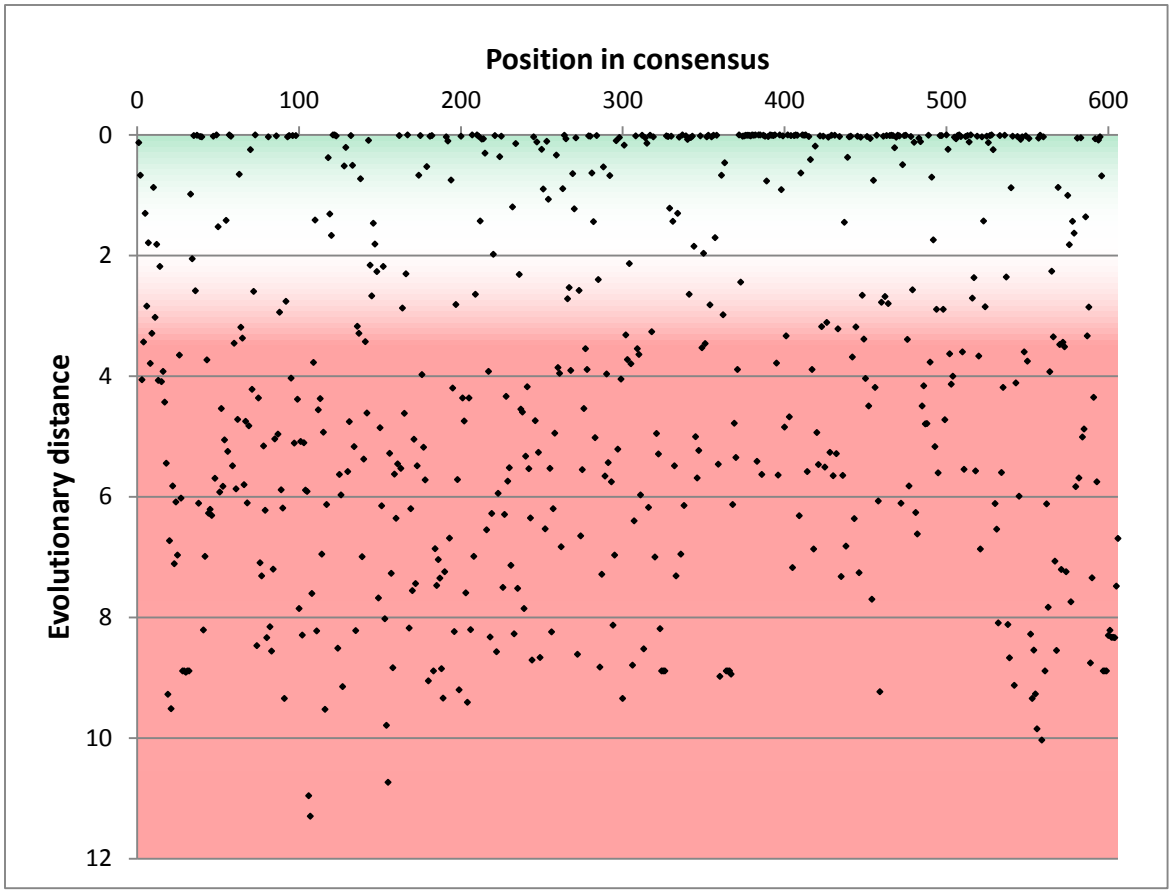

Figure 10. Variability map of the hemagglutinin sequence. Evolutionary distances were computed for each position as the average Euclidean length of the BLOSUM50 score vector for each alignment. The green region represents the band of lowest variability.

activated membrane anchor, targeting this region could inhibit this function and allow the lysosome to degrade the virus before it can empty its payload into the cell.

In addition to the 24 amino acids with zero evolutionary distance, there are several other positions with low evolutionary distance. Figure 9 shows several clusters of low-variability residues, denoted by points near the position axis. The gap around position 360 represents the interface of the two subunits. Most of the invariant residues are in the fusion-peptide subunit. Since this subunit is mostly located on the interior surface of the viral membrane, it is free from most selective forces, and the high degree of conservation in this region reflects that fact. The C-terminus of the globular head subunit is also buried on the interior of the protein, similarly shielded from external selective forces. However, there is a region of the sequence at position 217 that is both highly conserved and mostly polar. The sequence TDADQQSL forms a helix near the active site of the protein, with the first Q appearing in all but a few sequences (score of 0.0117). This helix is far more accessible to a small soluble peptide, and the polar acidic residues can form a strong interaction with a basic or positively-charged antibody. Moreover, there are three copies of each subunit in the functional hemagglutinin. It is therefore possible for an antibody to bind to all three copies of this helix, which dramatically increases the binding affinity relative to binding to just one copy.

The surface comparisons returned several candidates for a binding surface to hemagglutinin, even though none of them were shown to associate with the flu protein in a manner consistent with other subtypes or with the laws of physical chemistry. However, the search was performed in the low computational power environment of FAUST at the University of North Carolina at Chapel Hill, where probing the full query would have taken at least three weeks. To compensate, the query size was reduced significantly, retaining only every tenth point on each hemagglutinin surface in the query. As a result, the search duration dropped to just 2-3 days. With greater computational power such as that available on the HORNET supercluster, the running time on a full-surface query can be reduced to around 1-2 weeks. By further restricting the query to only surface points of interest, such as those

associated with the conserved polar helix, the search can be completed within a few hours, and the output may be more consistent across hemagglutinin subtypes.

Conclusion

Hemagglutinin has a few regions of low or no variability in its sequence, which may confer functional and structural stability for this protein. As expected, their locations in the observed quaternary structure of hemagglutinin are also highly conserved. Fingerprint comparisons using a sample of surface points returned some complementary surface candidates which may be strengthened by searching with the complete surface. In addition, a conserved helix found near the active site may prove significant in more directed fingerprint comparison searches. Further investigation in this direction may provide a candidate for a universal antibody against the flu virus which would prevent the onset of influenza and save thousands of lives.

Acknowledgements

I would like to take this opportunity to extend my thanks and gratitude to the following people for their help in my project and in my undergraduate career:

- To Dr. Greg Huber, my original research mentor, who helped connect me with the UNC group and the software resources with which I performed most of my research.
- To Dr. Nikolay Dokholyan and the Dokholyan Group, of the University of North Carolina at Chapel Hill, for sharing his fingerprint comparison software and fingerprint database.
- To Professors Craig Nelson and Ion Mandoiu, in whose class I was first exposed to the field and some techniques of computational biology and bioinformatics.

- To Dr. Lynne Goodstein, Dr. Jennifer Lease-Butts, and the rest of the Honors staff, for providing the program resources that led to this project happening.
- To Dr. Patrick McKenna, for mediating the final administrative steps leading to submission of this document.
- To Dr. Keith Conrad, for his instructive and moral support throughout my five years in the Department of Mathematics at the University of Connecticut.

Bibliography

1. National Library of Medicine, HBB, Available at http://ghr.nlm.nih.gov/gene/HBB (2014).

2. Shibayama, N., Sugiyama, K., Tame, J. & Park, S., Capturing the hemoglobin allosteric transition in a single crystal form. *J Am Chem Soc* **136** (13), 5097-105 (2014).

3. Harrington, D., Adachi, K. & Royer, W. J., The high resolution crystal structure of deoxyhemoglobin S. *J Mol Biol* **272** (3), 398-407 (1997).

4. Taubenberger, J. . M. M., 1918 Influenza: the Mother of All Pandemics, Available at http://wwwnc.cdc.gov/eid/article/12/1/05-0979_article.htm (2006).

5. Pandemic (H1N1) 2009, Available at http://www.who.int/csr/don/2010_05_14/en/index.html (2010).

6. Influenza (Seasonal), Available at http://www.who.int/mediacentre/factsheets/fs211/en/ (2009).

7. White, J., Hoffman, L. & Arevalo, J., in *Structural Biology of Viruses* (Oxford University Press, 1997), pp. 80–104.

8. Varghese, J., McKimm-Breschkin, J., Caldwell, J., Kortt, A. & Colman, P., The structure of the complex between influenza virus neuraminidase and sialic acid, the viral receptor. *Proteins*, 327-32 (1992).

9. Velumani, S. *et al.*, A Novel Peptide ELISA for Universal Detection of Antibodies to Human H5N1 Influenza Viruses, Available at http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0020737#pone.0020737-Prabakaran1 (2011).

10. Sahini, L., Tempczyk-Russell, A. & Agarwal, R., Large-Scale Sequence Analysis of Hemagglutinin of Influenza A Virus Identifies Conserved Regions Suitable for Targeting an Anti-Viral Response, Available at http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0009268 (2010).

11. Shisheng, S., Qinzhe, W., Fei, Z., Wentian, C. & Zheng, L., Glycosylation Site Alteration in the Evolution of Influenza A (H1N1) Viruses. *Public Library of Science ONE*, 0022844 (2011).

12. Henikoff, S. & Henikoff, J. G., Amino acid substitution matrices for protein blocks. *PNAS*, 10915-9 (1992).

13. Saitou, N. & Nei, M., The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 406-25 (1987).

14. PDB, Available at http://www.rcsb.org/pdb/.

15. Yin, S. & Dokholyan, N., Fingerprint-based structure retrieval using electron density. *PROTEINS: Structure, Function, and Bioinformatics*, 1002-1009 (2011).

16. Bao, Y. *et al.*, The Influenza Virus Resource at the National Center for Biotechnology Information. *Journal of Virology*, 596-601 (2008).

17. Gamblin, S. J. *et al.*, The structure and receptor binding properties of the 1918 influenza hemagglutinin. *Science*, 1838-1842 (2004).

18. Sauter, N. *et al.*, Binding of influenza virus hemagglutinin to analogs of its cell-surface receptor, sialic acid: analysis by proton nuclear magnetic resonance spectroscopy and X-ray crystallography. *Biochemistry*, 9609-21 (1992).

19. DuBois, R. M. *et al.*, Acid stability of the hemagglutinin protein regulates H5N1 influenza virus pathogenicity. *Plos Pathog*, e1002398 (2011).

20. Jmol: an open-source Java viewer for chemical structures in 3D, Available at http://www.jmol.org/.

21. Schrödinger, L., The PyMOL Molecular Graphics System, Version 1.5.0.4, Available at http://www.pymol.org.

22. Wyant, J., Zernike polynomials, Available at http://wyant.optics.arizona.edu/zernikes/zernikes.htm.

Appendix

Name	3-letter Code	1-letter Code
Alanine	Ala	A
Cysteine	Cys	C
Aspartate	Asp	D
Glutamate	Glu	E
Phenylalanine	Phe	F
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Lysine	Lys	K
Leucine	Leu	L
Methionine	Met	M
Asparagine	Asn	N
Proline	Pro	P
Glutamine	Gln	Q
Arginine	Arg	R
Serine	Ser	S
Threonine	Thr	T
Valine	Val	V
Tryptophan	Trp	W
Tyrosine	Tyr	Y

Table A-1. Amino acid nomenclature reference sheet

<table>
<tr><th></th><th>T</th><th>C</th><th>A</th><th>G</th><th></th></tr>
<tr><td rowspan="2">T</td><td rowspan="2">Phe</td><td rowspan="2">Ser</td><td rowspan="2">Tyr</td><td rowspan="2">Cys</td><td>T</td></tr>
<tr><td>C</td></tr>
<tr><td rowspan="2">Leu</td><td>STOP</td><td>STOP</td><td>A</td></tr>
<tr><td>STOP</td><td>Trp</td><td>G</td></tr>
<tr><td rowspan="4">C</td><td rowspan="4">Leu</td><td rowspan="4">Pro</td><td rowspan="2">His</td><td rowspan="4">Arg</td><td>T</td></tr>
<tr><td>C</td></tr>
<tr><td rowspan="2">Gln</td><td>A</td></tr>
<tr><td>G</td></tr>
<tr><td rowspan="4">A</td><td rowspan="2">Ile</td><td rowspan="4">Thr</td><td rowspan="2">Asn</td><td rowspan="2">Ser</td><td>T</td></tr>
<tr><td>C</td></tr>
<tr><td rowspan="2">Lys</td><td rowspan="2">Arg</td><td>A</td></tr>
<tr><td>Met</td><td>G</td></tr>
<tr><td rowspan="4">G</td><td rowspan="4">Val</td><td rowspan="4">Ala</td><td rowspan="2">Asp</td><td rowspan="4">Gly</td><td>T</td></tr>
<tr><td>C</td></tr>
<tr><td rowspan="2">Glu</td><td>A</td></tr>
<tr><td>G</td></tr>
</table>

Table A-2. The genetic code.

Table A-3. Conservation scores for each residue in the consensus sequence. Gaps have been removed prior to numbering.

M1	**0.128572778**	**A40**	0.034478833	**A79**	6.222403906	**N118**	0.377006111
K2	0.671863277	**N41**	8.208523694	**P80**	8.33446492	**G119**	1.312484278
T3	4.057843069	**N42**	6.987781771	**L81**	0.034535389	**T120**	1.666057222
I4	3.434033937	**N43**	3.732726651	**H82**	8.154426101	**C121**	0
I5	1.302704731	**T44**	6.269902183	**L83**	8.5598895	**Y122**	0
A6	2.840000484	**D45**	6.209418302	**G84**	7.200852769	**P123**	0.015019444
L7	1.789334226	**T46**	6.307986985	**N85**	5.041829797	**G124**	8.510521396
S8	3.787682701	**V47**	0.022211889	**C86**	0.017513333	**D125**	5.628055556
Y9	3.29060492	**D48**	5.692077551	**N87**	4.960730645	**F126**	5.970322167
I10	0.870018104	**T49**	0	**I88**	2.940466677	**P127**	9.146818452
L11	3.02347629	**V50**	1.524813625	**A89**	5.887010537	**D128**	0.517776111
C12	1.814741525	**L51**	5.921370319	**G90**	6.187366198	**Y129**	0.206126333
L13	4.068137407	**E52**	4.535245395	**W91**	9.343094444	**E130**	5.581401637
V14	2.183936206	**K53**	5.828332216	**I92**	2.757500808	**E131**	4.75286787
F15	4.093824334	**N54**	5.057239178	**L93**	0.034800278	**L132**	0.0112935
A16	3.922231746	**V55**	1.416130093	**G94**	0.011413889	**R133**	0.507121833
Q17	4.430351705	**T56**	5.24580296	**N95**	4.031651526	**E134**	5.16581787
K18	5.442741874	**V57**	0	**P96**	0.012074833	**Q135**	8.21721069
Y19	9.273277845	**T58**	0.020904444	**E97**	5.108539388	**L136**	3.17245406
P20	6.730051343	**H59**	5.484406214	**C98**	0.009735	**S137**	3.289413852
F21	9.515899815	**S60**	3.456172393	**D99**	4.385669931	**S138**	0.729356333
N22	5.820300338	**V61**	5.869385252	**G100**	7.853122676	**V139**	6.994643838
D23	7.111550264	**N62**	4.719662734	**F101**	5.085243847	**G140**	5.376061607
N24	6.087705893	**L63**	0.654009889	**Q102**	8.294240748	**S141**	3.428926126
N25	6.965494431	**L64**	3.187969865	**N103**	5.103455044	**F142**	4.612655556
A26	3.650734603	**E65**	3.368786278	**K104**	5.889472694	**E143**	0.0891545
D27	6.020909964	**D66**	5.797798836	**K105**	5.909719351	**R144**	2.158514237
N28	8.888888889	**S67**	4.74857001	**W106**	10.95628592	**F145**	2.666877778
N29	8.888888889	**H68**	6.102327778	**W107**	11.29414416	**E146**	1.463065592
S30	8.906651667	**N69**	4.825890732	**S108**	7.604447303	**I147**	1.808314488
T31	8.888888889	**G70**	0.246128333	**Y109**	3.772133974	**F148**	2.263787778
A32	8.888888889	**K71**	4.220416488	**I110**	1.411664771	**P149**	7.679777912
T33	0.980326778	**L72**	2.595397367	**V111**	8.223835208	**K150**	4.856890908
L34	2.052335556	**C73**	0	**E112**	4.556476132	**E151**	6.149324393
C35	0.014056667	**D74**	8.470157548	**T113**	4.37352079	**S152**	2.18367462
I36	2.585166677	**L75**	4.364209384	**K114**	6.951260834	**S153**	8.022024694
G37	0.008234778	**R76**	7.096520796	**S115**	4.929887823	**W154**	9.787683333
Y38	6.107064158	**G77**	7.311803763	**Y116**	9.524805899	**P155**	10.73180209
H39	0.027154111	**V78**	5.160272819	**D117**	6.129117077	**N156**	5.278184117

H157	7.271593585
D158	8.83382641
T159	5.624408156
N160	6.358072529
K161	5.454803976
G162	0.012576111
V163	5.528720404
S164	2.871428436
A165	4.617524963
A166	2.305742222
C167	0
P168	8.176305181
H169	6.195063621
R170	7.558145571
G171	5.048763073
N172	7.436641277
K173	5.485812315
S174	0.668730556
F175	0.010030111
Y176	3.975225952
K177	5.177418222
N178	5.720008327
L179	0.524773556
I180	9.052104067
W181	0.023956111
L182	0.011589889
L183	8.888888889
T184	6.85996408
H185	7.472903886
K186	7.041384965
G187	7.349902007
N188	8.849079456
Y189	9.3358302
K190	7.242817519
Y191	0.034237444
P192	0.101973
K193	6.686378284
L194	0.747478333
N195	4.195276651
K196	8.232710935
S197	2.814787127
Y198	5.713915735
P199	9.202751785
N200	0.022111278
N201	4.363640222
K202	4.741817231
G203	7.590543815
K204	9.407879442
D205	4.361342153
V206	8.20273338
L207	0
Y208	6.98701109
L209	2.6453307
W210	0
G211	0.020710444
V212	1.425951098
H213	0.071493333
H214	0.065219944
P215	0.300764444
S216	6.547491788
T217	3.922438889
D218	8.323192615
A219	6.276803465
D220	1.979472778
Q221	0.011699667
Q222	8.5710874
S223	5.941855556
L224	0.361744778
Y225	0.022766056
Q226	7.503773798
N227	6.294872718
A228	4.335838889
D229	5.739422809
G230	5.515913408
Y231	7.138737134
V232	1.194609444
F233	8.2718535
V234	0.142761889
G235	7.518300454
T236	2.315083333
S237	4.548328728
R238	4.594722222
Y239	7.854298463
Q240	5.327132836
Q241	4.176841515
K242	5.533040344
F243	6.350688833
I244	8.706897495
P245	0.030455722
E246	4.736573173
I247	0.119488333
G248	5.26573563
S249	8.664546999
R250	0.240467778
P251	0.898972222
K252	6.532881574
V253	0.105492111
R254	1.066755
D255	5.527416667
Q256	8.242264973
E257	6.198522807
G258	4.943690851
R259	0.335055556
M260	3.857466651
N261	3.954919438
Y262	6.830037826
Y263	0.892862222
W264	0.008073333
T265	0.066522944
L266	2.718717475
V267	2.531911111
E268	3.904822235
P269	0.642630556
G270	1.224436667
D271	0.050718944
K272	8.613159148
I273	2.580094455
L274	6.649339908
F275	5.548031924
E276	4.538206506
A277	3.546972393
T278	3.887455556
G279	0.014767778
N280	0.024452222
L281	0.632001667
I282	1.438842549
A283	5.019632774
P284	0.012074833
R285	2.400492222
Y286	8.821389453
A287	7.287472583
F288	0.530307222
A289	5.658608852
M290	3.961328877
E291	5.43558578
R292	0.673937778
N293	5.750252581
G294	8.131273264
G295	6.966845241
S296	0.095508556
G297	5.210361318
I298	0.044465889
M299	4.046561095
R300	9.340980651
S301	0.170059111
D302	3.316277778
A303	3.724189908
P304	2.134777222
I305	3.795941402
H306	8.793756826
D307	6.398211548
C308	0.016013333
N309	3.546977778
S310	3.641517434
E311	5.970892316
C312	0.007334444
Q313	8.519366291
T314	0.047833389
P315	0.138125833
N316	6.1767622
G317	0.004095556
A318	3.262043453
I319	0.032537333
N320	7.000253993

N321	4.948894797	**S362**	2.981716142	**S403**	4.676214905	**L444**	3.185231838
D322	5.288925782	**R363**	0.462715939	**G404**	0.011292	**E445**	0.023900278
L323	8.184501199	**A364**	8.888888889	**Y405**	7.175040017	**K446**	7.259821666
T324	8.888888889	**Y365**	8.888888889	**A406**	0.010259444	**R447**	0.038423056
S325	8.888888889	**Q366**	8.888888889	**A407**	0	**I448**	2.660739444
L326	8.888888889	**R367**	8.942569444	**D408**	0.00803	**E449**	3.383330723
P327	0.014839833	**R368**	6.129548586	**L409**	6.314422222	**N450**	4.03979597
F328	0.029039556	**K369**	4.779972158	**K410**	0.633968333	**L451**	0.025395833
Q329	1.214367778	**S370**	5.34995269	**S411**	0	**N452**	4.496573845
N330	0.020003278	**R371**	3.888888889	**T412**	0.004821111	**K453**	0.058062333
V331	1.434696654	**G372**	0	**Q413**	0	**K454**	7.701481678
H332	5.485628436	**L373**	2.441577788	**N414**	5.581314356	**V455**	0.753292333
P333	7.310970228	**F374**	0.023447778	**A415**	0.027000278	**D456**	4.186605367
I334	1.300108333	**G375**	0.016305	**I416**	0.406741667	**D457**	0.008293389
T335	0.038110611	**A376**	0.004919556	**D417**	3.889495111	**G458**	6.07197473
Y336	6.951909672	**I377**	0.008096889	**Q418**	6.863720865	**F459**	9.232028053
G337	0	**A378**	0.007993333	**I419**	0.188677389	**I460**	2.774736525
K338	6.142282513	**G379**	0.016319944	**T420**	4.935706277	**D461**	0.019713222
C339	0.017513333	**F380**	0	**N421**	5.463793164	**I462**	2.680183343
P340	0.077399278	**I381**	0.004401444	**K422**	0.014577611	**W463**	0.008912778
K341	2.643728496	**E382**	0	**V423**	3.179435819	**T464**	2.79582057
Y342	0.041250222	**G383**	5.414035077	**N424**	0.022350667	**Y465**	0.009839111
V343	0.019469222	**G384**	0	**S425**	5.505326301	**N466**	0.007102222
K344	1.84782	**W385**	0	**V426**	3.111780263	**A467**	0.009754333
S345	5.002190748	**T386**	5.629894782	**I427**	0.037702444	**E468**	0.212100389
N346	5.68561356	**G387**	0.00556	**E428**	5.265777778	**L469**	0.039032778
K347	5.232523678	**M388**	0.020951778	**K429**	0.0065675	**L470**	0.004210889
L348	0.014588778	**V389**	0.763895556	**M430**	5.650868417	**V471**	0.016781556
R349	3.530037725	**D390**	0.026213556	**N431**	0.007227167	**L472**	6.109910325
L350	1.963740593	**G391**	0	**T432**	5.285019626	**L473**	0.493794111
A351	3.461094444	**W392**	0.008073333	**Q433**	3.215271173	**E474**	0.013326111
T352	0.033251111	**Y393**	0.009642556	**F434**	0.029273111	**N475**	0.005955222
G353	0.005853889	**G394**	0	**H435**	7.320485903	**E476**	3.391449124
M354	2.818192176	**Y395**	3.784261751	**A436**	5.643915017	**R477**	5.820975521
R355	0.033781556	**H396**	5.641480287	**V437**	1.448061438	**T478**	0.029708222
N356	0.008612056	**H397**	0	**G438**	6.819436099	**L479**	2.568194455
V357	1.703927222	**Q398**	0.908530722	**K439**	0.372983889	**D480**	0.121878111
P358	0.005592222	**N399**	0.017987222	**E440**	0.030988444	**F481**	6.263340692
S359	5.458306266	**E400**	4.842312315	**F441**	0.019903278	**H482**	6.617772222
I360	8.977625108	**Q401**	3.330202165	**N442**	3.682130549	**D483**	0.051944778
Q361	0.666751248	**G402**	0	**H443**	6.360949599	**S484**	0.112125222

N485	4.494284956
M486	4.160690942
K487	4.79101313
N488	4.787795648
L489	0.005004056
Y490	3.766617307
E491	0.703066389
K492	1.741360932
V493	5.167793885
R494	2.891520252
S495	5.60367863
Q496	0.012520278
L497	0.012579333
R498	2.889863085
N499	4.723851623
N500	0.003265444
A501	0.241675556
K502	3.630340381
D503	4.133677326
M504	4.000867762
G505	0.03708
N506	0.062556611
G507	0.011670556
C508	0
F509	0.026683333
E510	3.598195009
F511	5.54608748
Y512	0.029298889
H513	0.011447222
K514	0.118553889
C515	0

D516	2.707994444
N517	2.368943889
A518	5.570786829
C519	0.022228889
M520	3.667131108
E521	6.864962948
S522	0.034256111
V523	1.427093625
K524	2.850194357
N525	0.020609222
G526	0.129634389
T527	0.013237611
Y528	0
D529	0.246137
Y530	6.110280824
P531	6.536969143
K532	8.09000538
Y533	0.010584778
S534	5.599790042
D535	4.188782882
E536	0.006585556
A537	2.358025
K538	8.11726614
L539	8.671355566
N540	0.875834833
R541	0.022928778
E542	9.127889842
Q543	4.112912926
I544	0.027959611
D545	5.988888662
G546	0.077711111

V547	0.031401389
K548	3.599033898
L549	0.014859167
E550	3.749451492
S551	0.057719722
G552	8.279648205
Y553	9.343627401
I554	8.544225055
Y555	9.267348476
W556	9.849000726
I557	0.055067722
L558	0.008421778
W559	10.03423835
I560	0.033170722
Y561	8.888888889
Y562	6.117509575
S563	7.832924928
T564	3.924843106
V565	2.261395093
A566	3.347321231
C567	7.067467634
S568	8.549436703
L569	0.872919665
V570	3.478413754
C571	7.203385634
V572	3.439599904
V573	3.515593845
S574	7.24502508
L575	1.002833944
G576	1.820103111
A577	7.741641274

I578	1.430639258
S579	1.631880556
F580	5.830137709
W581	0.054731667
M582	5.688151928
C583	0.050961667
S584	5.007746304
N585	4.877762315
G586	1.356844497
N587	3.333979536
L588	2.856377499
L589	8.755248877
L590	7.34575
Q591	4.352206885
C592	0.062356111
R593	5.752691661
I594	0.087081222
C595	0.034395
I596	0.680237118
L597	8.888888889
D598	8.888888889
Q599	8.888888889
N600	8.297106976
F601	8.21600591
R602	8.332647038
N603	8.332694531
F604	8.335752653
R605	7.480763521
I606	6.689617808