doi:10.1016/j.jmb.2011.04.037 J. Mol. Biol. (2011) 410, 1023–1034

Contents lists available at www.sciencedirect.com

Journal of Molecular Biology

journal homepage: http://ees.elsevier.com.jmb

Using Knowledge of Protein Structural Constraints to Predict the Evolution of HIV-1

Simon G. Williams[1], Rachit Madan[1], Matthew G. S. Norris[1], John Archer[1], Kenji Mizuguchi[2], David L. Robertson[1] and Simon C. Lovell[1]*

[1]*Faculty of Life Sciences, University of Manchester, Oxford Road, Manchester M13 9PT, UK*
[2]*National Institute of Biomedical Innovation, 7-6-8, Saito-asagi, Ibaraki, Osaka 5670085, Japan*

Received 31 January 2011; received in revised form 12 April 2011; accepted 13 April 2011

Edited by M. F. Summers

Keywords:
HIV evolution; structural constraint; single-nucleotide polymorphism; inverse protein folding

The high levels of sequence diversity and rapid rates of evolution of HIV-1 represent the main challenges for developing effective therapies. However, there are constraints imposed by the three-dimensional protein structure that affect the sequence space accessible to the evolution of HIV-1. Here, we present a strategy for predicting the set of possible amino acid replacements in HIV. Our approach is based on the identification of likely amino acid changes in the context of these structural constraints using environment-specific substitution matrices as well as considering the physical constraints imposed by local structure. Assessment of the power of various published algorithms in predicting the evolution of HIV-1 Gag P17 shows that it is possible to use these methods to make accurate predictions of the sequence diversity. Our own method, SubFit, uses knowledge of local structural constraints; it achieves similar prediction success with the best-performing methods. We also show that erroneous predictions are largely due to infrequently occurring amino acids that will probably have severe fitness costs for the protein. Future improvements; for example, incorporating covariation and immunological constraints will permit more reliable prediction of viral evolution.

Introduction

The greatest impediment to producing an effective treatment for HIV is the high rate of viral evolution. This is a consequence of HIV's high rate of mutation,[1] propensity to recombine,[2,3] high rate of viral turnover[4,5] and the actions of the immune response promoting diversifying selection.[6–8] The ability of HIV to change rapidly leads to high levels of sequence diversity both within and between infected individuals, enabling a persistent infection to be maintained despite the actions of the immune response.[9,10] A key challenge for HIV biology is, therefore, dealing with this evolutionary change.

HIV's rapid evolution represents a significant challenge for all types of HIV therapy. HIV has been shown to escape drug therapy via a multitude of mutations,[11] including the ability to acquire multiple mutations.[12] Additionally, the virus has a demonstrated ability to escape both the antibody response[13] and the cytotoxic T lymphocyte response.[14] Determining whether there are limits of HIV evolution, and determining what they might be, is therefore of prime therapeutic importance.[15]

*Corresponding author. E-mail address: simon.lovell@manchester.ac.uk.

Abbreviations used: nsSNP, nonsynonymous single-nucleotide polymorphism; MSA, multiple sequence alignment; ESST, environment-specific substitution tables; MCC, Matthews correlation coefficient; TP, true positive; FN, false negative.

Although the degree of sequence diversity is extremely high, we expect that there is an upper bound.[16] Single-nucleotide mutations within genes have the potential to alter amino acids and may ultimately disrupt the structure and function of the encoded protein molecule. Structural constraints on protein evolution have been studied across a wide range of systems. The local structural environment of amino acid residues has been shown to strongly affect not only the degree of conservation observed, but also the entire pattern of substitutions[17,18] and compatibility of amino acids in particular structural locations.[19] In addition, functional sites within the protein add evolutionary constraints.[20] These analyses have been used to predict the likely relationship between substitutions and disease.[21]

The likely effect of a given substitution can be predicted from the known characteristics of the structure and the particular side chain substituted.[22–29] Protein structural and functional constraints determine whether an amino acid altering change is neutral (of little or no fitness cost), prone to purifying selection (a high cost of fitness to the virus in terms of structure and/or function) or positively selected (immediately beneficial to viral fitness). Substitutions with a high "cost of fitness" would therefore be subject to purifying selection. Thus, the requirement to maintain the correct structure and function of the molecule imposes restrictions on which sites can change and the nature of the evolutionary change that can occur. For HIV proteins, we have previously found that the majority of sites are subject to evolutionary constraint, and this constraint derives in part from protein structure and function.[16] This is the case even for *env*, which is the most mutable of HIV-1's genes.

Many nonsynonymous single-nucleotide polymorphisms (nsSNPs) have been implicated in human disease phenotypes,[30] and several studies have attempted to predict how these changes can alter protein function. These methods use various aspects of protein structure and/or sequence to determine whether a substitution will be tolerated.

Methods utilising protein sequence information[31,32] use multiple sequence alignments (MSAs) to determine the likelihood of substitutions based on the normalised probabilities or biochemical similarity between the mutants and wild-type amino acids. Structure-based methods use information such as solvent accessibility and location of known binding sites[23] to determine the effect of a substitution, whilst neural networks[25,28] and support vector machines[27] trained on mutagenesis data have also been used. Changes to protein stability caused by nsSNPs have been demonstrated to be a major factor in disease mechanisms,[26] and many *in silico* methods have been developed to estimate the change in folding free energy upon mutation.[24,29] Although these algorithms have not been developed explicitly for the identification of nonacceptable snSNPs, they also provide a measure of how likely a substitution can be in terms of its influence on overall stability.

We propose that methods used to predict the effect of mutations on protein structure could be adapted to predict an upper bound on the diversity generated during the course of the evolution of rapidly evolving HIV-1 proteins. Here, in addition to determining the utility of such predictive methods, we also apply our own structural constraints-based method, SubFit, to determine the likely effects of nsSNPs. Our method takes an "inverse protein folding" approach; that is, it determines which sequences are compatible with a known structure.[33,34] Those residues that are compatible with a structure are those that can be substituted without compromising structural integrity to an extent that the folded state is destabilised. This can then be used to make predictions of the viable evolutionary trajectory of HIV-1.

The prediction of protein sequence evolution based on structural templates relies heavily on amino acid side-chain packing analysis and interpretation of interactions at the atomic level. Interactions between side chains are important for maintaining the stability of the protein as well as enabling correct folding of the molecule.[35,36] Substitutions that alter the amino acid side chain, particularly in the core of the protein where the structure is well packed, will have an impact on the surrounding residues. They may lead to van der Waals overlaps (see Fig. 1a for an example), or induce a conformational adjustment that may result in side-chain or main-chain movement of the protein leading to strained conformations with respect to local energy minima.[37] Substitutions could also occur that introduce voids into the well-packed core, resulting in decreased protein stability.[38,39] These potentially negative implications of amino acid substitutions will influence which residues can be accommodated at each position within a protein. Were the virus to synthesize proteins with these replacements, their presence may result in a fitness cost and, in extreme cases, possibly an unfolded nonfunctional protein. We hypothesise that such substitutions are infrequently observed in the genomes of viable viruses.

In order to understand this process, it is necessary to have accurate information in relation to the compatibility of amino acid side chains at each site within the structure. Environment-specific substitution tables (ESSTs)[17] account for the constraints of local structure on substitution probability and act to place limits on the number of residues that we predict can be accepted at each site. Such constraints have been shown to be important in determining evolutionary patterns in proteins.[40,41] Surrounding sites will also influence the likelihood of an amino acid substitution, and therefore we additionally

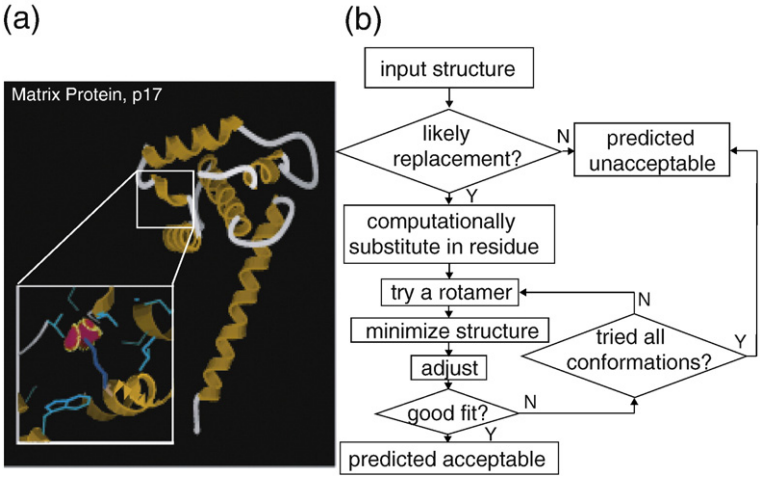

Fig. 1. Predicting HIV evolution at individual sites using SubFit. (a) Matrix protein P17's structure and illustration of the "goodness-of-fit" aspect of the method. Each site in P17 has every amino acid substituted computationally (b). In the inset, position 85 (leucine) is replaced with methionine (dark blue). In this conformation, the methionine has substantial van der Waals overlaps, as illustrated by the pink and yellow spikes, and so would be considered a constrained replacement. (b) The algorithm used for the predictive method. Proposed amino acid substitutions are first assessed for likelihood using the ESSTs. Acceptable residues are then tested for goodness of fit. All low-energy side-chain conformations ("rotamers") are tested following energy minimization, and the molecular interactions between this residue and the rest of the protein are calculated with PROBE to determine the goodness of fit at a specific site. See Methods for further details.

calculate a goodness-of-fit score for each substitution and integrate this to form our prediction of amino acid substitution (Fig. 1).

As a proof of concept we apply our strategy to HIV-1's matrix protein, P17. P17 is processed from the Gag polyprotein and is associated with the inner viral membrane, having a number of functions essential for viral assembly and entry into the infected cell.[42] We have chosen this protein primarily because of its known immunogenic importance.[43,44] Using our approach, we demonstrate that amino acid usage at individual sites is significantly constrained. Moreover, our method accurately predicts real viral diversity at the majority of sites in P17. We benchmark our method in a quantitative manner against a selection of published methods that use a variety of criteria to assess nsSNP viability and demonstrate that, to a certain extent, you can predict HIV evolution from knowledge of the constraints imposed by protein structure. Incorporation of further structural and functional information into our method has the potential to increase the accuracy of the predictions of viral evolution.

Results

Subtype B predictions

Genetic divergence between the various subtypes of HIV-1 will result in slight variations in both the sequence and the structural background of each position in the P17 proteins. Distinguishing "acceptable" from "unlikely" sequence variation will depend on the particular protein structure to which the method is applied. In addition, there is a direct relationship between the predictive power of the approach and the genetic distance between the sequence to be predicted and that of the structure used. The available P17 crystal structure [Protein Data Bank (PDB) code 1HIW][45] is from subtype B and, as such, is likely to be more accurate when used to predict the likely sequence variation in other subtype B sequences.

Predictive algorithms were assessed using the subtype B sequence and/or structure. Comparison of the amino acids predicted to be acceptable with those observed in the subtype B sequence alignment demonstrates performance of the various methods (Table 1). We determine the accuracy of the method in terms of the sensitivity [true positive (TP)/true positive (TP)+false negative (FN)], specificity [true negative/true negative +false positive (FP)] and Matthews correlation coefficient[46] (MCC) of predictions. In terms of sensitivity and specificity, there appears to be a trade-off with no one method excelling in both. Relatively high levels of specificity are obtained by some methods, but in many cases this is coupled with poor sensitivity, indicating an overly conservative set of acceptable amino acid predictions. The methods with the best MCC scores are SIFT, SNAP and our method SubFit. The sensitivity achieved by these methods range from 0.38 in SNAP to 0.65 in SubFit, indicating that the level of sequence divergence in the P17 protein is much more than these methods would predict to be viable.

Predictions using algorithms based on protein stability were initially made with substitutions leading to reduced stability disallowed (Table 1). This threshold is strict and prevents the prediction of many amino acids that may be viable. We varied

Table 1. Breakdown of predictions from each method for subtypes B and C

	Subtype B							Subtype C						
Method	TP	TN	FP	FN	SN	SP	MCC	TP	TN	FP	FN	SN	SP	MCC
SubFit	496	1220	218	266	0.65	0.84	0.51	374	1370	219	233	0.62	0.86	0.48
FoldX[a]	365	1034	424	397	0.48	0.71	0.18	278	1208	341	373	0.43	0.78	0.21
PoPMuSiC[a]	202	1337	121	560	0.27	0.92	0.25	189	1427	122	462	0.29	0.92	0.28
MUpro[a]	158	1334	124	604	0.21	0.91	0.17	145	1414	135	506	0.22	0.91	0.19
SDM[a]	596	482	953	169	0.78	0.34	0.11	628	444	1033	95	0.87	0.30	0.18
Contact-dependent[a]	274	1109	329	488	0.35	0.77	0.13	112	1538	11	539	0.17	0.99	0.33
Align-GVGD	463	969	466	302	0.61	0.68	0.27	542	999	478	181	0.75	0.68	0.40
PolyPhen	477	989	446	288	0.62	0.69	0.30	471	975	512	242	0.66	0.66	0.30
PMUT	322	1306	129	443	0.42	0.91	0.39	342	1292	441	125	0.73	0.75	0.41
SNAP	288	1375	60	477	0.38	0.96	0.44	437	1244	233	286	0.60	0.84	0.46
SIFT	395	1269	165	371	0.51	0.88	0.44	403	1330	148	319	0.56	0.90	0.50

[a] Stability-based methods: predictions made with destabilizing substitutions disallowed. This strict threshold is explored further in the Discussion and Supplementary Figs. 1 and 2.

this threshold for FoldX, PoPMuSiC and MUpro (Supplementary Fig. 1) as well as the contact-dependent method (Supplementary Fig. 2) to demonstrate the prediction success over a range of stability thresholds.

Subtype C predictions

SubFit is amongst the best-performing methods benchmarked (Table 1). In order to assess how well the methods perform in absolute terms, we have further characterised SubFit as an example of an accurately performing methodology. Following the subtype B predictions, we attempted to predict the sequence variation in subtype C of HIV-1 using SubFit with the subtype B reference structure (PDB code 1HIW). As the sequence differences between the subtype B structure and the subtype C protein being predicted increased, the number of sites incorrectly predicted also increased (Fig. 2). This demonstrates that increased divergence at specific sites alters the local structural environment and ultimately affects the sequence variation that can be accommodated. Therefore, to increase the accuracy of the bounding sequence variation in other subtypes, protein structures that are closely related to the sequence in question would have to be determined. To alleviate the subtype-specific limitations of the protein structure, we created an energy-minimized subtype C structural model using the subtype B structure as a template (see Methods). We repeated the analysis and found that using this subtype C structure resulted in significantly fewer sites not predicted ($P<0.0001$, t test) (Fig. 2). In the same way, the method performed worse when the modeled subtype C structure was used to predict subtype B sequence variation. When using the subtype C sequence and/or structure as input for the predictive methods, we obtain overall MCC scores similar to those achieved for subtype B.

Prediction assessment

Using SubFit, we find that 106 positions out of 110 are predicted to change, that is they exhibit more than one residue per site (Fig. 3 and Supplementary Table 1). The mean number of predicted amino acids per site is 6, with the maximum being 13. To determine the accuracy of the predictions, we compare them to the number of residues observed at each site in the subtype B sequence alignment

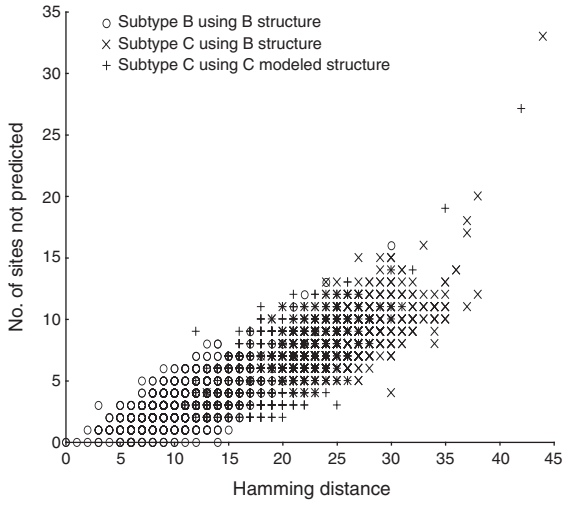

Fig. 2. The relationship between genetic distance and accuracy of the method. The genetic distance between the subtype B protein sequence used and all other sequences in the subtype B and subtype C alignments are represented by Hamming scores. As the genetic distance between sequences increases, the number of sites not predicted increases. Predictions for subtype B sequences (circles) are shown to be more accurate than those for subtype C (×) when using the subtype B structure. However, when the subtype C modeled structure is used, reducing the genetic distance to subtype C sequences, the accuracy of predictions is improved (+).

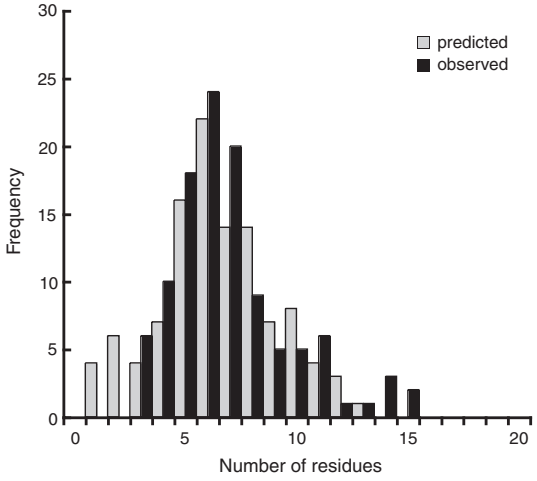

Fig. 3. Bar chart depicting the frequencies of residues predicted (grey) and observed (black) at sites in P17 subtype B by SubFit. The observed data are taken from the alignment of 2128 unique HIV-1 subtype B sequences.

(Fig. 3). Of the 110 sites in the alignment, none contained an amino acid conserved in 100% of the sequences, with the mean number of amino acids per site being 7 and the maximum 15. Comparing the distribution of these observed residues to those predicted (Fig. 3), we find that they are not significantly different ($P=0.6$, Mann–Whitney test); however, there is a tendency for the real data to include more amino acids than predicted (Figs. 3 and 4a). This underprediction of the sequence variation is a common feature of a number of the predictive methods used and is evident as low-sensitivity scores (Table 1) for both subtypes.

To investigate the statistical significance of our predictions, we assessed whether we correctly predict the residues at a site as compared with 10,000 random predictions (Supplementary Table 2). Despite the clear underprediction at the majority of sites (dark bars relative to light bars in Fig. 4a), sensitivity is greater than in random simulation per site ($P<0.0001$) (Supplementary Table 2). These values were statistically significant at 55% of sites ($P<0.05$, indicated above each site with white squares in Fig. 4a). The specificity per site (dark bars relative to light bars in Fig. 4b) was also significant over random simulation ($P<0.0001$), with 54% of sites being statistically significant ($P<0.05$, indicated above each site with white squares in Fig. 4b).

Underprediction and the subsequent occurrence of FNs may be either methodological (i.e., due to simplifying assumptions in developing the method) or, alternatively, biological (i.e., due to features of the biological systems upon which we are making predictions). The main simplification is that we treat all sites independently. This could clearly lead to underprediction of diversity at some positions due to covariation between neighbouring sites, accommodating mutations that are not allowed by our method (see Discussion). An interesting feature of HIV-1 biology is the common generation of sequences that represent evolutionary dead ends and make no contribution to the ongoing infection within an individual.[47] It is possible, therefore, that some sequences contain amino acids that may be considered irrelevant, as they would be so detrimental to the structure that viral fitness would be significantly reduced. If this is the case, then the FNs should occur at significantly lower frequencies than

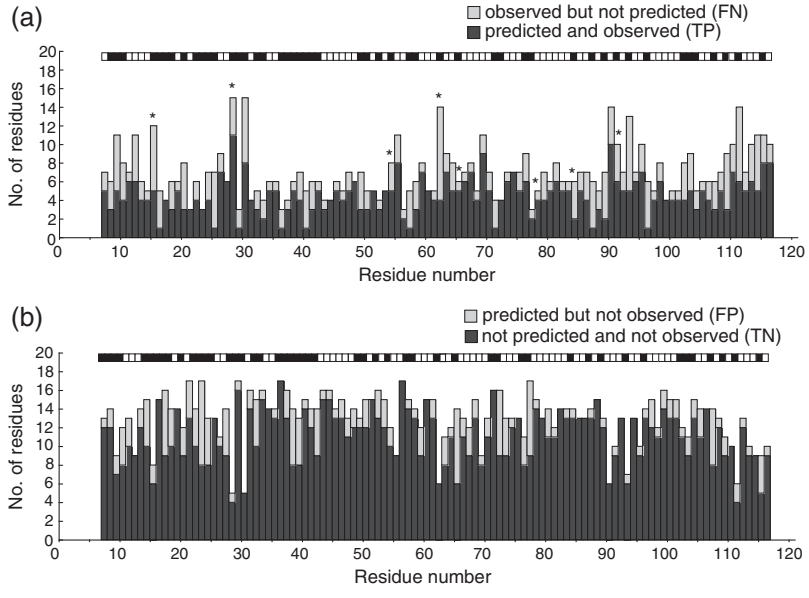

Fig. 4. (a) The number of residues observed (total bar height). The proportion of these that are predicted by SubFit and observed (TP, black) and the proportion observed but not predicted (FN, light grey) are shown. Stars indicate sites inferred to be under the influence of positive selection.[7] (b) The number of residues not observed in the alignment (total bar height). The number of these not predicted and not observed (TN, black) and the number predicted but not observed (FP, light grey) are shown. In (a) and (b), the white squares denote sites with predictions significantly different from random ($P<0.05$); black boxes are insignificant.

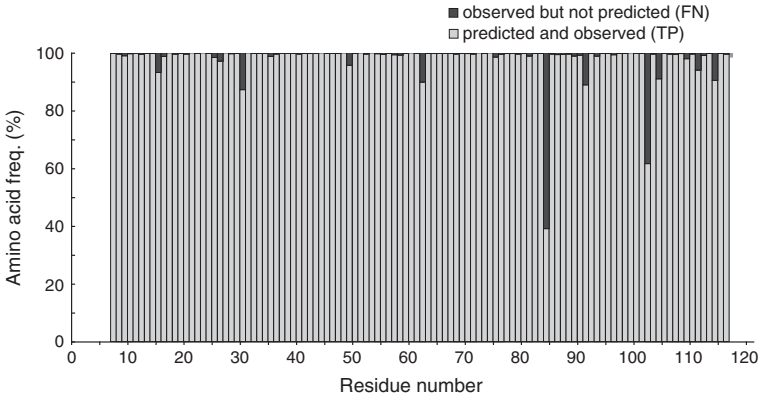

Fig. 5. The summed amino acid frequencies of all predicted by SubFit and observed residues (TP, light bars) *versus* all observed but not predicted by SubFit (FN, dark bars) at each site in P17.

those that are correctly predicted (TPs). We find this is the case with FNs occurring at significantly lower frequencies, 2% on average per site *versus* 98% for the TPs (Fig. 5; $P<0.0001$, Wilcoxon test). Indeed, of 554 amino acids that occur at a frequency of <1% in the subtype B alignment, 44% fall into the FN category and were found to have a greater predicted destabilizing effect in FoldX (average $\Delta\Delta G$ of 1.1 compared to 0.51 for low-frequency TPs; ($P=0.0044$), indicating that they are likely to be substitutions with large structural consequences for the protein.

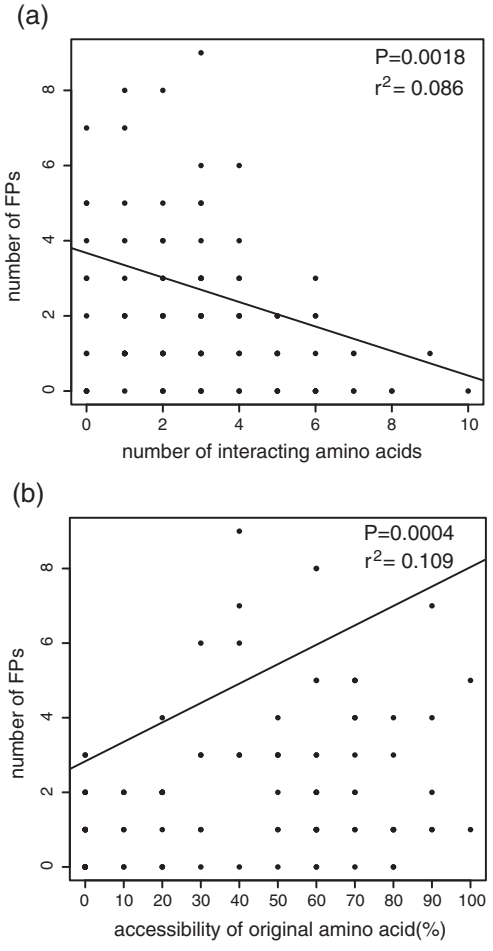

Fig. 6. (a) The relationship between the number of same-chain interacting residues and number of FPs. (b) The relationship between the solvent accessibility per site and number of FPs.

In addition, underprediction at sites can be attributed to the action of positive selection. Of eight such sites previously identified in the P17 region,[7] four correspond significantly to sites with high numbers of observed amino acids (Fig. 4a; $P=0.0019$, Wilcoxon test) and occurrence of FNs. At these sites, sequence variation is high, and it is possible that more structurally disrupting amino acids are being sampled in an attempt to escape the immune response. At only one site (position 84) did our approach fail to predict the consensus residue (Fig. 5). This rare erroneous prediction arises from the consensus amino acid failing the ESST likelihood test. Other erroneous predictions can come about by failure of the goodness-of-fit test, presumably the result of the P17 structure having been solved for one sequence only and a problem that would be resolved with the availability of more P17 structures.

Overprediction of SubFit at certain sites leads to high FPs, where the sequence variation is less varied than we predict. Analysis of the structural environments demonstrates that FPs increase with solvent accessibility of the site and at positions with fewer intrachain contact interactions (Fig. 6). These sites are on the surface of the protein where our method predicts many amino acids are likely to be accommodated. It is a proportion of these sites that are under additional constraints that are not accounted for by SubFit.

Discussion

Using algorithms for predicting the viability of nsSNPs in proteins, we have shown that it is

possible to make useful predictions of the evolutionary divergence of HIV P17 proteins. Such methods were originally designed for the analysis of energetic effects of amino acid substitutions to predict phenotypic effects. We have shown that they can also be useful for predicting the upper bounds on HIV-1 evolution, as they are essentially determining which sequences are compatible with a known structure.

The predictive algorithms tested use a variety of sequence and structure-based principles and have varying success in predicting the observed sequence variation.

SIFT,[32] an algorithm based on sequence homology, performs as well as any other method (Table 1). However, methods utilising sequence homology in this way will not necessarily identify amino acid changes with severe consequences for the protein structure but merely represent the sequence divergence that has already been observed within the family of proteins. Benchmarking the predictions using alignments of observed homologues becomes a rather circular process, and it is only due to the abundance of very low frequency amino acids in the alignment that predictions are not more accurate.

The Align-GVGD algorithm combines a positional conservation score generated from an MSA with a measure of the physicochemical differences between substituting amino acids. PolyPhen expands on this by creating a position-specific profile matrix based on an MSA and combines this with a variety of structure-based information. Features such as solvent accessibility, side-chain volume and *B*-factor of the site are considered and scored to predict the likelihood of the substitution. As with Align-GVGD, predictions are too permissive, and many substitutions are allowed that are likely to be constrained due to more complex structural interactions within the molecule. However, it should be noted that these methods are not trained on viral sequences, and their authors do not recommend them for this use.

Neural-network-based methods such as SNAP and PMUT perform comparatively well, but predictions include a high number of FNs for the subtype B data. A large number of factors are considered in their predictions including biochemical properties, local environment, chain flexibility, SWISS-PROT and Pfam information, PSI-BLAST profiles and sequence conservation. PMUT has been developed primarily with the aim of identifying disease-related SNPs in humans and potential sites for experimental mutagenesis. For these reasons, caution should be taken when this method is used to predict HIV evolution where certain substitutions may occur that push the boundaries of fitness to escape the immune response. Indeed, the high numbers of FNs included in the subtype B predictions indicates that many observed substitutions are classified as "non-neutral" or "pathological," suggesting that these are the more structurally damaging substitutions that may be detrimental to viral fitness.

Algorithms that predict the stability changes upon mutation such as FoldX.[29] MUpro,[27] PoPMuSiC[24] and SDM[48] were not designed to partition "likely" from "unlikely" amino acid substitutions and as such perform comparatively poorly when used in this manner (Table 1). In many cases, stability changes are likely to be subtle, and variations between the algorithms seem to either result in a trend of greatly overpredicting (many FPs) or underpredicting (many FNs). Here, substitutions that lead to a predicted decrease in stability, no matter how small, are deemed unacceptable. This assumption is likely to be somewhat flawed, as marginal decreases in stability can still result in correctly folded and functional proteins.[49] We attempted to allow for this by varying the thresholds for acceptable stability changes in FoldX, PoPMuSiC and MUpro (Supplementary Fig. 1). The FoldX algorithm uses an energy function using data from protein engineering experiments; PoPMuSiC uses database-derived potentials based on solvent accessibility; and MUpro uses a support vector machine to determine protein stability changes. Despite these methodological differences, results are similar. Optimal MCC scores are achieved where stability thresholds are very strict, and so while they achieved very high specificity scores, sensitivity is very low and predictions are not useful. Allowing for a small decrease in stability upon substitution (0 to 1 $\Delta\Delta G$) increases sensitivity in all of these methods and suggests that marginally destabilizing substitutions are often observed in the p17 data. Even allowing for a decrease in stability, these methods do not perform as well as others when used to predict sequence variation. There is a marked trade-off between specificity and sensitivity in this destabilizing region (Supplementary Fig. 1), indicating that there are many observed amino acids that cannot be partitioned correctly using stability alone. While many destabilizing substitutions can occur in this range, many cannot, and it is likely that other constraints will determine viability.

Interactions with surrounding amino acids will influence which substitutions are likely to occur. We implemented a contact-dependent energy-based method based on previous work by Bastolla *et al.*[50] and Rastogi *et al.*[51] Here, the energy from interactions between neighbouring side chains is calculated, and the native structure and substituted structures are compared. We found that subtype B predictions are similar to those from other stability-based methods (Supplementary Fig. 2a). Excluding a substitution that results in a decreased overall interaction energy leads to many FNs and hence poor MCC score and confirms that many of the

observed substitutions may be less stable than the native structure. Subtype C predictions appear improved at this threshold (Table 1), but sensitivity is very low. The "native" subtype C structure is itself a model that has been energy–minimized, and therefore it is unlikely that a simple model structure featuring a substitution will have lower interaction energy. This algorithm was developed for calculating optimum energy conformations for protein folding. Adapting it to predict likely sequence variation on a site-by-site basis demonstrates that while many single substitutions decrease stability, segregating amino acids based on this can lead to poor predictions. However, the derived sequences tested here are based on single substitutions that may have not occurred. In addition, modeled structures from which contact maps are determined may not represent the lowest energy conformations; this highlights the potential problems in adapting this algorithm for sequence prediction.

Our own approach to predict the evolutionary change works because there are strong structural constraints acting on viral proteins such that the amino acids that can occur at individual sites are restricted.[16] This link between sequence and structure imposes physical restrictions both on which sites will change and on the nature of the change that can occur; consequentially, structure affects the distribution of amino acids observed. As with all of the tested algorithms, there is a proportion of the observed sequence divergence that is not predicted (Fig. 3). Analysis of these FNs reveals that they are largely due to the occurrence of very low frequency amino acids in the alignments (Fig. 5) and that these substitutions are likely to have a greater destabilizing effect on the protein than low-frequency TPs. The nature of our method predicts that these residues are likely to be evolutionarily irrelevant, as their presence would result in severe fitness consequences and possibly unfolded proteins. Conversely, those amino acids correctly predicted to be present are found at much higher frequencies and represent the possible sequence and structural evolution within the limits of structural integrity.

Other erroneous predictions are FPs, where sites appear to be more conserved than we predict. Analysis of these sites reveals that there is a tendency to overpredict the variation at sites on the surface of the protein that make few same-chain interactions (Fig. 6). At these sites, there are likely to be additional constraints that restrict the sequence variation that are not accounted for by our method. HIV proteins, including P17, are known to have many interactions with human proteins.[52,53] It is therefore likely that certain sites, more likely on the surface of the protein, will be constrained due to these interactions leading to overprediction.

SubFit allows limited backbone flexibility due to the energy-minimization step following replacement of the rotameric side chains. This movement from energy minimization is small but likely to accurately represent the subtle backbone movements that occur following amino acid substitutions.[54] The importance of subtle structural variations to our method is demonstrated in Fig. 2. Subtype C predictions improve when a minimized modeled subtype C structure is used, highlighting the importance of local structural movements in determining amino acid acceptance. It has previously been suggested that approaches that combine a free-to-move backbone with energy minimization can often be far too permissive,[22] predicting that an inappropriately wide range of residues are compatible with the structure. In our case, we err on the side of conservative prediction. However, it is possible that the underprediction of diversity at some sites (Fig. 3) is due in part to this simplifying assumption.

The structural constraints-based method is conservative in that it is identifying replacements with severe fitness consequences (those having consequences for folding), whereas many mutations that will have consequences for fitness might not be considered constrained by our method. These may include the so-called "cost-of-fitness" residue replacements associated with escape mutations,[55] which, although associated with fitness decrease relative to wild-type, can have higher fitness in the context of the immune response. We have investigated replacements that have been reported to be associated with fitness in P17.[56,57] All of these fall into our predicted and observed set of amino acid residues, although at relatively low frequencies: E17K, 0.8% frequency; K26R, 6.8%; and Q28H, 0.3%. Despite our sample being small, this observation is consistent with our assertion that the method is identifying residues with dramatic consequences for folding. To check this, we also looked at the replacement, K30M, which had been associated with SIVcpz and, as a probable reversion, with subtype C.[58] Thus, we would expect this to be predicted as a constrained amino acid for HIV-1 and we find this is the case.

The inclusion of additional constraints at functional sites,[42] the requirement to maintain binding interfaces,[20] and the intra- and intermolecular interactions[18] will place further limits on the change that can occur in proteins and could be incorporated into future predictions. Additionally, more complex evolutionary models have taken tertiary structure into account to model substitutions based on site interdependence.[50,59] The quantification of these additional constraints on viral evolution will permit, with even greater accuracy than achieved here, the prediction of the evolutionary trajectory of HIV at specific sites.

Despite its current limitations, our approach based on prediction of structurally significant amino acid replacements is the first attempt to

predict the limits of viral evolution. Crucially, the high cost-of-fitness mutations associated with immune escape require compensatory mutations.[60–62] Indeed, it is the occurrence of compensatory mutation that changes an amino acid's replacement from improbable to probable. Thus, understanding in greater detail the nature of compensatory covariation in terms of both intra- and intermolecular interactions will permit the prediction of these important evolutionary changes. Ultimately, an in-depth understanding of the limits of HIV's ability to change will aid understanding of immune escape, drug resistance and design of any future vaccine.

Methods

P17 protein structures

The protein structure 1HIW was selected from the PDB[63] as being representative of HIV-1's matrix protein, P17[45] in subtype B. The subtype C model structure was generated using Modeller[64] with 1HIW as a template. The comparative model was subjected to energy minimisation using limited-memory L-BFGS minimization to a root-mean-square (RMS) gradient of 0.1 kcal mol^{-1} $Å^{-1}$ using the charmm27 force field[65] as implemented in Tinker.[66] Hydrogen atoms were added to the structure in optimum positions using REDUCE.[67]

Substitution matrices

Substitution tables were used to further filter the predicted amino acids at each position in the structure. Replacement tables were used in order to account for propensity issues such as the likelihood of buried charges and exposed hydrophobic side chains, which would not be taken into account in predictions using goodness of fit alone. ESSTs were created using alignments from the HOMSTRAD database[68] (16 August 2010). The structural environment of each amino acid position was calculated using the program JOY,[69] and structural alignments of homologous families were selected with a percent sequence identity of >85%. Sixty-four environment-specific substitution tables[17,18] were generated, utilising secondary structure (helix, sheet, coil, positive phi), solvent accessibility [accessible (>7% of its side-chain area accessible to a 1.4 Å probe) or buried] and hydrogen bonding to local side chains and main chains (eight classes). Substitution tables were calculated using the SEQSUBST program.[70,71] The percent sequence identity cutoff of >85% was applied, as this results in substitution tables that most accurately describe the level of sequence divergence observed in the P17 sequence alignment. The likelihood of each of the amino acids occurring at these structural positions was assessed using the replacement tables, and any residue with a log-odds score of greater than −11 was deemed suitable for replacement in the structure. This threshold was used as it was found to achieve optimal predictions; however, results were found to be largely insensitive to this cutoff (Supplementary Fig. 3).

Goodness of fit

Each position within the protein structure was analyzed by computational replacement of the existing side chain to all other amino acid side chains using a rotamer library.[72] In order to allow some flexibility of the backbone after each rotamer substitution, the structure was energy-minimized to an RMS gradient of 0.1 kcal mol^{-1} $Å^{-1}$ using the charmm27 force field[65] in Tinker.[66] The goodness of fit of these substituted amino acids was then assessed.[22] The side chains of each amino acid were rotated in steps of 5° around each of their χ angles to known rotamer boundaries in order to optimise efficiency and minimize computational time. After each rotation step, all-atom contact measurements were carried out using PROBE[37] to assess how well the residue would fit within the local structure of the molecule in the given conformation. PROBE uses the rolling probe algorithm[73] to recognize regions where steric clashes between atoms occur. Once this had been done for all amino acids at a particular position, those residues were selected with PROBE score greater than −1.[74] These were amino acids with at least one rotameric conformation that did not cause steric clashes within the existing local protein structure. We predict that replacements of these residues into the structure would not cause local structural disruption and are, therefore, more likely to occur than replacements requiring the local structural environment to shift in order to accommodate them.

Amino acids within the cutoffs of the ESSTs scores and goodness of fit formed our predicted set of residues for each site in the protein structure (Fig. 1b). These are amino acids that, if substituted into the protein, would cause minimal structural disruption.

Data sets

HIV-1 sequences for which a near full-length genome is available were downloaded from the LANL HIV Sequence Database†. Sequences corresponding to P17 were extracted, translated to amino acids and aligned using Muscle.[75] For subtype B, this resulted in a data set of 2128 aligned sequences after strains with identical names and nonsense mutations were removed. Sites that included more than 95% gaps were removed from all amino acid sequence alignments. Sequences as subtype C were also downloaded and processed, resulting in 1945 sequences for analysis. These alignments were used to assess how well the predictive methods matched the observed data. The overall percentage sequence identity in the alignments was 86% and 85% for subtype B and subtype C, respectively.

Parameters for other predictive methods

Contact-dependent method

Model structures of all possible substitutions at each position were generated using Modeller.[64] Contact maps were calculated with all amino acid side-chain heavy

† http://www.hiv.lanl.gov

atoms within 4.5 Å considered interacting. The effective free energy of each model structure was calculated by summing the interaction energy of all contacting residues. The interaction matrix used was derived by Bastolla *et al.*[50] and has been implemented in this way by Rastogi *et al.*[51] It is an optimized energy function based on the maximization of the thermodynamic average of the overlap between protein native structures and a Boltzmann ensemble of alternative structures. The total free energy of contact interactions (in units of k_BT) of the substituted proteins was compared to that of the native protein. Supplementary Fig. 2 shows the variation of the threshold with prediction accuracy.

Other methods

SIFT (version 4.0.3)[32] was run by submitting the sequence of the subtype B and subtype C proteins with known structure. The high-quality Uniprot–SwissProt 56.6 database was selected to search for homologous sequences to avoid potentially nonfunctional HIV proteins present in trEMBL from affecting the predictions. The PSI-BLAST generated alignment from SIFT was then used as the input for Align-GVGD.[31] Sequences were submitted to PMUT[25] and the small neural network was selected, as suggested for nonhuman mutations. The reference P17 sequence was also submitted to PolyPhen[76] and SNAP[28] using default settings.

PoPMuSiC (version 2.0)[29] was run using the systematic approach given the PDB code 1HIW for subtype B and the modeled PDB structure for subtype C. FoldX (v3.0 Beta5.1)[24] was downloaded and run with default parameters. MUpro (v1.1)[27] was downloaded and run using the regression model over all substitutions. SDM[48] was run by submitting both the wild-type PDB files and those containing the mutant amino acids.

Supplementary materials related to this article can be found online at doi:10.1016/j.jmb.2011.04.037

Acknowledgements

S.G.W. and J.A. were supported by Engineering and Physical Sciences Research Council and Biotechnology and Biological Sciences Research Council studentships, respectively.

References

1. Gao, F., Chen, Y., Levy, D. N., Conway, J. A., Kepler, T. B. & Hui, H. (2004). Unselected mutations in the human immunodeficiency virus type 1 genome are mostly nonsynonymous and often deleterious. *J. Virol.* **78**, 2426–2433.
2. Robertson, D. L., Hahn, B. H. & Sharp, P. M. (1995). Recombination in AIDS viruses. *J. Mol. Evol.* **40**, 249–259.
3. Jetzt, A. E., Yu, H., Klarmann, G. J., Ron, Y., Preston, B. D. & Dougherty, J. P. (2000). High rate of recombination throughout the human immunodeficiency virus type 1 genome. *J. Virol.* **74**, 1234–1240.
4. Ho, D. D., Neumann, A. U., Perelson, A. S., Chen, W., Leonard, J. M. & Markowitz, M. (1995). Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature*, **373**, 123–126.
5. Wei, X., Ghosh, S. K., Taylor, M. E., Johnson, V. A., Emini, E. A., Deutsch, P. *et al.* (1995). Viral dynamics in human immunodeficiency virus type 1 infection. *Nature*, **373**, 117–122.
6. Wolinsky, S. M., Korber, B. T., Neumann, A. U., Daniels, M., Kunstman, K. J., Whetsell, A. J. *et al.* (1996). Adaptive evolution of human immunodeficiency virus-type 1 during the natural course of infection. *Science*, **272**, 537–542.
7. Yang, W., Bielawski, J. P. & Yang, Z. (2003). Widespread adaptive evolution in the human immunodeficiency virus type 1 genome. *J. Mol. Evol.* **57**, 212–221.
8. Choisy, M., Woelk, C. H., Guegan, J. F. & Robertson, D. L. (2004). Comparative study of adaptive molecular evolution in different human immunodeficiency virus groups and subtypes. *J. Virol.* **78**, 1962–1970.
9. Ciurea, A., Hunziker, L., Martinic, M. M., Oxenius, A., Hengartner, H. & Zinkernagel, R. M. (2001). $CD4^+$ T-cell–epitope escape mutant virus selected in vivo. *Nat. Med.* **7**, 795–800.
10. Draenert, R., Le Gall, S., Pfafferott, K. J., Leslie, A. J., Chetty, P., Brander, C. *et al.* (2004). Immune selection for altered antigen processing leads to cytotoxic T lymphocyte escape in chronic HIV-1 infection. *J. Exp. Med.* **199**, 905–915.
11. Rhee, S. Y., Gonzales, M. J., Kantor, R., Betts, B. J., Ravela, J. & Shafer, R. W. (2003). Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.* **31**, 298–303.
12. Molla, A., Korneyeva, M., Gao, Q., Vasavanonda, S., Schipper, P. J., Mo, H. M. *et al.* (1996). Ordered accumulation of mutations in HIV protease confers resistance to ritonavir. *Nat. Med.* **2**, 760–766.
13. Wei, X., Decker, J. M., Wang, S., Hui, H., Kappes, J. C., Wu, X. *et al.* (2003). Antibody neutralization and escape by HIV-1. *Nature*, **422**, 307–312.
14. Leslie, A. J., Pfafferott, K. J., Chetty, P., Draenert, R., Addo, M. M., Feeney, M. *et al.* (2004). HIV evolution: CTL escape mutation and reversion after transmission. *Nat. Med.* **10**, 282–289.
15. McMichael, A. J. (2006). HIV vaccines. *Annu. Rev. Immunol.* **24**, 227–255.
16. Woo, J., Robertson, D. L. & Lovell, S. C. (2010). Constraints on HIV-1 diversity from protein structure. *J. Virol.* **84**, 12995–13003.
17. Overington, J., Donnelly, D., Johnson, M. S., Sali, A. & Blundell, T. L. (1992). Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci.* **1**, 216–226.
18. Overington, J., Johnson, M. S., Sali, A. & Blundell, T. L. (1990). Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc. Roy. Soc. London Ser. B*, **241**, 132–145.

19. Luthy, R., Bowie, J. U. & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature*, **356**, 83–85.
20. Chelliah, V., Chen, L., Blundell, T. L. & Lovell, S. C. (2004). Distinguishing structural and functional restraints in evolution in order to identify interaction sites. *J. Mol. Biol.* **342**, 1487–1504.
21. Burke, D. F., Worth, C. L., Priego, E. M., Cheng, T., Smink, L. J., Todd, J. A. & Blundell, T. L. (2007). Genome bioinformatic analysis of nonsynonymous SNPs. *BMC Bioinformatics*, **8**, 301.
22. Word, J. M., Bateman, R. C., Presley, B. K., Lovell, S. C. & Richardson, D. C. (2000). Exploring steric constraints on protein mutations using MAGE/PROBE. *Protein Sci.* **9**, 2251–2259.
23. Sunyaev, S., Ramensky, V., Koch, I., Lathe, W., III, Kondrashov, A. S. & Bork, P. (2001). Prediction of deleterious human alleles. *Hum. Mol. Genet.* **10**, 591–597.
24. Guerois, R., Nielsen, J. E. & Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* **320**, 369–387.
25. Ferrer-Costa, C., Orozco, M. & de la Cruz, X. (2004). Sequence-based prediction of pathological mutations. *Proteins*, **57**, 811–819.
26. Yue, P., Li, Z. & Moult, J. (2005). Loss of protein structure stability as a major causative factor in monogenic disease. *J. Mol. Biol.* **353**, 459–473.
27. Cheng, J., Randall, A. & Baldi, P. (2006). Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins*, **62**, 1125–1132.
28. Bromberg, Y., Yachdav, G. & Rost, B. (2008). SNAP predicts effect of mutations on protein function. *Bioinformatics*, **24**, 2397–2398.
29. Dehouck, Y., Grosfils, A., Folch, B., Gilis, D., Bogaerts, P. & Rooman, M. (2009). Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics*, **25**, 2537–2543.
30. Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shiel, J. A., Thomas, N. S. *et al.* (2003). Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.* **21**, 577–581.
31. Mathe, E., Olivier, M., Kato, S., Ishioka, C., Hainaut, P. & Tavtigian, S. V. (2006). Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. *Nucleic Acids Res.* **34**, 1317–1325.
32. Ng, P. C. & Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863–874.
33. Drexler, K. E. (1981). Molecular engineering: an approach to the development of general capabilities for molecular manipulation. *Proc. Natl Acad. Sci. USA*, **78**, 5275–5278.
34. Ponder, J. W. & Richards, F. M. (1987). Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193**, 775–791.
35. Dahiyat, B. I. & Mayo, S. L. (1997). Probing the role of packing specificity in protein design. *Proc. Natl Acad. Sci. USA*, **94**, 10172–10177.
36. Sandberg, W. S. & Terwilliger, T. C. (1989). Influence of interior packing and hydrophobicity on the stability of a protein. *Science*, **245**, 54–57.
37. Word, J. M., Lovell, S. C., LaBean, T. H., Taylor, H. C., Zalis, M. E., Presley, B. K. *et al.* (1999). Visualizing and quantifying molecular goodness of fit: small-probe contact dots with explicit hydrogen atoms. *J. Mol. Biol.* **285**, 1711–1733.
38. Cuff, A. L. & Martin, A. C. (2004). Analysis of void volumes in proteins and application to stability of the p53 tumour suppressor protein. *J. Mol. Biol.* **344**, 1199–1209.
39. Xu, J., Baase, W. A., Baldwin, E. & Matthews, B. W. (1998). The response of T4 lysozyme to large-to-small substitutions within the core and its relation to the hydrophobic effect. *Protein Sci.* **7**, 158–177.
40. Fornasari, M. S., Parisi, G. & Echave, J. (2002). Site-specific amino acid replacement matrices from structurally constrained protein evolution simulations. *Mol. Biol. Evol.* **19**, 352–356.
41. Parisi, G. & Echave, J. (2001). Structural constraints and emergence of sequence patterns in protein evolution. *Mol. Biol. Evol.* **18**, 750–756.
42. Fiorentini, S., Marini, E., Caracciolo, S. & Caruso, A. (2006). Functions of the HIV-1 matrix protein p17. *New Microbiol.* **29**, 1–10.
43. Iversen, A. K., Stewart-Jones, G., Learn, G. H., Christie, N., Sylvester-Hviid, C., Armitage, A. E. *et al.* (2006). Conflicting selective forces affect T cell receptor contacts in an immunodominant human immunodeficiency virus epitope. *Nat. Immunol.* **7**, 179–189.
44. Allen, T. M. & Altfeld, M. (2008). Crippling HIV one mutation at a time. *J. Exp. Med.* **205**, 1003–1007.
45. Hill, C. P., Worthylake, D., Bancroft, D. P., Christensen, A. M. & Sundquist, W. I. (1996). Crystal structures of the trimeric human immunodeficiency virus type 1 matrix protein: implications for membrane association and assembly. *Proc. Natl Acad. Sci. USA*, **93**, 3099–3104.
46. Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.
47. Shankarappa, R., Margolick, J. B., Gange, S. J., Rodrigo, A. G., Upchurch, D., Farzadegan, H. *et al.* (1999). Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J. Virol.* **73**, 10489–10502.
48. Topham, C. M., Srinivasan, N. & Blundell, T. L. (1997). Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. *Protein Eng.* **10**, 7–21.
49. Matthews, B. W. (1995). Studies on protein stability with T4 lysozyme. *Adv. Protein Chem.* **46**, 249–278.
50. Bastolla, U., Farwer, J., Knapp, E. W. & Vendruscolo, M. (2001). How to guarantee optimal stability for most representative structures in the Protein Data Bank. *Proteins*, **44**, 79–96.
51. Rastogi, S., Reuter, N. & Liberles, D. A. (2006). Evaluation of models for the evolution of protein sequences and functions under structural constraint. *Biophys. Chem.* **124**, 134–144.
52. Pinney, J. W., Dickerson, J. E., Fu, W., Sanders-Beer, B. E., Ptak, R. G. & Robertson, D. L. (2009). HIV–host

interactions: a map of viral perturbation of the host system. *AIDS*, **23**, 549–554.
53. Ptak, R. G., Fu, W., Sanders-Beer, B. E., Dickerson, J. E., Pinney, J. W., Robertson, D. L. *et al.* (2008). Cataloguing the HIV type 1 human protein interaction network. *AIDS Res. Hum. Retroviruses*, **24**, 1497–1502.
54. Williams, S. G. & Lovell, S. C. (2009). The effect of sequence evolution on protein structural divergence. *Mol. Biol. Evol.* **26**, 1055–1065.
55. Kent, S. J., Fernandez, C. S., Dale, C. J. & Davenport, M. P. (2005). Reversion of immune escape HIV variants upon transmission: insights into effective viral immunity. *Trends Microbiol.* **13**, 243–246.
56. Goepfert, P. A., Lumm, W., Farmer, P., Matthews, P., Prendergast, A., Carlson, J. M. *et al.* (2008). Transmission of HIV-1 Gag immune escape mutations is associated with reduced viral load in linked recipients. *J. Exp. Med.* **205**, 1009–1017.
57. Sanchez-Merino, V., Farrow, M. A., Brewster, F., Somasundaran, M. & Luzuriaga, K. (2008). Identification and characterization of HIV-1 $CD8^+$ T cell escape variants with impaired fitness. *J. Infect. Dis.* **197**, 300–308.
58. Wain, L. V., Bailes, E., Bibollet-Ruche, F., Decker, J. M., Keele, B. F., Van Heuverswyn, F. *et al.* (2007). Adaptation of HIV-1 to its human host. *Mol. Biol. Evol.* **24**, 1853–1860.
59. Rodrigue, N., Lartillot, N., Bryant, D. & Philippe, H. (2005). Site interdependence attributed to tertiary structure in amino acid sequence evolution. *Gene*, **347**, 207–217.
60. Allen, T. M., Altfeld, M., Yu, X. G., O'Sullivan, K. M., Lichterfeld, M., Le Gall, S. *et al.* (2004). Selection, transmission, and reversion of an antigen-processing cytotoxic T-lymphocyte escape mutation in human immunodeficiency virus type 1 infection. *J. Virol.* **78**, 7069–7078.
61. Peyerl, F. W., Bazick, H. S., Newberg, M. H., Barouch, D. H., Sodroski, J. & Letvin, N. L. (2004). Fitness costs limit viral escape from cytotoxic T lymphocytes at a structurally constrained epitope. *J. Virol.* **78**, 13901–13910.
62. Crawford, H., Prado, J. G., Leslie, A., Hue, S., Honeyborne, I., Reddy, S. *et al.* (2007). Compensatory mutation partially restores fitness and delays reversion of escape mutation within the immunodominant HLA-B*5703-restricted Gag epitope in chronic human immunodeficiency virus type 1 infection. *J. Virol.* **81**, 8346–8351.
63. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242.
64. Sali, A. & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815.
65. MacKerell, A. D., Bashford, J. D., Bellott, M., Dunbrack, R. L., Jr, Evanseck, J. D, Field, M. J. *et al.* (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, **102**, 3586–3616.
66. Ren, P. & Ponder, J. W. (2002). Consistent treatment of inter- and intramolecular polarization in molecular mechanics calculations. *J. Comput. Chem.* **23**, 1497–1506.
67. Word, J. M., Lovell, S. C., Richardson, J. S. & Richardson, D. C. (1999). Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* **285**, 1735–1747.
68. Mizuguchi, K., Deane, C. M., Blundell, T. L. & Overington, J. P. (1998). HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.* **7**, 2469–2471.
69. Mizuguchi, K., Deane, C. M., Blundell, T. L., Johnson, M. S. & Overington, J. P. (1998). JOY: protein sequence-structure representation and analysis. *Bioinformatics*, **14**, 617–623.
70. Mizuguchi, K., Sele, M. & Cubellis, M. V. (2007). Environment specific substitution tables for thermophilic proteins. *BMC Bioinformatics*, **8**(Suppl. 1), S15.
71. Mokrab, Y., Stevens, T. J. & Mizuguchi, K. (2010). A structural dissection of amino acid substitutions in helical transmembrane proteins. *Proteins*, **78**, 2895–2907.
72. Lovell, S. C., Word, J. M., Richardson, J. S. & Richardson, D. C. (2000). The penultimate rotamer library. *Proteins: Struct. Funct. Genet.* **40**, 389–408.
73. Connolly, M. L. (1983). Solvent-accessible surfaces of proteins and nucleic acids. *Science*, **221**, 709–713.
74. Thusberg, J. & Vihinen, M. (2009). Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. *Hum. Mutat.* **30**, 703–714.
75. Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797.
76. Ramensky, V., Bork, P. & Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* **30**, 3894–3900.