A Study of the Coevolutionary Patterns Operating within the *env* Gene of the HIV-1 Group M Subtypes

Simon A. A. Travers,[1] *Damien C. Tully,*†[1] *Grace P. McCormack,*‡ *and Mario A. Fares**†

*Molecular Evolution and Bioinformatics Laboratory, Department of Biology, National University of Ireland, Maynooth, Ireland; †Evolutionary Genetics and Bioinformatics Laboratory, Department of Genetics, Smurfit Institute of Genetics, Trinity College Dublin, Dublin, 2, Ireland 4; and ‡Molecular Evolution and Systematics Laboratory, Martin Ryan Institute, National University of Ireland, Galway, Ireland

The *env* gene of human immunodeficiency virus (HIV) is a functionally important gene responsible for the production of protein products (gp120 and gp41) involved in host cell recognition, binding, and entry. This occurs through a complex and, as yet, not fully understood process of protein–protein interaction and within and between protein functional communication. Exposure on the surface of active HIV virions means the gp120–gp41 complexes are subjected to intense immune system pressure and have, therefore, evolved mechanisms to avoid neutralization. Using protein-coding sequences representing all the HIV type-1 (HIV-1) group M subtypes, we have identified amino acids within the *env* gene whose evolution is inextricably linked over the entire HIV-1 group M epidemic. We identified 848 pairs of coevolving residues (involving 263 out of 764 amino acid sites), which represent 0.29% of all possible pairs. Of the coevolving pairs, 68% were significantly correlated by hydrophobicity, molecular weight, or both hydrophobicity and molecular weight. Subsequent grouping of coevolving pairs resulted in the identification of 290 groups of amino acid residues, with the size of these groups ranging from 2 to 10 amino acid residues. Many of these dependencies are correlated by function including CD4 binding, coreceptor binding, glycosylation, and protein–protein interaction. This analysis provides important information regarding the functional dependencies observed within all the HIV-1 group M subtypes and may assist in the identification of functional protein domains and therapeutic targets within the HIV-1 *env* gene.

Introduction

Human immunodeficiency virus type 1 (HIV-1) enters the host cell through a specific binding of the viral envelope glycoprotein gp120 to CD4 and a coreceptor (either CCR5 or CXCR4) (Deng et al. 1996; Dragic et al. 1996; Feng et al. 1996). The HIV *env* gene expresses a 160-kD protein (gp160) that is cleaved to produce gp120 and gp41, which exist as a trimeric spike on the surface of the HIV virion containing 3 exterior gp120 and 3 gp41 transmembrane glycoproteins (Bernstein et al. 1995). The surface of the gp120 trimer is highly glycosylated with as much as 50% of the surface of the gp120 molecule being covered by carbohydrates, which enables evasion of immune system recognition (Kwong et al. 1998; Wyatt et al. 1998; Chen et al. 2005; Pantophlet and Burton 2006). A glycan shield model has been proposed whereby the gp120 glycans are continuously repositioned so as to escape neutralizing antibodies (Wei et al. 2003). It has also been proposed that the repositioning of glycans may compensate for conformational changes due to amino acid replacements occurring in virus escape from neutralizing antibodies (Pantophlet and Burton 2006).

The binding of gp120 to the host cell CD4 receptor induces conformational changes that enable binding of a coreceptor, generally CCR5 or CXCR4, to enable host cell entry (Chen et al. 2005; Huang et al. 2005). Studies have suggested that, as well as the V3 loop, amino acid residues in the gp120 core around the bridging sheet are important in coreceptor binding (Rizzuto et al. 1998; Otto et al. 2003). However, it is thought that the V3 loop is responsible for determining which coreceptor, CCR5 or CXCR4, is used (Hwang et al. 1991; Resch et al. 2001). It has been proposed that, following coreceptor binding, gp120 disassociates from gp41, thereby allowing access of the gp41 fusion peptide to the target cell membrane, enabling membrane fusion between the virion and the host cell membranes (Caffrey et al. 1998). gp41 is known to comprise 4 functional domains: an N-terminal fusion peptide, an ectodomain, a transmembrane domain, and a cytoplasmic domain (Freed and Martin 1995). Recently, however, it has been suggested that the C-terminal tail of gp41 may exist in 2 conformations, with gp41 molecules incorporated into active virions actually containing 2 ectodomains (termed the major and minor ectodomains) and 3 membrane-spanning domains (Hollier and Dimmock 2005).

Such complexities of function and communication within and between gp120 and gp41 are reflected in the complex evolutionary patterns observed in the *env* gene (Yamaguchi-Kabata and Gojobori 2000; Yang 2001; Yang et al. 2003; Choisy et al. 2004; de Oliveira et al. 2004; Travers et al. 2005).

Many methods have been devised to detect selective constraints in linear multiple sequence alignments. However, intramolecular functional relationships between amino acid sites or domains can be better understood by studying the evolutionary dependence among sites. This test in combination with other methods can yield biologically meaningful results because the dependency among amino acid sites becomes a measurable parameter when testing for coevolution (Fares 2006). This dependence can highlight intraprotein patterns of variation used as an evolutionary strategy of the virus to escape immune response of the host and yet recognize the host cell receptor (Tully and Fares 2006). Studying evolution within the *env* gene using coevolution/covariation analysis has been suggested to be useful for identifying potential functional domains for mutagenesis analysis and also as selection for peptides to be used in vaccine design (Korber et al. 1993). Identifying coevolving amino acids within *env*

[1] S.A.A.T. and D.C.T. contributed equally to this study.

Key words: coevolution, HIV-1, env, M subtypes.

E-mail: faresm@tcd.ie.

Mol. Biol. Evol. 24(12):2787–2801. 2007
doi:10.1093/molbev/msm213
Advance Access publication October 5, 2007

may also aid in the identification of domains important in intra-/interprotein communication as well as domains important in protein–protein interaction. Although such identification of protein–protein interaction interfaces within or between proteins is theoretically possible using coevolution analyses, it is currently difficult to distinguish between the various classes of coevolving residues. The development of mathematical/statistical analytical models to distinguish between these classes of coevolving residues would be of immense benefit when testing more specific hypothesis-driven coevolutionary studies. A number of methods have been developed to detect the presence of coevolution between amino acid residues (Korber et al. 1993; Gobel et al. 1994; Shindyalov et al. 1994; Taylor and Hatrick 1994; Tillier and Collins 1995; Chelvanayagam et al. 1997; Pollock and Taylor 1997; Lockhart et al. 1998; Tuffley and Steel 1998; Pollock et al. 1999; Pritchard et al. 2001; Tillier and Lui 2003; Galtier 2004; Ane et al. 2005; Dutheil et al. 2005; Gloor et al. 2005). Many of these methods, however, are limited in that they cannot accurately distinguish phylogenetic linkage from true coevolution, they do not take into account random noise within a multiple sequence alignment, or they require extremely large numbers of sequences to tackle the problem of the high rates of false positives detected. We have recently described a method that exhibits high levels of sensitivity and specificity in the detection of coevolution (Fares and Travers 2006), and here, we present the application of this method to the detection of coevolving residues within the HIV *env* gene.

In previous studies examining coevolution within HIV-1, the *env* gene has been limited to the gp120 V3 loop (Korber et al. 1993; Bickel et al. 1996; Gilbert et al. 2005). Korber et al. (1993) studied 308 subtype B sequences, whereas Bickel et al. (1996) reanalyzed the Korber data as well as a new data set containing 440 sequences that represented a number of HIV-1 group M subtypes (A, C, D, and E) (Bickel et al. 1996). Upon reanalyzing the subtype B data set of Korber et al. (1993), Bickel et al. (1996) identified 4 of the 7 coevolving pairs identified by Korber as well as a number of other coevolving pairs. However, there was no overlap of coevolving pairs identified between the analysis of the data set containing multiple subtypes and the subtype B data set. The lack of overlap between the two sets of analyses was probably due to the use of a more relaxed strategy in the case of Bickel's et al. study, which led to the identification of greater percentage of coevolving pairs. Gilbert et al. (2005) observed significant differences between the number of coevolving pairs identified in their *env* subtype B (26 pairs) and subtype C (1 pair) data sets. From the results presented in these 3 studies, it is obvious that the HIV-1 group M subtypes are exhibiting different levels of coevolution within the *env* V3 loop (Korber et al. 1993; Bickel et al. 1996; Gilbert et al. 2005). Interestingly, it is thought that the ability of the HIV strains to make the transition to CXCR4 coreceptor usage during infection may vary by subtypes. Subtype C, in particular, exhibits a lower frequency of CXCR4 usage when compared with other subtypes (Abebe et al. 1999; Ping et al. 1999; Peeters and Sharp 2000; Cilliers et al. 2003). It is quite possible that the coevolution differences observed in the V3 loop between different subtypes represent biologically functional differences. An elegant study has been recently published that searches evolutionary convergencies (evolutionary interactions) in HIV-1 envelope (Poon et al. 2007). In this study authors applied a "covarion" like phylogenetic model to show that potential N-glycosylated sites (PNGSs) are evolutionarily linked and that exclusive interactions occur significantly more frequently between co-localised PNGSs. We have previously observed heterogeneous selective pressures operating in the evolution of the *env* gene over the HIV-1 group M subtypes (Travers et al. 2005), and more recently, we have observed functional and coevolutionary divergence within *env* gene amino acids between the group M subtypes (Tully DC, Travers SAA, McCormack GP, Fares MA, in preparation). Although the identification of subtype-specific coevolution is important in identifying subtype-specific evolutionary events, it is important to identify coevolving pairs/groups that are present across the entire HIV-1 group M phylogeny. The identification of such residues will provide evidence of functional, structural, or interacting constraints that are conserved over the entire group M epidemic and may, therefore, identify potential functional domains for mutagenesis analysis or peptides that may potentially be used in vaccine design.

Materials and Methods

Taxon Selection

The HIV *env* gene data set used in this study was previously described (Travers et al. 2005). For each HIV-1 group M subtype, all available full-genome sequences were retrieved from the Los Alamos HIV database (http://hiv-web.lanl.gov) and aligned to each other using MacClade 4.08 (Maddison WP and Maddison DR 1992). Representative sequences were selected to represent the spread of diversity throughout the subtype. This procedure was followed to avoid biased representation of the real intrasubtype diversity. Random selection of sequences from each subtype would increase the likelihood of selecting phylogenetically close sequences. For this reason, sequences were carefully selected as to comprise a set of sequences spread throughout the complete evolutionary history of that subtype, based on the reconstructed phylogeny of all full-genome *env* sequences for that subtype. Only a maximum of 4 sequences were finally selected to represent the evolutionary history of each subtype. The selected representative sequences were then manually aligned using MacClade 4.08 (Maddison WP and Maddison DR 1992). Ambiguous regions of the alignment were removed to avoid false positives due to erroneous alignment of nonhomologous sites (residues removed are as follows: 6K-7Y, 12R-16R, 32E-33K, 132T-154I, 172E, 183P-190S, 310Q-311R, 320I, 354G-358T, 386N-413T, 459G-465S, and 782V-788R; numbering is based on the HXB2 reference sequence). The final alignment contained 36 taxa, which represents the extent of diversity present over the entire HIV-1 group M subtypes and was 2292 nucleotides in length. A Neighbor-Joining tree for the resulting data sets was reconstructed using PAUP* 4.0b10 (Swofford 1998). We used this data set to examine coevolving pairs present over the entire HIV-1 group M epidemic.

Table 1
Mean Nucleotide Pairwise Substitutions per Site for the Different Subtypes Used in This Study

	Representative Data Set		All Available Sequences	
	Mean	SE	Mean	SE
A	0.11684119	0.0051538	0.11765014	0.0005955
B	0.07518906	0.00590288	0.0981569	0.00014368
C	0.09584061	0.00503124	0.10160873	3.8526E-05
D	0.11271088	0.00455046	0.10666579	0.00061208
F	0.11092412	0.0053079	0.10947092	0.0033051
G	0.09416812	0.00251751	0.10147994	0.0014638
H	0.11113401	0.00836918	0.11113401	0.00836918
J	0.03589074	NA	0.03589074	NA
K	0.10280576	NA	0.10280576	NA

NOTE.—The mean nucleotide distances were estimated by maximum likelihood using the model TVM + I + G for the entire set of available HIV-1 *env* full-genome sequences and for the representative set of sequences used in this study for coevolution analyses. Whenever only 2 sequences were available for a particular subtype, SE could not be calculated. (NA).

In order to ensure that no biases were introduced regarding the subtype divergence levels by selecting particular sequences, we estimated the mean pairwise nucleotide divergence for each subtype in the data set of representative sequences and compared the divergence levels between subtypes and between the representative alignment and an alignment containing all available full-genome (700) *env* sequences. Pairwise nucleotide divergences were estimated under a maximum-likelihood criterion using the model TVM + I + G, which has been estimated using the program Modeltest (Posada and Crandall 1998). The mean pairwise nucleotide distance of the full alignment of 700 sequences ($0.147 \pm 7.35 \times 10^{-5}$ nucleotide substitutions per site) and the subset used in this study for coevolution analyses ($0.155 \pm 8.0 \times 10^{-5}$) were very similar, indicating no bias in the divergence levels between both data sets. We also compared the nucleotide mean pairwise nucleotide distances between the full HIV-1 data set and the representative data set in each one of the subtype, and the results show no significant differences (table 1). To ensure that the number of sequences is not introducing any bias regarding coevolution detection, we used also 2 data set, with one containing all the sequences available for 3 subtypes (A, B, and F) and the other containing the representative sequences of these subtypes. We used this approach instead of testing coevolution in the full alignment data set due to computational limitations of the program to run over 700 sequences' alignment. Finally, to discard any effect of the number of sequences in the coevolutionary analyses, we have also conducted these analyses on different subsets of the 700 based multiple sequence alignment. These subsets were built sampling randomly from each subtype the same number of sequences as in the original analyses and always those sequences showing equal divergence levels as the original set.

Analysis of Intra- and Interprotein Molecular Coevolution

To test for intra- and intermolecular coevolution, we used our recently published method for the coevolution analysis of protein sequences (Fares and Travers 2006). We have previously demonstrated that the sensitivity of CAPS in detecting significant coevolving pairs is statistically significant for multiple sequence alignments containing 20 or more sequences (Fares and Travers 2006). The method has previously been used with good effect to study coevolution within data sets containing numbers of sequences similar to those used in this study (Fares and Travers 2006; Travers and Fares 2007). Briefly, CAPS compares the correlated variance of the evolutionary rates at 2 sites corrected by the time since the divergence of the 2 sequences they belong to. This method compares the transition probability scores between 2 sequences at 2 particular sites, using the blocks substitution matrix (Henikoff and Henikoff 1992). The significance of the CAPS correlation values was assessed by randomization of pairs of sites in the alignment, calculation of their correlation values, and comparison of the real values with the distribution of 10,000 randomly sampled values. To correct for multiple tests and for non-independence of data, we implemented the step-down permutation procedure in both methods and corrected the probabilities accordingly (Westfall and Young 1993). CAPS is implemented in the program CAPS v1.0 (Fares and McNally 2006). For coevolution analyses, we used the protein-coding sequence, corrected for type I error using an alpha value of 0.001. The HXB2 reference sequence was used to identify the amino acid positions, and all amino acid numberings presented here correspond to HXB2. To correct for the divergence levels on each amino acid site, we weighted the correlated variability between amino acid sites by the level of substitutions per synonymous sites estimated by Li (1993). We only used full-genome representative sequences from each subtype. The selection of these sequences allowed us to definitively exclude any intersubtype recombinant sequences that may bias the results obtained from the coevolution analyses. We have also attempted to run CAPS on the complete HIV-1 sequence data set. However, because CAPS is a very computationally intensive method, computers could not run this program on such a large data set. To ensure that the numbers of sequences included were not biasing the analyses, we ran CAPS on an alignment, which included subtypes A (63 sequences), B (161 sequences), and F (13 sequences). We then compared the coevolutionary results on these subtypes with those obtained when we ran CAPS on a data set comprising the representative sequences of these 3 subtypes (6 sequences from subtype A, including A1 and A2, 4 sequences from subtype B, and 7 sequences from subtype F, including F1 and F2). The same pairs of coevolving residues were detected in both although the correlation coefficients were slightly lower in the alignment containing the full list of sequences for the 3 subtypes. These coefficients were nevertheless significant at a 0.001 alpha value. The conclusion from this analysis is then that the size of the alignment does not influence the sensitivity of the coevolution analysis as far as the multiple sequence alignment contains more than 10 sequences, something already shown in a previous work (Fares and Travers 2006).

Molecular coevolution can be divided into many different types including structural, functional, interaction, phylogenetic, and stochastic coevolution (Atchley et al.

Table 2
Details of the Coevolving Pairs Whose Coevolution Was Correlated by Hydrophobicity, Molecular Weight, or Both

	Number of Significant Pairs	Mean Correlation	Mean Probability	gp120	gp41
Hydrophobicity	311	0.459403 (−0.1660–0.9877)	0.0182053 (0.0015–0.0500)	123	69
Molecular weight	268	0.396898 (0.1653–1.000)	0.0217441 (0.0015–0.0494)	113	64
Hydrophobicity and molecular weight	194	Hydro: 0.525354 (0.1511–0.9877) Mw: 0.431891 (0.1653–1.000)	0.014571 (0.0015–0.0482) 0.0190494 (0.0015–0.0494)	95	56

NOTE.—The number of significant pairs is shown as are the mean values and ranges for the correlation and probability statistics. Also shown is the number of correlated residues located within the *env* gp120 and gp41 domains.

2000). Disentangling the different types of coevolution is anything but straightforward. In our previous work, however, we attempted to distinguish between phylogenetic, stochastic, and the other components of coevolution through a phylogenetic-based coevolution analysis procedure (Fares and Travers 2006). Distinguishing between structural, functional, and interaction coevolution requires biological information in addition to the mathematical adjustments made to the method. Estimating the correlated variation in hydrophobicity, molecular weight, or combination of both parameters may introduce further information regarding the coevolutionary relationships (functional, structural, or functional and structural) among covarying sites. We therefore conducted an analysis of correlation between coevolving amino acid sites, taking into account these biological parameters.

Mapping Significant Amino Acid Residues onto *env* Protein 3D Structures

Many 3-dimensional (3D) structures have been resolved for the gp120 and gp41 proteins, and in this study, we used structures representing gp120 in complex with a CD4 receptor and a neutralizing antibody (PDB accession number 1G9M), an unliganded simian gp120 core structure (2BF1), the V3 loop from a V3 loop-containing gp120 structure (2B4C), and a structure representing the SIV gp41 ectodomain (1IF3) (Caffrey et al. 1998; Kwong et al. 1998; Chen et al. 2005; Huang et al. 2005). The 3D structure viewing and manipulation was performed using iMOL (http://www.pirx.com/iMol/).

We used a conservative mean distance of 8 Å in determining whether 2 amino acid residues were significantly proximal in the 3D structure. The relative distance between 2 amino acids was calculated by taking the mean 3D atomic coordinates for each amino acid in the structure. We then calculated the distance between 2 amino acids as the distance between their mean coordinates as follows:

$$d=\sqrt{(x-x')^2+(y-y')^2+(z-z')^2}$$

where x, y, and z are the mean atomic coordinates for residue 1 and x', y', and z' are the mean atomic coordinates for residue 2.

Results

Coevolution analysis resulted in the identification of 848 pairs of coevolving residues, representing 0.29% of all possible pairs. The observation of coevolving pairs constituting 0.29% of all possible pairs in this study is significantly lower than those previously reported for HIV-1, for example, 1.33% by Korber et al. (1993), 5.24% observed by Gilbert et al. (2005) for their subtype B data set, and 12.69% observed by (Bickel et al. 1996) in both their 308 and 440 data sets. The mean correlation coefficient for these pairs was 0.5962 (range 0.5000–0.9944). These pairs represented 233 amino acid residues within the *env* gene, 145 and 88 within gp120 and gp41, respectively. Subsequent grouping of all pairs resulted in 290 groups of coevolving residues, with the size of these groups ranging from 2 to 10 amino acid residues. The majority of these groups (72%), however, contained either 2 or 3 residues (supplementary table 1, Supplementary Material online).

We also applied further filters to identify coevolving pairs whose coevolution was correlated by hydrophobicity, molecular weight, or both (table 2). This analysis would enable identification of compensatory mutations and/or mutations at structurally related amino acid sites. Of the 848 pairs of coevolving residues (involving about 263 out of 763 amino acid sites) identified in *env*, 311 and 268 of these were correlated by hydrophobicity and molecular weight, respectively, whereas 194 residues were correlated by both hydrophobicity and molecular weight. The 848 coevolving pairs together with their co-evolutionary parameters are shown in table 3 of supplementary information.

In order to visualize the spread of coevolving pairs throughout the *env* gene, we plotted a matrix exhibiting coevolving pairs and also pairs whose coevolution was correlated by hydrophobicity, molecular weight, or both hydrophobicity and molecular weight (fig. 1). Although coevolving pairs were spread throughout *env*, 2 distinct regions exhibit high levels of coevolution with many residues in *env*: the end of C2 with V3 and C3 as well as a portion of the gp41 cytoplasmic domain (fig. 1).

Coevolution within the gp120 V3 Loop and Proposed Coreceptor-Binding Domains

Previous coevolution studies in *env* focused on identifying coevolution within the V3 loop (Korber et al. 1993; Bickel et al. 1996; Gilbert et al. 2005). We have expanded upon these studies and investigated the presence of coevolution throughout the entire *env* gene. Within the V3 loop, however, we have identified 24 pairs of coevolving residues comprising 14 residues of the V3 loop (fig. 2). Of the 24 pairs of coevolving residues, 4 (17%) of these were significantly correlated by hydrophobicity, whereas 13 (54%)

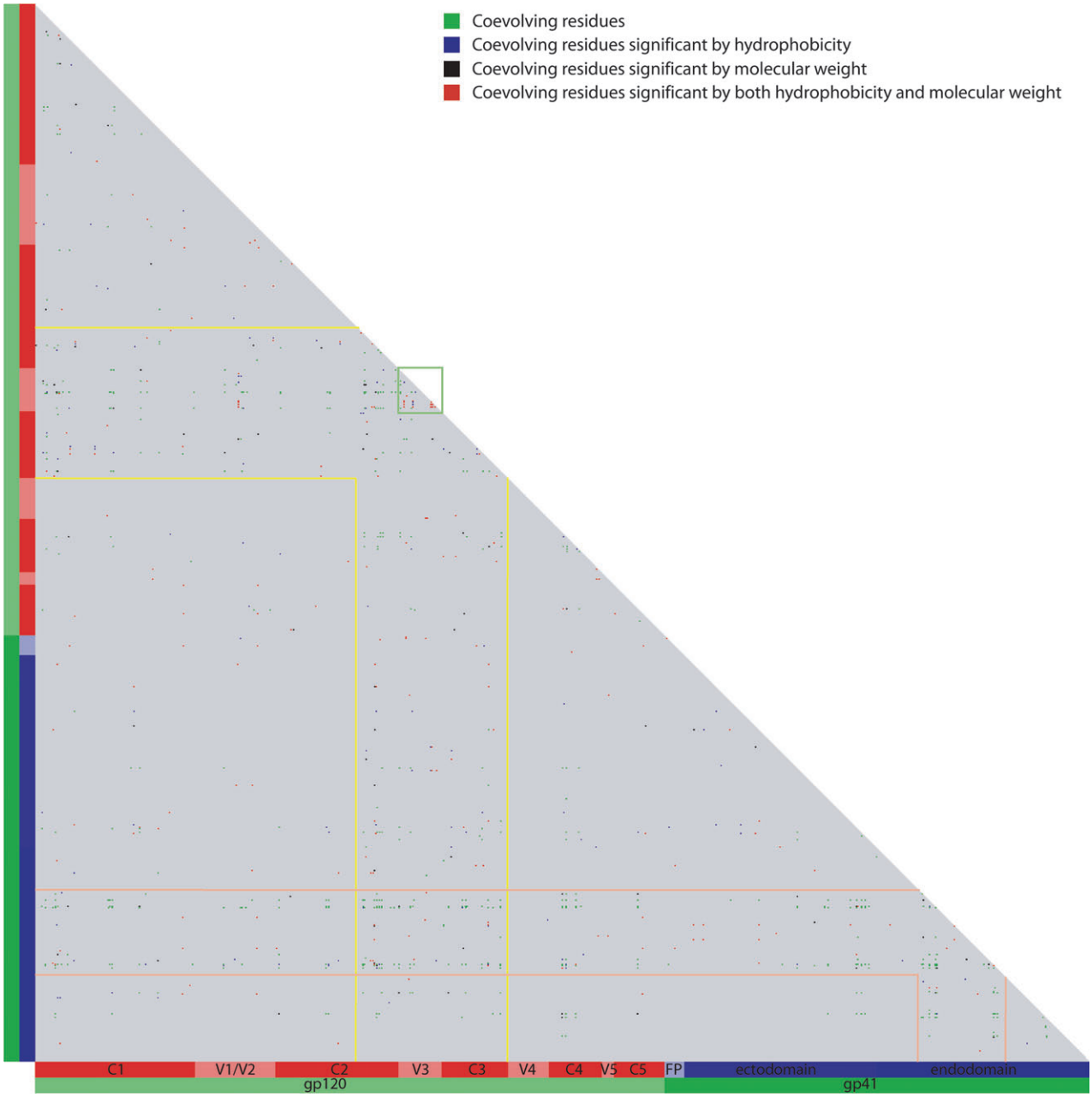

FIG. 1.—Coevolving pairs of amino acid residues over the complete *env* gene. Pairs whose coevolution is correlated by hydrophobicity, molecular weight, or both hydrophobicity and molecular weight are marked. Also shown between colored lines are the 2 regions that appear to exhibit higher levels of coevolution than that observed over the entire *env* gene. Coevolving residues within the V3 loop are also shown (green box).

were significantly correlated by both hydrophobicity and molecular weight. Five of these 13 pairs of coevolving residues had also been identified as coevolving by Bickel et al. (1996) in their 440 data set, which contained representative sequences from multiple subtypes.

Upon solving the structure of a V3-containing HIV-1 gp120 core, Huang et al. (2005) proposed that, following CD4 binding, the N-terminus of the CCR5 receptor binds the gp120 core and V3 base, whereas the V3 tip binds the coreceptor's second extracellular loop. Complimentary to this, Rizzuto et al. (1998) identified a number of residues within the gp120 core, the mutation of which significantly affects CCR5 coreceptor binding by gp120. Many of these residues were not identified as coevolving within *env*, most likely because of functional conservation. Rizzuto et al. (1998) identified 2 residues, P437 and Q442, that, when mutated, result in a $\geq$50% increase in CCR5 binding with respect to wild-type gp120. Both these residues were identified as coevolving with residues within the V3 loop stem (N302 with P437 and R306 with Q442). Also within the proposed gp120 core, CCR5-binding domain residue V200 was identified as coevolving, with 3 residues in the V3 loop, 2 in the tip (R315 and A316), and 1 in the base (Q328). Mutation of V200 showed a slight decrease

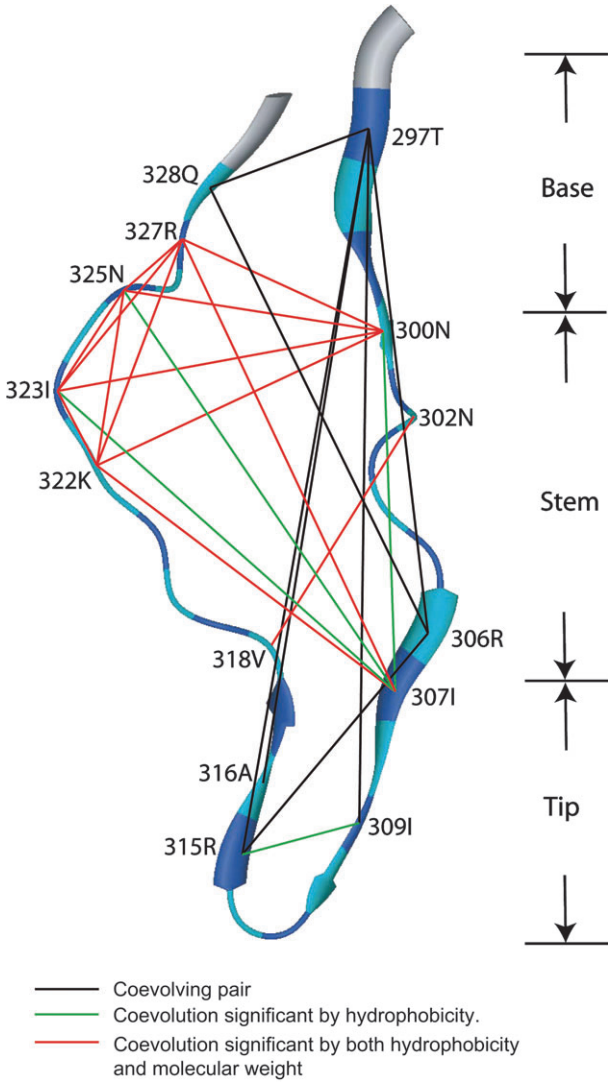

FIG. 2.—Pairwise coevolving residues observed within the gp120 V3 loop (Huang et al. 2005). Numbering is per the HXB2 reference sequence, and residues are colored alternately to ease visualization.

(16%) in CCR5 binding when compared with wild type (Rizzuto et al. 1998). Residue G379, although not tested by Rizzuto et al. (1998), is directly adjacent in the gp120 3D structure to E381 (5.38 Å), which, when mutated, decreases CCR5 binding by 93%. G379 was observed in this study as coevolving with T297 in the base of the V3 loop.

CD4 binding has been shown to induce conformational changes in gp120 (Chen et al. 2005), which have been proposed to expose domains within gp120 responsible for coreceptor binding (Chen et al. 2005; Huang et al. 2005). Therefore, because of the structural dependencies between the CD4 and coreceptor-binding domains, one would expect to see a certain degree of coevolution between these domains or their neighbor peptide regions to maintain function. We have observed 8 residues within the V3 loop and 5 residues within the proposed coreceptor-binding domain on the gp120 core that coevolve with residues that either bind directly

Table 3
Residues in the V3 Loop and Proposed CCR5-Binding Domain That Coevolve with Residues Involved in CD4 Binding

	Bind Directly to CD4	Directly Adjacent to CD4-Binding Pocket
V3 loop		
T297	T283, K429	S364 (S365, 4.36451 Å)
R306	—	S364 (S365, 4.36451 Å)
I307	—	T278 (D279, 5.34946 Å)
I309	T283	—
R315	D279, A281, T283	—
A316	D279	—
R327	—	S274 (T283, 6.72668 Å), N276 (D279, 5.35118 Å)
Q328	D279, A281, T283	—
CCR5-binding domain		
K121	S365	—
P437	—	T373 (I371, 6.34689 Å)
R440	D279, K429	—
Q442	D279	S364 (S365, 4.36451 Å), K432 (N425, 5.96993 Å)
R444	—	K432 (N425, 5.96993 Å)

NOTE.—Also shown in brackets are the closest adjacent CD4-binding residues and the pairwise distance in angstroms (Å).

to CD4 or are proximally contained within the CD4-binding pocket (<8 Å from residues that bind CD4 directly, table 3).

Networks of Coevolving Amino Acids between gp120 and CD4

Core to the function of HIV is the binding of gp120 to the CD4 receptor on the host cell surface. We have investigated the coevolution network present within amino acid residues involved in CD4 binding by gp120. Included in this were residues that bind directly to CD4 (Kwong et al. 1998), residues that comprise the epitope for BMS-806, the binding of which interferes with gp120-CD4 binding (Pantophlet and Burton 2006), as well as other residues contained within conserved CD4-binding site epitopes detailed by Wyatt et al. (1998). We have also included amino acid residues, which may be functionally proximal (<8 Å) to residues responsible in CD4 binding in both the HIV liganded and SIV unliganded gp120 structures (Kwong et al. 1998; Chen et al. 2005). The CD4 coevolution network contained 32 amino acid residues (fig. 3), 8 of which bind CD4 directly, one that maps to the BMS-806 epitope, and 5 residues that are directly glycosylated in the HIV or SIV structures (these are located <8 Å from residues important in CD4 binding). Of the 37 coevolving pairs present in the CD4 network, the coevolution of 65% of these was correlated by hydrophobicity (5 pairs), molecular weight (2 pairs), or both hydrophobicity and molecular weight (17 pairs).

The "glycan shield" model suggests that domains within the gp120 structure are protected from neutralizing antibodies by the presence of carbohydrate molecules bound to the surface (Wei et al. 2003). This shield is formed within the gp120 tertiary structure, bringing linearly distant domains into close proximity to form the shield. Therefore,

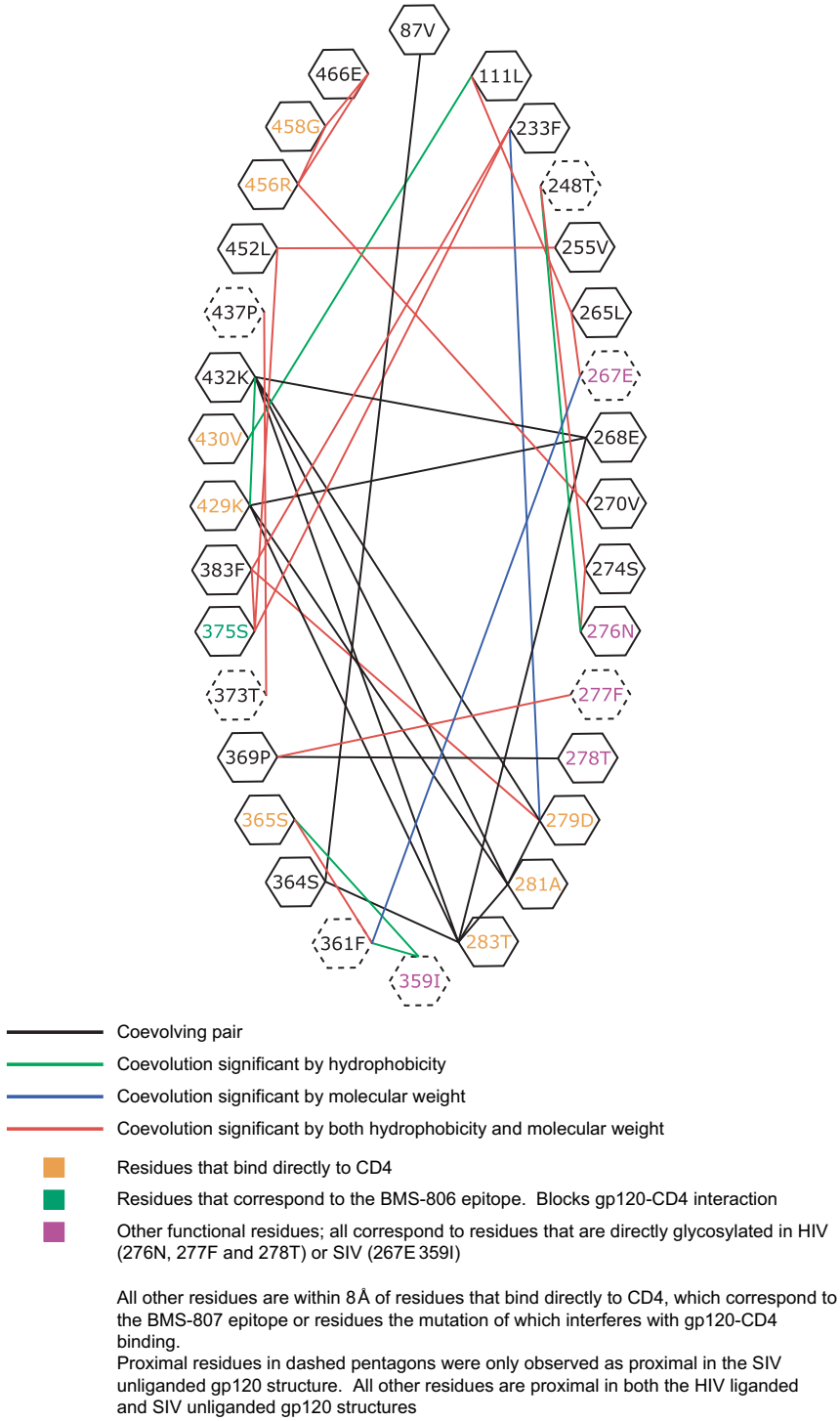

FIG. 3.—The CD4 coevolution network. Included are residues that bind CD4 directly and residues that correspond to the BMS-806 epitope, the binding of which interferes with gp120-CD4 interaction. All other residues are within 8 Å of CD4 functional residues. A number of these proximal residues correspond to known glycosylation residues, and these are also marked.

one would expect that, in order to maintain the overall structure of the glycan shield, there would be a degree of coevolution between directly glycosylated residues and also residues directly proximal to glycosylated residues (as mutation at these residues could affect the overall structure of the shield). This is, in fact, the case with an extensive network of coevolution observed between 41 directly glycosylated and glycosylation-related residues (fig. 4). The coevolution of 67% of the 42 coevolving pairs in the glycosylation network was correlated by hydrophobicity (8 pairs), molecular weight (7 pairs), or both hydrophobicity and molecular weight (13 pairs). These results support that

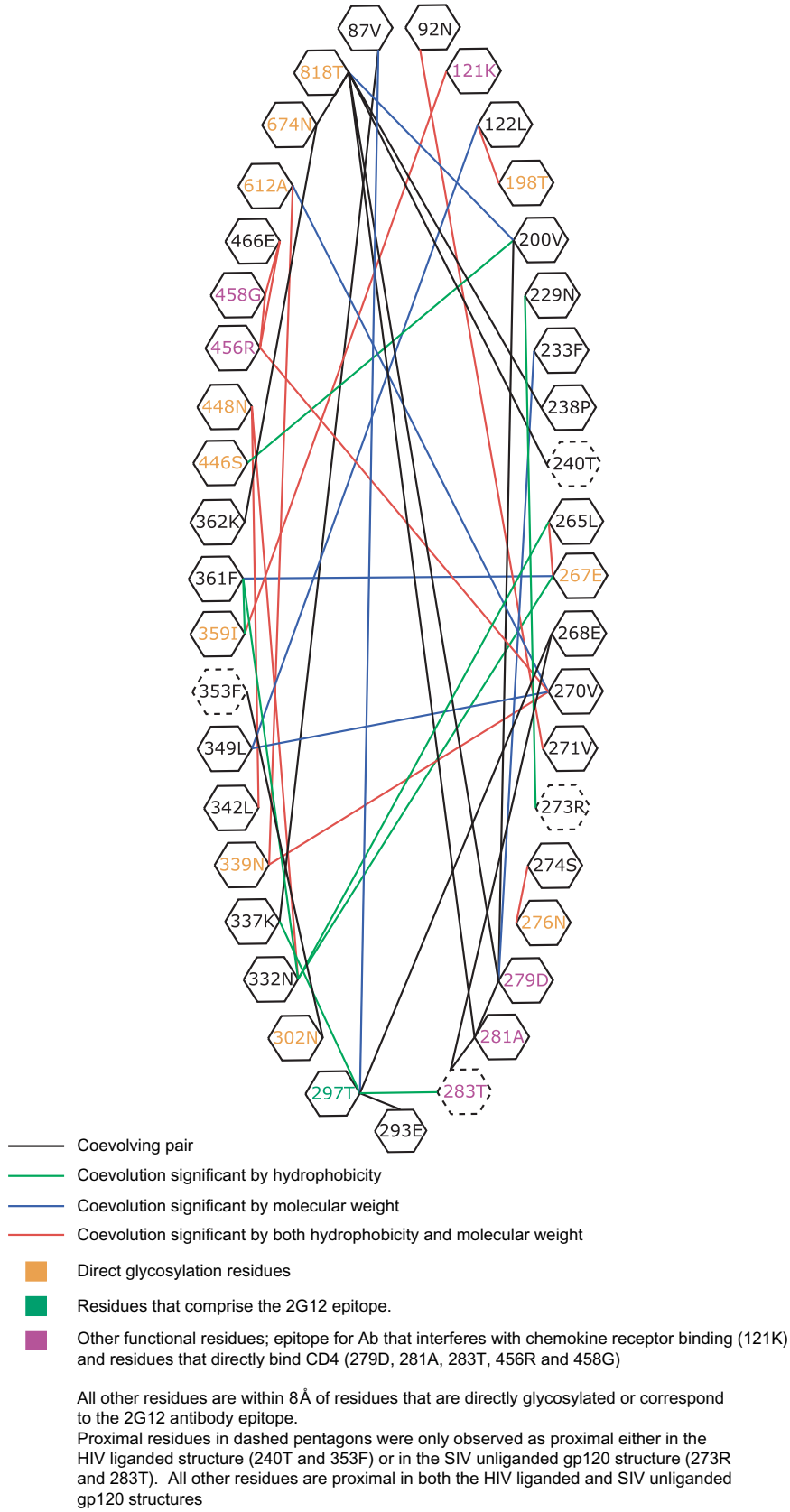

FIG. 4.—The glycosylation coevolution network. Residues that are directly glycosylated are shown as are residues that comprise the 2G12 epitope. Similar to the CD4 network, all other residues are within 8 Å of glycosylation functional residues. A number of these residues are functional residues, and these are marked.

Table 4
Coevolving Amino Acid Residues That Exhibit Marked Differences in Their Mean Pairwise Distances between the Liganded HIV gp120 and the Unliganded SIV gp120 Structures

Coevolving Pair	Mean Pairwise Distance in Liganded gp120 (Å)	Mean Pairwise Distance in Liganded gp120 (Å)
Coevolving residues proximal in liganded gp120		
92N[a] and 271V[b]	19.0935	28.7078
111L[a] and 265L[b]	24.1784	32.5983
122L[a] and 471G[b]	24.5348	34.9453
240T[a] and 348K	22.4094	31.5596
369P[a] and 379G[a]	19.143	32.4727
369P[a] and 440S[c]	23.4603	35.4941
Coevolving residues proximal in unliganded gp120		
87V and 106E[a]	31.0536	22.5208
87V and 364S[a]	43.9368	29.8071
102E[a] and 364S[a]	26.6992	16.1491

NOTE.—Distances are shown in angstroms (Å).

[a] Residues that exhibit movement between the liganded HIV and unliganded SIV gp120 structures.

[b] Pairs whose coevolution is correlated by both hydrophobicity and molecular weight.

[c] Pairs whose coevolution is correlated by molecular weight.

coevolution between or nearby Nglycosylated sites is important to maintain the structure of the glycosyl shield against the defense system of the host. Also, most of the coevolving pairs included sites that were not directly proximal in the structure supporting the results previously reported (Poon et al. 2007). We also observed a large degree of overlap between the CD.

We also observed a large degree of overlap between the CD4 and glycosylation networks, with 63 pairs of coevolving residues observed between these. Of the 63 coevolving pairs, 12 were present in both networks and 42 were present in one of the networks, whereas 9 novel pairs had not been identified in either the CD4 or the glycosylation coevolution networks.

Coevolution within gp120 Moving Domains

Recently, the structure of an unliganded SIV gp120 core was resolved (Chen et al. 2005) and showed marked differences with a structure of a gp120 core liganded with CD4 (Kwong et al. 1998). Chen et al. observed large displacements within the gp120 core inner domain and the absence of the bridging sheet in the unliganded structure. With the exception of 2 regions, the orientation of the outer domain remained essentially the same between the 2 structures. Using both the liganded and unliganded gp120 structures, we looked for coevolving residues that showed a significant difference (>8 Å difference) in their mean pairwise distances between the 2 structures. We identified 6 coevolving pairs that were significantly more proximal in CD4-bound gp120 than in unliganded gp120 and 3 coevolving pairs that were significantly closer in the unliganded gp120 structure (table 4). With the exception of 2 pairs (P369 and G379, and E102 and S364), only one of the coevolving residues in each pair is located within a domain identified by Chen et al. (2005) as moving significantly following CD4 binding.

Table 5
Residues within the Proposed gp120-Binding Hydrophobic Patch of the gp41 Ectodomain and Residues Elsewhere in *env* That They Coevolve with

gp41 Hydrophobic Patch Residues	Coevolving Residues
588K	17W[a], 277F[b], 369P, 373T[a], 535M[b], 543Q[c], 564H[a], 662E[a], 726G, 746I[a], 758D[a], 777I[a]
605T	270V[c], 339N[c], 683K[a]
607A	778V
612A	270V[b], 339N[a]

[a] Pairs whose coevolution correlates by both hydrophobicity and molecular weight.

[b] Pairs whose coevolution correlates by molecular weight.

[c] Pairs whose coevolution correlates by hydrophobicity.

Inter gp120–gp41 Coevolution

The 3D structure of the ectodomain of SIV gp41 has been solved, and combining the properties of this structure with previous mutagenesis analyses, Caffrey et al. (1998) proposed that the gp120-binding domain is located within a hydrophobic patch within the gp41 loop domain. We observed 4 residues that map to this hydrophobic patch (K588, T605, A607, and A612) as coevolving with residues elsewhere within both gp120 and gp41 (table 5 and fig. 5). Residue K588 coevolves with 3 residues (M535, Q543, and H564), all of which map to the fusion peptide/N-terminal heptad repeat (NHR) domain within the gp41 ectodomain (fig. 5*A*), which is proposed to move out from gp120, and enable fusion of the virion and target cell membranes (Caffrey et al. 1998). The only other residue in the gp41 ectodomain that K588 coevolves with is E662 that, although located on the C-terminal heptad repeat (CHR), appears to be located in the corresponding position on the CHR as Q543 is on the NHR (fig. 5*A*).

Both T605 and A612 coevolve with V270 and N339, which map to the highly glycosylated outer domain of the gp120 core structure, whereas K588 coevolves with F277, P369, and T373, all which map to the CD4-binding domain in gp120 (fig. 5*B*). With the exception of W17 and K683, the remaining 5 residues that coevolve with the gp41 hydrophobic domain (G726, I746, D758, I777, and V778) are located within the gp41 cytoplasmic domain.

Discussion

In this study, we have evaluated the coevolution operating within the *env* gene across all the HIV-1 group M subtypes. The identification of pairs or groups of coevolving residues provides a wealth of information with regard to amino acid residues or protein domains that exhibit dependency in their evolution. We have, where possible, connected the coevolution results to biological knowledge. The remaining coevolution pairs/groups presented here (supplementary table 1, Supplementary Material online) should be viewed as potentially biologically significant pairings, and we suggest that many of these results should be further examined experimentally to determine the biological significance of the observed coevolution. We must

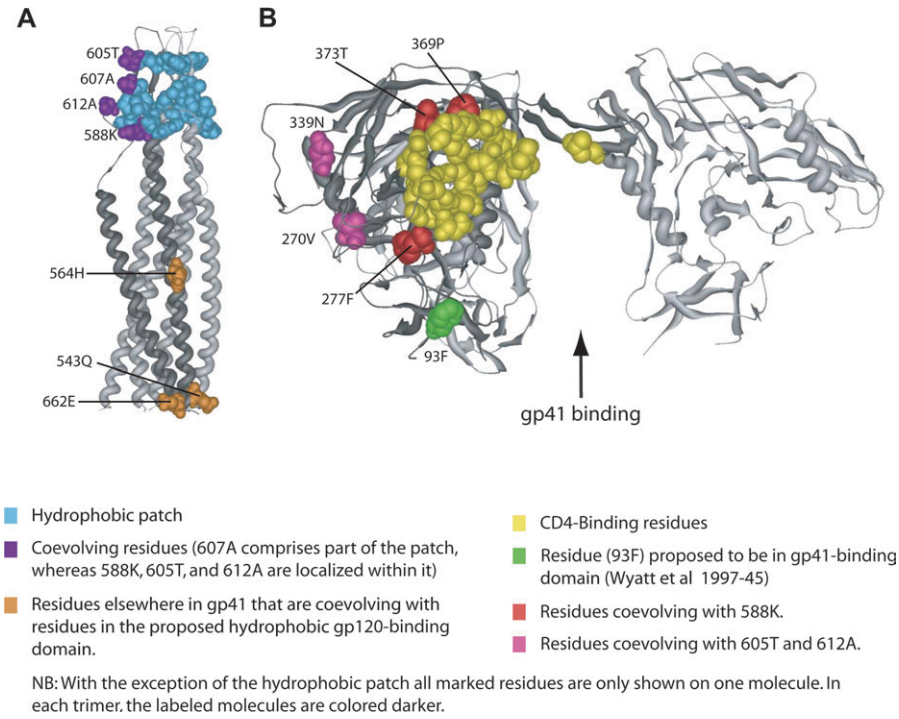

FIG. 5.—Interprotein coevolution between gp41 and gp120. (*A*) Coevolving residues within the gp41 ectodomain and (*B*) the residues within the gp120 core with which they coevolve.

emphasize, however, that the observation of coevolution between 2 domains does not indicate protein–protein interaction between these domains. We have suggested other reasons for the observation of coevolution, both here and in previous works (Fares and Travers 2006; Travers and Fares 2007). Representative sequences were selected in such a way as to sample a complete cross section of the diversity observed within each subtype (Travers et al. 2005).

Although still not fully understood, the functional complexities of the gp120–gp41 complex have been well documented (Wyatt et al. 1997; Kwong, Wyatt, Sattentau, et al. 2000; Poignard et al. 2001; Chen et al. 2005; Hartley et al. 2005; Pantophlet and Burton 2006). Coupled with this, the intense selective pressures known to operate on *env* have resulted in a gene with incredibly complex, multifaceted evolutionary dynamics (Holmes et al. 1992; Seibert et al. 1995; Yang 2001; Choisy et al. 2004; de Oliveira et al. 2004; Travers et al. 2005). In this study, we have attempted to improve the understanding of the coevolutionary selective pressures operating on the *env* gene across all the HIV-1 group M subtypes. Although previous studies have concentrated on examining coevolution within the gp120 V3 loop (Korber et al. 1993; Bickel et al. 1996; Gilbert et al. 2005), and others have observed subtype specific patterns of evolution (Korber et al. 1994; Gaschen et al. 2002; Gnanakaran et al. 2007) the advent of more sensitive and accurate methods has enabled us to examine coevolution across the entire *env* gene. The use of a highly sensitive method such as CAPS has allowed us to identify a significantly smaller subset of coevolving pairs in HIV-1 *env* gene compared with previous works. The sensitivity of CAPS has been estimated to be around 95% in multiple sequence alignments comprising around 40 sequences (Fares and Travers 2006). Our confidence on the low proportion of false positives based on previous works allows us to confirm that most of the amino acid site pairs detected are true-positive coevolving pairs. This work has been also applied in other case studies, showing the high accuracy of the method in detecting true-positive results (Travers and Fares 2007). In contrast to previous studies, we have studied a data set containing representative sequences from all HIV-1 group M subtypes as opposed to single subtypes or a number of subtypes together. The identification of such coevolving residues can provide insights into domains of proteins or pairs of amino acid residues within a protein whose evolution is inextricably linked by structural, functional, or interacting constraints. In this study, 46% of all coevolving pairs were correlated by hydrophobicity, molecular weight, or both hydrophobicity and molecular weight. Some of these correlated pairs are linearly proximal, for example, 456R and 458G. However, some of these correlated pairs are linearly distant but structurally proximal, for example, 122L and 198T are separated by 76 amino acids on a linear level yet are only 6.7 Å apart in the HIV gp120 3D structure (Kwong et al. 1998). Such correlations of coevolving pairs are testament to the evolutionary complexities operating within HIV. The absence, however, of complete structural data and deeper comprehension on the mode of virus operation makes the distinction of the type of amino acid site dependency anything but straightforward.

We cannot exclude the effect of recombination in our results of coevolution. Even though the selection of representative full-genome sequences of each subtype allowed us to avoid the effects of intersubtype recombination, excluding intrasubtype recombination remains a problem.

However, currently there is no way to identify intrasubtype recombination and, as with all analyses with HIV-1 group M multiple sequence alignments; therefore, some intrasubtype recombinants may be present in the data.

Discussion of the biological significance of the complete set of coevolving pairs (848 pairs) is impossible in the manuscript because there is no reported functional data on each one of the pairs. Also, only simulation studies can provide a measure of the sensitivity, and thus of the amount of positive results, of the method to detect real coevolution. Several lines of evidence indicate that these pairs are not false positive resulting from a limited statistical power of the method used. First, comparison of the correlation coefficients of each nondiscussed pair of coevolving sites with that for the pairs with biological information show no difference in their values. Second, our previous analysis of the performance of the method to detect coevolution (Fares and Travers 2006), using a simulation approach developed in other works and not related to our algorithm to detect coevolution, showed that the sensitivity of the method can be as high as 90% when the number of sequences in the multiple sequence alignment is above 20, although we have used 36 sequences in our study. Despite this fact, we still believe that a minor fraction may be false positives, although this fraction is dramatically smaller than in other studies performed so far.

The Distribution of Coevolving Residues throughout the *env* Gene

Although coevolving residues are spread throughout the *env* gene, we did observe 2 regions that present a higher density of coevolution with residues throughout the *env* gene (fig. 1). The first of these regions covers the latter part of C2 as well as the V3 loop and C3 domain. Observing such a density of coevolution in this region is not at all surprising as it contains a large proportion of amino acid residues involved in glycosylation as well as CD4 and chemokine receptor binding (Kwong et al. 1998; Rizzuto et al. 1998; Wyatt et al. 1998; Kwong, Wyatt, Majeed, et al. 2000; Wei et al. 2003; Chen et al. 2005). Coevolution between residues within this domain is most likely occurring to maintain the overall structural properties required for optimum protein function. The binding of gp120 to CD4 is known to induce conformational changes within gp120 (Chen et al. 2005), which have been proposed to make further gp120 domains accessible for coreceptor binding (Chen et al. 2005; Hartley et al. 2005; Huang et al. 2005). Following coreceptor binding, it has been proposed that gp120 disassociates from gp41 to enable gp41-facilitated cell membrane fusion (Caffrey et al. 1998). This complex mechanism requires a large degree of intra- and interdomain communication within and between the gp120 and gp41 molecules. The large degree of coevolution present between the C2–V3–C3 region and residues throughout *env* supports this claim.

The second region exhibiting a high level of coevolution with residues throughout *env* is interesting as it is located within the gp41 cytoplasmic domain. Hollier and Dimmock (2005) detailed a number of studies that have shown that antibodies specific to an antigenically active motif (Kennedy sequence) in the gp41 cytoplasmic domain can neutralize HIV-1 virions (Hollier and Dimmock 2005). As antibodies cannot cross the lipid bilayer of the cell membrane, this suggests that a portion of the gp41 cytoplasmic domain is exposed on the cell surface. Based on this evidence, Hollier and Dimmock (2005) proposed a structural model for gp41 that consists of 3 membrane-spanning domains and 2 ectodomains, a major and a minor. The Kennedy sequence is exposed on the outer face of the proposed minor ectodomain. Hollier and Dimmock (2005) suggested that this gp41 structure is evident only in a minority of cell-associated gp41 molecules that are destined for incorporation into active virions. Of the 9 residues within gp41 that show high levels of coevolution with residues elsewhere in *env*, 5 of these (L721, G726, E731, G732, and I746) are located in the minor ectodomain, with 3 of them (G726, E731, and G732) being located within the Kennedy sequence. It has been suggested that there may be interactions between the minor ectodomain and the major ectodomain as well as with elements of gp120 and also with other gp41 monomers that form the gp41 trimer (Hollier and Dimmock 2005). This level of functional dependency among domains within the gp120–gp41 complex would explain the large degree of coevolution observed between the 5 residues located within the minor ectodomain and residues elsewhere in *env*. The remaining 4 residues that comprise the region of gp41 coevolving with a large number of residues throughout *env* (L774, V778, T779, and I781) are located within the cytoplasmic domain of gp41. All these residues are directly adjacent to the second tyrosine-dependent sorting signal in $gp41^{768}$ $YHRL^{771}$ (Hollier and Dimmock 2005), a peptide of which has been shown to interact with an adaptor protein-2 complex (Ohno et al. 1997; Boge et al. 1998); however, it is not known whether this signal is functional within gp41 (Rowell et al. 1995; Boge et al. 1998).

Coevolution within the gp120 V3 Loop

The gp120 V3 loop is critical for coreceptor binding and is also responsible in determining coreceptor usage (Hwang et al. 1991). It has also been shown to be a target for the host immune response and can somehow affect the sensitivity of virions to neutralization (Hartley et al. 2005). Recent structural analysis has shown that the V3 loop protrudes by as much as 30 Å from the gp120 trimer, suggesting that perhaps the N-terminus of the CCR5 receptor binds the gp120 core and V3 base, whereas the V3 tip binds the coreceptor's second extracellular loop (Huang et al. 2005). We have shown the presence of coevolution both within and between amino acid residues within the V3 loop base and tip regions (fig. 2). There is also a large degree of coevolution involving residues within the V3 loop stem (fig. 2), the majority of which correlate by hydrophobicity or by both hydrophobicity and molecular weight, probably as a result of the functional and structural constraints imposed on this functional domain. Resch et al. (2001) proposed that coreceptor usage is directed by positions 11 and 25 within the V3 loop (R306 and K322) and that positively charged amino acids at these positions direct CXCR4

usage, whereas others direct CCR5 usage. These positions exhibit a high degree of coevolution, with residues both within the V3 loop (fig. 2) and elsewhere in *env*. For example, R306 coevolves with 21 residues within *env*, 3 within the V3 loop as well as residues located in the CD4-binding pocket (S364 and K432), residues the mutation of which greatly reduces CCR5 binding (Q442), residues located in the proposed minor ectodomain (E731 and G732), and residues adjacent to the second tyrosine-dependent sorting signal in gp41 (V778 and I781). K322 coevolves with 8 residues in *env*, 5 within the V3 loop (fig. 2), and of the remaining 3, 1 is located in the gp120 V2 loop (R166) and 2 are located within the loop region in gp41, which connects the NHR and CHR and has been associated with gp120 association (L602 and Q621).

The functions of the V3 domain are closely linked with domains elsewhere within the gp120 core (Rizzuto et al. 1998; Poignard et al. 2001; Hartley et al. 2005; Huang et al. 2005). The observation of coevolution between residues within the CD4-binding domain, residues involved in glycosylation, and amino acid residues outside the V3 loop suggested to be involved in coreceptor binding corroborates the level of evolutionary functional dependency operating within gp120.

Although previous studies have examined coevolution within the *env* V3 loop (Korber et al. 1993; Bickel et al. 1996; Gilbert et al. 2005), it is not possible to perform a direct comparison between them and this study for a number of reasons. The methods used vary between each of the studies, and we have previously shown extreme differences in the sensitivities of a number of methods used to identify the presence of coevolution based on the properties of the data set (Fares and Travers 2006).

Coevolution Networks Demonstrate the Complexity of Evolution Operating within *env*

The extent of coevolution identified between amino acid residues throughout the *env* gene reflects the functional codependence of the gp120–gp41 trimer. We have shown that residues involved in both CD4 binding and in glycosylation showed a large degree of coevolution with residues throughout *env* (figs. 3 and 4). We have included amino acid residues that are directly proximal (<8 Å) to functional residues in these networks as changes within proximal residues can affect the structure, and therefore functionality, of essential residues (Gloor et al. 2005). The CD4 coevolution network (fig. 3) contains 8 residues that bind directly to CD4 as well as 1 residue that maps to the BMS-806 epitope, which interferes with gp120-CD4 binding (Pantophlet and Burton 2006). All the other residues within the network are directly proximal to CD4-binding sites, 5 of which are residues that are directly glycosylated. Similarly, within the glycosylation network (fig. 4), 11 residues are directly glycosylated and 1 residue corresponds to part of the 2G12 epitope (Trkola et al. 1996), with the remaining residues being directly proximal to directly glycosylated residues. Of these residues proximal to directly glycosylated sites, one of them when mutated interferes with chemokine receptor binding and 5 directly bind CD4 (Kwong et al. 1998; Rizzuto et al. 1998). In both the networks, 56% of residues are not directly functional yet exhibit a high level of coevolution within the network (figs. 3 and 4). This "backbone" of coevolving residues may maintain protein functionality through facilitating movement and communication throughout the gp120–gp41 trimer. The significant overlap of coevolving pairs between the CD4 and glucosylation networks further indicates the dependencies of evolution operating throughout the structure. Examining residues outside the CD4 network that coevolve with direct CD4-binding residues also shows a large degree of overlap with 6, 4, and 5 residues coevolving with 2, 3, and 4 CD4-binding residues, respectively (supplementary table 2, Supplementary Material online). Our results are overlapping with those presented by Poon et al. (2007), where they detect coevolution between N-linked glycosylated sites in the envelope protein of HIV-1 using a phylogenetic model and a Bayesian graphical model of evolution. Proximal coevolving amino acid sites can also indicate compensatory epistatic effects. Epistatis effects is an important factor to take into account when studying coevolution and in the understanding of the dynamics of adaptation (Shapiro et al. 2007). Compensatory mutations have been observed in the HIV genome with the majority associated in pol with drug resistance (Piana, Carloni, and Rothlisberger 2002; Menendez-Arias et al. 2003; Perno, Svicher, and Ceccherini-Silberstein 2006) as well as a number of compensatory mutations identified in gag (Friedrich et al. 2004; Yeh et al. 2006). A recent study by Gorry and colleagues, however, proposed the presence of compensatory mutations between residues 308R/317F and 308R/321G that affect coreceptor binding (Gorry et al. 2007). Our study did not observe direct covariation/coevolution between either of these pairs although residues proximal to both of these pairs were observed as coevolving (Figure 2). Similarly, Baldwin and Berkhout proposed a number of potential compensatory mutations in both gp120 and gp41 which enabled escape from T20-dependent replication (Baldwin and Berkhout 2006). The initial mutations that caused T20 dependency occurred as V549A and N637K. Multiple occurrences of a G431R mutation enabled T20-dependency escape were observed suggesting that compensatory mutations within the CD4 binding domain may affect T20 dependency. We have observed high levels of coevolution within the CD4 binding domain (Figure 3) and have also observed strong coevolution between 430V located in the CD4 binding domain and 567Q identified by Baldwin and Berkhout as an escape mutant from T20-dependent replication (Baldwin and Berkhout 2006). In addition to proximal coevolving glycosylated sites, we observed many of the coevolving pairs of sites to present distances in the structure above 4.5 Å. Coevolution between distant N-glycosylated sites may be convenient to ensure an efficient shielding through glycosylation of sites recognized by the host defense system as previously pointed out (Poon et al. 2007).

Although detection of coevolution is an interesting problem per se, the pragmatic value of detecting coevolution transcends many areas of research. The understanding of the molecular communication between the different proteins involved in infectivity and spread in HIV-1 is essential to identify functionally/structurally important protein

domains and hence to design proper therapeutics against the virus. This communication is only tractable from the evolutionary point of view, and in this sense, coevolution analysis can easily highlight such dependencies. In this study, we aimed at identifying these covariation dependencies in order to understand the evolutionary dynamic of the 2 most important proteins of the HIV-1 infection machinery.

Although we have been able to assign biological significance to many of the coevolving pairs and groups identified in this study, this is not always the case. Many of the residues in the gp120–gp41 trimer, although not directly functionally important, may be important in the maintenance of the protein in a functional conformation or may be involved in intra- or interprotein communication. Highly significant coevolving residues may provide ideal targets for future site-directed mutagenesis analysis in the identification of functional domains and have also been suggested as a strategy for the design of broadly neutralizing vaccines (Korber et al. 1993).

We propose not only that pairs/groups of coevolving amino acids are seen across the entire HIV-1 group M phylogeny but also that subtype-specific pairs/groups exist. In fact subtype specific patterns of evolution have been previously identified (Korber et al. 1994; Gaschen et al. 2002). The association of these subtypes specific evolutionary patterns and the structure characteristics of the protein have been elegantly examined in a recent work (Gnanakaran et al. 2007). However, coevolutionary patterns in specific subtypes have to be as yet comprehensively studied. Residues observed as coevolving across group M have been functionally/structurally constrained throughout the evolution of group M, whereas subtype-specific coevolving residues may represent novel dependencies within *env* for a particular subtype. Analysis of such subtype-specific dependencies may provide clues as to subtype-specific mechanisms' immune escape or infectivity.

Supplementary Material

Supplementary tables 1,2 and 3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Peter Kwong for supplying the coordinates of the gp120 trimer and Christopher J. Creevey, Jennifer Commins, and Christina Toft for assistance with the software used to produce figure 1. This study is supported by the President of Ireland Young Researcher Award program of Science Foundation Ireland awarded to M.A.F (04/YI1/M518). S.A.A.T. is supported by a Health Research Board Research Project Grant (RP/2006/141). We are also grateful to the editor and to two anonymous reviewers for their insightful comments on the manuscript.

Literature Cited

Abebe A, Demissie D, Goudsmit J, Brouwer M, Kuiken CL, Pollakis G, Schuitemaker H, Fontanet AL, Rinke de Wit TF. 1999. HIV-1 subtype C syncytium- and non-syncytium-inducing phenotypes and coreceptor usage among Ethiopian patients with AIDS. AIDS. 13:1305–1311.

Ane C, Burleigh JG, McMahon MM, Sanderson MJ. 2005. Covarion structure in plastid genome evolution: a new statistical test. Mol Biol Evol. 22:914–924.

Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW. 2000. Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. Mol Biol Evol. 17:164–178.

Bernstein HB, Tucker SP, Kar SR, McPherson SA, McPherson DT, Dubay JW, Lebowitz J, Compans RW, Hunter E. 1995. Oligomerization of the hydrophobic heptad repeat of gp41. J Virol. 69:2745–2750.

Bickel PJ, Cosman PC, Olshen RA, Spector PC, Rodrigo AG, Mullins JI. 1996. Covariability of V3 loop amino acids. AIDS Res Hum Retrovir. 12:1401–1411.

Boge M, Wyss S, Bonifacino JS, Thali M. 1998. A membrane-proximal tyrosine-based signal mediates internalization of the HIV-1 envelope glycoprotein via interaction with the AP-2 clathrin adaptor. J Biol Chem. 273:15773–15778.

Caffrey M, Cai M, Kaufman J, Stahl SJ, Wingfield PT, Covell DG, Gronenborn AM, Clore GM. 1998. Three-dimensional solution structure of the 44 kDa ectodomain of SIV gp41. EMBO J. 17:4572–4584.

Chelvanayagam G, Eggenschwiler A, Knecht L, Gonnet GH, Benner SA. 1997. An analysis of simultaneous variation in protein structures. Protein Eng. 10:307–316.

Chen B, Vogan EM, Gong H, Skehel JJ, Wiley DC, Harrison SC. 2005. Structure of an unliganded simian immunodeficiency virus gp120 core. Nature. 433:834–841.

Choisy M, Woelk CH, Guegan JF, Robertson DL. 2004. Comparative study of adaptive molecular evolution in different human immunodeficiency virus groups and subtypes. J Virol. 78:1962–1970.

Cilliers T, Nhlapo J, Coetzer M, Orlovic D, Ketas T, Olson WC, Moore JP, Trkola A, Morris L. 2003. The CCR5 and CXCR4 coreceptors are both used by human immunodeficiency virus type 1 primary isolates from subtype C. J Virol. 77:4449–4456.

Deng H, Liu R, Ellmeier W, et al. (15 co-authors). 1996. Identification of a major co-receptor for primary isolates of HIV-1. Nature. 381:661–666.

de Oliveira T, Salemi M, Gordon M, Vandamme AM, van Rensburg EJ, Engelbrecht S, Coovadia HM, Cassol S. 2004. Mapping sites of positive selection and amino acid diversification in the HIV genome: an alternative approach to vaccine design? Genetics. 167:1047–1058.

Dragic T, Litwin V, Allaway GP, et al. (11 co-authors). 1996. HIV-1 entry into CD4+ cells is mediated by the chemokine receptor CC-CKR-5. Nature. 381:667–673.

Dutheil J, Pupko T, Jean-Marie A, Galtier N. 2005. A model-based approach for detecting coevolving positions in a molecule. Mol Biol Evol. 22:1919–1928.

Fares MA. 2006. Computational and statistical methods to explore the various dimensions of protein evolution. Curr Bioinformatics. 1:207–217.

Fares MA, McNally D. 2006. CAPS: coevolution analysis using protein sequences. Bioinformatics. 22:2821–2822.

Fares MA, Travers SAA. 2006. A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses. Genetics. 173:9–23.

Feng Y, Broder CC, Kennedy PE, Berger EA. 1996. HIV-1 entry cofactor: functional cDNA cloning of a seven-transmembrane, G protein-coupled receptor. Science. 272:872–877.

Freed EO, Martin MA. 1995. The role of human immunodeficiency virus type 1 envelope glycoproteins in virus infection. J Biol Chem. 270:23883–23886.

Galtier N. 2004. Sampling properties of the bootstrap support in molecular phylogeny: influence of nonindependence among sites. Syst Biol. 53:38–46.
Gilbert PB, Novitsky V, Essex M. 2005. Covariability of selected amino acid positions for HIV type 1 subtypes C and B. AIDS Res Hum Retrovir. 21:1016–1030.
Gloor GB, Martin LC, Wahl LM, Dunn SD. 2005. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. Biochemistry. 44:7156–7165.
Gobel U, Sander C, Schneider R, Valencia A. 1994. Correlated mutations and residue contacts in proteins. Proteins. 18:309–317.
Hartley O, Klasse PJ, Sattentau QJ, Moore JP. 2005. V3: HIV's switch-hitter. AIDS Res Hum Retrovir. 21:171–189.
Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA. 89:10915–10919.
Hollier MJ, Dimmock NJ. 2005. The C-terminal tail of the gp41 transmembrane envelope glycoprotein of HIV-1 clades A, B, C, and D may exist in two conformations: an analysis of sequence, structure, and function. Virology. 337:284–296.
Holmes EC, Zhang LQ, Simmonds P, Ludlam CA, Brown AJ. 1992. Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a single infected patient. Proc Natl Acad Sci USA. 89:4835–4839.
Huang CC, Tang M, Zhang MY, et al. (12 co-authors). 2005. Structure of a V3-containing HIV-1 gp120 core. Science. 310:1025–1028.
Hwang SS, Boyle TJ, Lyerly HK, Cullen BR. 1991. Identification of the envelope V3 loop as the primary determinant of cell tropism in HIV-1. Science. 253:71–74.
Korber BT, Farber RM, Wolpert DH, Lapedes AS. 1993. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. Proc Natl Acad Sci USA. 90:7176–7180.
Kwong PD, Wyatt R, Majeed S, Robinson J, Sweet RW, Sodroski J, Hendrickson WA. 2000. Structures of HIV-1 gp120 envelope glycoproteins from laboratory-adapted and primary isolates. Structure Fold Des. 8:1329–1339.
Kwong PD, Wyatt R, Robinson J, Sweet RW, Sodroski J, Hendrickson WA. 1998. Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. Nature. 393:648–659.
Kwong PD, Wyatt R, Sattentau QJ, Sodroski J, Hendrickson WA. 2000. Oligomeric modeling and electrostatic analysis of the gp120 envelope glycoprotein of human immunodeficiency virus. J Virol. 74:1961–1972.
Li WH. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol. 36:96–99.
Lockhart PJ, Steel MA, Barbrook AC, Huson DH, Charleston MA, Howe CJ. 1998. A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. Mol Biol Evol. 15:1183–1188.
Maddison WP, Maddison DR. 1992. MacClade. Sunderland (MA): Sinauer.
Ohno H, Aguilar RC, Fournier MC, Hennecke S, Cosson P, Bonifacino JS. 1997. Interaction of endocytic signals from the HIV-1 envelope glycoprotein complex with members of the adaptor medium chain family. Virology. 238:305–315.
Otto C, Puffer BA, Pohlmann S, Doms RW, Kirchhoff F. 2003. Mutations in the C3 region of human and simian immunodeficiency virus envelope have differential effects on viral infectivity, replication, and CD4-dependency. Virology. 315:292–302.
Pantophlet R, Burton DR. 2006. GP120: target for neutralizing HIV-1 antibodies. Annu Rev Immunol.
Peeters M, Sharp PM. 2000. Genetic diversity of HIV-1: the moving target. AIDS. 14(Suppl 3):S129–S140.
Ping LH, Nelson JA, Hoffman IF, et al. (15 co-authors). 1999. Characterization of V3 sequence heterogeneity in subtype C human immunodeficiency virus type 1 isolates from Malawi: underrepresentation of X4 variants. J Virol. 73:6271–6281.
Poignard P, Saphire EO, Parren PW, Burton DR. 2001. gp120: biologic aspects of structural features. Annu Rev Immunol. 19:253–274.
Pollock DD, Taylor WR. 1997. Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. Protein Eng. 10:647–657.
Pollock DD, Taylor WR, Goldman N. 1999. Coevolving protein residues: maximum likelihood identification and relationship to structure. J Mol Biol. 287:187–198.
Poon AF, Lewis FI, Pond SL, Frost SD. 2007. Evolutionary interactions between N-linked glycosylation sites in the HIV-1 envelope. PLoS Comput Biol. 3:e11.
Posada D, Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics. 14:817–818.
Pritchard L, Bladon P, MOMJ. JDM. 2001. Evaluation of a novel method for the identification of coevolving protein residues. Protein Eng. 14:549–555.
Resch W, Hoffman N, Swanstrom R. 2001. Improved success of phenotype prediction of the human immunodeficiency virus type 1 from envelope variable loop 3 sequence using neural networks. Virology. 288:51–62.
Rizzuto CD, Wyatt R, Hernandez-Ramos N, Sun Y, Kwong PD, Hendrickson WA, Sodroski J. 1998. A conserved HIV gp120 glycoprotein structure involved in chemokine receptor binding. Science. 280:1949–1953.
Rowell JF, Stanhope PE, Siliciano RF. 1995. Endocytosis of endogenously synthesized HIV-1 envelope protein. Mechanism and role in processing for association with class II MHC. J Immunol. 155:473–488.
Seibert SA, Howell CY, Hughes MK, Hughes AL. 1995. Natural selection on the gag, pol, and env genes of human immunodeficiency virus 1 (HIV-1). Mol Biol Evol. 12:803–813.
Shindyalov IN, Kolchanov NA, Sander C. 1994. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Eng. 7:349–358.
Swofford DL. 1998. PAUP*: phylogenetic analysis using parsimony (*and other methods). Sunderland (MA): Sinauer Associates.
Taylor WR, Hatrick K. 1994. Compensating changes in protein multiple sequence alignments. Protein Eng. 7:341–348.
Tillier ER, Collins RA. 1995. Neighbor joining and maximum likelihood with RNA sequences: addressing the interdependence of sites. Mol Biol Evol. 12:7–15.
Tillier ER, Lui TW. 2003. Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. Bioinformatics. 19:750–755.
Travers SA, Fares MA. 2007. Functional coevolutionary networks of the Hsp70-Hop-Hsp90 system revealed through computational analyses. Mol Biol Evol. 24:1032–1044.
Travers SAA, O'Connell MJ, McCormack GP, McInerney JO. 2005. Evidence for heterogeneous selective pressures in the evolution of the env gene in different human immunodeficiency virus type 1 subtypes. J Virol. 79:1836–1841.
Trkola A, Purtscher M, Muster T, Ballaun C, Buchacher A, Sullivan N, Srinivasan K, Sodroski J, Moore JP, Katinger H. 1996. Human monoclonal antibody 2G12 defines a distinctive neutralization epitope on the gp120 glycoprotein of human immunodeficiency virus type 1. J Virol. 70:1100–1108.

Tuffley C, Steel M. 1998. Modeling the covarion hypothesis of nucleotide substitution. Math Biosci. 147:63–91.

Tully DC, Fares MA. 2006. Unravelling selection shifts among foot-and-mouth disease virus (FMDV) serotypes. Evol Bioinform Online. 2:223–337.

Wei X, Decker JM, Wang S, et al. (15 co-authors). 2003. Antibody neutralization and escape by HIV-1. Nature. 422: 307–312.

Westfall PH, Young SS. 1993. Resampling-based multiple testing. New York: John Wiley & Sons.

Wyatt R, Desjardin E, Olshevsky U, Nixon C, Binley J, Olshevsky V, Sodroski J. 1997. Analysis of the interaction of the human immunodeficiency virus type 1 gp120 envelope glycoprotein with the gp41 transmembrane glycoprotein. J Virol. 71:9722–9731.

Wyatt R, Kwong PD, Desjardins E, Sweet RW, Robinson J, Hendrickson WA, Sodroski JG. 1998. The antigenic structure of the HIV gp120 envelope glycoprotein. Nature. 393:705–711.

Yamaguchi-Kabata Y, Gojobori T. 2000. Reevaluation of amino acid variability of the human immunodeficiency virus type 1 gp120 envelope glycoprotein and prediction of new discontinuous epitopes. J Virol. 74:4335–4350.

Yang W, Bielawski JP, Yang Z. 2003. Widespread adaptive evolution in the human immunodeficiency virus type 1 genome. J Mol Evol. 57:212–221.

Yang Z. 2001. Maximum likelihood analysis of adaptive evolution in HIV-1 gp120 env gene. Pac Symp Biocomput. 226–237.

Edward Holmes, Associate Editor

Accepted September 26, 2007