Beilstein-Institut

Bozen 2006, May 15th – 19th, 2006, Bozen, Italy

Changing Patterns of Selective Pressure in Human Influenza H3 Haemagglutinin

Benjamin P. Blackburne and Richard A. Goldstein

Division of Mathematical Biology, National Institute of Medical Research (MRC), Mill Hill, London, NW7 1AA, U.K.

E-Mail: goldst@nimr.mrc.ac.uk

Received: 1st August 2006 / Published: 5th November 2007

Abstract

We analyse the evolution of haemagglutinin from human influenza H3 using a model that allows for variations in selective pressure, both at different locations in the protein as well as during the course of evolution. It has been observed that, in contrast to the steady rate of sequence change, the antigenic properties of the haemagglutinin changes in a punctuated manner between well-defined clusters [1]. We find that the changes in antigenic properties correspond to increased rate of change in selective pressure, as if these antigenic clusters correspond to different interactions between the virus and the immune system. Conversely, despite a large increase in glycosylation during the past 40 years, these changes in glycosylation do not generally seem to be correlated either with changes in antigenic properties or with significantly more rapid changes in selective pressure.

Introduction

Influenza, as a single-strand RNA virus, evolves rapidly. This rapid evolution provides a challenge both to the immune system as well as to the developers of new vaccines. The time-lag between the choice of a new vaccine and its production and availability means that the prediction of the next dominant strain must be made quite early. This choice is assisted by the presence of a world-wide monitoring system that is on constant watch for new and expanding strains, but it is still difficult to evaluate and predict the nature of the emerging and changing threat. This difficulty is exacerbated by the lack of good evolutionary models for understanding how these viruses evolve.

Most of the insights and models underlying protein evolution are based on a number of assumptions the applicability of which to viruses can be questioned. Among these common assumptions are that the changes in the virus are conservative, that they are reversible, that the amino acids are in equilibrium, that the selective pressure does not depend upon the location in the protein, nor upon the time at which the changes occur. While there has been some work relaxing some of these assumptions in some ways, the relaxation of many of these assumptions is still at an early stage. For instance, it is increasingly common to allow differences in the *magnitude* of the selective pressure at different locations [2 – 4], but rarer to allow differences in the *nature* of the selective pressure. Measuring changes in the selective pressure over time is rarer still. Many evolutionary analyses are done with conventional substitution models that encode the conservative, reversible nature of the evolutionary process as it generally occurs in prokaryotic and eukaryotic proteins. Whether these models are appropriate for viruses is again an open question.

Evolutionary analyses rest on the validity of the assumptions used. Conversely, the breakdown of these assumptions can give us insight into the nature of viral evolution. Observing how and when there are variations in the selective pressure can give hints to the changing interactions between the virus and the immune system.

We have developed a mixture-, hidden states-, or 'site-class'- model [5 – 8]. This model tries to include variations in both the magnitude and nature of the selective pressure at different locations. The underlying premise is that, *a priori*, we cannot know the different types of selective pressure acting on the protein, nor which locations are under which type. Instead, we let the available sequence data, through a maximum-likelihood analysis, develop the optimal set of models for the different types of selective pressure. We can then calculate *a posteriori* which locations evolve under which models. This approach allows us to calculate the different types of selective pressure, to group different locations which are under similar forms of selective pressure, and analyse changes in the magnitude and nature of the selective pressure, without making potentially-limiting assumptions about what is behind these differences and changes.

In this paper, we describe preliminary results on the modelling of haemagglutinin (HA), a membrane-bound glycoprotein present on the surface of the influenza A virus which is responsible for receptor-binding and membrane-fusion [9]. Sixteen different subtypes of HA have been identified in influenza A (H1 to H16) of which two forms (H1 and H3) currently circulate widely in humans [10]. Before membrane fusion can occur, the HA precursor (HA_0) must be cleaved into two polypeptides (HA_1 and HA_2) that are linked by a disulphide bond. Five antigenic regions have been identified on the HA_1 polypeptide of H3 [11, 12]. Because HA_1 is under particular selective pressure from the immune system, it has a higher replacement rate for amino acids than HA_2 [13]. While a significant amount of attention has been focused on H5 (so-called 'Bird Flu'), the greatest current health problems are the result of H3. In this paper, we focus on the molecular evolution of H3 in humans.

One interesting observation that has been made is that the antigenic properties of H3 move in a discontinuous, step-wise manner between well-defined clusters [1], in contrast to the smoother changes in the sequence. There are two possible interpretations of these punctuated changes. The first is that some changes in some locations simply have more of an effect than others. These changes in these locations correspond to the transitions between clusters, while other changes in other locations correspond to movements within clusters. As an example, it is observed that two different cluster changes are caused by the same single amino acid substitution. A second interpretation is that the clusters represent a particular type of interaction between the virus and the immune system, while a transition between clusters would represent changes in the nature of this interaction. If so, it is likely that these cluster transitions would also involve changes in the selective pressure.

It has also been noted that there has been a rapid increase in the glycosylation of H3. Again, we might expect that changes in glycosylation would both affect the antigenic properties as well as the selective pressures.

In this paper, we describe an extension of our site-class model to include time-variation, the idea that the selective pressure can be different at different evolutionary times. We then address the question about the relationship between changes in antigenic properties, changes in glycosylation, and changes in selective pressure. We find that the changes in antigenic properties correspond to faster rate of change in selective pressure, indicating that transitions between antigenic clusters involve a changing relationship between the virus and the immune system. Surprisingly, we find that there is little correlation between changes in glycosylation and either changes in selective pressure or changes in antigenic properties.

Methods

Substitution models

We call locations under the same type of selective pressure as belonging to the same 'site class'. Each site class is associated with a substitution matrix, representing the rate at which the 380 possible amino acid substitutions occur, thus encoding the degree and nature of the selective pressure. We do not assign different locations in the protein to different site classes. Rather, each site class k is associated with an *a priori* probability $P(k)$ of representing any location in the protein.

In this preliminary study, we assume that the substitution matrices are reversible, that is, the probability of observing a substitution of amino acid x for y is equal to the probability of observing its inverse. In this case, we can represent the substitution matrix by an overall rate υ, a symmetric matrix $\mathbf{S}$ (S_{ij}=S_{ji}), and the equilibrium frequencies of the twenty amino acids $\{\pi_i\}$. In order to reduce the number of adjustable parameters, we assume that $\mathbf{S}$ is the same for all site classes, so that the differences in selective pressure can be represented by differences in the overall rate of substitution and in differences in the amino acid equilibrium frequencies. The adjustable parameters of the site-class model then include, for each site class, the *a priori* probabilities $P(k)$, the site-class dependent overall substitution rate

υ_k, and the equilibrium amino acid frequencies. As all sites must belong to some site class, $\sum_k P(k) = 1$. Similarly, the equilibrium frequencies of the amino acids for each site class must sum to one. In addition, the averaged substitution rates (weighted by the *a priori* probabilities) are normalized to a constant number of substitutions per evolutionary distance.

Following Holmes and Rubin [14], we consider that there can be changes in site class, corresponding to changes in selective pressure. We do this by constructing a $20N_s$ x $20N_s$ substitution matrix, where N_s is the number of site classes. The N_s 20 x 20 blocks on the main diagonal represent the different amino acid substitution matrices corresponding to each of the site classes, while other entries represent changes between site classes. For simplicity, we assume that there are no simultaneous changes of both site classes and amino acid. We also assume that the rate of change of site class is reversible. Q_{ii}^{kl}, the rate of change from site class k to l for a location currently occupied by amino acid A_i is given by $Q_{ii}^{kl} = P(l)\ \pi_i^l\ Z_{kl}$ where **Z** is a N_s x N_s symmetric matrix, with $(N_s^2 - N_s)/2$ adjustable parameters.

We used the dataset of Smith *et al.* [1] which contains 254 Human H3 HA_1 sequences sampled from 1968 to 2003. An avian H3 sequence (A/Duck/Hokkaido/33/80, M16739) was used as an outgroup to root the tree. Sequences were extracted from the Influenza Database [15]. Various sets of sequences were assigned to different antigenic clusters following Smith *et al.* [1]. In contrast to the assignments of Smith *et al.*, better results were obtained with the assignment of A/Shanghai/24/90 to the BE92 cluster rather than the SI87 cluster. In terms of antigenic properties, A/Shanghai/24/90 is intermediate between these two antigenic clusters (A. Lapedes, personal communication.) Following the maximum likelihood formalism, we first constructed a phylogenetic tree of the various sequences using standard substitution models and the program PHYML [16]. We then used PAML [17] with the WAG substitution model [18] and a Gamma-distributed rate [4] to construct an optimal symmetric substitution matrix **S**. We then encoded the site-class model and adjusted the parameters (including branch-lengths) in order to maximize the calculated log-likelihood.

We are interested in considering the possibility that there are certain branches where site-classes change more quickly than others. We do this by multiplying **Z** by an unknown factor γ for these particular branches. The appropriateness of this factor can be evaluated by calculating how it changes the log-likelihood values.

Ancestral glycosylation states were determined by searching for locations containing the sequence Asn-Xaa-Ser/Thr with probability > 0.95. Homology models of a representative set of ML ancestral sequences were made with SwissModel [19] based on the 1MQN structure [20]. When glycosylation states were predicted by the GlyProt server [21], all potential locations were predicted to be glycosylated.

Model choice

In general, more complicated models will yield better fits to the data, as represented by log-likelihood scores. The question often is, does the resulting increase in log-likelihood justify the increasing complexity of the model? We can distinguish between two different cases. In the first case, the simple model is 'nested' within the more complex model, that is, fixing some of the adjustable parameters of the more complex model results in the simple model. In this case, we can use the likelihood-ratio test to distinguish whether the additional complexity is justified. This is the situation when we are seeking to justify the presence of changes between site classes, or whether some branches have a faster rate of site-class change than others.

When the models are not nested, the likelihood-ratio test is no longer appropriate. In this case we consider the Akaike Information Criterion (AIC) [22], which is defined as:

$$\mathrm{AIC} = 2\mathrm{Np} - 2\Lambda \qquad (1)$$

where N_p is the number of adjustable parameters and Λ is the log-likelihood. The preferred model is that which minimizes the resulting AIC.

RESULTS

Model development and optimization

The phylogenetic tree is drawn schematically in Fig. 1. In contrast to the tree derived by Smith *et al.* [1], we have transitions from antigenic clusters EN72 to VI75 and TX77, with VI75 representing an evolutionary dead-end, as well as transitions from SI87 to both BE89 and BE92, with BE89 representing a dead-end.

The standard optimized single substitution-model (Gamma-distributed rate class) yielded a log-likelihood of -4674.1, for AIC = 10492.2 (with N_p= 572). The four site-class model, prohibiting changes between site-classes, achieved a log-likelihood value of -4339.4 for an AIC = 9946.8 (with N_p= 634), significantly lower. This indicates that the distribution of substitution rates cannot be adequately represented by a Gamma distribution of absolute rates but is better modelled by a set of substitution matrices that includes qualitative differences in the selective pressure at different locations, such as the site-class model. Site-class models with greater or fewer site classes resulted in higher values of AIC, indicating that the four site-class model was optimal for these data.

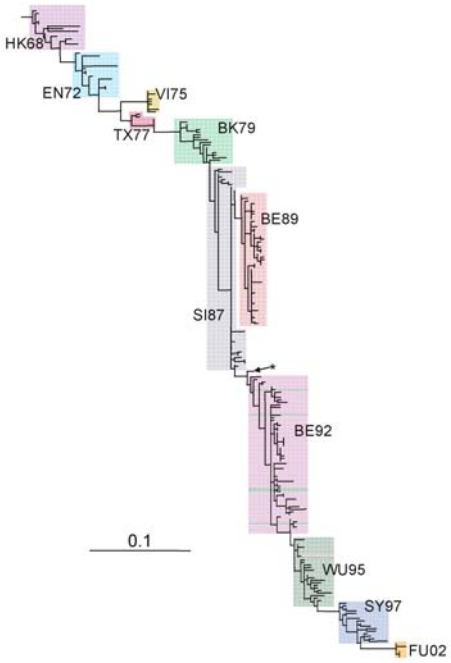

Figure 1. Phylogenetic tree of influenza H3 HA_1 sequences. Regions of the tree are colour-coded and labelled according to their antigenic cluster, as defined in (Smith, Lapedes *et al*.[1]); labels represent the location (Hong Kong (HK), England (EN), Victoria (VI), Texas (TX), Bangkok (BK), Sichuan (SI), Beijing (BE), Wuhan (WU), Sydney (SY), Fujian (FU)) and year of the first identification. * indicates A/Shanghai/24/90, which we assigned to the BE92 cluster rather than the SI87 cluster.

Glycosylation changes

We used this model to predict glycosylation states of the various ancestral nodes. We observed a substantial loss in glycosylation sites in the terminal branches of the tree, suggesting possible sequencing error, selective pressure differences in the culturing of viruses previous to sequencing, or a reduced fitness of these viruses consistent with their lack of progeny. For this reason, we restricted our analysis of glycosylation sites to internal nodes. We observe a sharp increase in the amount of glycosylation from six sites (HK68) to 11 (FU02), as shown in Fig. 2.

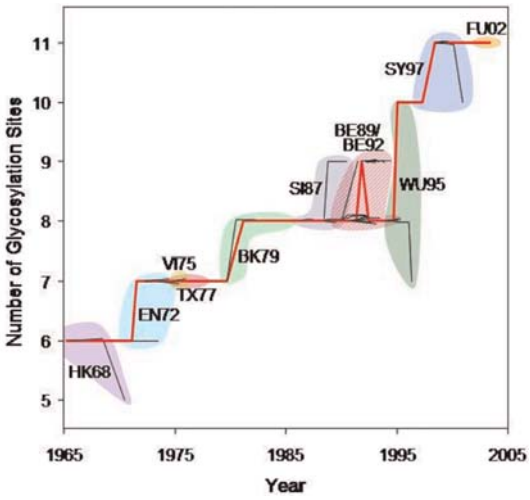

Figure 2. Changes in the number of glycosylation sites with time. Clusters are indicated as in Fig. 1. Red line represents the main branch of the tree. Dates for the internal nodes were estimated based on a least-squares fit to the acquisition time of the available sequences.

Incorporating changes in selective pressure

Allowing changes of site class increased the log-likelihood to -4325.1. In this case, the previous model (no changes in site class) is nested in this more general model, so we can use the likelihood ratio test to evaluate if the increase in log-likelihood justifies the additional six parameters, which it can ($P < 10^{-4}$). This indicated that the selective pressure acting on different locations was changing during the time period represented by the data. An additional question is whether the changes in selective pressure occurred randomly throughout the phylogenetic tree, or whether it was concentrated on certain branches. We investigated this issue with a series of more elaborate models: we considered that there were two different categories of branches, with the second category more likely to experience changes in selective pressure. This was performed by multiplying the rates of change of site class for the second category by an adjustable parameter γ. We first investigated whether there was a statistical tendency for changes in selective pressure to occur during changes in antigenic cluster. Considering such branches as a separate category with a different rate of site class change increased the log-likelihood by 2.28 ($\gamma = 6.6 \pm 2.2$), corresponding to $P = 0.03$ with the likelihood-ratio test. This indicates that we could reject the hypothesis that intra-cluster branches and inter-cluster branches could be represented by the same matrix. We also examined a model where branches involving changes in glycosylation state were proposed to involve faster changes of site class. The resulting model had a negligible increase in likelihood (0.04, $P = 0.78$), indicating lack of evidence that changes in glycosylation were associated with changes in selective pressure.

Discussion

Interestingly, there seems to be no correlation between changes in glycosylation and changes in antigenic properties, with no transition between antigenic clusters corresponding to a difference in glycosylation. Conversely, some antigenic clusters contain a multiplicity of different glycosylation sites; WU95 viruses, for instance, contain between 7 and 10 glycosylation sites. This suggests that the rapid change of glycosylation patterns are disjoint from the changes in antigenic properties.

Haemagglutinin evolves under different sources of selective pressure. Some locations are under purifying (negative) selection based on its need to recognize and bind to the target cell and initiate membrane fusion, while there is also positive selective pressure to change to avoid detection by the immune system. The conflicting nature of these pressures can be seen by the large overlap between antigenic sites and the binding regions. Only some locations in the antigenic regions are able to undergo rapid sequence evolution, with many of the antigenic sites changing slowly.

The immune system and the influenza virus co-evolve, each responding to measures and counter-measures developed by the other. There is an 'historical record' of their interactions encoded in the immunological memory as well as in the sequence of the viral proteins. This historical record provides the foundation for future sequence changes, and the changing nature of the accumulating historical record suggests that the selective pressure acting on HA might also change. If the dominant antibodies shift their binding to a new antigenic region, we would expect a corresponding shift in the region of HA under positive selective pressure. Glycosylation changes might hide part of the protein surface from immune recognition, reducing the selective pressure at this location. These different mechanisms are related, with the virus adapting its strategy to the changes in the nature of the immune response. Substitutions that assist the HA in avoiding immune recognition might have structural and functional consequences. This might induce selective pressure on other regions of the protein to undergo compensatory substitutions. Similarly, changes that have structural and functional consequences can either allow or restrict the ability of certain parts of the virus to evolve to avoid the immune response.

In fact, there is significant evidence of such changes in selection pressure. For instance, while changes in the 18 locations identified by Bush *et al.* as under positive selection from 1983 to 1997 [23] seemed to be correlated with the subsequent phylogenetic trajectories [24] and changes in antigenic properties [1] during this same time, changes in these 18 locations over a longer time-range, however, were only weakly correlated with changes in antigenic properties [25, 1]. An analysis of sequence change and sequence variability suggested that antigenic drift involved changing of a single epitope, but that the identity of the epitope varied from transition to transition [26]. All of this suggests that positive selection is a feature of influenza evolution, but that the *locations* undergoing positive selection may change and new 'antigenic sites' may emerge.

We also observe direct evidence for these changes in selective pressure. Interestingly, we find that these shifts in selective pressure occur preferentially along certain branches of the phylogenetic three. This suggests that, rather than a gradually shifting relationship between immune system and influenza virus, there are certain 'rules of engagement' that last for a period of time, and then are replaced by a new set.

These types of 'punctuated' changes have been observed at the phenotypic level previously, as Smith and co-workers observed that the antigenic properties of haemagglutinin do not change smoothly, but rather jump between well-described clusters [1]. This latter observation does not necessarily indicate that the relationship between immune system and virus is changing. An alternative explanation is that a small fraction of sequence changes have significant impact on the antigenic properties while most do not. In this alternative, the selective pressure might still be relatively constant, or change in a way not correlated with the changes in antigenic properties.

We find that our observed changes in selective pressure correspond closely to the changes in antigenic properties observed by Smith *et al.* [1] suggesting that these phenotypic changes correspond to adjustments in the relationship between virus and immune system – either changes in the mechanism of immune-avoidance for the influenza, or changes in the nature of the antibodies that must be avoided.

Changes in site class occur preferentially in antigenic regions, as well as in locations identified as under selective pressure. Interestingly, the site-class changes that occur in antigenic sites are preferentially found in coil and beta-sheet structures, and are less frequent in alpha-helices and loops. We also observe rapid substitutions, as well as rapid changes of site class, in the central 'pore' of the protein. These locations are not identified either as being under positive selection, or as changing during shifts in antigenic property.

Smith *et al.* [1] identified amino acids that were different in the different antigenic clusters. It is important to note that these are not necessarily *cluster-defining* changes, in that some of these changes might have occurred independently of any changes in antigenic properties. Still, there is a strong correlation between sites undergoing such amino acid changes between antigenic clusters and locations where there are corresponding changes in selective pressure. Often, however, cluster-difference mutations take place at locations that remain in the slowest two site classes, with nearby locations undergoing changes in site class. Cluster changes might involve a change of amino acid at an antigenic site that is structurally or functionally conserved. The change in antigenic properties caused by this rare event may then result in a change in the antibodies that can bind successfully (hence the jump to a new cluster) leading to in a change in selective pressure on the surrounding locations.

Interestingly, while changes in selective pressure occur preferentially at locations where there are changes in glycosylation states, we do not detect a significant increase in the rate of site class change during changes in glycosylation. This might represent a lack of sufficient data to identify such a rate increase, or it may indicate that changes in glycosylation cause local changes in selective pressure, in contrast to more wide-spread and sig-

nificant changes. It is interesting to note that there is no observed tendency for changes in antigenic cluster to be associated with changes in glycosylation. No antigenic cluster change involves a glycosylation change, while each antigenic cluster contains a wide range of different glycosylation states. As indicated in Fig. 2, there has been an appreciable increase in HA glycosylation during the past decades. Given the lack of connection between these changes and changing selective pressure or changing antigenic properties, the role of these glycosylation changes remains puzzling.

We find strong support for a model where the selective pressure changes at different locations. We find that a model that allows for different rates of site class changes during transitions between clusters is significantly better than a uniform model. From this we can infer that evolution of human H3 consists of periods of amino acid variation according to a relatively constant set of rules, interspersed with periods where the rules governing variation change. As the selective pressure changes during the evolutionary process, this indicates that, in addition to modelling how amino acids change during time, we also may need to develop models for how the selective pressure changes if we wish to be able to identify or predict future significant variations.

References

[1] Smith, D.J., Lapedes, A.S. *et al*. (2004) Mapping the antigenic and genetic evolution of influenza virus. *Science* **305**(5682)**:**371**-**376**.**

[2] Nei, M. (1987) *Molecular Evolutionary Genetics.* Columbia University Press, New York.

[3] Yang, Z. (1993) Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* **10**(6)**:**1396 – 1401.

[4] Yang, Z. (1994) Maximum likelihood estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39:**306**-**314**.**

[5] Koshi, J.M., Mindell, D.P.*et al.* (1997) Beyond mutation matrices: Physical chemistry based evolutionary models. In: *Genome Informatics1997.* (Miyano, S., Takagi, T., Eds) pp.80 – 89. Universal Academy Press, Tokyo.

[6] Koshi, J.M., Goldstein, R.A. (1998) Mathematical models of natural site mutations including site heterogeneity. *Proteins* **32:**289 – 295.

[7] Koshi, J.M., Mindell, D. *et al.* (1999) Using physical-chemistry based mutation models in phylogenetic analyses of HIV-1 subtypes. *Mol. Biol. Evol.* **16:**173 – 179.

[8] Dimmic, M.W., Mindell, D.P. *et al.* (2000) Modeling evolution at the protein level using an adjustable amino acid fitness model. *Pac. Symp. Biocomput.* 18 – 29.

[9] Skehel, J., Wiley, D. (2000) Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. *Annu. Rev. Biochem.* **69:**531 – 569.

[10] Lin,Y.P., Gregory, V. *et al.* (2004) Recent among human influenza viruses. *Virus Res.* **103**(1 – 2)**:**47**-**52**.**

[11] Wiley,D., Wilson, I. *et al.* (1981) Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. *Nature* **289**(5796)**:**373 – 378.

[12] Wilson, I., Cox, N. (1990) Structrual basis of immune recognition of influenza virus hemagglutinin. *Annu. Rev. Immunol.* 8:737 – 771.

[13] Hay, A.J., Gregory, V. *et al.* (2001) The evolution of human influenza viruses. *Phil. Trans R. Soc. Lond. B. Biol. Sci.* **356**(1416)**:**1867**-**1870.

[14] Holmes, I., Rubin, G.M. (2002) An expectation maximization algorithm for training hidden substitution models. *J. Mol. Biol.* **317**(5)**:**753**-**764**.**

[15] Macken**,** C., Lu**,** H. *et al.* (2001) The value of a database in surveillance and vaccine selection. In: *Options for the Control of Influenza.* (Osterhaus, A.D.M.E., Cox, N., Hampson, A.W., Eds), **IV**, pp.103 – 106. Elsevier, Amsterdam.

[16] Guindon, S., Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52:**696 – 704.

[17] Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *CABIOS* **15:**555 – 556.

[18] Whelan, S., Goldman, N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18:**691 – 699.

[19] Schwede, T., Kopp, J. *et al.* (2003) SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res.* **31**(13)**:**3381 – 3385**.**

[20] Ha, Y., Stevens, D.A. *et al.* (2003) X-ray structure of the hemagglutinin of a potential H3 avian progenitor of the 1998 Hong Kong pandemic influenza virus. *Virology* **309**(2)**:**209 – 218.

[21] Bohne-Lang, A., von der Lieth, C.W. (2005) GlyProt: in silico glycosylation of proteins.*Nucleic Acids Res.* **33**(Web Server Issue)**:** W214 – 219.

[22] Akaike, H. (1978) A Bayesian analysis of the minimum AIC procedure. *Annals Inst. Stat. Math.* 30:9 – 14.

[23] Bush, R.M., Fitch, W.M. *et al.* (1999) Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Mol. Biol. Evol.* **16**(11)**:**1457**-**1465**.**

[24] Bush**,** R.M., Bender**,** C.A. *et al.* (1999) Predicting the evolution of human influenza virus A. *Science* **286**(5446)**:**1921**-**1925.

[25] Lee, M.-S., Chen, J.S.-E. (2004) Predicting antigenic variants of influenza A/H3N2 viruses. *Emerg. Infect. Dis.* **10:**1385 – 1390.

[26] Plotkin, J.B., Dushoff, J. *et al.* (2002) Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus. *Proc. Natl Acad. Sci. U S A* **99**(9)**:**6263**-**6268.