COMPUTATIONAL DESIGN OF COMBINATORIAL PEPTIDE LIBRARY FOR MODULATING PROTEIN-PROTEIN INTERACTIONS [a]

XIANG LI[1,2] and JIE LIANG[2]
[1] *Graduate Program in Bioinformatics and* [2] *Department of Bioengineering, SEO, MC-063*
University of Illinois at Chicago
851 S. Morgan Street, Room 218
Chicago, IL 60607–7052 U.S.A.
(Accepted by Pacific Symposium of Biocomputing, 2005)

Screening phage-displayed combinatorial peptide library is an effective approach for discovery of peptide modulators for protein-protein interactions. However, as peptide length increases, the chance of finding active peptides in a finite size library diminishes. To increase the likelihood of finding peptides that bind to a protein, we develop statistical potential for computational construction of biased combinatorial antibody-like peptide libraries. Based on the alpha shapes of antibody-antigen complexes, we developed an empirical pair potential for antigen-antibody interactions that depends on local packing. We validate this potential and show that it can successfully discriminate the native interface peptides from a simulated library of 10,000 random peptides for 34 antigen-antibody complexes. In addition, we show that it can successfully recognize the native binding surface patch among all possible surface patches taken from either the antibody or the antigen for seven antibody-antigen protein complexes contained in the CAPRI (Critical Assessment of Predicted Interactions) dataset. We then develop a Weighted Amino Acid Residue sequence Generator (WAARG) for design of biased peptide library. When compared with a random peptide library, WAARG libraries contain more native-like binding peptides at a significantly smaller size. Our method can be used to construct peptide library for screening of antibody variants with improved specificity and affinity to a target antigen. It can also be used for screening of antibody-like antagonist peptides modulating other protein-protein interactions.

1 Introduction

Modulating protein-protein interactions has the promise of obtaining many novel therapeutic agents. An effective approach for discovery of such modulators is through screening of combinatorial libraries of peptides. To identify peptides that interact with an antigen, the technique of phage display is effective [1,2], because it can produce synthetic peptides that may have the target-recognition qualities of natural antibodies. By fusing the DNA sequence

[a] This work is supported by grants from NSF(CAREER DBI0133856 and DBI0078270), NIH(GM68958), and ONR(N000140310329). We thank Dr. Brian Kay for many stimulating discussions.

encoding a particular peptide to the gene of a coat protein of a bacterial virus (called phage), the peptide is displayed on the virus coat. The coding sequence for the selected peptide inside the phage can be readily retrieved for further analysis and amplification. A phage library is formed by a collection of a large number of antibody-covered phages. The new technique of trinucleotide-phosphoramidite-based synthesis enables the design of peptide library at individual residue level [3].

Random phage-display peptide libraries have been applied to identify binding peptides of a specific target. They can also be used to predict binding sites on a 3D structure [4]. However, random combinatorial libraries meet their limitations because of the huge sequence space. The entire mutated fragment of each peptide phage libraries can contain up to a billion different peptides, a size comparable to that of the repertoire of human immune system. It is still too small to cover the space of 3×10^{19} possible 15-mers. When a typical random phage library containing 10^9 unique peptide sequences is screened, there is only a minuscule chance that a peptide of length 15 with reasonably high affinity for an intended protein target is actually contained in the library. It is necessary to generate biased libraries that are enriched with active peptides.

In this study, we develop a method for computational design of phage display libraries, with the goal to improve the likelihood of finding effective peptide modulators. Our method requires a known protein target structure of antigen to which a modulating peptide will bind. A critical ingredient is a potential function that can be computed efficiently to guide the generation of promising candidate peptides. We develop such an empirical potential function based on statistical analysis of contact interactions across protein-protein interfaces in protein database. Specifically, we select a set of protein-protein complexes from Protein Data Bank and compute their alpha shapes. Statistical models are then developed for estimating propensity for two residues on two proteins to interact.

An important consideration of our model for contact interactions is the local environment. It is well known that protein-protein interface is not evenly packed: some regions are tightly packed, but others contain voids and pockets [5]. Both energetically important interfacial residues, termed as "hot spots" [6], and structurally conserved residues are more likely to be located at tightly packed regions [7]. The importance of hot spots in such tightly packed region are due to not only the numerous contact interactions with the binding partner, but also the dehydrated environment where H-bonding interactions are enhanced because of reduced dialectic constant [8]. A parameter related to local packing environment is explicitly included in our model.

We organize this paper as follows. First, we describe the geometric model

for contact interactions and for the coordination shell surrounding a contacting residue pair, and how they can be computed using alpha shape. We then discuss the probabilistic model of packing-dependent empirical interface contact potential. This empirical potential is then validated by testing its ability to discriminate native antibody interfaces from random peptides, and its ability to recognize native binding surface patches from other surface patches for antibody and antigen complexes used in the CAPRI competition. We then describe how this scoring function can be used to generate biased peptide library. We conclude with discussion.

2 Model and Methods

Empirical potential of residue interactions. Due to its simplicity and fast evaluation, empirical potential based on statistics of protein database is well-suited for rapid generation of peptide library of tens of thousands peptides. Molecular mechanics and other methods based on potential functions derived from physical modes are difficult to use for this purpose.

Empirical potential can be derived based on either description of protein structure at residue level[9–11] or at atomic level[12,13]. It has been applied with success in fold recognition and structure prediction. Because the atomic details of protein-protein interactions are difficult to obtain, we develop potential function based on residue level representation for designing peptide library.

We use a simplified residue model for protein structure. We follow the union of ball model and represent the i-th amino acid residue as a ball b_i [14], whose center $\boldsymbol{x}_i \in \mathbb{R}^3$ coincides with the geometric center of its side-chain. Because Gly residue has no side chain, the position of C_α atom is taken as the center of the ball $\boldsymbol{x}_i$. The radius r_i of each ball depends on the size of side chain, and is taken from values by Levitt listed in [15], with an added 0.5 Å increment to account for uncertainty due to side-chain flexibility. This is necessary to reduce spurious contacts.

Alpha contacts and coordination shell. We are interested in identifying contacting residues that are spatial nearest neighbors. We use the dual complex calculated by the alpha shape software to identify such residues [14,16–20]. Briefly, the Voronoi diagram decomposes the space and the union of residue balls $\bigcup B = \bigcup b_i$ into convex regions V_B, and the dual complex $\mathcal{K}$ or the alpha shape of the molecule records the overlap pattern among these regions [14]: $\mathcal{K} = \{\sigma = \mathrm{conv}\boldsymbol{x}_B \mid \bigcap V_B \cap \bigcap B \neq \emptyset\}$, where $\boldsymbol{x}_B$ is the set of residue centers $\{\boldsymbol{x}_i\}$ of a set of balls B, V_B is the set of Voronoi cells of balls B, whose intersection $\bigcap V_B$ overlap with the intersection $\bigcap B$ of the balls. $\mathrm{conv}\boldsymbol{x}_B$ is the

convex hull of residue centers $\boldsymbol{x}_B$, which forms a simplex σ. In this study, we only make use of a subset of 1-simplices σ_{ij} (or alpha edges), such that the corresponding two residues i and j are from two proteins. Denote the name of the protein of residue i as $\mathbb{I}(i)$, we have $\mathbb{I}(i) \neq \mathbb{I}(j)$.

For a pair of such balls $B = \{b_i, b_j\}$ located on the protein interface, we examine the set of residues S_{ij} that are connected by an alpha edge to either residue b_i or residue b_j: $S_{ij} = \{\ b_k | \sigma_{ki,k\neq j} \in \mathcal{K}$ or $\sigma_{kj,k\neq i} \in \mathcal{K}\}$. We call this set of residues the *coordination shell* of the interacting residue pair i and j. The number of such residues $z_{ij} = |S_{ij}|$ is termed the coordination number of contacting residue pair ij.

The Delaunay triangulation is computed using the DELCX program, and the alpha shapes computed using the MKALF program [17, 19]. Both can be downloaded from the web-site at (`http://www.alphashape.org`).

Probabilistic model. The propensity $p(k, l, z)$ for residue of type k interacting with residue of type l with coordination number z is modeled as an odds ratio. We first estimate the probability $q(k, l, z)$ of residues of type k and type l interact across protein-protein interface with a coordination number z. The random probability $q_R(k, l, z)$ of a pairwise contact involving both residue k and l with coordination number z is calculated from a null model (or reference state). Specifically, we have: $p(k, l, z) = \frac{q(k,l,z)}{q_R(k,l,z)}$, where $q(k, l, z) = \frac{n(k,l,z)}{\sum_{k',l',z} n(k',l',z)}$. Here, $n(k, l, z) = |\{\sigma_{ij} | \sigma_{ij} \in \mathcal{K}$ and $\mathbb{I}(i) \neq \mathbb{I}(j)\}|$ is the number count of alpha edge contacts on protein interface involving residue type k and residue type l when coordination number is z. $\sum_{k',l',z} n(k', l', z)$ is the total number of all interfacial alpha contacts of any residue types with the same coordination number z. The random probability $q_R(k, l, z)$ is the probability that a pair of contacting residues is selected from surface residue of type k and type l, when chosen randomly and independently. Here a surface residue is defined as that with more than 15% of its total solvent accessible surface area exposed in the model of a tri-peptide Gly-X-Gly [21]. We divide the range of coordination number z into five intervals $[0, 3], [4, 6], [7, 9], [10, 12]$ and $[13, \infty)$ for all pair contact interactions.

The choice of random model or reference state for estimating $q_R(k, l, z)$ is critical for empirical potential. We use a random model or reference state, where there are no preferred contacts between any residue type k and any residue type l, no preference for location of k or l to be on interface or on the rest of surface, and no preferential coordination number z for any interfacial contact pair. For our random model, packing plays no direct roles for protein-

protein interactions. We have:

$$q_R(k,l,z) = q_R(k,l) = 2 \cdot n_k n_l \cdot (\frac{1}{n(n-1)}), \quad \text{when} \quad k \neq l \tag{1}$$

and

$$q_R(k,l,z) = q_R(k,l) = n_k(n_k-1) \cdot \frac{1}{n(n-1)}, \quad \text{when} \quad k = l, \tag{2}$$

where n_k is the number of surface residues of type k, and n is the total number of surface residues.

The alpha contact potential $U(i,j)$ of protein-protein interaction between residue i and residue j with coordination number z is obtained as $U(\sigma_{ij}) = -\ln p(a(i), a(j), z_{ij})$ using kT unit, where $a(i)$ and $a(j)$ are the residue types of residue i and j, respectively. The overall energy of a protein-protein interface is calculated as:

$$E = \sum_{\substack{\sigma_{ij}, \\ \mathbb{I}(i) \neq \mathbb{I}(j)}} U(\sigma_{ij}), \tag{3}$$

To assess the importance of local packing environment as reflected by the coordination number z, we also develop a simpler scoring function which does not consider the local packing environment: $p(k,l) = \frac{q(k,l)}{q_R(k,l)}$, where $q(k,l) = \frac{n(k,l)}{\sum_{k',l'} n(k',l')}$, and $n(k,l) = |\{\sigma_{ij} | \sigma_{ij} \in \mathcal{K}, a(i) = k, a(j) = l \text{ and } \mathbb{I}(i) \neq \mathbb{I}(j)\}|$ is the number count of interfacial contact pairs between residue types k and l. The random probability $q_R(k,l)$ is calculated using Equation (1) and (2).

Dataset of nonredundant antibody-antigen complexes. Contact potentials derived from one dataset may not be universal and fully transferable to other systems [22]. Therefore, the selection of a representative set of proteins is important for developing empirical potential. Because our goal is to design synthetic antibody for enhanced binding affinity or for creating novel binding, we collect a dataset of antibody-antigen complex structures. We select co-crystallized complex structures from the Protein Data Bank that satisfy the following criteria: the resolution of each chain should be less than 2.5 Å; each chain should have more than 30 amino acids; no pair of chains in protein complexes have a sequence identity larger than 25% to any other chain in the data set. Based on these three criteria, we collected a set of 34 antibody-antigen complexes.

3 Results

Discriminating native antibody interfaces. We first validate the empirical potential by testing whether it can identify interface residues on the native antibody. For each antibody-antigen complex, we first locate the interface residues on the antibody and on the antigen, respectively. These are residues connected by alpha edges across the two proteins. To model a random phage library where each residue has equal probability to be generated at each position of the peptide, we randomly substitute uniformly all the interfacial residues on the antibody with any of the 20 amino acid residues. The interface residues on the antigen are unchanged. For simplicity, we assume that the interface contacts and hence the coordination numbers for all contacting pairs remain unchanged. Given m interface residues on the antibody, there are 20^m possible different sequences, where m ranges typically from 4 to 26. We randomly generate a sample of 10,000 sequences to test the empirical potential.

We performed 34 leave-one-out tests. In each case, 33 of the antibody-antigen complexes are used to construct the empirical potential. The remaining antibody-antigen complex is used for testing. Table 1 shows that among 34 antibody-antigen complexes, 28 native interfaces rank among the top 10 interfaces among the corresponding 10,000 random interfaces. The median and average rankings of 34 native sequences among the 34 sets of 10,000 generated sequences are 1 and 20, respectively, and their z-scores are 4.79 and 4.38, respectively. The incorporation of the coordination number for local packing environment is important. Without such consideration, the median and mean rankings of the 34 native sequences are only 9 and 236, respectively, and their z scores are 3.44 and 3.01, respectively. In terms of the mean value of ranking of native sequences, the performance improves more than 10 times when local-packing is considered. For comparison, results of discrimination using Miyazawa-Jernigan potential are also listed in Table 1.

Recognition of binding surface patch of CAPRI targets. The CAPRI (Critical Assessment of PRedicted Interactions) competition is designed to evaluate current protein docking algorithms. A blind docking prediction starts from two known crystallographic or NMR structures of unbound proteins and ends with a comparison to a solved structure of the protein complex, to which the participants did not have access. Since its inception in 2001, a total of 19 protein complexes have been used for blind docking. Among these, seven are antibody or antibody related proteins (*e.g.*, Fab fragment, T-cell receptor). We use these seven complex structures to evaluate the effectiveness of the empirical potential.

Table 1: Discrimination of Native Antibody Interfaces.

Ab - Ag Complexes	Local packing dependent	Local packing independent	Miyazawa-Jernigan potential	$^bN_{Ab}$	$^cN_{Ag}$
1i8k	[a]375/2.29	4770/0.04	3523/0.39	18	8
1nmb	137/ 2.47	111/2.37	406/1.84	15	18
1e6j	54/2.72	123/2.34	463/1.66	18	13
1jps	35/2.93	364/1.86	142/2.31	21	22
1iqd	15/3.19	1022/1.29	1764/0.94	26	17
1nsn	7/3.30	502/1.69	410/1.70	20	21
1osp	20/3.38	69/2.63	215/2.08	15	21
1nca	10/3.47	23/2.81	1192/1.18	22	24
1qfu	3/3.63	326/1.86	47/2.55	24	21
1kb5	5/3.78	50/2.54	1399/1.10	24	26
2jel	1/3.86	181/2.19	1508/1.02	20	16
1eo8	5/3.90	201/2.11	861/1.39	19	19
1ai1	3/4.29	53/2.73	6582/-0.42	17	6
1dvf	1/4.32	7/3.25	272/1.92	19	18
1wej	1/4.35	2/3.97	2/4.16	13	11
1mpa	1/4.60	4/3.57	4781/0.04	16	7
1ktr	1/4.71	157/2.25	1486/1.04	17	4
1fe8	1/4.79	4/3.67	480/1.71	23	20
3hfm	1/4.80	4/3.55	262/1.93	21	17
1f58	1/4.81	4/3.60	1022/1.28	18	10
3hfl	1/4.84	7/3.64	45/2.80	19	16
1nby	1/4.86	3/3.93	33/2.90	22	18
1jhl	1/4.89	2/4.16	9/3.43	16	11
1fns	1/4.94	6/3.80	191/2.16	16	12
1iai	1/5.01	3/3.46	1147/1.21	21	23
1gc1	1/5.10	29/3.09	225/2.04	14	12
1g9n	1/5.16	16/3.46	22/2.76	13	12
1a2y	1/5.20	1/4.38	153/2.29	14	14
1jrh	1/5.27	1/4.54	4/3.97	20	15
2iff	1/5.28	10/3.42	28/3.03	20	16
2hrp	1/5.63	7/3.46	303/1.88	17	8
1cu4	1/5.65	4/3.52	223/2.04	22	9
1a3r	1/5.69	2/4.35	48/2.57	31	14
2ap2	1/5.83	1/4.22	1581/1.02	17	8
Average	20/4.38	236/3.01	907/1.88	19	15
Median	1/4.75	9/3.44	286/1.90	19	16

[a] The first number in each cell is the rank of native antibody interface and the second number is the Z score. Z score $= \overline{E} - E_{native}/\sigma$; $\overline{E}$ and σ are the mean and standard deviation of the scores of 10,000 randomly generated peptides, respectively. $^bN_{Ab}$: number of interfacial residues on the interface of antibody side. $^cN_{Ag}$: number of interfacial residues on the interface of antigen side.

In docking, a *cargo* protein is docked to a *seat* protein. All surface patches as candidate at binding interface are sampled from the surface of an unbound structure. We therefore generate candidate sequences of binding peptide from

the surfaces of *cargo* protein structure. Since our goal is not docking but to evaluate the performance of the potential function, we assume the knowledge of the binding interface on the *seat* protein. We further assume the knowledge of the coordination number for interface residues.

We first partition the surface of the unbound *cargo* protein into candidate surface patches, each has the same size as the native binding surface of m residues. A candidate surface patch is generated by starting from a surface residue on the *cargo* protein, and following alpha edges on the boundary of the alpha shape by breadth-first search, until m residues are found. We construct n candidate surface patches by starting in turn from each of the n surface residue on the *cargo* protein. None of the candidate patches is identical to the native binding surface patch.

Second, we assume that a candidate surface patch on *cargo* protein has the same set of contacts as that of the native binding surface. The coordination number for each hypothetical contacting residue pair is also assumed to be the same. We replace the m residues of the native surface with the m residues from the candidate surface patch. There are $\frac{m!}{\prod_{i=1}^{20} m_i!}$ different ways to permute the m residues of the candidate surface patch, where m_i is the number of residue type i on the candidate surface patch. A typical candidate surface patch has about 20 residues, therefore the number of possible permutation is very large. For each candidate surface patch, we take a sample of 1,000 random permutations. The expected binding energy $\bar{E}$ for a candidate surface patch is estimated as $\bar{E} = \sum_{k=1}^{1,000} E_k$, where E_k is calculated using Equation (3) for the k-th permutation. The value of $\bar{E}$ is used to rank the candidate surface patches.

We assess the empirical potential by taking antibody/antigen protein in turn as the seat protein, and the antigen/antibody as cargo protein. The native interface on the seat protein is fixed and we test if our empirical potential can identify the correct surface patch on the *cargo* protein from the set of candidate surface patches plus the native surface patch. The results are listed in Table 2. Among the 14 native binding surfaces for 7 protein complexes, we can rank 11 native binding surfaces successfully as the top surface of the rank ordered list. The remaining 3 native binding surfaces all rank among the top 5, and the best ranking candidate surface patches for these three proteins all have over 50% native interfacial residues. By this criteria, our potential function can correctly recognize the native or near native binding surface patches of antibody and antigen complexes.

Weighted Amino Acid Residue sequence Generator (WAARG) We then develop an method to generate candidate antibody sequences for a given

Table 2: Recognition of Native Binding Surface of CAPRI Targets

		[a]Antibody			Antigen		
Target	Complex	[b]R_{native}	[c]O	[d]N	R_{native}	o	m
T02	Rotavirus VP6-Fab	1	0.65	283	1	0.72	639
T03	Flu hemagglutinin-Fab	1	0.55	297	1	0.68	834
T04	α-amylase-camelid Ab VH 1	2	0.53	89	1	0.47	261
T05	α-amylase-camelid Ab VH 2	1	0.43	90	5	0.56	263
T06	α-amylase-camelid Ab VH 3	1	0.63	88	1	0.56	263
T07	SpeA superantigen TCRβ	1	0.57	172	1	0.64	143
T13	SAG1-antibody complex	3	0.64	286	1	0.68	249

[a] "Antibody": surface patches on the antibody molecule are scored, while the native binding surface on the antigen is kept unchanged. "Antigen": similarly defined as "Antibody". [b]R_{native}: Ranking of native binding surface among all n candidate surface patches. [c]O: Percentage of overlap of residues from the best candidate patch with that of the native binding surface patch. [d]m: Number of surface residues. It is also the number of partitioned candidate surface patches.

antigenic protein (called WAARG for Weighted Amino Acid Residue sequence Generator) based on the empirical potential. Again, we assume the knowledge of the binding surface on the seat protein, the size m of the binding interface on the cargo protein, which is also taken as the length of peptide that needs to be generated. We further assume the same contact patterns as observed in the complex structure. The unknowns are the identities and sequence of the residues that would best bind to the binding surface on the seat protein.

To generate biased sequences, the probability $\pi(i, a(i))$ of placing a residue of type $a(i)$ at position i of the length m sequence is set to be proportional to $\exp(\sum_{j,\sigma_{ij}} U(i, j, z_{ij}))$, where σ_{ij} is an interfacial alpha edge across two proteins, and $U(i, j, z_{ij})$ is the empirical energy score. The value of $\pi(i, a(i))$ therefore depends on the residue types of the contacting residue pair i and j, and the local packing environment reflected by the coordination number z_{ij}.

One way to specify the design of a biased peptide library is to provide a profile listing the favorable residues at each peptide position, along with the bias (or weight). Table 3 shows an example of such a profile for constructing a Protein A binding peptide library of length 7. The known native binding sequence is listed in the first row, followed by the profile consisting of the top 10 amino acid residues ranked by their weights at each position. At three residue positions, the wild type residues in the native binding surface are ranked first. For other positions, the wild type residues are ranked among the top 6, except for the position of GLY.

Table 3: Example of Weighted Sequence Library for Complex of Protein A and Antibody Fab (*1osp*).

[a]Native Seq.	Y	S	D	Y	G	Y	R
1	[b]Y(0.47)	H(0.34)	Y(0.34)	Y(0.51)	R(0.32)	Y(0.58)	W(0.36)
2	W(0.27)	N(0.26)	E(0.25)	N(0.19)	K(0.23)	N(0.12)	S(0.22)
3	F(0.06)	M(0.09)	H(0.16)	W(0.12)	Y(0.12)	S(0.06)	T(0.12)
4	N(0.04)	Y(0.06)	D(0.09)	V(0.06)	V(0.08)	V(0.06)	Y(0.10)
5	S(0.03)	S(0.04)	R(0.02)	Q(0.05)	I(0.05)	W(0.05)	K(0.06)
6	V(0.02)	W(0.04)	N(0.02)	C(0.01)	W(0.04)	Q(0.02)	R(0.03)
7	C(0.02)	Q(0.03)	I(0.01)	R(0.01)	A(0.03)	R(0.01)	M(0.01)
8	R(0.02)	R(0.03)	L(0.01)	K(0.01)	C(0.03)	K(0.01)	G(0.01)
9	K(0.01)	L(0.02)	F(0.01)	G(0.01)	M(0.02)	G(0.01)	N(0.01)
10	G(0.01)	V(0.02)	S(0.01)	M(0.01)	H(0.02)	M(0.01)	A(0.01)

[a]Native Seq.: The interfacial sequence from the heavy chain. [b]Y(0.47): residue type Y is to be chosen with a weight of 0.47. Underline: the chosen residue is the same as the residue at native interface.

Assessing WAARG performance. To assess the overall quality of the peptide library generated computationally, we calculate similarity score of a designed sequence to the corresponding native sequence using the BLOSUM62 substitution scoring matrix. We found that in most cases, candidate peptides generated by the Weighted Amino Acid Residue sequence Generator (WAARG) have significantly higher sequence similarity than random sequences (Figure 1). To generate random sequences, we sample uniformly each of the 20 amino acid residues for each of the m positions. The average similarity between a native sequence of antibody interface and 1,000 random sequences $\overline{S}_{random}$ ranges from -30.14 to -11.95. The average similarity between a native sequence of antibody interface and 1,000 biased sequences generated by WAARG $\overline{S}_{weighted}$ ranges from -6.73 to 38.36. Figure 1(b) shows the distributions of similarity scores for designed and random sequences for N10-staphylococcal nuclease-antibody complex (*1nsn*). The $\overline{S}_{random}$ and $\overline{S}_{weighted}$ is -16.22 and 25.12, respectively. The overall distribution of similarity scores by WAARG has much higher similarity compared to the distribution of random sequences. These results shows the peptide library generated by WAARG will have significantly more enriched native-alike peptides than random.

Another method to assess the performance in generating biased library is to compare the number of sequences appeared before a sequence similar to that of the wild-type binding interface first occurs for both WAARG and random generators. This evaluation provides indication of the appropriate size of a peptide library to ensure inclusion of a number of good candidate sequences. We illustrate with the example of the binding interface between the heavy chain of NC10 antibody and influenza virus neuraminidase (*1nmbHN*) as a

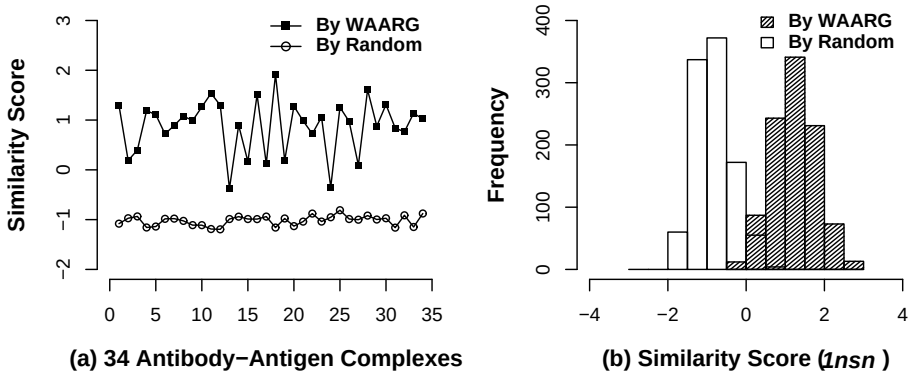

Figure 1: Evaluation of biased library. (a) Average similarity scores between one of the 34 native sequences of antibody interfaces and 1,000 sequences generated by random ($\overline{S}_{random}$) and 1,000 sequences by empirical potential ($\overline{S}_{weighted}$) ; (b) The similarity score distribution of $\overline{S}_{random}$ and $\overline{S}_{weighted}$ for antibody N10-staphylococcal nuclease complex (*1nsn*), one example among the 34 complexes. Every similarity score is normalized by the sequence length.

testing example. On the interface of this complex, there are seven residues on the binding surface of the antibody, and ten residues on the binding surface of the antigen. We design a library of peptides of length 7 that would bind to the antigen, and record the number of sequences appeared for the two methods before a candidate sequence with 4, 5, 6, and 7 identical residues to that of the wild-type binding surface peptide occur. If the sequence identity is 7, the candidate sequence is exactly the same as the native sequence. Table 4 shows that the WAARGG can generate reasonably good candidate sequences with a much smaller library size than the random generator.

Table 4: Efficiency of WAARG

Identity	Num by WAARG	Num by Random	Candidate Seq. by WAARG
4	170	1721	N N Y Y D W H
5	2854	19417	S N Y F Y Y G
6	13645	$^a 20^7/(7 \times 19)$	S N Y Y Y Y G
7	367288	$^a 10^7$	^{b}S N Y Y D Y G

^{a}Expected number of sequences generated before an active peptide occurs. This number follows exponential distribution with the expectation $1/\lambda$, where λ is the probability of a random sequence to be a required one. $\lambda = 7 \cdot (1/20)^6 \cdot (19/20) = (7 \times 19)/20^7$ when the identity is required to be six, and $\lambda = (1/20)^7$ when the identity is required to be seven. *b*: Wild type.

4 Discussion

We have developed a method for computational design of peptide library that can introduce useful bias to increase the efficiency in discovery of peptides binding to a target antigen protein. The key elements of our method is the alpha shape method to identify precise contact interactions, and an empirical potential for antibody-antigen interactions. We show that such a potential can be obtained by analyzing the alpha edges of known protein complexes. We find that it is important to consider the local packing environment, and the introduction of the coordination number in the empirical potential significantly improves the performance of the designed peptide library. Further development will need to consider the codon usage of the bacteria where the phage library is expressed.

References

1. D.J. Rodi, L. Makowski, and B.K. Kay, *Curr Opin in Chem Biol* **6**, 92 (2002).
2. R.H. Hoess, *Chem. Rev.* **101(10)**, 3205 (2001).
3. M.D. Hughes, D.A. Nagel, A.F. Santos, A.J. Sutherland, and A.V. Hine, *J. Mol. Biol.* **331(5)**, 973 (2003).
4. I. Halperin, H. Wolfson, and R. Nussinov, *Protein Sci.* **12**, 1344 (2003).
5. A.T. Binkowski, L. Adamian, and J. Liang, *J. Mol. Biol.* **332**, 505 (2003).
6. T. Clackson and J.A. Wells, *Science* **267**, 383 (1995).
7. I. Halperin, H. Wolfson, and R. Nussinov, *Stucture* **12**, 1027 (2004).
8. X. Li, O. Keskin, B. Ma, R. Nussinov, and J. Liang, *J. Mol. Biol.* **In press.** (2004).
9. S. Miyazawa and R. Jernigan, *J. Mol. Biol.* **256**, 623 (1996).
10. X. Li, C. Hu, and J. Liang. *Proteins* **53** 792 (2003).
11. B. Krishnamoorthy and A. Tropsha. *Bioinformatics* **19(12)** 1540 (2003).
12. R. Samudrala and J. Moult. *J. Mol. Biol.* **275** 895 (1998).
13. H.Y. Zhou and Y.Q. Zhou, *Protein Sci.* **11** 2714 (2002).
14. H. Edelsbrunner. *Discrete Comput. Geom.* **13** 415 (1995).
15. M. Levitt. *J Mol. Biol.* **104** 59 (1976).
16. H. Edelsbrunner and P. Fu. *Rept. UIUC-BI-MB-94-01, Molecular Biophysics Group, Beckman Inst. Univ. Illinois, Urbana, IL*, (1994).
17. H. Edelsbrunner and E.P. Mücke. *ACM Trans. Graphics*, 13:43–72, (1994).
18. H. Edelsbrunner, M. Facello, P. Fu, and J. Liang. In *Proc. 28th Ann. Hawaii Int'l Conf. System Sciences*, volume 5, pages 256–264, Los Alamitos, California. IEEE Computer Scociety Press (1995).
19. M.A. Facello. *Computer Aided Geometric Design* **12** 349 (1995).
20. J. Liang, H. Edelsbrunner, P. Fu, P.V. Sudhakar, and S. Subramaniam. *Proteins* **33** 1 (1998).
21. G.D. Rose, A.R. Geselowitz, G.J Lesser, R.H. Lee, and M.H. Zehfus. *Science*

229 834 (1985).
22. J. Khatun, S. Khare, and N. Dokholyan. *J. Mol. Biol.* **336** 1223 (2004).