Geometric Filtering of Pairwise Atomic Interactions Applied to the Design of Efficient Statistical Potentials

Afra Zomorodian [a], Leonidas Guibas [a], Patrice Koehl [b,*]

[a] *Department of Computer Science*
Stanford University
Stanford, CA 94305, USA

[b] *Department of Computer Science and Genome Center*
University of California, Davis
Davis, CA 95616, USA

Abstract

Distance-dependent, pairwise, statistical potentials are based on the concept that the packing observed in known protein structures can be used as a reference for comparing different 3D models for a protein. Here, packing refers to the set of all pairs of atoms in the molecule. Among all methods developed to assess three-dimensional models, statistical potentials are subject both to praise for their power of discrimination, and to criticism for the weaknesses of their theoretical foundations. Classical derivations of pairwise potentials assume statistical independence of all pairs of atoms. This assumption, however, is not valid in general. We show that we can filter the list of all interactions in a protein to generate a much smaller subset of pairs that retains most of the structural information contained in proteins. The filter is based on a geometric method called *alpha shapes* that captures the packing in a conformation. Statistical scoring functions derived from such subsets perform as well as scoring functions derived from the set of all pairwise interactions.

Key words: Protein structure, Delaunay, Alpha Shape, Geometric Filtering, Statistical Potentials

* Corresponding author.
Email addresses: afra@cs.stanford.edu (Afra Zomorodian), guibas@cs.stanford.edu (Leonidas Guibas), koehl@cs.ucdavis.edu (Patrice Koehl).

1 Introduction

It is not feasible to determine experimentally the structure of the millions of proteins whose corresponding genes have been sequenced as part of the multiple genome projects. However, there is hope that biologically useful models can be derived by inference from the databases of known protein structures. The main reason for this hope is the common knowledge that proteins with homologous sequences have similar structures. In such cases, models for the unknown structure of a new protein can be built from the structure of a homologous protein using comparative modeling techniques [1]. In many cases however, there is no significant sequence similarity between the sequence of the unknown protein, and the sequences of the proteins whose structures are known. In these cases, it is possible to generate models for the protein whose structure is unknown using either fold recognition techniques [2,3] or *ab initio* protein structure prediction methods. The main challenge faced by the latter approaches is the definition of an approximation of the potential energy function describing the stability of a protein. This approximation would ideally have the native structure of a protein as a minimum. Should such a function be known, protein structure prediction would proceed either by optimizing the conformation of a protein such that its energy is minimum, or use the function for identifying native-like models in a large set of non-native models, also called *decoys* [4]. There are two major types of potentials currently developed for that purpose: the atomic "physical" potentials, and the database-derived potentials [5].

Atomic force-fields are derived from physical models for intra-molecular interactions and for solvent effects. The force-fields are widely used for precise molecular dynamics and molecular mechanics simulations. The parameters of these energy functions are traditionally optimized with respect to the properties of small molecules. In some cases, the parameters are extracted from quantum chemical calculations [6]. These force-fields have proven useful for studying the structures and dynamics of molecules near their native states. They are, however, not (yet) appropriate for studying complex processes over large time scales. They have large computational complexities and are unable to discriminate reliably between native and non-native models of proteins. Consequently, there has been a growing interest in developing simplified, specifically designed energy functions that are tractable for simulations [7,8,9,10,11,12,13]. Two types of potentials have been designed, reflecting their respective usage. Firstly, folding potentials are optimized for simulation of the protein folding process. The common approach is to choose a functional form for the potential and optimize a folding criterion over the parameter space [14, 15, 16, 17, 18]. Secondly, potential energy functions are designed to provide reliable separation between near-native and non-native structures for a protein. Most of these potential are knowledge-based, that is, they extract information from

the database of known protein structures [9].

The basic assumption for designing knowledge-based potentials is that the packing observed in known protein structures can be used as a reference point when comparing different models for a protein. The packing information is extracted using either a physical or statistical model. This information is then utilized in a *potential of mean force* or *statistical potential*, respectively. It was found that residue pairing frequencies [7, 8, 9, 19, 4, 11, 20, 21] (with or without distance dependence), residue triplet [22] and residue quadruplet frequencies [23], solvent accessibility and atomic contacts [24, 25, 26], dihedral angle preferences [27], ion pair distributions [28], as well as combinations of these properties [29, 4] provide useful information for discriminating near native structural models from misfolded models of proteins. Despite significant progress however, there is not yet a "universal" potential that reliably solves the discrimination problem.

Further, the theoretical basis of statistical potentials has been challenged [30, 31, 32]. For example, pair-wise potentials rely on the assumption of the independence of the different pairs of atoms considered. This assumption is certainly not valid for a protein in a compact state: it was shown, for instance, that statistical potentials based on residue pair frequencies remember the topology of the proteins included in the database from which they were derived [32]. In addition, most of these potentials are defined with respect to a reference state. There is yet no consensus on how to define this state [33,34].

It is worth mentioning that decoys sets, i.e. collections of structural models for a protein structure, play an essential role in the development of energy functions: the latter are optimized for detecting near native conformations in the former which, in turn, are constantly improved to provide stringent and meaningful test sets to assess the quality of energy functions [4,35,36,37]. This connection can in certain circumstances be put in good use. For example Baker and co-workers have shown that if the distribution of structures in a decoy set reflects the distribution of conformations expected from energetics consideration, theere are a greater number of decoys near a native-like conformation than there are surrounding a misfolded conformation [38]. This observation has lead to the widespread strategy of clustering a decoy set using structural similarity as a distance measure, and choosing a candidate model for the protein whose structure is predicted among the conformations found in the largest cluster [38, 39]. As another example, Keasar and Levitt [37] have shown that by refining iteratively the set of decoys and the energy function, they could reduce substantially the conformational space that need to be sampled for efficient structure prediction. The use of this connection however comes with a price: it leads to development of decoy dependent energy functions [40], that are likely to fail on decoy sets generated under different conditions.

In this paper, we focus on pair-wise, residue-level, contact information in proteins, and their translation into statistical scoring functions. Our aim is not to derive a *better* potential than existing ones. Rather, we investigate the extent to which all pairs of residues within a protein are required for proper discrimination. We also examine the possibility of extracting subsets of all pairs, without loss of information. We test our potentials on well established decoy sets, and provide comparison with published potentials.

2 Statistical Potentials

General atomic force-fields used in molecular simulations are designed to provide good approximations of the physical interactions that stabilize a proteins. Since the native state of a protein is at a minimum of its free energy, it can theoretically be derived by minimization of its total energy defined from these force-fields. Statistical potentials, on the other hand, are designed to be optimal for native protein structures. The native state of a protein is assumed, and the parameters describing the potential are derived such that this native state becomes optimal. Pairwise residue-specific contacts found in known protein structures are used to define the relative probability of finding atom pairs at specified distances. We use the above principle to develop a statistical potential from probability distribution functions (PDFs) of distances between atoms. Our formalism is residue-based: each residue in the protein is represented by a single atom. However, it is easily extendable to a full-atom potential.

Let M be a protein of n residues. We denote its sequence as S, where S_i is the ith amino acid in the sequence. We are interested in examining possible conformations for this protein. Our main assumption is that a conformation of a protein is fully characterized by all its inter-residue distances. Let (i, j) denote the pair of residues at location i and j in the sequence S, and Ω be the set of all $\binom{n}{2}$ residue pairs. A proposed conformation C places each residue pair (i, j) at distance d_{ij}. For each conformation, the space Ω can be described by two variables, corresponding to structure and sequence, respectively. Let X be the state variable for the distance between the two residues, and let T be the state variable defining the amino acid types of a pair of residues. We now wish to evaluate the probability that a given conformation C has sequence S. This is the conditional probability $P(S \mid C)$. We expect this probability to attain its maximum when C is the native conformation of M. Therefore, we define our scoring function to be the negative of this probability, linearized by taking a log.

$$E(S \mid C) = -\ln \mathrm{P}(S \mid C) \tag{1}$$

Note that our scoring function E behaves like a traditional energy function, reaching its minimum at the native conformation.

We now examine the computation of the scoring function. Assuming that all residue pairs in Ω are independent, we have

$$\mathrm{P}(S \mid C) = \prod_{(i,j)\in\Omega} P_{ij}, \tag{2}$$

where

$$P_{ij} = \mathrm{P}\left(T = (S_i, S_j) \mid X = d_{ij}\right). \tag{3}$$

In other words, knowing that a pair of residues lie at distance d_{ij}, we compute the probability that their types matches the corresponding residues in the sequence of our protein M. For each pair (i, j), we apply Bayes's rule to (3) to get

$$P_{ij} = \frac{\mathrm{P}\left(X = d_{ij} \mid T = (S_i, S_j)\right)}{\mathrm{P}(X = d_{ij})} \times \mathrm{P}\left(T = (S_i, S_j)\right). \tag{4}$$

In Section 4, we describe our method for estimating the probabilities in (4) using frequencies computed on known protein structures. Substituting (4) into (2) and (1), we arrive at our scoring function:

$$\begin{aligned} E(S \mid C) &= -\ln \mathrm{P}(S \mid C) \\ &= -\ln \left[\prod_{(i,j)\in\Omega} P_{ij}\right] \\ &= -\sum_{(i,j)\in\Omega} \ln \mathrm{P}\left(X = d_{ij} \mid T = (S_i, S_j)\right) \\ &\quad + \sum_{(i,j)\in\Omega} \ln \mathrm{P}(X = d_{ij}) \\ &\quad - \sum_{(i,j)\in\Omega} \ln \mathrm{P}\left(T = (S_i, S_j)\right). \end{aligned} \tag{5}$$

The first two sums on the right side of Equation (5) depend on the conformation C, while the last sum depends only on the sequence S. If we use the scoring function $E(S \mid C)$ to compute model structures for a fixed sequence, we can ignore the last term as long as all residue pairs are included in the computation. Indeed, dropping this term gives us the distance-dependent potential of Sippl, derived through a physical argument based on the Boltzmann

law [9, 21, 32]. This simplification is not valid, however, if we utilize a different subset of the residue pairs for each conformation. This is precisely the case when we use a distance cutoff or a geometric filter to obtain a subset for computation.

3 Alpha Shapes

In this paper, we examine a method for selecting a significant subset of Ω, the set of all residue pairs, for computing scoring functions. Our approach is to use a geometric method called *alpha shapes* that captures the packing in a conformation. The method of alpha shapes has a natural affinity to space-filling models of molecules, such as the van der Waals model [41]. It models an atom as a hard ball with its van der Waals radius. In reality, we should view atoms as having fuzzy boundaries. We model a molecule's interactions with solvents by growing or shrinking the balls. Generalizing, we could grow and shrink the balls to capture all the possible "shapes" of a molecule. The alpha shapes model formalizes this idea. The full mathematical exposition of this method is beyond the scope of this paper, and we refer the interested reader to Edelsbrunner [42]. In this section, we hope to familiarize the reader with two main ideas of this method: representing molecules via a combinatorial object called the *dual complex*, and growing the representation of atoms via a model that allows for efficient computation.

3.1 Dual Complex

We view a molecule as a union of spherical balls. We show a simple example in two dimensions in Figure 1. To discover how the union of balls is connected, we decompose the region into convex pieces using a convenient distance metric that takes the radius of the atoms into account. Specifically, the *weighted square distance* of a point x from a ball centered at c with radius r is $\|x - c\|^2 - r^2$. The *Voronoi region* of a ball is the set of points closest to it according to this metric. The boundaries of the Voronoi regions decompose the union of balls into convex cells, as illustrated in Figure 1. Any two regions are either disjoint or they overlap along a shared portion of their boundary.

The alpha shapes model captures the connectivity of the convex regions in a combinatorial object called the *dual complex*. The complex is composed of *simplices* (singular *simplex*): vertices, edges, triangles, and tetrahedra, in dimensions 0 to 3, respectively. Given a molecule, the vertices of the dual complex are simply the centers of the atoms. If a pair of convex cells have a common intersection, we place an edge between their respective atom centers.

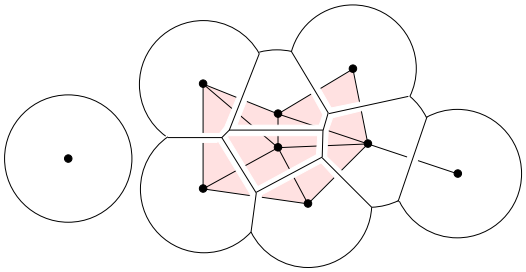

Fig. 1. The space filling diagram and its dual complex. As common in biology, we model the atoms of a molecule as intersecting balls whose union forms the space filling diagram. In the two-dimensional example shown here, the solid line segments inside the disk separate the Voronoi regions, decomposing the union of discs into convex cells. The dual of this decompostion is called the *dual complex*. Duality means that we draw an edge between two atom centers if their convex cells share an edge, and include the triangles formed by these edges in the dual complex if their corresponding atoms (disks) share a common intersection (shaded triangles). The dual complex contains all information about a molecule required to compute its surface and volume.

Similarly, we generate a triangle between the centers of three intersecting cells, and a tetrahedra for four intersecting cells. We show the dual complex for our toy molecule in Figure 1. In general, at most four Voronoi regions (three in two dimensions) can have a non-empty common intersection. Degenerate cases arise in practice, however, but we may eliminate them via computational techniques [43].

The dual complex is a *simplicial complex*: any two simplices are either disjoint or they intersect among a common simplex of lower dimension. For example, two triangles always intersect along an edge or by sharing a common vertex. The dual complex is *dual* to the union of balls, as it is connected the same way. Therefore, we may use it to represent any union of balls, be it a molecule, or a molecule undergoing growth.

3.2 Alpha Complex

We now describe the growth model that the alpha shapes method utilizes. The idea is to increase the squared radius of every ball by any real number α^2. This strategy ensures that the Voronoi regions do not change and the complex only grows as the balls are expanded. We call the dual complex K_α of balls expanded by α^2 an *alpha complex* [43]. When $\alpha^2 = -\infty$, the balls are imaginary and the alpha complex $K_{-\infty}$ is empty. When $\alpha^2 = \infty$, the alpha complex K_∞ is the *Delaunay complex*, D, a well-studied object in computational geometry [44]. In between, we always get a subset of the Delaunay complex, called an alpha complex. This is the key to the efficiency of the alpha shapes method: we only need to compute the Delaunay complex once, and then determine the α at

which a simplex enters the alpha complex. We show a few alpha complexes for the protein 1ctf in Figure 2. Alpha complexes are a multi-scale representation

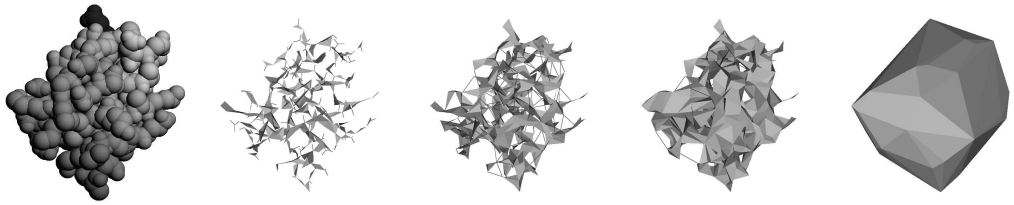

Fig. 2. The van der Waals model of 1ctf (right) and four of its 13,081 alpha complexes with increasing alpha. All alpha complexes are subcomplexes of the Delaunay complex D of the balls, shown on the right.

of the connectivity of a molecule and give us effective control over the size of the features we wish to consider in computing the scoring function.

4 Method

Our aim is to study the influence of the size and quality of the database of residue pairs on the performance of knowledge-based scoring functions. In this section, we begin by describing the set of protein structures we utilize as input. We then describe our method for extracting different databases of residue pairs from this input, as well as computing scoring functions based on these databases. Finally, we discuss our verification procedures for the scoring functions.

4.1 Protein Structure Database

The majority of the sequences of the proteins included in the Protein Data Bank (PDB) [45] are very similar to each other. To minimize the effect of this redundancy, we select a subset of high-quality unique representatives using the BLAST [46] E-value as a measure of protein sequence similarity. The representative subset is a selection of protein domains defined at the 0.0001 level of sequence similarity. That is, the similarity between any two sequences in this subset has an E-value greater than 0.0001. This subset includes 2,145 protein domains from SCOP 1.55 [47]. It can be retrieved from the ASTRAL compendium [48].

4.2 Databases of Residue Pairs

We use the C_α atom to represent each residue in a protein. The protein structure is consequently modeled as a union of balls centered at the C_α, with radius $r = 2$Å. For each protein domain in the database, we compute the Delaunay and alpha complexes using the alpha shapes software [43]. We then extract the edges, triangles, and tetrahedra of an alpha complex. Currently, we only use the edges to compute pair potentials, but we plan to examine the viability of three-body and four-body potentials in the future. We establish two databases. The $\Omega_{<d}$ database contains all residue pairs with distance lower than a cutoff distance d. The Ω_α database contains the subset of pairs that occur as edges in the alpha complex K_α. We do not include pairs consisting of same or adjacent residues in either database. This is equivalent to eliminating pairs at topological levels zero and one, using Sippl's terminology [9]. For our experiments, we choose $d = 20$Å and set $\alpha = 10$Å to generate pairs of similar maximum distance.

We store the data raw, binning only up to the significant digits, using a simple and flexible hashing scheme [49]. Each pair (i, j) in a database contributes a data point specifying its type (S_i, S_j) and distance d_{ij}. There are a total of $\binom{20}{2} + 20 + 1 = 211$ pair types, the last being the "unspecified" type. We store each data point twice, once for its type and once for the unspecified type. While this doubles the size of the database, it allows for efficient calculation of probabilities.

4.3 Random Databases

To verify that our geometric filter has selection power, we also perform experiments using randomly designed databases. Using a random selection procedure, we establish databases of approximately the same size as Ω_α. We then derive the score for a protein by computing Equation (5) over a set of randomly chosen residue pairs, using the random database as reference. Again, we choose approximately the same number of pairs as selected by the alpha shapes method.

4.4 Computing the Scoring Function

Our databases contain probability density distributions for each type of residue pairs, including the unspecified one. When computing probabilities, we use a bucket of size 0.1Å. We have experimented with other bucket sizes to check the robustness of the method. Given a sequence S in conformation C, we first

obtain two list of interacting pairs, representing the all pair and alpha complex pairs, respectively. We then compute Equation (5) over the pairs in each list using the corresponding databases $\Omega_{<d}$ and Ω_{α} to arrive at the scores for that conformation.

4.5 Testing the Potentials

Protein decoy sets greatly facilitate the verification of a newly developed energy function [4]. A decoy set usually contains a few to thousands computer-generated models for a given protein structure. We test a scoring function by measuring its ability to identify near-native conformations among all non-native decoys. In this study, we use two measures of performance, namely the rank scores and correlation coefficients. The *rank score (RS)* is simply the rank in cRMS of the structure with the best score. A perfect energy function would always give the lowest score to the structure that is the closest to the native structure (RS = 1.) The *correlation coefficient (CC)* is Pearson's coefficient between the scores and cRMS values of the model structures being tested. Again, a good scoring function should have a high correlation coefficient, either negative or positive.

In this study, we focus on the 4state-reduced set of decoys [4] that can be obtained from the Decoys 'Я' Us website: http://dd.stanford.edu [35]. This set has been, and remains widely used as a test set for new potential (see for example [50], and references therein). It contains decoys for seven proteins: 1ctf, 1r69, 1sn3, 2cro, 3icb, 4pti and 4rxn. The decoy set for each protein contains the native structure of the protein and between 600 and 700 computer-generated models covering a large range of cRMS distance to the native structure, including many conformations within 4Å cRMS and several within 2Å cRMS. The decoys were generated geometrically through random modifications of the dihedral angles in the inter secondary structure regions of the native protein structures [4]. As such, these decoys are unlikely to be biased toward any particular energy function.

5 Results and Discussion

Recall that pairwise knowledge-based energy functions rely on an independence assumption on all pairs of atoms in a protein. This assumption has been proven not to be correct [32]. We aim to learn how much bias this assumption introduces in energy functions. We are also interested in defining subsets of residue pairs that contain the same information as the full set of pairs. In this section, we begin by characterizing the database of all residue

pairs. We then discuss the different subsets obtained by geometric filtering using the concept of alpha shapes, described in Section 3. Finally, we extract energy functions from our databases and compare their discrimination power using decoy sets.

5.1 Databases: All Pairs versus Alpha Shapes

All distance-dependent, pairwise potentials ignore interactions beyond a certain distance cutoff value d on the basis that long range interactions are not residue specific and are mainly determined by solvation effects. Sippl and Jaritz studied the dependence of the performance of their pairwise statistical potentials on the distance cutoff [51]. They showed that the predictive power of their potential was low for short cutoff distances ($d < 10$Å), and increased slowly and steadily for larger cutoff distances. They concluded that setting $d = 20$Å yields the best results for successful discrimination of the native fold of a protein from non-native conformations. This cutoff distance has been subsequently used for residue based potentials [51] as well as for all-atom potentials [21]. The number of residue pairs increases quadratically as we increase the cutoff distance d (illustrated in Figure 3). For $d = 20$, $\Omega_{<20}$ contains 29,654,812 pairs. We expect however that the dependence among residue pairs also increases as the cutoff distance is increased. Furuichi and Koehl [32] have shown that statistical potentials have a memory of the size of the proteins included in their databases for $d > 8$Å. Jones et al. [3] and Melo and Feytmans [20] recognized this implicitly by defining their statistical potentials for distances up to 10Å. We wish to filter the set of all pairs of residues so that we maintain a minimal set of nearly independent pairs that characterize the whole structure. At the same time, we hope to avoid the loss of information that occurs when using short cutoffs. We propose to use the method of alpha shapes to achieve both goals.

Alpha shapes select relatively independent pairs by construction as it is always a subset of the Delaunay complex D. An important geometric property of D is that any edge can be surrounded by an empty sphere, that is, one that does not contain atoms uninvolved in the creation of the edge. The Delaunay simplices can then be seen as local independent descriptors of the geometry of the protein of interest. The different alpha complexes define approximations of the shape of the protein, with increasing level of accuracy as we decrease the value of α (see Figure 2.) For the simple case where all atom balls have the same radius r, the length d_α of the longest possible edge contained in the alpha complex K_α is defined by:

$$d_\alpha^2 = 4\left(r^2 + \alpha^2\right).$$

For large α, $d_\alpha \simeq d/2$. Note that the edges of the alpha complex K_α constitute a subset of the set of all residue pairs with length smaller than d_α. In Figure 3 we compare the size of the databases $\Omega_{<d}$ and Ω_α when $d/2 = \alpha$. The figure

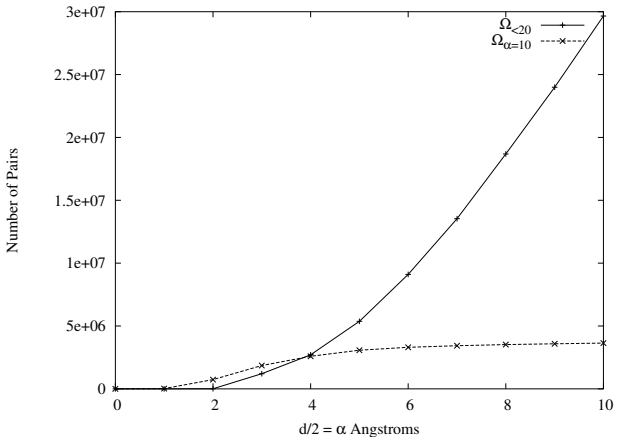

Fig. 3. A comparison of the size of databases $\Omega_{<d}$ and Ω_α with $d/2 = \alpha$.

shows that the number of alpha complex edges in Ω_α increases at a much smaller rate than the corresponding set of all pairs $\Omega_{<d}$. The set $\Omega_{\alpha=10}$ contains 3,643,018 pairs, representing only 12.8% of the full set $\Omega_{<20}$.

We assess the impact of our geometric filtering by comparing the probability density distributions extracted from the databases $\Omega_{\alpha=10}$ and $\Omega_{<20}$, as shown in Figure 4. Intrinsically, the alpha shapes method captures local information, that is, pairs with a small inner-distance, and few pairs with a longer inner-distance. These few pairs however capture significant information about the overall structure of the protein. In the rest of this section, we investigate the quality of this filtering, and the degree to which geometric filtering distills structural information contained in distance pairs.

5.2 Predictive Power of Filtered Statistical Potentials

There are two questions that we can ask about scoring functions: what is the best structure they can pick in a large set of computer generated decoys, and more generally, how well can they identify near native conformations among these decoys? The rank score (RS) of the energy of the native structure provides a quantitative measure to answer the first question, while the correlation coefficient (CC) measures how well a scoring function discriminates the near native conformations. We utilize these two measures to compare the quality of the scoring functions derived from the all-pair database $\Omega_{<20}$, and the geometrically filtered database $\Omega_{\alpha=10}$.

We employ the 4state-reduced decoys of Park and Levitt [4] we introduced in

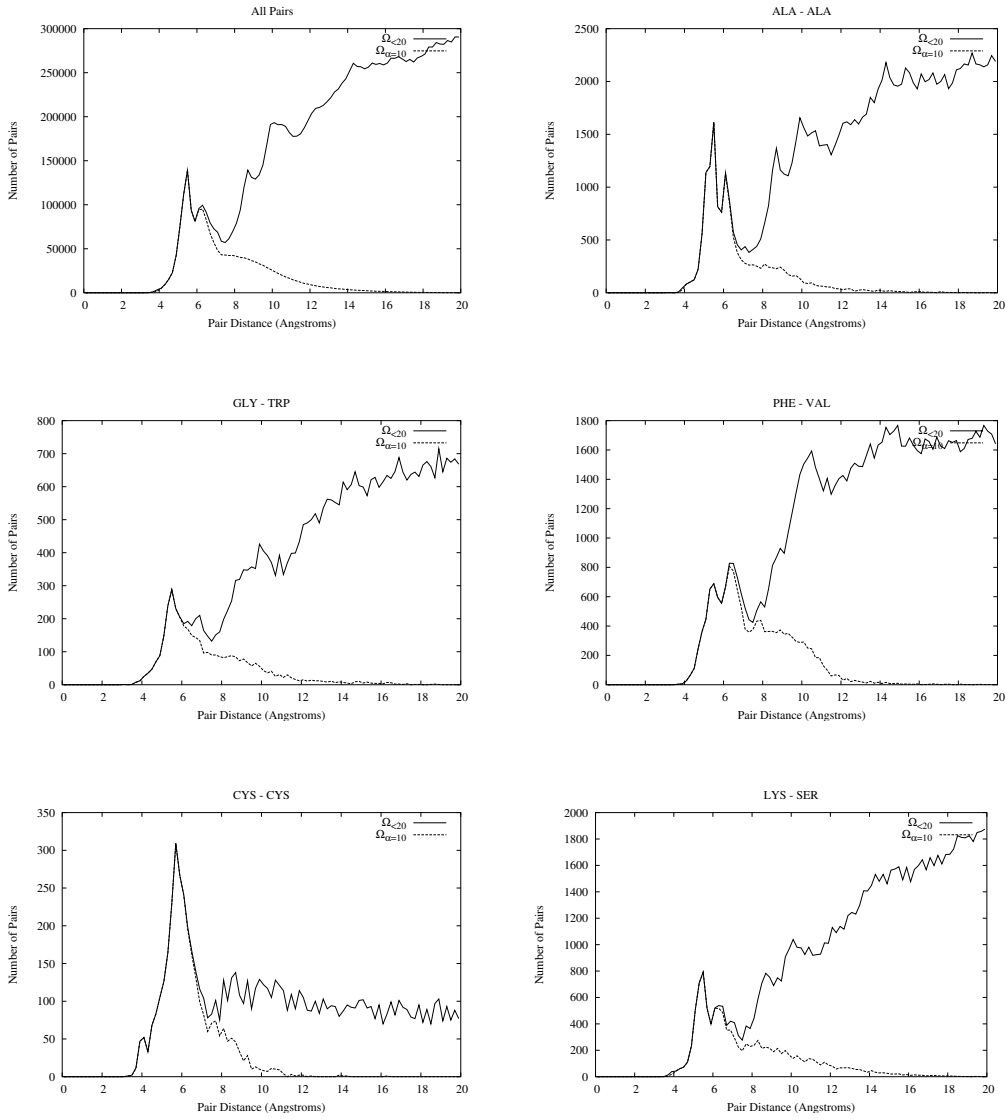

Fig. 4. Comparison of a few probability density distributions contained in the pairs databases $\Omega_{<20}$ and $\Omega_{\alpha=10}$.

Section 4 as our test sets. Figure 5 shows the rank score/cRMS scatter plots. Table 1 summarizes the results for correlation coefficients, and Table 2 lists the rank scores of the native conformation, as well as our two scoring functions.

We compared the discriminative power of the two scoring functions presented in this paper with those of ΔG^{env}, a physics-based free energy function that quantify the environment of each atom [52], of RAPDF, an all-atom pairwise statistical potential [21], and of PAT/ME, a statistical atomic environment potential [50] (Table 1 and table 2). The two scoring functions presented in this paper have only average discrimination powers. The Score versus cRMS plots shown in Figure 5 have the classical "sheep" appearance, and the difference

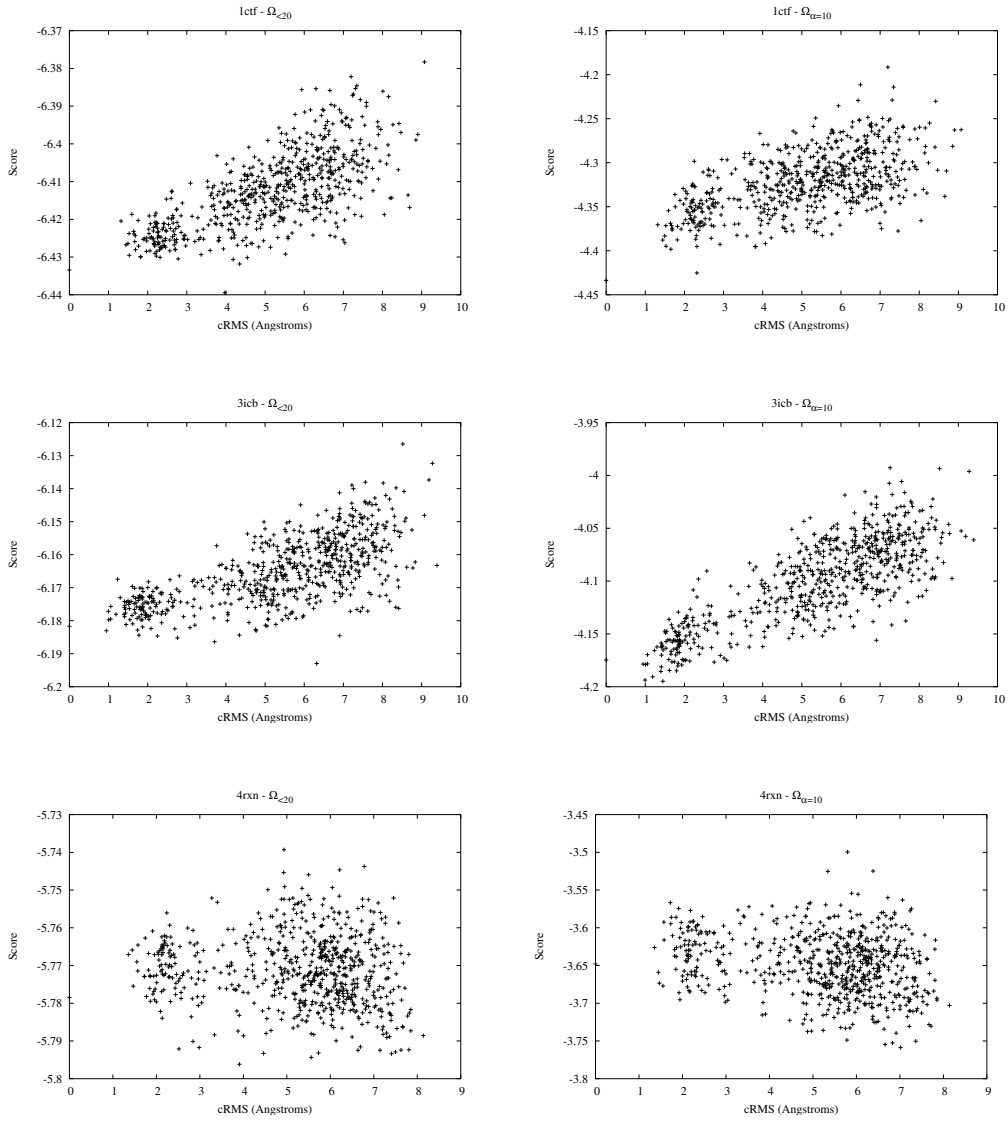

Fig. 5. Rank score/cRMS correlation plots for three proteins of the 4state-reduced decoy set. We compute scores using the $\Omega_{<20}$ database on the left, and $\Omega_{\alpha=10}$ on the right.

between near-native, and non native structures is not optimal. Both RAPDF and PAT/ME performs better than the scoring functions considered here; several other statistical potentials also perform better (see [53,26,54] and [50] for an overview of the performance of statistical potentials on the 4state-reduced decoy sets). The correlation coefficients between score and cRMS (Table 1) are good but not optimal for 1ctf and 3icb, average for 1r69 and 2cro, and not significant for 1sn3, 4pti and 4rxn. We observe the same behavior when we compare the decoys with lowest energy selected by the two scoring functions (Table 2). Neither of the two functions identify the best decoys (in terms of cRMS distance to the native structure) as the structure with the lowest

protein	size	# decoys	type	correlation (CC)				
				$\Omega_{<20}$	$\Omega_{\alpha=10}$	ΔG^{env}	RAPDF	PAT/ME
1ctf	74	630	$\alpha+\beta$	0.70	0.56	0.28	0.73	0.72
1r69	69	675	α	0.31	0.43	0.25	0.7	0.68
1sn3	65	660	α	-0.04	0.002	-0.15	0.47	0.36
2cro	75	674	α	0.32	0.52	0.19	0.76	0.65
3icb	75	653	α	0.66	0.79	0.41	0.85	0.8
4pti	58	687	small	0.18	0.04	0.1	0.47	0.36
4rxn	54	677	small	-0.08	-0.22	0.26	0.58	0.39

Table 1
Protein size, type, number of decoys, and score/cRMS correlation for our two databases for the 4state-reduced decoy sets. For comparison, we provide the score/cRMS correlations obtained using the functions ΔG^{env} [52], RAPDF [21], and PAT/ME [50]

protein	best	cRMS		rank (RS)				
	cRMS	$\Omega_{<20}$	$\Omega_{\alpha=10}$	$\Omega_{<20}$	$\Omega_{\alpha=10}$	ΔG^{env}	RAPDF	PAT/ME
1ctf	1.32	3.97	**2.32**	156	57	1	2	4
1r69	0.88	5.35	**4.47**	410	268	8	8	72
1sn3	1.31	6.56	**6.36**	428	383	354	24	69
2cro	0.81	4.19	**1.87**	246	24	314	6	6
3icb	0.94	6.32	**1.45**	387	16	5	26	26
4pti	1.41	6.28	**4.75**	424	169	187	158	129
4rxn	1.36	3.91	**7.01**	140	606	302	14	9

Table 2
The best (lowest) cRMS for each decoy set, and the cRMS and rank for the decoy selected by our scoring functions, computed using the full $\Omega_{<20}$ and filtered $\Omega_{\alpha=10}$ databases. For comparison, we provide the cRMS rank of the best decoy identified by ΔG^{env} [52], RAPDF [21], and PAT/ME [50]

energies, for all seven proteins decoy sets. The scoring function defined from the filtered database $\Omega_{\alpha=10}$ selects better structures than the scoring function defined from the all pair database, but we expect this difference to be marginal, as the corresponding correlation coefficients are not significantly different. We have included a comparison over a wide range of scoring functions: from physics-based (ΔG^{env}), pairwise atom-based (RAPDF) or residue-based (the two functions developed in this study, to multi-body, atom-based (PAT/ME). It is worth mentioning that none of these functions clearly out-

perform the other functions on all test cases included in the 4state decoy sets. This was also observed by Summa and colleagues [50]. The current state of the art in applying statistical potentials to analyse large decoy sets calls for the application of multiple functions and averaging out the results.

Our goal in this study, however, is not to derive a new potential that outperforms existing potentials. We are interested, rather, in understanding the effect of including correlated residue pairs in the databases used to generate statistical energy functions. The main result we observe from Figure 5 and Tables 1 and 2 is that the scoring function derived from $\Omega_{\alpha=10}$ performs competitively with the one derived $\Omega_{<20}$, even though the former uses only 12.8% of the data. To substantiate this result, we have to show that the same is not true for *any* random subset. To do so, we perform a number of random trials as follows. We establish ten databases using residue pairs chosen uniformly at random from all pairs, so that each database contains approximately the same number of pairs as the alpha shape database Ω_α. For each protein structure in a decoy set, we compute ten sets of randomly chosen pairs, once again of the same size as the set filtered by alpha shapes. We then compute scores for each set using each of the ten random databases, leading to 100 trials in total. We show the correlation coefficients (CC) for each trial in Figure 6 for the decoy set 1ctf. Note that the random statistical scores achieve an average correlation coefficient of 0.23, less than half the correlation coefficient obtained for the scoring function derived from the geometric filtered database. Therefore, the random tests show that our geometric filtering method does have selection power.

5.3 Databases and Statistical Scoring Functions

Distance-dependent pair potentials are the most common potentials used for fold recognition [3, 19], protein structure assessment [21, 55], and for ab initio protein structure prediction. The critical parameters that usually define such potentials include the set of proteins used to derive the databases of pairs, the range of distances considered, and the resolution used to represent the protein structures (from C_α only to all atoms) [56].

All of the experimental protein structures included in the PDB are legitimate sources of information for statistical potentials, as they all correspond to native protein structures. There is however a high level of redundancy in the PDB, and protein sets used to build statistical potentials are usually filtered to include only proteins with low levels of sequence similarity. For example, in this study we use a subset of all SCOP domains in which no two sequences have a similarity better than an E-value of 0.0001. Filters based on sequence similarity are not enough, however. We know that the discrimination power of

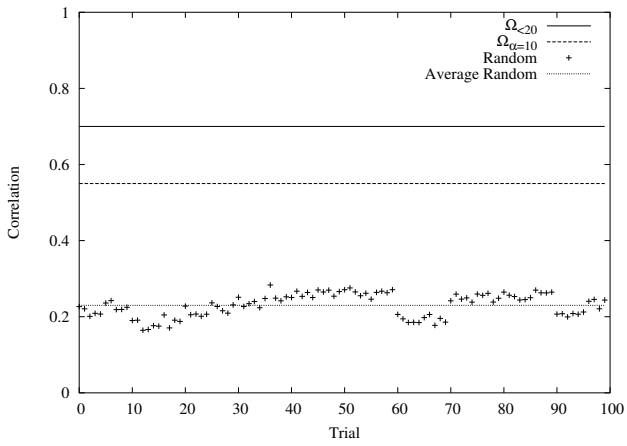

Fig. 6. Random scoring functions. We establish ten databases Ω_i of the same size as $\Omega_{\alpha=10}$ using pairs chosen uniformly at random from the database containing all residue pairs of proteins. We then obtain ten random sets Ω_j of residue pairs for 1ctf such that each set has the same size as the set of edges of the alpha complex for 1ctf with $\alpha = 10$. We plot the correlation coefficient of the score and cRMS values for each of the 100 trials $\{\Omega_i, \Omega_j\}$. For comparison, we also plot the correlation coefficient obtained for the scoring functions derived from the all pair database $\Omega_{<20}$ (solid) and from the alpha-complex database $\Omega_{\alpha=10}$ (dashed).

statistical potentials depends both on the number of structures used in their derivation, as well as on the size of structures themselves [30, 32]. The dependence on the number of proteins can be explained by statistical consideration as the sampling improves with larger protein sets. The dependence on the length of the proteins, on the other hand, is more complex and is related to the definition of the range (or cutoff distance) used to compute the statistical potential. Most of the short range interactions of an atom in a protein involve other atoms of the same proteins, while long range interactions include mostly solvent interactions. The separation distance d_s between "short range" and "long range" interactions depends on the size of the protein considered: small proteins will have a smaller d_s than large proteins. This is consistent with the fact that the optimal cutoff d is smaller for small proteins (around 20Å) than for long proteins (around 30Å). The difference between potentials derived from small proteins and from large proteins disappears when a short cutoff distance ($d < 8$) is used, at a significant cost however in discrimination power [32].

While long distance cutoffs define potentials with better abilities to discriminate between good and bad models, they raise the question of the validity of the hypothesis on which these potentials are designed. Probably the strongest assumption considered is the independence of residue pairs in proteins, which has already been shown not to be valid [30, 32]. We have shown in this study that we can reduce the effect of data dependence on the potential by using a geometric filter on the sets of all residue pairs representing a protein structure.

By construction, the alpha shape complex selects the minimal set of residue pairs that characterizes the protein shape. The size of the set of all pairs of atoms is on the order of n^2, where n is the number of atoms in the protein. The size of the alpha complex, in contrast, is generally on the order of n. Interestingly, the reduction in size does not translate into loss of discrimination power (see Tables 1 and 2.)

We have presented results for scoring functions derived from a representation of protein structure based on C_α only. We have, however, repeated all the same studies with more detailed representations of the proteins (inclusion of all atoms of the backbone, as well as all-atom models), and reached similar conclusions (data not shown).

6 Conclusion

The recent literature on distance-dependent, pairwise statistical potentials and their application to protein structure modeling makes no secret of their limitations and problems. In particular, there are known concerns on the influence of the sets of proteins used to derive the potentials, on the range of distances included in the definition of the potentials, as well as on the resolution used to represent the protein structures. Most attempts to derive potentials that are mathematically or physically sound have unfortunately led to potentials with less discriminative power. As a consequence, a pragmatic approach has emerged, in which the problems are acknowledged, but the potentials are designed for efficiency rather than generality. For example, it was suggested that size should be taken into account explicitly when testing a protein model, by deriving the scoring potential from native proteins of similar size [56]. Similarly, it was shown that statistical potentials derived from mainly α proteins will work better on assessing mainly α proteins [32].

In this paper, we focused on the definition of databases of residue pairs for deriving distance-dependent, pairwise scoring functions. We proposed to filter the set of all residue pairs using the concept of alpha shapes, which provides hierarchical shape descriptors for protein structures. We showed that this filtering leads to a significant reduction of the number of pairs without loss of information and discrimination power. The filtering is simple to implement and may be used automatically for all types of pairwise statistical potentials.

Filtering the database of all residue pairs using the alpha shape complex does not deteriorate, but does not improve the discrimination power of pairwise potentials. The alpha shape descriptors provide information on residue triplets (measured through the surface of the corresponding triangles), and residue quadruplets (measured through the volume of the corresponding tetrahedra).

We are currently working on integrating all this information into a general scoring function that captures the geometry of the protein structures.

Acknowledgments

Research of all authors is partially supported by NSF under grant CCR-00-86013.

References

[1] W. Browne, A. North, D. Philipps, K. Brew, T. Vanaman, R. Hill, A possible three dimensional structure of bovine alpha-lactalbumin based on that of hen's egg white lysozyme, J. Mol. Biol. 42 (1969) 65–86.

[2] J. Bowie, R. Luthy, D. Eisenberg, A method to identify protein sequences that fold into a known three-dimensional structure, Science 235 (1991) 164–170.

[3] D. Jones, W. Taylor, J. Thornton, A new approach to protein fold recognition, Nature 358 (1992) 86–89.

[4] B. Park, M. Levitt, Energy functions that discriminate x-ray and near-native folds from well constructed decoys, J. Mol. Biol. 258 (1996) 367–392.

[5] J. Moult, Comparison of database potentials and molecular mechanics force fields, Curr. Opin. Struct. Biol. 7 (1997) 194–199.

[6] T. Halgren, Potential energy functions, Curr. Opin. Struct. Biol. 5 (1995) 205–210.

[7] S. Tanaka, H. Scheraga, Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins, Macromolecules 9 (1976) 954–960.

[8] S. Miyazawa, R. Jernigan, Estimation of effective inter-residue contact energies from protein crystal structures: quasichemical approximations, Macromolecules 18 (1986) 534–552.

[9] M. Sippl, Calculation of conformational ensembles from potentials of mean force, J. Mol. Biol. 213 (1990) 859–883.

[10] M. Sippl, Knowledge-based potentials for proteins, Curr. Opin. Struct. Biol. 5 (1995) 229–235.

[11] S. Miyazawa, R. Jernigan, Residue-residue potentials with a favorable contact pair term and an unfavorable high packing term, for simulation and threading, J. Mol. Biol. 256 (1996) 623–644.

[12] R. Jernigan, I. Bahar, Structure-derived potentials and protein simulations, Curr. Opin. Struct. Biol. 6 (1999) 195–209.

[13] T. Laziridis, S. Gill, Effective energy functions for protein structure prediction, Curr. Opin. Struct. Biol. 10 (2000) 139–145.

[14] V. Maiorov, G. Crippen, Contact potential that recognizes the correct folding of globular proteins, J. Mol. Biol. 227 (1992) 876–888.

[15] L. Mirny, E. Shakhnovich, How to derive a protein folding potential? a new approach to an old problem, J. Mol. Biol. 264 (1996) 1164–1179.

[16] K. Koretke, Z. Luthey-Schulten, P. Wolynes, Self-consistently optimized energy functions for protein structure prediction by molecular dynamics, Proc. Natl. Acad. Sci. (USA) 95 (1998) 2932–2937.

[17] M. Hao, H. Scheraga, Designing potential energy functions for protein folding, Curr. Opin. Struct. Biol. 9 (1999) 184–188.

[18] Y. Xia, M. Levitt, Extracting knowledge-based energy functions from protein structures by error rate minimization: comparison of methods using lattice model, J. Chem. Phy. 113 (2000) 9318–9330.

[19] M. Sippl, S. Weitckus, Detection of native like models for amino acid sequences of unknown three dimensional structure in a data base of known protein conformations, Proteins: Struct. Func. Genet. 13 (1992) 258–271.

[20] F. Melo, E. Feytmans, Novel knowledge-based mean force potential at atomic level, J. Mol. Biol. 267 (1997) 207–222.

[21] R. Samudrala, J. Moult, An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction, J. Mol. Biol. 275 (1998) 895–916.

[22] A. Kolinski, A. Godzik, J. Skolnick, A general method for the prediction of the three dimensional structure and folding pathway of globular proteins: application to designed helical proteins, J. Chem. Phy. 98 (1993) 7420–7433.

[23] R. Singh, A. Tropsha, I. Vaisman, Delaunay tessellation of proteins: Four body nearest-neighbor propensities of amino acid residues, J. Comp. Biol 3 (1996) 213–221.

[24] M. Delarue, P. Koehl, Atomic environment energies in proteins defined from statistics of accessible and contact surface areas, J. Mol. Biol. 249 (1995) 675–690.

[25] F. Eisenhaber, Hydrophobic regions on protein surfaces: derivation of the solvation energy from their area distribution in crystallographic protein structures, Protein Sci. 5 (1996) 1676–1686.

[26] B. McKonkey, V. Sobolev, M. Edelman, Discrimination of native protein structures using atom-atom contact scoring, Proc. Natl. Acad. Sci. (USA) 100 (2003) 3215–3220.

[27] R. DeWitte, E. Shakhnovich, Pseudodihedrals: simplified protein backbone representation with knowledge based energy, Protein Sci. 3 (1994) 1570–1581.

[28] S. Bryant, C. Lawrence, The frequency of ion pair substructures is quantitatively related to the electrostatic potential: a statistical model for nonbonded interactions, Proteins: Struct. Func. Genet. 9 (1991) 108–119.

[29] J. Kocher, M. Rooman, S. Wodak, Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches, J. Mol. Biol. 235 (1994) 1598–1613.

[30] P. Thomas, K. Dill, Statistical potentials extracted from protein structures: How accurate are they?, J. Mol. Biol. 257 (1996) 457–469.

[31] A. BenNaim, Statistical potentials extracted from protein structures: are they meaningful potentials?, J. Chem. Phy. 107 (1997) 3698–3706.

[32] E. Furuichi, P. Koehl, Influence of protein structure databases on the predictive power of statistical pair potentials, Proteins: Struct. Func. Genet. 31 (1998) 139–149.

[33] B. Reva, A. Finkelstein, M. Sanner, A. Olson, Residue-residue mean force potentials for protein structure recognition, Protein Eng. 10 (1997) 865–876.

[34] S. Miyazawa, R. Jernigan, An empirical energy potential with a reference stat for protein fold and sequence recognition, Proteins: Struct. Func. Genet. 36 (1999) 357–369.

[35] R. Samudrala, L. M., Decoys 'r' us: A database of incorrect protein conformations for evaluating scoring functions, Protein Sci. 9 (2000) 1399–1401.

[36] J. Tsai, R. Bonneau, A. Morozov, B. Kuhlman, C. Rohl, D. Baker, An improved protein decoy set for testing energy functions for protein structure prediction, Proteins: Struct. Func. Genet. 53 (2003) 76–87.

[37] C. Keasar, M. Levitt, A novel approach to decoy set generation: Designing a physical energy function having local minima with native structure characteristics, J. Mol. Biol. 329 (2003) 159–174.

[38] D. Shortle, K. T. Simons, D. Baker, Clustering of low-energy conformations near the native structures of small proteins, Proc. Natl. Acad. Sci. (USA) 95 (1998) 11158–62.

[39] Y. Zhang, J. Skolnick, Spicker: A clustering approach to identify near-native protein folds, J. Comput. Chem. 25 (2004) 865–871.

[40] K. Wang, B. Fain, M. Levitt, R. Samudrala, Improved protein structure selection using decoy dependent discriminatory functions, BMC Struct. Biol. 8 (2004) 4.

[41] T. Creighton, Proteins. Structures and Molecular Principles, Freeman, New York, NY, 1984.

[42] H. Edelsbrunner, The union of balls and its dual shape, Discrete Comput. Geom. 13 (1995) 415–440.

[43] H. Edelsbrunner, E. P. Mücke, Three-dimensional alpha shapes, ACM Trans. Graphics 13 (1994) 43–72.

[44] M. de Berg, M. van Kreveld, M. Overmars, O. Schwarzkopf, Computational Geometry: Algorithms and Applications, Springer, New York, 1997.

[45] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, J. Shindyalov, P. Bourne, The protein data bank, Nucl. Acids. Res. 28 (2000) 235–242.

[46] S. Altschul, W. Gish, W. Miller, E. Myers, D. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.

[47] A. Murzin, S. Brenner, T. Hubbard, C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures, J. Mol. Biol. 247 (1995) 536–540.

[48] J. Chandonia, N. Walker, L. Lo conte, P. Koehl, M. Levitt, S. Brenner, Astral compendium enhancements, Nucl. Acids. Res. 30 (2002) 260–263.

[49] T. H. Cormen, C. E. Leiserson, R. L. Rivest, Introduction to Algorithms, The MIT Press, Cambridge, MA, 1994.

[50] C. M. Summa, M. Levitt, W. F. DeGrado, An atomic environment potential for use in protein structure prediction, J. Mol. Biol. 352 (2005) 986–1001.

[51] M. Sippl, M. Jaritz, Predictive power of mean force pair potentials., in: H. Bohr, S. Brunak (Eds.), Proteins Structure by Distance Analysis, IOS Press, Amsterdam, 1994, pp. 113–134.

[52] P. Koehl, M. Delarue, Polar and non-polar atomic environment in the protein core: implications for folding and binding, Proteins: Struct. Func. Genet. 20 (1994) 264–278.

[53] H. Zhou, Y. Zhou, Distance scaled, finite ideal-gas reference state improves structure derived potentials of mean force for structure selection and stability prediction, Protein Sci. 11 (2002) 2714–2726.

[54] H. Zhou, Y. Zhou, Quantifying the effect of burial of amino acid residues on protein stability, Proteins: Struct. Func. Genet. 54 (2004) 315–322.

[55] F. Melo, E. Feytmans, Assessing protein structures with a non local atomic interaction energy, J. Mol. Biol. 277 (1998) 1141–1152.

[56] F. Melo, R. Sanchez, A. Sali, Statistical potentials for fold assessment, Protein Sci. 11 (2002) 430–448.