OptCDR: a general computational method for the design of antibody complementarity determining regions for targeted epitope binding

R.J. Pantazes and C.D. Maranas[1]

Department of Chemical Engineering, The Pennsylvania State University, University Park, PA 16802, USA

[1]To whom correspondence should be addressed. E-mail: costas@psu.edu

Edited by Anthony Rees

Antibodies are an important class of proteins with many biomedical and biotechnical applications. Although there are a plethora of experimental techniques geared toward their efficient production, there is a paucity of computational methods for their *de novo* design. OptCDR is a general computational method to design the binding portions of antibodies to have high specificity and affinity against any targeted epitope of an antigen. First, combinations of canonical structures for the antibody complementarity determining regions (CDRs) that are most likely to be able to favorably bind the antigen are selected. This is followed by the simultaneous refinement of the CDR structures' backbones and optimal amino acid selection for each position. OptCDR is applied to three computational test cases: a peptide from the capsid of hepatitis C, the hapten fluorescein and the protein vascular endothelial growth factor. The results demonstrate that OptCDR can efficiently generate diverse antibody libraries of a pre-specified size with promising antigen affinity potential as exemplified by computationally derived binding metrics.

Keywords: antibody design/computational protein design/ fluorescein/hepatitis C/vascular endothelial growth factor

Introduction

Antibodies are proteins in vertebrate immune systems that are able to bind a diverse set of molecules, ranging from proteins and peptides to haptens (i.e. small molecules), with high specificity and affinity. They are composed of pairs of heavy and light chains folded into the well-known 'Y' shape. Two identical heavy chains form the stem and each branch is constructed from a light chain and the end of one of the heavy chains. The entire stem and approximately half of each branch are collectively referred to as the constant region while the ends of each branch are the variable regions. Each variable region is composed of a highly conserved framework and six complementarity determining regions (CDRs), also known as hypervariable regions. The six CDRs, three on the light chain (L1, L2 and L3) and three on the heavy chain (H1, H2 and H3), are known to be responsible for the majority of antibody-binding interactions. Humanization (Almagro and Fransson, 2008) is a common experimental technique where the CDRs from a non-human antibody are attached to the framework of a human antibody, thereby retaining the binding properties of the non-human antibody while decreasing or eliminating its immunogenicity. It is well established that for all CDRs, except H3, there are discrete sets of conformations that their secondary structures assume, known as canonical structures (Chothia and Lesk, 1987). Notably, there is a subset of antibodies found in camelids that lack light chains and have shorter than normal heavy chains (Hamers-Casterman *et al.*, 1993).

Ever since antibody-producing mouse B cells were first immortalized by fusion with cancer cells to create hybridomas (Kohler and Milstein, 1975), antibodies have been extensively used in a variety of contexts. Experimentally, their ability to recognize molecules with high specificity and affinity has been leveraged in popular assays such as ELISA (Reen, 1994) and ELISPOT (Czerkinsky *et al.*, 1983). Medicinally, they are a profitable drug class, with the five top-selling antibodies [Rituxan (Plosker and Figgitt, 2003), Remicade (Feldmann and Maini, 2001), Herceptin (Vogel *et al.*, 2001), Humira (Kempeni, 1999) and Avastin (Chen *et al.*, 2001)] having revenues in excess of $11 billion in 2006 (Dimitrov and Marks, 2009).

Therefore, it is not surprising that over the years, there have been extensive efforts geared toward designing antibodies and libraries thereof. A number of experimental techniques (Hanes and Pluckthun, 1997; Chen *et al.*, 1999; Boder *et al.*, 2000; De Pascalis *et al.*, 2003; Rajpal *et al.*, 2005; Reiersen *et al.*, 2005; Fukuda *et al.*, 2006; Walker *et al.*, 2009) have been developed and successfully applied to design antibodies to bind desired antigens or to improve the binding characteristics of an existing antibody. For example, Chen *et al.* (1999) used monovalent phage display to increase the affinity of an antibody against vascular endothelial growth factor (VEGF) by 100-fold. While libraries have been computationally designed for other protein targets (Meyer *et al.*, 2003; Saraf *et al.*, 2004; Pantazes *et al.*, 2007), computational antibody design methods have so far focused on altering existing antibodies to improve their properties (Clark *et al.*, 2006; Lippow *et al.*, 2007; Barderas *et al.*, 2008; Clark *et al.*, 2009). Two notable examples are the use of a variety of computational methods to predict beneficial mutations for an antibody–antigen complex (Clark *et al.*, 2006) and the rational redesign of an anti-VLA1 antibody through modification of a single CDR (Clark *et al.*, 2009). Other computational efforts have focused on predicting antibody structures from only their amino acid sequences (Morea *et al.*, 2000; Whitelegg and Rees, 2000; Sivasubramanian *et al.*, 2009).

Existing experimental and computational antibody design methods have made key contributions; however, there is still

a need for a general computational method that can rapidly design libraries of antibodies to bind difficult-to-target epitopes. The proposed general computational method, Optimal Complementarity Determining Regions or OptCDR, is used for the *de novo* design of novel antibodies to bind any selected antigen with high affinity and specificity. OptCDR expediently generates multiple distinct and original libraries of user-specified size of antibody CDRs that can be grafted into the framework of an antibody using standard humanization techniques (Almagro and Fransson, 2008). Benchmarking against three test cases reveals OptCDR's ability to expediently generate multiple original and diverse antibody libraries with computational binding metrics at least as good as the ones for naturally occurring antibodies.

Methods

OptCDR is a four-step workflow to design libraries of antibody CDRs to bind a specified antigen epitope. The first step is the selection of the combination of CDR canonical structures (only backbones) that are most likely to have favorable binding with the antigen. During the second step, the amino acid sequences of the selected structures are initialized one at a time using a rotamer library, energy functions, and a mixed-integer linear programming (MILP) optimization formulation. This is followed by several thousand iterations of a modified version of the previously developed iterative protein redesign and optimization (IPRO) (Saraf *et al*., 2006) procedure, which simultaneously refines the backbone structures and amino acid sequences of the CDRs. The fourth step of OptCDR is library generation by accumulation of the most-promising mutations for the CDRs.

Canonical structure identification

The first step of OptCDR requires the use of a library of canonical structures that spans the full range of structural diversity exhibited by CDRs in known antibodies. This library was constructed by performing a clustering analysis on the CDRs identified from 883 antibody structures downloaded from the Protein Data Bank (PDB) (Berman *et al*., 2000). Sequence-based rules, given in Table I, were employed to identify the CDRs in each antibody. The rules were based on those from

the Web Antibody Modeling (WAM) (Whitelegg and Rees, 2000) method and then expanded to include amino acids with similar chemical properties. When these sequence-based rules were not sufficient (∼25% of cases), the framework of the antibody variable region was aligned to a consensus framework and structural analysis was used to identify the CDRs. Through utilization of sequence and structural analyses, at least one CDR was identified in 833 of the 883 structures and all six in 700 of them.

A clustering procedure inspired by the work of Martin and Thornton (1996) was used to group similar CDRs into canonical clusters. At the end of the clustering, a canonical structure was selected from each cluster as the structure with the smallest deviations in the sines and cosines of its phi, psi and omega dihedral angles from the averages of the cluster. The clustering was carried out such that all members of a cluster had a backbone atom (N, Cα and C) RMSD of no more than 1.5 Å from the canonical structure. A number of modifications were made *a posteriori* to resolve structural inconsistencies, such as improper attachment to the variable region framework or significant clashing (i.e. at least one pair of atoms less than the sum of the Van der Waals radii apart) with other canonical structures. Clusters containing only one or two members for all CDRs except H3 were discarded. Ultimately, this process resulted in 13 L1, 1 L2, 9 L3, 8 H1, 4 H2 and 124 H3 CDR canonical structures. The four clusters of H2 CDRs are shown in Fig. 1 as an example of how similar structures are grouped together and dissimilar structures are disaggregated.

Step 1: selection of CDR canonical structures. Canonical structures for CDRs describe only their backbone conformation with no information about the side chains. Owing to the absence of residue type information, we developed a scoring system to distinguish between structures that have the potential to exhibit favorable binding with the antigen using only backbone atom–antigen distance information. The hypothesis is that the antigen should be placed within reach of the antibody CDRs (i.e. <8 Å) but avoid detrimental clashes (i.e. closer than the sum of two atoms Van der Waals radii). Specifically, if the distance between a backbone atom in the structure and the closest atom in the antigen is less

Table I. The sequence-based rules for identifying CDRs

CDR loop	Starting position	Preceded by	Length	Followed by
L1	1 residue after first CYS	CYS	10–17 residues	TRP - TYR/LEU/PHE/VAL - GLN/LEU/PHE
L2	14–19 residues after L1	ILE/VAL/MET - TYR/LYS/PHE/HIS/GLY	7 residues	GLY/GLU
L3	1 residue after first CYS after L2	CYS	7–13 residues	PHE - GLY - XXX - GLY
H1	4 residues after first CYS	CYS - XXX - XXX - XXX	10–12 residues	TRP - ILE/VAL/MET/ALA/PHE - ARG/LYS - GLN/ASN/LYS/GLU/HIS
H2	Approximately 15 residues after H1	TRP/TYR/LEU - ILE/LEU/VAL/MET - GLY/ALA/SER	9–15 residues	(A) TYR - XXX - XXX - XXX - XXX - LYS (B) TYR/PHE - XXX - XXX - XXX - XXX - GLN/LYS/ARG
H3	3 residues after CYS. Approximately 40 residues after H2	CYS - XXX - XXX	3–22 residues	TRP - GLY - XXX - GLY

XXX corresponds to any permitted amino acid. '/' separate amino acids that are acceptable at a given position while '-' distinguish between different positions. For the end of the H2 CDR, rule (A) was applied first and if that failed then rule (B) was used.
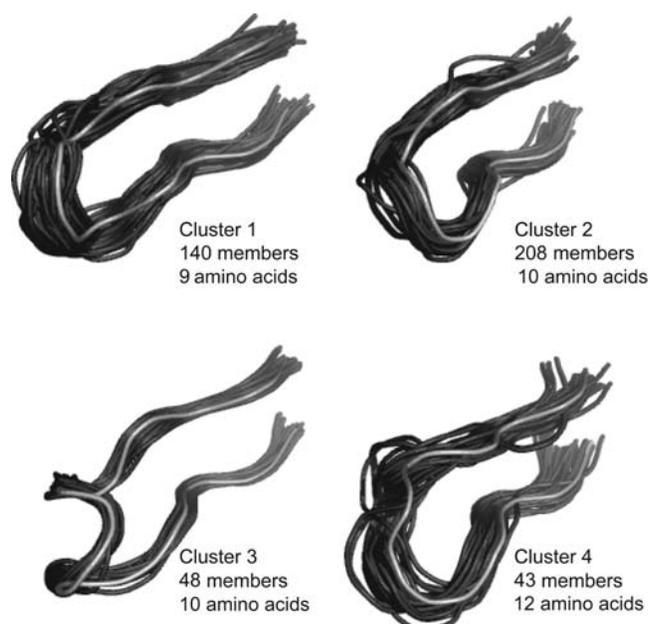
Fig. 1. The four clusters of H2 CDRs. The members of the clusters are shown in dark gray, and the canonical structure of each cluster is shown in light gray. All members of each cluster have an RMSD of no more than 1.5 Å with the canonical structure of that cluster.

than the sum of their Van der Waals radii, then the contact is penalized with a negative score. Alternatively, if the two atoms are more than 8 Å apart, then the contact is considered to be unlikely to contribute to binding and a zero score is assigned. Only if the distance is between the sum of the two atoms Van der Waals radii and 8 Å is a positive score assigned. A penalty score of -5 for steric clashes and a reward score of $+1$ for potential to contribute favorably to binding were used. The score for a given canonical structure is computed as the sum of the scores of the atoms in its backbone. It is worth noting that all steric clashes between the backbone atoms of the CDRs and the antigens are equally penalized. The reason for this uniform treatment of clashes is that if an antigen is clashing with the backbone atoms of a CDR, then most likely there would be significant clashes with the side chain.

The problem of selecting the highest scoring combination of non-clashing canonical structures can be mathematically represented using an MILP representation. This requires the definition of the index set $I = \{i \mid L1, L2, L3, H1, H2, H3\}$ denoting the six CDRs and sets $C^i = \{c \mid 1, \ldots, C^i\}$ encoding the number of canonical structures for a given CDR (i.e. 13, 1, 9, 8, 4 and 124, respectively). Set IC^{clash} contains all pairwise canonical structure combinations (i_1, c_1) and (i_2, c_2) that share at least one pair of atoms that are closer than the sum of the two atoms Van der Waals radii (steric clash). The importance of excluding sterically clashing canonical structure pairs is bolstered by the fact that out of 521 antibody structures with resolution no worse than 2.5 Å, there were only two such clashing pairs (PDB codes 1LGV, 1OCW) and both such clashes were small (<0.2 Å). Parameter $S_{i,c}$ encodes the score contribution of the cth structure of the ith CDR. Binary variable $y_{i,c}$ encodes which structure c is selected for the ith CDR. Specifically, $y_{i,c}$ is equal to one if structure c has been selected for CDR i and zero otherwise.

The resulting MILP optimization problem is posed as:

$$\text{Maximize} \sum_{i=1}^{6} \sum_{c=1}^{C} y_{i,c} S_{i,c} \qquad (1)$$

$$\text{Subject to} : \sum_{c=1}^{C} y_{i,c} = 1, \quad \forall i \in I \qquad (2)$$

$$y_{i1,c1} + y_{i2,c2} \le 1, \quad \forall (i1, c1, i2, c2) \in IC^{\text{clash}} \qquad (3)$$

Equality constraint (2) ensures that exactly one structure is selected for each CDR while inequalities (3) preclude the simultaneous presence of two canonical structures that sterically clash. The optimization formulation described collectively by Equations (1–3) is solved to global optimality using CPLEX version 11 (ILOG, 2007) accessed in the GAMS modeling environment. The solution of the above MILP formulation yields the highest scoring combination of canonical structures for a given position of the antigen.

The scoring system was tested on 254 native antibody–antigen complexes with resolutions no worse than 2.5 Å in which all six CDRs had been identified and 25 400 decoy complexes. The decoy complexes were generated by randomly changing the CDR canonical structures of the native complexes 100 times each. As shown in Fig. 2, the distribution of scores for the native complexes was notably better than those for the decoy complexes. Using a cutoff score of 52, the scoring system achieves a sensitivity of 85% and a specificity of 80%, where sensitivity is the percentage of native complexes above the cutoff and specificity is the percentage of decoy complexes below the cutoff. Notably, a slightly better separation between native and decoy complexes could be achieved with a cutoff of 10 Å, but it was not adopted because it created an abnormal bias for longer canonical structures. The highest scoring combination of CDR canonical structures was also identified for each one of the 254 antibody–antigen complexes. On average, the native complexes scored within 23% of the highest scoring combination of CDR canonical structures for their antigen positions and in 10 cases, the specific CDRs of an antibody scored better than the best combination of canonical structures. This is consistent with the expectation that native antibodies are generally very good but rarely optimal at recognizing a particular antigen.

Proper position of the antigen in the antibody binding pocket is addressed by iteratively solving the MILP for different antigen locations. For each type of antigen (hapten, peptide and protein), average Cartesian coordinates and standard deviations for the center of mass of the portion of the antigen being bound were calculated. Several thousand antigen positions and orientations are randomly generated with normally distributed departures for the three translational and uniformly distributed departures for the three rotational degrees of freedom. When it is desired to target a specific epitope, the rotational degrees of freedom can be limited to ensure that the epitope is the only portion of the antigen interacting with the CDRs. Afterwards, several antigen positions with the highest scoring combinations of
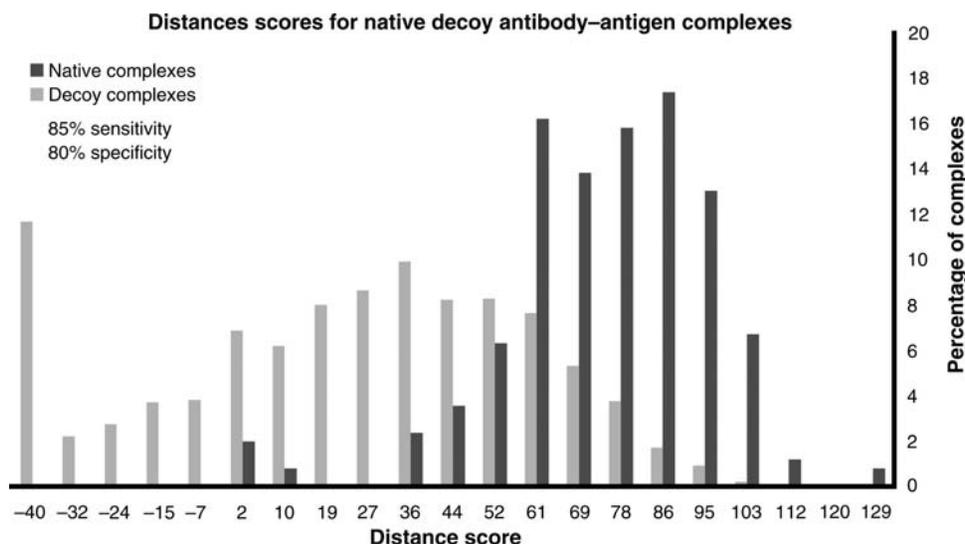
Fig. 2. The score distributions of native and decoy antibody–antigen complexes. Using the scoring function, 254 native and 25 400 decoy antibody–antigen complexes were examined and good separation between them was observed. Using a cutoff of 52, a sensitivity of 85% and a specificity of 80% were achieved, where sensitivity is the percentage of native complexes above the cutoff and specificity is the percentage of decoy complexes below it.

canonical structures are retained. It is well established that antibodies typically form differently shaped binding pockets depending on the type of antigen: flat surfaces for proteins, grooves for peptides and cavities for haptens (MacCallum et al., 1996; Collis et al., 2003; Persson et al., 2006). Interestingly, the first step of OptCDR recapitulates these types of binding pockets for the different classes of antigens without the inclusion of any specific restraints to bias the selection toward these types of pockets.

Step 2: initialization of CDR amino acid sequences. The second step of OptCDR is the initialization of the amino acid sequences of the selected canonical structures one at a time in an order that is indicative of their typical proximity to antigens (e.g. H3, L3, H2, H1, L1 and L2) and can be specified by the user of OptCDR. This is done using energy functions [Van der Waals and electrostatics from CHARMM (MacKerell et al., 1998) and Lazaridis–Karplus solvation (Lazaridis and Karplus, 1999)], a rotamer library and a previously presented MILP rotamer optimization formulation (Saraf et al., 2006) that selects the lowest-energy combination of rotamers. Three additional sets of constraints are implemented in the MILP to ensure that the amino acids selected are consistent with the chosen canonical structures.

The first two sets of constraints limit the total percentage usage of each amino acid and all charged amino acids in a given CDR to below one standard deviation greater than their average percentage usage in the downloaded database of natural antibodies. The third set of constraints limits the specific amino acids permitted at each position in a canonical structure. If a canonical structure was observed at least 30 times, then only the amino acids that were encountered before at a given position are permitted. If a canonical structure was observed fewer than 30 times, then only the amino acids of the same 'type' are permitted. The chemical types used are charged (D, E, K, R and H), aliphatic (G, L, I, A, V and M), aromatic (F, W and Y) and polar (S, T, C, N and Q). Proline is treated as an aliphatic residue, but no rotamers of

it exist in the rotamer library, so it is never a selected amino acid.

After the amino acid sequences of all six CDRs have been initialized, a rigid-body docking step is carried out to refine the position of the antigen in the antibody binding pocket. This is followed by sequential solution of the MILP optimization problem for each CDR in the same order as before and then another rigid-body docking procedure.

Step 3: structure and sequence refinement with IPRO. In the third step of OptCDR, several thousand iterations of a modified version of IPRO (Saraf et al., 2006; Fazelinia et al., 2007) are carried out. In each iteration, a CDR is randomly selected and the phi and psi dihedral angles of each residue in the CDR are perturbed in CHARMM (MacKerell et al., 1998). Next, the amino acid sequence of the CDR is re-determined using the rotamer selection MILP. This is followed by an energy minimization in CHARMM, during which weak harmonic constraints (0.05 kcal/mol) are used on all CDRs except H3 to ensure that their canonical structures are not changed. During this energy minimization, similar constraints may also be used on the antigen to prevent or allow antigen conformational rearrangements. The Metropolis criterion, as in simulated annealing, is used to determine whether or not to retain the results of the iteration. If they are retained, a rigid-body docking procedure is carried out followed by another CHARMM energy minimization. If the post-docking interaction energy is lower than the pre-docking energy, then the post-docking results are retained. Otherwise, the pre-docking results are kept and the retained structures serve as the starting points for the next iteration. Over the course of many iterations, IPRO progressively identifies sets of backbone perturbations and amino acid sequence mutations that improve the interaction energy between the CDRs and the antigen. The results from applying IPRO to specific, experimentally examined systems are discussed in the Results section. IPRO has previously been used to design several protein-substrate systems that were experimentally verified to have the desired properties

(Fazelinia *et al*., 2009; Khoury *et al*., 2009). The concept of iterating between structure perturbation and sequence/rotamer redesign was pioneered by Kuhlman *et al*. (2003) in the design of Top7, a computationally designed protein with a novel fold. The backbone perturbation followed by optimal rotamer selection allows IPRO to simultaneously mutate every amino acid in a CDR. This can uncover multiple non-additive mutations that would be unlikely to be sampled using directed evolution experiments. The use of simulated annealing in IPRO safeguards against being trapped in local energy minima, which allows for the identification of multiple solutions with very different sequences.

Step 4: library generation. In the final step of OptCDR, the rotamer selection MILP is again employed to identify multiple low-energy amino acid sequences for each CDR. Starting with the final structures from step 3, the MILP formulation is solved multiple times for each CDR to identify not only the best but also the second, third, etc. best solutions by accumulating integer cuts that exclude previously visited solutions. This identifies multiple amino acid sequences for each CDR and a library of targeted size can be constructed from the lowest-energy combinations of CDRs. By using these four steps, OptCDR can expediently generate multiple novel and diverse libraries of antibody CDRs to bind any specified antigen.

Energy function testing

As a validation of the energy functions used in OptCDR, two tests were carried out. First, we used IPRO to examine the 95 mutants to an anti-VLA1 antibody (PDB: 1MHP) that had been computationally predicted to improve affinity with a 12% experimental success rate (Clark *et al*., 2006). Following the experimental verification step, the authors *a posteriori* refined their methods, so that 26% of their predicted favorable mutants were correct. Using the rotamer selection and energy minimization using IPRO, we calculated the change in interaction energy compared with wild-type for all 95 mutants. We observed an overall accuracy of 78% at predicting the experimental effect of the mutations and 50% of our predicted favorable mutants were in agreement with the experiment. These results demonstrate the efficacy of the energy functions used within OptCDR to distinguish between mutants that improve/decrease affinity.

As a follow-up, we tested the ability of OptCDR to pinpoint the correct amino acid sequence for a given CDR structure. We chose 38 high-resolution antibody–antigen complexes and attempted to find even lower energy amino acid sequences for their CDRs. For each antibody, we used IPRO's energy minimization and energy calculation functions to refine the initial crystal structures and determine the wild-type CDR-antigen interaction energies. Subsequently, the rotamer selection MILP was used to generate the five lowest-energy, non-wild-type amino acid sequences for each CDR in each complex. The corresponding rotamers were patched into the CDRs' backbones and IPRO's energy minimization and energy calculation functions were used to determine the interaction energy of the predicted low-energy CDRs. For 67% of the CDRs, no sequence with a lower energy than the native sequence could be found. Only 24% of the non-native CDR amino acid sequences had lower interaction energies than the native sequences, and the differences between the improved non-native and native sequences were minor (<9%). These results indicate that OptCDR recognizes the wild-type sequences known to lead to effective binding, for a given antigen and CDR combination, as the best or near-best binders.

Results

Three systems were selected to test OptCDR's efficacy: a peptide from the capsid of hepatitis C (PDB: 1N64) (Menez *et al*., 2003), fluorescein (PDB: 1FLR) (Whitlow *et al*., 1995) and VEGF (PDB: 1CZ8) (Chen *et al*., 1999). Several computational metrics are used to draw comparisons, including interaction energy defined as the minimized energy of the antigen–CDRs complex minus the energy of the CDRs and the energy of the antigen individually. It is approximated

Table II. Computational and experimental binding data for the antibodies

Antibody	Antigen	Interaction energy (kcal/mol)	Contacts	Polar contacts	Experimental K_d
19D9D6	Hepatitis C capsid peptide	−62.6	74	8	1.3 nM
IPRO affinity maturation of 19D9D6	Hepatitis C capsid peptide	−78.1 to −80.1	81–87	14–15	NA
OptCDR designs	Hepatitis C capsid peptide	−88.2 to −104.8	88–115	18–23	NA
OptCDR design with antigen rearrangment	Hepatitis C capsid peptide	−123.6 to −175.8	88–112	23–31	NA
4-4-20	Fluorescein	−49.5	22	4	0.7 nM
Boder *et al*. best design	Fluorescein	−78.7	23	4	48 fM
Fukuda *et al*. consensus design	Fluorescein	−67.8	22	5	~1 nM (wt 32 nM)
Fukuda *et al*. best design	Fluorescein	−70.4	24	3	0.88 nM (wt 32 nM)
Jermutus *et al*. consensus design	Fluorescein	−74.9	24	3	~37 pM
IPRO affinity maturation of 4-4-20	Fluorescein	−77.8	17	1	NA
OptCDR designs	Fluorescein	−55.2 to −57.3	71–79	8–9	NA
PDB 1CZ8	VEGF-epitope 1	−110.8	86	21	0.11 nM
IPRO affinity maturation of 1CZ8	VEGF-epitope 1	−111.0 to −116.0	89–99	21–22	NA
OptCDR design	VEGF-epitope 1	−82.0 to −109.0	73–110	11–25	NA
OptCDR design	VEGF-epitope 2	−88.6 to −98.4	57–85	15 to 20	NA
OptCDR nanobodies	VEGF-epitope 1	−82.2 to −92.6	79–86	20–22	NA

The various computational binding metrics are defined in the text at the start of the Results Section and all experimental K_d values were taken from the appropriate publications. VEGF-epitope 1 is the epitope bound by bevacizumab while epitope 2 is on the opposite side of VEGF. The nanobody designs have only three CDRs, all other antibodies have six.
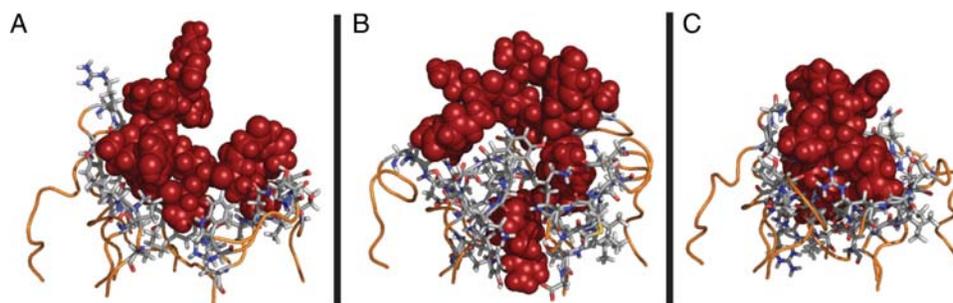
Fig. 3. CDR–hepatitis C capsid peptide complexes. The capsid peptides are shown as red spheres, the CDRs are shown as orange ribbons, and CDR residues within 4 Å of the peptide are explicitly shown. All images are from the same perspective. (**A**) The natural antibody–peptide complex (PDB: 1N64). (**B**) An OptCDR design with no peptide conformational change. (**C**) An OptCDR design with peptide conformational change.

Table III. A library of hepatitis C capsid binding antibody CDRs

L1	L2	L3	H1	H2	H3
QGTTQKMVAASGMD	GGTSRRT	TQYSTWGGKYQ	GFSIRRSNVGVT	EIGSRGNNYTMK	ARSGKTQEAQGTRGGWMDEDDN
SQN M YD A	SN KS	NS Q F	LT D W	DF K F T	DKA EQTISA KT NQKR QY
N F					E NM M AD M
					N

The CDR sequences and predicted mutations to them to form a library of up to 1.1×10^{14} antibodies that can all bind the peptide from the capsid of hepatitis C.

within CHARMM (MacKerell *et al.*, 1998) using the Van der Waals, electrostatics, bonds, angles, dihedral angles, improper dihedral angles and generalized Born with molecular volume integration implicit solvation energy functions. Contacts are defined as the number of CDR atoms within 3 Å of the antigen and polar contacts are determined using PyMOL (DeLano, 2008). All computations were carried out on 3.0 GHz Intel Xeon processors with 4 GB of RAM. Each complete antibody library design was generated on its own processor and all were completed in <12 days of computations, with an average of about 9 days.

Hepatitis C capsid peptide

Hepatitis C is a virus that infects approximately 3.2 million people in the USA (http://www.cdc.gov/hepatitis/C/cFAQ. htm#statistics). The CDRs of antibody 19D9D6 (PDB: 1N64) (Menez *et al.*, 2003) that bind a peptide (residues 13–40) from the capsid of the hepatitis C virus with a K_d of 1.3 ± 0.1 nM are shown in Fig. 3A. This system was selected as a general test of OptCDR to generate promising design alternatives. We first examined the extent of improvements that can be achieved in the computationally accessible metrics of binding quality by only mutating the original antibody structure without altering the CDR canonical structures. The results obtained using IPRO show improvements in all three binding metrics (Table II). Interestingly, the mutations identified are confined to the CDRs with the fewest antigen contacts. This trend of predicting mutations in the periphery of the antibody binding pocket is consistent with a previous computational study that was experimentally validated (Lippow *et al.*, 2007). In this case, since 19D9D6 is already a high-affinity antibody, the results indicate that the dominant interactions in the center of the binding pocket are already effective and binding could only be improved further through repacking of the edges of the antibody–antigen interface.

We next used OptCDR to design three sets of antibody CDRs to bind the peptide instead of relying on only adding

point mutations to 19D9D6. We first assumed a conservative posture by imposing harmonic constraints that insured that the antigen conformation did not change significantly upon binding. The three generated designs exhibit highly diverse antigen locations/orientations, canonical structure selections and amino acid sequences, but all share the groove that is typically observed for peptide-binding antibodies. Significant improvements in all computational binding metrics are observed (Table II) over the case of using only mutations. Table III depicts the predicted lowest-energy amino acid sequences for the CDRs of the structure shown in Fig. 3B. By combining the predicted mutations, a library of CDRs can be generated (maximum library size = 1.1×10^{14}). By prioritizing mutations based on their binding scores, libraries of any smaller size can be culled from the original.

Finally, we removed the harmonic constraints on the antigen, allowing for its conformation to radically change in response to the interaction energy minimization step, and used OptCDR to generate three additional sets of antibody CDRs to bind the peptide. As seen in Table II, allowing conformational changes to the peptide in response to energy minimization led to additional improvements in interaction energy and polar contacts, although no effect on the number of contacts. It should be noted that the conformational changes were not forced or pre-specified, but came about as a response to interactions with the CDRs during the IPRO step (i.e. step 3) of OptCDR. In this case, the conformational changes were all between 4.2 and 5.2 Å RMSD from the initial peptide conformation. These results demonstrate how antigen conformational changes upon antibody binding may be an important contributor in informing antibody–antigen interactions. OptCDR allows for a user-specified presence, absence and modulation of the strength of any imposed harmonic constraints.

Fluorescein

We next turned our attention to the hapten fluorescein for a comparison with experimental results. Anti-fluorescein
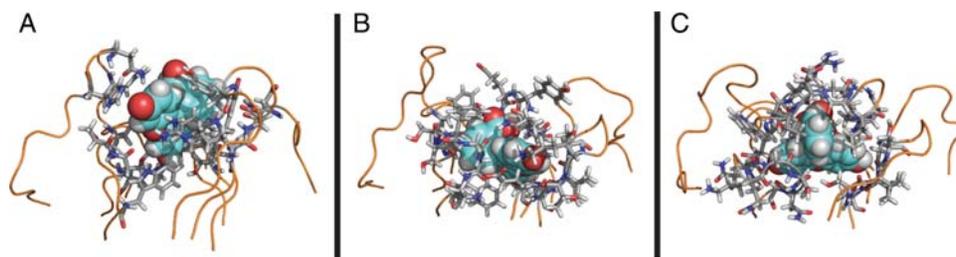
Fig. 4. CDR–fluorescein complexes. Fluorescein is shown as cyan spheres with oxygens and hydrogens shown in red and white, respectively. The CDRs are shown as orange ribbons and all CDR residues within 4 Å of fluorescein are explicitly shown. All the complexes are shown from the same perspective relative to the antibody binding pocket. (**A**) The structure of antibody 4-4-20. (**B** and **C**) The two OptCDR fluorescein-binding designs.

Table IV. Two libraries of fluorescein-binding antibody CDRs

	L1	L2	L3	H1	H2	H3
Design 1	ETQSSGGSNTDKSWVE S N Q M T N I	GGTKSRT S S N A	TQWKEQGSNGA N S AG A N M	EVTGTQFGVN A N F Y Y A	SVSGDAKTG I AGGNN D K	KKNGMAEGDTEGGRDSSTNRAQ D A ST NDD QLAANKQEE S D DD E N DQ SN MR L
Design 2	ESSQRLVHGSGKTFMS AT N LDKN S	EGTNSKG N T A Y S	QNFSSYGGKFQ N Y T N	GFSITRTNEGVS L Y W F	GVSSTGKTE I AS N	AGGQYESASKKE S N A TR

The amino acid sequences of the two OptCDR designed fluorescein-binding antibodies and the predicted mutations to the CDRs to form libraries of antibodies.

antibodies have been used as the test system by several experimental directed evolution efforts (Boder *et al.*, 2000; Jermutus *et al.*, 2001; Fukuda *et al.*, 2006) with different display technologies to improve the binding of an scFv derived from the anti-fluorescein antibody 4-4-20. This is a particularly interesting system because 4-4-20's affinity for fluorescein is near the affinity ceiling of the tertiary immune response (0.7 ± 0.3 nM) (Boder *et al.*, 2000). Boder *et al.* (2000) used yeast surface display to identify an scFv with 14 mutations exhibiting a ∼6500-fold improvement in K_d, while Fukuda *et al.* (2006) and Jermutus *et al.* (2001) used mRNA and ribosome display, respectively, to identify mutants with ∼30-fold improvement in K_d.

First, we used IPRO to evaluate these experimentally identified mutants: the best mutants identified by Boder *et al.* (2000) and Fukuda *et al.* (2006) and the consensus mutants identified by Fukuda *et al.* (2006) and Jermutus *et al.* (2001). Since some of the mutations were in framework regions, the entire variable domains were used to create the mutants, not just the CDRs. SwissParam (http://swissparam.ch/) was used to create the topology and parameter files needed in CHARMM for fluorescein. Once the mutant antibodies were modeled, we extracted their CDRs, so the calculated interaction energies could be directly compared with OptCDR results. Table II shows the computational and experimental improvements over the wild-type scFv of the four mutants. Note that the rank order of the mutants in terms of improvement over wild-type using K_d and interaction energy, as quantified in OptCDR, match. Furthermore, the H3 CDR in the Jermutus *et al.* (2001) mutant shows the highest RMSD from the wild-type structure, which matched the experimental observation of increased flexibility of this CDR due to the removal of a salt bridge.

Next, IPRO was used to computationally affinity mature the CDRs of antibody 4-4-20, leading to numerous mutations

(40 total) in all CDRs except H2. The computational binding results are detailed in Table II. Although the improvement in interaction energy over the wild-type antibody is not quite as large as that of the best experimental mutant, it does surpass the calculated energies for all other mutants. We also used OptCDR to design two sets of CDRs to bind fluorescein. Their computational binding metrics are given in Table II, structures in Fig. 4 and amino acid sequences in Table IV. Interestingly, the interaction energies of the OptCDR designs do not reach the same levels as those of the computationally and experimentally affinity-matured versions of 4-4-20, but they do surpass the wild-type antibody. Both designs share a number of features that appear favorable to binding. First, in both cases, fluorescein is positioned within a deep cavity between the L3 and H2 CDRs on one side and the H3 CDR on the other. Both designs have long H3 CDRs folded mostly over the top of the fluorescein molecules to trap them in place, and it is this position of the H3 CDRs that lead to the notable increase in contacts between the experimental and OptCDR designs (Table II). For both designs, the edges of the cavity have polar residues with each fluorescein oxygen involved in at least one polar contact. Finally, the sides of both cavities are composed of aliphatic and aromatic residues stabilizing the core hydrophobic portion of fluorescein. We hypothesize that the H3 CDRs of the unbound designs are sufficiently flexible to allow fluorescein access to the binding pocket.

Vascular endothelial growth factor

Finally, OptCDR designs for binding VEGF are contrasted against an affinity-matured antecedent of the antibody medication bevacizumab (Chen *et al.*, 2001) to examine OptCDR's epitope targeting abilities. VEGF has been shown (Willett *et al.*, 2004) to promote tumor proliferation and growth. A
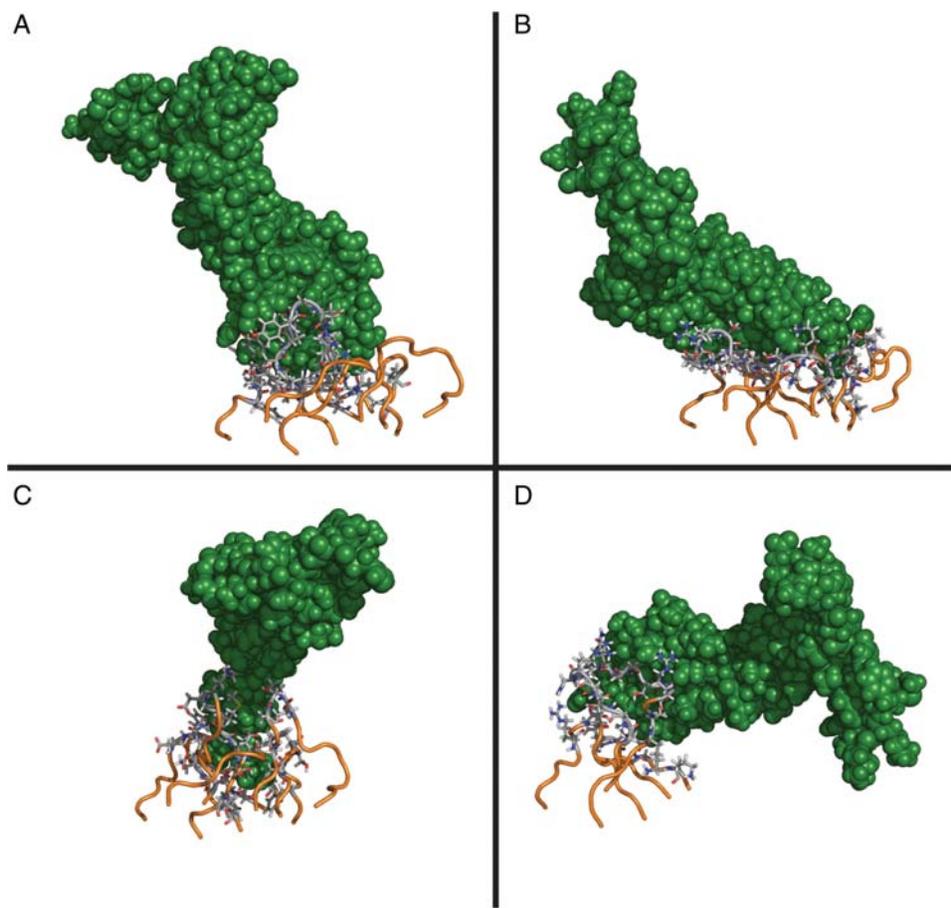
Fig. 5. CDR–VEGF complexes. VEGF is shown as green spheres, the CDRs are shown as orange ribbons, and CDR residues that are within 4 Å of VEGF are explicitly shown. All images are from the same perspective relative to the antibody binding pocket. (**A**) The structure of PDB 1CZ8. (**B**) The best OptCDR design generated to bind the same portion of VEGF as bevacizumab using all six CDRs while (**C**) is the best design to bind an epitope on the opposite side of VEGF. (**D**) The best nanobody OptCDR design.

number of anti-VEGF antibody-based drugs, including bevaci-zumab (Chen *et al.*, 2001), with low nanomolar affinity are available. The resolved structure (PDB: 1CZ8) of an affinity-matured bevacizumab antecedent shows VEGF situated within a pocket primarily formed by the heavy chain CDRs (Chen *et al.*, 1999). In all OptCDR results, a harmonic constraint was used to prevent the structure of VEGF from changing significantly during calculations.

IPRO-based computational affinity maturation of 1CZ8 led to only minimal improvements in the interaction energy, as well as the number of contacts and polar contacts, which is a testament to the thoroughness of the experimental affinity maturation that 1CZ8 has already undergone. When we used OptCDR to predict novel CDRs to bind the same epitope of VEGF targeted by 1CZ8, the best design, shown in Fig. 5B, has binding metrics that are comparable with the existing antibody (slightly greater interaction energy and a few more contacts and polar contacts). Thus, the range of binding metrics for OptCDR (Table II) only reaches the levels of 1CZ8, although the predicted structures are all notably different and most of which exhibit the planar binding pocket expected for protein-binding antibodies.

Next, we explored another set of OptCDR designs by targeting an epitope on the opposite side of VEGF from the portion bound by 1CZ8 and our other designs. To the best of

our knowledge, this portion of VEGF is not recognized by cellular VEGF receptors or any designed antibodies. One of the predicted designs is shown in Fig. 5C and the computational binding metrics are given in Table II. Even though the obtained designs are very different as they target a completely different epitope of VEGF, the computational binding metrics achieved are quite similar in value. The results obtained demonstrate the efficacy of OptCDR to generate designs that bind VEGF with equivalent computational binding characteristics as the wild-type antibody with novel CDRs or targeting a different epitope of the VEGF molecule, alluding to the built-in redundancy of molecular recognition.

We decided to further explore VEGF-binding designs by focusing on only three out of the six CDRs, as in nanobodies. Nanobodies are single-domain proteins derived from the variable domain of heavy chains from a special subset of antibodies in camelids that lack light chains and thus have only three CDRs instead of six. OptCDR was used to generate nanobody CDR designs by considering only the H1, H2 and H3 CDRs. Despite the reduction in the number of structural degrees of freedom, OptCDR identified designs based solely on the heavy chain CDRs (Table II) that had similar computational binding metrics to the six CDR designs. This surprising finding is consistent with the experimental observation that nanobodies can have binding affinities that are

equivalent to antibodies despite their smaller size. We believe that OptCDR achieved this through the selection of longer than typical canonical structures, especially for the H3 CDR, as is typical in experimental nanobodies. In addition, the absence of the light chain allows the antigen to assume positions and orientations that are normally prohibited due to steric clashes. A representative design plotted in Fig. 5D illustrates how the selection of longer structures for the H3 domain counteracts the loss of the light chain.

Discussion

Although there are many experimental techniques to design and redesign antibodies to bind target antigens, to date computational methods have focused only on redesigning existing antibodies to have improved binding affinities. To the best of our knowledge, OptCDR is the first general computational method for the *de novo* design of antibodies to bind any specified epitope of an antigen. Its four-step procedure works by first selecting appropriate CDR canonical structures to bind the antigen, then filling in their amino acid sequences, followed by a simultaneous structural and sequence refinement and finally a library generation step through accumulation of the best mutations to each CDR. By using these four steps, OptCDR can generate multiple novel and diverse libraries of antibodies to bind any specified antigen. We recognize that the framework regions of antibody variable domains can contribute to antibody affinity (Almagro and Fransson, 2008). In this first effort, however, we chose to exclusively focus on the CDRs as they are the most important factors in antibody recognition. This fact is manifested by the largely conserved affinity upon humanization of antibodies by grafting the CDRs onto human antibody frameworks. Nevertheless, we believe that the OptCDR workflow is open to future extensions that will target the design of the entire antibody variable domains.

OptCDR can be thought of as a computational analogue to the human immune system (or directed evolution experiments). At the start of an infection, B cells produce initial antibodies that bind the pathogen. As the infection progresses, the generated antibodies undergo an evolutionary process where their pathogen-binding affinities are improved by sequential identification of favorable mutations. The first two steps of OptCDR, the identification of appropriate combinations of canonical structures and the filling in of their amino acid sequences, can be viewed as the identification of an initial antibody to bind the antigen. The third step of OptCDR is the sequential identification of perturbation/mutation combinations that lead to improved antigen binding. The use of a methodology similar to the immune system gives OptCDR the same flexibility as the immune system: the ability to generate antibodies to bind a wide range of possible epitopes of antigens. A strength and limitation of this paradigm is that it selects antibodies on the basis of having adequate but not necessarily optimal binding affinities. This leads to an initial pool of structurally diverse antibody designs whose affinity for the targeted antibody can be ratcheted up further through the accumulation of additional mutations. The diversity of possible antibody design space in response to an antigen challenge suggests that OptCDR identified solutions are unlikely to converge to the amino acid choices of existing antibodies.

We have tested OptCDR on three antigens that span the range of antigen sizes bound by antibodies: the hapten fluorescein, a peptide from the capsid of hepatitis C, and the protein VEGF. All three antigens are bound by their wild-type antibodies with low nanomolar affinities ($K_d = 0.7$, 1.3 and 0.11 nM, respectively), making them challenging test cases. Nonetheless, for all three cases, OptCDR arrived at multiple novel sets of CDRs that form the appropriate types of antibody binding pockets for the given antigen, have interaction energies comparable to or better than the native antibodies and share many features that appear to be conducive to high-affinity binding.

Finally, we believe that OptCDR also makes an important contribution to the *de novo* protein design challenge. While there exist a large number of protein engineering methods, only a relative small number have successfully achieved the *de novo* design of proteins and enzymes for a particular task (Kraemer-Pecore *et al.*, 2003; Kuhlman *et al.*, 2003; Offredi *et al.*, 2003; Bender *et al.*, 2007). By taking advantage of the highly conserved structural features of antibodies, OptCDR achieves the *de novo* design for a special class of protein molecules.

OptCDR is available for download on our website, http://maranas.che.psu.edu. Each of the four steps is run by an interactive python script that asks for and validates all of the necessary inputs for that step and then creates and, if desired, submits a PBS script to run the step. A readme file with the codes explains all of the terminology used by the python scripts as well as highlighting the few lines of code that will need to be changed to ensure the codes all work properly on each user's system.

References

Almagro,J.C. and Fransson,J. (2008) *Front. Biosci.*, **13**, 1619–1633.
Barderas,R., Desmet,J., Timmerman,P., Meloen,R. and Casal,J.I. (2008) *Proc. Natl Acad. Sci. USA*, **105**, 9029–9034.
Bender,G.M., Lehmann,A., Zou,H., *et al.* (2007) *J. Am. Chem. Soc.*, **129**, 10732–10740.
Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) *Nucleic Acids Res.*, **28**, 235–242.
Boder,E.T., Midelfort,K.S. and Wittrup,K.D. (2000) *Proc. Natl Acad. Sci. USA*, **97**, 10701–10705.
Chen,Y., Wiesmann,C., Fuh,G., Li,B., Christinger,H.W., McKay,P., de Vos,A.M. and Lowman,H.B. (1999) *J. Mol. Biol.*, **293**, 865–881.
Chen,H.X., Gore-Langton,R.E. and Cheson,B.D. (2001) *Oncology (Williston Park)*, **15**, 1017, 1020, 1023–1016.
Chothia,C. and Lesk,A.M. (1987) *J. Mol. Biol.*, **196**, 901–917.
Clark,L.A., Boriack-Sjodin,P.A., Eldredge,J., *et al.* (2006) *Protein Sci.*, **15**, 949–960.
Clark,L.A., Boriack-Sjodin,P.A., Day,E., Eldredge,J., Fitch,C., Jarpe,M., Miller,S., Li,Y., Simon,K. and van Vlijmen,H.W. (2009) *Protein Eng. Des. Sel.*, **22**, 93–101.
Collis,A.V., Brouwer,A.P. and Martin,A.C. (2003) *J. Mol. Biol.*, **325**, 337–354.
Czerkinsky,C.C., Nilsson,L.A., Nygren,H., Ouchterlony,O. and Tarkowski,A. (1983) *J. Immunol Methods*, **65**, 109–121.

DeLano,W.L. (2008) *The PyMOL Molecular Graphics System*. DeLano Scientific LLC, Palo Alto, CA, USA.

De Pascalis,R., Gonzales,N.R., Padlan,E.A., Schuck,P., Batra,S.K., Schlom,J. and Kashmiri,S.V. (2003) *Clin. Cancer Res.*, **9**, 5521–5531.

Dimitrov,D.S. and Marks,J.D. (2009) In Dimitrov,A.S. (ed.), *Methods in Molecular Biology—Therapeutic Antibodies: Methods and Protocols*. Humana Press, Baltimore, MD, pp. 1–27.

Fazelinia,H., Cirino,P.C. and Maranas,C.D. (2007) *Biophys. J.*, **92**, 2120–2130.

Fazelinia,H., Cirino,P.C. and Maranas,C.D. (2009) *Protein Sci.*, **18**, 180–195.

Feldmann,M. and Maini,R.N. (2001) *Annu. Rev. Immunol.*, **19**, 163–196.

Fukuda,I., Kojoh,K., Tabata,N., Doi,N., Takashima,H., Miyamoto-Sato,E. and Yanagawa,H. (2006) *Nucleic Acids Res.*, **34**, e127.

Hamers-Casterman,C., Atarhouch,T., Muyldermans,S., Robinson,G., Hamers,C., Songa,E.B., Bendahman,N. and Hamers,R. (1993) *Nature*, **363**, 446–448.

Hanes,J. and Pluckthun,A. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 4937–4942.

ILOG. (2007) *ILOG CPLEX 11.0 User's Manual*. ILOG Inc., Mountain View, CA, USA.

Jermutus,L., Honegger,A., Schwesinger,F., Hanes,J. and Pluckthun,A. (2001) *Proc. Natl Acad. Sci. USA*, **98**, 75–80.

Kempeni,J. (1999) *Ann. Rheum. Dis.*, **58**(Suppl. 1), I70–I72.

Khoury,G.A., Fazelinia,H., Chin,J.W., Pantazes,R.J., Cirino,P.C. and Maranas,C.D. (2009) *Protein Sci.*, **18**, 2125–2138.

Kohler,G. and Milstein,C. (1975) *Nature*, **256**, 495–497.

Kraemer-Pecore,C.M., Lecomte,J.T. and Desjarlais,J.R. (2003) *Protein Sci.*, **12**, 2194–2205.

Kuhlman,B., Dantas,G., Ireton,G.C., Varani,G., Stoddard,B.L. and Baker,D. (2003) *Science*, **302**, 1364–1368.

Lazaridis,T. and Karplus,M. (1999) *Proteins*, **35**, 133–152.

Lippow,S.M., Wittrup,K.D. and Tidor,B. (2007) *Nat. Biotechnol.*, **25**, 1171–1176.

MacCallum,R.M., Martin,A.C. and Thornton,J.M. (1996) *J. Mol. Biol.*, **262**, 732–745.

MacKerell,A.D., Jr, Brooks,B., Brooks,C.L., III, Nilsson,L., Roux,B., Won,Y. and Karplus,M. (1998) In Schleyer,P.v.R., *et al.* (eds), *The Encyclopedia of Computational Chemistry*. John Wiley & Sons, Chichester, pp. 271–277.

Martin,A.C. and Thornton,J.M. (1996) *J. Mol. Biol.*, **263**, 800–815.

Menez,R., Bossus,M., Muller,B.H., Sibai,G., Dalbon,P., Ducancel,F., Jolivet-Reynaud,C. and Stura,E.A. (2003) *J. Immunol.*, **170**, 1917–1924.

Meyer,M.M., Silberg,J.J., Voigt,C.A., Endelman,J.B., Mayo,S.L., Wang,Z.G. and Arnold,F.H. (2003) *Protein Sci.*, **12**, 1686–1693.

Morea,V., Lesk,A.M. and Tramontano,A. (2000) *Methods*, **20**, 267–279.

Offredi,F., Dubail,F., Kischel,P., *et al.* (2003) *J. Mol. Biol.*, **325**, 163–174.

Pantazes,R.J., Saraf,M.C. and Maranas,C.D. (2007) *Protein Eng. Des. Sel.*, **20**, 361–373.

Persson,H., Lantto,J. and Ohlin,M. (2006) *J. Mol. Biol.*, **357**, 607–620.

Plosker,G.L. and Figgitt,D.P. (2003) *Drugs*, **63**, 803–843.

Rajpal,A., Beyaz,N., Haber,L., Cappuccilli,G., Yee,H., Bhatt,R.R., Takeuchi,T., Lerner,R.A. and Crea,R. (2005) *Proc. Natl Acad. Sci. USA*, **102**, 8466–8471.

Reen,D.J. (1994) *Methods Mol. Biol.*, **32**, 461–466.

Reiersen,H., Lobersli,I., Loset,G.A., *et al.* (2005) *Nucleic Acids Res*, **33**, e10. First published,

Saraf,M.C., Horswill,A.R., Benkovic,S.J. and Maranas,C.D. (2004) *Proc. Natl Acad. Sci. USA*, **101**, 4142–4147.

Saraf,M.C., Moore,G.L., Goodey,N.M., Cao,V.Y., Benkovic,S.J. and Maranas,C.D. (2006) *Biophys. J.*, **90**, 4167–4180.

Sivasubramanian,A., Sircar,A., Chaudhury,S. and Gray,J.J. (2009) *Proteins*, **74**, 497–514.

Vogel,C.L., Cobleigh,M.A., Tripathy,D., *et al.* (2001) *Oncology*, **61**(Suppl. 2), 37–42.

Walker,L.M., Bowley,D.R. and Burton,D.R. (2009) *J. Mol. Biol.*, **389**, 365–375.

Whitelegg,N.R. and Rees,A.R. (2000) *Protein Eng.*, **13**, 819–824.

Whitlow,M., Howard,A.J., Wood,J.F., Voss,E.W., Jr and Hardman,K.D. (1995) *Protein Eng.*, **8**, 749–761.

Willett,C.G., Boucher,Y., di Tomaso,E., *et al.* (2004) *Nat. Med.*, **10**, 145–147.