Journal of General Virology (2004), **85**, 3249–3259 DOI 10.1099/vir.0.80021-0

Homology model of the structure of influenza B virus HA1

Chang-Shung Tung, Joshua L. Goodman,† Henry Lu‡ and Catherine A. Macken

Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

Correspondence
Catherine A. Macken
cmacken@lanl.gov

Received 5 February 2004
Accepted 19 July 2004

Influenza B virus is one of two types of influenza virus that cause substantial morbidity and mortality in humans, the other being influenza A virus. The inability to provide lasting protection to humans against influenza B virus infection is due, in part, to antigenic drift of the viral surface glycoprotein, haemagglutinin (HA). Studies of the antigenicity of the HA of influenza B virus have been hampered by lack of knowledge of its structure. To address this gap, two possible models have been inferred for this structure, based on two known structures of the homologous HA of the influenza A virus (subtypes H3 and H9). Statistical, structural and functional analyses of these models suggested that they matched important details of experimental observations and did not differ from each other in any substantive way. These models were used to investigate two HA sites at which viral variants appeared to carry a selective advantage. It was found that each of these sites coevolved with nearby sites to compensate for either size or charge changes.

INTRODUCTION

Influenza B virus causes substantial morbidity and mortality in humans. The inability to provide lasting protection to humans against influenza B virus infection is due, in part, to the rapid evolution of the viral surface glycoprotein, haemagglutinin (HA), which leads to a change in its antigenic nature.

Solution of the structure of the HA of influenza A virus (subtype H3) strain A/Aichi/2/68 (Wilson *et al.*, 1981) and studies of variant viruses enabled mapping of the antigenic sites on this protein (reviewed by Wilson & Cox, 1990). Similar mapping of the antigenic sites on the influenza B virus HA has been hampered by the lack of a known crystal structure. Studies of variants of influenza B viruses have allowed inference of some of the antigenic nature of the HA (Berton *et al.*, 1984; Hovanec & Air, 1984; Berton & Webster, 1985; Rivera *et al.*, 1995). These authors inferred the structure of antigenic sites on influenza B virus HA based on the assumption, proposed by Krystal *et al.* (1982), that the three-dimensional structures of the HAs of influenza A (H3) and B viruses are very similar. Qualitatively, this assumption is reasonable because of the functional similarity of these proteins; in reality, it leaves out details in structural differences that may occur due to mutations, insertions and deletions.

†Present address: Department of Biology, Indiana University, 1001 E. Third Street, Bloomington, IN 47403, USA.

‡Present address: 454 Life Science Corporation, 20 Commercial Street, Branford, CT 06405, USA.

More recently, the structure of the HA esterase (HEF) protein from C/Johannesburg/1/66, an influenza C virus (the third type of influenza virus that infects humans), was solved (Rosenthal *et al.*, 1998), followed by structures of the influenza A virus HA for three additional subtypes: H1 (A/South Carolina/1/18 and others; Gamblin *et al.*, 2004; Stevens *et al.*, 2004), H5 (A/duck/Singapore/3/97; Ha *et al.*, 2001) and H9 (A/swine/Hong Kong/9/98; Ha *et al.*, 2001). Whilst the functions of the HEF of influenza C virus are not equivalent to those of the HAs of influenza A and B viruses (the HEF has a combined HA and receptor-destroying esterase activity), nevertheless, the structures of the HEF and HA of A/Aichi/2/68 (H3) are remarkably similar (Rosenthal *et al.*, 1998). We decided to utilize this structural and functional knowledge to model the structure of the HA of influenza virus B/Lee/40, the oldest influenza B virus with an HA sequence in public databases.

Two classes of structure-prediction methods, *ab initio* and 'knowledge-based', have attracted great interest in recent decades and objective evaluation of their performance has received public scrutiny (http://PredictionCenter.llnl.gov). *Ab initio* methods have made notable progress toward predicting structures of small (<60 aa) proteins (e.g. Huang *et al.*, 1999; Moult, 1999; Bonneau & Baker, 2001), but have not yet been successful on proteins as large as HA. The 'knowledge-based' class of methods uses available information from protein-structure databases to develop models of proteins of unknown structure. Within this class, comparative modelling (or homology modelling) works at its best when the protein with unknown structure shares

substantial sequence homology with a protein of known structure. Under these circumstances, the known structural features of one protein (the template) can be used to model the unknown structural features of the other (the target).

The quality of a homology model depends critically on the extent of identity between the template and target sequences. A rule of thumb is that one needs $\geqslant 35\,\%$ sequence identity between the template and target proteins in order for homology modelling to be successful. An identity of $>50\,\%$ ensures that most of the features of a homology model will be accurate and that the root mean square error will be <2 Å, regardless of the size of the protein (Sánchez & Šali, 1997), an accuracy equivalent to that of a medium-resolution crystal structure. Low sequence identity ($<25\,\%$) leads to much less accurate predictions. Still, low sequence identity can lead to a useful model, as described by Tramontano (1998).

As *ab initio* predictions of the structure of the influenza B virus HA are not yet possible, we chose a homology modelling approach. We selected the HA1 polypeptide chain, one of two chains into which HA is cleaved during virus replication, as our target of interest, as it shows most of the variation of the HA.

Identity between the HA sequence of B/Lee/40 and the HA sequence from any of the influenza A viruses with known HA structure is substantially below the desired minimum of 35 %. However, the HA of A/Aichi/2/68 (H3) and the HEF of C/Johannesburg/1/66 have lower sequence identity than any of the B/A comparisons, yet their structures are similar (Rosenthal *et al.*, 1998). Hence, we were optimistic about the prospects for modelling the structure of the HA1 of B/Lee/40 successfully. To improve our chances of success, we included knowledge of the structural similarity between the HEF of C/Johannesburg/1/66 and the known HA structures of the influenza A viruses in our analysis. By the criteria described below, our modelling of the B-HA1 structure produced results of high quality that are consistent with the available sequence/function information on the HA of B/Lee/40, thereby providing a basis for investigating the evolution of this protein.

METHODS

Numbering system. For all sequences, we defined the Met that starts the protein-encoding sequence as position 1. Where insertions and deletions occurred in the influenza B virus HA molecule, we numbered residues in the HA1 according to the B/Lee/40 sequence.

Choice of template. The function of the HA of B/Lee/40 is closer to that of the HA of influenza A viruses than to that of the HEF of influenza C viruses. Hence, for comparative purposes, we selected two templates from known A-HA structures by assessing pairwise distances among A-HA structures and pairwise identity between B/Lee/40 and A-HA sequences in the multiple alignment of A-HA, B-HA and C-HEF sequences that we derived by using the procedure described below.

Sequence alignments. We chose the HA1 sequence of B/Lee/40 (GenBank accession no. NP_056660) as the representative of influenza B virus HA, for which we inferred structural models. The profile alignment option of CLUSTAL_W (Thompson *et al.*, 1994) with default parameters was used for initial alignments.

To generate an alignment between the HA1 sequences of B/Lee/40 and influenza A viruses, we began with a structure-informed alignment of the HEF sequence from C/Johannesburg/1/66 with one sequence from each of the 15 HA subtypes of influenza A virus, provided to the Influenza Sequence Database (ISD) (http://www.flu.lanl.gov; Macken *et al.*, 2001) by Dr J. Skehel, and available at http://www.flu.lanl.gov/review/review.html. To this profile, we added sequences manually, while maintaining the alignment, to capture the variation among types and subtypes by year and host species. We aligned this augmented A/C profile to an alignment of influenza B HA sequences that was used as a profile, thus generating a multiple alignment of A, B and C sequences. Insertions in the HA sequence of B/Lee/40 were then adjusted manually to preserve secondary-structure features in the A-HA sequences. Elements of secondary structure were assigned according to the crystal structure of the HA from A/Aichi/2/68.

Confirmation of alignment of HA/HEF sequences from influenza A, B and C viruses. For each site that was conserved in the A/B/C alignment, we counted the sequences in which that site was conserved within an alignment of sequences from each single type/subtype.

Homology modelling. We used the homology modelling techniques of Tung (1999). Briefly, we first matched the main-chain structures of the target to those of the template in the aligned regions. Insertions in the target relative to the template were treated as loops with known end-structure. Stretching the predicted structure accommodated insertions in the template relative to the target. The main-chain structures of the loops were modelled by using an efficient Monte Carlo loop-sampling method (Tung, 1997; Ryu *et al.*, 1998). Once the main-chain structure was modelled, the side-chain atoms were attached. As the head of the HA molecule is compact, limited space is available to place the side-chain atoms. Hence, in this analysis, side-chain torsional angles were initialized to equal or be close to those in the template structure. This consideration is particularly useful in avoiding clashes between side chains in the modelled structure. Finally, the all-atom models were subjected to a short run of energy minimization (1000 cycles) by using AMBER (Weiner *et al.*, 1986) to relieve unfavourable steric interactions and to optimize the stereochemistry.

Quality of homology model. We used PROCHECK (Laskowski *et al.*, 1993) to calculate the main-chain torsional angles, i.e. the Ramachandran plot (Ramachandran *et al.*, 1963), for our predicted structures.

Predicted model functionality. We asked whether our model structures supported three important functionalities of the HA1 molecule: binding sialic acid on the surface of the host cell (receptor binding), evolution to escape immune pressure and attachment of carbohydrates. To examine receptor-binding potential, we used the known structure of sialic acid as it binds to the HA of A/Aichi/2/68 (PDB accession code 1HGG) and tested the fit of this sialic acid structure to the predicted B-HA receptor-binding site. To check the structural plausibility of escape mutants, we used NACCESS (Hubbard & Thornton, 1993) to calculate the proportion of the surface areas of mutating residues that is solvent-accessible in our models. The same approach was used to check the plausibility of carbohydrate binding at potential glycosylation sites.

RESULTS

Choice of template

Structures of the HA of influenza A virus H1, H3, H5 and H9 subtypes are very similar. All pairwise root mean square distances (measured between C_α atoms for the aligned regions) among these structures are $<1{\cdot}52$ Å. For the purposes of homology modelling, all structures will lead to the same fold. That is, the overall shape of the model will be the same, regardless of the choice of template. However, local details of the model, such as loops at sites of insertions, are affected by the differences between the sequences of the target and template. In our alignment, the closest influenza A virus HA sequence to that of B/Lee/40 comes from A/swine/Hong Kong/9/98 (H9) (24 % identity). The next closest comes from A/Aichi/2/68 (H3) and A/duck/Singapore/3/97 (H5) (21 % identity). The HA sequence from the influenza A (H1) virus is the most distinct, having only 18 % identity. We decided to generate homology models based on known structures of the HA from A/swine/Hong Kong/9/98 (H9) (PDB accession code 1JSD) and A/Aichi/2/68 (H3) (PDB accession code 1HGF). Comparison of the models from these two templates would show the robustness of the model to the sequence of the template. An alternative A-H3 template is PDB accession code 1HGE; this structure has slightly higher resolution (2·6 Å) than 1HGF (3·0 Å), but corresponds to a G135R mutant. We chose 1HGF in order to work with the wild-type virus, as the difference in resolution would make no difference to our predictions. We did not choose the H1 subtype of the A-HA as a template because the details of a model based on this template will be predictably inferior to models from the other templates, even though the overall fold will be the same.

Sequence alignments

Our alignment of HA sequences from A/Aichi/2/68, A/swine/Hong Kong/9/98 and B/Lee/40, extracted from our A/B/C multiple alignment, is shown in Fig. 1. Twenty-one sites were conserved across these sequences and the HEF sequence of C/Johannesburg/1/66.

Krystal *et al.* (1982) published an alignment between the sequences of the HAs from A/PR/8/34 (H1N1) and B/Lee/40 viruses that has been utilized subsequently to infer the structure of influenza B virus HA. Our A/B/C alignment included an alignment of HA1 sequences from A/South Carolina/1/18 (H1) and B/Lee/40, which we compared with that of Krystal *et al.* (1982). The most significant difference lay in the region that spans residues 53–70 (B/Lee/40 numbering). In the paper by Krystal *et al.* (1982), in order to align Cys-69 in B/Lee/40 and Cys-68 in A/PR/8/34, the authors introduced a 12 aa insertion and a 7 aa deletion. As the Cys-293 of A/swine/Hong Kong/9/98, which forms a disulphide bond with Cys-68, is not matched in the B sequence, introducing a large insertion and deletion in order to align these two Cys is not energetically favourable. Our alignment involved only one 6 aa insertion in this region, as shown in Fig. 1.

Confirmation of the alignment

The majority of the 21 conserved residues in our A/B/C alignment were highly or completely conserved in single-type alignments of influenza A and B virus HA1 sequences (Table 1). Positions 83 and 269 (B/Lee/40 numbering) were interesting exceptions; variation at these positions will be discussed later in the context of model structure.

```
                                          *                                *
A/Aichi/2/68 (H3)     MKTIIALSYI FCLALG..QD LPGNDNSTAT LCLGHHAVPN GTLVKTITDD QIEVTNATEL VQSSSTGKIC N.NPHRILDG    77 [ 61]
A/Sw/HK/9/98 (H9)     -EAASLITIL LVVTASNA.. ........DK I-I-YQSTNS TET-D-L-ET NVP--H-K-- LHTEHN-ML- AT-LGHP-IL    70 [ 52]
B/Lee/40              --A--.VLLM VVTSNA.... ........DR I-T-ITSSNS PHV---A-QG EVN--GVIP- TTTPTKSHFA -.LKGTQTR-    66 [ 51]

                                  * *                 *   *     *
A/Aichi/2/68 (H3)     ......IDCT LIDALLGDPH CDVFQ.NETW DLFVERSKAF S.NCYPYDVP DYASLRSLVA SSGTLEFITE .GF...TWTG   145 [129]
A/Sw/HK/9/98 (H9)     ......DT-- IEGLIY-N-S --LLLGGRE- SYI---PS-V NGM---GN-E NLEE----FS -ASSYQR-QI ..-PDTI-.N   141 [123]
B/Lee/40              KLCPNCFN-- DL-VA--R-K -MGNTPSAKV SILH-VKP-T -.G-F-.IMH -RTKI-Q-PN LLRGY-N-RL STSNVINT.E   143 [128]

                         *     *         *
A/Aichi/2/68 (H3)     V.....TQNG GSNACKRG.P GSGFFSRLNW LTKSGSTYPV .....LNVTM PNNDNFDKLY IWGIHHPSTN QEQTSLYV.Q   213 [197]
A/Sw/HK/9/98 (H9)     -.....SYS- T-K--S.... .DS-YRSMR- --QKNNA--I .....QDAQY T--RGKS-IF M---N--P-D TV--N--T.R   205 [187]
B/Lee/40              TAPGGPYKV- T-GS-PNVAN -N---NTMA- VIPKDNNKTA INPVTVE-PY ICSEGE-QIT V--F-S.DDK TQMER--GDS   222 [207]

                                                        ***              *
A/Aichi/2/68 (H3)     ASGRVTVSTR RSQQTIIPNI GSRPWV...R GL..SSRISI YWTIVKPGDV LVINSNGNLI APR.GYFKMR TGKSSIMRSD   287 [281]
A/Sw/HK/9/98 (H9)     TDTTTS-T-E DINR-FK-V- -P--L-...N --..HG--DY --SVL---QT -RVR------ --WY-HILSG ESHGR-LKT-   280 [262]
B/Lee/40              NPQKF-S-AN GVTTHYVSQ- -GF-NQTEDE --KQ-G--VV DYMVQ---KT GT-VYQRGIL L-QKVWCAS. .-R-KVIKGS   300 [285]

                         *       *         *          * **                  *
A/Aichi/2/68 (H3)     APIDTCISEC ITPN.GSIPN DKPFQNV.NK ITYGACPKYV KQNTLKLATG MRNVPEKQT. .R   345 [329]
A/Sw/HK/9/98 (H9)     LNSGN-VVQ- Q-ER.-GLNT TL--H--.S- YAF-N----- GVKS----V- L----ARSS. .--  338 [320]
B/Lee/40              L.PLIGEAD- LHEKY-GLNK S--YYTGEHA KAI-N--IW- -T.P----N- TKYR-PAKLL KER  361 [346]
```

Fig. 1. Alignment of HA1 sequences of A/Aichi/2/68 (H3), A/swine/Hong Kong/9/98 (H9) and B/Lee/40 derived from a multiple alignment of influenza A virus and influenza B virus HA1 and influenza C virus HEF1 sequences. '*' indicates a site that is conserved in the A/B/C multiple alignment; '.' represents a gap; '-' represents a match with the A/Aichi/2/68 sequence. Numbering in square brackets refers to the mature HA1.

Table 1. Variability of the 21 sites that are conserved in the A/B/C multiple alignment

Position* (A-H9/A-H3/B)	Amino acid†	A (H9)‡ ($n=140$)	A(H3)‡ ($n=997$)	B‡ ($n=599$)
22/30/19	Cys		V, S, L	W
45/53/42	Thr	K	P	
81/88/83	Gly			46A
83/90/85	Pro		L	
99/105/101	Glu			
103/109/105	Ala		S	I, G, 594V
108/113/109	Cys		S	
146/150/153	Gly		3E	
151/155/158	Cys		F	F
155/163/167	Phe			L
246/254/268	Lys		T, 2E	
247/255/269	Pro		S, 2R	174S
248/256/270	Gly			E, R
262/270/284	Pro			
290/297/309	Cys			
295/302/315	Gly	D		
302/309/322	Pro		H	
312/319/333	Gly		R	
314/321/335	Cys		F	W
315/322/336	Pro		L	
333/340/353	Pro			

*Position of the conserved site in the respective numbering system of A/Aichi/2/68 (H3), A/swine/Hong Kong/9/98 (H9) and B/Lee/40.
†Residue at the conserved site.
‡Site-specific variation in alignments of *n* sequences of a single type/subtype. For example, 2R at position 247/255/269 means that, among 997 H3 sequences, Arg occurred twice and Pro 994 times at position 255 (H3 HA1 numbering). Blanks indicate complete conservation.

Quality of homology models

Both A-H3 and A-H9 templates could accommodate the required insertions and deletions. The site of the 6 aa insertion after residue 66 (A/swine/Hong Kong/9/98 numbering; Fig. 1) was not on the surface, but could nonetheless accommodate the additional residues. Most other insertions occurred in association with loops on the surface. According to PROCHECK, the model derived by using the A-H3 template has 73, 20, 7 and 0 % of residues in the core, allowed, generously allowed and disallowed categories, respectively. The comparable figures for the model that was derived by using the A-H9 template were 67, 31, 2 and 0 %. Furthermore, the Ramachandran plot is generally considered to be a good indicator of the quality of a folded protein structure (Wilson *et al.*, 1998); it compares the main-chain torsional angles of a predicted structure with those of known structures. The most frequent main-chain angles in the databases of known structures comprise the core, and the next most frequent comprise the allowed, categories. A model structure with a high percentage of residues in the combined core and allowed categories, and none in the disallowed category, is likely to represent a good protein fold. Thus, PROCHECK validated the folding integrity of our two models and indicated that the model structure that was derived from the A-H9 template is of higher quality in terms of the protein fold than that derived from the A-H3 template (98 and 93 % of residues are placed in the combined core and allowed categories for the H9 and H3 templates, respectively).

Homology models

Fig. 2 shows the energy-minimized predicted structure of the HA1 of B/Lee/40 based on the A-H9 template, together with the crystal structures of the influenza A (H9) virus HA1 and influenza C virus HEF1 (PDB accession code 1FLC). All of the secondary-structural elements, as well as the overall shape, of the influenza A virus HA1 were preserved in the model structure of influenza B virus HA. The comparisons are essentially the same when the crystal structure and template refer to the A-H3 subtype (data not shown).

We asked how well our models could accommodate naturally occurring variations. Fig. 3 shows an alignment of a segment of four influenza B virus HA1 sequences, representing the variation observed in a region that experiences recurring insertion and deletion events (McCullers *et al.*, 1999), together with details of our model structure based on the A-H9 template in this region, showing a loop with fixed ends that varies in length from 1 to 4 aa, thus easily accommodating the observed variations. The model derived from the A-H3 template predicted the same structural behaviour.

Conserved residues and their roles in folding stability

Of the 21 conserved residues in the A/B/C alignment, five (24 %) were Gly, six (29 %) were Pro and five (27 %) were Cys (see Table 1). The occurrences of Gly, Pro and Cys among the conserved residues were substantially more frequent than their occurrences overall in the HA1 of B/Lee/40 (9·4, 6·9 and 3·0 %, respectively).

All five conserved Cys residues in the A/B/C alignment (Fig. 1) are involved in disulphide bridges of the influenza A virus (H9) HA and influenza C virus HEF structures (Rosenthal *et al.*, 1998; Ha *et al.*, 2001), as well as those of the modelled influenza B virus HA1 structure. Four of the Cys residues are involved in disulphide bridges within the HA1 or HEF1 polypeptide chain, whereas the fifth forms a bridge with the other polypeptide chain of HA. An additional two Cys residues were conserved between the A and B sequences of this alignment (Fig. 1). They form a disulphide bridge between Cys-80 and Cys-92 in the HA1 of A/swine/Hong Kong/9/98 (A/swine/Hong Kong/9/98 numbering) (Ha *et al.*, 2001) and aligned with Cys-75 and Cys-87 (B/Lee/40 HA1 numbering) to form a disulphide

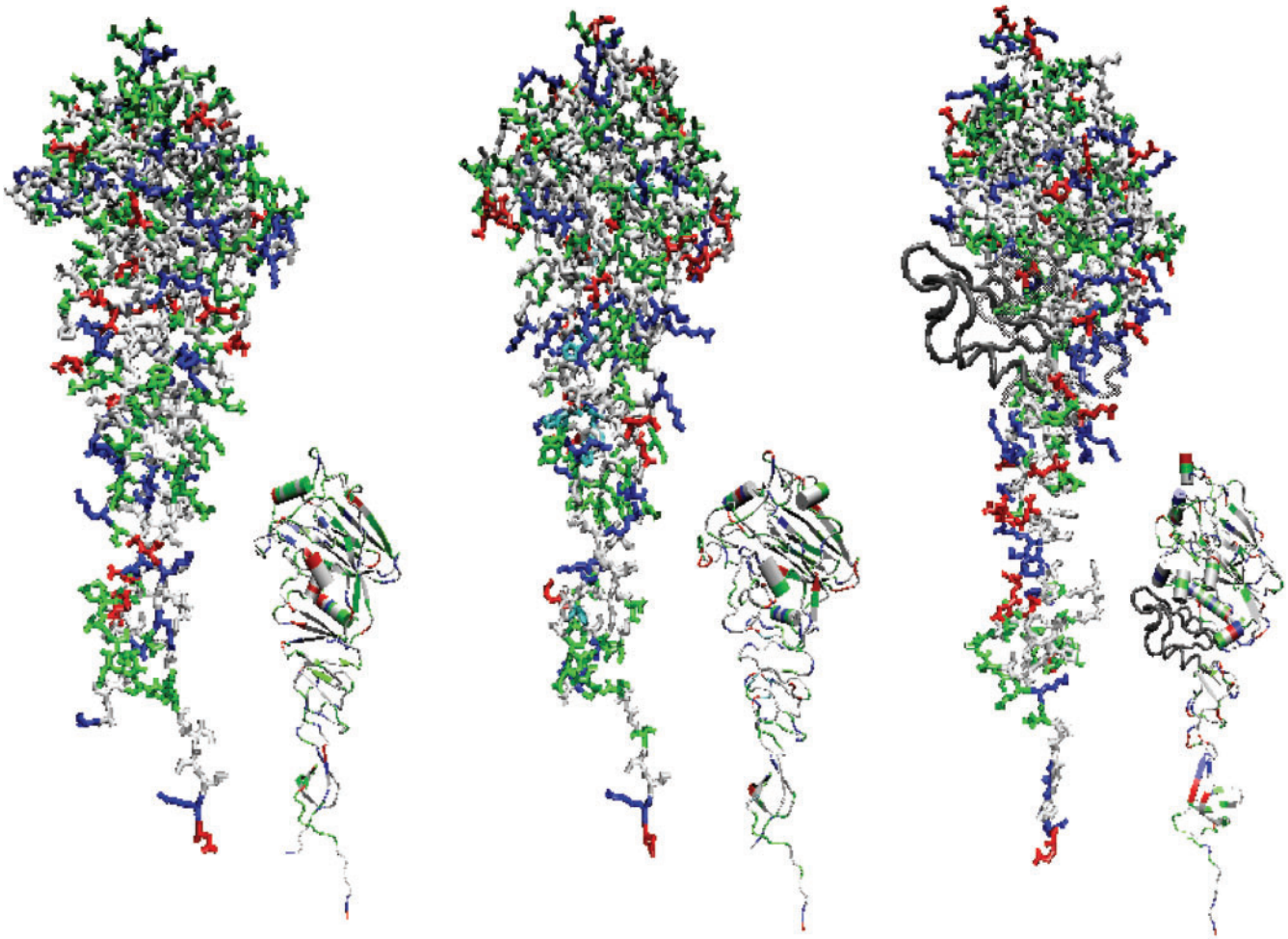

Fig. 2. Energy-minimized predicted structure of the HA1 polypeptide chain of B/Lee/40 (centre) is plotted with the crystal structure of the HA1 of A/swine/Hong Kong/9/98 (left) and the crystal structure of the HEF1 of C/Johannesburg/1/66 (right). All three molecules were aligned to show the same view. The smaller images alongside each of the molecules show the secondary structures. The part of HEF1 that corresponds to esterase activity is drawn as a grey tube. This figure was generated by using VMD with the ResName colouring option.

bridge in our structural model. A matching bridge is not observed in the crystal structure of the HEF1 of C/Johannesburg/1/66. The observed disulphide bridge between Cys-68 and Cys-293 in the HA1 of A/swine/Hong Kong/9/98 (A/swine/Hong Kong/9/98 numbering) did not have analogues in either the modelled influenza B virus HA1 structure or the crystal structure of influenza C virus HEF1. It is well-known that disulphide bridges serve the critical function of stabilizing protein folds and, hence, the predominance of conserved Cys residues in these functionally related proteins is not surprising.

The high rate of conservation of Gly and Pro residues was interesting. The small side chains of Gly and the main-chain ring structures of Pro enable these two residue types to take specialized roles in protein folding. Gly and Pro are 'special' amino acids (Tramontano, 1998) that play important roles in the protein fold. Indeed, in both the crystal structure of A(H9)-HA1 and the modelled structure of B-HA1, three out of 11 of these conserved Gly and Pro residues are involved in forming β-turns (i.e. four consecutive residues that are not involved in a helical conformation and $C_\alpha(i)$ to $C_\alpha(i+3)$ distance no larger than 7 Å; Cai *et al.*, 1999). Three additional conserved Pro and Gly residues are involved in both the crystal structure of A-HA1 and the B-HA1 model in borderline β-turns [i.e. $C_\alpha(i)$ to $C_\alpha(i+3)$ distance <8 Å]. The formation of a β-turn is a crucial step in protein folding and an important motif in protein stability (Zimmerman & Scheraga, 1977; Fasman, 1989; Fetrow *et al.*, 1998).

Taken together, the prevalence of Cys, Pro and Gly residues among the conserved residues of our A/B/C alignment showed that conserved sites were typically involved in the folding stability of the HA protein.

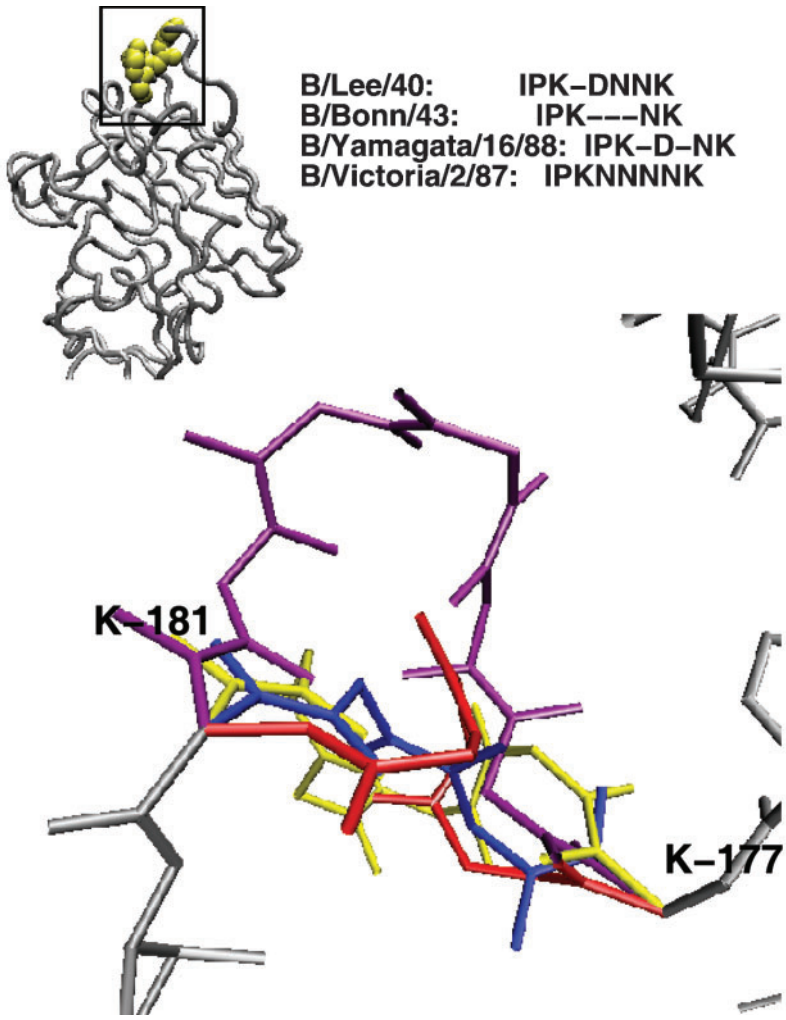

Fig. 3. An alignment of a segment (between K-177 and K-181 in B/Lee/40 numbering) of four influenza B virus HA1 sequences, chosen to represent the range of variation observed in a region that experiences recurring insertion and deletion events. The modelled loop structures (for clarity, only main-chain atoms are shown) derived from the four influenza B virus HA1 sequences are plotted in blue, red, yellow and purple for B/Bonn/43, B/Yamagata/16/88, B/Lee/40 and B/Victoria/2/87, respectively. The relationship of the region of interest (shown in the box) to the rest of the molecule is displayed in the upper left corner of the figure.

Predicted model functionality

The three-dimensional structure of the HA of A/Aichi/2/68 complexed with the natural receptor analogue sialyllactose has been solved (Weis *et al.*, 1988). Matrosovich *et al.* (1993) inferred the receptor-binding site of B/Lee/40 from an alignment of influenza A (H1 and H3) virus and B virus HA1 proteins. Their alignment required two 2 aa insertions into the left-hand edge and a 4 aa insertion adjacent to the right-hand edge of the receptor-binding site of A/Aichi/2/68 HA. In contrast, our A/B/C alignment required only one 2 aa insertion into the left-hand edge.

To study the receptor binding of our model, we used the structure of sialyl-2-3-lactose in its complex with the HA1 of A/Aichi/2/68 (PDB accession code 1HGG) and examined the contacts that this sialic acid structure would make with our modelled influenza B virus HA1. The results are shown in Fig. 4 for the model based on the A-H9 template. Both the crystal structure of HA1 from A/Aichi/2/68 and the modelled structure of HA1 from B/Lee/40 contain a deep binding pocket (see the top panel of Fig. 4). Whilst the shapes of these two pockets are not identical, they each show a number of points of contact with the target molecule (sialic acid) and, hence, we predict that the modelled HA1 of B/Lee/40 would bind with an affinity approximating that of the binding of A-HA1 to its target molecule. Free-energy calculations can provide an estimate of binding energy between two interacting molecules, but such calculations are beyond the scope of this work.

Table 2 lists residues that are associated with mutational escape from mAbs, together with their side-chain percentage exposures, based on our modelled B-HA1 structures. Although there are clearly some local differences between the two models, overall they give similar results, illustrating their robustness to choice of template. In our preferred model (that derived from the A-H9 template), almost all of the residues at which mutations occur are readily accessible to the solvent, as evidenced by a minimum of 30 % exposure. The exception is the Gly residue at position 156. However, Gly has no side chain and therefore calculations of side-chain percentage exposure would not seem to be relevant.

Table 3 lists residues of B-HA that are potential Asn-linked glycosylation sites for one or more influenza B viruses in the

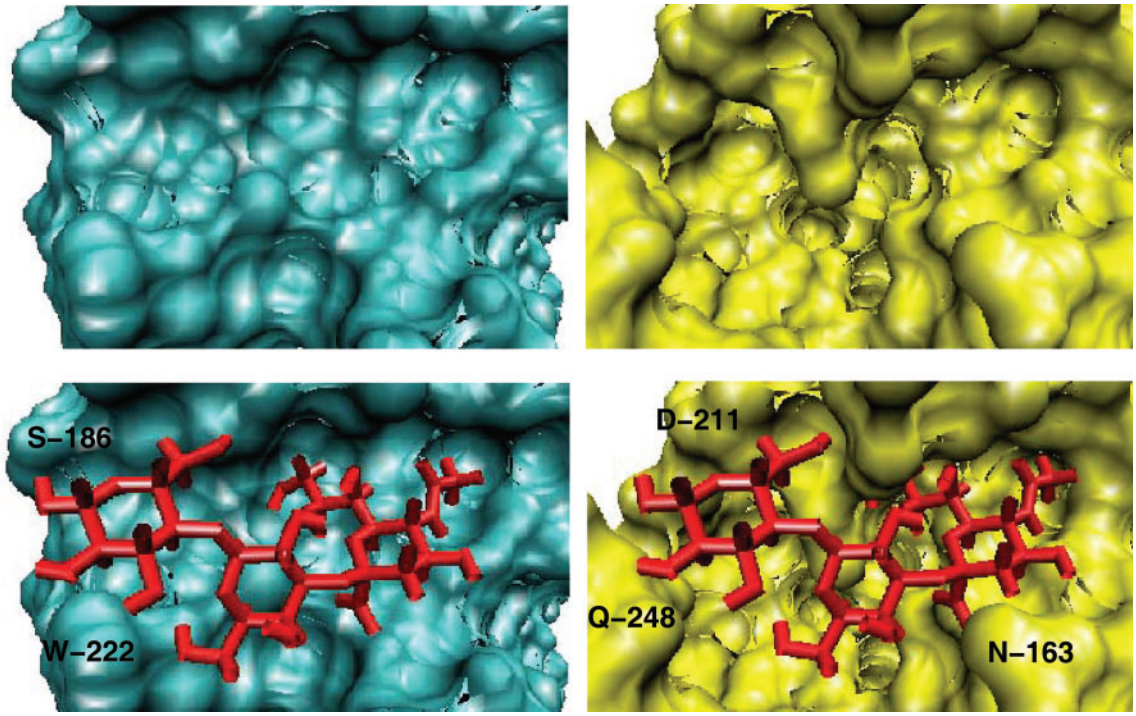

Fig. 4. The top panel shows the receptor-binding region of the crystal structure of A/Aichi/2/68 HA1 (plotted in cyan) and the modelled structure of B/Lee/40 HA1 (plotted in yellow). The lower panel shows the crystal structure of A/Aichi/2/68 HA1 bound to sialyl-2-3-lactose and the modelled structure with the structure of the sialic acid analogue superimposed. Residues S-186 and W-222 of A/Aichi/68 HA1, as well as residues N-163, D-211 and Q-248 of B/Lee/40 HA1, are labelled.

Table 2. Surface exposure (%) in the predicted influenza B virus HA1 structure of sites of escape from mAbs

Position*	Exposure (%)†		Escape mutation‡	Variation in wild-type§
	H3 template	H9 template		
156	76·3	15·0	G156V	G completely conserved
162	97·1	85·7	T162A	T completely conserved
165	66·9	42·2	N165S	S, E, K, N, T
177	48·8	32·8	K177I	K, R
177/178	90·5	54·4	Insert N	Insert N
181	62·7	81·7	K181I/N	K completely conserved
211	30·8	56·5	N211D/K	S, N, D
212	90·2	51·2	E212K	K, E
214	21·0	41·9	Q214K	Q completely conserved
216	57·6	40·9	V216L/E/A	E, G, V, K, A
217	81·7	71·6	K217G/R/T/N	R, I, T, K, N
222	83·0	72·0	S222P/L	S completely conserved
255	72·2	68·3	P255T/Q	K, P, Q

*B/Lee/40 numbering.

†Relative surface accessibility as calculated by NACCESS.

‡Compilation from Berton *et al.* (1984), Hovanec & Air (1984) and Rivera *et al.* (1995) of sites of escape from mAbs in the HA1 molecule.

§Observed site-specific variation in wild-type viruses.

Table 3. Surface exposure (%) in the predicted influenza B virus HA1 structure of potential Asn-linked glycosylation sites

Position*	Exposure (%)†		Comment‡
	H3 template	H9 template	
40	34·5	34·5	599
74	58·6	30·7	599
160	63·9	93·1	599
180	84·7	37·7	571
182	85·8	61·6	28
211	30·8	56·5	417
247	107·4	75·9	189
318	71·7	54·4	599
347	68·7	84·7	599

*B/Lee/40 numbering.

†Relative surface accessibility as calculated by NACCESS.

‡No. influenza B viruses with potential glycosylation site in HA1 (out of 599 in sample).

database. Again, the two models give similar results, illustrating their robustness to choice of template. In our preferred model, all residues in the table were exposed substantially and therefore available for the attachment of a carbohydrate moiety.

Characterization of evolutionary changes in influenza B virus sublineages

Although influenza B virus HA evolves more slowly than influenza A virus HA, some sites do exhibit distinct patterns of variation, which we investigated with our models. We examined positions 83 and 269, which are highlighted in Table 1. We examined two additional sites, 71 and 144, which were selected for their relatively high variability.

The amino acid at position 269 is a signature of the sublineages 'Yamagata' and 'Victoria', into which the influenza B virus HA bifurcated in 1976 (Hay *et al.*, 2001). Residue 269 is Pro in the Yamagata sublineage, unchanged since B/Lee/40, but Ser in all viruses of the Victoria sublineage. The change from Pro to Ser is a non-conservative change, involving charge distributions (neutral to polar) and shapes of the amino acids. We asked why the amino acid at position 269 showed little variation, as a single nucleotide mutation in Pro can lead to seven different amino acids, namely Ala, Ser, Thr, Leu, His, Gln and Arg. Fig. 5 shows residue 269 together with residues in close proximity to 269 in our model. Thr is significantly larger than Pro and would interfere with Gly-198 and Glu-199 unless the local structure is disturbed. Hence, mutations from Pro to Thr or to the larger Leu, His, Gln or Arg were not favourable. In contrast, both Ser and Ala are somewhat smaller than Pro. Therefore, Pro-269-Ser and Pro-269-Gly mutations will not cause any significant alterations in the general fold of the protein. Both Pro-269-Ser and Pro-269-Ala mutations resulted in the loss of some favourable contacts, due to their smaller sizes. However, the Pro-269-Ser mutation enabled a hydrogen bond with Glu-197 to stabilize the local fold; this bond was not possible with the Pro-269-Ala mutation. Hence, our model provides simple reasons, based on considerations of contacts and hydrogen bonds, for only two amino acid types at the variable position 269.

Position 83 showed a pattern of distinct but limited variation, similar to that at position 269; we wondered whether similar considerations could explain the variation at each position. Position 83 was completely buried. Although a single mutation in the Gly of B/Lee/40 can lead to eight different residues, substitution of Ala (one of the eight options) was predicted to cause minimum disturbance to the surrounding structure. Indeed, we observed only Ala or Gly residues in this position. Both B/Victoria/2/87 and B/Yamagata/16/88 have an Ala at position 83 and, hence, the variation must reflect effects other than phylogeny.

Positions 71 and 144 were selected for additional study as examples of sites at which changes occurred that appeared to confer a selective advantage, as the new amino acid quickly predominated in the majority of viruses of a sublineage in which the change occurred. Both of these positions had residues in close proximity (within 9 Å) and were not in a loop. Our model structure showed clear evidence of size-compensating changes (i.e. covariation) between position 71 and a neighbouring site; an Asn to Thr change (decrease in size) at position 71 is compensated by a Thr to Ile change (increase in size) at position 90, 7·8 Å away.

Position 144 interacted with the residues in the loop starting at position 177 that contained the pattern of insertions and deletions that was noted by McCullers *et al.* (1999). The C_α–C_α distance between residues 144 and 177 is 5·0 Å in our model. The residue types at position 144 can be non-polar (Ala, Met, Trp), polar (Gly, Thr, Tyr), acidic (Glu) or basic (Lys, Arg). Whilst the changes in the loop involved insertions, mutations and deletions, an overwhelming 85 % of 500 sequences that are currently in the ISD had an overall neutral charge when the charge at positions 144 and the charge in the loop were summed. It seems likely that charge compensation is an important factor in the variations observed in the loop at position 177.

DISCUSSION

Detailed structural information is a great asset in deciphering evolutionary signals. Unfortunately, some molecules, particularly surface glycoproteins, resist crystallization, thus making high-resolution structural determination very difficult. Molecular modelling is a promising alternative and homology modelling offers potential for predicting the structures of polypeptides of the size typically produced by viruses.

We used the known structures of H3 and H9 subtypes of influenza A virus HA to predict, by using homology

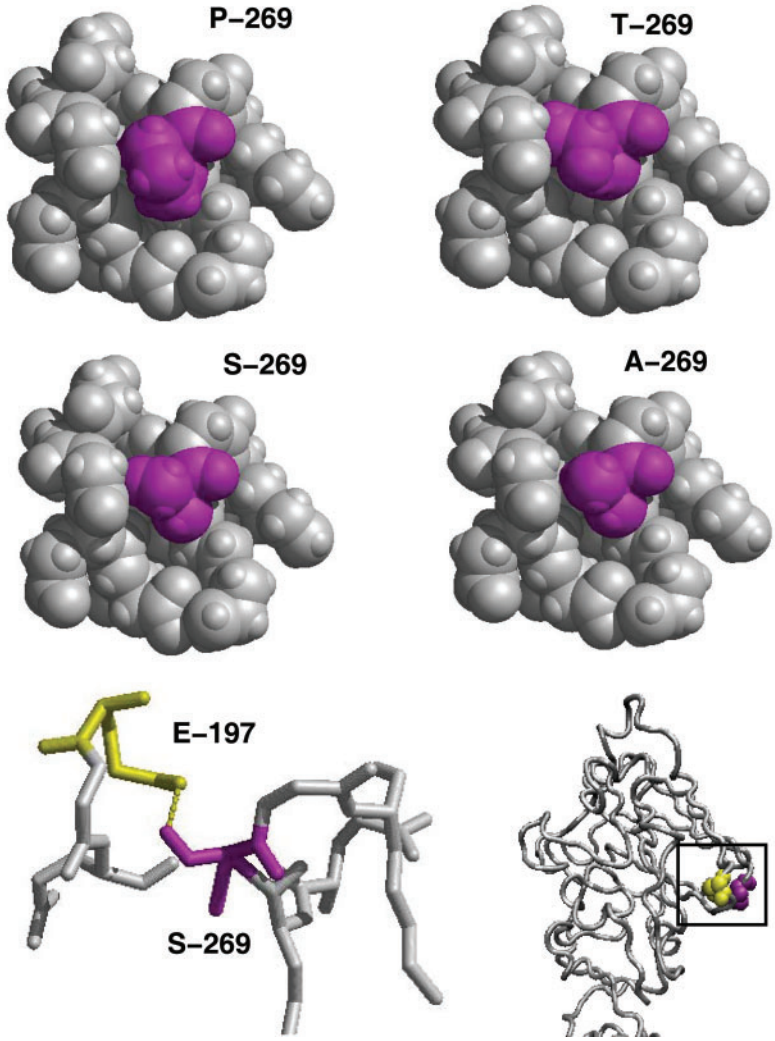

Fig. 5. Model structure in the vicinity of residue 269 (depicted in magenta) with residue 269 being Pro, Thr, Ser and Ala. Residue Ser-269 forms a hydrogen bond (dotted yellow) with residue Glu-197 (solid yellow) to stabilize the structure. The region surrounding residue 269 is boxed and shown in the lower right corner of the figure.

modelling, two structures of influenza B virus HA1. These homology models depend critically on our alignment of highly diverged HA1 and HEF1 sequences from influenza A, B and C viruses. We assessed the quality of our three-way alignment in two ways: statistically, by looking at conservation of aligned sites, and structurally, by observing the effects of insertions and deletions on secondary structure.

We examined our models to ensure that they supported an observed, persistent pattern of insertions and deletions and several essential functionalities of the HA molecule. Differences between the two predicted structures were insignificant for the purposes of this study, although by the criteria described here, the structure predicted from the H9 subtype of A-HA was slightly better.

Whilst the accuracy of a structure predicted by homology modelling can approach that of a medium-resolution structure determined experimentally (Martí-Renom *et al.*, 2000) and our B/Lee/40 HA1 models are consistent with the functionality of the molecule, some of the structural details (e.g. regions involving deletions/insertions) remain to be tested experimentally. Although supported strongly for the uses described here, the predicted structures are not sufficiently accurate to study specific molecular interactions (i.e. number of hydrogen bonds, hydrophobic patches etc.). Nevertheless, our models are likely to be a valuable tool for studying many functional aspects of the molecule, based on its structure.

In influenza A virus, amino acid variation at hypervariable sites in the HA molecule is attributed largely to evolution under pressure to escape the immune system. However, analysis of varying sites in the head region of our models of influenza B virus HA1 and the crystal structures of influenza A virus HA1 suggests that excess amino acid variation in these molecules is by no means restricted to surface positions. Whilst it was largely true that buried (<20 % surface-accessible) positions in the head region of influenza B virus HA1 varied little, above this level of accessibility, there was no apparent relationship between degree of exposure and variability.

Our homology models allowed us to examine structural

details at two of the more variable sites of influenza B virus HA1. By looking in the immediate neighbourhood of each of these sites, we found clear evidence of size- and charge-compensating co-variation. The case of charge compensation involved a loop in which a recurring pattern of insertions, mutations and deletions has been observed over the years. Knowledge of the compensatory changes to maintain neutral charge may lead to insight into possible functions of this region of the influenza B virus HA molecule.

The presence of co-varying sites has implications for statistical analysis of sequence data. Inferences of antigenicity based on variability alone may lead to erroneous interpretation of the antigenic sites of the influenza B virus HA. Careful examination of a crystal or model structure can help in deeper understanding of the capacity for variation in the influenza virus HA.

ACKNOWLEDGEMENTS

We are grateful to J. Skehel for providing the original, structurally informed alignment of influenza A virus HA and influenza C virus HEF protein sequences. We appreciate the helpful comments of Dr J. McCauley and two anonymous referees. The co-ordinates of the model of the HA1 of B/Lee/40, based on the A-HA1 (subtype H9) template, are available in the Protein Database (accession code 1TX1).

REFERENCES

Berton, M. T. & Webster, R. G. (1985). The antigenic structure of the influenza B virus hemagglutinin: operational and topological mapping with monoclonal antibodies. *Virology* **143**, 583–594.

Berton, M. T., Naeve, C. W. & Webster, R. G. (1984). Antigenic structure of the influenza B virus haemagglutinin: nucleotide sequence analysis of antigenic variants selected with monoclonal antibodies. *J Virol* **52**, 919–927.

Bonneau, R. & Baker, D. (2001). Ab initio protein structure prediction: progress and prospects. *Annu Rev Biophys Biomol Struct* **30**, 173–189.

Cai, Y.-D., Li, Y.-X. & Chou, K.-C. (1999). Classification and prediction of β-turn types by neural network. *Adv Eng Softw* **30**, 347–352.

Fasman, G. D. (1989). Protein conformational prediction. *Trends Biochem Sci* **14**, 295–299.

Fetrow, J. S., Spitzer, J. S., Gilden, B. M., Mellender, S. J., Begley, T. J., Haas, B. J. & Boose, T. L. (1998). Structure, function, and temperature sensitivity of directed, random mutants at proline 76 and glycine 77 in Ω-loop D of yeast iso-1-cytochrome *c*. *Biochemistry* **37**, 2477–2487.

Gamblin, S. J., Haire, L. F., Russell, R. J. & 9 other authors (2004). The structure and receptor binding properties of the 1918 influenza hemagglutinin. *Science* **303**, 1838–1842.

Ha, Y., Stevens, D. J., Skehel, J. J. & Wiley, D. C. (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.

Hay, A. J., Gregory, V., Douglas, A. R. & Lin, Y. P. (2001). The evolution of human influenza viruses. *Philos Trans R Soc Lond B Biol Sci* **356**, 1861–1870.

Hovanec, D. L. & Air, G. M. (1984). Antigenic structure of the hemagglutinin of influenza virus B/Hong Kong/8/73 as determined from gene sequence analysis of variants selected with monoclonal antibodies. *Virology* **139**, 384–392.

Huang, E. S., Samudrala, R. & Ponder, J. W. (1999). *Ab initio* fold prediction of small helical proteins using distance geometry and knowledge-based scoring functions. *J Mol Biol* **290**, 267–281.

Hubbard, S. J. & Thornton, J. M. (1993). NACCESS: atomic solvent accessible area calculations. Department of Biochemistry and Molecular Biology, University College London, UK (http://wolf.bms.umist.ac.uk/naccess/).

Krystal, M., Elliott, R. M., Benz, E. W., Jr, Young, J. F. & Palese, P. (1982). Evolution of influenza A and B viruses: conservation of structural features in the hemagglutinin genes. *Proc Natl Acad Sci U S A* **79**, 4800–4804.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Crystallogr* **26**, 283–291.

Macken, C., Lu, H., Goodman, J. & Boykin, L. (2001). The value of a database in surveillance and vaccine selection. In *Options for the Control of Influenza IV*. Edited by A. D. M. E. Osterhaus, N. Cox & A. W. Hampson. Amsterdam: Elsevier.

Martí-Renom, M. A., Stuart, A. C., Fiser, A., Sánchez, R., Melo, F. & Šali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* **29**, 291–325.

Matrosovich, M. N., Gambaryan, A. S., Tuzikov, A. B., Byramova, N. E., Mochalova, L. V., Golbraikh, A. A., Shenderovich, M. D., Finne, J. & Bovin, N. V. (1993). Probing of the receptor-binding sites of the H1 and H3 influenza A and influenza B virus hemagglutinins by synthetic and natural sialosides. *Virology* **196**, 111–121.

McCullers, J. A., Wang, G. C., He, S. & Webster, R. G. (1999). Reassortment and insertion-deletion are strategies for the evolution of influenza B viruses in nature. *J Virol* **73**, 7343–7348.

Moult, J. (1999). Predicting protein three-dimensional structure. *Curr Opin Biotechnol* **10**, 583–588.

Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *J Mol Biol* **7**, 95–99.

Rivera, K., Thomas, H., Zhang, H., Bossart-Whitaker, P., Wei, X. & Air, G. M. (1995). Probing the structure of influenza B hemagglutinin using site-directed mutagenesis. *Virology* **206**, 787–795.

Rosenthal, P. B., Zhang, X., Formanowski, F., Fitz, W., Wong, C.-H., Meier-Ewert, H., Skehel, J. J. & Wiley, D. C. (1998). Structure of the haemagglutinin-esterase-fusion glycoprotein of influenza C virus. *Nature* **396**, 92–96.

Ryu, K., Lee, H., Kim, S., Beauchamp, J., Tung, C.-S., Isaacs, N. W., Ji, I. & Ji, T. H. (1998). Modulation of high affinity hormone binding. Human choriogonadotropin binding to the exodomain of the receptor is influenced by exoloop 2 of the receptor. *J Biol Chem* **273**, 6285–6291.

Sánchez, R. & Šali, A. (1997). Advances in comparative protein-structure modelling. *Curr Opin Struct Biol* **7**, 206–214.

Stevens, J., Corper, A. L., Basler, C. F., Taubenberger, J. K., Palese, P. & Wilson, I. A. (2004). Structure of the uncleaved human H1 hemagglutinin from the extinct 1918 influenza virus. *Science* **303**, 1866–1870.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673–4680.

Tramontano, A. (1998). Homology modeling with low sequence identity. *Methods* **14**, 293–300.

Tung, C. S. (1997). A computational approach to modeling nucleic acid hairpin structures. *Biophys J* **72**, 876–885.

Tung, C. S. (1999). Structural study of homeodomain protein-DNA complexes using a homology modeling approach. *J Biomol Struct Dyn* **17**, 347–354.

Weiner, S. J., Kollman, P. A., Nguyen, D. T. & Case, D. A. (1986). An all atom force field for simulations of proteins and nucleic acids. *J Comput Chem* **7**, 230–252.

Weis, W., Brown, J. H., Cusack, S., Paulson, J. C., Skehel, J. J. & Wiley, D. C. (1988). Structure of the influenza virus haemagglutinin complexed with its receptor, sialic acid. *Nature* **333**, 426–431.

Wilson, I. A. & Cox, N. J. (1990). Structural basis of immune recognition of influenza virus hemagglutinin. *Annu Rev Immunol* **8**, 737–771.

Wilson, I. A., Skehel, J. J. & Wiley, D. C. (1981). Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 Å resolution. *Nature* **289**, 366–373.

Wilson, K. S., Butterworth, S., Dauter, Z. & 17 other authors (1998). Who checks the checkers? Four validation tools applied to eight atomic resolution structures. *J Mol Biol* **276**, 417–436.

Zimmerman, S. S. & Scheraga, H. A. (1977). Local interactions in bends of proteins. *Proc Natl Acad Sci U S A* **74**, 4126–4129.