

# Ligand design by a combinatorial approach based on modeling and experiment: application to HLA-DR4

Erik Evensen · Diane Joseph-McCarthy ·  
Gregory A. Weiss · Stuart L. Schreiber ·  
Martin Karplus

Received: 15 November 2006 / Accepted: 19 April 2007 / Published online: 27 July 2007  
© Springer Science+Business Media B.V. 2007

**Abstract** Combinatorial synthesis and large scale screening methods are being used increasingly in drug discovery, particularly for finding novel lead compounds. Although these “random” methods sample larger areas of chemical space than traditional synthetic approaches, only a relatively small percentage of all possible compounds are practically accessible. It is therefore helpful to select regions of chemical space that have greater likelihood of

yielding useful leads. When three-dimensional structural data are available for the target molecule this can be achieved by applying structure-based computational design methods to focus the combinatorial library. This is advantageous over the standard usage of computational methods to design a small number of specific novel ligands, because here computation is employed as part of the combinatorial design process and so is required only to determine a propensity for binding of certain chemical moieties in regions of the target molecule. This paper describes the application of the Multiple Copy Simultaneous Search (MCSS) method, an active site mapping and de novo structure-based design tool, to design a focused combinatorial library for the class II MHC protein HLA-DR4. Methods for the synthesizing and screening the computationally designed library are presented; evidence is provided to show that binding was achieved. Although the structure of the protein-ligand complex could not be determined, experimental results including cross-exclusion of a known HLA-DR4 peptide ligand (HA) by a compound from the library. Computational model building suggest that at least one of the ligands designed and identified by the methods described binds in a mode similar to that of native peptides.

E. Evensen · S. L. Schreiber · M. Karplus  
Committee on Higher Degrees in Biophysics, Harvard  
University, Cambridge, MA, USA

E. Evensen · D. Joseph-McCarthy · G. A. Weiss ·  
S. L. Schreiber · M. Karplus  
Department of Chemistry and Chemical Biology, Harvard  
University, Cambridge, MA, USA

*Present Address:*

E. Evensen  
Sunesis Pharmaceuticals, South San Francisco, CA, USA  
e-mail: ee@post.harvard.edu

D. Joseph-McCarthy (✉)  
Chemical and Screening Sciences Department, Wyeth Research,  
Cambridge, MA, USA  
e-mail: djoseph@wyeth.com

*Present Address:*

G. A. Weiss  
University of California, Irvine, CA, USA  
e-mail: gweiss@uci.edu

S. L. Schreiber  
Howard Hughes Medical Institute, Cambridge, MA, USA  
e-mail: stuart-schreiber@harvard.edu

M. Karplus (✉)  
Laboratoire de Chimie Biophysique, ISIS, Université Louis  
Pasteur, 8 allée Gaspard Monge, 67000 Strasbourg, France  
e-mail: marci@tammy.harvard.edu

**Keywords** Combinatorial library · MCSS ·  
Fragment docking · Structure-based drug design ·  
Active site map

## Introduction

Combinatorial synthesis combined with efficient screening procedures provides a powerful method for generating large numbers of compounds and selecting those that have high affinity for a particular target [1, 2]. Practical

considerations, however, limit these methods to sampling a relatively small portion (approximately  $10^2$  to  $10^6$  compounds) of the chemical space (on the order  $10^{200}$  compounds) of small molecules (molecular weight less than 850 Da) [3]. Recent developments in computational chemistry aid in designing libraries to focus the sampling of chemical space. This approach is particularly suited to finding ligands for a receptor with unknown structure [4, 5]. If the structure of the receptor is known, an alternative approach is to use computed functional group binding preferences to focus the combinatorial library for the specific target [6–9]. Diversity estimates of computed binding preferences can aid in this process by analyzing the distribution of possible library constituents in chemical space and designing libraries to target specific coverage. This corresponds to using the computed interaction maps (as one would use the initial iterations of a library) to narrow experimental sampling to the chemical space which is likely to yield candidate ligands. In this way, targeted libraries can be designed and higher discovery efficiency may be achieved. Subsequent libraries may be used to fill in gaps in the information derived from the first generation results and to probe more finely the most promising portions of chemical space [10–12].

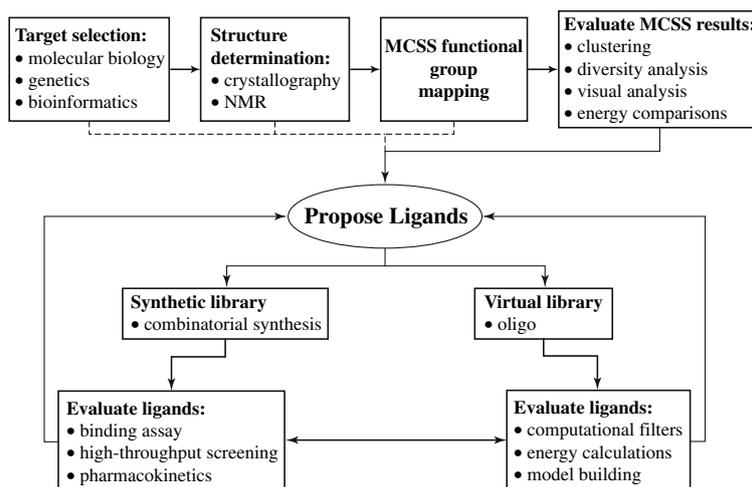
The MCSS method [13, 14] computes preferred binding positions and orientations for a chosen set of molecular probes in the binding region of a given target structure. In previous uses of MCSS for combinatorial ligand design [15, 16], the fragments placed by MCSS were connected to form synthetically accessible candidate ligands by other programs such as HOOK [17] and DLD [18]. An alternative is to use the molecular probe preferences to focus a combinatorial library that is synthesized and tested with an appropriate assay. The inaccuracies of the computational methods are less important for such an application because it is necessary only to predict enhanced interaction probabilities for biasing combinatorial libraries, rather than to design a specific ligand for synthesis. The computational and experimental approaches are complementary in that information from one can be used to refine models derived from the other. MCSS makes it possible to rapidly (in comparison to experiment) explore chemical space with respect to possible library fragments and to propose populations that have a higher probability of containing promising ligands. Combinatorial chemistry and large scale screening then can be used to generate and evaluate experimentally these ensembles of compounds. The design strategy proposed here is summarized in Fig. 1.

Other fragment-based computational ligand design methods have been described and applied with promising results. These include the GrowMol method which builds ligands complementary to a protein binding site de novo from a manually selected “root” or seed point by

combinatorially adding and evaluating atoms and fragments [19, 20]. The CombiSMoG method generates combinatorial libraries in silico and evaluates them using a knowledge-based potential [21]. Another recently reported method is SkelGen [22] which grows compounds in a target active site de novo. It is difficult to compare the performance of these and other methods because they have been applied to very different biological targets. It should be noted that the targets in the GrowMol, CombiSMoG and SkelGen publications contain more compact binding sites in comparison to the large, open, and relatively promiscuous MHC class II binding site targeted here. MCSS has been applied previously to guide the design of much smaller libraries targeted to picornavirus [23]. Other groups have shown that pharmacophore approaches incorporating active site interaction information derived from protein structures improve combinatorial library design [24–28]. Indeed, MCSS molecular probe maps can be used to derive active site pharmacophore models [29]. Somewhat analogous experimental molecular probe mapping and fragment-based ligand discovery approaches have shown success [30–36].

This paper describes improvements to the MCSS method and its use in conjunction with combinatorial synthesis and screening to identify potential ligands for HLA-DR4, an MHC Class II peptide binding protein. The target is of interest because it is involved in modulating the humoral immune response. Aberrant recognition of self peptide-MHC complexes as foreign can lead to autoimmune pathologies; specifically, individuals possessing the HLA-DR4 allotype are observed to have a greater likelihood of developing rheumatoid arthritis [37, 38]. In addition HLA-DR4 is linked to autoimmune hepatitis [39].

The MCSS method was used to determine favorable binding positions and orientations with respect to the MHC Class II allotypes HLA-DR3 and HLA-DR4 for 23 small molecule probes representing candidate functionalities for the combinatorial library. The resulting molecular probe maps were examined using clustering, diversity analysis, and computer graphics. The skeleton of the chosen library which includes a novel branched architecture was suggested by the molecular probe map for propane. The propane map showed how to link several regions in the binding site which bound proportionately larger numbers of functional groups. Most of these regions corresponded to known peptide side chain binding pockets, but there was also one pocket that had not been observed previously to be involved in ligand binding. Once the overall architecture of the library had been selected, the specific combinatorial library was developed based on the MCSS derived data, information on the peptide binding preferences of HLA-DR4 and the availability of chemical starting materials. Synthesis and screening of the library showed that a number of ligands generated by the combinatorial library



**Fig. 1** Ligand discovery strategy: this paper describes one iteration of the strategy outlined here. Starting with a crystal structure, MCSS and the analysis of MCSS results were used to propose a set of molecules that should contain compounds that would bind HLA-DR4. This proposal was tested by generating the set of compounds using

combinatorial synthesis and screening them *in vitro*. A representative compound selected by the experimental screen was modeled into the binding site to illustrate how it could bind the target protein in the predicted binding mode

bound to HLA-DR4 under stringent assay conditions. Characterization of these molecules identified in the selection assay, followed by building and refining a model structure suggested how the putative ligand could bind in the peptide binding pocket. Experimental studies demonstrated that peptides compete with the ligand for binding to HLA-DR4. This suggests that the ligand is located in the MHC binding site. Crystals suitable for X-ray analysis, however, could not be obtained to verify the results.

## Background

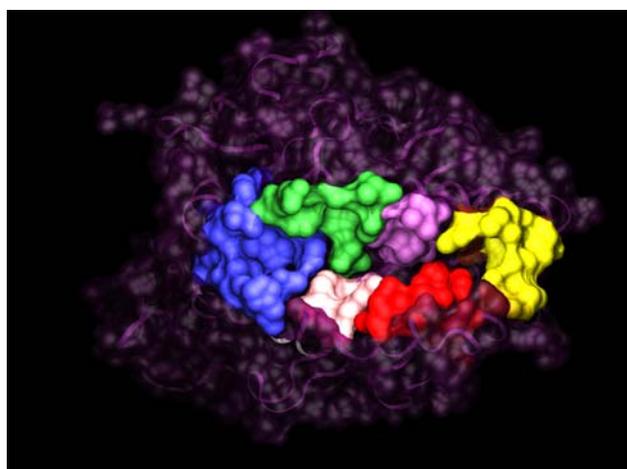
The major histocompatibility complex (MHC) spans approximately  $4 \times 10^6$  base pairs and contains at least 50 genes [40]. It contains a polymorphic set of genes coding for cell-surface glycoproteins (MHC Class I and Class II) responsible for presenting peptide fragments to T cells as well as genes for transporters associated with antigen processing (TAP1 and TAP2), the 70kD heat shock protein (HSP-70), and other proteins involved in immunity [41, 42]. In the case of class I MHC proteins, found on the surface of all nucleated cells, the peptides originate from endogenously expressed proteins and provide a sample of the cell interior to CD8+ cytotoxic T lymphocytes (CTLs); cells recognized as presenting non-self proteins (e.g., viral proteins) are killed by the CTLs. Class II MHC proteins are present on the surface of antigen presenting cells (APCs) such as B cells and macrophages and bind antigenic peptide fragments from endocytosed proteins. Recognition by CD4+ helper T cells of non-self peptides presented by class

II MHCs precipitates a humoral immune response [40]. Aberrant recognition of self peptides can lead to autoimmune disorders [39, 43–49].

The MHC class I and II proteins are encoded by highly variable loci of a limited number of genes in the MHC. Each varying gene is termed an allele, of which there may be more than 100 for some the loci. Different alleles give rise to variable immunocompatibility between individuals and MHC mismatches cause graft rejection [50]. Genetic permutations in the MHC produce a set of polymorphic proteins with different peptide sequence binding preferences. Because each individual has a small number of MHC proteins (six class I and eight class II) in comparison to the number of peptide sequences to be presented ( $10^7$ – $10^8$ ), each protein must be able to bind a large variety of peptides to give sufficient immuno-surveillance [51, 52]. This promiscuous peptide binding implies that many different peptide ligands bind a given MHC protein with a narrow range of affinities [51].

The class I and class II MHC proteins are not primary sequence homologous and have different domain organization, yet they exhibit remarkable structural similarity [51, 53]. Both proteins are membrane bound with a characteristic long binding groove that accommodates a peptide. In both class I and class II proteins, the peptide binds in an extended conformation. The binding groove is formed by two antiparallel  $\alpha$ -helices atop a platform comprising an eight strand  $\beta$ -sheet. In class I proteins, the  $\alpha_1$  and  $\alpha_2$  subunits of the protein comprise the binding site; in contrast, the binding site of class II proteins is made up entirely of the  $\alpha_1$  subunit [53]. The binding site of class I proteins is

closed on the ends with several residues forming a network of hydrogen bonds with the termini of the bound peptides; with the ends of the peptide held firmly in place, small variability in peptide length (total length 8–10 amino acid residues) is accommodated by a characteristic kink in the bound peptide [54, 55]. Class II proteins have a binding site that is open on the ends and therefore accommodates longer peptides (up to at least 13 amino acid residues), which can extend beyond the ends of the groove [53]. The binding sites of both class I and class II proteins have pockets that are specific for certain amino acid side chains and other positions that allow sequence variation [56–59]. These pockets are illustrated in Fig. 2. Bound peptides also project a number of side chains away from the MHC protein into the extracellular environment. The protruding side chains, together with the upper surface of the MHC receptor, serve as recognition elements for T cells [53, 63–66]. In both classes of proteins, extensive hydrogen bonding between the backbone of the bound peptide and the binding site residues helps to stabilize the protein in its folded state [67]. Newly synthesized class II MHC proteins in the endoplasmic reticulum (ER) are stabilized by binding the invariant chain (Ii), another MHC encoded protein [60, 68]. This also prevents class II proteins from binding intracellularly derived peptide fragments as well as the cell's own nascent, unfolded polypeptides in the ER [40]. Proteolysis of Ii in the endosome yields a peptide core called CLIP bound to the MHC [60]. Prior to the class II MHC translocating to the cell surface CLIP is exchanged for a peptide from an exogenous, endocytosed protein. In



**Fig. 2** Solvent accessible surface representation of HLA-DR3: the peptide binding pockets are highlighted as solid surfaces of various colors: *blue*: in contact with peptide position 1 (P1), *green*: contacts P3, *pink*: contacts P4, *purple*: contacts P6, *red*: contacts P7, and *yellow*: contacts P9 (same numbering as in reference [60]). The remainder of the protein is shown as a ribbon covered by the transparent purple surface to give the structural context for the binding site. Figures produced using VMD [61] and POV-Ray [62]

human cells, this exchange is catalyzed by the MHC-encoded protein HLA-DM [60, 69, 70].

The MHC proteins, also called human leukocyte antigens (HLAs) in humans, are named by the allotype from which they arise. The most studied disease association is that of HLA-DR1 with type 1 diabetes [47]. Further, HLA-DR3 is associated with susceptibility to the autoimmune disease myasthenia gravis [71]. Similarly, HLA-DR4 is associated with autoimmune hepatitis [39] and HLA-DR2 with multiple sclerosis [38, 49]. The key role of MHC proteins in modulating immune system activity make them attractive targets for therapeutic agents of autoimmune and inflammatory disorders [72–75].

Other groups have approached the determination of ligands for MHC class II proteins, specifically HLA-DR1, through traditional methods, including peptidomimetic design and structure-based modifications of known ligands. Smith and co-workers substituted a peptidomimetic pyrrolinone for a four amino-acid segment in the known thirteen amino acid peptide ligand (virus hemagglutinin (HA) peptide) and reported favorable binding ( $IC_{50} = 137$  nM in comparison to  $IC_{50} = 89$  nM for the native ligand). Further attempts to modify the four residue substitution based on molecular modeling results yielded poorer affinity ligands ( $IC_{50} > 100$   $\mu$ M) [76, 77]. Woulfe, et al. reported a seven amino acid long peptidomimetic inhibitor that competes favorably with the HA peptide ( $IC_{50} = 50$  nM compared to  $IC_{50} = 62$  nM for HA as reported in their work) [78]. Another group found heptapeptide peptidomimetics that they designed based on sequence preferences discovered through phage display and studies of random heptapeptides. Selections of mimetic functionalities were based on molecular modeling. They discovered compounds with  $IC_{50}$ s as low as 50 nM [72]. The same group reported designing, based on phage display and X-ray crystallographic data, peptide and peptide-mimetic HLA-DR ligands with significant binding affinity and robust activity inhibiting T-cell activation in vitro [79]. In contrast to these studies, the approach presented in this paper uses computational de novo design tools in conjunction with chemical and computational diversity methods to discover novel ligands for HLA-DR4. Other groups have reported designing ligands for MHC class I proteins [80, 81].

As previously mentioned, an increasingly popular way to identify starting points for drug discovery is combinatorial synthesis and high throughput screening. However, because of the relatively non-specific peptide binding of the MHC class I and II proteins described above, discriminating hits from a naive (non-focused) combinatorial library is challenging. It has proven difficult to use a standard strategy based on a diverse library that covers chemical space to first identify the region that should yield a specific high affinity ligand because the protein is likely

to bind a large number of compounds with low affinity. Consequently, the exploratory library is likely to yield an uninformatively large number of hits which do not limit sufficiently the chemical space that needs to be searched. The procedure presented in this study addresses this difficulty. First, an active site map is generated via a computational virtual screen of the binding site using the MCSS method. Clustering, diversity methods, and visualization are employed to augment chemical intuition in analyzing the MCSS results. The information derived from this analysis is used to determine a set of functionalities that when linked to a common molecular framework or backbone architecture generate a collection of compounds (i.e., a combinatorial library) that samples a portion of ligand space that is thought to be more likely to contain compounds that have affinity and specificity for the target. This focused library is then synthesized combinatorially and evaluated experimentally. Even though the computational method does not treat protein flexibility explicitly, the complete approach presented here probes binding site plasticity experimentally.

## Materials and methods

### Preparation of protein coordinates

Coordinates of the MHC Class II proteins HLA-DR3[60] and HLA-DR4[82] (accession codes 1A6A and 2SEB, respectively) were supplied by D. Wiley (personal communication) in Protein Data Bank (PDB) format. The co-crystallized peptide ligands were removed from the structures and the coordinates were translated into CHARMM format coordinate files. Polar hydrogens were added using the HBUILD [83] facility in the CHARMM program [84, 85]. To facilitate specifying a conservative box shaped binding region (for more efficient MCSS calculations) the coordinates were transformed so that the major axis of the binding region coincides with the Cartesian *z*-axis. The structures were not minimized prior to MCSS molecular probe mapping. Later minimization of the HLA-DR4 structure for 300 steps of steepest descents showed a minimal (0.25 Å root mean square) difference in all-atom coordinates. Selected maps were rerun and compared and they exhibited no significant differences (unpublished results).

Molecular probe mapping using the method described below was first performed for the HLA-DR3 structure because HLA-DR4 coordinates were not available. HLA-DR4 was chosen as the experimental library target because it was available in sufficient quantity for use in screening and assaying the library, as well as for its pharmaceutical interest. Coordinates for HLA-DR4 became available

before the library design was finalized and the molecular probe maps were recalculated. HLA-DR3 and HLA-DR4 are nearly sequence identical in the region for which the ligand was targeted and so very few differences between the molecular probe maps for the two allotypes were expected or found. Structural alignment reveals that the two structures are similar with root mean square difference in backbone atom coordinates of less than 0.6 Å.

### The enhanced MCSS method

The MCSS method efficiently finds favorable positions and orientations for molecular probes in the binding site of macromolecules. The original MCSS algorithm [13] was enhanced and fully reimplemented in the Expect scripting language [86] which is an extension of the Tcl scripting language [87]; a version implemented in the C programming language is now available (see below). Version 2.1 of MCSS [14] was developed and applied for generating molecular probe maps for HLA-DR3 and HLA-DR4.

In MCSS, a large number of copies of individual molecular probes (replicas) are simultaneously energy minimized in the forcefield of a macromolecule according to the Time Dependent Hartree (TDH) approximation [88] as implemented in the REPLICA facility in CHARMM. In the TDH approximation, the system is divided into two parts *i* and *j*, atoms belonging to molecules in *i* are influenced by the average field of all replicas in *j*, and vice versa. Replicas in the same partition do not interact. The standard application of the MCSS method is a special case of the TDH approximation in that one partition contains the rigid target macromolecule and the second is comprised of freely moving replicas. Under these conditions, the field influencing the replicas (i.e., that of the target molecule) is constant and explicitly determined; the replicas do not see each other and the target molecule does not move.

The new MCSS script was developed using the Expect language which was designed for automating interactive procedures. With Expect's ability to parse output from CHARMM and to access Tcl language constructs it was possible to rapidly prototype and implement the flexibility and ease of use desired in MCSS. The novel features of MCSS version 2.1 include: (i) incorporation of a standard version of CHARMM; (ii) simplified creation of new molecular probes; (iii) faster distance-based generation of the initial distribution of molecular probe replicas; (iv) ability to generate molecular probe maps in box shaped regions; (v) user specified minimization protocols, constraints, replication specifications, non-bond conditions, and target macromolecule structure generation; (vi) non-covalently linked molecular probes (e.g., hydrated *N*-methylacetamide); and (vii) hybrid (i.e., all hydrogen protein and polar hydrogen molecular probe) representation.

The simplified procedure for creating molecular probes was particularly useful for the present project because it allows convenient incorporation of chemical functionalities for specific ligand design studies. The new customizability allows the user to modify MCSS without changing the core source code by altering templates that the MCSS script reads and uses to generate input to the CHARMM program. Because the Expect language provides access to Tcl constructs the user is able to use values read from the input (e.g., the nonbond cutoff) in their input templates; this makes it easier to keep modifications to the default protocols coherent with the rest of the MCSS script. In addition, because Tcl is a well documented, rich, and accepted scripting language, the MCSS protocol may be used as part of a larger script and, through the use of the Tk extensions to Tcl, may be integrated to a cross platform graphical user interface. Recently, the Expect script was translated to a C language program by Accelrys (formerly Molecular Simulations, Inc.) and the code was back-ported and is distributed as a version 2.5 of MCSS; all features described for version 2.1 are available in version 2.5.

Generating molecular probe maps in box shaped regions allows the binding site of some target molecules to be mapped more efficiently than by using spherical regions. This is because spherical mapping regions often include large volumes that are not of interest or not in contact with the target. Since the binding region of MHC proteins is long and narrow, for example, a sphere sufficiently large to encompass the length of the binding site would include a large amount of empty space above the binding groove.

The program RMOL generates initial placements of molecular probes as follows. First the user specified binding region is discretized into a grid. A random gridpoint is chosen and a replica is placed with random position and orientation with its center of mass within half a grid space of the selected gridpoint. The placement is evaluated by checking whether any atom of the replica is within a user specified distance of any atom of the target macromolecule. To accomplish this check quickly, a list of the ten closest target molecule atoms to each gridpoint is computed and stored. To check the distance from the closest target molecule atom to each replica atom, the pair list for the gridpoint closest to the replica atom is looked up and the distance is checked for only the ten atoms in this pair list. In the previous version of MCSS, generating the initial placements and orientations of the molecular probes was time consuming because each candidate placement was evaluated using a pseudo-energy function which necessitated examining a larger number of atoms in the target than the simple distance-based check requires.

Batches of initially placed molecular probe replicas are energy minimized using the CHARMM program. The REPLICAS facility in CHARMM is used to generate the

copies of the molecular probe in the internal CHARMM protein structure file and to normalize the forces between the protein and replicas. In this study the default minimization protocol for MCSS was used. This involves an initial 800 steps of steepest descents minimization to reduce pathological bad contacts followed by up to 10000 steps of conjugate gradient minimization with culling (gathering) of high energy replicas (the energy assigned to each replica is the sum of its internal energy and its interaction with the target) and duplicate replicas (i.e., replicas that converged to the same position and orientation) every 500 steps. The script decreases geometrically the energy criterion for removing high energy minima after each gather step by computing the average of the current energy criterion and the final energy criterion. Initial and final energy cutoffs of 500 kcal/mol and 3 kcal/mol, respectively, were used. The script monitors the minimization and does no further minimization of the current batch if it has converged (gradient less than 0.01 kcal/Å) or there are no remaining novel replicas (i.e., all replicas have converged to local minima that were found in previous minimization batches). This monitoring allows requesting a large number of minimization steps to ensure that the replicas will find local minima while maintaining high efficiency since no excess minimization is performed. Minimization protocols and convergence criteria can be readily adjusted by editing configuration files that are read by the MCSS script—no modification of the script (v2.1) or source code (v2.5) is required.

Replicas are gathered using a root mean square difference (RMSD) between atoms criterion. Pairs of replicas are deemed redundant if they satisfy two criteria: first, their all atom RMSD is less than a specified cutoff (typically 0.2 Å); second, the RMSD decreases during minimization (otherwise, they are considered diverging and are not clustered). The algorithm is optimized by initially checking the distance between the centers-of-mass of replica pairs and only computing the all atom RMSD and checking for divergence if the centers-of-mass distance is less than the cutoff. If a set of active replicas are converging to a novel local minimum, one replica is kept as a representative and the duplicates are removed. All replicas converging to previously found local minima are removed.

The minimize-gather cycles are repeated until all batches have been minimized. The results are stored in a CHARMM coordinate file that contains the coordinates and interaction energies of the functional group minima. The interaction energies for each replica are calculated using the BLOCK facility in CHARMM [89]. The interaction energy for a replica is the sum of internal energy terms and Coulombic and van der Waals interactions between the replica and the target macromolecule. Energies are normalized by subtracting the ground state in vacuo energy of

the molecular probe. In effect, this includes the internal strain caused by the replica adopting a conformation to fit the binding region in the MCSS reported interaction energy. This more accurately rank orders replicas of a given type by accounting for internal strain each replica experiences to form favorable interactions with the protein. The normalization also, to a first approximation, puts sets of minima calculated for different groups (or for the same group using different representations) on the same energy scale so that they can be compared.

By changing the default constraint specification, MCSSv2.1 can be used to study systems with flexible target macromolecules and systems in which parts of the target macromolecule are replicated. In a specific study of the HLA system, five residues of the protein were allowed to move while computing a map to reproduce known binding preferences for the aspartate side chain. In addition, alternate protocols for quenching the molecular probe replica in the binding site such as Monte Carlo annealing (R. Putzer and D. Joseph-McCarthy, personal communication) or molecular dynamics [90] may be incorporated with small modifications to the script and configuration files. These features are under development. Monte Carlo annealing of molecular probe replica into binding regions can be useful for placing larger probes or ligands since it should allow for more conformational freedom or greater sampling of internal degrees of freedom of the probe. Dynamics has not been used in production runs with a flexible or partially flexible target because the behavior of systems with large numbers of replicas in one partition and a relatively small number in the other exhibit difficult dynamics properties such as unequal distributions of kinetic energy between the partitions. In more focused studies where the ratio of replicas in each partition is close to one, molecular dynamics protocols can be useful for probing phenomena such as induced fit in ligand association [90]. When using MCSS results to bias or focus combinatorial libraries, it is not strictly necessary to treat binding site flexibility because it possible to include elements of varying size and flexibility in the library to experimentally probe protein dynamics. Coupling MCSS with pharmacophore methods may prove to be an efficient way to implement such loose constraints in library design.

### Functionality mapping

The binding positions of 23 small molecule probes representing functionalities chosen because of their availability for inclusion in the combinatorial library were determined for HLA-DR3 and HLA-DR4 using the MCSS algorithm. The mapped binding region encompassed the entire binding groove of each protein and measured  $15 \text{ \AA} \times 36 \text{ \AA} \times 37 \text{ \AA}$  (19980  $\text{\AA}^3$ ). For each protein 10000 initial

placements were made with an initial distribution distance criterion of 1.2  $\text{\AA}$  (see above). Batches of 500 to 1000 replicas were simultaneously minimized using the default MCSS protocol with a shifted potential, 7.5  $\text{\AA}$  nonbond cutoff, constant dielectric ( $\epsilon = 1.0$ ), and group based nonbond list generation; nonbond list updates were performed periodically according to the heuristic in CHARMM (INBFRq = -1). The replicas were gathered such that pairs with less than 0.2  $\text{\AA}$  RMSD in their coordinates were considered identical. The molecular probes used for HLA-DR3 and HLA-DR4 in the polar hydrogen representation are shown in Table 1.

### Clustering molecular probe maps

The molecular probe maps were clustered spatially to facilitate analysis of the results. Clustering was performed using the ART-2' [91] algorithm as implemented in CHARMM. To compare differing molecular probes for the purpose of assigning them to clusters, the center-of-mass for each molecular probe replica minimized position was computed. The centers-of-mass of all minima for all groups were combined into a single dataset and clustered simultaneously. This was done to produce a map of spatially distinct areas of favorable interaction with the protein, the clusters were later analyzed with respect to the number and the chemical and spatial diversity of the minima contained within them. The goal of clustering the minima in this way was to derive a semi-quantitative picture of where functionalities should be placed with respect to the protein.

The ART-2' algorithm iteratively classifies elements into clusters. In each iteration, after all molecular probe minima have been assigned to a cluster (as described below), the centroid of each cluster is recalculated as the average of the centers-of-mass of the minima belonging to that cluster. Assigning the minima to clusters (binning) and computing the centroids is repeated until there is no residual error, as defined as any center-of-mass lying further than a given radius from the centroid, and there is no change in cluster membership upon updating the centroid positions. Binning is performed using a neural network to classify each minimum as belonging to one cluster. The result of clustering is a list of molecular probe replica minima that belong to each cluster as well as the coordinates of the center of the cluster, the radius of the cluster, and the number of minima contained in the cluster. Minima were not filtered energetically; that is, all minima computed using the MCSS method (with final energy cutoff of 3 kcal/mol) were used in computing the cluster positions and were assigned to a cluster. A cluster radius of 2.5  $\text{\AA}$  was chosen empirically to provide a balance between reducing data to a reasonable form and retaining sufficient information. Clusters containing relatively small numbers

**Table 1** Functional group probes for which maps were computed for HLA-DR3 and HLA-DR4. The “side chain” groups represent the amino acid R-groups with  $C_{\beta}$  treated as a terminal  $CH_3$  group

functional group name	description
2BTN	2-butene
ACAM	<i>N</i> -methyl acetamide
ASPR	aspartate side chain, acetate ion
ARGR	arginine side chain, propylguanidinium
ASNR	asparagine side chain, acetamide
BENZ	benzene, computed in polar and all hydrogen representations
CHEX	cyclohexane
CYSR	cysteine side chain, methanethiol
ETHR	ether
ILER	isoleucine side chain, butane
LEUR	leucine side chain, isobutane
LYSR	lysine side chain, 1-amino pentane ion
MAMM	methylammonium
MEOH	methanol
NITR	acetonitrile
PHEN	phenol
PHER	phenylalanine side chain, toluene
PPTD	peptoid, $CH_3CON(CH_3)_2$
PRPN	propane
TFME	trifluoromethane
THRR	threonine side chain, ethanol
VNLG	vinyllogous peptide backbone unit, $CH_3CONH-CH_2CHCH_2$
WATR	water

of minima (less than 50 out of approximately 7000–10000 total minima) were discarded to facilitate the visual analysis and to improve the statistical analysis of the clusters. The discarded low occupancy clusters generally were located outside of the binding region.

#### Analysis of the clusters

*Visual analysis.* The clusters were analyzed visually in the context of the proposed framework for the combinatorial library and with respect to the protein structure. Clusters were visualized in VMD [61] as spheres colored by occupancy (i.e., the number of minima contained in a cluster) and centered at the centroid of each cluster. Clusters corresponding to elements of the combinatorial library were selected by superimposing the cluster spheres with the propane molecular probe map and manually picking the representatives.

*Numerical analysis.* The clusters were also analyzed quantitatively in terms of the proportion of each functionality present and the corresponding energies of these

replicas. After clustering, the molecular probe minima were filtered based on several energy criteria described below. For each molecular probe type in a given cluster, the ratio of the number of minima of that type in the cluster to the total number of minima of that type remaining after energy filtering were used for comparing the relative contribution of each group to the population of each cluster. Ratios were compared to normalize the numbers of groups with respect to the potential energy surface for each group. In other words absolute numbers of minima should not be compared because the character of the energy surface (and accordingly, the number of local minima) varies for each group. If the energy surface for a particular functionality is relatively flat more local minima are found in comparison to a group for which the energy surface is more sharply defined.

Five energy cutoff schemes were examined to determine which yielded a reasonable balance between reducing the number of minima to be examined and retaining sufficient numbers of minima for use in focussing the combinatorial library. For each of the cutoff schemes, minima with MCSS interaction energies greater than the cutoff energy associated with the group for that scheme were removed from the data set. The first cutoff considered was one half the solvation enthalpy of the group which is intended to indicate the relative importance of the interaction of the group with solvent in comparison to the binding site [13]. A similar cutoff scheme is to use the free energy of solvation for nonionic probes and one half the solvation enthalpy for charged groups [92]. Caffisch introduced using the electrostatic free energy of solvation of the molecular probe as computed using the Linearized Poisson–Boltzmann equation as the energy criterion [16]. This may not be an effective filter for nonpolar probes, however, because these have computed electrostatic solvation free energies of zero. The fourth and fifth cutoff schemes considered, the average energy and one standard deviation less than the average energy for each molecular probe, are intended to retain minima that are favorable on the energy scale determined by the MCSS results for each molecular probe; that is, to keep the minima with energy better or significantly better than others of that type. The energy values used in each cutoff method are summarized in Table 2 and the numbers of minima retained by the various cutoff are listed in Table 3.

In designing a combinatorial library it is useful to have an indication of which parts of the molecule should be populated with diverse substituents and which with relatively similar substituents. This information can be derived from MCSS results by computing a score that indicates the spatial and chemical similarity of the minima belonging to the cluster or clusters corresponding to a library position. This is a somewhat different problem than what is usually referred to as similarity scoring because in this case the

spatial similarity/diversity is information uniquely available from MCSS functionality maps. It is important to note the MCSS maps provide a collection of likely positions for molecular probes in the binding site of the target. This is a key distinction of the application of chemical diversity in this context: MCSS performs a combinatorial docking; that is, it provides a set of molecular probe positions that could be assembled to create ligand molecules rather than providing information on the placement of complete ligands. Therefore, the distinct positions of the molecular probes are a key element of information to be used in designing combinatorial libraries. Any two-dimensional similarity estimate and most three-dimensional similarity scores lose the spatial information provided by MCSS. This follows from the fact that typically these methods overlap the molecules for maximum similarity prior to scoring them. For whole-molecule comparisons this is appropriate because of the expectation that the molecules would be

oriented upon binding to place similar functionalities or pharmacophores in similar positions with respect to the binding site; in other words, molecules should be aligned as closely as possible prior to comparing them [101]. By contrast, when considering molecular probes or fragments of molecules it is important to retain the alternate binding modes represented by each unique molecular probe replica position. There are two reasons for this. First, the unique positions provide a basis for building up molecules and second, aligning the molecular probe replicas essentially destroys the information content available from the three-dimensional structure of the target. The molecular probes have more orientational and positional freedom alone than they would as part of a larger molecule so the precise positions of the probe atoms are likely not to reflect their positions in a bound ligand. Nonetheless the probes' orientations and positions encode valuable information about the structure and composition of the active site and thus

**Table 2** Energetic cutoff criteria for filtering various functional group probes. All quantities are in units of  $kcal \cdot mol^{-1}$

Functional group	$\Delta H_{solv}/2$	Experimental $\Delta G_{solv}$	Electrostatic $\Delta G_{solv}$	Average MCSS energy (DR4)	$\sigma$ MCSS energy (DR4)	Max/min MCSS energy (DR4)	Average MCSS energy (DR3)	$\sigma$ MCSS energy (DR3)	Max/min MCSS energy (DR3)	Energy of isolated group
2btn	-2.98 <sup>i</sup>	1.4 <sup>a</sup>	-0.25	-5.51	2.67	1.54/-12.96	-5.80	2.51	-0.01/-14.82	0.97
acam	-8.80 <sup>h</sup>	-10.1 <sup>b</sup>	-9.01	-15.07	9.62	-0.24/-48.46	-18.47	10.40	0.48/-62.67	-3.49
acet	-46.5 <sup>f</sup>	-80.7 <sup>b</sup>	-71.1	-18.92	27.43	2.57/-114.53	-19.92	27.00	2.07/-129.48	0.00216
argr		-75.0 <sup>c</sup>	-82.2	-41.30	29.75	-3.04/-137.85	-45.30	27.69	-0.04/-116.87	17.10
asnr	-8.45 <sup>h</sup>	-9.72 <sup>b</sup>	-10.8	-15.90	11.16	2.93/-49.82	-17.34	9.77	-0.02/-63.32	-12.44
aspr	-46.5 <sup>f</sup>	-80.7 <sup>b</sup>	-71.1	-16.32	23.28	2.57/-99.36	-26.24	28.72	0.62/-129.48	0.00216
benz	-3.80 <sup>d</sup>	-0.87 <sup>b</sup>	0.00	-5.88	1.51	-3.39/-10.79	-6.68	2.81	-3.11/-11.49	0.08943
benzh	-3.80 <sup>d</sup>	-0.87 <sup>b</sup>	-1.95	-15.01	3.89	-7.01/-27.37	-6.99	4.09	2.34/-23.35	12.19348
chex	-3.97 <sup>d</sup>	1.24 <sup>a</sup>	0.00	-3.08	2.95	2.71/-10.16	-6.67	2.81	1.82/-17.07	0.01137
cysr	-1.55 <sup>g</sup>	-1.24 <sup>b</sup>	-1.04	-6.03	4.03	2.84/-16.50	-6.28	3.37	3.24/-20.49	0.00
ethr	-3.83 <sup>g</sup>	-1.90 <sup>g</sup>	-1.77	-8.58	5.22	2.58/-24.35	-10.23	5.37	1.48/-32.12	0.00
iler	-3.10 <sup>e</sup>	2.18 <sup>a</sup>	0.00	-4.91	1.51	1.62/-8.79	-5.48	1.64	-0.68/-10.51	-0.0938
leur	-2.69 <sup>d</sup>	2.32 <sup>d</sup>	0.00	-4.84	1.45	-0.96/-8.61	-4.85	1.89	1.61/-9.63	0.04536
lysr		-62.4 <sup>b</sup>	-96.4	-48.61	34.31	-0.27/-165.94	-55.87	28.64	-0.55/-138.77	-0.47921
mamm	-37.5 <sup>f</sup>	-81.5 <sup>b</sup>	-98.9	-41.94	35.06	2.19/-162.96	-44.66	28.42	2.88/-133.86	0.00006
meoh	-5.32 <sup>d</sup>	-5.11 <sup>d</sup>	-7.17	-9.34	6.30	2.33/-30.63	-12.87	7.43	0.94/-42.98	0.00000
nitr	-4.15 <sup>d</sup>	-3.94 <sup>a</sup>	-3.78	-11.28	7.25	2.99/-32.82	-12.27	6.68	1.22/-38.97	0.00000
phen	-6.80 <sup>d</sup>	-6.62 <sup>b</sup>	-6.61	-11.80	7.50	1.10/-38.34	-14.36	7.68	-0.09/-48.23	-0.39347
pher	-4.34 <sup>d</sup>	-0.90 <sup>a</sup>	0.00	-6.77	1.79	-2.05/-12.57	-7.01	2.08	0.25/-15.14	-0.15422
pptd	-8.55 <sup>h</sup>	-8.23 <sup>d</sup>	-5.68	-13.06	8.83	0.09/-41.99	-12.34	9.04	-0.62/-60.46	-1.29105
prpn	-2.69 <sup>d</sup>	2.02 <sup>a</sup>	0.00	-4.32	1.30	0.06/-7.39	-4.57	1.69	5.69/-9.22	0.1799
tfme		0.82 <sup>a</sup>	-2.16	-9.90	5.23	1.64/-32.06	-10.82	5.62	2.16/-29.92	0.18318
thrr	-6.26 <sup>d</sup>	-5.01 <sup>d</sup>	-6.63	-10.93	6.68	0.00/-34.10	-13.47	7.47	3.46/-44.44	-0.06996
vnlg	-10.6 <sup>d</sup>		-8.80	-20.45	9.54	-7.09/-52.02	-21.19	12.39	-0.39/-65.66	-2.8664
watr	-4.99 <sup>g</sup>	-6.32 <sup>g</sup>	-13.3	-11.32	8.45	1.01/-38.05	-13.24	9.43	4.43/-46.05	0.00000

<sup>a</sup> From Reference [93]. <sup>b</sup> From Reference [94]. <sup>c</sup> From Reference [92]. <sup>d</sup> From Reference [95]. <sup>d</sup> Calculated according to Reference [95]. <sup>e</sup> From Reference [96]. <sup>f</sup> From Reference [97]. <sup>g</sup> From Reference [98]. <sup>h</sup> From Reference [99]. <sup>i</sup> From Reference [100]

**Table 3** Number of minima retained under various energy cutoff criteria

Functional group	Total number of minima DR3/DR4	Number of minima retained by cutoff						
		$\Delta H_{\text{soln}}/2$ DR3/DR4	Experimental $\Delta G_{\text{soln}}$ DR3/DR4	Electrostatic $\Delta G_{\text{soln}}$ DR3/DR4	Average MCSS energy (DR4)	Average- $\sigma$ MCSS energy (DR4)	Average MCSS energy (DR3)	Average- $\sigma$ MCSS energy (DR3)
2btn	490/270	452/231	490/269	486/263	130	44	219	59
acam	361/251	314/193	292/168	311/190	100	38	163	46
acet	287/197	49/30	12/15	19/19	50	30	90	49
argr	504/835	278/386	93/86	55/62	357	132	210	101
asnr	552/627	462/457	436/422	415/394	271	113	249	89
aspr	142/125	36/14	7/6	11/7	35	17	47	27
benz	38/84	34/78	38/84	38/84	39	11	16	7
benzh	188/124	158/124	168/124	163/124	57	22	85	25
chex	602/250	547/114	601/231	601/203	134	42	188	113
cysr	153/182	143/155	144/163	144/168	85	37	72	23
ethr	213/110	206/93	213/101	213/102	51	17	87	41
iler	296/356	284/322	296/356	296/355	165	58	106	46
leur	203/199	191/194	203/199	198/199	84	33	80	30
lysr	670/498	462/234	258/93	42/41	194	79	302	121
mamm	104/96	60/40	16/13	4/9	30	13	50	19
meoh	334/237	333/171	334/171	310/141	99	28	174	68
nitr	176/197	162/163	166/168	167/169	73	36	73	33
phen	682/700	568/485	573/490	573/490	311	107	309	95
pher	534/613	497/570	532/613	532/613	291	100	206	64
pptd	264/663	174/446	179/454	233/527	242	90	107	25
prpn	205/224	193/211	203/224	203/223	108	34	78	33
tfme	318/163	316/161	318/162	294/150	83	18	151	52
thrr	657/513	558/400	579/428	552/390	209	66	290	113
vnlg	822/910	703/788	na/na	755/849	383	182	353	73
watr	208/209	164/154	158/144	88/72	87	33	89	35

should be preserved and leveraged where possible in analysis.

That the molecular probe minima are not aligned prior to computing the dissimilarity in a given cluster makes the calculation fast and simple to implement. The form chosen for the dissimilarity measure is derived from the work of Chapman in which the dissimilarity score is based on interatomic distances between the closest atoms and between the atoms of the closest polarity in the two molecules being compared [102]. The function was developed originally for use in aligning and comparing whole molecules, for example in evaluating a library of compounds. In Chapman's work, polarity is determined by rules; in the present application, the partial charge of the atoms is used as an indicator of polarity; that is, the distance between atoms of closest charge is taken as a measure of electrostatic dissimilarity. Thus, the dissimilarity between two molecular probe placements is expressed as shown in Eq. 1 where  $i$  and  $j$  are the groups in the clusters for which dissimilarity is being computed.

$$\begin{aligned}
 D_{\text{overlap}} &= \sum_{i=0}^{N_{\text{groups}}-1} \sum_{j=i+1}^{N_{\text{groups}}} \sum_{\text{atoms in } i} r_{\text{closest atom in } j} \\
 D_{\text{electrostatic}} &= \sum_{i=0}^{N_{\text{groups}}-1} \sum_{j=i+1}^{N_{\text{groups}}} \sum_{\text{atoms in } i} r_{\text{atom with closest charge in } j} \\
 \text{Dissimilarity} &= D_{\text{overlap}} + D_{\text{electrostatic}}
 \end{aligned}
 \tag{1}$$

Because this sum is expected to scale with the square of the number of groups, it is normalized by dividing by the square of the number of groups for the purpose of comparing the dissimilarity of clusters containing differing numbers of molecular probe minima. Note that this measure of dissimilarity contains information on the differences between binding modes and is applicable to molecules with differing numbers of atoms since it relies only on finding the closest atoms in two molecular probe placements. One other difference from the function described by Chapman is that this function does not employ a

so-called “soft-threshold” function which puts an upper bound on the distance dissimilarity contribution for each atom. In Chapman’s application there was a point beyond which pharmacophores were considered maximally dissimilar and it was counterproductive to penalize compounds that had dissimilarities in distant parts of the molecules being compared; for example, if two pharmacophores overlap well and are similar but the third pharmacophore in each compound occupy different regions, it may not be useful to score these compounds as very dissimilar. The upper bound was not used because it was desired to have a function that would measure the spatial diffuseness of each cluster.

While sufficient studies have not been performed to completely validate the application of the above formulation for cluster diversity, the proposed measure provides, at the very least, a qualitative measure of the heterogeneity of binding site molecular probe preferences. This information helps complete the picture of the overall character of the chemical space that is more likely to yield hits for the targeted binding region, and may be used to assist in designing a combinatorial library to interrogate the binding site. It should be noted that the numerical trends observed for the diversity measure reflect those seen upon visual and manual inspection of the molecular probe maps.

### Monomer selection

Based on the MCSS molecular probe maps and experimental observations that tetrapeptides thought to occupy pockets P1 through P4 of MHC Class II proteins inhibit binding of full length peptides [103], the library was designed to focus on P1 through P4 of the binding groove as well as a region near the break in the  $\alpha_1$ -helix bounding the binding groove (HLA-DR4 residues S53, F54, and E55) and adjacent to P1 and P2. Initial visual inspection of the MCSS molecular probe maps suggested a framework for the library. Monomers were selected for each site using criteria of similarity to MCSS molecular probes observed at that position, chemical intuition, and availability. In selecting the library composition, it was assumed that the sites were relatively independent; in other words, there was no account for possible interaction between monomers. While additivity of the monomer binding energies is a desired outcome, further computational assessment of the binding energetics of the complete ligands was not performed; thus, no additivity was considered. A six monomer branched library was designed. The details of the library design are described below, in Sect. ‘[MCSS maps and design of combinatorial library](#)’. Retrospective quantitative analysis of the maps via clustering, applying energy cut-offs, and computing chemical dissimilarity within clusters in conjunction with knowledge of the sequence binding

preferences of the protein were used to validate the methods and verify the architecture and monomer selection for the combinatorial library.

### Model building and refinement

The consensus ligand selected by experimentally screening the combinatorial library was modeled into the binding site according to the binding mode hypothesized for the library. The chemical structure of the ligand was built in Quanta using the 2-D builder. This structure was transferred to the 3-D builder in Quanta where initial coordinates were assigned and chirality was examined. Upon exiting the 3-D builder a CHARMM residue topology file for the ligand was written out. In the modeling mode of Quanta, key functionalities of the ligand were manually aligned with their corresponding MCSS minima. The coordinates for the ligand were exported from Quanta in the CHARMM coordinate file and Merck file formats. Partial atomic charges were assigned to the ligand atoms using the Merck Molecular Force Field (MMFF) [104] with CHARMM. Missing bond, angle, and dihedral parameters for the ligand were taken from the Quanta 4.1 PARM.PRM file or derived from corresponding parameters for chemically similar atom types.

The initial model was refined to rationalize the internal geometry and to optimize the interactions with the protein using molecular dynamics (MD) simulated annealing and molecular mechanics minimization. The model built structure was first minimized (10000 steps Powell minimizer) in the field of a fixed protein. The C26 and C59 atoms of the ligand were also constrained. This minimized structure was annealed using MD (LEAPFROG integrator in CHARMM) starting at 3000 K and cooling in 25 K decrements every 300 dynamics steps to a final temperature of 300K over 16.35 picoseconds using 0.5 fs timesteps. Following annealing, the structure was minimized to a gradient of 0.0001 kcal/Å using the Powell minimizer. The structures were then reannealed using the same annealing and quenching schedule as above but with two different schemes for constraining the ligand. In the first, no constraints were used to hold the ligand near the protein; in the second, the C26 and C59 atoms were harmonically constrained using a mass weighted force constant of 1.0 kcal/mol/Å<sup>2</sup> and the ligand was allowed to relax at 300 K. In all cases the portion of the protein within 6.5 Å of the ligand was flexible but subjected to a mass weighted harmonic constraint with force constant of 5.0 kcal/mol Å<sup>2</sup>. Both annealing/constraint protocols were carried out twenty-five times with varying random number seeds and the final structures with the best energies were selected. The interaction energy between the protein and the ligand ranged from −73.15 to −133.94 kcal/mol.

## Experimental protocols

### *Split and pool synthesis*

Split and pool combinatorial synthesis was used to generate 243040 potential ligands for HLA-DR4 bound individually to solid support (TentaGel resin). The branch position of the library consisted of diamine monomers which could be chemoselectively deprotected by exposure to orthogonal conditions.

**Library initialization, deprotection, and split.** A linker of Glycine-(6-aminocaproic acid)- $\beta$ -Alanine was coupled onto TentaGel S NH<sub>2</sub> (RAPP, Polymere GmbH, 1.0 g, 0.25 mmol) using an Applied Biosystems Model 431A peptide synthesizer running standard Fmoc-HBTU chemistry. This initialized library was deprotected by twice dissolving the resin in 20% piperidine solution in DMF for 10 min. The library was synthesized using a modified “split and pool” strategy [105], which allows incorporation of “skip codons” or blank positions resulting in sub libraries [106, 107]. After extensive washing in CH<sub>2</sub>Cl<sub>2</sub> (5 mL  $\times$  5), DMF (5 mL  $\times$  5), NMP (5 mL  $\times$  5), the library was split into 31 portions by suspending the resin in 31 mL DMF and transferring 1 mL aliquots to numerically labeled reaction vessels.

**Library tag coupling.** Six binary encoding tags were used to encode the first split of 31 monomers. Solutions of each tag were made by first dissolving the tag in DMF (50  $\mu$ L) and then CH<sub>2</sub>Cl<sub>2</sub> (to a final molarity of 1.6 mM). Solutions of HOBT (100.5 mg, 0.74 mmol, 0.24 M) and 1,3-diisopropylcarbodiimide (DIPC, 117  $\mu$ L, 0.74 mmol, 0.48 M) were made in DMF and CH<sub>2</sub>Cl<sub>2</sub>, respectively. To each of the 31 reaction vessels resulting from the first split, HOBT solution [150 molar equivalent (equiv), 100  $\mu$ L] was added, followed by a unique combination of tag solutions (0.02 equiv, 100  $\mu$ L each). After the addition of CH<sub>2</sub>Cl<sub>2</sub> (500  $\mu$ L), DIPC solution (150 equiv, 50  $\mu$ L) was added, and the vessel mixed carefully, while being vented. The vessels were shaken overnight, before being washed extensively with NMP (5 mL  $\times$  5) and DMF (5 mL  $\times$  5). To verify the integrity of the encoding tag, approximately 20% of the vessels were tested by removing a single bead, washing it 10 times in DMF and 10 times in decane, then photolyzing it for 1–2 h and analyzing the tags via gas chromatography.

**Library monomer coupling.** The vessel encoded with the skip codon was set aside and stored in a desiccator. To each of the remaining vessels, a DMF solution of a specific monomer (500  $\mu$ L, 16 mmol, 2 equiv) was added, followed by treatment with DMF solutions of HATU (100  $\mu$ L, 6.8 mg, 18 mmol, 2.2 equiv) and DIEA (100  $\mu$ L, 2.7 mg, 27 mmol, 3.3 equiv). After shaking for 3 h, the vessels were drained and washed with CH<sub>2</sub>Cl<sub>2</sub> (5 mL  $\times$  3), DMF (5 mL  $\times$  3), and NMP (5 mL  $\times$  3). Coupling reagents

(monomer, HATU, DIEA) were re-added to each vessel, and the vessels were shaken overnight to ensure thorough completion of the coupling reaction. The resin was washed again with CH<sub>2</sub>Cl<sub>2</sub> (5 mL  $\times$  5), DMF (5 mL  $\times$  5), and NMP (5 mL  $\times$  5). With exception of the resin in the skip codon vessel, the resin was pooled by being rinsed with DMF into a large polypropylene peptide synthesis tube. The pooled resin was capped two times with a 20% solution of acetic anhydride (Ac<sub>2</sub>O) in CH<sub>2</sub>Cl<sub>2</sub> for 10 min each. After washing the resin with CH<sub>2</sub>Cl<sub>2</sub> (5 mL  $\times$  5), DMF (5 mL  $\times$  5), and NMP (5 mL  $\times$  5), the resin was deprotected by two treatments with 20% piperidine solution in DMF for 10 min each. The resin was washed extensively with CH<sub>2</sub>Cl<sub>2</sub> (5 mL  $\times$  10), DMF (5 mL  $\times$  10), and NMP (5 mL  $\times$  10), before being recombined with the contents of the skip codon vessel. The library was split again, and the tag and monomer coupling procedure repeated, as described above. At the branch position, the BOC protecting group was removed by treatment two times with a solution of 20% TFA in CH<sub>2</sub>Cl<sub>2</sub> for 30 min. After washing the resin with CH<sub>2</sub>Cl<sub>2</sub> (5 mL  $\times$  5), DMF (5 mL  $\times$  5), and NMP (5 mL  $\times$  5), the resin was neutralized by dissolution in a 20% solution of DIEA in CH<sub>2</sub>Cl<sub>2</sub>. Deprotection of the *t*-butyl ester on the aspartic acid side chain was accomplished by shaking the resin in a solution of 95% TFA in CH<sub>2</sub>Cl<sub>2</sub> for 2 h. After being washed with CH<sub>2</sub>Cl<sub>2</sub> (5 mL  $\times$  5), DMF (5 mL  $\times$  5), NMP (5 mL  $\times$  5), and methanol (5 mL  $\times$  5), the resin was neutralized as above.

### *Screening with antibody based assay*

Ligands with affinity for HLA-DR4 were selected using an on-bead enzyme linked immunoassay (ELISA). A portion of the completed, deprotected library (750 mg, estimated to contain three beads bearing each individual compound) was screened in parallel by the following protocol. The 750 mg sample was placed in a polypropylene peptide synthesis vessel and washed five times with water and the screening buffer which consists of 0.5 M NaCl, 75 mM sodium acetate (pH = 4.5) and 1% glycerol. The resin was suspended in 10 mL screening buffer to which HLA-DR4 solution (42  $\mu$ L, 428.4 pmol, 42 nM) was added. The suspension was held upright in a rotary-mixing incubator at 37 °C for 36 h. Anti-HLA-DR antibody (L-243, 0.6  $\mu$ g, approximately 4 pmol, 2 nM) and Goat anti-Mouse IgG conjugated to alkaline phosphatase (Gibco BRL, 0.3  $\mu$ g, 1.5 pmol, 1 nM) were mixed in a buffer of phosphate buffered saline (pH = 7.4), 0.1% Tween 20, and 1% bovine serum albumin (PBS-T + BSA, 1.5 mL), and left at room temperature for one hour. The vessel containing the library was drained, and the resin was washed with PBS-T+BSA (5 mL  $\times$  4). The antibody solution was then added to the library resin and the suspension was mixed on a rotator for

one hour at room temperature. The vessel containing the library was drained, and the resin was washed with PBS-T + BSA (5 mL  $\times$  3). The resin was treated with substrates for alkaline phosphatase [5-bromo-4-chloro-3-indolyl phosphate, *p*-toluidine salt (BCIP) and nitro-blue tetrazolium (NBT)] in a buffer of 100 mM Tris (pH = 9.5), 100 mM NaCl, and 5 mM MgCl<sub>2</sub>. This would release dye in the vicinity of and stain beads presenting compounds with affinity for HLA-DR4. After 5 to 20 min, the reaction was quenched by draining the vessel and resuspending the resin four times in 6 M guanidinium hydrochloride (pH = 7.0). The beads were placed in a petri dish containing 0.01% Triton X-100 in water. The darkest stained beads (approximately 70) were physically removed with a 50  $\mu$ L gastight syringe, and washed with NMP (3  $\times$ ), methanol (3  $\times$ ), and water (10  $\times$ ). These beads were re-assayed, as described above, in the absence of HLA-DR4, and false positives (i.e., beads presenting ligands that bind antibody rather than HLA, a total of 4) were removed.

The above procedure was repeated with 7.5 nM concentration of HLA-DR4 in the presence of a competing peptide ligand to be more stringent in selecting beads. The sequence of the resulting hits were determined by reading the binary encoding tags. The frequency of occurrence of monomers at each position in the selected compounds were graphed as a histogram and consensus ligands representing the most frequently represented monomers were selected.

**Individual ligand synthesis.** The consensus ligands were synthesized on Rink amide methyl benzhydrylamine (MBHA) resin, using manual solid phase synthesis, and conditions identical to those described above for synthesizing the library. All ligands were cleaved from the resin by application of 95% TFA in water for 2.5 h. Ligands were purified by reverse phase-HPLC, running a gradient of 100:0 to 50:50 0.1% TFA/H<sub>2</sub>O:CH<sub>3</sub>CN over 30 min. Low-resolution mass spectra (LRMS) for the isolated compounds were obtained on a JEOL AX-505 or JEOL SX-102 mass spectrometer. Data were acquired using fast atom bombardment (FAB). Positive ion mass spectrometry required *m*-nitrobenzyl alcohol (NBA) as a matrix with NaI added; while negative ion mass spectrometry relied on a matrix of glycerol. Ligand **6A**: LRMS (FAB, NBA/NaI) calculated for C<sub>45</sub>H<sub>51</sub>F<sub>2</sub>N<sub>7</sub>O<sub>9</sub> + H<sup>+</sup> = 872, found 872. Ligand **6B**: LRMS (FAB, negative ion, glycerol) calculated for C<sub>36</sub>H<sub>47</sub>F<sub>2</sub>N<sub>7</sub>O<sub>9</sub> - H<sup>+</sup> = 758, found 758.

#### Characterization of putative ligand

The following assays were used to verify binding by the selected ligands.

**On-bead ELISA.** Ligands identified during screening were re-synthesized, both as conjugates to Tentagel resin and in un-conjugated form. The resin-conjugated ligands

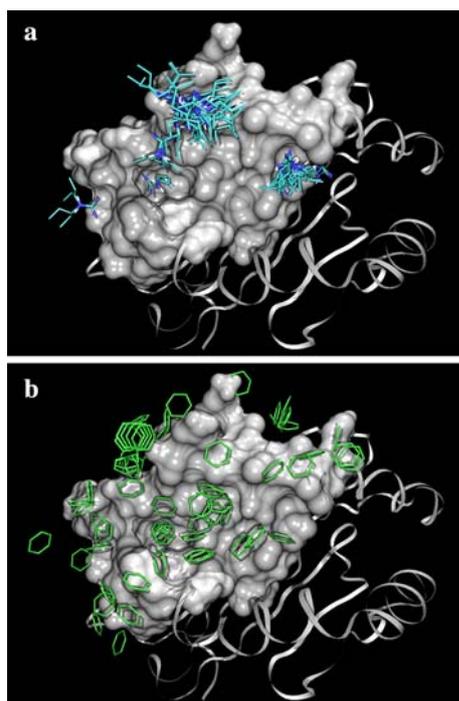
were subjected to the on-bead ELISA described above, and shown to bind HLA-DR4, but not the anti-mouse antibody. This experiment verifies that ligands identified during library screening bind HLA-DR4, when conjugated to a solid support. This on-bead assay was repeated, using HLA-DR4 that was pre-loaded with biotinylated HA peptide. Peptide-loaded HLA-DR4 failed to bind ligand conjugated to resin. This experiment demonstrates that peptide binding and ligand binding to HLA-DR4 are mutually exclusive. This suggests that the selected ligands occupy the peptide binding groove; however, other modes of ligand binding to HLA-DR4, consistent with this data, are also possible.

**Further binding studies.** 6A, unconjugated to resin, was shown to bind HLA-DR4. First, 6A stabilizes empty HLA-DR4 homodimers during native gel electrophoresis (data not shown). Secondly, 6A forms a complex with HLA-DR4 that can be isolated by gel filtration chromatography. Empty or non-ligand-bound HLA-DR4 is unstable to both electrophoresis and chromatography conditions [108]. These experiments demonstrate that 6A binds to HLA-DR4 to form stable complexes, with properties similar to HLA-DR4 peptide complexes.

## Results and discussion

### MCSS maps and design of combinatorial library

MCSS molecular probe active site maps were computed for the entire binding cleft of the HLA-DR3 and HLA-DR4 proteins (see Methods). The total numbers and ranges of energies of the minima found for the various molecular probes are summarized in Tables 2 and 3; representative molecular probe maps are depicted in Fig. 3. Because the library is focused on the region roughly corresponding to peptide binding pockets P1 through P4 and the segment of the  $\alpha_1$  helix near P1 and P2, only the molecular probe distributions in this region are discussed. The resulting library contained three diversity positions, two specificity positions, and a novel branching point. The library was built on a peptide-like skeleton, it is summarized in Figs. 4 and 5. To highlight how the molecular probe occupancies as determined by MCSS correspond to monomer selections for the library, the molecular probe maps are analyzed with respect to the positions of the library elements in the proposed binding mode in the protein. That is, the position of clusters (of different group types) whose centroids are closest to each monomer attachment point in the proposed ligand framework are discussed. The computed diversities of the clusters occupying the elements of the library are summarized in Table 4. Residues are referred to as numbered

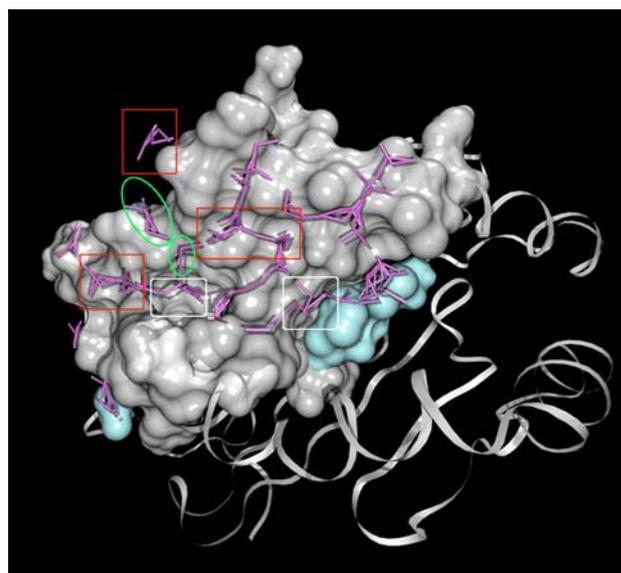


**Fig. 3** Representative molecular probe maps for HLA-DR4: (a) arginine side-chain (propyl guanidinium) and (b) benzene. The protein is represented by a surface in the region for which the library was targeted and by ribbons elsewhere to provide context. The molecular probes are represented by thin stick figures colored by atom type. Note that there are few propyl guanidinium minima in the design region, this preference is reflected by having almost no basic functionalities in the designed library

in the DR4 structure with corresponding residues from the DR3 structure in parenthesis.

**Scaffold.** The propane molecular probe (PRPN) map served as the basis for designing the scaffold onto which the combinatorial library was built. The PRPN map gives an approximate shape for the library. In fact, PRPN may prove to be a good general purpose group for defining library shape. That is, PRPN may serve as a “groove finding” probe which helps highlight the shape of the binding site into which a ligand could be built. This is true because PRPN is small but has a sufficient number of atoms to have orientational preference and because all PRPN atoms are charge neutral so the probe will not exclude regions because of electrostatic repulsion.

The propane replicas depict a well defined skeleton with tetrahedral geometry near a possible branching position for a library. The map traces out the peptide binding groove as well as a groove near a break in the  $\alpha_1$  helix. The resulting chemical library is peptide-like because the chemistry for coupling monomers using a peptide bond is well established and reagents were readily available. Thus, the overall library design results from using the PRPN map to define the approximate shape of the library skeleton and



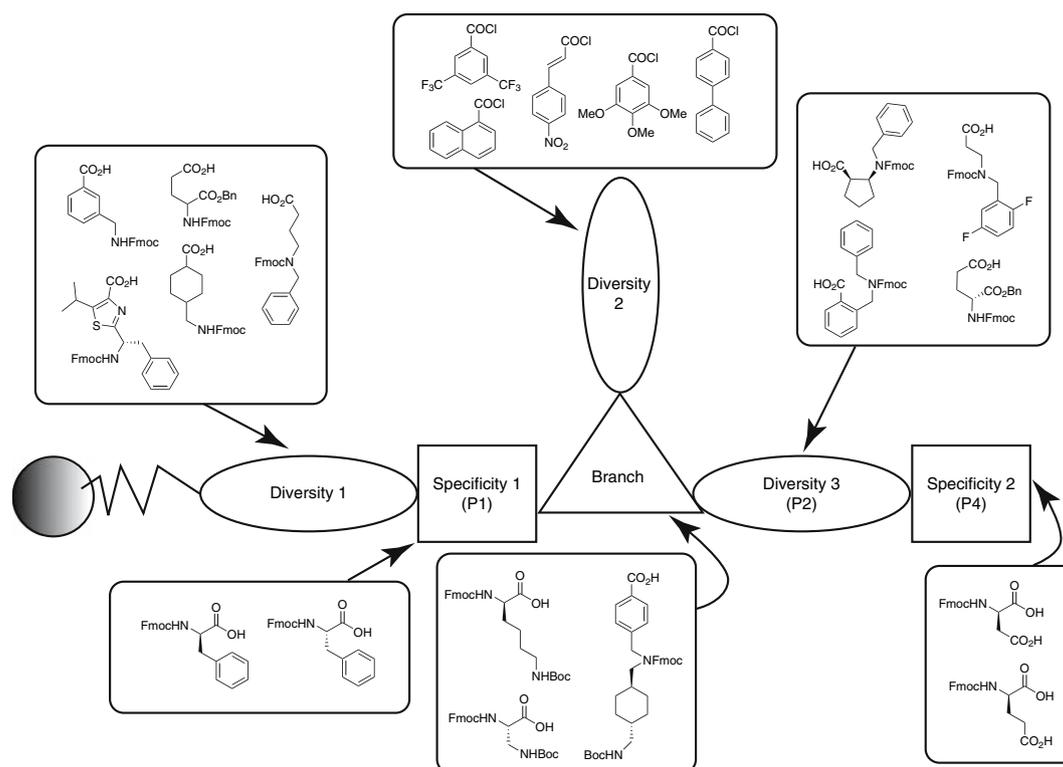
**Fig. 4** Solvent accessible surface representation of the part of the binding site of HLA-DR4 targeted with the combinatorial library. The propane minima defining the library architecture are shown in purple sticks. Positions targeted by combinatorial library elements are denoted by red boxes for diversity positions, white rounded corner boxes for specificity elements, and green ovals for the branch positions. The residues that differ between HLA-DR3 and HLA-DR4 are colored light blue, the ribbon diagram shows the rest of the protein for context

other MCSS groups to define the chemical character of specific parts of the library. Combinatorial chemistry was then used to explore a region of chemical space that fits within the MCSS-derived shape.

The propane map, along with the locations of clusters with relatively high occupancy and the knowledge of established binding pockets defined the positions of the library elements described below.

**Diversity Position 1.** This position lies between the linker from the bead and peptide pocket one (P1), it is within 6.5 Å of atoms in residues Phe 51, Ala 52, Ser 53, Phe 54, His 260, and Val 264 (F47, A48, S49, F50, H253, V257 in DR3) of the protein. Of these residues, the backbone atoms of DR4 residues F51, S53, F54 would contact a group placed in this position; side chain atoms of A52, H260, and V264 are near this position. In the HLA-DR4 active site map, cluster 12 (clusters 48 and 112 for the HLA-DR3 map) is closest to this position. Cluster 12 from the DR4 map contains hydrophobic, polar, and negatively charged probes. This preference is mirrored in cluster 48 from the DR3 map; cluster 112 has a smaller population of probes and shows a stronger preference for polar probes.

The cluster nearest this position (12 in HLA-DR4) has a relatively high dissimilarity score so a diverse selection of monomers was incorporated at this site. These monomers included aromatic, aliphatic, and carboxylate moieties



**Fig. 5** Schematic illustration of the design of the combinatorial library. Chemical structures shown are a sample of monomers chosen for inclusion in the various parts of the library. P1, P2, and P4 indicate the position on the protein for which the respective elements of the

library are targeted. For complete lists of library components see Fig. 6 for “Diversity 1”, Fig. 7 for “Specificity 1”, Fig. 8 for “Branch”, Fig. 9 for “Diversity 2”, Fig. 10 for “Diversity 3”, and Fig. 11 for “Specificity 2” monomers

**Table 4** Summary of clusters of functional group minima for HLA-DR4 that correspond to proposed combinatorial library elements

Library element	Cluster number	Number of minima in cluster	Computed dissimilarity	Dissimilarity normalized by number of groups
Diversity 1	12	135	$2.625 \times 10^5$	14.40
Specificity 1	6	306	$1.005 \times 10^6$	10.74
Diversity 2	102	71	$8.414 \times 10^4$	16.69
Diversity 3	11	240	$5.934 \times 10^6$	11.30
	49	129	$1.980 \times 10^5$	11.90
	173	103	$1.662 \times 10^5$	15.66
Specificity 2	3	175	$3.368 \times 10^5$	11.00

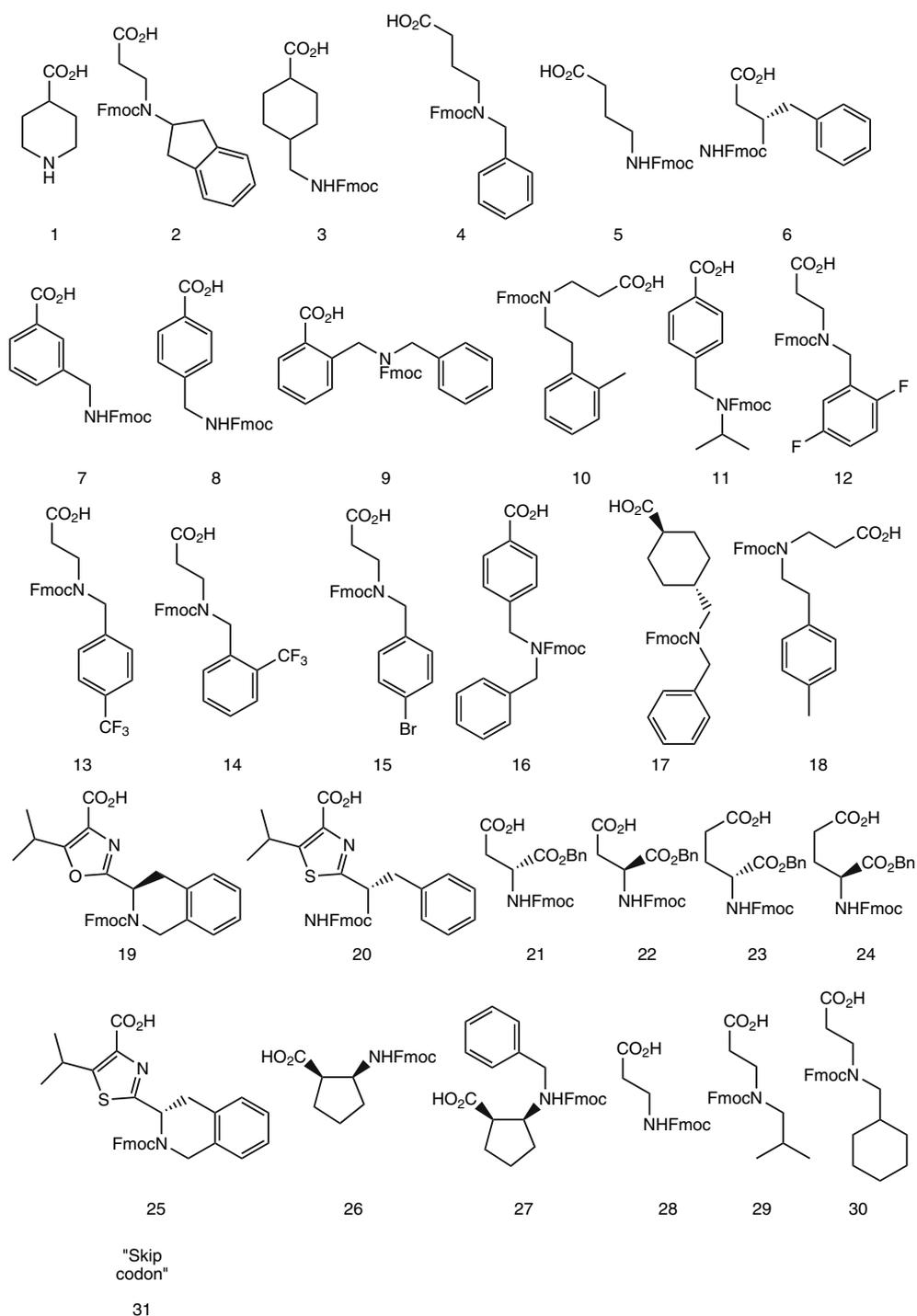
which matches the collection of molecular probes occupying this region of the binding site. There were 31 reagents specified for this portion of the library (Fig. 6).

**Specificity Position 1.** This position corresponds to P1. Both the DR3 and DR4 allotypes of HLA have a preference for phenylalanine at this position [38, 58, 71]. The entrance to this pocket is occupied by cluster 10 in HLA-DR4 (82 in HLA-DR3) and the pocket itself by cluster 6 in DR4 (45 in HLA-DR3) maps. DR4 residues within 6.5 Å of cluster 6 are Phe 24, Ile 31, Phe 32, Trp 43, Ala 52, Ser 53, Phe 54,

His 260, Asn 261, Tyr 262, Val 264, Gly 265, and Phe 268 (F20, I27, F28, W39, A48, S49, F50, N254, V257, V258, F261 in HLA-DR3). The side chains of all of these except residues 53, 260, 262, and 265 (HLA-DR4 numbering) potentially contact ligand atoms in this pocket. They create a predominantly hydrophobic pocket. The contents of cluster 6 in DR4 (45 in HLA-DR3) show an enriched population of hydrophobic and aromatic probes (2BTN, ILER, PHER) as well as the hydrophobic tails of the ARGR probe in the pocket. These results agree with the observed preference for hydrophobic side chains in peptides that bind HLA-DR4 (e.g., [37, 58]).

In particular, the peptide sequence binding motifs for HLA-DR4 indicate that phenylalanine is preferred at P1. Since this preference is mirrored in the MCSS functional group maps, only the R and S enantiomers of phenylalanine were incorporated at library position (Fig. 7). The MCSS results do not provide a stereochemical preference for this position. In light of this, both enantiomers were included in the library design. This exemplifies the synergy of experimental and computational methods: where the computational result is ambiguous library elements can be incorporated to test the possible hypotheses. The relatively low diversity score computed for cluster 6 is consistent

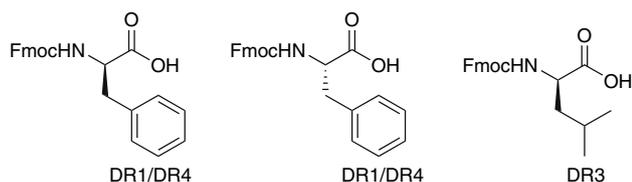
**Fig. 6** Monomers included for Diversity 1 position



with its use as a conserved, specific position; i.e., library position was intended to confer specificity for HLA-DR4 as well as to provide “register” for the library, aligning the other library elements in their proposed positions.

**Branch Position.** This position originates near  $C_{\alpha}$  of Arg 369 of the CLIP peptide in the complex with HLA-DR3 and leads out toward Diversity Position 2 described below. The side chain of Phe 54 in HLA-DR4 (F50 in HLA-DR3)

forms a hydrophobic patch connecting the peptide binding groove to Diversity Position 2. The Branch Position lies above and between P1 and P2. In the HLA-DR4 functional group maps the centroid of cluster 9 lies at the branch point, in the HLA-DR3 maps, cluster 52 occupies the branch point and 104 occupies the region leading out of the binding groove. In HLA-DR4, residues Gln 9, Phe 24, Phe 54, Thr 256, Tyr 257, His 260, and Asn 261 are within



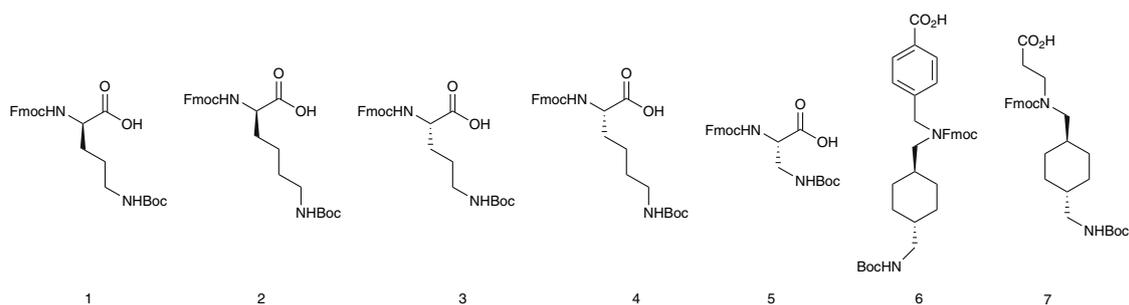
**Fig. 7** Monomers included for Specificity 1 position. The DR1/DR4 and DR3 annotations indicate predicted specificities based on known peptide binding preferences of the respective HLA allotypes

6.5 Å of the centroid of cluster 9. (In HLA-DR3, residues Q5, F20, F50, E51, N249, Y250, H253, and N254 are within 6.5 Å of the centroid of cluster 52 and residues N249, Y250, H253, and N254 are within 6.5 Å of the centroid of cluster 104.) Cluster 9 for HLA-DR4 (cluster 52 for HLA-DR3) shows a fairly even population of molecular probes with a preference for aliphatic moieties (ILER and PRPN probes).

The exact spacing and conformational requirements of the Branch monomer necessary to connect to Diversity Position 2 and Diversity Position 3 were not known. To accommodate this ambiguity, a range of linker sizes (1–4 carbon straight chains and cyclohexyl) were included to connect Diversity Position 2 at the branch point. The 1–4 carbon linkers were conformationally free (no double

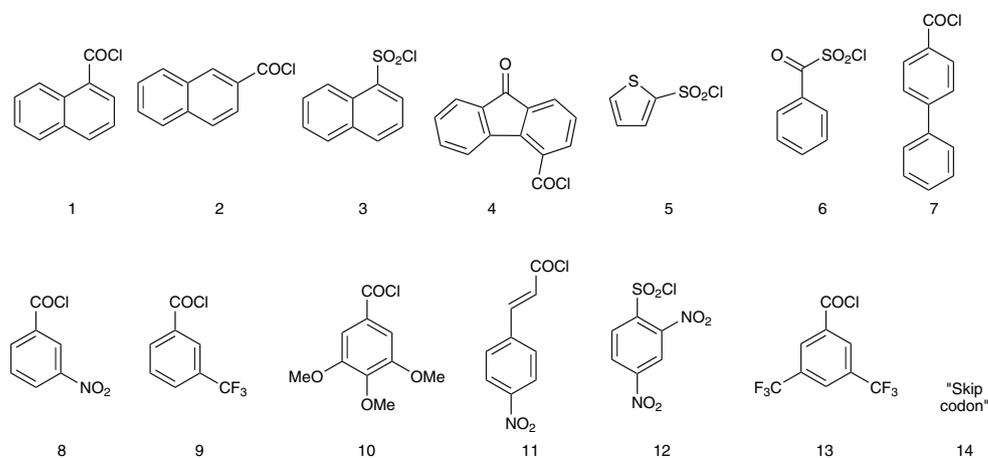
bonds) and stereoisomers were included. In addition, a range of linker sizes (single carbon and benzene ring) were allowed at the other limb of the Branch Point to connect to the monomer at Diversity Position 3. In total, there were seven reagents included in the library for this position (Fig. 8). All were aliphatic which reflects the preferences observed in the MCSS molecular probe maps; but a variety of components with varying length, shape, and flexibility were included to account for the multiple possibilities represented in the MCSS results.

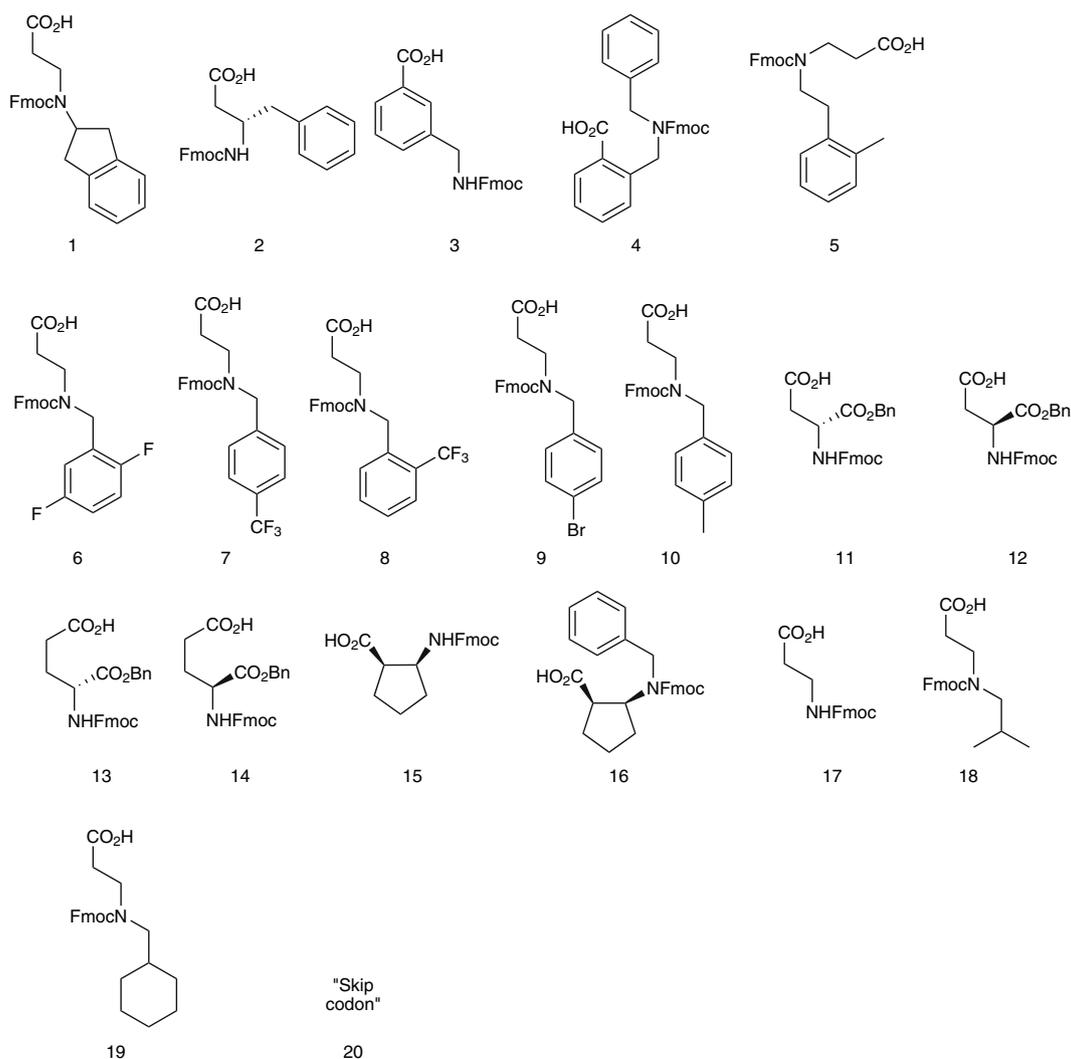
**Diversity Position 2.** This position corresponds to a groove or notch adjacent to the peptide binding groove at a break in the helical structure of the  $\alpha_1$  helix bounding the binding groove. It is formed primarily by residues Ser 53, Phe 54, and Glu 55 (S49, F50, and E51 of HLA-DR3). The region is a relatively flat and wide valley into which a part of the ligand may bind. The floor of the valley is formed primarily by the backbone atoms of these residues. The aliphatic parts of the S53 and E55 side chains forming the sides of the valley and the F54 side chain comprising the hydrophobic connection to the Branch Position and the binding groove. Cluster 102 for HLA-DR4 145 for HLA-DR3) occupies this region. The cluster (in both proteins) has a low occupancy and a high dissimilarity score. Inspection of the cluster contents shows energetically



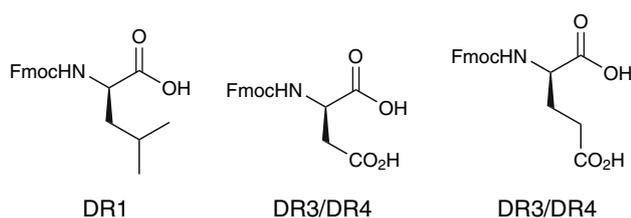
**Fig. 8** Monomers included for Branch position

**Fig. 9** Monomers included for Diversity 2 position





**Fig. 10** Monomers included for Diversity 3 position

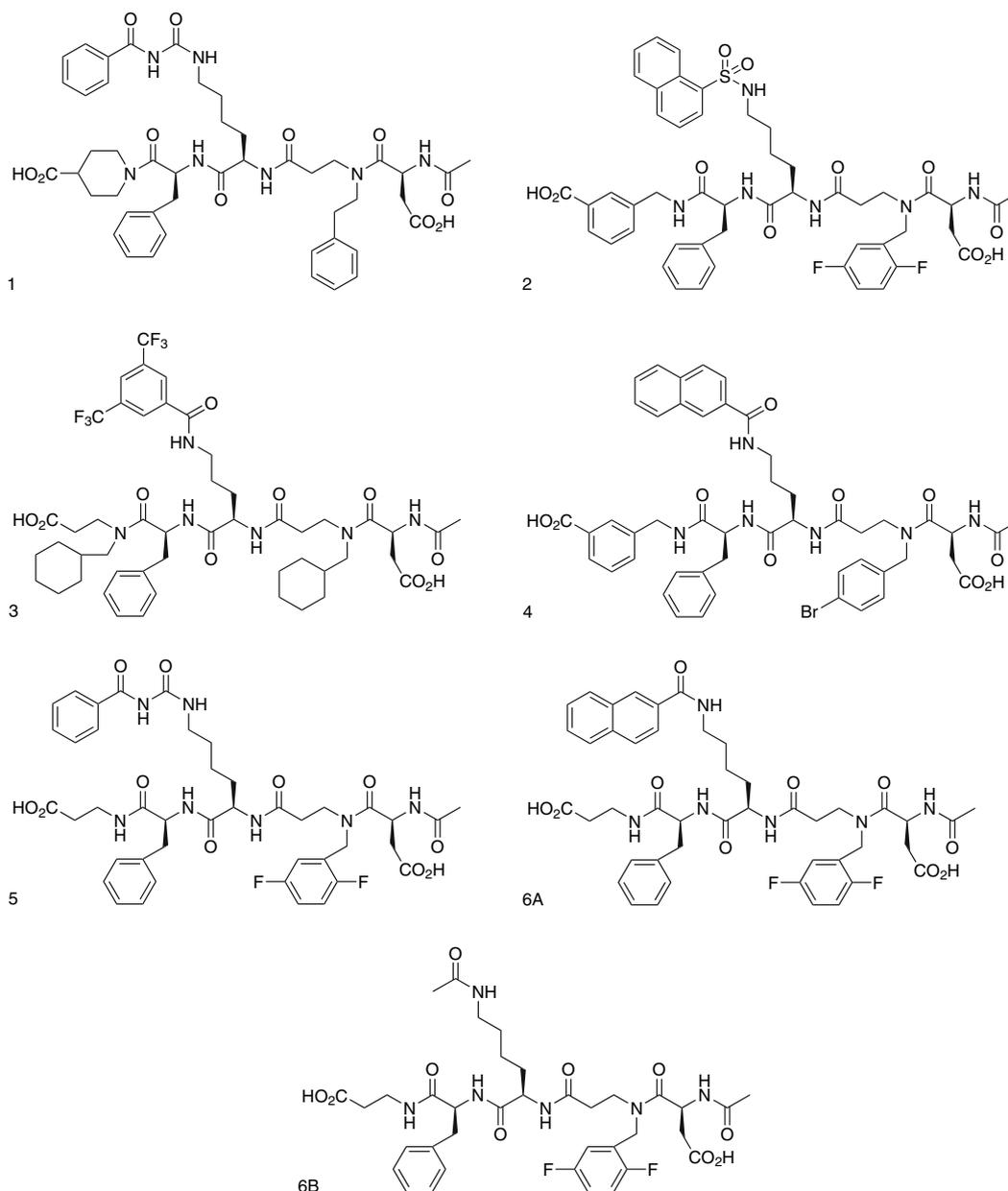


**Fig. 11** Monomers included for Specificity 2 position. The DR1 and DR3/DR4 annotations indicate predicted specificities based on known peptide binding preferences of the respective HLA allotypes

favorable cyclohexane minima as well as a somewhat enriched number of aliphatic and aromatic probes (2BTN, ILER, PRPN, PHER, LEUR). This region is a curious feature of the HLA-DR structure because of the break in the bounding  $\alpha_1$  helix. In addition, the template established by the PRPN maps leads into this region. Therefore, the ligand was designed into this region and diversity in the

synthetic library was used to probe experimentally the binding preferences for this part of the HLA-DR structure. Because of the enriched aliphatic and aromatic cluster contents and visual inspection of the region and maps which indicated that larger aromatic functionalities such as naphthyl groups could occupy this space, benzene, mono-, di-, and tri-substituted benzene, naphthalene, and similar monomers were included at this library position (Fig. 9). In total, 14 monomers were chosen for Diversity Position 2.

*Diversity Position 3.* Diversity Position 3 roughly occupies P2 and overlaps, to a degree, with the sidechain of Met 370 of the CLIP peptide in the HLA-DR3 co-crystal structure. Cluster 49 in HLA-DR4 (60 in HLA-DR3) occupies the substituent pocket for this position; cluster 11 (181 in HLA-DR3) lies at the backbone position for this part of the library. In HLA-DR4, residues Phe 54, Glu 55, Gln 57, Gly 58, Ala 59, and Asn 62 are within 6.5 Å of the centroid of cluster 49 (in HLA-DR3, F50, E51, Q53, G54,



**Fig. 12** Chemical structures of a sampling of the compounds from the combinatorial library that were identified from positively reporting beads in stringent on-bead ELISA: compounds 1-6A are hits from the library, compound 6B is a derivative of 6A lacking the

naphthoyl moiety in Diversity 2. The figures are drawn oriented to the proposed binding mode discussed in Sect. 'Proposed structure and binding mode' as opposed to the conventional orientation of amino terminus to carboxyl terminus, left to right

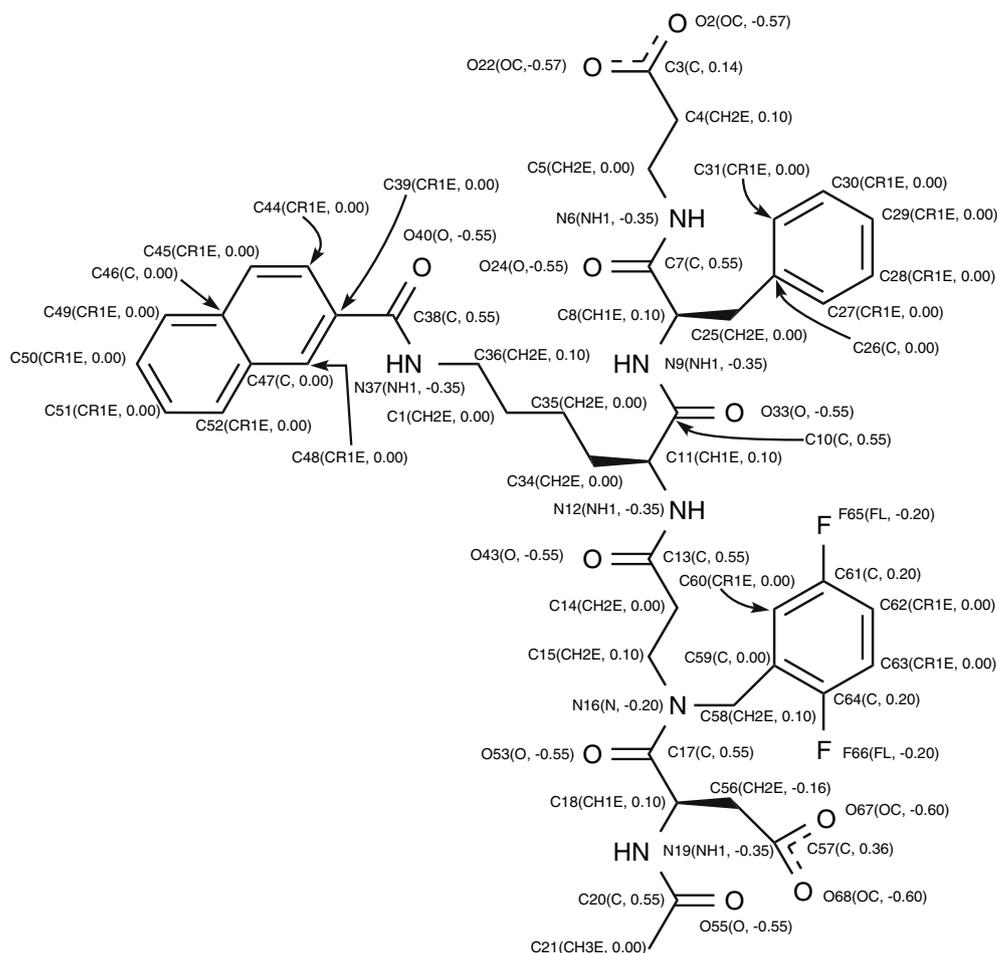
and A55 are within 6.5 Å of the centroid of cluster 60). DR4 residues Gln 9, Phe 22, Phe 24, Phe 54, Gly 58, Ala 59, Asn 62, Tyr 257, and Asn 261 are within 6.5 Å of the centroid of cluster 11 (in DR3, Q5, E7, F18, F20, F50, N58, S185, R246, Y250, and N254 within 6.5 Å of the centroid of cluster 181).

The dissimilarity scores of the clusters in this region suggest a conserved portion of the ligand near the backbone, as expected, with more diversity in the ligand side chain pocket. The library monomers allowed at this posi-

tion had varying length backbone linkers to the previous monomer as well as diverse sidechains including aliphatics, aromatics, halogenated aromatics, and carboxylates (Fig. 10). In all, there were 20 monomers selected for inclusion in the library at this position.

*Specificity Position 2.* This position corresponds to P4. It overlaps with the position of Ala 371 of the CLIP peptide in the HLA-DR3 structure. In the HLA-DR4 structure, residues Gln 9, Asn 62, His 192, Phe 205, Asp 207, Lys 250, Ala 253, and Tyr 257 lie within 6.5 Å of cluster 3 and

**Fig. 13** Chemical structure of compound 6A showing atom names, CHARMM atom types, and partial charges

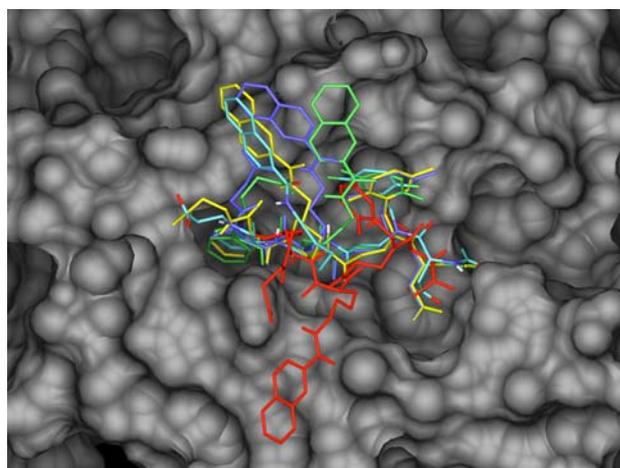


form P4. HLA-DR4 has been observed to have a preference for binding peptides with negative charge [37, 58] at this position and there are favorable (ranked in the top 30) acetate ion (ASPR probe) minima in the cluster corresponding to this position. Because of these observations, aspartate and glutamate monomers were included in the library at this position (Fig. 11).

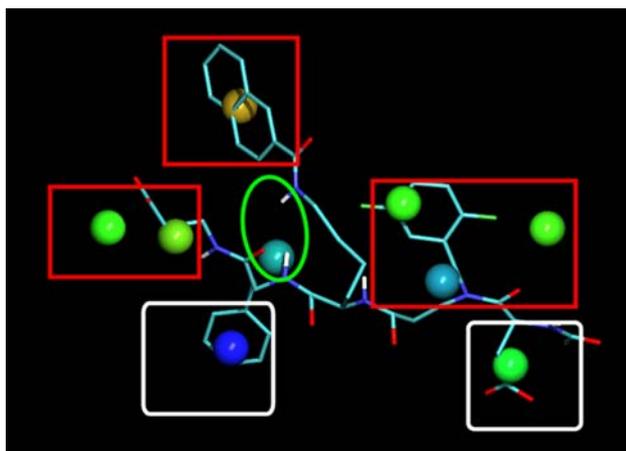
*Overall.* In total the combinatorial library contained 243040 compounds. These compounds were synthesized on solid phase (see Methods), with three resin beads displaying each library member.

#### Identity of ligands

This section discusses the chemical structure of the compounds selected by screening the synthetic combinatorial library. Under the more stringent assay conditions of 7.5 nM protein concentration, 14 ligands were identified from approximately 60 positively reporting (stained dark in the on-bead ELISA) beads. The majority of these compounds have the same enantiomeric preferences at the first specificity and branch positions (R and S, respectively). A



**Fig. 14** Model built structure of compound 6A in the binding site of HLA-DR4. The original handbuilt structure is colored by atom types and a sample of structures resulting from simulated annealing refinement are shown in solid colors: best energy structure with no constraints on final reannealing (blue), second best structure with no constraints (yellow), structure with poorest interaction energy with the protein after refining with no constraints in second annealing step (red), best energy structure refined with decreasing harmonic constraints during reannealing (green); see text for more details on ranges of energies



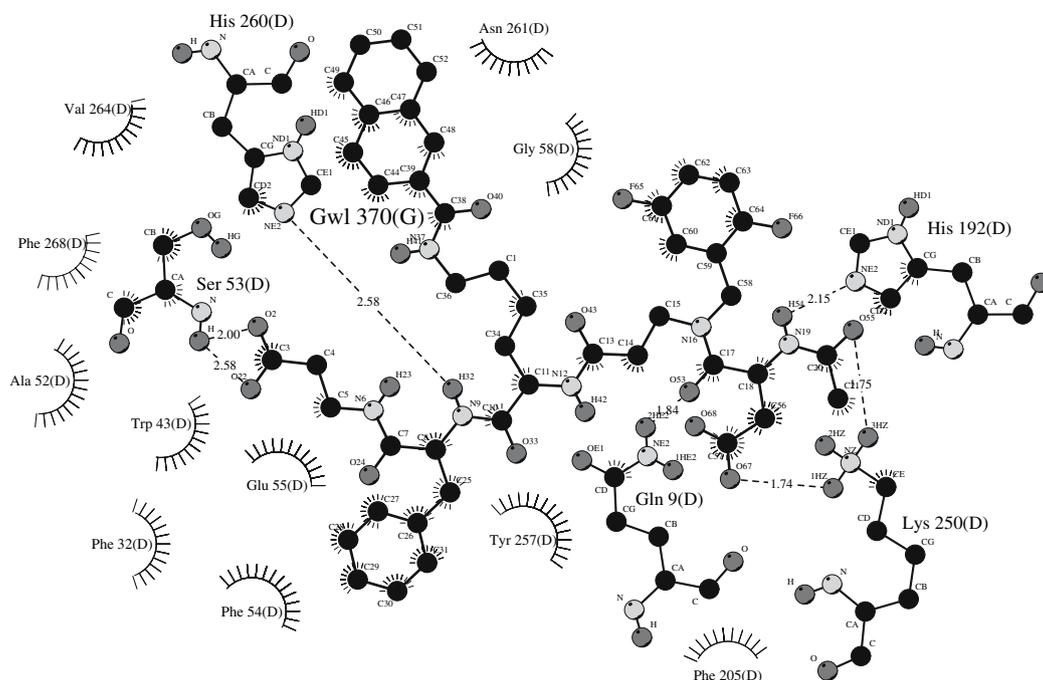
**Fig. 15** Model built structure of compound 6A. Clusters of MCSS minima corresponding to elements of the combinatorial library are shown as spheres with blue and green spheres having relatively higher occupancies. This shows how the model built structure fits the clusters used for developing the combinatorial library

sample of the compounds identified as HLA-DR4 ligands is shown in Fig. 12. The selected compounds display a preference for aliphatic or aromatic groups at Diversity Position 1. They have straight linkers of three or four carbon atom lengths at the Branch Position and naphthyl, 3,5-di-trichloromethyl benzyl, benzamido, and naphthyl sulfonyl functionalities in Diversity Position 2, with a small preference for the naphthyl group. Diversity Position 3

favors halogenated benzene rings, benzene, or cyclohexane. Specificity Position 2 shows a preference for aspartate over glutamate (only R enantiomers were incorporated at this position). Compound 6A was selected as a representative of the identified ligands because it contained the functionalities that occurred most frequently in the ligands. The atom names and partial charges for 6A are illustrated in Fig. 13. Compound 6A and a derivative containing a methyl group in place of the naphthyl group at Diversity Position 2 (compound 6B in Fig. 12) were synthesized on solid phase for further binding studies.

#### Proposed structure and binding mode

Ligand 6A was model built into its proposed binding position and orientation as described in Sect. ‘Model building and refinement’. Three key elements were aligned with their corresponding MCSS functional group maps for their respective positions: the benzyl group of Specificity 1 was aligned to BENZ and PHER in P1, the difluorobenzyl group with BENZ and TFME minima near P2 and P3 (Diversity 3), and the acetyl group with an ASPR minimum at P4 (Specificity 2). During the initial minimization and annealing, the C26 and C59, corresponding to the  $\gamma$  carbon of the phenylalanine in P1 and the  $\gamma$  carbon on the difluorobenzyl functionality near P3, respectively, of the ligand were fixed to maintain the position of the ligand relative to the protein. The resulting structure of the complex was



**Fig. 16** Two-dimensional schematic representation of compound 6A (labeled Gwl 370) showing the hydrogen bond and nonbond interactions between 6A and HLA-DR4. Figure created using LIGPLOT [109]

reannealed first with and then without constraints on the ligand. In the structures with the best CHARMM interaction energy between the ligand and the protein, the ligand remained in the position in which it was built; that is, the key elements that were aligned as described above remained in their respective locations (Figs. 14 and 15). The details of the interactions between compound 6A and the protein are shown schematically in Fig. 16.

## Conclusions

Chemical space is a vast resource for pharmaceutical discovery. Combinatorial synthesis methods enable an increased rate of sampling of this space. Nevertheless, large-scale random or uninformed exploration of chemical space remains an inefficient approach to identifying leads for pharmaceutical development. The coupling of computational structure-based ligand design methods, combinatorial synthesis, and large scale screening presented in this paper illustrate a procedure for efficiently identifying regions of chemical space that yield compounds with affinity and specificity for a given target molecule.

By using the MCSS method to compute the functional group binding preferences for a target with known structure, it is possible to generate rapidly hypotheses about the types of compounds that are likely to bind to the target. As a de novo ligand design method, MCSS is a valuable tool because it not only aids chemical intuition but also inspires novel ideas in formulating a scaffold and selecting substituents with which to populate the combinatorial library. In the case of the HLA-DR4 ligands described in this paper, the MCSS results led directly to the creation of the branched library architecture. Also, the clustering and diversity methods described assisted in clarifying which positions should contain diverse substituents and which should be conservatively populated.

One of the limitations of computational ligand design is the difficulty in determining the binding affinity of ligands which makes it hard to discriminate promising compounds. Because combinatorial synthesis and large-scale screening methods enable the evaluation of a large number of compounds they complement the computational methodology; that is, the computational method is able to narrow the search for promising ligands to a particular neighborhood of chemical space and the experimental methods are able to search that neighborhood more efficiently and accurately than available computational methods.

The results suggest that the mixed computational-experimental approach presented here works in practice. In the first synthetic combinatorial library several ligands were identified that compete favorably with known peptide ligands for binding to HLA-DR4. While direct

crystallographic verification of the designed ligand complexed with the protein was not possible, the peptide binding exclusion results indicate that the designed ligand occupies the peptide binding groove of HLA-DR4.

Designing a strong binding ligand is an iterative process in which a lead is identified and hypotheses for improving binding and specificity are generated and tested. In the hybrid combinatorial approach the computational results were used to formulate a first generation library that was focused for the HLA-DR4 binding site. The calculations replaced the experimental screen of an exploratory library which would have been difficult given the promiscuous ligand binding characteristics of the target. Experimental methods were used to test the proposed library and computational analyses rationalized the experimental data. In addition, the computational models provide a framework for using the experimental results to guide refinement of the library to optimize binding; that is, to more finely explore the chemical space centered around the identified lead molecules. Even without crystallographic data, model building results provide a context for using structure-based design methods to suggest refinements in the library such as rigidifying the framework to fix the favored library elements in their positions.

## References

1. Gallop MA, Barrett RW, Dower WJ, Fodor SPA, Gordon EM (1994) *J Med Chem* 37(9):1233–1251
2. Gordon EM, Barrett RW, Dower WJ, Fodor SPA, Gallop MA (1994) *J Med Chem* 37(10):1385–1401
3. Czarnick AW (1996) *Acc Chem Res* 29(3):112–113
4. Beroza PP, Feinstein R, Greene J, Goldman B, Mount J, Smellie A (1999) The design of maximum information libraries as a highly efficient tool for lead generation. Theory and methods. presentation at acs national meeting, anaheim, ca., CombiChem, Inc., Palo Alto, CA
5. Bradley EK, Beroza P, Eksterowicz J, Genin H, Grootenhuys P, Miller D, Miller J, Penzotti J, Shirley W, Spellmeyer D, Srinivasan J, Stanton R (1999) The design of maximum information libraries as a highly efficient tool for lead generation. Application of the method. presentation at acs national meeting, anaheim, ca., CombiChem, Inc., Palo Alto, CA
6. Kick EK, Roe DC, Skillman AG, Liu G, Ewing TJA, Sun Y, Kuntz ID, Ellman JA (1997) *Chem Biol* 4:297–307
7. Kubinyi H (1998) *Curr Op Drug Disco Devel* 1(1):16–27
8. Murray CM, Cato SJ (1999) *J Chem Inf Comput Sci* 39(1):46–50
9. Tondi D, Slomczynska U, Costi MP, Watterson DM, Ghelli S, Shoichet BK (1999) *Chem Biol* 6(5):319–331
10. Martin EJ, Critchlow RE (1999) *J Comb Chem* 1(1):32–45
11. Koehler RT, Dixon SL, Villar HO (1999) *J Med Chem* 42(22):4695–4704
12. Mason JS, Morize I, Menard PR, Cheney DL, Hulme C, Labaudiniere RF (1999) *JMC* 42(18):3557–3571
13. Miranker A, Karplus M (1991) *Proteins* 11(1):29–34
14. Evensen E, Joseph-McCarthy D, Karplus M (1997) MCSS version 2.1. Harvard University, Cambridge, MA, USA

15. Caffisch A, Miranker A, Karplus M (1993) *J Med Chem* 36(15):2142–2167
16. Caffisch A (1996) *J Comput-Aided Mol Des* 10(5):372–396
17. Eisen MB, Wiley DC, Karplus M, Hubbard RE (1994) *Proteins* 19:199–221
18. Miranker A, Karplus M (1995) *Proteins: Structure, Function, and Genetics* 23:472–490
19. Bohacek RS, McMartin C (1994) *J Am Chem Soc* 116:5560–5571
20. Ripka AS, Satyshur KA, Bohacek RS, Rich DH (2001) *Org Lett* 3:2309–2313
21. Grzybowski BA, Ishchenko AV, Kim C-Y, Topalov G, Chapman R, Christianson DW, Whitesides GM, Shakhnovich EI (2002) *Proc Natl Acad Sci* 99:1270–1273
22. Firth-Clark S, Willems HMG, Williams A, Harris W (2006) *J Chem Inf Model* 46:642–647
23. Joseph-McCarthy D, Tsang SK, Filman DJ, Hogle JM, Karplus M (2001) *J Am Chem Soc* 123:12758–12769
24. Mason JS, Morize I, Menard PR, Cheney DL, Hulme C, Labaudiniere RF (1999) *J Med Chem* 42:3251–3264
25. Mason JS, Cheney DL (2000) *Pac Symp Biocomput* 5:573–584
26. Eksterowicz JE, Evensen E, Lemmen C, Brady GP, Lancot JK, Bradley EK, Saiah E, Robinson LA, Grootenhuis PDJ, Blaney JM (2002) *J Mol Graph Model* 20:469–477
27. Evensen E, Eksterowicz JE, Stanton RV, Oshiro C, Grootenhuis PDJ, Bradley EK (2003) *J Med Chem* 46:5125–5128
28. Deng Z, Chuaqui C, Singh J (2006) *J Med Chem* 49:490–500
29. Joseph-McCarthy D, Alvarez JC (2003) *Proteins* 51:189–202
30. Hajduk PJ, Meadows RP, Fesik SW (1999) *Q Rev Biophys* 32:211–240
31. Petros AM, Dignes J, Augeri DJ, Baumeister SA, Betebenner DA, Bures MG, Elmore SW, Hajduk PJ, Joseph MK, Landis SK, Nettlesheim DG, Rosenberg SH, Shen W, Thomas S, Wang X, Zanze I, Zhang H, Fesik SW (2006) *J Med Chem* 49:656–663
32. Card GL, Blasdel L, England BP, Zhang C, Suzuki Y, Gillette S, Fong D, Ibrahim PN, Artis DR, Bollag G, Milburn MV, Kim SH, Schlessinger J, Zhang KY (2005) *Nat Biotechnol* 23:201–207
33. Gill A, Cleasby A, Jhoti H (2005) *Chembiochem* 6:506–512
34. Hartshorn MJ, Murray CW, Cleasby A, Frederickson M, Tickle IJ, Jhoti H (2005) *J Med Chem* 48:403–413
35. Erlanson DA, McDowell RS, O'Brien T (2004) *J Med Chem* 47:3463–3482
36. Erlanson DA (2006) *Curr Opin Biotechnol* 17:643–652
37. Hammer J, Gallazi F, Bono E, Karr RW, Guenot J, Valasini P, Nagy ZA, Sinigaglia F (1995) *J Exp Med* 181:1847–1855
38. Wucherpfennig KW, Strominger JL (1995) *J Exp Med* 181:1597–1601
39. Czaja AJ (2005) *Ann Hepatol* 4:6–24
40. Janeway CA Jr, Travers P (1996) *Immunobiology: The Immune System in Health and Disease*. Garland Publishing Inc., New York, NY
41. Powis SH, Geraghty DE (1995) *Immunol Today* 16:466–468
42. Campbell RD, Trowsdale J (1993) *Immunol Today* 14:349–352
43. Nepom GT, Erlich H (1991) *Annu Rev Immunol* 9:493–525
44. Theofilopoulos AN (1995) *Immunol Today* 16:90–98
45. Theofilopoulos AN (1995) *Immunol Today* 16:150–158
46. Mueller DL, Jenkins MK (1997) *Curr Biol* 7:R255–R257
47. Wong FS, Wen L (2003) *Curr Mol Med* 3:1–15
48. Larsen CE, Alper CA (2004) *Curr Opin Immunol* 16:660–667
49. Li Y, Huang Y, Lue J, Quandt JA, Martin R, Mariuzza RA (2005) *EMBO J* 24:2968–2979
50. George AJT, Ritter MA, Lechler RI (1995) *Immunol Today* 16:209–211
51. Wilkinson AJ (1996) *Chem Biol* 3:519–524
52. Abbas AK, Lichtman AH, Pober JS (1997) *Cellular and Molecular Immunology*, 3rd ed. Saunders, Philadelphia, PA, USA
53. Stern LJ, Wiley DC (1994) *Structure* 2(4):245–251
54. Madden DR, Gorga JC, Strominger JL, Wiley DC (1991) *Nature* 353:321–325
55. Madden DR, Garboczi DN, Wiley DC (1993) *Cell* 75:693–708
56. Garrett TPJ, Saper MA, Bjorkman PJ, Strominger JL, Wiley DC (1989) *Nature* 342:692–696
57. Gou H-C, Madden DR, Silver ML, Jardetzky TS, Gorga JC, Strominger JL, Wiley DC (1993) *Proc Natl Acad Sci USA* 90:8053–8057
58. Sette A, Sidney J, Oseroff C, del Guercio M-F, Southwood S, Arrhenius T, Powell MF, Colón SM, Gaeta FCA, Grey HM (1993) *J Immunol* 151(6):3163–3170
59. Rognan D, Krebs S, Kuonen O, Lamas JR, López de Castro JA, Folkers G (1997) *JCAMD* 11:463–478
60. Ghosh P, Amaya M, Mellins E, Wiley DC (1995) *Nature* 378:457–462
61. Humphrey WF, Dalke A, Schulten K (1996) *J Mol Graphics* 14:33–38
62. POV-Team. <http://www.povray.org>
63. Almagro JC, Vargas-Madrado E, Lara-Ochoa F, Horjales E (1995) *Prot Sci* 4:1708–1717
64. Sant'Angelo DB, Waterbury G, Preston-Hurlburt P, Yoon ST, Medzhitov R, Hong S-C, Janeway CA Jr (1996) *Immunity* 4:367–376
65. Garbozci DN, Ghosh P, Utz U, Fan QR, Biddison WE, Wiley DC (1996) *Nature* 384:134–141
66. Vasmatazis G, Cornette J, Sezerman U, DeLisi C (1996) *J Mol Biol* 261:72–89
67. Bouvier M, Wiley DC (1994) *Science* 265:398–402
68. Cresswell P (1996) *Cell* 84:505–507
69. Jensen PE, Weber DA, Thayer WP, Westerman LE, Dao CT (1999) *Immunol Rev* 172:229–238
70. Busch R, Rinderknecht CH, Roh S, Lee AW, Harding JJ, Burster T, Hornell TM, Mellins ED (2005) *Immunol Rev* 207:242–260
71. Malcherek G, Falk K, Röttschke O, Rammensee H-G, Stevanovic S, Gnau V, Jung G, Melms A (1993) *Int Immunol* 5(10):1229–1237
72. Falcioni F, Ito K, Vidovic D, Belunis C, Campbell R, Berthel SJ, Bolin DR, Gillespie PB, Huby N, Olson GL, Sarabu R, Guenot J, Madison V, Hammer J, Sinigaglia F, Steinmetz M, Nagy ZA (1999) *Nature Biotech* 17:562–561
73. Eldor R, Cohen IR, Raz I (2005) *Int Rev Immunol* 24:327–339
74. Haque A, Blum JS (2005) *J Biol Regul Homeost Agents* 19:93–104
75. Dzhabazov B, Nandakumar KS, Kihlberg J, Fugger L, Holmdahl R, Vestberg M (2006) *J Immunol* 176:1525–1533
76. Smith AB III, Benowitz AB, Guzman MC, Sprengeler PA, Hirschmann R, Schweiger EJ, Bolin DR, Nagy Z, Campbell RM, Cox DC, Olson GL (1998) *J Am Chem Soc* 120:12704–12705
77. Smith AB III, Benowitz AB, Sprengeler PA, Barbosa J, Guzman MA, Hirschmann R, Schweiger EJ, Bolin DR, Nagy Z, Campbell RM, Cox DC, Olson GL (1999) *J Am Chem Soc* 121:9286–9298
78. Woulfe SL, Bono CP, Zacheis ML, Welply JK, Kirschmann DA, Baudino TA, Wang Y, Stone DA, Hanson GJ, Vuletic JL, Bedell LJ, Schwartz BD, Howard SC (1997) *J Pharm Expt Therapeutics* 281:663–669
79. Bolin DR, Swain AL, Sarabu R, Berthel SJ, Gillespie P, Huby NJS, Makofske R, Orzechowski L, Perrotta A, Toth K, Cooper JP, Jiang N, Falcioni F, Campbell R, Cox D, Gaizband D, Belunis CJ, Vidovic D, Ito K, Crowther R, Kammlott U, Zhang

- X, Palermo R, Weber D, Guenot J, Nagy Z, Olson GL (2000) *J Med Chem* 43:2135–2148
80. Dedier S, Krebs S, Lamas JR, Poenaru S, Folkers G, Lopez de Casto JA, Seebach D, Rognan D (1999) *J Recept Signal Transduction Res* 19:645–657
81. Madurga S, Belda I, Llorà X, Giralt E (2005) *Protein Sci* 14:2069–2079
82. Dessen A, Lawrence MC, Cupo S, Zaller DM, Wiley DC (1997) *Immunity* 7:473–481
83. Brünger AT, Karplus M (1988) *Proteins* 4:148–156
84. Brooks BR, Brucoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M (1983) *J Comp Chem* 4:187–217
85. MacKerell AD Jr, Brooks B, Brooks CL III, Nilsson L, Roux B, Won Y, Karplus M (1998) In: Schleyer PvR et al (eds) *The encyclopedia of computational chemistry*, vol 1. John Wiley & Sons, Chichester, pp 271–277
86. Libes D (1995) *Exploring Expect: A Tcl-Based Toolkit for Automating Interactive Programs*. O'Reilly & Associates Inc., Sebastopol, CA, USA
87. Ousterhout JK (1994) *Tcl and the Tk toolkit*. Addison-Wesley Publishing Company, Reading, MA, USA
88. Elber R, Karplus M (1990) *J Am Chem Soc* 112:9161–9175
89. Tidor B, Karplus M (1991) *Biochemistry* 30:3217–3228
90. Stultz CM, Karplus M (1999) *Proteins* 37(4):512–529
91. Karpen ME, Tobias DJ, Brooks CL III (1993) *Biochemistry* 32:412–420
92. Caffisch A, Karplus M (1995) *Persp Drug Discov Des* 3:51–84
93. Viswanadhan VN, Ghose AK, Chandra Singh U, Wendoloski JJ (1999) *J Chem Inf Comput Sci* 39:405–412
94. Sitkoff D, Sharp KA, Honig B (1994) *J Phys Chem* 98:1978–1988
95. Cabani S, Gianni P, Mollica V, Lepori L (1981) *J Sol Chem* 10(8):563–595
96. Jorgensen WL, Gao J, Ravimohan C (1985) *J Phys Chem* 89:3470–3473
97. Jorgensen WL, Gao J (1986) *J Phys Chem* 90:2174–2182
98. Kubo MM, Gallicchio E, Levy RM (1997) *J Phys Chem B* 101:10527–10534
99. Della Gatta G, Barone G, Elia G (1986) *J Sol Chem* 15(2):157–167
100. Dec SF, Gill SJ (1984) *J Sol Chem* 13(1):27–41
101. Mestres J, Rohrer DC, Maggiora GM (1997) *J Mol Graphics Mod* 15:114–121
102. Chapman D (1996) *J Comp Aided Mol Des* 10:501–512
103. Cunningham BR, Rivetna M, Tolman RL, Flattery SJ, Nichols A, Elizabeth, Schwartz CD, Wicker LS, Hermes JD, Jones AB (1997) *Bioorg Med Chem Lett* 7(1):19–24
104. Halgren TA (1996) *J Comp Chem* 17(5–6):490–519
105. Furka Á, Sebestyén F, Asgedom M, Dibó G (1991) *Int J Peptide Protein Res* 37:487–493
106. Combs AP, Kapoor TM, Feng S, Chen JK, Daudé-Snow L, Schreiber SL (1996) *J Am Chem Soc* 118:287–288
107. Morken JP, Kapoor TM, Feng S, Shirai F, Schreiber SL (1998) *J Am Chem Soc* 120:30–36
108. Kozono H, White J, Clements J, Marrack P, Kappler J (1994) *Nature* 369:151–154
109. Wallace AC, Laskowski RA, Thornton JM (1995) *Prot Eng* 8:127–134