Rapid Evolution by Positive Darwinian Selection in T-Cell Antigen CD4 in Primates

Zhengdong D. Zhang · George Weinstock · Mark Gerstein

Received: 9 September 2007 / Accepted: 28 February 2008 / Published online: 15 April 2008

Abstract CD4, an integral membrane glycoprotein, plays a critical role in the immune response and in the life cycle of simian and human immunodeficiency virus (SIV and HIV). Pairwise comparisons of orthologous human and mouse genes show that *CD4* is evolving much faster than the majority of mammalian genes. The acceleration is too great to be attributed to a simple relaxation of the action of purifying selection alone. Here we show that the selective pressure acting on CD4 is highly variable between regions in the protein and identify codon sites under strong positive selection. We reconstruct the coding sequences for ancestral primate CD4s and model tertiary structures of all ancestral and extant sequences. Structural mapping of positively selected sites shows they distribute on the surface of the D1 domain of CD4, where the exogenous SIV gp120 protein binds. Moreover, structural models of the ancestral sequences show substantially larger variation in the interfacial electrostatic charge on CD4 and in the surface complementary between CD4 and gp120 in CD4 lineages from primates with natural SIV infections than those without. Thus, positive selection on CD4 among primates may reflect forces driven by SIV infection and could provide a link between changes in sequence and structure of CD4 during evolution and the interaction with the immunodeficiency virus.

Keywords CD4 · gp120 · Human immunodeficiency virus · Simian immunodeficiency virus · Lentivirus · Molecular adaptation · Positive selection · Protein evolution

Introduction

CD4 is an integral membrane glycoprotein of 55 kDa. The mature human protein, encoded by the *CD4* gene, composed of 9 introns and 10 exons on chromosome 12 (Hanna et al. 1994), consists of 433 amino acids, which, based on their cellular locations, can be divided into the extracellular, the transmembrane, and the cytoplasmic segments, of 371, 24, and 38 amino acids, respectively. Characterized as a recombinant soluble protein (sCD4), the extracellular region contains four domains (D1-D4), all of which show sequence and structure similar to those of immunoglobulin superfamily domains. The crystal structures of the human D1D2 fragment (Ryu et al. 1990; Wang et al. 1990), the rat D3D4 fragment (Brady et al. 1993), and, later, the intact human sCD4 (Wu et al. 1997) have been determined. These structural studies suggest that the four extracellular domains of CD4 have a linear arrangement, with D1 at the NH_2-terminus farthest from the cell surface.

CD4 is involved in thymocyte development and plays a central role in T-cell activation by binding to the non-polymorphic regions of MHC-II as a coreceptor for the T-cell antigen receptor (TCR) to increase the affinity

Electronic supplementary material The online version of this article (doi:10.1007/s00239-008-9097-1) contains supplementary material, which is available to authorized users.

Z. D. Zhang · M. Gerstein (✉)
Department of Molecular Biophysics and Biochemistry,
Yale University, New Haven, CT 06520, USA
e-mail: mark.gerstein@yale.edu

G. Weinstock
Human Genome Sequencing Center, Baylor College of
Medicine, One Baylor Plaza, Houston, TX 77030, USA

M. Gerstein
Interdepartmental Program in Computational Biology and
Bioinformatics and Department of Computer Science,
Yale University, New Haven, CT 06520, USA

between thymocytes and antigen-presenting cells (Marrack et al. 1983; Reinherz and Schlossman 1980). Moreover, CD4 is the primary cell surface receptor for SIV/HIV (Dalgleish et al. 1984). Binding of the gp120 component of the viral envelope protein to CD4—by structural mimicry of the binding interface of MHC-II and with a binding affinity orders of magnitude stronger than that of the CD4/MHC-II complex (Kwong et al. 1998)—initiates the cellular entry of the primate immunodeficiency viruses.

The nonsynonymous/synonymous substitution rate ratio ($\omega = d_N/d_S$) of *CD4*, which measures the selective pressure exerted at the protein level and thus indicates its evolutionary characteristic, is 0.77 by pairwise comparison of the human and mouse CD4 coding sequences (Ansari-Lari et al. 1998). The median ω ratio of protein-coding genes is 0.12 using 12,615 pairs of human and mouse orthologues (Waterston et al. 2002). Although the ω ratio of *CD4* is <1, its elevation from the median value is evident and too significant to be attributed to a decrease in the intensity of the purifying selective pressure (Hughes 1997), which can be sufficiently discerned from the usually limited positive selection by the novel phylogenetic analysis method using maximum likelihood (PAML [Yang 1997]). To gain new insights into the evolution of CD4 and the evolutionary effect of the host-virus interaction in the form of the binding between CD4 and the surface protein gp120 of the simian and human immunodeficiency virus (SIV and HIV), the evolutionary characteristic of CD4 in the primate lineage was investigated. In this study we characterized the molecular adaptation of CD4 and identified the amino acid sites at which this protein has been positively selected during evolution.

Methods

Sequence Data

All CD4 mRNA sequences used in this study came from the GenBank database. After the removal of the redundant and fragmental sequences, 11 primate CD4 mRNA sequences were analyzed together in this paper: AF452616.1 (*Callithrix jacchus*, white-tufted-ear marmoset), X73327.1 (*Cercocebus atys*, sooty mangabey), AF001221.1 (*Chlorocebus tantalus*, tantalus monkey), X73324.1 (*Erythrocebus patas*, patas monkey), BC025782.1 (*Homo sapiens*, human), D63349.1 (*Macaca fascicularis*, crab-eating macaque), D63348.1 (*Macaca fuscata*, Japanese macaque), D63347.1 (*Macaca mulatta*, rhesus monkey), D63346.1 (*Macaca nemestrina*, pigtailed macaque), M31135.1 (*Pan troglodytes*, chimpanzee), and AF452617.1 (*Saimiri sciureus*, common squirrel monkey). Mouse CD4 mRNA sequence (BC039137.1) was used as the outgroup sequence.

Modeling Variable Selective Pressure Among Codon Sites

The alignments of the CD4 mRNA sequences were created based on the alignment of their corresponding protein sequences that were generated using CLUSTALX (Thompson et al. 1997) and improved manually. The mRNA alignment, after complete deletion of the gaps and missing data, contains 397 codons in each sequence. The phylogeny of CD4 was constructed by Bayesian Markov chain Monte Carlo analysis using MRBAYES 3.0B4 (Ronquist and Huelsenbeck 2003) with the general time-reversible model of the nucleotide substitution, a four-category gamma distribution of rates across sites, four heated chains with temp = 0.20, 80,000 iterations, and an initial 10,000 “burn-in” trees to be discarded prior to the steady state. The summarization of all the poststationarity trees generated a Bayesian CD4 consensus tree, and its primate subtree is in accordance with the composite estimate of the phylogeny of 202 primates (Purvis 1995).

Site-specific d_N/d_S analysis was carried out with CODEML, an application from the PAML software package that can apply different evolution models to sequences. CODEML takes the Bayesian tree and the multiple sequence alignment as the input, and uses maximum likelihood to predict sites in a group of coding sequences that have been subject to positive selection. Null models where sites are predicted to have a d_N/d_S ratio (ω) between 0 and 1 can be compare with more general alternative models that also allow for $\omega > 1$. For each model a log likelihood value (ℓ) is calculated by maximum likelihood, which enables a likelihood ratio test (LRT) to determine whether ω is significantly different from 1 for the pairwise comparison between a null model H_0 and an alternative nested model H_1. If the alternative model indicates that an estimated $\omega > 1$ and the test statistic [$2\Delta\ell = 2(\ell_1 - \ell_0)$] is greater than critical values of the chi-square distribution with the appropriate degree of freedom (df), then positive selection can be inferred. In this circumstance, the Bayesian theorem is used to predict which sites (codons) from the original data are most likely to have been under diversifying selection.

Three pairs of null and alternative models were compared: M0 (one-ratio) with M3 (discrete), M1a (near neutral) with M2a (positive selection), and M7 (β) with M8 (β&ω). M0 assumes the same ω for all sites in the protein, whereas M3 attempts to fit the data into three site classes with the proportions and ω values estimated from the data. M1a assumes two site classes, with one ω value fixed at 1 and the other estimated from data under constraint to be between 0 and 1, whereas M2a adds a third site class with ω as a free parameter estimated from the data, thus allowing for sites with $\omega > 1$. M7 attempts to fit the data to a beta distribution $\beta(p, q)$ of ω values between 0 and 1 with 10 site classes, whereas M8 includes an additional site class whose proportion and ω ratio are estimated from the data (Supplementary Table 2). Since the M0/M3 LRT is compared with χ^2 with 4 df, whereas both the M1a/M2a and the M7/M8 LRTs have 2 df, the latter two pairs are more statistically powerful. Because the local minima can trap the iterative estimation of ω values under both M2a and M8, and thus thwart the finding of the best numerical solutions, M2a and M8 CODEML runs were consequently started with different initial ω values (0.03, 0.3, and 1.3), and only the results with the greatest log likelihood values were presented here.

Simulations for Performance Assessment

We used sequence simulation to assess the identifiability of the sites under positive selection in the CD4 sequences being analyzed. One hundred replicates of sequence sets were simulated with the maximum likelihood estimates of parameters (including the unrooted phylogenetic tree) calculated with the actual CD4 sequence data and included a small fraction of sites evolving under positive selection as estimated under the model M2a (Table 2). The sequence simulation was carried out using the program evolver in the PAML package. An ω ratio and an ancestral codon state at the root of the tree are randomly drawn according to the given multinomial distributions. Then the program "evolves" each site along the branches of the tree independently, according to the Markov process of codon substitution. Sites evolving by positive selection (with $\omega > 1$) are listed in a file and are later compared with the predicted sites.

Modeling Variable Selective Pressure Among Partitions

Partition-specific d_N/d_S analysis was carried out with six models, A–F, which accommodate different levels of site heterogeneity (Yang and Swanson 2002). They are specified by the option variable *G* in the sequence data file and the variable *Mgene* in the control file in the PAML program package. Model A, the simplest model, assumes that all sites in the sequence have the same substitution pattern with identical parameters, while model F, the most complex model, assumes that all site partitions have different substitution patterns with independent substitution parameters. Lying in between these two extremes, models B–E all assume different substitution rates among the site partitions. Apart from the different substitution rates, model B assumes homogeneity among partitions in the transition/transversion rate ratio κ, the nonsynonymous/synonymous rate ratio ω, and the codon frequencies. Model C assumes identical κ and ω but different codon frequencies among partitions. Model D assumes different κ and ω but identical codon frequencies among partitions. Model E assumes different κ and ω and different codon frequencies among partitions (Supplementary Table 2).

Sequence Reconstruction and Structural Analysis of the Ancestral CD4

CODEML was used to reconstruct the CD4 sequences for extinct ancestral nodes in the phylogeny (Yang et al. 1995): with the marginal reconstruction algorithm, the F61 (free-parameters) model of codon substitution, a gamma distribution with four categories of rates across sites, and the Bayesian tree topology inferred using the method described above, it calculated maximum likelihood estimates of branch lengths, codon frequencies, and site-specific codon reconstructions. For the nine ancestral sequences for nodes from 12 to 20, the reconstruction accuracies, each of which is averaged over all sites in each sequence, are 0.97919, 0.98800, 0.99838, 0.99918, 0.99999, 0.99998, 0.99998, 0.99940, and 0.99977.

Inferred nucleotide sequences were translated into protein sequences using the standard genetic code. Accurate prediction of the three-dimensional (3D) structures of the ancestral CD4 is possible because of the significant similarity among the extant and the reconstructed primate CD4 sequences. As a comparative modeling technique based on the principle that proteins with homologous sequences have usually similar structures (Browne et al. 1969; Greer 1990), the homology modeling of the reconstructed ancestral CD4 proteins was carried out with DEEPVIEW 3.7. The Protein Data Bank was searched for the structural modeling template, which is 1RZJ for all the CD4s to be modeled, and into which the CD4 sequences would be threaded. The basal structural models were then optimized by the Swiss-Model server (Schwede et al. 2003) and were in turn subjected to 50 iterations of the steepest descent energy minimization with the 43B1 parameter set of the GROMOS96 force field (Scott et al. 1999). The program DelPhi (Gilson and Honig 1988) was used to calculate the electrostatic charges at interfacial atoms of CD4 that are less than 8 Å away from HIV gp120, and then the absolute charges were summed up to approximate the total electrostatic charges on the interface of CD4. The surface complementarity (*Sc*) between various CD4s and HIV gp120 was analyzed by the program SC in the CCP4 software package (Collaborative Computational Project 1994).

Results

Modeling Variable Selective Pressures Among Sites Indicates the Adaptive Evolution of Primate CD4 and Identifies Sites Evolving Under Positive Selection

The pairwise approach for the ω ratio estimation has a low sensitivity to detect positive selection, as it averages selective pressure over the entire evolutionary history separating the two lineages and over all codon sites in the sequences. In most functional genes, however, the majority of amino acid sites will be subject to strong purifying selection, with only a small fraction of the sites potentially targeted by adaptive evolution.

To detect positive selection in CD4, we first tested for variability in the selective pressure among its sites. The phylogenetic analysis using maximum likelihood (PAML) of 11 primate CD4 coding sequences was conducted using the null model that allows only one ω ratio for all sites and the alternative one that allows several different ones. The LRT comparing these two models indicates extreme variation in selective pressure among amino acid sites in CD4 (Table 1). With the selective pressure variation among sites established, next we tested for positive selection among CD4 sites by two model comparisons. In each comparison, the null model does not allow the presence of sites under positive selection, while the alternative does. The test statistics of both comparisons are highly significant (Table 1), which, by rejecting the null models, indicates the presence of sites evolving under positive selection in CD4.

Estimates of ω under models that allow for sites under positive selection (M2a and M8) indicate a fraction of sites evolving under positive selective pressure (Table 2). To identify codon sites in CD4 under positive selection, the posterior probability of every site belonging to a particular site class is calculated using the empirical Bayes approach. The posterior probabilities for the only site class that allows $\omega > 1$ under model M8 are plotted in Fig. 1A. Even though many sites appear to be under positive selection, only a handful of them meet or exceed a stringent criterion. Both model M2a and model M8 identify six sites in CD4—T42, S48, N64, P73, N77, and A80—to be under positive selection with significant probabilities ($p_{\omega>1} > 0.95$). Two

Table 1 Likelihood ratio test statistics ($2\Delta\ell$) for model comparisons

Alternative model vs. Null model			$2\Delta\ell$[a]	df[b]	$\chi^{2}_{1\%}$[c]	p
Test for variability in the selective pressure among CD4 codon sites						
M3 (discrete)	vs.	M0 (one-ratio)	71.82	4	13.28	9.33×10^{-15}
Tests for positive selection among CD4 codon sites						
M2a (positive selection)	vs.	M1a (nearly neutral)	37.18	2	9.21	8.44×10^{-9}
M8 ($\beta\&\omega$)	vs.	M7 (β)	37.08	2	9.21	8.88×10^{-9}

[a] The test statistic of the likelihood ratio test

[b] Degrees of freedom

[c] The critical χ^2 value at the 1% level with the specified degrees of freedom

Table 2 Modeling variable selective pressures among sites

Model	N[a]	ℓ[b]	Estimated parameter(s)[c]		Positively selected sites[d]
M0: one-ratio	1	−3040.84	$\omega = 0.9563$		Not modeled
M3: discrete	5	−3004.93	$\omega_0 = 0.324$, $\omega_1 = 2.920$, $\omega_2 = 9.915$	$p_0 = 0.746$, $p_1 = 0.226$, ($p_2 = 0.028$)	Not used
M1a: nearly neutral	2	−3023.92	$\omega_0 = 0.00$, ($\omega_1 = 1$)	$p_0 = 0.40$, ($p_1 = 0.60$)	Not allowed
M2a: positive selection	4	−3005.33	($\omega_0 = 0.165$), ($\omega_1 = 1.000$), $\omega_2 = 5.489$	$p_0 = 0.473$, $p_1 = 0.403$, ($p_2 = 0.123$)	$\underline{42}$, 48, 64, 73, 77, $\underline{80}$
M7: β	2	−3023.92	$p = 0.008$, $q = 0.005$		Not allowed
M8: $\beta\&\omega$	4	−3005.38	$p = 0.292$, $q = 0.242$, $p_0 = 0.876$, (p_1= 0.124), $\omega = 5.501$		$\underline{42}$, 48, 49, $\underline{64}$, 73, $\underline{77}$, $\underline{80}$, 91

[a] Number of free parameters for the ω distribution

[b] Log-likelihood value

[c] Parameters not in parentheses are free parameters

[d] Codon/amino acid sites with the posterior probability of the positively selected site class > 0.95 (underlined if > 0.99). The numbering is based on the human CD4 sequence

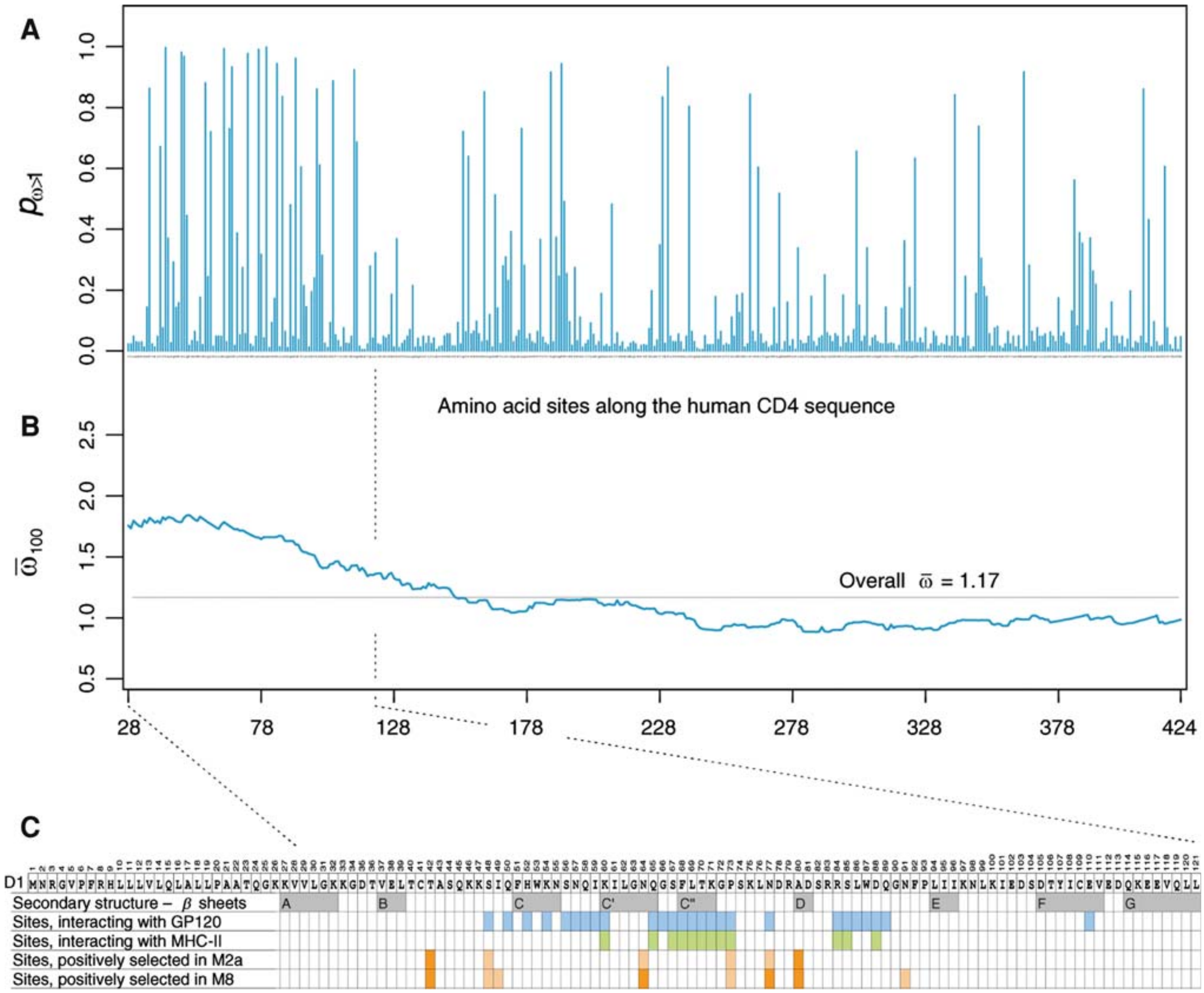

Fig. 1 The identification of sites under positive selection from 11 primate CD4 genes. (**A**) Posterior probabilities of the positively selected site class (with $\omega > 1$) for amino acid sites from V28 to H424 along the CD4 sequence, the majority of the extracellular portion of CD4. (**B**) Posterior mean of ω averaged over a sliding 100-aa window. The light-gray line indicates the overall average ω (= 1.17). For both panels, maximum likelihood estimates under the random-sites model M8 (β&ω) were used. (**C**) Domain 1 (D1) of human CD4 and its relevant properties. Nine strands in D1 form two β-sheets: ABED and GFCC′C″ (Ryu et al. 1990; Wang et al. 1990). The sites in CD4 that interact with HIV gp120 (PDB ID: 1RZJ) and a class II MHC molecule (1JL4), respectively, are highlighted in gray. The sites identified to be under diversifying selection in this domain under models M2a and M8 are also highlighted—ones with significant probability ($p > 0.95$) are in light orange, and ones with highly significant probability ($p > 0.99$) in orange. For each amino acid site, model M8 estimates the ω ratios and the prior and the posterior probabilities of 11 site classes. The first 10 ω ratios, ω_0 to ω_9, were derived from the beta distribution $\beta(p, q)$ and, as a result, limited to the range (0, 1). The ratio ω_{10} of the eleventh site class was freely estimated from the data and thus can be >1. In this study, $\omega_{10} = 5.5$. The posterior probabilities, dramatically altered from the prior probabilities by the codon configuration at a site in gene orthologues, could be quite different from the prior probabilities. The sum of the first 10 posterior probabilities for site G34 is 0.9946, and this site is very likely to be under purifying selection. The posterior probability of the eleventh site class for site T42 is 0.9954, and this site is almost certainly under diversifying selection

of them, T42 and A80, with highly significant probabilities ($p_{\omega>1} > 0.99$) are also identified as sites under positive selection when we apply the sitewise likelihood-ratio (SLR) method (Massingham and Goldman 2005) to this data site.

To evaluate the quality of our specific sequence-based inferences, we assessed the identifiability of the sites under positive selection in the sequence set being analyzed. Even though some rules of thumb on the number of sequences and overall sequence divergence could be used for a quick rough assessment, there is no analytical solution to this problem. However, it can be addressed to some extent by sequence simulation. Previous studies (Anisimova et al. 2001, 2002) have shown that the log-likelihood ratio test is very robust for detecting the existence of positive selection in a set of sequences but the subsequent Bayes prediction of amino acid sites under positive selection is a much harder problem. Thus, we focused the assessment on the Bayes prediction in our study. To do so, we first simulated 100 sets of coding sequences using parameters estimated from the actual CD4 sequences. In each simulation, there is a small fraction of sites evolving under diversifying selection, which are the true positives. We then analyzed each replicate set using the same method that was applied to the actual CD4 sequences and inferred a list of sites evolving under diversifying selection. The comparison between these two kinds of lists—the true positives and the inferred positives—shows that when 0.95 is used as the threshold on the posterior probability of sites under positive selection, as we did above, about 93% of sites classified as sites under positive selection are indeed evolving as so in a sequence set that closely resembles the real one being analyzed (Supplementary Table 1).

Sites Evolving Under Positive Selection in CD4 Are Located in its D1 Domain on the Interface with gp120

To identify the location under diversifying selection in CD4 on a larger scale than individual amino acid sites, the posterior means of ω were averaged over a window of 100 amino acids (aa)—the average size of the four domains, D1–D4, of CD4—sliding along the CD4 sequence (Fig. 1B). The plot in Fig. 1A and the horizontal line indicative of the overall average ω in Fig. 1B are in fact the two extreme cases of averaging over 1-aa and 397-aa windows, respectively. As expected, the sliding widow size increases and the plotted line becomes smoother with the regions—not individual amino acid sites—under diversifying selection accentuated. Figure 1B clearly shows that the first domain (D1), with the maximum ω in this region nearly twice greater than the overall average, is the region where the positive selection has exerted its effect on CD4.

To formally test whether the D1 domain of CD4 is evolving under positive selection, we segmented residues of CD4 into two partitions: those in the D1 domain and those in D2–D4. We modeled variable selective pressures among partitions using models A–F and calculated separate substitution rates and ω ratios for these two partitions from the sequence data (Table 3). First, we test the overall heterogeneity in selective pressures as measured by kappa and omega between these two partitions of CD4. Comparison between model B and model D by the LRT rejects equal transition/transversion rate ratios (κ_s) and nonsynonymous/synonymous substitution rate ratios (ωs) for these two partitions ($2\Delta\ell = 16.02$, $p = 3.32 \times 10^{-4}$, df = 2). The parameter estimation indicates that the substitution rate in D1 is twice ($r_2 = 2$) as high as in the rest of CD4 and the sites in D2–D4 are under purifying selection with $\omega_1 = 0.66$, whereas the sites in D1 are under diversifying selection with $\omega_2 = 2.30$. Then we test whether the ω ratio in the D1 domain is significantly different from 1 and recalculate the log-likelihood value in models D, E, and F with fixed $\omega_2 = 1$. Under model D, the log-likelihood is −3023.89 when ω_2 is a free parameter and −3019.09 when ω_2 is = 1 fixed. As the log LRT statistic is $2\Delta\ell = 9.6$, with $p = 1.9 \times 10^{-3}$ at df = 1, the test rejects the null hypothesis and suggests that the ω_2 ratio in D1 is significantly >1. Similar comparison under model E or F also indicates positive selection in the D1 domain.

The respective binding sites on CD4 for gp120 (Kwong et al. 1998) and MHC-II (Kwong et al. 1998; Wang et al.

Table 3 Modeling variable selective pressures among partitions

Model	Type	N[a]	ℓ[b]	r_2[c]	κ[d]		ω[e]	
A	Homogeneous model	30	−3040.84	1	4.036		0.956	
B	Different r[f]	31	−3027.10	2.01	4.013		0.971	
C	Different r, π[g]	40	−2998.62	2.03	3.998		0.978	
					κ_1	κ_2	ω_1	ω_2
D	Different r, κ, ω	33	−3019.09	2.03	3.992	4.323	0.658	2.299
E	Different r, κ, ω, π	42	−2989.03	2.03	4.085	4.292	0.642	2.533
F	Separate analysis	60	−2969.74	1.98	4.101	4.244	0.643	2.512

[a] The number of parameters, including $b = 19$ branch lengths and $g = 2$ partitions

[b] The log-likelihood value

[c] The substitution rate of the second site partition (D1) relative to the rate of the first partition (D2–D4)

[d] The transition/transversion rate ratio

[e] The nonsynonymous/synonymous substitution rate ratio

[f] Two substitution rates (r_1 and r_2) of the two site partitions. r_2 is a free parameter in models B–F, estimated from the data, while r_1 is fixed at 1

[g] The codon frequencies at equilibrium (πs), calculated using the empirical nucleotide frequencies observed at the three codon positions, with nine parameters used

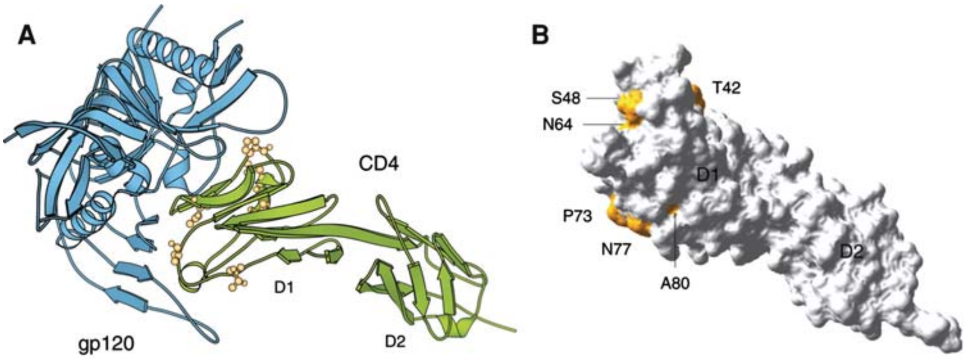

Fig. 2 Sites under positive selection in the tertiary structure of human CD4. (**A**) Ribbon diagram of the human CD4 and HIV gp120 complex. The antigen-binding fragment of the antibody 17b in the original crystal structure (PDB ID: 1RZJ, an update of 1GC1 from Kwong et al. [1998]) is omitted from this presentation. Sites under diversifying selection identified under both M2a and M8 (T42, S48, N64, P73, N77, and A80) are highlighted with their side chains shown. This figure was prepared with MOLSCRIPT (Kraulis 1991). (**B**) Molecular surface of CD4. The same six sites under diversifying selection, accented in black, are distributed on the molecular surface at the top of domain D1 of CD4, which forms part of the interface between CD4 and gp120 and between CD4 and MHC-II

2001) are both fully contained in D1, and the former is a superset of the latter. These sites are mapped along with the positively selected sites under different models to the peptide sequence of D1 of the human CD4 in Fig. 1C. Juxtaposition of the binding sites and the sites predicted to be under positive selection for a comparison of their distribution in CD4 manifests a substantial correspondence between the two types of sites. Sites inferred to be under diversifying selection under both M2a and M8 are mapped onto the α-carbon ribbon diagram of human CD4 bound to HIV gp120 (Kwong et al. 1998) in Fig. 2. Scattered from strand B to strand D in D1 of CD4 in the primary sequence, the six sites under positive selection are distributed on the molecular surface at the top of the domain D1 of CD4, which forms part of the interface between CD4 and gp120 or between CD4 and MHC-II.

Structural Modeling of Ancestral Primate CD4s Shows Significant Structural Variation on Their Interface with gp120

To examine how sequence changes in CD4 during primate evolution might affect its chemical properties and its interaction with gp120, we reconstructed the codon and amino acid sequences of ancestral primate CD4s. In addition, we modeled the three-dimensional (3D) structures (limited to the D1 and D2 domains) of the ancestral and nonhuman primate CD4s. The actual (human) and the modeled CD4 tertiary structures are overall highly similar but have subtle yet consequential differences in their interface with gp120. For example, the molecular surface area of the D1 and D2 domains ranges from 7647 to 8159 Å^2, with an average of 7827 Å^2, compared with the 7883-Å^2 surface area of D1D2 of the human CD4.

To evaluate possible effects of natural SIV infections on the CD4 structure, we compared structural changes in CD4s over four terminal lineages from primates with natural SIV infections to their immediate ancestors with those over the remaining eight lineages descended from the latest ancestor nodes 15 and 20 (Fig. 3). For this evaluation, we considered three structural properties: the electrostatic potential on the interface of CD4 to gp120, the shape complementarity of the interface between CD4 and gp120, and the accessible surface area occluded by the binding of CD4 and gp120. For each structural property, we first calculated the fold increase of the average structural change in the first group of lineages to the average in the second group and then used the one-side Wilcoxon rank sum test to assess the statistical significance of the observed increase. Due to the small number of data points available for such a comparison, however, the test result should be viewed with caution.

The surface electrostatic potential measures the distribution of electrostatic charges on the surface of a molecule and plays an important role in specific protein-protein recognition. With the development of the continuum electrostatics model for proteins, it has become possible to calculate and display the electrostatic potential of protein structures by numerically solving the Poisson-Boltzmann equation for the protein-solvent system (as implemented in the program DelPhi) (Gilson and Honig 1988). Even though all of the CD4 sequences show a generally similar pattern of the surface electrostatic potential, there are differences in the potentials around the interface between the ancestral form and its immediate descendant CD4 along lineages with natural SIV infections (Fig. 3). There is a more than threefold increase in the average absolute difference of the total interfacial electrostatic charge on CD4 over lineages involving natural SIV hosts (the Wilcoxon rank sum test statistic $W = 29$, $p = 0.01371$; Supplementary Fig. 1A). (One notable exception, for example, is the change of electrostatic charge from node 18 to node 5, the pigtailed macaque, which is not naturally SIV-infected.)

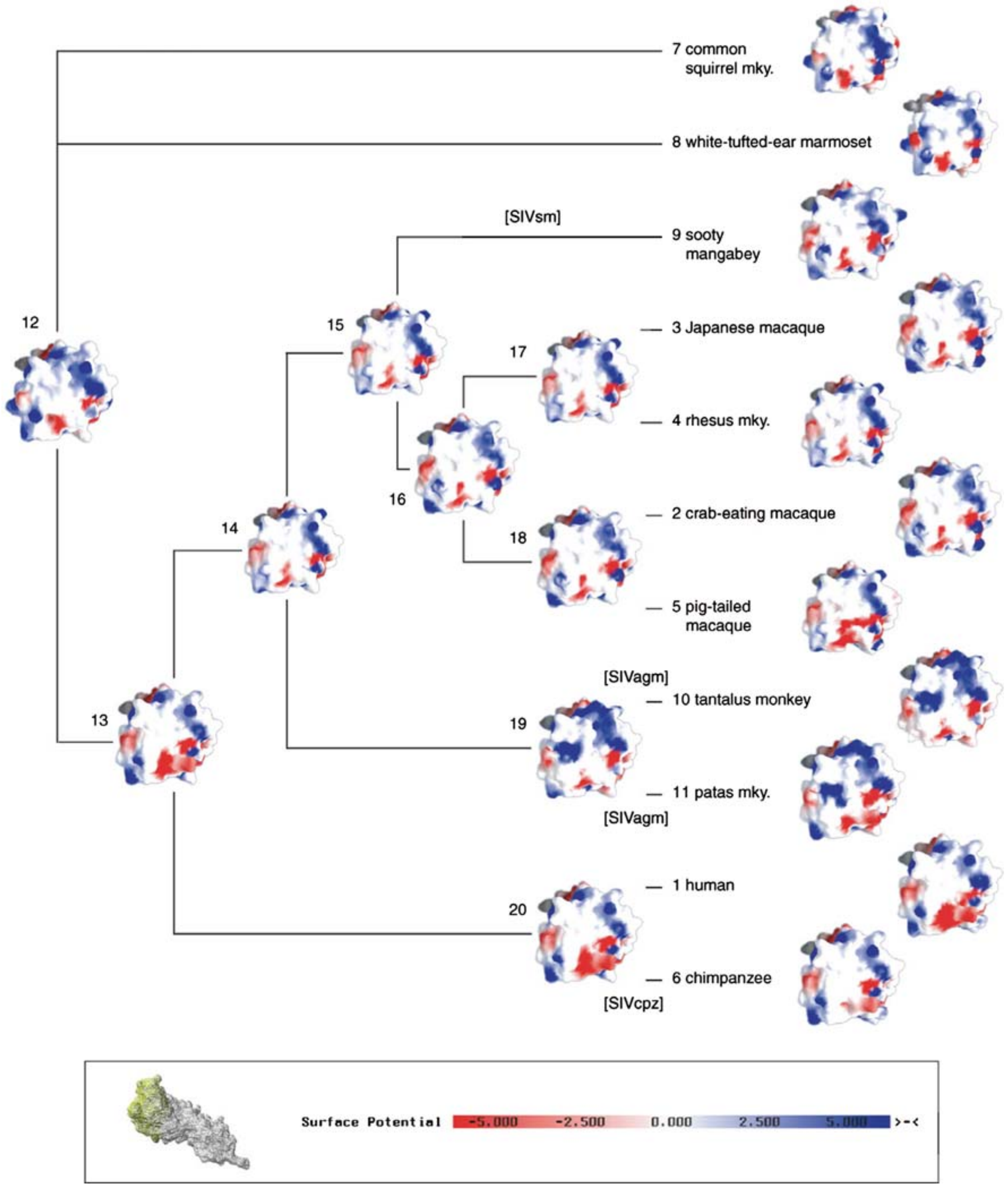

Fig. 3 Electrostatic surface of primate CD4s. The surface of D1 domain where gp120 binds is facing directly toward the viewer (see the green area on the surface of a CD4 molecule in the legend box for orientation). The electrostatic potential is shown at the molecular surface, which is colored according to the local electrostatic potential. This unrooted tree topology was used in the reconstruction of the ancestral primate CD4 sequences. GRASP 1.3.6 (Nicholls et al. 1991) was used to calculate the electrostatic potential and generate the isopotential surface maps. The most striking example of changes in surface electrostatic potential is the lineage leading from the root (the node 12) to the tantalus monkey and the patas monkey. In this case, along the branch from the root to node 13 the negative electrostatic potential intensifies surrounding the contacting surface on CD4, while the positive potential decreases; along the next branch from node 13 to its direct descendant at node 14, the opposite is observed, mostly the diminishment of the negative potential; finally, at node 19, the negative charge on the surface of CD4 not only increases in the peripheral areas but also appears *de novo* in this otherwise neutral interface. The SIV strains from Sooty mangabeys (SIVsm), African green monkeys (SIVagm), and chimpanzees (SIVcpz) are indicated along the last tree branches leading to their natural hosts

Nucleotide substitutions at four codon sites, 42, 48, 64, and 77, all of which are inferred to be under positive selection by both model M2a and model M8 (Table 2), contribute to the variations of the electrostatic potential in or near the interface of CD4. Resulted from such evolutionary sequence changes, these variations could potently influence its interaction with gp120, although the precise effect is difficult to quantify.

The *Sc* value, a statistic used to measure the shape complementarity of the interfaces between the complexed HIV gp120 and various CD4s (Lawrence and Colman 1993), ranges from 0.684 to 0.724, with 0.714 as the average, compared with 0.716 measured between gp120 and human CD4. It shows patterned variations among the interactions between gp120 and different CD4s. There is also a threefold increase in the average absolute difference of *Sc* over lineages involving natural SIV hosts (the Wilcoxon rank sum test statistic $W = 28$, $p = 0.01949$; Supplementary Fig. 1B). A similar pattern is also present in the variation of the accessible surface area occluded by the binding of gp120 and CD4.

Clearly, some of the sequence changes in the primate CD4 during evolution could cause substantial alterations in the electrostatic potential on or near its interface contacting gp120 and the shape complementarity of CD4/gp120 interfaces. When the structural changes in CD4 and the lineages leading to naturally infected hosts during primate evolution are considered together, the apparent correlation between these two suggests that the structural changes were a result of molecular adaptation driven by the viral infection. These radical changes could affect the binding between the viral envelope protein and its primary cell surface receptor.

Discussion

Previous similar studies based on d_N/d_S calculation have reported positive selection in a variety of genes related to immune response (Endo et al. 1996; Hughes 1997; Sawyer et al. 2004), viral genomes (Bush et al. 1999; Fitch et al. 1997; Nielsen and Yang 1998), and sexual reproduction (Swanson et al. 2001, 2003). Our study shows that in primates CD4, a T-cell antigen, has been subjected to adaptive evolution and identify the codon sites under positive selection. Such sites are distributed on the surface of the D1 domain where both the exogenous SIV gp120 and the endogenous MHC-complex II proteins bind. Since the interface is bound by both molecules, it seems undetermined which interaction causes the adaptive evolution of CD4. However, careful examination of the different circumstances in which binding of CD4 to MHC-II or gp120 occurs sheds light on the cause of the positive selection observed. In a normal circumstance, CD4 engages in T-cell activation by binding to the nonpolymorphic regions of MHC-II as a coreceptor for the T-cell antigen receptor. In the course of HIV infection, however, it is also the primary cell surface receptor for the virus, and its binding of the gp120 initiates the cellular entry of the virus. Thus, CD4 enables the virus to infect and, by causing a fatal disease, exert selective pressure on an infected primate population. Advantageous mutations, which increase the fitness of their carriers, are fixed by natural selection with positive selective coefficients.

The interplay between CD4 and gp120 or between CD4 and MHC-II suggests that the advantageous mutations in CD4 are those that either interfere with the binding of CD4 to gp120 or facilitate the interaction between CD4 and MHC-II. Multiple sequence alignment of various primate and murine MHC-II molecules shows that the region where CD4 binds is, unlike the polymorphic antigen-presenting region of MHC molecules, conserved (Supplementary Fig. 2). As a result, the D1 domain of CD4, which interacts with this conserved region of MHC-II, should be subjected to selective constraints. A molecular clock-based method has estimated that the primate lentivirus divergence is much more recent than that of its hosts (Sharp et al. 2000). However, the host-pathogen adaptation between SIV and its natural host species cannot be precluded, because despite a high seroprevalence of SIV in wild sooty mangabeys and African green monkey populations, there is no evidence that infection is associated with immunodeficiency (Heeney et al. 1993; Jolly et al. 1996; Rey-Cuille et al. 1998; Silvestri et al. 2003). Thus, we hypothesize that it is the binding of CD4 to gp120 and its potential pernicious consequences that have elicited the molecular adaptation in its D1 domain since the lentivirus infection started in the primate lineages.

Calculated from the pairwise comparison of the human and mouse gene sequences, the d_N/d_S ratio of *CD4* has long been recognized to be curiously high. While the pairwise sequence comparison of CD4 genes is only suggestive of positive selection due to the innate methodological limitations, our phylogenetic analysis of the same dataset using maximum likelihood not only detected the molecular adaptation of CD4 but also identified the amino acid sites at and lineages along which this protein has been positively selected during evolution. As the tertiary structure of CD4 shows, these amino acid sites distribute on the surface of its D1 domain where class II MHC molecules and gp120 bind. The biological function of gp120 suggests that the driving force behind the positive selection of CD4 is the immunodeficiency viral infection, which exerts selective pressure on an infected primate population. This hypothesis is supported by the effects that the evolutionary sequence changes of primate CD4 had on its tertiary structure and its interaction with gp120.

Acknowledgments This work was supported by the NIH/NLM grant T15 LM07056. Additional funding was provided by a grant from NIH/NLM (1K99LM009770-01) to Z.D.Z. and grants from NIH/NHGRI to G.W. and M.G.

References

Anisimova M, Bielawski JP, Yang Z (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol Biol Evol 18:1585–1592

Anisimova M, Bielawski JP, Yang Z (2002) Accuracy and power of bayes prediction of amino acid sites under positive selection. Mol Biol Evol 19:950–958

Ansari-Lari MA, Oeltjen JC, Schwartz S, Zhang Z, Muzny DM, Lu J, Gorrell JH, Chinault AC, Belmont JW, Miller W, Gibbs RA (1998) Comparative sequence analysis of a gene-rich cluster at human chromosome 12p13 and its syntenic region in mouse chromosome 6. Genome Res 8:29–40

Brady RL, Dodson EJ, Dodson GG, Lange G, Davis SJ, Williams AF, Barclay AN (1993) Crystal structure of domains 3 and 4 of rat CD4: relation to the NH2-terminal domains. Science 260:979–983

Browne WJ, North AC, Phillips DC, Brew K, Vanaman TC, Hill RL (1969) A possible three-dimensional structure of bovine alpha-lactalbumin based on that of hen's egg-white lysozyme. J Mol Biol 42:65–86

Bush RM, Fitch WM, Bender CA, Cox NJ (1999) Positive selection on the H3 hemagglutinin gene of human influenza virus A. Mol Biol Evol 16:1457–1465

Collaborative Computational Project (1994) The CCP4 suite: programs for protein crystallography. Acta Crystallogr D Biol Crystallogr 50:760–763

Dalgleish AG, Beverley PC, Clapham PR, Crawford DH, Greaves MF, Weiss RA (1984) The CD4 (T4) antigen is an essential component of the receptor for the AIDS retrovirus. Nature 312:763–767

Endo T, Ikeo K, Gojobori T (1996) Large-scale search for genes on which positive selection may operate. Mol Biol Evol 13:685–690

Fitch WM, Bush RM, Bender CA, Cox NJ (1997) Long term trends in the evolution of H(3) HA1 human influenza type A. Proc Natl Acad Sci USA 94:7712–7718

Gilson MK, Honig B (1988) Calculation of the total electrostatic energy of a macromolecular system: solvation energies, binding energies, and conformational analysis. Proteins 4:7–18

Greer J (1990) Comparative modeling methods: application to the family of the mammalian serine proteases. Proteins 7:317–334

Hanna Z, Simard C, Laperriere A, Jolicoeur P (1994) Specific expression of the human CD4 gene in mature CD4+ CD8- and immature CD4+ CD8+ T cells and in macrophages of transgenic mice. Mol Cell Biol 14:1084–1094

Heeney J, Jonker R, Koornstra W, Dubbes R, Niphuis H, Di Rienzo AM, Gougeon ML, Montagnier L (1993) The resistance of HIV-infected chimpanzees to progression to AIDS correlates with absence of HIV-related T-cell dysfunction. J Med Primatol 22:194–200

Hughes AL (1997) Rapid evolution of immunoglobulin superfamily C2 domains expressed in immune system cells. Mol Biol Evol 14:1–5

Jolly C, Phillips-Conroy JE, Turner TR, Broussard S, Allan JS (1996) SIVagm incidence over two decades in a natural population of Ethiopian grivet monkeys (*Cercopithecus aethiops aethiops*). J Med Primatol 25:78–83

Kraulis PJ (1991) MOLSCRIPT: Aa program to produce both detailed and schematic plots of protein structures. J Appl Crystallogr 24:946–950

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

Lawrence MC, Colman PM (1993) Shape complementarity at protein/protein interfaces. J Mol Biol 234:946–950

Marrack P, Endres R, Shimonkevitz R, Zlotnik A, Dialynas D, Fitch F, Kappler J (1983) The major histocompatibility complex-restricted antigen receptor on T cells. II. Role of the L3T4 product. J Exp Med 158:1077–1091

Massingham T, Goldman N (2005) Detecting amino acid sites under positive selection and purifying selection. Genetics 169:1753–1762

Nicholls A, Sharp KA, Honig B (1991) Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. Proteins 11:281–296

Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–936

Purvis A (1995) A composite estimate of primate phylogeny. Philos Trans R Soc Lond B Biol Sci 348:405–421

Reinherz EL, Schlossman SF (1980) The differentiation and function of human T lymphocytes. Cell 19:821–827

Rey-Cuille MA, Berthier JL, Bomsel-Demontoy MC, Chaduc Y, Montagnier L, Hovanessian AG, Chakrabarti LA (1998) Simian immunodeficiency virus replicates to high levels in sooty mangabeys without inducing disease. J Virol 72:3872–3886

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574

Ryu SE, Kwong PD, Truneh A, Porter TG, Arthos J, Rosenberg M, Dai XP, Xuong NH, Axel R, Sweet RW et al (1990) Crystal structure of an HIV-binding recombinant fragment of human CD4. Nature 348:419–426

Sawyer SL, Emerman M, Malik HS (2004) Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. PLoS Biol 2:E275

Schwede T, Kopp J, Guex N, Peitsch MC (2003) SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res 31:3381–3385

Scott WRP, Hunenberger PH, Tironi IG, Mark AE, Billeter SR, Fennen J, Torda AE, Huber T, Kruger P, van Gunsteren WF (1999) The GROMOS biomolecular simulation program package. J Phy Chem A 103:3596–3607

Sharp PM, Bailes E, Gao F, Beer BE, Hirsch VM, Hahn BH (2000) Origins and evolution of AIDS viruses: estimating the time-scale. Biochem Soc Trans 28:275–282

Silvestri G, Sodora DL, Koup RA, Paiardini M, O'Neil SP, McClure HM, Staprans SI, Feinberg MB (2003) Nonpathogenic SIV infection of sooty mangabeys is characterized by limited bystander immunopathology despite chronic high-level viremia. Immunity 18:441–452

Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF (2001) Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in Drosophila. Proc Natl Acad Sci USA 98:7375–7379

Swanson WJ, Nielsen R, Yang Q (2003) Pervasive adaptive evolution in mammalian fertilization proteins. Mol Biol Evol 20:18–20

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25:4876–4882

Wang JH, Yan YW, Garrett TP, Liu JH, Rodgers DW, Garlick RL, Tarr GE, Husain Y, Reinherz EL, Harrison SC (1990) Atomic structure of a fragment of human CD4 containing two immunoglobulin-like domains. Nature 348:411–418

Wang JH, Meijers R, Xiong Y, Liu JH, Sakihama T, Zhang R, Joachimiak A, Reinherz EL (2001) Crystal structure of the human CD4 N-terminal two-domain fragment complexed to a class II MHC molecule. Proc Natl Acad Sci USA 98:10799–10804

Waterston RH, Lindblad-Toh K, Birney E et al (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562

Wu H, Kwong PD, Hendrickson WA (1997) Dimeric association and segmental variability in the structure of human CD4. Nature 387:527–530

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13: 555–556

Yang Z, Swanson WJ (2002) Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. Mol Biol Evol 19:49–57

Yang Z, Kumar S, Nei M (1995) A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141:1641–1650