Detecting site-specific physicochemical selective pressures: Applications to the class-I HLA of the human major histocompatibility complex and the SRK of the plant sporophytic self-incompatibility system

Raazesh Sainudiin,* Wendy Shuk Wan Wong,† Krithika Yogeeswaran,‡
June B. Nasrallah‡, Ziheng Yang,§ and Rasmus Nielsen†

January 14, 2004

*Department of Statistical Science, Cornell University, Ithaca, NY 14853

†Department of Biological Statistics and Computational biology, Cornell University, Ithaca, New York 14853

‡Department of Plant Biology, Cornell University, Ithaca, New York 14853

§Department of Biology, University College London, London, UK

Running Head: physicochemical selective pressures

Key Words: Codon-based Markov models, likelihood ratio tests, MHC, physicochemical selection, SRK.

Corresponding Author:

Raazesh Sainudiin

Department of Statistical Science

Cornell University

301 Malott Hall

Cornell University

Ithaca, NY 14853, USA

Phone: (607) 255-8066

Fax: (607) 255-9801

rs228@cornell.edu

Abstract

Models of codon substitution are developed that incorporate physicochemical properties of amino acids. When amino acid sites are inferred to be under positive selection, these models suggest the nature and extent of the physicochemical properties selected for. This is accomplished by first partitioning the codons based on some property of the amino acids they code for, and then using this partition to parametrize the rates of property-conserving and property-altering base substitutions at the codon level by means of finite mixtures of Markov models that also account for codon and transition:transversion biases. Here, we apply this method to two positively-selected receptors involved in ligand-recognition; the class-I alleles of the human Major Histocompatibility Complex (MHC) of known structure and the S-locus Receptor Kinase (SRK) of the sporophytic self-incompatibility system (SSI) in cruciferous plants (*Brassicaceae*), whose structure is unknown. Through likelihood ratio tests we demonstrate that the positively selected MHC and SRK proteins are under physicochemical selective pressures to alter polarity, volume, polarity &/or volume and charge to various extents at some sites. An empirical Bayes approach is used to identify such sites which may be important for ligand recognition in these proteins.

The extent to which an amino-acid (AA) residue may freely change depends on its structural and functional role within a protein. The neutral theory of molecular evolution posits that most of the observed polymorphism at the molecular level is due to the random fixation of selectively neutral mutations (KIMURA 1983). At some codon sites, no nonsynonymous (AA-altering) mutations are tolerated, due to strong purifying selection against deleterious mutations, while at other sites only a restricted class of nonsynonymous mutations that preserve the overall structure and/or function of the protein may be permitted. Therefore, selectively neutral mutations at a codon site that ensure a protein's functionality are of two types: all synonymous (AA-conserving) mutations and some nonsynonymous mutations. In most proteins, only a few residues at regions of low variability are functionally crucial (for example, those residues with catalytic function at or close to the active site), whereas, the majority of other residues play a role in maintaining the appropriate three-dimensional structure of the protein to ensure its functionality (PAKULA and SAUER 1989). Therefore, most AA sites are either unlikely to change or only preferentially change over time, depending on the strength of site-specific purifying selection. It has been well established that at such preferentially changing sites, AA substitutions among physicochemically similar amino acids are more frequent than those between dissimilar ones (ZUCKERKANDL and PAULING 1965; SNEATH 1966; EPSTEIN 1967; CLARKE 1970; GRANTHAM 1974; MIYATA *et al.* 1979).

However, in certain proteins, highly variable regions turn out to be functionally important. Such hypervariable regions that are targeted by selection tend to have an excess of nonsynonymous substitutions compared to what would be expected if both synonymous and nonsynonymous substitutions occur at the same rate. Under the assumption that the rate ratio of nonsynonymous to synonymous substitutions being 1 is reflective of neutral evolution, nonparametric as well as parametric codon-based statistical methods exist in the literature to test for the presence, and quantify the strength, of positive selection. Since it is highly unlikely that a positively selected codon site will blindly accept any nonsynonymous

substitution and since different sites may evolve at different rates, parametric models have been proposed to allow for such heterogeneities. Such methods also explicitly model transition:transversion as well as codon biases operating at the DNA level. For instance, YANG *et al.* (1998) assume a parametric form based on an amino acid metric of MIYATA *et al.* (1979) for this heterogeneity in rates between distinct pairs of nonsynonymous codons, while SCHADT and LANGE (2002) partition the amino acids into similarity classes and parametrize acceptance probabilities for transitions within or between classes. Thus, in the presence of positive selection, one can learn more about its nature and not merely its strength, by deciphering any physicochemically meaningful patterns in the nonsynonymous substitutions at these hypervariable regions.

The nature and extent of site-specific physicochemical selective pressures on two positively selected receptors are quantified in this study by superimposing finite mixtures of Markov models of codon substitution, induced by the corresponding physicochemical partition of the codon state space, upon the phylogenetic tree of homologous codon sequences in a likelihood framework. The methods of NIELSEN and YANG (1998) provide a likelihood ratio test to detect elevated rates of nonsynonymous substitutions relative to synonymous substitutions acting at least upon some of the sites along an evolving protein. Building on these methods we devise a novel strategy that detects elevated rates of property-altering substitutions relative to property-conserving substitutions that can be applied to any physicochemical property of interest. The empirical Bayes method described by (NIELSEN and YANG 1998) is used to compute the posterior probability that a particular receptor site has been subject to an elevated ratio (> 1) of a particular physicochemical property-altering substitution rate to this property-conserving substitution rate.

In receptor-ligand interactions, the two proteins come in intimate contact with each other to allow recognition. At this interface, the side-chain residues from one protein interact with their counterparts on the other to form hydrogen bonds, salt bridges, etc. in physicochemically feasible ways. Tight contact is ensured when complementary residues accomodate

each other by their size and shape and when repulsion due to similar charge and/or dissimilar polarity are avoided (CREIGHTON 1996). In co-evolving protein systems, like the ones chosen for this study, residues on one protein may be under selective pressure to alter their physicochemical properties to accomodate changes in the complementary residues of the interacting protein. Since volume, polarity, and charge are important properties in protein-protein interactions, we design partitions to detect selective forces driving changes in these. We show that SRK, the female receptors involved in recognizing self-pollen in the sporophytic self-incompatibility system (SSI) of cruciferous plants, as well as HLA-A and HLA-B, the class-I glycoproteins involved in peptide recognition in the human major histocompatibility complex (MHC), are positively selected as reflected by a high rate ratio of nonsynonymous to synonymous substitutions at some sites. Furthermore, we explore the physiochemical nature of this positive selection through the reflections of elevated rate ratios of property-altering to property-conserving substitutions at some sites. All the investigated properties, namely, polarity, volume, polarity &/or volume, and charge, show elevated rate ratios at some sites. Specific residues of these proteins that may be the targets of positive selection for changes in at least one of these physicochemical properties are identified and have been shown or surmised to be important for ligand recognition.

THEORY

The methods of NIELSEN and YANG (1998) and YANG *et al.* (2000) are based on modeling the evolution of the coding sequence as a continuous time Markov chain with state space on the set of sense codons. The rate of a nonsynonymous nucleotide substitution on a codon at site h relative to that of a synonymous substitution at any site is parametrized as

$\omega^{(h)} \in (0, \infty)$ in the rate matrix as follows:

$$q_{uv} = \begin{cases} \pi_v \ , & \text{if } u \text{ and } v \text{ differ by a synonymous transversion} \\ \kappa \ \pi_v \ , & \text{if } u \text{ and } v \text{ differ by a synonymous transition} \\ \omega^{(h)} \ \pi_v \ , & \text{if } u \text{ and } v \text{ differ by a nonsynonymous transversion} \\ \omega^{(h)} \ \kappa \ \pi_v \ , & \text{if } u \text{ and } v \text{ differ by a nonsynonymous transition} \\ 0 \ , & \text{if } u \text{ and } v \text{ differ at more than one position} \end{cases} \tag{1}$$

The parameter κ is the transition/transversion rate ratio and π_v is the stationary frequency of codon v. In order to allow rate variation along the protein sequence, $\omega^{(h)}$ at site h is considered unknown and drawn from some finite mixture of rate classes with possibly distinct rates. The parameters are estimated in the maximum likelihood framework. The likelihood function is calculated via the pruning algorithm of FELSENSTEIN (1981) subsequent to the superimposition of the codon substitution process along the branches of the phylogenetic tree (NIELSEN and YANG 1998). To test if there is any evidence for positive selection, a likelihood ratio test may be performed between model M7, which uses a discretized beta distribution taking values between 0 and 1 to draw values of ω from, and model M8, which draws ω from a mixture of a discretized beta distribution (as in M7) and an extra rate class allowed to be larger than 1 (YANG *et al.* 2000). When, ω, the ratio of nonsynonymous to synonymous substitution rates, is 0 or 1 at a site then that site is thought to be under strong purifying or neutral selection, respectively. When ω is > 1 at a site then it is thought that that site is under positive selection. Thus, M7 allows nonsynonymous rate variation without positive selection ($0 \leq \omega \leq 1$), while M8 allows for positive selection, in terms of allowing ω to be larger than 1, at some of the sites. Therefore, rejecting M7 is interpreted as rejecting the hypothesis that no sites are undergoing such a positive selection. Furthermore, if M7 is rejected, then the positively selected sites can be detected by computing the posterior

probability that a particular site has a value of $\omega > 1$ under the parameter estimates of the M8 model through an empirical Bayes approach (NIELSEN and YANG 1998).

In the above formulation, the space of codons is partitioned into codons that code for the same amino acids. Ignoring the modeling aspects of codon and transition:transversion biases (κ and π_v's) for the moment, the substitution rate between codons that code for the same amino acid is set at 1 for all sites, while the substitution rate between codons that code for different amino acids is defined as ω. Thus, the nonsynonymous substitution rate is scaled relative to the synonymous substitution rate.

A simple generalization of the above framework which allows one to probe into the nature of different kinds of selective pressures acting on the protein is proposed in this study. Instead of partitioning the codons on the basis of the amino acids they code for, they are partitioned according to some physicochemical property. For instance, if polarity were chosen to be the property of interest, the codons would be partitioned into those that code for polar amino acids and those that do not. The polarity conserving rate, *i.e.*, the rate of substitution between codons that code for polar amino acids and that between codons that code for nonpolar amino acids, is set at 1. In other words, the rate of a substition at the DNA level of a codon that conserves the polarity or nonpolarity of the encoded amino acid is set at 1. However, the rate of substitution between codons that code for a polar amino acid and those that code for a nonpolar amino acid and vice versa is defined to be γ_p. One may model γ_p to vary among sites through mixture models as before. Let model M7p and M8p under the polarity-based partition of the codons be the analogs of models M7 and M8 of YANG *et al.* (2000) under the original AA-based partition. If model M7p under the polarity partition is rejected in favor of model M8p, then one may similarly conclude that there is evidence in favor of an elevated rate of polarity-altering substitutions compared to the rate of polarity-conserving substitutions at some sites. Subsequently, site-specific posterior probabilities of being subject to an elevated polarity-altering rate when compared to the polarity-conserving rate ($\gamma_p > 1$) can be computed through an analog of the empirical Bayes

approach of NIELSEN and YANG (1998). In general, by specifying a partition of codons based on some set of properties of the amino acids they code for, one can gain insight into the nature of specific physicochemical selective pressures acting along the primary sequence of a protein.

The rate matrix for such a codon substitution process with a pre-specified partition based on some physicochemical property and with the codon and transition:transversion biases accounted for is given by,

$$q_{uv} = \begin{cases} \pi_v \ , & \text{if } u \text{ and } v \text{ differ by a property-conserving transversion} \\ \kappa \ \pi_v \ , & \text{if } u \text{ and } v \text{ differ by a property-conserving transition} \\ \gamma^{(h)} \ \pi_v \ , & \text{if } u \text{ and } v \text{ differ by a property-altering transversion} \\ \gamma^{(h)} \ \kappa \ \pi_v \ , & \text{if } u \text{ and } v \text{ differ by a property-altering transition} \\ 0 \ , & \text{if } u \text{ and } v \text{ differ at more than one position.} \end{cases} \tag{2}$$

Observe that property-conserving substitutions include synonymous substitutions. Also, note that $\gamma^{(h)} = \omega^{(h)}$ when the codons are partitioned based on the amino acids they code for.

Suppose $\mathcal{A} = \{A_1, A_2, \cdots, A_i, \cdots, A_m\}$ is a partition of the set A of 20 amino acids, where $m \in \{1, \cdots, 20\}$ and $A_i = \{a_{i1}, a_{i2}, \cdots, a_{iz_i}\}$ denotes the set of z_i distinct amino acids. Each partition $\mathcal{A}$ of the amino acids induces a corresponding partition $\mathcal{C}$ of the 61 sense codons, such that $\mathcal{C} = \{C_1, \cdots, C_m\}$, where $C_i = \bigcup_{j=1}^{z_i} c_{ij}$ and c_{ij} is the set of codons that code for amino acid $a_{ij} \in A_i$. Thus C_i is the set of codons which code for the amino acids in the set A_i. This partition $\mathcal{C}$ induces a Markov chain model which parametrizes the substitution rate between codons $u \in C_i$ and $v \in C_j$ as the property-altering rate $\gamma_{\mathcal{A}}$, whenever $i \neq j$, relative to the property-conserving rate between codons $u, v \in C_i$ for all i. Thus all possible codon pairs may be divided into two groups corresponding to the conserving and altering rates of substitution they specify according to some partition $\mathcal{A}$. There are more

than 6×10^{18} models corresponding to the number of ways to partition the set A of amino acids. Once again, the finest partition $\mathcal{A}_\omega$ arises when $m = 20$ to induce the familiar model with the nonsynonymous rate ω relative to the synonymous rate.

Out of such a large class of models, some models with physicochemically meaningful partitions are of interest (see Figure 1). Besides the familiar model with its AA-based partition quantifying the nonsynonymous pressure through ω, we study four other models that arise from four partitions of codons based on the following properties of the amino acids they code for (1) polarity; polar (Y,W,H,K,R,E,Q,T,D,N,S,C) and non-polar (M,F,I,V,L,A,G,P), (2) volume; large (L,I,F,M,Y,W,H,K,R,E,Q) and small (A,G,C,S,T,D,N,P,V), (3) polarity and/or volume; large polar (Y,W,H,K,R,E,Q), small polar (T,D,N,S,C), large non-polar (L,I,F,M) and small non-polar (A,G,P,V), and (4) charge; positively charged (H,K,R), negatively charged (D,E), and uncharged (A,N,C,Q,G,I,L,M,F,P,S,T,W,Y,V). One may similarly look at models with other partitions, such as hydrophobicity, aromaticity, electrostatic potential, etc. Furthermore, appropriate partitions may be designed in light of empirical evidence for the particular system under investigation to test the extent of the hypothesized physicochemical pressures targeted by positive selection.

[Figure 1 about here.]

MATERIALS AND METHODS

Gene Sequences: Gene name, source organism, and GenBank accession numbers for the analyzed sequences are as follows:

MHC Class I HLA genes:

Homo sapiens: HLA-B*3902 (M94053), HLA-B18 (M24039), HLA-Bw42 (M24034), HLA-A2 SLU (Z27120), HLA-A11E (X13111), HLA-Aw74 (X61701), and PDB reference sequence 1QSE.

S-locus Receptor protein Kinases:

Brassica oleracea: SRK6 (M76647), SRK3 (X79432), SRK29 (Z30211) ,SRK60 (AB032474),

SRK18 (AB032473), SRK13 (SEG_AB024419S), SRK23 (AB013720), SRK15 (Y18260), SRK5 (Y18259), SRK2 (AB024416).

Brassica campestris (syn. *B. rapa*): SRK22 (AB054061), SRK29 (E15797), SRK45 (E15795), SRK46 (SEG_AB013717S), SRK12 (D38564), SRK9 (D30049), SRK8 (D38563).

Brassica napus: U00443, M9766.

Arabidopsis lyrata: SRKa (AB052755), SRKb (AB052756).

Sequence Alignment and Phylogeny Reconstruction: Amino acid sequences from each data set were aligned by using TCOFFEE (http://www. ch.embnet.org/software/TCoffee.html). The multiple alignment of AA sequences was used to obtain the corresponding codon alignment (http://bioweb.pasteur.fr/seqanal/interfaces/ protal2dna-simple.html). The maximum likelihood tree from the nucleotide sequences for each data set was obtained using the DNAML program (version 3.5c) in PHYLIP (http:// evolution.gs.washington.edu/phylip.html). For the SRK data set, a 95% confidence set of trees under a Bayesian framework was obtained using PHYBAYES (http://statgen.ncsu.edu/ stephane/softs.htm).

Likelihood Ratio Tests and Posterior Probabilities: First, a likelihood ratio test is performed by comparing models M7 and M8 under the standard AA-based partition to detect an elevation in ω, the nonsynonymous rate relative to synonymous rate for each data set. If one can reject model M7 in favor of model M8, then there is strong evidence against the ω rate ratio being confined to the interval $[0, 1]$ for all sites. This is interpreted as rejecting the absence of positive selection at all sites in favor of its presence at some of the sites as embodied by model M8. Second, we investigate the extent to which various physicochemical properties are targeted by positive selection in homologous proteins evolving with a high ω rate ratio. This is accomplished by performing similar likelihood ratio tests using the models M7$\mathcal{A}$ and M8$\mathcal{A}$ induced by the partition $\mathcal{A}$ based on some physicochemical property of the amino acids as described earlier. If M7$\mathcal{A}$ is rejected in favor of M8$\mathcal{A}$ then there is strong

evidence against the $\gamma_{\mathcal{A}}$ rate ratio being strictly confined to the interval $[0, 1]$ for all sites. This is interpreted as rejecting the absence of selective pressure to alter the physicochemical property corresponding to $\mathcal{A}$ at all sites in favor of its presence at some of the sites. Four such tests, namely, M7p versus M8p, M7v versus M8v, M7pv versus M8pv, and M7c versus M8c, are performed. These four pairs of models are induced by the above described partitions based on polarity, volume, polarity and/or volume, and charge, respectively. If one can reject the null model in each of these four tests, then there is evidence for an elevation in γ_p, the polarity-altering rate relative to the polarity-conserving rate, γ_v, the volume-altering rate relative to the volume-conserving rate, γ_{pv}, the polarty &/or volume-altering rate relative to the the polarty &/or volume-conserving rate, and γ_c, the charge-altering rate relative to the charge-conserving rate, at least at a few sites. The source code of the *codeml* program in PAML (http://abacus.gene.ucl.ac.uk/software/paml.html) was modified to accomplish the task. When M7$\mathcal{A}$ is rejected in favor of M8$\mathcal{A}$ we use the empirical Bayes approach described earlier to obtain the posterior probability that a given site has a ratio of the property-altering substitution rate to the property-conserving substitution rate larger than 1. If this posterior probability is > 0.95, then we say that the site is under pressure to change that property or that the rate ratio of property-altering to property-conserving substitutions is > 1 at the site. By sequentially examining such posterior probabilities under different M8$\mathcal{A}$ models, each of which rejects its M7$\mathcal{A}$ submodel, one obtains insight into the extent of various physicochemical pressures targeted by positive selection at each site.

APPLICATIONS

MHC class-I: A protein encoded by a gene at any one of the three highly polymorphic MHC class I loci, HLA-A, HLA-B, and HLA-C, is expressed on the cell surface, in conjunction with β_2-microglobulin, to present peptides to T-cell receptors (TCRs). This presentation leads to a cascade of cellular immune responses culminating in the destruction of the infected cell. The polymorphism is believed to be maintained by selection for resistance to pathogens

evolving to escape immune detection by avoiding HLA presentation to T-cells (DOHERTY and ZINKERNAGEL 1975). The structure of HLA is well understood (for a review see MADDEN 1995). The peptide binding region (PBR) is a highly polymorphic cleft formed between two α-helices lying side by side on a β-pleated sheet. The side chain residues directed towards the interior of this cleft form ridges that demarcate six peptide-binding pockets (A-F). Pockets A and F have a number of highly conserved residues responsible for anchoring the N and C terminal ends of the peptide antigen. All the pockets, especially B-E, contain variable residues that determine the specificity of the interaction by determining the size and shape of the pockets and thereby the size and shape of side-chain residues of the proximal region of the peptide that can be accomodated. It has been reported that even a small number of amino acid changes in the cleft will result in drastic changes in the types of peptides the HLA protein can bind to (BARBER *et al.* 1997). Amino acid replacements in this cleft appear to be driven by balancing selection from several lines of empirical evidence, including observations of allele frequency deviations from neutral expectations based on Ewens' sampling theory (HEDRICK and THOMSON 1983), elevated rates of nonsynonymous substitutions (HUGHES and NEI 1988), high levels of heterozygosity at AA sites (HEDRICK *et al.* 1991) in the antigen-binding cleft, and an excess of heterozygotes compared with Hardy-Weinberg proportions (MARKOW *et al.* 1993) or Mendelian expectations (BLACK and HEDRICK 1997). The six class I MHC alleles from HLA-A and HLA-B loci chosen for this study were already shown by SWANSON *et al.* (2001) to be under positive selection as reflected by an elevated nonsynonymous to synonymous substitution rate ratio at some sites. We further analyze this data set with models induced by other physicochemical partitions in order to gain insight into the nature and extent of various physicochemical properties targeted by positive selection at each site.

The results of the likelihood ratio tests to detect an elevation in ω, the nonsynonymous to synonymous substitution rate ratio, and an elevation in each of the four property-altering to property-conserving substitution rate ratios, γ_p, γ_v, γ_{pv}, and γ_c, for the physicochemical properties of polarity, volume, polarity and/or volume and charge, respectively, at least at a

few sites of the MHC are shown in Table 1. In each of the five tests, the null model is strongly rejected and the maximum likelihood estimate of property-altering to property-conserving substitution rate ratio under the alternative model is larger than 1 at some sites. Thus, there is strong evidence against the absence of positive selection at all sites, in terms of the ω rate ratio being confined by the discretized Beta distribution of the null model to the interval $[0, 1]$ for all sites. Furthermore, there is just as strong an evidence against each of the four property-altering to property-conserving rate ratios being confined to the interval $[0, 1]$ for all sites.

The amino acid sites of the MHC protein that have a high posterior probability (> 0.95) of being under at least one of the elevated substitution rate ratios are identified. We find that under this stringent posterior probability cut-off, $\omega > 1$ at sites 114 and 156, $\gamma_p > 1$ at site 116, $\gamma_v > 1$ at sites 63, 67, and 97, $\gamma_{pv} > 1$ at sites 45, 63, 67, and 97, and $\gamma_c > 1$ at sites 45, 114 and 156. These sites are mapped onto the crystal structure of an MHC class I HLA-A2 protein (PDB file 1QSE) using RASMOL v2.7.2.1 (http://www.bernstein-plus-sons.com/software/rasmol/) as shown in Figure 2. All 7 selected sites are found to be present in one or more of the 6 peptide-interacting pockets (A-F) of the antigen binding cleft. Residues 45 and 67 are found in Pocket B, 63 in A and B, 97 in C and E, 114 in C, D and E, 116 in C and F and 156 in D and E.

[Figure 2 about here.]

Approximately 1.5% of all sites in the MHC proteins are under polarity-altering pressure ($\gamma_p > 1$), whereas 6%, 7%, and 14% are under charge-altering ($\gamma_c > 1$), AA-altering ($\omega > 1$), and volume-altering ($\gamma_v > 1$) pressures, respectively. Thus, different proportions of sites are under different kinds of physicochemical selection pressures. The small proportion of sites with $\gamma_p > 1$ in the MHC is consistent with the observation that most nonsynonymous codon substitutions involving a single mutation are polarity-conserving (EPSTEIN 1967). The larger proportion of sites with $\gamma_v > 1$ may partly reflect that substitutions altering

the volume of an amino acid in the antigen-binding cleft more often improve its ability to physically accomodate novel peptides, compared to substitutions that alter other properties, and are thereby selected and retained in the population. As expected, these sites are all found in pockets B-E which are thought to play an important role in determining ligand specificity (HASHIMOTO *et al.* 1999).

Out of the 43 polymorphic sites in our alignment of 6 MHC proteins, only 7 sites are seen to be targeted by positive selection as reflected by an elevation in at least one of the five substitution rate ratios under study. All these sites are within an Ångstrom away from active Van der Waal interactions with a peptide residue (within 5 Ångstroms) based on an analysis of several structurally resolved peptide-MHC complexes (RECHE and REINHERZ 2003). Though no structural studies on the 6 HLA variants selected for this study exist, x-ray crystallographic studies of other HLA-A and HLA-B proteins (GUO *et al.* 1993; MACDONALD *et al.* 2003), site-specific mutagenesis studies (DOMENECH *et al.* 1991) and studies on varying specificity resulting from single site variations in HLA-B subtypes (HULSMEYER *et al.* 2002; MACDONALD *et al.* 2003) report all 7 of these selected sites (or regions containing them) to be critical in peptide recognition and peptide repertoire determination. This may either be through direct biochemical interactions with peptide side-chain residues or by indirect interactions with other residues and water molecules in the pocket in a manner that defines the physical dimensions and the physicochemical environment of the PBR resulting in specificity for certain peptide conformations.

The only two sites, 114 and 156, with a high posterior probability of $\omega > 1$ are located in Pocket E. Site 156 was determined to be buried at the base of this pocket, but still critical in influencing the extent of cleft opening through its interactions with adjacent residues in HLA-B44 variants (MACDONALD *et al.* 2003). D156 (negatively-charged) forms a salt bridge with R97 (positively-charged) which allows a H-bond with D114, a strong interaction that indirectly results in the whole cleft tightening. When AA156 is neutral, a shallow pocket is presented for residue-3 of the ligand to interact with. It also results in a change

in conformation that forces apart residues D116 and Y74. As a result, these residues must interact via H-bonds to a water molecule, rather than directly, that causes the cleft to open up. We found that residues 114 and 156 were under pressure to alter charge ($\gamma_c > 1$) but not polarity as shown in Figure 3.

[Figure 3 about here.]

Each of the 5 other critical residues, namely 45, 63, 67, 97, and 116, has a posterior probability > 0.80 that $\omega > 1$. Clearly, these sites have escaped detection under the standard AA-based partition due to the stringent cut-off of 0.95 which was chosen to increase accuracy in the possible presence of any unaccounted recombination (as discussed below). Site 97 , also found in pocket E has been shown to interact with either residues 156 or 114 interactions which as previously mentioned are responsible for cleft tightening or loosening(MACDONALD *et al.* 2003) . In our analysis this residue was found to have the highest posterior probability of being targeted for changes in volume ($\gamma_v > 1$) compared to the other physicochemical changes, a property that is perhaps essential to allow these alternative interactions within the pocket.

Sites 45, 63 and 67 all lie within pocket B which is critical for specificity. The side chain of residue 2 of the peptide (P2) interacts with these residues in this pocket. In B*3501, F67 forms a pocket sterically favorable for a proline at P2 (SMITH *et al.* 1996), and in the two HLA-A alleles the V at this position may either block or permit access to the bottom of the pocket depending on the orientation of its side chain. In the former case this leads to selection of small side-chained P2 residues (A*6801), and in the latter case, to the selection of larger, non-polar side chains at site P2 to interact with the non-polar M45 in A*0201 (GUO *et al.* 1993). Site 45 is located at the bottom of the B pocket. This site is a conserved M in all HLA-A alleles but variable in HLA-B alleles (RECHE and REINHERZ 2003). In the latter case variants observed include E (negatively-charged) and K (positively-charged) which select for R (positively-charged) and E (negatively-charged), respectively at site P2

in the peptide resulting in the formation of salt bridges between them (MACDONALD *et al.* 2003; GUO *et al.* 1993), aside from neutral residues like T and M (RECHE and REINHERZ 2003). We observe this site to be targeted for changes in charge ($\gamma_c > 1$) in our analysis. Since this site also has an elevated substitution rate ratio under the polarity and/or volume partition ($\gamma_{pv} > 1$), both these properties may be critical in residue selection at P2, especially for HLA-A and others B alleles with neutral AA residues at this position.

Site 116 is the only site with an elevated rate ratio exclusively under the polarity partition ($\gamma_p > 1$). This site is located at the bottom of the F Pocket and is important in determining 'anchoring residues' in the C-terminal end of the peptide ligand. Polarity appears to be critical in allele HLA-B*4402 where D116 (polar) forms a direct H-bond to Y74 (MACDONALD *et al.* 2003) resulting in the tightening of the entire binding cleft (MACDONALD *et al.* 2003). Selection for change in polarity at site 116 may therefore be pivotal in tightening or relaxing the binding cleft and consequently narrowing or expanding the range of peptide specificity. However, studies in mammals demonstrate that bulky aromatic residues (Y or F) that select L, I or V in the peptide, or small residues (D or S) which accomodate the long side chain-bearing peptide residues (R or K), are commonly seen at this position (YOUNG *et al.* 1995; FALK *et al.* 1991). Therefore in this particular case, polarity does not seem to be the only property that might be subject to selective pressures, but a potentially crucial one, at least for this data set. A closer look at our data shows a posterior probability of 0.88 for this site under the volume-altering partition, below the threshold of 0.95 and is likely to be due to the small size of the data set or a problem inherent to the method itself. Namely, partitions may have several overlapping pairs of codons defining their property-altering rate, so interpretations of physicochemical pressures needs to be made cautiously (as discussed below). It is helpful to look at the posterior probabilities under all five partitions for the sites selected by at least one partition (Figure 3).

SRK: The self/nonself discriminating sporophytic self-incompatibility (SSI) system in hermaphroditic flowering plants of the crucifer family (*Brassicaceae*) prevents self-fertilization by inhibiting the germination and growth of self-related pollen tubes (reviewed in NASRALLAH 2002). Specificity in SSI is determined by two highly polymorphic genes that are tightly linked within the *S*-locus complex. One gene encodes the *S*-locus receptor kinase protein (SRK), a single-pass transmembrane serine/threonine kinase displayed on the surface of the stigmatic epidermal cells that cap the female pistil (STEIN *et al.* 1996). The second gene encodes the *S*-locus cysteine-rich protein (SCR), a 50-59 amino-acid long, hydrophilic, and positively charged protein located in the outer coat of pollen grains (SCHOPFER *et al.* 1999). SCR has been shown to be the ligand for SRK, and specificity in the SSI response derives from allele-specific interactions between the SRK and SCR encoded within the same *S*-locus haplotype (KACHROO *et al.* 2002). During pollination, pollen grains released from the male anthers come in contact with the stigmatic epidermal cells, allowing the pollen-specific SCRs to interact with the stigma-specific SRKs. Upon self-pollination, SCR binds the SRK ectodomain, thereby activating the receptor and triggering a signalling cascade that leads to inhibition of pollen tube development at the stigma surface. In cross-pollinations, SCR does not bind SRK, the SRK-mediated signalling cascade is not activated, and non-self pollen germinates and grows to effect fertilization. The implication of this mechanism of self-recognition is that the SRK and SCR genes must coevolve to maintain the specific interaction of their products.

Crucifer species typically exhibit a large number of *S* haplotypes, each with unique *SRK* and *SCR* alleles. Several of these *S* haplotypes predate speciation events (DWYER *et al.* 1991; UYENOYAMA 1995) and are thought to be under balancing selection. NISHIO and KUSABA (2000) have established three hypervariable regions (HVR1, HVR2, HVR3) and a C-terminal variable region (CVR) in the SRKs based on amino acid variation. By computing the average numbers of synonymous (π_S) and nonsynonymous (π_N) nucleotide differences per site between two randomly chosen sequences, SATO *et al.* (2002) show that the ratio of

π_N:π_S in the hypervariable regions is > 1, indicating the operation of positive selection.

Likelihood ratio tests identical to those done on the MHC are carried out on the SRK proteins and summarized in Table 2. All five null hypotheses are strongly rejected for the SRK data set as well. Thus, there is strong evidence against the absence of positive selection at all sites, in terms of the substitution rate ratios, ω, γ_p, γ_v, γ_{pv}, and γ_c induced by their respective partitions being confined to the interval $[0, 1]$ for all sites. Figure 4 shows all the 40 amino acid sites with a high posterior probability (> 0.95) that at least one of these substitution rate ratios is greater than 1. The numbering corresponds to the protein sequence of *B. oleracea* SRK60. Intuitively, the different partitions of the codon space and the corresponding Markov models they induce may be thought of as different sieves, each efficient at filtering out sites with certain patterns of substitutions relative to certain other patterns of substitution. Thus, by using several sieves one can filter out several kinds of sites. Figure 4 would thus show the results of several sieves applied to filter out SRK sites under different physicochemical pressures.

[Figure 4 about here.]

All these selected sites lie in the S domain of SRK which functions in recognition of self-SCR. A majority of the selected sites are under all five pressures. However, some sites are only under specific combinations of pressures. For instance, sites 206, 217, and 105 all have high posterior probabilities (> 0.95) of being under AA-altering pressure ($\omega > 1$). However, only the first two sites have high posterior probabilities of being under polarity-altering pressure ($\gamma_p > 1$), whereas site 105 is under volume-altering pressure ($\gamma_v > 1$). Several sites that did not have a posterior probability > 0.95 of being under an elevated nonsynonymous to synonymous rate ratio ($\omega > 1$) have high posterior probabilities of being under an elevated property-altering to property-conserving rate ratio (sites outside the solid ellipse in Figure 4). Sites 208 and 215 of SRK have posterior probabilities > 0.95 of being under AA-altering pressure ($\omega > 1$), but not for being under polarity-altering, volume-altering, or polarity &/or

volume-altering pressures. However, they are under charge-altering pressure ($\gamma_c > 1$). This additional information about these positively selected sites being exclusively targeted for charge alteration is useful since the coevolving SCR is a highly-charged ligand. There are several more sites only under charge-altering pressure and several other sites under pressure to alter some set of properties but not others. Figure 5 shows the posterior probabilities of an elevation in the property-altering to property-conserving substitution rate ratio, under each of the five partitions, for these 40 selected sites. Each selected sites is subject to different physicochemical pressures to different extents as evident from Figure 5. Thus, this approach sheds light on the extent of various site-specific physicochemical selective forces acting on a protein evolving with $\omega > 1$ at some of the sites.

[Figure 5 about here.]

Sites 205-220, 270-306, 328-343, and 413-423 of *B. oleracea* SRK60 correspond respectively to the hypervariable regions HVR1, HVR2, HVR3, and CVR established by (NISHIO and KUSABA 2000), which together comprise 80 sites in these four hypervariable regions. Our analysis presents 40 sites with high posterior probabilities of being under at least one of the five physicochemical selective pressures studied. Out of these 40 sites, only 27 belong to the hypervariable regions of (NISHIO and KUSABA 2000), and 14 have posterior probabilities < 0.95 of being under an elevated rate ratio of $\omega > 1$. Sites 215-219 are close to an indel mutation and the alignment ambiguity in this region could give rise to false signals of positive selection. This analysis used 21 SRK sequences which included receptors from both class-I and class-II Brassica alleles as well as two *A. lyrata alleles.* An identical analysis using only 15 Brassica class-I alleles yielded similar conclusions (results not shown).

The likelihood analyses presented here are only done on one tree topology, namely the maximum likelihood topology under the DNA substitution process of the SRK coding sequences using *dnaml.* In order to ensure that uncertainty in tree topology does not affect the conclusions drawn from hypothesis tests, the likelihood ratio tests were repeated on each

of 17 distinct but similar topologies that contained 95% of the posterior probability mass on the space of possible topologies for 21 SRKs using PHYBAYES. Rejection of model M7 in favor of M8 occurred under each topology, thus establishing the presence of positive selection, as reflected by $\omega > 1$ at some of the sites (results not shown). In order to evaluate the performance of the posterior predictions that a given site is under a certain physicochemical pressure, data were simulated under the charge and polarity partitions, with the maximum likelihood parameter estimates of the SRK data, and evaluated. The false positive rates were found to be extremely low ($< 1\%$).

DISCUSSION

In positively selected proteins, one can learn more about the nature and strength of various physicochemical forces shaping its evolution by studying the patterns in nonsynonymous substitutions at hypervariable sites. Some sites on a positively selected protein may have only undergone neutral nonsynonymous mutations and therefore may not be the true targets of natural selection, although they would elevate the estimated nonsynonymous rate. At other sites, there may have been a comparable number of nonsynonymous substitutions but these may have been of a particularly advantageous physicochemical nature, e.g. volume-altering substitutions, thus making them the true targets of selection. Both types of sites would have similar posterior probabilities of being under an elevated rate ratio of nonsynonymous to synonymous substitutions ($\omega > 1$). However, the latter type of sites will have higher posterior probabilities of being under volume-altering pressure ($\gamma_v > 1$) than of being under AA-altering pressure ($\omega > 1$). Thus they may be overlooked by an analysis that exclusively focuses on the rate of generic nonsynonymous rates relative to synonymous rates of substitution at the codon level.

Most parametric models of molecular evolution assume that the only source of diversity is point mutations. However, both intra-locus and inter-locus recombination are thought to play a role in the generation of polymorphisms at the MHC (for a review see CARRINGTON

1999). Ignoring recombination partly amounts to compromising for one"average" tree for all sites, which is at best a projection of the true ancestral recombination graph residing in a correlated product treespace. ANISIMOVA *et al.* (2003) reported that while LRTs in a maximum likelihood framework are robust to low levels of recombination, at higher levels the tests may mistake recombination as evidence of positive selection. They find that the M7-M8 likelihood ratio test to detect positive selection is the least affected by recombination. Based on simulations of the Hepatitis D small antigen gene, they also found that the Bayes' prediction of sites under positive selection, based on their posterior probabilities being > 0.95 under the parameter estimates of the M8 model, had a high accuracy of 91% even when there were on average 46.7 recombination events in the history of a sample of 30 sequences. Since SSI is observed to segregate as a single Mendelian trait, the *S*-locus genes are thought to be tightly linked (CASSELMAN *et al.* 2000). Sequence divergence between *S*-haplotypes, extensive genic rearrangements, and the presence of repetitive elements are factors thought to contribute to the suppression of both intergenic and intragenic recombination at this locus (BOYES and NASRALLAH 1993; BOYES *et al.* 1997). A recombination analysis of a population segregating for the S6 haplotype, in which the *SRK* and *SCR* genes are separated by more than 200 Kb of sequence (J.B. Nasrallah, unpublished observations), failed to uncover crossovers between these genes in 800 chromosomes analysed (CASSELMAN *et al.* 2000). This result indicates that recombination in the *S* locus is a rare event. A recent study by CHARLESWORTH *et al.* (2003) looked at patterns of linkage disequillibrium in SRK alleles from *A. lyrata* and found no evidence supportive of recombination occuring in these genes.

Inspite of the possibility for recombination in MHC, conclusions of physicochemical pressures operating on the six class-I MHC proteins, inferred from LRTs as well as the posterior prediction of sites under selective pressure to alter various physicochemical properties, are in agreement with other empirical studies (see previous section). The robustness is partly due to lower levels of recombination in the class-I telomeric region of HLA-A and lack of

crossovers in regions between HLA-B and HLA-C, unlike the class-II region with its recombination hotspots (CARRINGTON 1999). Variation in genomic arrangement that disturbs homology upon pairing is thought to be the cause of reduced recombination. Our very stringent posterior probability cut-off of 0.95 also assures a higher accuracy of detecting positively selected sites (ANISIMOVA *et al.* 2003). We expect our analysis of the SRK to be at least as robust as that of the MHC due to much lower levels of recombination at the S locus. Nonetheless, it is possible for just a few recombination events of the right size and at the right time, to cause distortions to these methods which assume their absence.

As with the class I MHC protein, the selected sites in the SRK protein may be involved in ligand (SCR) recognition. Therefore, identification of positively selected sites may also help with the identification of regions that play a functional role in female-male (SRK-SCR) interaction and coevolution. In the absence of a 3D structure, this information may be useful in proposing hypotheses for empirical studies aimed at deciphering the regions responsible for SRK-SCR interaction. Ultimately protein structure and empirical studies are necessary to corroborate the functional significance of the sites identified in this study.

The novelty of this approach stems from looking at several different partitions of the codon state space. Each such partition is based on some physicochemical property of the encoded amino acids and parametrizes a rate ratio of the corresponding property-altering to property-conserving substitutions. The maximum likelihood estimates of finite mixtures of each such rate ratio shed light on its variation with or without the possibility of elevation (> 1) across different sites. However, the weakness of the approach comes from analyzing these partition-based models one at a time. This formulation precludes any formal comparison of models induced by different partitions through their likelihood ratios. Moreover, in the current approach, the rate of one type of nonsynonymous substitution is scaled relative to the rate of all other types of nonsynonymous substitutions as well as all synonymous substitutions pooled together. This may not be biologically meaningful. For these reasons we consider these methods to be tools for data exploration more than a complete framework for testing

hypotheses regarding positive selection on physicochemical properties.

Clearly, we have made two drastic simplifications for reasons of estimability and computational speed. First, we assume that no recombination is occuring. Second, we assume that a collection of time-homogeneous finite mixture of Markov chains approximates the time-nonhomogeneous spatially correlated substitutional process. Improvements should be made on both fronts. The most natural solution toward addressing the latter would be to construct models that simultaneously allow more than one partition to be parametrized. For instance, such mixture models at the codon level would allow the estimation of volume-altering substitution rate ($\gamma_v \geq 0$) as well as the nonsynonymous substitution rate ($\omega \geq 0$) relative to the synonymous substitution rate which could be set to 1. The corresponding site-specific posterior probabilities over the positive orthant with γ_v and ω as its axes would shed more light. Moreover, such models would provide a natural setting to test hypotheses of physicochemical subfunctionalization especially when recast in the lineage as well as site-specific framework of YANG and NIELSEN (2002).

[Table 1 about here.]

[Table 2 about here.]

ACKNOWLEDGMENTS

R.S. is supported by the IGERT fellowship from National Science Foundation grant DGE-9870631, National Science Foundation grant DEB-0089487 to Rasmus Nielsen, and National Science Foundation grant DEB-9602229 to Carlos Castillo-Chavez. K.Y. is supported in part by National Institutes of Health grant GM57527 to June Nasrallah. Z.Y. is supported by grants from the Biotechnology and Biological Sciences Research Council (UK) and the Human Frontier Science Program (EU). R.S. thanks Willie Swanson for guiding him through various phylogenetic tools, Wa Yang and Joel Bielewski for the most insightful discussions on physicochemical selection, Stephane Aris-Brosou for discussions on Bayesian confidence

sets on tree spaces, Aardra Pontis for discussions on the SSI system, and Richard Durrett and Matt Dimmic for comments on earlier drafts.

LITERATURE CITED

ANISIMOVA, M., R. NIELSEN, and Z. YANG, 2003 Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics **03:** 1229–1236.

BARBER, L. D., L. PERCIVAL, K. L. ARNETT, K. E. GUMPERZ, L. CHEN, and P. PARHAM, 1997 Polymorphism in the α1 helix of the HLA-B heavy chain can have an overriding influence on peptide-binding specificity. Jnl. Immunol. **158:** 1660–1669.

BLACK, F. L. and P. W. HEDRICK, 1997 Strong balancing selection at HLA loci: Evidence from segregation in South Amerindian families. Proc. Natl. Acad. Sci. USA **94:** 12452–12456.

BOYES, D. C. and J. B. NASRALLAH, 1993 Physical linkage of the SLG and SRK genes at the self-incompatibility locus of *Brassica oleracea.* Mol. Gen. Genet. **236:** 369–373.

BOYES, D. C., M. E. NASRALLAH, J. VREBALOV, and J. B. NASRALLAH, 1997 The self-incompatibility (S) haplotypes of *Brassica* contain highly divergent and rearranged sequences of ancient origin. Plant Cell **9:** 237–247.

CARRINGTON, M., 1999 Recombination within the human MHC. Immunol. Rev. **167:** 245–256.

CASSELMAN, A. L., J. VREBALOV, J. A. CONNER, A. SINGHAL, J. GIOVANNONI, M. E. NASRALLAH, and J. B. NASRALLAH, 2000 Determining the physical limits of the *Brassica* S locus by recombinational analysis. Plant Cell **12:** 23–33.

CHARLESWORTH, D., C. BARTOLOMÉ, M. H. SCHIERUP, and K. MABLE, 2003 Haplotype structure of the stigmatic self-incompatibility gene in natural populations of *Arabidopsis lyrata.* Mol. Biol. Evol. **20:** 1741–1753.

CLARKE, B., 1970 Selective constraints on amino-acid substitutions during the evolution of proteins. Nature **228:** 159–160.

CREIGHTON, T. E., 1996 *Proteins: structures and molecular properties.* New York: W. H. Freeman and co. .

DOHERTY, P. C. and R. M. ZINKERNAGEL, 1975 Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. Nature **256:** 50–52.

DOMENECH, N., J. SANTOS-AGUADO, and J. A. LOPEZ DE CASTRO, 1991 Antigenicity of HLA-A2 and HLA-B7. Loss and gain of serological determinants induced by site-specific mutagenesis at residues 62-80. Hum. Immunol. **30:** 140–146.

DWYER, K. G., M. A. BALENT, J. B. NASRALLAH, and M. E. NASRALLAH, 1991 DNA sequences of self-incompatibility genes from *Brassica campestris* and *B. oleracea*: polymorphism predating speciation. Plant Mol. Biol. **16:** 481–486.

EPSTEIN, C. J., 1967 Non-randomness in amino-acid changes in the evolution of homologous proteins. Nature **215:** 355–359.

FALK, K., O. ROTZSCHKE, S. STEVANOVIC, G. JUNG, and R. H-G., 1991 Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. Nature **351:** 290–296.

FELSENSTEIN, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. Jnl. Mol. Evol. **17:** 368–376.

GRANTHAM, R., 1974 Amino acid difference formula to help explain protein evolution. Science **185:** 862–864.

GUO, H. C., D. R. MADDEN, M. L. SILVER, T. S. JARDETZKY, J. C. GORGA, J. L. STROMINGER, and D. C. WILEY, 1993 Comparison of the P2 specificity pocket in the three human histocompatibility antigens: HLA-A*6801, HLA-A*0201, and HLA-B*2705. Proc. Natl. Acad. Sci. USA **90:** 8053–8057.

HASHIMOTO, K., K. OKAMURA, H. YAMAGUCHI, M. OTOTAKE, T. NAKANISHI, and Y. KUROSAWA, 1999 Conservation and diversification of MHC class I and its related molecules in vertebrates. Immunol. Rev. **167:** 81–100.

HEDRICK, P. W. and G. THOMSON, 1983 Evidence for balancing selection at HLA. Genetics **104:** 449–456.

HEDRICK, P. W., T. S. WHITTAM, and P. PARHAM, 1991 Heterozygosity at Individual Amino Acid Sites: Extremely High Levels for HLA-A and -B Genes. Proc. Natl. Acad. Sci. USA **88:** 5897–5901.

HUGHES, A. L. and M. NEI, 1988 Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335:** 167–170.

HULSMEYER, M., R. C. HILLIG, A. VOLZ, M. RUHL, W. SCHRODER, W. SAENGER, A. ZIEGLER, and B. UCHANSKA-ZIEGLER, 2002 HLA-B27 subtypes differentially associated with disease exhibit subtle structural alterations. Jnl. Biol. Chem. **277:** 47844–47853.

KACHROO, A., M. E. NASRALLAH, and J. B. NASRALLAH, 2002 Self-incompatibility in the Brassicaceae: receptor-ligand signaling and cell-to-cell communication. Plant Cell **14:** S227–S238.

KIMURA, M., 1983 *The neutral theory of molecular evolution.* Cambridge, England: Cambridge University Press.

MACDONALD, W. A., A. W. PURCELL, N. A. MIFSUD, L. K. ELY, D. S. WILLIAMS, L. CHANG, J. J. GORMAN, C. S. CLEMENTS, L. KJER-NIELSEN, D. M. KOELLE, S. R. BURROWS, B. D. TAIT, R. HOLDSWORTH, A. G. BROOKS, G. O. LOVRECZ, L. LU, J. ROSSJOHN, and J. MCCLUSKEY, 2003 A naturally selected dimorphism within the HLA-B44 supertype alters class I structure, peptide, repertoire, and T-cell recognition. Jnl. Exp. Med. **198:** 679–691.

MADDEN, D. R., 1995 The three-dimensional structure of peptide-MHC complexes. Annu. Rev. Immunol. **13:** 587–622.

MARKOW, T., P. W. HEDRICK, K. ZUERLEIN, D. J., J. MARTIN, T. VYVIAL, and C. ARMSTRONG, 1993 HLA polymorphism in the Havasupai: evidence for balancing selection. Am. Jnl. Hum. Genet. **53:** 943–952.

MIYATA, T., S. MIYAZAWA, and T. YASUNAGA, 1979 Two types of amino acid substitution in protein evolution. Jnl. Mol. Evol. **12:** 219–236.

NASRALLAH, J. B., 2002 Recognition and rejection of self in plant reproduction. Science **296:** 305–308.

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

NISHIO, T. and M. KUSABA, 2000 Sequence diversity of SLG and SRK in *Brassica oleracea L.* Ann. Botany **85:** 141–146.

PAKULA, A. A. and R. T. SAUER, 1989 Genetic analysis of protein stability and function. Annu. Rev. Genet. **23:** 289–310.

RECHE, P. A. and E. L. REINHERZ, 2003 Sequence variability analysis of human class I and class II MHC molecules: functional and structural correlates of amino acid polymorphisms. Jnl. Mol. Biol. **331:** 623–641.

SATO, K., T. NISHIO, R. KIMURA, M. KUSABA, T. SUZUKI, K. HATAKEYAMA, D. J. OCKENDON, and Y. SATTA, 2002 Coevolution of the S-locus genes SRK, SLG and SP11/SCR in *Brassica oleracea* and *B. rapa.* Genetics **162:** 931–940.

SCHADT, E. and K. LANGE, 2002 Codon and rate variation models in molecular phylogeny. Mol. Biol. Evol. **19:** 1534–1549.

SCHOPFER, C. R., M. E. NASRALLAH, and J. B. NASRALLAH, 1999 The male determinant of self-incompatibility in *Brassica.* Science **286:** 1697–1700.

SMITH, K. J., S. W. REID, K. HARLOS, A. J. MCMICHAEL, D. I. STUART, J. I. BELL, and E. Y. JONES, 1996 Bound water structure and polymorphic amino acids act together to allow the binding of different peptides to MHC class I HLA-B53. Immunity **4:** 215–228.

SNEATH, P. H. A., 1966 Relations between chemical structure and biological activity. Jnl. Theor. Biol. **12:** 157–195.

STEIN, J. C., R. DIXIT, M. E. NASRALLAH, and J. B. NASRALLAH, 1996 SRK, the stigma-specific S locus receptor kinase of Brassica, is targeted to the plasma membrane in transgenic tobacco. Plant Cell **8:** 429–445.

SWANSON, W. J., Z. YANG, M. F. WOLFNER, and C. F. AQUADRO, 2001 Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proc. Natl. Acad. Sci. USA **98:** 2509–2514.

UYENOYAMA, M., 1995 A genaralized least-squares estimate for the origin of sporophytic self-incompatibility. Genetics **139:** 975–992.

YANG, Z. and R. NIELSEN, 2002 Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. **19:** 908–917.

YANG, Z., R. NIELSEN, N. GOLDMAN, and A.-M. K. PEDERSEN, 2000 Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155:** 431–449.

YANG, Z., R. NIELSEN, and M. HASEGAWA, 1998 Models of amino acid substitution and applications to mitochondrial protein evolution. Mol. Biol. Evol. **15:** 1600–1611.

YOUNG, A. C. M., S. G. NATHENSON, and J. C. SACCHETTINI, 1995 Structural studies of class I major histocompatibility complex protein: insights into antigen presentation. FASEB Jnl. **9:** 26–36.

ZUCKERKANDL, E. and L. PAULING, 1965 *Evolutionary divergence and convergence in proteins. In: Bryson V. , Vogel, J. (eds) Evolving genes and proteins.* New York: Academic press.

LIST OF FIGURES

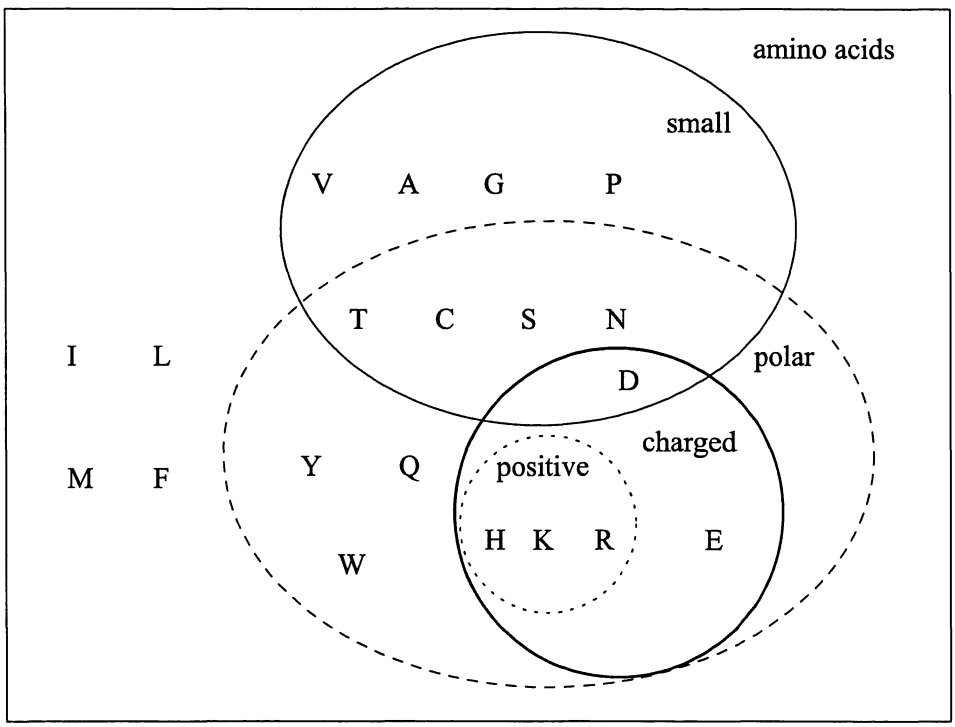

Figure 1: Partition of the amino acids based on volume (small and large), polarity (polar and nonpolar) and charge (uncharged, positively charged, and negatively charged).

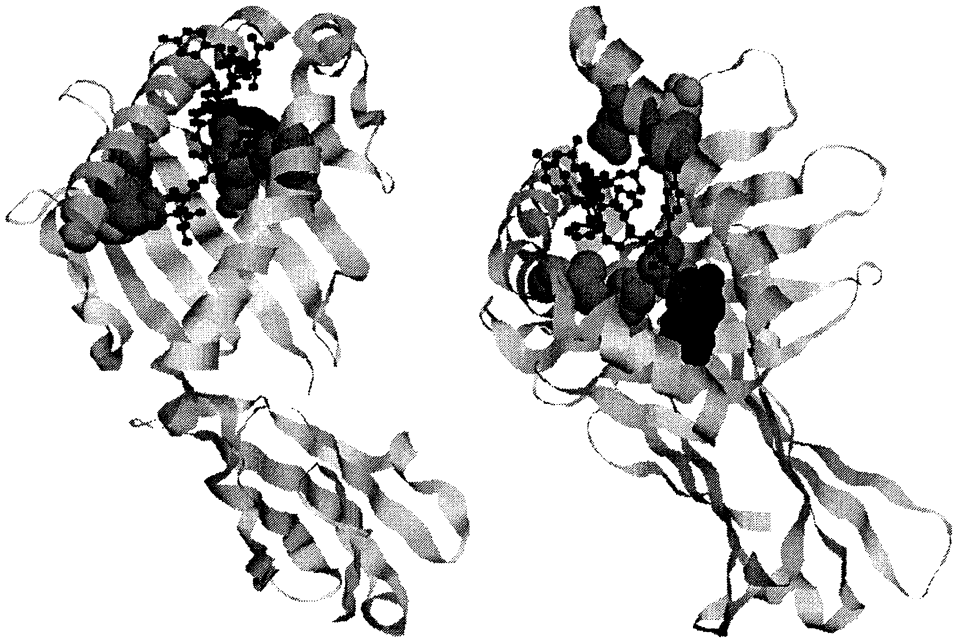

Figure 2: Sites in the MHC CLASS I protein with high posterior probability (> 0.95) of being under various physicochemical selective pressures. Sites 114 and 156 (both colored brown) are under AA-altering ($\omega > 1$) and charge-altering ($\gamma_p > 1$) pressures. Site 45(orange) is under charge-altering ($\gamma_c > 1$) as well as polarity &/or volume-altering ($\gamma_{pv} > 1$) pressures. Sites 63, 67, and 97 (all green) are under volume-altering ($\gamma_v > 1$) as well as polarity &/or volume-altering ($\gamma_{pv} > 1$) pressures. Finally, site 116 (blue) is under polarity-altering ($\gamma_p > 1$) pressure. The numbering of sites corresponds to the HLA-A2 sequence in the protein data bank file 1QSE. The viral peptide is shown in red.

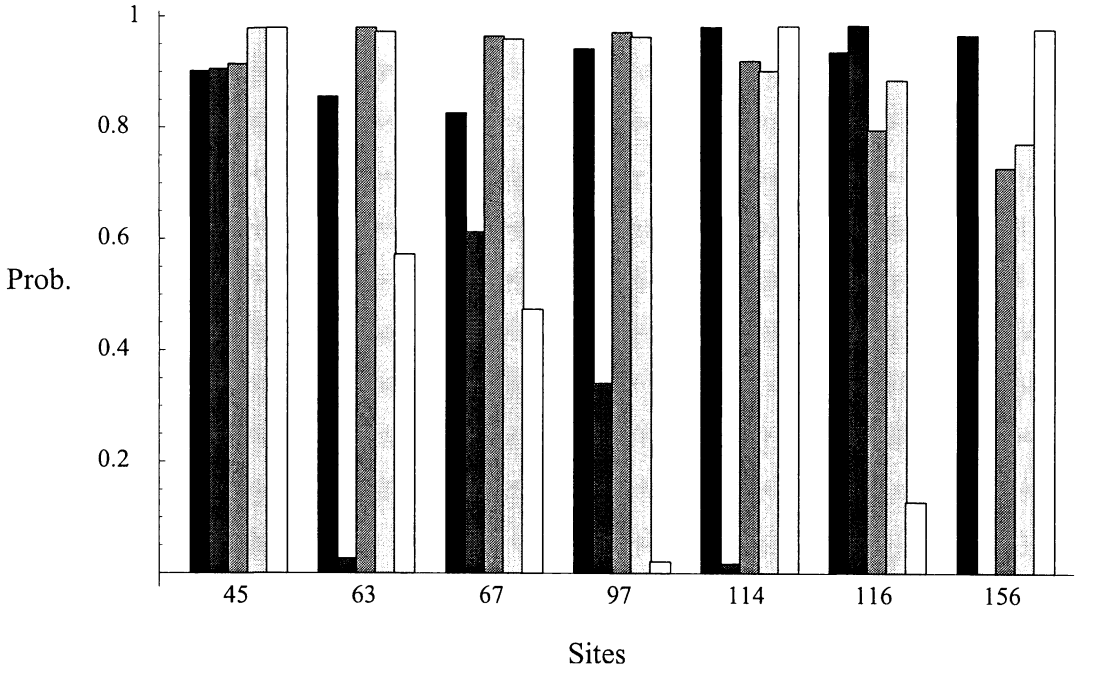

Figure 3: Posterior probabilities of $\omega > 1$ (black), $\gamma_p > 1$ (dark gray), $\gamma_v > 1$ (gray), $\gamma_{pv} > 1$ (light gray), and $\gamma_c > 1$ (white) at the 7 selected sites in MHC.

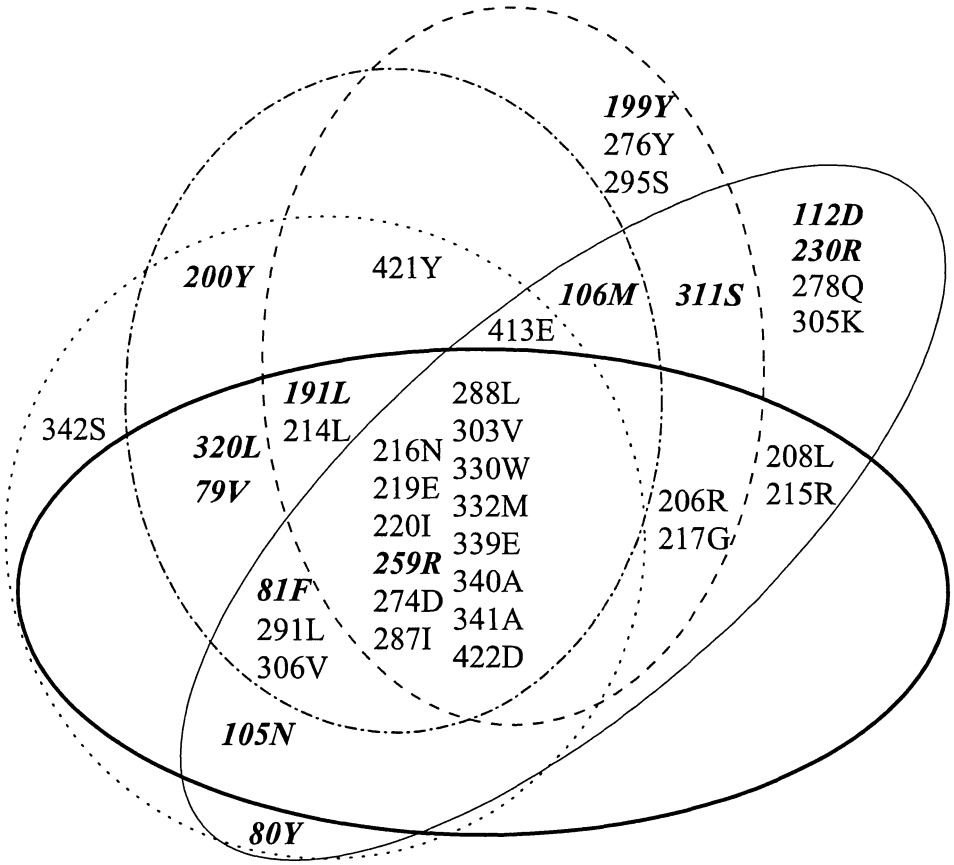

Figure 4: Sites in SRK with high posterior probability (> 0.95) of being under selective pressure to alter (i) polarity (dashed ellipse), (ii) volume (dotted circle), (iii) polarity and/or volume (dash-dotted ellipse), (iv) charge (thin solid ellipse) and (v) amino acid (thick solid ellipse). The sites shown in bold face italics are outside the HVR and CVR regions. The numbering of sites with their corresponding amino acids is that of *B. oleracea* SRK60.

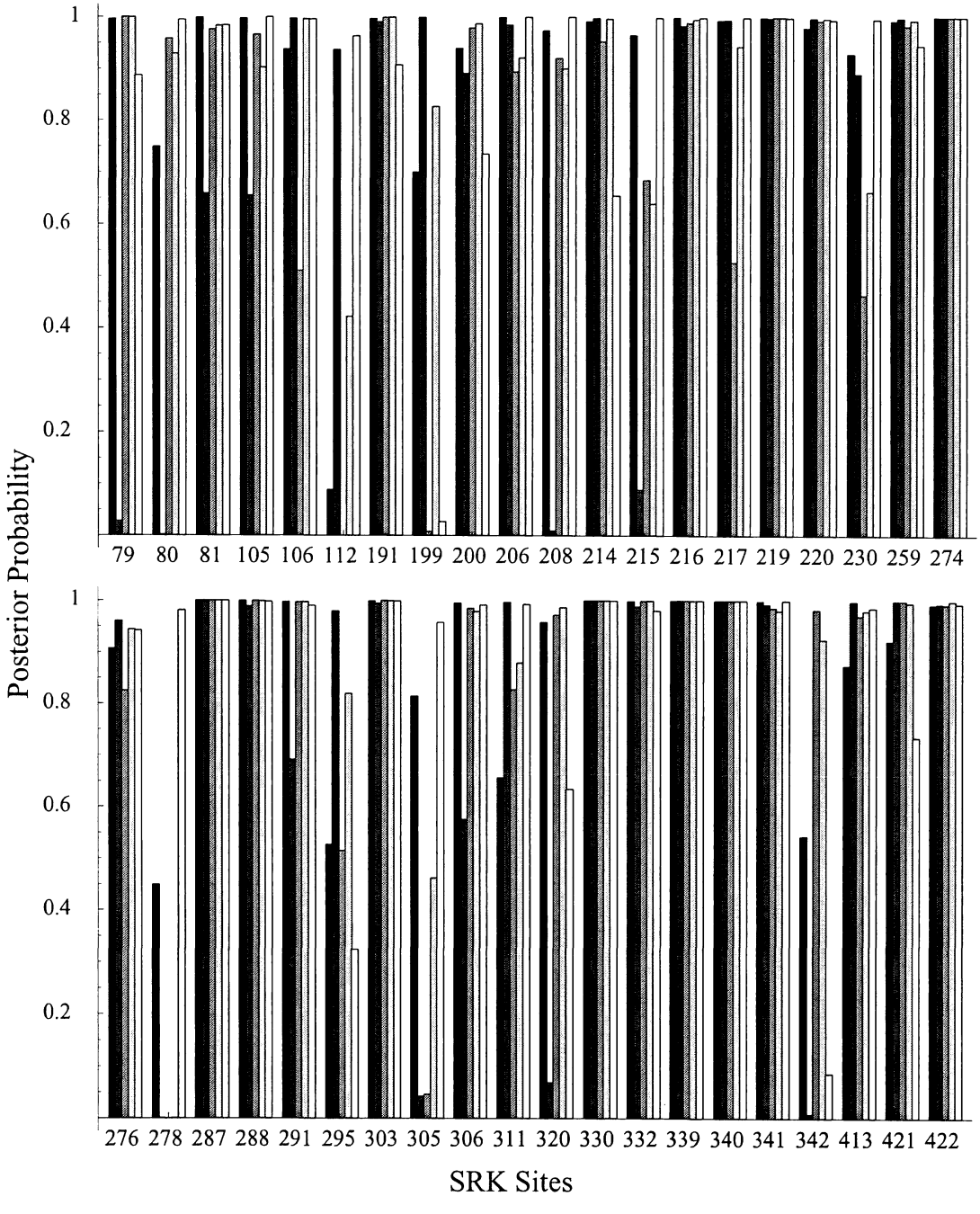

Figure 5: Posterior probabilities of $\omega > 1$ (black), $\gamma_p > 1$ (dark gray), $\gamma_v > 1$ (gray), $\gamma_{pv} > 1$ (light gray), and $\gamma_c > 1$ (white) at the 40 selected sites in SRK.

LIST OF TABLES

Table 1: Likelihood Ratio Tests for MHC Class-I HLA

Selective Pressure	Model	ℓ	Parameter estimates	$-2\Delta\ell$	P
nonsynonymous (dn/ds)	M7	-2416.94	p=0.011, q=0.022, κ=2.41		
	M8	-2409.56	p=0.104, q=0.27, κ=2.56	14.76	$\ll$0.01
			p_1=0.073, ω=4.06		
Polarity-altering	M7p	-2442.18	p=0.19, q=0.49, κ=2.42		
	M8p	-2437.25	p=0.47, q=1.46, κ=2.69	9.86	$\ll$0.01
			p_1=0.015, γ_p=44.85		
Volume-altering	M7v	-2439.71	p=0.18, q=0.35, κ=2.41		
	M8v	-2434.95	p=0.41, q=2.15, κ=2.47	9.52	$\ll$0.01
			p_1=0.14, γ_v=2.77		
Polarity &/or	M7pv	-2435.96	p=0.18, q=0.39, κ=2.42		
Volume-altering	M8pv	-2430.59	p=0.46, q=2.14, κ=2.52	10.74	$\ll$0.01
			p_1=0.12, γ_{pv}=3.04		
Charge-altering	M7c	-2439.02	p=0.19, q=0.42, κ=2.54		
	M8c	-2431.70	p=0.48, q=1.69, κ=2.68	14.64	$\ll$0.01
			p_1=0.06, γ_c=6.45		

Table 2: Likelihood Ratio Tests for SRK

Selective Pressure	Model	ℓ	Parameter estimates	$-2\Delta\ell$	P
nonsynonymous (dn/ds)	M7	-17627.95	p=0.30, q=0.48, κ=2.03		
	M8	-17507.17	p=0.36, q=0.62, κ=2.21	241.6	$\ll$0.01
			p_1=0.059, ω=3.28		
Polarity-altering	M7p	-17955.09	p=0.18, q=0.41, κ=2.04		
	M8p	-17858.48	p=0.44, q=1.25, κ=2.16	193.2	$\ll$0.01
			p_1=0.065, γ_p=5.00		
Volume-altering	M7v	-17873.72	p=0.24, q=0.44, κ=1.99		
	M8v	-17760.51	p=0.45, q=0.97, κ=2.11	226.4	$\ll$0.01
			p_1=0.054, γ_v=4.68		
Polarity &/or	M7pv	-17771.62	p=0.25, q=0.46, κ=1.98		
Volume-altering	M8pv	-17656.94	p=0.47, q=1.06, κ=2.12	229.4	$\ll$0.01
			p_1=0.068, γ_{pv}=3.58		
Charge-altering	M7c	-17917.87	p=0.22, q=0.37, κ=2.07		
	M8c	-17788.99	p=0.40, q=0.83, κ=2.21	257.76	$\ll$0.01
			p_1=0.07, γ_c=5.12		