MBE Advance Access published October 5, 2007

A study of the coevolutionary patterns operating within the *env* gene of the HIV-1 group M subtypes

Travers S. A. A.[1]†^, Tully D. C.[2]†, McCormack G. P.[3] and Fares M. A[2*]

[1] Molecular Evolution and Bioinformatics Laboratory, Department of Biology, National University of Ireland, Maynooth, Ireland

[2] Evolutionary Genetics and Bioinformatics Laboratory, Department of Genetics, Smurfit Institute of Genetics, University of Dublin, Trinity College, Ireland

[3] Molecular Evolution & Systematics Laboratory, Martin Ryan Institute, National University of Ireland, Galway, Ireland

*Corresponding Author: Dr. Mario A. Fares

Tel: +353 01 8963521
Fax: +353 01 6798558

Email: faresm@tcd.ie

† S.A.A.T and D.C.T. are joined first authors of this work

^ Present address: Martin Ryan Institute, Department of Zoology, National University of Ireland Galway, Ireland

Abstract

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 communications. 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 of the HIV-1 group 1 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 by 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 two to ten 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 of 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.

Abbreviations used

NHR: N-terminal heptad repeat
CHR: C-terminal heptad repeat
FP: fusion peptide
HIV: Human Immunodeficiency Virus

Introduction

Human immunodeficiency virus type I (HIV-I) 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 160kD protein (gp160) that is cleaved to produce gp120 and gp41, which exist as a trimeric spike on the surface of the HIV virion containing three exterior gp120 and three 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, Hoffman, and Swanstrom 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 host cell membranes (Caffrey et al.

1998). gp41 is known to comprise of four 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 two conformations, with gp41 molecules incorporated into active virions actually containing two ectodomains (termed the major and minor ectodomains) and three 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, Bielawski, and Yang 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, intra-molecular 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 intra-protein 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* may also aid in the identification of domains important in intra/inter protein communication as well as domains important in protein-protein

interaction. While 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, Kolchanov, and Sander 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, Taylor, and Goldman 1999; Pritchard et al. 2001; Tillier and Lui 2003; Galtier 2004; Ane et al. 2005; Dutheil et al. 2005; Gloor et al. 2005). However, many of these methods 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.

Previous studies examining coevolution within HIV-1 the *env* gene have been limited to the gp120 V3 loop (Korber et al. 1993; Bickel et al. 1996; Gilbert, Novitsky, and Essex 2005). Korber and colleagues studied 308 subtype B sequences while Bickel and colleagues reanalyzed the Korber data as well as a new dataset containing 440 sequences that represented a number of HIV-1 group M subtypes (A, C, D and E) (Korber et al. 1993; Bickel et al. 1996). Upon reanalyzing Korber and colleagues'

subtype B dataset Bickel and colleagues identified four of the seven 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 dataset containing multiple subtypes and the subtype B dataset (Bickel et al. 1996). 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 and colleagues observed significant differences between the number of coevolving pairs identified in their *env* subtype B (26 pairs) and subtype C (one pair) datasets (Gilbert, Novitsky, and Essex 2005). From the results presented in these three 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, Novitsky, and Essex 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 to 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 represents 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 (manuscript in preparation). While 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 dataset 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 and Maddison 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 intra-subtype 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 four 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 and Maddison 1992). Ambiguous regions of the alignment were removed to avoid false positives due to erroneous alignment of non-homologous 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, 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 neighbour joining tree for the resulting datasets was reconstructed using PAUP* 4.0b10 (Swofford 1998). We used this dataset to examine coevolving pairs present over the entire HIV-1 group M epidemic.

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 dataset 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 datasets. We also compared the nucleotide mean pairwise nucleotide distances between the full HIV-1 dataset and the representative dataset 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 two dataset, with one containing all the sequences available for three 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 dataset due to computational limitations of the program to run over 700 sequences alignment. Finally, to discard any effect of he number of sequences in the coevolutionary analyses, we have also conducted these analyses on different subsets of the 705 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 inter-protein molecular coevolution

To test for intra- and inter-molecular 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 datasets containing similar numbers of sequences 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 two sites corrected by the time since the divergence of the two sequences they belong to. This method compares the transition probability scores between two sequences at two particular sites, using the blocks substitution matrix (BLOSUM) (Henikoff and Henikoff 1992). The significance of the CAPS correlation values was assessed by randomisation of pairs of sites in the alignment, calculation of their correlation values and comparison of the real values with the distribution of ten thousand 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 numbering presented here corresponds 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 inter-subtype recombinant sequences that may bias the results obtained from the coevolution analyses. We have also attempted to run CAPS on the complete HIV-I sequence dataset. However, because CAPS is a very computationally intensive method, computers could not run this program on such a large dataset. 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 to those obtained when we ran CAPS on a dataset comprising the representative sequences of these three 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 three 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. 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 co-evolution 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 three-dimensional 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). 3D structure viewing and manipulation was performed using iMOL (http://www.pirx.com/iMol).

We used a conservative mean distance of 8Å in determining whether two amino acid residues were significantly proximal in the 3D structure. The relative distance between two amino acids was calculated by taking the mean three-dimensional atomic coordinates for each amino acid in the structure. We then calculated the distance between two 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 an 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 and colleagues (Korber et al. 1993), 5.24% observed by Gilbert and colleagues for their subtype B dataset (Gilbert, Novitsky, and Essex 2005) and 12.69% observed by Bickel and colleagues in both their 308 and 440 datasets (Bickel et al. 1996). 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 two to ten amino acid residues. The majority of these groups (72%), however, contained either two or three residues (**Supplementary Table 1**).

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 764 amino acid sites) identified in *env*, 311 and 268 of these were correlated by hydrophobicity and molecular weight respectively, while 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 (**Figure 1**). Whilst coevolving pairs were spread throughout *env*, two 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 (**Figure 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, Novitsky, and Essex 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 (**Figure 2**). Of the 24 pairs of coevolving residues, four (17%) of these were significantly correlated by hydrophobicity, while 13 (54%) 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 and colleagues (Bickel et al. 1996) in their 440 dataset, which contained representative sequences from multiple subtypes.

Upon solving the structure of a V3-containing HIV-1 gp120 core, Huang and colleagues proposed that, following CD4 binding, the N-terminus of the CCR5 receptor binds the gp120 core and V3 base while the V3 tip binds the coreceptor's second extracellular loop (Huang et al. 2005). Complimentary to this, Rizzuto and colleagues identified a number of residues within the gp120 core the mutation of which significantly affects CCR5 coreceptor binding by gp120 (Rizzuto et al. 1998).

Many of these residues were not identified as coevolving within *env*, most likely because of functional conservation. Rizzuto and colleagues identified two residues, P437 and Q442, that, when mutated, result in a ≥50% increase in CCR5 binding with respect to wild-type gp120 (Rizzuto et al. 1998). Both of 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 three residues in the V3 loop, two in the tip (R315 and A316) and one in the base (Q328). Mutation of V200 showed a slight decrease (16%) in CCR5 binding when compared to wild type (Rizzuto et al. 1998). Residue G379, while not tested by Rizzuto and colleagues, is directly adjacent in the gp120 three-dimensional structure to E381 (5.38Å) 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 eight residues within the V3 loop and five residues within the proposed coreceptor binding domain on the gp120 core that coevolve with residues that either bind directly to CD4 or are proximally contained within the CD4 binding pocket (<8Å 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 and colleagues (Wyatt et al. 1998). We have also included amino acid residues, which may be functionally proximal (<8Å) 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 (**Figure 3**), eight of which bind CD4 directly, one that maps to the BMS-806 epitope and five residues that are directly glycosylated in the HIV or SIV structures (these are located <8Å 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 (five pairs), molecular weight (two 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 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 (**Figure 4**). The coevolution of 67% of the 42 coevolving pairs in the glycosylation network was correlated by hydrophobicity (eight pairs), molecular weight (seven pairs) or by both hydrophobicity and molecular weight (13 pairs). These results support that coevolution between or nearby N-glycosylated 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 CD4 and glycosylation networks with 63 pairs of coevolving residues observed between these. Of the 63 coevolving pairs, 12 were present in both networks, 42 were present in one of the networks while nine novel pairs had not been identified in either the CD4 or 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 and colleagues observed large displacements within the gp120 core inner domain and the absence of the bridging sheet in the unliganded structure. With the exception of two regions, the orientation of the outer domain remained essentially the same between the two structures. Using both the liganded and unliganded gp120 structures we looked for coevolving residues that showed a

significant difference (>8Å difference) in their mean pairwise distances between the two structures. We identified six coevolving pairs that were significantly more proximal in CD4 bound gp120 than in unliganded gp120 and three coevolving pairs that were significantly closer in the unliganded gp120 structure (**Table 4**). With the exception of two pairs (P369&G379 and E102&S364), only one of the coevolving residues in each pair is located within a domain identified by Chen and colleagues as moving significantly following CD4 binding (Chen et al. 2005).

Inter gp120-gp41 coevolution

The three-dimensional structure of the ectodomain of SIV gp41 has been solved and combining the properties of this structure with previous mutagenesis analyses Caffrey and colleagues proposed that the gp120-binding domain is located within a hydrophobic patch within the gp41 loop domain (Caffrey et al. 1998). We observed four residues that map to this hydrophobic patch (K588, T605, A607 and A612) as coevolving with residues elsewhere within both gp120 and gp41 (**Table 5, Figure 5**). Residue K588 coevolves with three residues (M535, Q543 and H564), all of which map to the FP/NHR domain within the gp41 ectodomain (**Figure 5A**) 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, while located on the CHR, appears to be located in the corresponding position on the CHR as Q543 is on the NHR (**Figure 5A**).

Both T605 and A612 coevolve with V270 and N339 which map to the highly glycosylated outer domain of the gp120 core structure, while K588 coevolves with F277, P369 and T373 all of which map to the CD4 binding domain in gp120 (**Figure**

5B). With the exception of W17 and K683 the remaining five 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 of 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)** 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 emphasize, however, that the observation of coevolution between two 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).

While still not fully understood, the functional complexities of the gp120-gp41 complex have been well documented (Wyatt et al. 1997; Kwong et al. 2000b; 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, multi-faceted 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 of the HIV-1 group M subtypes. While previous studies have concentrated on examining coevolution within the gp120 V3 loop (Korber et al. 1993; Bickel et al. 1996; Gilbert, Novitsky, and Essex 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 to 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 (Fares and Travers 2006). In contrast to previous studies we have studied a dataset 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 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 by 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 6.7Å 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 sites 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 inter-subtype recombination, excluding intra-subtype recombination remains a problem. However, currently there is no way to identify intra-subtype recombination and, as with all analyses with HIV-1 group M multiple sequence alignments, therefore some intra-subtype recombinants may be present in the data.

Discussion of the biological significance of the complete set of coevolving pairs (848 pairs) is impossible in this 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 non-discussed 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, while 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

While coevolving residues are spread throughout the *env* gene we did observe two regions that present a higher density of coevolution with residues throughout the *env* gene (**Figure 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 et al. 2000a; 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 inter-domain 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 detailed a number of studies which 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 proposed a structural model for gp41 that consists of three membrane spanning domains (MSDs) and two ectodomains, a major and a minor. The Kennedy sequence is exposed on the outer face of the proposed minor ectodomain. Hollier and Dimmock 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 nine residues within gp41 that show high levels of coevolution with residues elsewhere in *env*, five of these (L721, G726, E731, G732 and I746) are located in the minor ectodomain with three 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 five residues located within the minor ectodomain and residues elsewhere in *env*. The remaining four 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 of 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 (AP-2) complex (Ohno et al. 1997; Boge et al. 1998), however it is not known whether this

signal is functional within gp41 (Rowell, Stanhope, and Siliciano 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 30Å from the gp120 trimer suggesting that perhaps the N-terminus of the CCR5 receptor binds the gp120 core and V3 base while 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 (**Figure 2**). There is also a large degree of coevolution involving residues within the V3 loop stem (**Figure 2**), the majority of which correlate by hydrophobicity or by both hydrophobicity and molecular weight, probably as result of the functional and structural constraints imposed on this functional domain. Resch and colleagues 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 while others direct CCR5 usage (Resch, Hoffman, and Swanstrom 2001). These positions exhibit a high degree of coevolution with residues both within the V3 loop (**Figure 2**) and elsewhere in *env*. For example, R306 coevolves with 21 residues within *env*, three 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 eight residues in *env*, five within the V3 loop (**Figure 2**) and of the remaining three one is located in the gp120 V2 loop (R166) and two 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 of the V3 loop suggested to be involved in coreceptor binding corroborates the level of evolutionary functional dependency operating within gp120.

While previous studies have examined coevolution within the *env* V3 loop (Korber et al. 1993; Bickel et al. 1996; Gilbert, Novitsky, and Essex 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 dataset (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 co-dependence 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* (**Figures 3 & 4**). We have included amino acid residues that are directly proximal (<8Å) 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 (**Figure 3**) contains eight residues that bind directly to CD4 as well as one residue that maps to the BMS-806 epitope, which interferes with gp120-CD4 binding (Pantophlet and Burton 2006). All of the other residues within the network are directly proximal to CD4 binding sites, five of which are residues that are directly glycosylated. Similarly, within the glycosylation network (**Figure 4**) 11 residues are directly glycosylated and one 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 five directly bind CD4 (Kwong et al. 1998; Rizzuto et al. 1998). In both of the networks 56% of residues are not directly functional yet exhibit a high level of coevolution within the network (**Figures 3 & 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 glycosylation networks further indicates the dependencies of evolution operating throughout the structure. Examining residues outside of the CD4 network that coevolve with direct CD4 binding residues also shows a large degree of overlap with six, four and five residues coevolving with two, three and four CD4 binding residues respectively (**Supplementary Table 2**). Our results are overlapping with those presented by Poon and coworkers, where they detect coevolution between

N-linked glycosylated sites in the envelope protein of HIV-1 using a phylogenetic and a Bayesian graphical models of evolution (Poon et al. 2007). 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 two most important proteins of the HIV-I infection machinery.

While 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, while not directly functionally important, may be important in the maintenance of the protein in a functional conformation or may be involved in intra- or inter-protein 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 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 while 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.

Acknowledgements

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.

REFERENCES

Abebe, A., D. Demissie, J. Goudsmit, M. Brouwer, C. L. Kuiken, G. Pollakis, H. Schuitemaker, A. L. Fontanet, and T. F. Rinke de Wit. 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., J. G. Burleigh, M. M. McMahon, and M. J. Sanderson. 2005. Covarion structure in plastid genome evolution: a new statistical test. Mol Biol Evol 22:914-924.

Atchley, W. R., K. R. Wollenberg, W. M. Fitch, W. Terhalle, and A. W. Dress. 2000. Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. Mol Biol Evol 17:164-178.

Baldwin, C. E., and B. Berkhout. 2006. Second site escape of a T20-dependent HIV-1 variant by a single amino acid change in the CD4 binding region of the envelope glycoprotein. Retrovirology 3:84.

Bernstein, H. B., S. P. Tucker, S. R. Kar, S. A. McPherson, D. T. McPherson, J. W. Dubay, J. Lebowitz, R. W. Compans, and E. Hunter. 1995. Oligomerization of the hydrophobic heptad repeat of gp41. J Virol 69:2745-2750.

Bickel, P. J., P. C. Cosman, R. A. Olshen, P. C. Spector, A. G. Rodrigo, and J. I. Mullins. 1996. Covariability of V3 loop amino acids. AIDS Res Hum Retroviruses 12:1401-1411.

Boge, M., S. Wyss, J. S. Bonifacino, and M. Thali. 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., M. Cai, J. Kaufman, S. J. Stahl, P. T. Wingfield, D. G. Covell, A. M. Gronenborn, and G. M. Clore. 1998. Three-dimensional solution structure of the 44 kDa ectodomain of SIV gp41. Embo J 17:4572-4584.

Chelvanayagam, G., A. Eggenschwiler, L. Knecht, G. H. Gonnet, and S. A. Benner. 1997. An analysis of simultaneous variation in protein structures. Protein Eng 10:307-316.

Chen, B., E. M. Vogan, H. Gong, J. J. Skehel, D. C. Wiley, and S. C. Harrison. 2005. Structure of an unliganded simian immunodeficiency virus gp120 core. Nature 433:834-841.

Choisy, M., C. H. Woelk, J. F. Guegan, and D. L. Robertson. 2004. Comparative study of adaptive molecular evolution in different human immunodeficiency virus groups and subtypes. J Virol 78:1962-1970.

Cilliers, T., J. Nhlapo, M. Coetzer, D. Orlovic, T. Ketas, W. C. Olson, J. P. Moore, A. Trkola, and L. Morris. 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.

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

Deng, H., R. Liu, W. Ellmeier, S. Choe, D. Unutmaz, M. Burkhart, P. Di Marzio, S. Marmon, R. E. Sutton, C. M. Hill, C. B. Davis, S. C. Peiper, T. J.

Schall, D. R. Littman, and N. R. Landau. 1996. Identification of a major co-receptor for primary isolates of HIV-1. Nature 381:661-666.

Dragic, T., V. Litwin, G. P. Allaway, S. R. Martin, Y. Huang, K. A. Nagashima, C. Cayanan, P. J. Maddon, R. A. Koup, J. P. Moore, and W. A. Paxton. 1996. HIV-1 entry into CD4+ cells is mediated by the chemokine receptor CC-CKR-5. Nature 381:667-673.

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

Fares, M. A. 2006. Computational and Statistical Methods to Explore the Various Dimensions of Protein Evolution. Current Bioinformatics 1:207-217.

Fares, M. A., and D. McNally. 2006. CAPS: coevolution analysis using protein sequences. Bioinformatics 22:2821-2822.

Fares, M. A., and S. A. A. Travers. 2006. A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses. Genetics 173:9-23.

Feng, Y., C. C. Broder, P. E. Kennedy, and E. A. Berger. 1996. HIV-1 entry cofactor: functional cDNA cloning of a seven-transmembrane, G protein-coupled receptor. Science 272:872-877.

Freed, E. O., and M. A. Martin. 1995. The role of human immunodeficiency virus type 1 envelope glycoproteins in virus infection. J Biol Chem 270:23883-23886.

Friedrich, T. C., C. A. Frye, L. J. Yant, D. H. O'Connor, N. A. Kriewaldt, M. Benson, L. Vojnov, E. J. Dodds, C. Cullen, R. Rudersdorf, A. L. Hughes, N. Wilson, and D. I. Watkins. 2004. Extraepitopic compensatory substitutions partially restore fitness to simian immunodeficiency virus variants that escape from an immunodominant cytotoxic-T-lymphocyte response. J Virol 78:2581-2585.

Galtier, N. 2004. Sampling properties of the bootstrap support in molecular phylogeny: influence of nonindependence among sites. Syst Biol 53:38-46.

Gaschen, B., J. Taylor, K. Yusim, B. Foley, F. Gao, D. Lang, V. Novitsky, B. Haynes, B. H. Hahn, T. Bhattacharya, and B. Korber. 2002. Diversity considerations in HIV-1 vaccine selection. Science 296:2354-2360.

Gilbert, P. B., V. Novitsky, and M. Essex. 2005. Covariability of selected amino acid positions for HIV type 1 subtypes C and B. AIDS Res Hum Retroviruses 21:1016-1030.

Gloor, G. B., L. C. Martin, L. M. Wahl, and S. D. Dunn. 2005. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. Biochemistry 44:7156-7165.

Gnanakaran, S., D. Lang, M. Daniels, T. Bhattacharya, C. A. Derdeyn, and B. Korber. 2007. Clade-specific differences between human immunodeficiency virus type 1 clades B and C: diversity and correlations in C3-V4 regions of gp120. J Virol 81:4886-4891.

Gobel, U., C. Sander, R. Schneider, and A. Valencia. 1994. Correlated mutations and residue contacts in proteins. Proteins 18:309-317.

Gorry, P. R., R. L. Dunfee, M. E. Mefford, K. Kunstman, T. Morgan, J. P. Moore, J. R. Mascola, K. Agopian, G. H. Holm, A. Mehle, J. Taylor, M. Farzan, H. Wang, P. Ellery, S. J. Willey, P. R. Clapham, S. M. Wolinsky, S. M. Crowe, and D. Gabuzda. 2007. Changes in the V3 region of gp120

contribute to unusually broad coreceptor usage of an HIV-1 isolate from a CCR5 Delta32 heterozygote. Virology 362:163-178.

Hartley, O., P. J. Klasse, Q. J. Sattentau, and J. P. Moore. 2005. V3: HIV's switch-hitter. AIDS Res Hum Retroviruses 21:171-189.

Henikoff, S., and J. G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89:10915-10919.

Hollier, M. J., and N. J. Dimmock. 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, E. C., L. Q. Zhang, P. Simmonds, C. A. Ludlam, and A. J. Brown. 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 U S A 89:4835-4839.

Huang, C. C., M. Tang, M. Y. Zhang, S. Majeed, E. Montabana, R. L. Stanfield, D. S. Dimitrov, B. Korber, J. Sodroski, I. A. Wilson, R. Wyatt, and P. D. Kwong. 2005. Structure of a V3-containing HIV-1 gp120 core. Science 310:1025-1028.

Hwang, S. S., T. J. Boyle, H. K. Lyerly, and B. R. Cullen. 1991. Identification of the envelope V3 loop as the primary determinant of cell tropism in HIV-1. Science 253:71-74.

Korber, B. T., R. M. Farber, D. H. Wolpert, and A. S. Lapedes. 1993. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. Proc Natl Acad Sci U S A 90:7176-7180.

Korber, B. T., K. MacInnes, R. F. Smith, and G. Myers. 1994. Mutational trends in V3 loop protein sequences observed in different genetic lineages of human immunodeficiency virus type 1. J Virol 68:6730-6744.

Kwong, P. D., R. Wyatt, S. Majeed, J. Robinson, R. W. Sweet, J. Sodroski, and W. A. Hendrickson. 2000a. Structures of HIV-1 gp120 envelope glycoproteins from laboratory-adapted and primary isolates. Structure Fold Des 8:1329-1339.

Kwong, P. D., R. Wyatt, J. Robinson, R. W. Sweet, J. Sodroski, and W. A. Hendrickson. 1998. Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. Nature 393:648-659.

Kwong, P. D., R. Wyatt, Q. J. Sattentau, J. Sodroski, and W. A. Hendrickson. 2000b. Oligomeric modeling and electrostatic analysis of the gp120 envelope glycoprotein of human immunodeficiency virus. J Virol 74:1961-1972.

Li, W. H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol 36:96-99.

Lockhart, P. J., M. A. Steel, A. C. Barbrook, D. H. Huson, M. A. Charleston, and C. J. Howe. 1998. A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. Mol Biol Evol 15:1183-1188.

Maddison, W. P., and D. R. Maddison. 1992. MacClade. Sinauer, Sunderland, MA.

Menendez-Arias, L., M. A. Martinez, M. E. Quinones-Mateu, and J. Martinez-Picado. 2003. Fitness variations and their impact on the evolution of

antiretroviral drug resistance. Curr Drug Targets Infect Disord 3:355-371.

Ohno, H., R. C. Aguilar, M. C. Fournier, S. Hennecke, P. Cosson, and J. S. Bonifacino. 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., B. A. Puffer, S. Pohlmann, R. W. Doms, and F. Kirchhoff. 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., and D. R. Burton. 2006. GP120: Target for Neutralizing HIV-1 Antibodies. Annu Rev Immunol.

Peeters, M., and P. M. Sharp. 2000. Genetic diversity of HIV-1: the moving target. Aids 14 Suppl 3:S129-140.

Perno, C. F., V. Svicher, and F. Ceccherini-Silberstein. 2006. Novel drug resistance mutations in HIV: recognition and clinical relevance. AIDS Rev 8:179-190.

Piana, S., P. Carloni, and U. Rothlisberger. 2002. Drug resistance in HIV-1 protease: Flexibility-assisted mechanism of compensatory mutations. Protein Sci 11:2393-2402.

Ping, L. H., J. A. Nelson, I. F. Hoffman, J. Schock, S. L. Lamers, M. Goodman, P. Vernazza, P. Kazembe, M. Maida, D. Zimba, M. M. Goodenow, J. J. Eron, Jr., S. A. Fiscus, M. S. Cohen, and R. Swanstrom. 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., E. O. Saphire, P. W. Parren, and D. R. Burton. 2001. gp120: Biologic aspects of structural features. Annu Rev Immunol 19:253-274.

Pollock, D. D., and W. R. Taylor. 1997. Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. Protein Eng 10:647-657.

Pollock, D. D., W. R. Taylor, and N. Goldman. 1999. Coevolving protein residues: maximum likelihood identification and relationship to structure. J Mol Biol 287:187-198.

Poon, A. F., F. I. Lewis, S. L. Pond, and S. D. Frost. 2007. Evolutionary Interactions between N-Linked Glycosylation Sites in the HIV-1 Envelope. PLoS Comput Biol 3:e11.

Posada, D., and K. A. Crandall. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817-818.

Pritchard, L., P. Bladon, M. O. M. J, and J. D. M. 2001. Evaluation of a novel method for the identification of coevolving protein residues. Protein Eng 14:549-555.

Resch, W., N. Hoffman, and R. Swanstrom. 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, C. D., R. Wyatt, N. Hernandez-Ramos, Y. Sun, P. D. Kwong, W. A. Hendrickson, and J. Sodroski. 1998. A conserved HIV gp120 glycoprotein structure involved in chemokine receptor binding. Science 280:1949-1953.

Rowell, J. F., P. E. Stanhope, and R. F. Siliciano. 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, S. A., C. Y. Howell, M. K. Hughes, and A. L. Hughes. 1995. Natural selection on the gag, pol, and env genes of human immunodeficiency virus 1 (HIV-1). Mol Biol Evol 12:803-813.

Shapiro, B., A. Rambaut, O. G. Pybus, and E. C. Holmes. 2007. A phylogenetic method for detecting positive epistasis in gene sequences and its application to RNA virus evolution. Mol Biol Evol 23:1724-1730.

Shindyalov, I. N., N. A. Kolchanov, and C. Sander. 1994. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Eng 7:349-358.

Swofford, D. L. 1998. PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods). Sinauer Associates, Sunderland, Massachusetts.

Taylor, W. R., and K. Hatrick. 1994. Compensating changes in protein multiple sequence alignments. Protein Eng 7:341-348.

Tillier, E. R., and R. A. Collins. 1995. Neighbor Joining and Maximum Likelihood with RNA Sequences: Addressing the Interdependence of Sites. Mol Biol Evol 12:7-15.

Tillier, E. R., and T. W. Lui. 2003. Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. Bioinformatics 19:750-755.

Travers, S. A., and M. A. Fares. 2007. Functional Coevolutionary Networks of the Hsp70-Hop-Hsp90 System Revealed through Computational Analyses. Mol Biol Evol 24:1032-1044.

Travers, S. A. A., M. J. O'Connell, G. P. McCormack, and J. O. McInerney. 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., M. Purtscher, T. Muster, C. Ballaun, A. Buchacher, N. Sullivan, K. Srinivasan, J. Sodroski, J. P. Moore, and H. Katinger. 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., and M. Steel. 1998. Modeling the covarion hypothesis of nucleotide substitution. Math Biosci 147:63-91.

Tully, D. C., and M. A. Fares. 2006. Unravelling selection shifts among foot-and-mouth disease virus (FMDV) serotypes. Evolutionary Bioinformatics Online 1:223-337.

Wei, X., J. M. Decker, S. Wang, H. Hui, J. C. Kappes, X. Wu, J. F. Salazar-Gonzalez, M. G. Salazar, J. M. Kilby, M. S. Saag, N. L. Komarova, M. A. Nowak, B. H. Hahn, P. D. Kwong, and G. M. Shaw. 2003. Antibody neutralization and escape by HIV-1. Nature 422:307-312.

Westfall, P. H., and S. S. Young. 1993. Resampling-Based Multiple Testing. John Wiley & Sons, New York.

Wyatt, R., E. Desjardin, U. Olshevsky, C. Nixon, J. Binley, V. Olshevsky, and J. Sodroski. 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., P. D. Kwong, E. Desjardins, R. W. Sweet, J. Robinson, W. A. Hendrickson, and J. G. Sodroski. 1998. The antigenic structure of the HIV gp120 envelope glycoprotein. Nature 393:705-711.

Yamaguchi-Kabata, Y., and T. Gojobori. 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., J. P. Bielawski, and Z. Yang. 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.

Yeh, W. W., E. M. Cale, P. Jaru-Ampornpan, C. I. Lord, F. W. Peyerl, and N. L. Letvin. 2006. Compensatory substitutions restore normal core assembly in simian immunodeficiency virus isolates with Gag epitope cytotoxic T-lymphocyte escape mutations. J Virol 80:8168-8177.

Table 1. Mean nucleotide pairwise substitutions per site for the different subtypes used in this study. 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 two sequences were available for a particular subtype, standard errors could not be calculated (NA).

	Representative Dataset		**All Available Sequences**	
	Mean	***Std Error***	***Mean***	***Std Error***
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

Table 2: Details of the coevolving pairs whose coevolution was correlated by hydrophobicity, molecular weight or by both. 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.

	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

Table 3: Residues in the V3 loop and proposed CCR5 binding domain that coevolve with residues involved in CD4 binding. Also shown in brackets are the closest adjacent CD4 binding residues and the pairwise distance in angstroms (Å).

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

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. Distances are shown in angstroms (Å). Also shown are residues which exhibit movement between the liganded HIV and unliganded SIV gp120 structures (~) as are pairs whose coevolution is correlated by molecular weight (*) or by both hydrophobicity and molecular weight (^)

Coevolving pair	Mean pairwise distance in liganded gp120 (Å)	Mean pairwise distance in liganded gp120 (Å)
Coevolving residues proximal in liganded gp120.		
92N~ & 271V^	19.0935	28.7078
111L~ & 265L^	24.1784	32.5983
122L~ & 471G^	24.5348	34.9453
240T~ & 348K	22.4094	31.5596
369P~ & 379G~	19.143	32.4727
369P~ & 440S*	23.4603	35.4941
Coevolving residues proximal in unliganded gp120.		
87V & 106E~	31.0536	22.5208
87V & 364S~	43.9368	29.8071
102E~ & 364S~	26.6992	16.1491

Table 5: Residues within the proposed gp120 binding hydrophobic patch of the gp41 ectodomain and residues elsewhere in *env* that they coevolve with. Pairs whose coevolution correlates by hydrophobicity[1], molecular weight[2] and both hydrophobicity and molecular weight[3] are marked.

gp41 hydrophobic patch residues	Coevolving residues
588K	17W[3], 277F[2], 369P, 373T[3], 535M[2], 543Q[1], 564H[3], 662E[3], 726G, 746I[3], 758D[3], 777I[3]
605T	270V[1], 339N[1], 683K[3]
607A	778V
612A	270V[2], 339N[3]

Figure Legends

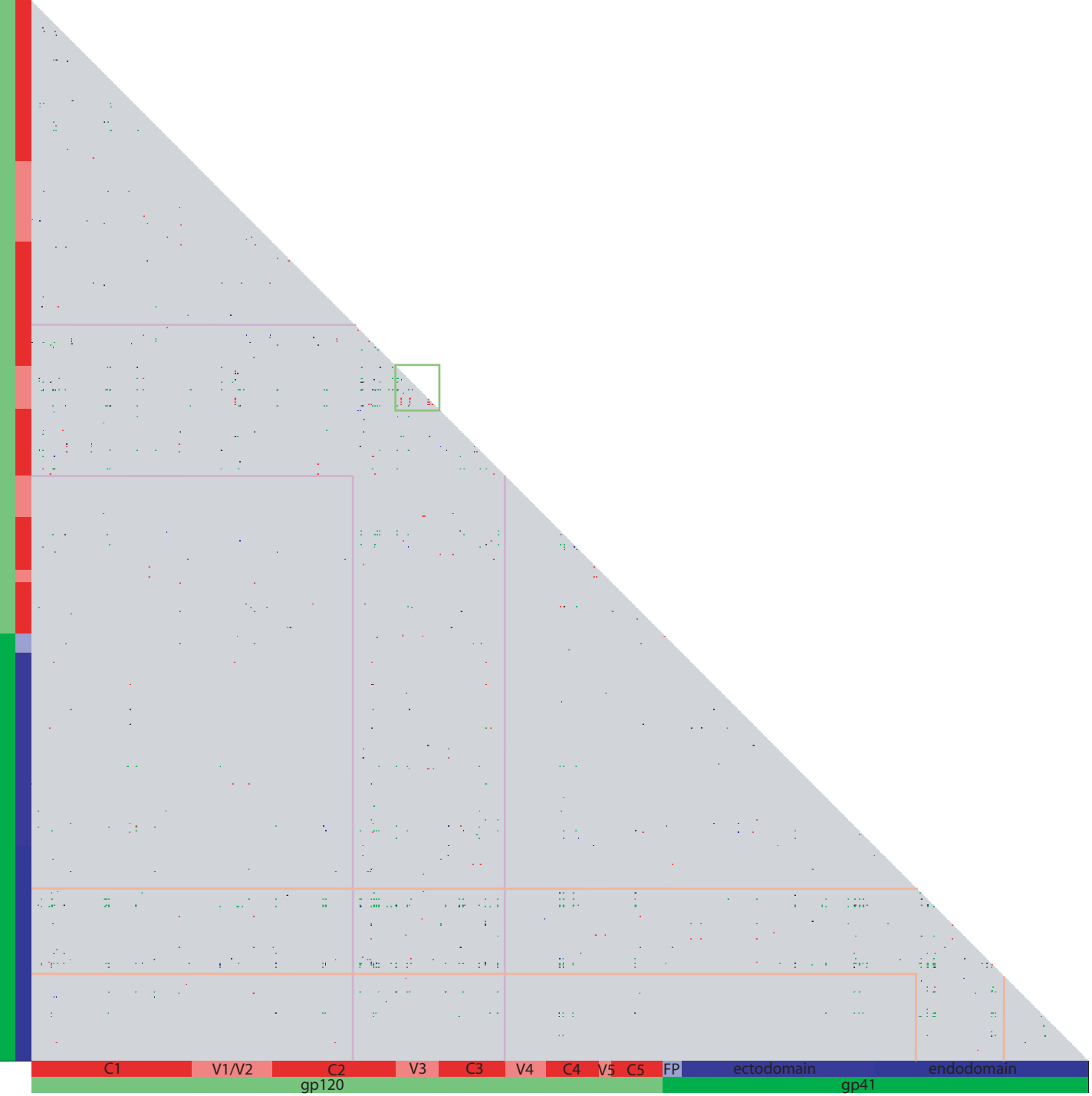
Figure 1: Coevolving pairs of amino acid residues over the complete *env* gene. Pairs whose coevolution is correlated by hydrophobicity, molecular weight or both hydrophobicity and molecular weight are marked. Also shown between coloured lines are the two regions that appear to exhibit higher levels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

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

Figure 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 8Å of CD4 functional residues. A number of these proximal residues correspond to known glycosylation residues and these are also marked.

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

Figure 5: Inter-protein coevolution between gp41 and gp120. (A) Coevolving residues within the gp41 ectodomain and (B) the residues within the gp120 core with which they coevolve.

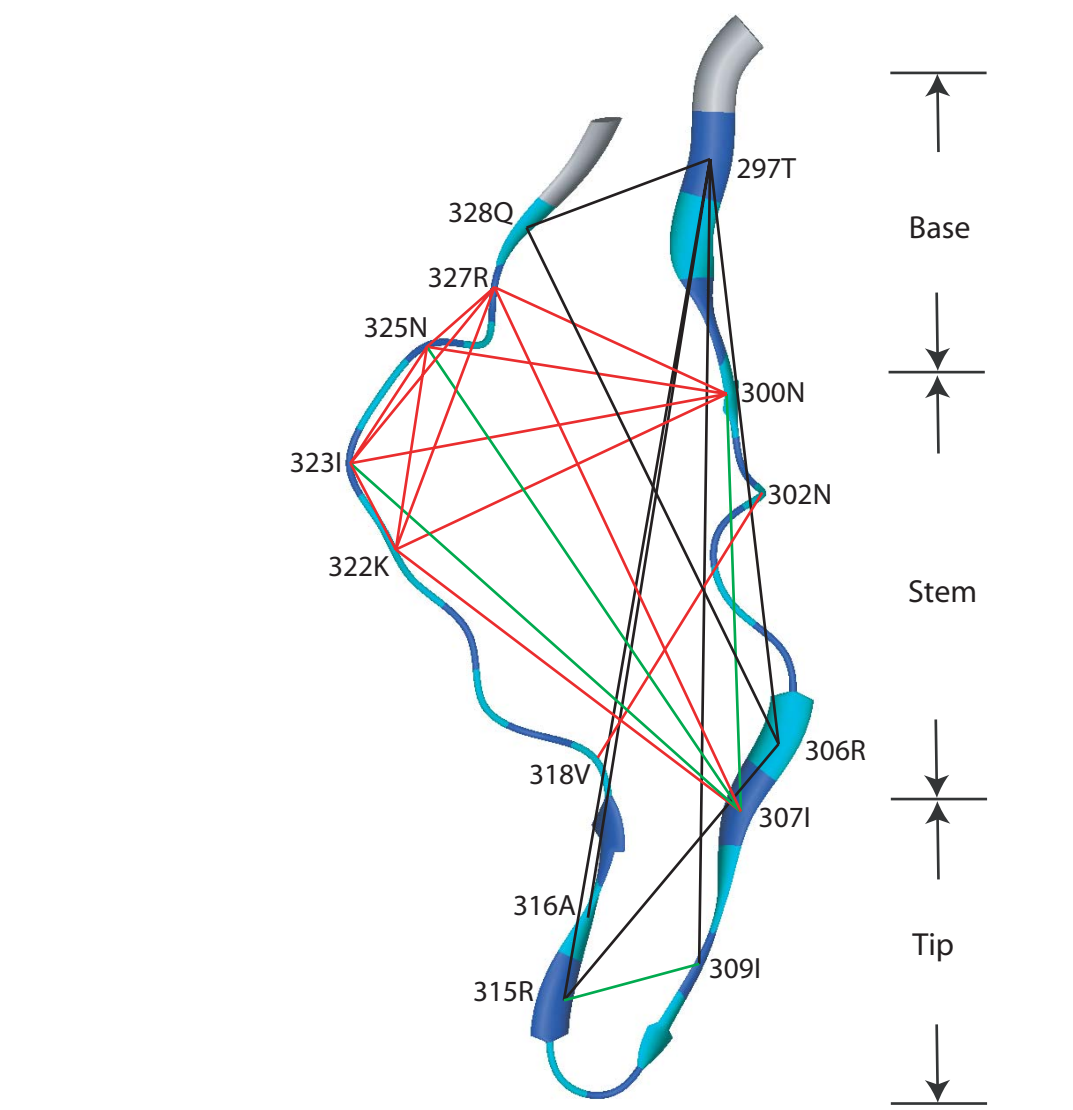


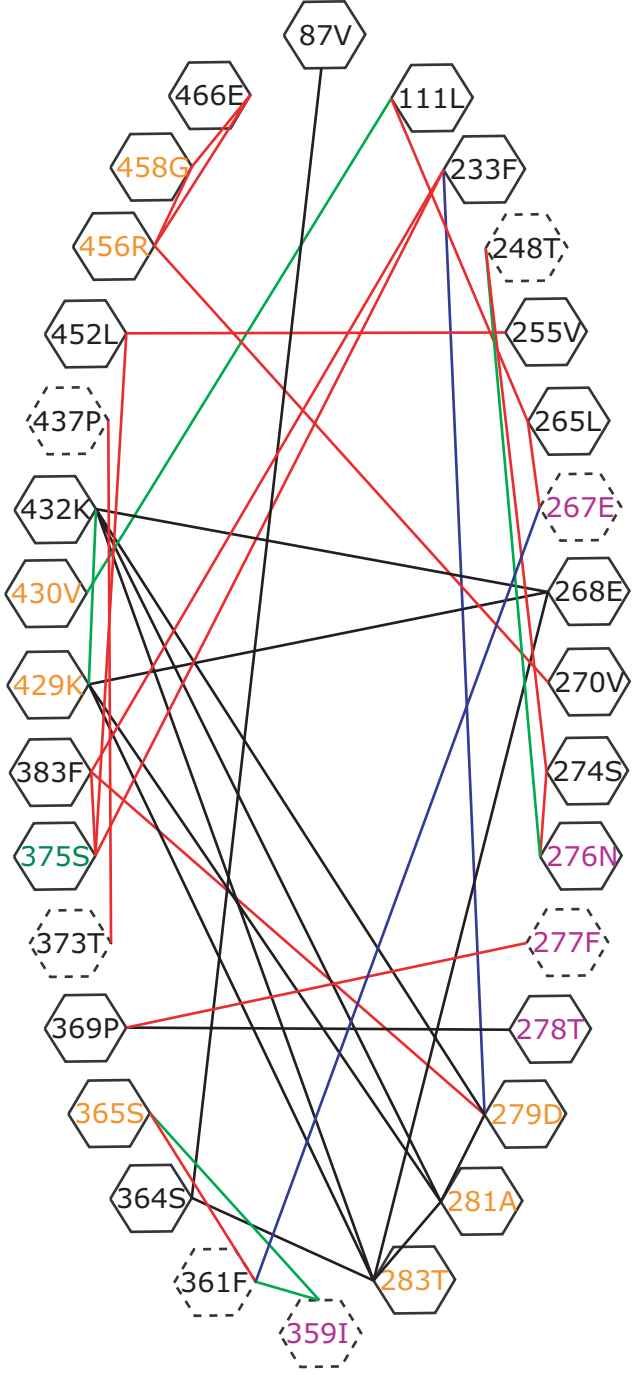
Coevolving Residues

Coevolving Residues Significant by Hydrophobicity

Coevolving Residues Significant by Molecular Weight

Coevolving Residues Significant by both Hydrophobicity and Molecular Weight


297T
328Q
327R
325N
300N
323I
302N
322K
306R
318V
307I
316A
309I
315R
Base
Stem
Tip
Coevolving Pair
Coevolution Significant by Hydrophobicity.
Coevolution Significant by both Hydrophobicity and Molecular Weight



Coevolving Pair

Coevolution Significant by Hydrophobicity.

Coevolution Significant by Molecular Weight

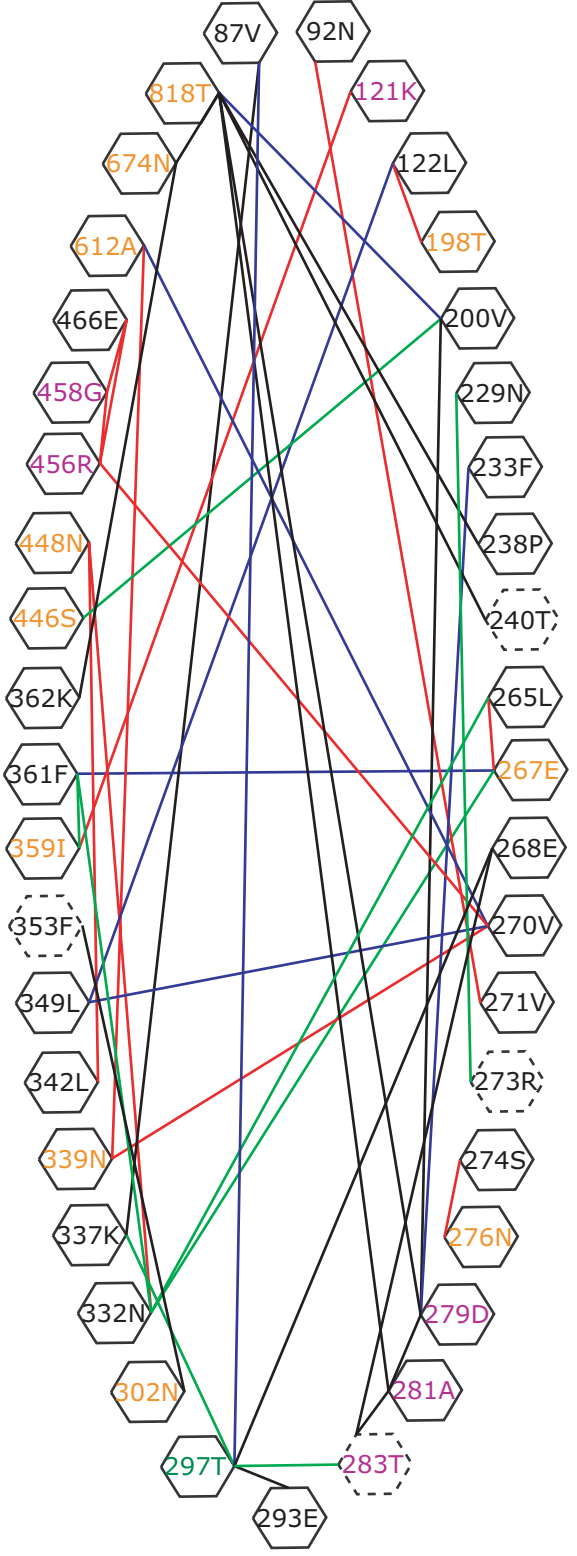
Coevolution Significant by both Hydrophobicity and Molecular Weight

Residues that bind directly to CD4

Residues that correspond to the BMS-806 epitope. Blocks gp120-CD4 interaction

Other functional residues; all correspond to residues that are directly glycosylated in HIV (276N, 277F, 278T) or SIV (267E, 359I)

All other residues are within 8Å of residues that bind directly to CD4, that correspond to the BMS-807 epitope or residues the mutation of which interferes with gp120-CD4 binding.
Proximal residues in dashed pentagons were only observed as proximal in the SIV unliganded gp120 structure. All other residues are proximal in both the HIV liganded and SIV unliganded gp120 structures



Coevolving Pair

Coevolution Significant by Hydrophobicity.

Coevolution Significant by Molecular Weight

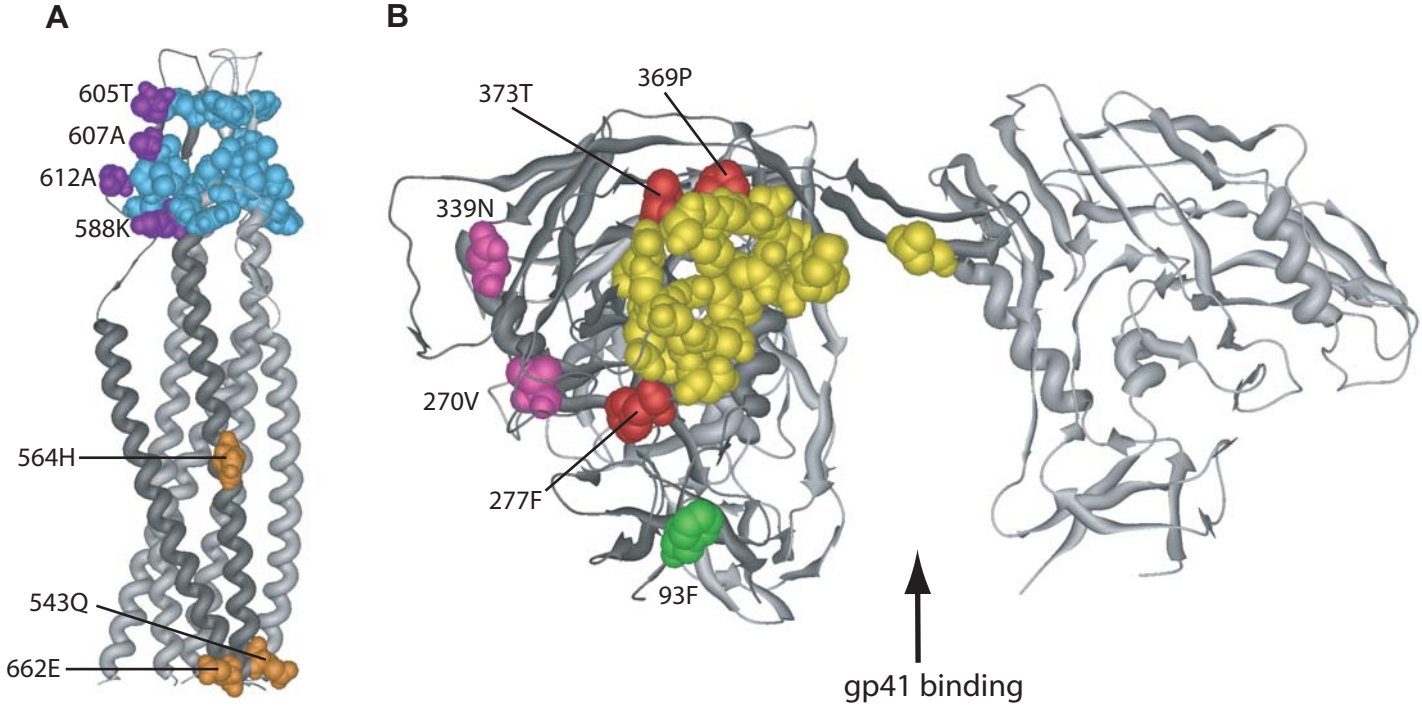
Coevolution Significant by both Hydrophobicity and Molecular Weight

Direct glycosylation residues

Residues that comprise the 2G12 epitope.

Other functional residues; epitope for Ab that interferes with chemokine receptor binding (121K) and residues that directly bind CD4 (279D, 281A, 283T, 456R and 458G)

All other residues are within 8Å of residues that are directly glycosylated or correspond to the 2G12 antibody epitope.
Proximal residues in dashed pentagons were only observed as proximal in either the HIV liganded structure (240T, 353F) or in the SIV unliganded gp120 structure (273R, 283T). All other residues are proximal in both the HIV liganded and SIV unliganded gp120 structures



- Hydrophobic patch
- Coevolving residues (607A comprises part of the patch while 588K, 605T and 612A are localised within it)
- Residues elsewhere in gp41 that are coevolving with residues in the proposed hydrophobic gp120 binding domain.
- CD4 Binding Residues
- Residue (93F) proposed to be in gp41 binding domain (Wyatt et al, 1997-45)
- Residues Coevolving with 588K.
- Residues Coevolving with 605T and 612A.

NB With the exception of the hydrophobic patch all marked residues are only shown on one molecule. In each trimer the labeled molecules are coloured darker.