BIOINFORMATICS

Vol. 00 no. 00 2008
Pages 1–7

A Hierarchy of Delaunay Tessellation-based Scoring Functions for Protein Fold Recognition

Andrew H. Fowler†, Bala Krishnamoorthy*, Kelly Stratton*, and Christopher Deutsch‡

†- Schweitzer Engineering Laboratories, Pullman; * Department of Mathematics, Washington State University; ‡ Department of Chemistry, Portland State University.

ABSTRACT

Motivation: The key advantage of scoring functions based on the Delaunay tessellation (DT) of proteins used for fold recognition and mutagenesis is the use of higher order contact terms (four rather than two). The use of lower order contacts along with four-body terms could improve the accuracy of such scoring functions. Analysis of surface exposure of the contacts *within* the DT framework is of independent interest, and could also be critical for the performance of DT-based scoring functions.

Results: We propose a unified framework for defining two, three, and four body amino acid (AA) contacts in proteins based on their Delaunay tessellations. Similar to four body scoring functions defined previously, we represent each AA by a single point, and distinguish all contacts based on their amino acid composition and back-bone chain connectivity. In addition, we define *degrees of buriedness* for the two and three body contacts under the same framework. There are four degrees of buriedness for two body contacts, and nine degrees for three body contacts, varying from completely non-buried (i.e., fully exposed to the surface) to completely buried. On an average, the accuracies of scoring functions for distinguishing native structures from decoys are higher when using higher order terms. More interestingly, the scoring function that combines three body terms with degrees of buriedness *and* four body terms is the most accurate, thus demonstrating the importance of lower order contacts and buriedness. We also use the three body contacts to define a scoring function for predicting the effects of mutagenesis on solubility of proteins.

Availability: Executables of programs, datasets of mutants, and summaries of scores for decoy sets, are available from the author's web page: http://www.wsu.edu/~kbala/DelaunayPots.html.

Contact: kbala@wsu.edu

1 INTRODUCTION

Correlations between sequence and structure are widely believed to be the key determinants of how proteins fold, and also how they function. Working under this premise, most computational methods used for structure and function prediction employ *scoring functions* that quantify the propensities of groups of amino acids to form specific structural or functional units. For instance, the accuracy of most *in silico* protein folding techniques such as homology modeling, molecular dynamics simulations, and threading, depends critically on an underlying potential that distinguishes correct structures from incorrect conformations. Many such potential functions have been proposed and studied over the last three decades – Sippl (1995), Wodak and Rooman (1993), Park and Levitt (1996), and Park et al. (1997) review several of them. Scoring functions for mutagenesis predict the effects of changing one or more amino acids (AAs) on various aspects of protein function – stability (Gilis and Rooman, 1997; Cheng et al., 2006; Deutsch and Krishnamoorthy, 2007), activity (Masso and Vaisman, 2007), solubility (Idicula-Thomas et al., 2006), etc.

Typically, the more detailed the definition of the energy function is, the more computationally expensive it is to score large number of conformations. Potentials using quantum mechanical calculations are highly accurate, but are not tractable when applied to large proteins (Halgren, 1995). Most computationally efficient potentials analyze proteins at the atomic level or at the AA level. Frequencies of AA contact pairs (i.e., two body contacts) have been used as the key factor to define several such potentials (Miyazawa and Jernigan, 1985, 1996; Sippl, 1990; Melo and Feytmans, 1997; Samudrala and Moult, 1998; Seno et al., 1998; Li et al., 2003; Zomorodian et al., 2006). Three body AA contacts have also been used in a few cases (Banavar et al., 2002; Li and Liang, 2005). Independent of two and three body contacts, four body AA contacts have been used to define such potential energies, mainly by employing the concept of Delaunay tessellation (DT) (Edelsbrunner, 2001) of proteins (Singh et al., 1996; Gan et al., 2001; Krishnamoorthy and Tropsha, 2003; Reck and Vaisman, 2007), and more recently, by an approach based on distance calculations (Feng et al., 2007). More local interactions at the atomic level have also been used (by defining atomic *environments*, which often capture solvent accessibility) to build accurate potentials (Delarue and Koehl, 1995; McConkey et al., 2003; Summa et al., 2005). Other properties used to define similar potentials include dihedral angles (DeWitte and Shakhnovich, 1994), ion pair interactions (Bryant and Lawrence, 1991), residue orientations (Buchete et al., 2004), and combinations of several such factors (Park and Levitt, 1996; Park et al., 1997). Correspondences of such potentials derived from databases of proteins to physically meaningful quantities, e.g., the potential energy of the protein or potentials of mean force for AA pairs, are debatable (Thomas and Dill, 1996; Ben-Naim, 1997). Still, these empirical functions are most often justified by their effectiveness.

In spite of the improvements reported in the performances of these scoring functions over the years, most of them suffer from a critical shortcoming – they work well on several classes of proteins, but fail on other classes. This is also the main reason why researchers should continue to focus attention on the design of potential functions. In this paper, we concentrate on scoring functions defined

using the DT of proteins. The main advantage of employing this concept from geometry is that it provides a more robust definition of nearest neighbors than pairwise distance calculations. While many of the erstwhile potential functions concentrated on pairwise contacts, it is natural to expect higher order contacts to carry more information than two body contacts. Further, it has been demonstrated that higher order contacts cannot be modeled by summing up the component pairwise contacts (Ben-Naim, 1997; Li and Liang, 2005). DT of proteins defines clusters of four AAs in contact, thus directly modeling higher order contacts. DT-based four body scoring functions have been shown to be competitive to other two body potentials for decoy discrimination (Zheng et al., 1997; Krishnamoorthy and Tropsha, 2003; Reck and Vaisman, 2007) and for folding simulations (Gan et al., 2001; Krishnamoorthy and Tropsha, 2003). In fact, the robustness of defining contacts using *alpha shapes* of proteins, which is a generalization of its DT, has shown to increase the accuracy of potentials for fold recognition even when using only two body contacts (Li et al., 2003; Zomorodian et al., 2006). DT of proteins has also been widely used as a generic computational tool to analyze various aspects of protein structure: secondary structure assignment (Taylor et al., 2005), structural classification (Bostick et al., 2004; Huan et al., 2005), analysis of small-world nature of protein contacts (Taylor and Vaisman, 2006), computational mutagenesis for protein stability (Carter, Jr et al., 2001; Masso et al., 2006; Deutsch and Krishnamoorthy, 2007) and enzyme activity (Masso and Vaisman, 2007).

Even though the all-atom structure of a protein is more accurate than representing each AA by a single point, the latter approach has its advantages. Apart from being simpler, the unified residue representation can be applied even when the full-atom structure is not available. This representation is also more well-suited for predicting mutagenesis and other similar processes, which result in structural changes (note that the all-atom structure of the mutant is usually not known). In this paper, we introduce a general framework for defining AA contacts using the DT of the unified residue representation of proteins. While four body contacts have been studied using DT (as they are naturally defined by the Delaunay tetrahedra), two and three body contacts have not been studied under the same framework (we do note that two and three body contacts have been studied using alpha shapes (Li et al., 2003; Li and Liang, 2005), which uses the all-atom model of proteins). We also introduce the concept of *degrees of buriedness* for two and three body contacts, which estimates the extent of surface exposure or buriedness of contacts using the DT framework (without measuring the actual surface areas). Once again, we note that the most efficient method for calculating solvent accessible surface areas uses alpha shapes (Edelsbrunner and Koehl, 2005) when working on all-atom models of proteins. At the same time, such surface area calculations do not consider the sequence identity of the AAs involved. On the other hand, some previous studies that included AA identities of the contacts have used arbitrary cut-off values on the associated solvent accessible surface areas to label the contacts as exposed or not (Feng et al., 2007). The degrees of buriedness provides a convenient middle ground for analyzing the AA composition and the buriedness of contacts in the same setting.

We define two and three body scoring functions under the DT framework for decoy discrimination, and compare their performances to that of the previously defined four body scoring function (Krishnamoorthy and Tropsha, 2003). We also consider combinations of these scoring functions. We test the hierarchy of DT-based scoring functions on the Decoys 'R' Us database (Samudrala and Levitt, 2000). Our results show that the accuracy of these scoring functions generally increase with the use of higher order contact terms. Interestingly, a combined three and four body scoring function with degrees of buriedness is the most accurate DT-based scoring function. We also demonstrate the usefulness of the new DT-based framework for predicting mutagenesis effects on protein solubility (the detailed treatment of this topic will be presented in a separate paper).

2 METHODS

Delaunay tessellation is a construct from geometry that defines clusters of nearest neighbor points based on their relative proximity (rather than using distances between them). We refer the reader to one of the textbooks in the area of computational geometry (Edelsbrunner, 2001) for the details of DT, and its dual construct, the *Voronoi diagram*, which defines convex polyhedral regions of space that are closer to the parent point than to other points. With each AA represented by a single point in 3D space, the DT describes the structure of the protein as a collection of space-filling, non-overlapping tetrahedra (see Figure 1 for an illustration in 2D). These tetrahedra naturally define four body AA contacts. We now describe how to define and analyze two and three body Delaunay contacts.

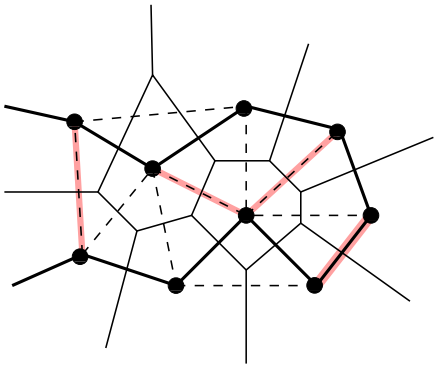

Fig. 1. Delaunay tessellation of a protein in 2D. The dots represent amino acids, and the thick solid line connecting the dots is the backbone. Dotted lines are Delaunay triangles and thin solid lines represent the Voronoi cells. The four shaded edges illustrate the four degrees of buriedness for two body contacts (see Figure 3 and Section 2.2).

2.1 Delaunay Contacts

Each Delaunay tetrahedron naturally defines six edges and four triangles. We define two and three body AA contacts using the Delaunay edges and triangles, respectively. We differentiate the contacts based on their AA composition, without considering the order in which the AAs occur along the protein sequence. The motivation for this definition is evident when one considers contacts formed by AAs that are distant along the backbone chain, but are close to each other in 3D space. Backbone chain connectivity is still an important aspect of the contacts, as demonstrated by the performance of four body scoring functions (Krishnamoorthy and

Tropsha, 2003; Deutsch and Krishnamoorthy, 2007). Similar to the case of four body contacts, we include backbone chain connectivity as a separate factor in the definition of the two and three body contacts. We define two connectivity classes for two body contacts – non-bonded and bonded. Extending the definition to three body contacts, we get three connectivity classes, having zero, one, or two bonded edges in the triangle (see Figure 2). We appropriately index the three body connectivity classes $0, 1, 2$ (and use $0, 1$ for two body connectivity classes). Notice that for the three body connectivity class 1, the bonded edge could either be lower down or higher up along the sequence, i.e., the residue numbers can be $(i, i+1, j)$ or $(i, j, j+1)$, with $i < j$. The reader can guess the extension of this definition to four body contacts, which gives five connectivity classes. This is indeed how the four body contacts were defined previously (Krishnamoorthy and Tropsha, 2003).

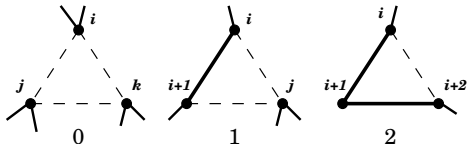

Fig. 2. Backbone connectivity classes for three body contacts. i, j, k etc. are residue numbers. The connectivity indices $(0, 1, 2)$ are ordered from most non-bonded to most bonded, or connected.

2.2 Delaunay Buriedness of Contacts

Surface exposure of AA contacts is typically determined by solvent accessible surface area calculations (Feng et al., 2007). Since we use a unified residue representation, it is more natural to consider levels of surface exposure from a combinatorial point of view. Any two Delaunay tetrahedra from the tessellation are non-intersecting, or intersect at a triangle, or an edge, or just a vertex (residue). Thus, each Delaunay triangle is shared by at most two tetrahedra. We define a triangle to be *Delaunay buried*, or simply *buried*, if it is part of two tetrahedra in the DT of the protein. A triangle that is part of at most one tetrahedron is hence non-buried, or *on the surface*. When a triangle is non-buried, we define each of its three component edges and three vertices as non-buried. To complete the definition, we say that an edge (or a vertex) is buried if it is not non-buried. Notice that the buriedness of two body contacts is defined using the buriedness of the three body contacts of which the former is a component. Thus, a vertex or an edge is non-buried if it is part of at least one non-buried triangle.

Once we have determined whether each vertex, edge, and triangle are buried or non-buried, we can define various levels of buriedness for two and three body contacts. For two body contacts, we define four levels of Delaunay buriedness, based on how many of the three simplices – two vertices and the edge connecting them – are buried. We appropriately index these four buriedness classes by $0, 1, 2$, and 3, based on the number of component simplices that are buried (see Figure 3). We illustrate the occurrences of the two body buriedness classes in 2D in Figure 1. Interestingly, we can define the same four buriedness classes for two body contacts in 3D as well.

We extend the definition of buriedness classes to three body contacts. This classification describes the various ways in which the vertices, edges, and face of each triangle can be located on the surface of the protein (as it has been tessellated). Altogether,

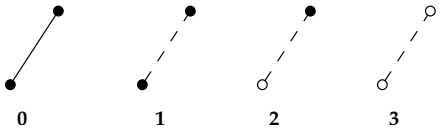

Fig. 3. Buriedness classes for two body contacts. White/dotted elements are buried and black/solid elements are on the surface.

there are nine buriedness classes for three body contacts (Figure 4), indexed 0-8, which range from completely non-buried (class 0) to completely buried (class 8).

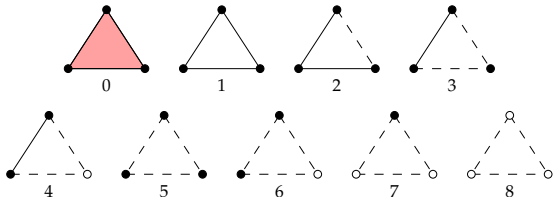

Fig. 4. Three body Buriedness classes. White/dotted elements are buried and black/solid elements are on the surface.

2.2.1 Number of Contacts Since there are twenty AAs, the number of AA pairs is 210 (20 multichoose 2), and the number of triplets is 1540 (20 multichoose 3). Reduced residue representations are possible as well – for instance, Feng et al. (2007) classify the AAs into eight groups to define triplets. We use the default twenty AA representation. Including backbone connectivity classes and buriedness classes, we get $210 \cdot 2 \cdot 4 = 1680$ two body contacts, and $1540 \cdot 3 \cdot 9 = 41580$ three body contacts. This number of three body contacts is comparable to the $8855 \cdot 5 = 44275$ four body contacts used by Krishnamoorthy and Tropsha (2003).

One may consider extending the definition of buriedness classes to four body contacts. First of all, the definition of the tetrahedron itself as buried or non-buried is ambiguous. Hence we could ignore the buriedness of the tetrahedron, and define levels of buriedness for four body contacts based on those of the four component triangles. In this case, we need to analyze *all* possible buriedness combinations of 4 triangles, 6 edges, and 4 vertices. Combined with the number of AA quadruplets (8855) and backbone connectivity types (5), such a buriedness classification results in an impractically large number of four body contacts. At the same time, since buriedness depends inherently on the surface of the protein, it is reasonable to expect the various three body buriedness classes to capture most of the relevant information. As a convenient middle ground, we use a combination of three body contacts with buriedness levels and the original four body contacts (Section 3).

2.2.2 Distance Cutoffs Distance cutoffs are not used to construct the DT. Still, we need to screen the tetrahedra using a preset distance cutoff in order to define biochemically relevant AA contacts. We used a distance cut-off of 9 Angstroms for the 3-body contacts, in order to capture all the relevant surface features of the protein. We developed the entire scoring function using a data-base of sequentially diverse (at most 25% pairwise sequence identity) set of 3988 protein chains, selected by the PISCES server (Wang and

Dunbrack, Jr., 2003). The distributions of various buriedness classes as well as backbone connectivity classes in this data base are shown in Table 1. Notice that the fully surface triangle type (buriedness class 0) is the most frequent buriedness class (24.6%), followed by class 4, which has one edge exposed with the remaining parts of the triangle buried (17.2%). Also notice that the non-bonded class (class 0) is the most common backbone connectivity class.

Table 1. Fractions of triangle counts by backbone and buriedness classes.

$b\backslash c$	0	1	2	sum
0	0.105	0.109	0.032	0.246
1	0.005	0.008	0.001	0.013
2	0.055	0.074	0.015	0.144
3	0.062	0.053	0.009	0.124
4	0.079	0.079	0.014	0.172
5	0.020	0.011	0.002	0.032
6	0.063	0.039	0.006	0.109
7	0.068	0.043	0.006	0.117
8	0.026	0.015	0.002	0.043
sum	0.482	0.431	0.086	1.000

2.3 Defining the Pseudo-Potentials

We generalized the log-likelihood formula presented earlier by Krishnamoorthy and Tropsha (2003) to subsume the three-body case, and added buriedness classes. The new formula is given as follows:

$$Q_{ijk}^{cb} = \log\left[\frac{f_{ijk}^{cb}}{p_{ijk}^{cb}}\right]. \qquad (1)$$

The frequency term

$$f_{ijk}^{cb} = \frac{\text{\# of } (ijk)\text{-triplets of backbone class } c \text{ and buriedness class } b}{\text{total \# of type } cb \text{ triplets}}$$

represents the *observed* frequency of triangles in backbone class c and buriedness class b consisting of amino acids i, j, and k. The expected frequency term

$$p_{ijk}^{cb} = C a_i a_j a_k p_{cb}$$

represents the statistical expectation of encountering various triangle types, where

$$a_i = \frac{\text{\# of amino acids of type } i \text{ in data set}}{\text{total \# of amino acids in data set}},$$

and

$$p_{cb} = \frac{\text{\# of type } cb \text{ triplets in data set}}{\text{total \# of triplets in data set}}.$$

Note that the index c takes values $0, 1, 2$, while the index b takes values from 0-8 (See Table 1 for p_{cb} values.) The combinatorial factor C accounts for certain duplicate versions of triplets that may occur (Krishnamoorthy and Tropsha, 2003).

One interesting exception was used when a certain triangle type was never observed. Because the log-likelihood formula (1) gave pseudo-potentials ranging from -5 to 6, we selected -8 as the score for a non-appearing triangle type. The idea here is to have a penalty in the eventual scoring function for *bad* triangles, but not too much of a penalty. This is because the reason for the non-appearance of the triangle may or may not have statistical significance. That is, it would be incorrect to greatly penalize a triangle type simply because the training set used was too small to observe it. On the other hand, some penalty is needed, because the purpose of the scoring function is to weed out proteins containing rare and strange triangle types. Therefore, -8 seemed to be a rational, if conservative, choice.

To complete the hierarchy of DT-based scoring functions, we also define *two-body* scoring functions. The most basic two-body scoring function does not discriminate (pairwise) contacts based on buriedness. Next in the hierarchy is the pairwise scoring function with contacts distinguished based on their degrees of buriedness (as depicted in Figure (3)). Next, we define three-body scoring functions without buriedness distinctions, which is followed by the three-body scoring function with buriedness classes (Equation 1). In order to obtain the three-body scoring function without buriedness distinctions, we sum up the numerators in the definitions of f_{ijk}^{cb} and p_{cb} over all values of the index b. Next in the hierarchy comes the four-body scoring function, as defined previously (Krishnamoorthy and Tropsha, 2003). Finally, we also consider a weighted combination of three- and four-body scoring functions, where the three-body contacts are distinguished based on degrees of their buriedness. The three-body score is added with a weight of 0.25 to the four-body score so as to balance the total contribution from each class – as there are many more subclasses of three-body contacts (27 counting connectivity and buriedness, against the five connectivity classes for four body contacts).

2.4 Assigning Buriedness Classes

The DT is first computed using the quickhull algorithm (using code adapted from the program of Watson (1992)). The triangles are listed by running through the list of tetrahedra (four per tetrahedron). It is a non-trivial task to fix the buriedness classes of triangles, edges, and vertices. We do the same by running through the complete list of triangles thrice, while following the definitions of buriedness described earlier (Subsection 2.2). The buriedness class of a simplex (face, edge, or point) is subsequently determined as per the definitions illustrated in Figures (3) and (4) . We maintain two lists of faces – one of buried faces and the other of surface faces. We first make a run through the tetrahedra, marking the occurrences of each face (triangle). If a face is spotted for the first time, we set the buriedness class of the face as well as its sub-simplices (edges and points) as *non-buried* (i.e., on the surface), and add the face to the list of surface faces. Instead, if we spot a face for the *second time*, we update the buriedness classes of the face and its sub-simplices to *buried*. At this point, we also move this face from the list of surface faces to the list of buried faces. We then make a second run through the two lists of faces in order to assign the buriedness classes of component simplices (triangles, edges, and points). Consistent with the definition of buriedness of edges and points, we first run through the list of buried faces and mark each subsimplex (edges and points) as buried. We then run through the list of surface faces, and repeat the process of marking subsimplices as surface ones. The buriedness classes of all simplices can be fixed once we have run through both the lists of faces. Hence we fix the buriedness classes of individual simplices when we run through the list of faces *again* for calculating the total scores for the protein. As such, we can assign the buriedness classes for all simplices and

calculate scores for them in *three* passes through the lists of all faces. Since each tetrahedron in the DT contributes at most *four* triangles (typically much less, once we account for buried triangles), we can assign the buriedness classes of all simplices in $O(T)$ time, where T is (an upper bound on) the number of tetrahedra in the DT of the protein. Notice that the space required for storing all the information pertinent to the faces is also $O(T)$.

3 DECOY DISCRIMINATION

We had previously reported (Krishnamoorthy and Tropsha, 2003) on the performance of the DT-based, four-body scoring function on certain sets of decoy structures from the Decoys 'R' Us database (Samudrala and Levitt, 2000), which contains a wide range of proteins and decoy types. The database typically offers single proteins grouped with sets of various numbers of decoys for that protein. The decoys are generated by computational methods based on several ideas from the literature. A root mean square deviation (RMSD) from the native structure for each of the decoys is provided, in order to measure its closeness to the correct structure. Assuming higher scores represent more native-like structures, a good scoring function should show a negative correlation between the total scores and RMSD values. In this study, we compare the performances of several DT-based scoring functions, ranging from simple two-body scoring function to a combination of three- and four-body scoring functions. Two staple measures used to evaluate methods for decoy discrimination are rank of the native (or correct) structure among a set of decoys, and the *Z-score* of the native structure according to the used scoring function (Lazaridis and Karplus, 1999; McConkey et al., 2003; Li and Liang, 2005; Reck and Vaisman, 2007). We define a new measure of performance for decoy discrimination that combines these two standard measures. The new measure, termed *power of native discrimination* and denoted by p_N, is defined as follows:

$$p_N = z_N \left(1 + \log \frac{n_D}{r_N}\right), \tag{2}$$

where z_N is the z-score of the native structure, n_D is the number of decoy structures (including the native structure), and r_N is the rank of the native structure according to the scoring function used. Note that the z-score is close to zero when the scoring function assigns a score to the native structure that is close to the average value (for the decoy set). A negative z-score indicates even poorer ranking. Similarly, the worst native rank among all decoys is the number of decoys itself, i.e., $r_N = n_D$. As such, the contribution of the $\log$ term to p_N is smaller when the native rank is poorer. Similarly, the value of p_N is largest when the native rank is smallest and the native z-score is largest. The idea of defining p_N as we did is to get a value close to zero (or a small negative value) when the discrimination is the worst. On the other end of the spectrum, the value of p_N is the largest when the performance on native discrimination is the best (i.e., high positive native z-score and low native rank).

The entire set of multiple decoys from the Decoys 'R' us database was scored using a hierarchy of DT-based scoring functions – from two-body scoring function without buriedness distinction to a combined 3- and 4-body scoring function with buriedness distinctions. The average powers of native discrimination for each decoy set were tabulated. A few sets of decoys – *semfold, ig_structal, ig_structal_hires,* and *vhp_mcmd* – turned out to be hard to discriminate by any of the DT-based scoring functions considered. All remaining decoy sets, including *4state_reduced, fisa, fisa_casp3, lattice_ssfit, lmds,* and *hg_structal* proved easy for native discrimination by the best DT-based scoring function, which turned out to be the combined 3- and 4-body scoring function with buriedness distinctions. The general trend for most sets of decoys is that using higher order terms (3-body v/s 2-body, and 4-body v/s 3-body) in general helps to increase the power of native discrimination, with the best results typically achieved for the 4-body scoring function *combined* with the 3-body scoring function with buriedness distinctions. Notice that buriedness is a factor not captured by the original 4-body scoring function.

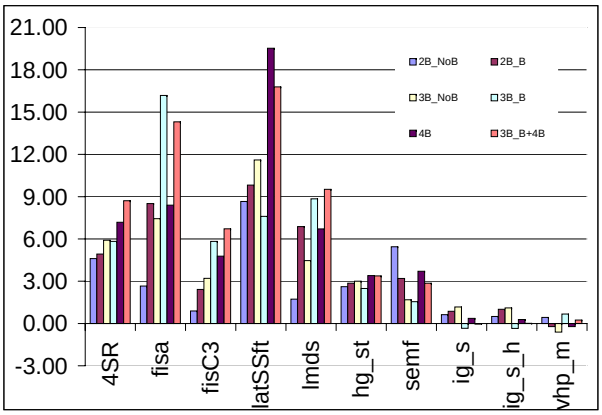

Fig. 5. Performances (as measured by average p_N values) of *six* DT-based scoring functions on *ten* multiple decoy sets from the Decoys 'R' Us database. The names on the x-axis correspond to various decoy sets, such as *4state-reduced, fisa, fisa-CASP3,* etc. (Samudrala and Levitt, 2000). The six different scoring functions considered are the two-body without buriedness, two-body with buriedness distinctions, three-body without buriedness, three-body with buriedness distinctions, four-body, and finally, combined three- and four-body scoring function.

The performance of the combined 3- and 4-body scoring function is markedly better than all other DT-based scoring functions considered for the decoy sets *4state_reduced, fisa, fisa_casp3, lattice_ssfit,* and *lmds.* We use native ranks to demonstrate this result, in Figure (6) . Out of the 35 multiple decoy sets considered, the combined scoring function ranked the native structure among the top 5% of the corresponding decoys for 29 decoy sets. The native structure was ranked as the top conformation in 16 decoy sets, and was ranked in the top 1% of the decoys in 20 decoy sets. The average numbers of decoys per correct structure for each of the five classes of decoys considered here were 665, 500, 1438, 2000, and 440, for 4state_reduced, fisa, etc., respectively. It is also interesting to note that the trend of higher accuracy with higher order contact terms is clearly visible for this subset of decoys.

4 PREDICTING SOLUBILITY MUTAGENESIS

DT-based scoring functions have been used for predicting the effects of single- and multiple-point mutations on the stability (Carter, Jr et al., 2001; Masso et al., 2006; Deutsch and Krishnamoorthy, 2007), and on the reactivity of proteins (Masso and Vaisman, 2007). Other (non DT-based) computational approaches have also been used for these purposes (see papers cited in the above references).

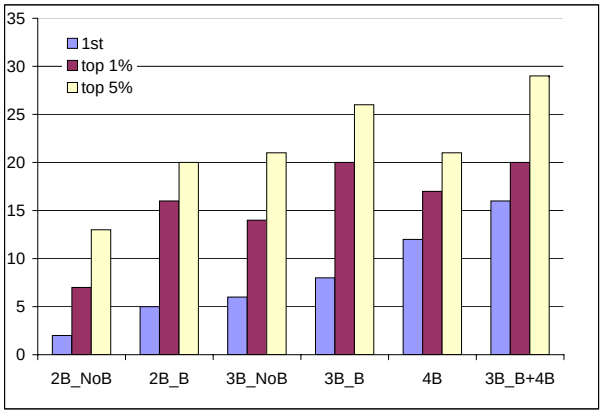

Fig. 6. Number of decoy sets from among the decoy classes *4state_reduced, fisa, fisa_casp3, lattice_ssfit,* and *lmds,* for which the native structure was ranked *the top, among the top 1%,* and *among the top 5%* of all the decoys using the six different DT-based scoring functions. The total number of decoy sets is 35. The combined 3- and 4-body scoring function with buriedness distinctions out-performs other DT-based scoring functions.

Yet, a far fewer number of similar computational approaches have been reported for predicting the effects of mutagenesis on protein *solubility*. It is natural to expect *surface* amino acids of proteins, and propensities of various amino acids to be on the surface of a protein, to play vital roles in determining its solubility. With the definition of buriedness classes of triplets, we have a convenient way to explore the use of DT-based scoring functions for predicting the effects of mutagenesis on solubility of proteins. At the same time, there are no known substantial datasets of mutants with the changes to solubility characterized (similar to the ProTherm database (Kumar et al., 2006) for thermostability mutagenesis, for instance). We present the initial version of such a database here, consisting of 67 mutants assembled from the literature, along with data on changes in the associated solubilities (see Supplementary Materials; also available on the web page for this paper - see Abstract for URL). We are currently in the process of enlarging this database of solubility mutations. The main purpose of introducing the same here is to demonstrate the usefulness of the hierarchy of DT-based scoring functions for purposes other than decoy discrimination. As such, we present only the results from the use of a 3-body scoring function without any training. An expanded version of the solubility mutations database, along with improved methods of predicting the changes to solubility using DT-based scoring functions will be published in the near future.

4.1 Score for solubility mutagenesis

We calculate the change in the total score of the five most non-buried classes of triangles (b classes 0-4; see Figure (4)) that see any change in residue composition due to the mutation. We assume the WT structure (in terms of the sidechain centers of residues) for the mutant protein as well, but the identity of the mutated residues are changed accordingly. Thus we find mutant total score minus the WT total score, where *total score* counts the log-likelihood scores of the five most non-buried (i.e., on the surface) classes of triangles in the DT of the protein. We define the *score of the mutation* as the fraction (or percentage) of this difference to the WT total score. We use a cutoff value of 0.1 in the score of the mutation to count a change (i.e., mutation scores that are less than 0.1 in absolute value are treated as indicating no changes in solubility). Finally, we correlate a positive (negative) score of mutation with an increase (decrease) in solubility of the protein. We use a distance cutoff of 10 Å for screening the DT in this scoring function.

We have assembled a dataset of 59 single-point and 8 multiple (2- or 3-) point mutants along with data on changes to their solubilities. The mutants were assembled from the works of Trevino et al. (2007); Idicula-Thomas and Balaji (2005); Sim and Sim (1999); Maxwell et al. (1999); and references therein. Using the DT-based mutation score, we successfully predict the direction of change of solubility for 47 mutants (out of 67), giving a success rate of 70.2%. In a recent study, Smialowski et al. (2007) have summarized the accuracies of related approaches so far. They reported an overall accuracy of 70%, while Idicula-Thomas et al. (2006) reported a slightly higher accuracy of 72%, which was the best reported accuracy (these authors used a different dataset of 64 mutants). At the same time, it is interesting to note that these two methods involved training the scoring function using a subset of the mutants before using for prediction, while our DT-based scoring function has not been trained on a set of mutants. We hope to include mutants from these and other studies to enlarge our dataset, and then use a subset of the same to train a DT-based scoring function, which will hopefully achieve a much higher accuracy of prediction.

5 CONCLUSIONS

It is interesting to note that some of the DT-based scoring functions defined by Reck and Vaisman (2007) using hydrated proteins could equivalently be defined using the more general framework for buriedness and surface features introduced here. Similarly, one could define and test a DT-based scoring function equivalent to the four-body scoring function defined by Feng et al. (2007), by using the DT to define contacts rather than absolute distance cutoffs, but carrying over the rest of the definition. Further, We have not compared side-by-side the performances of DT-based hierarchy of scoring functions to other fold recognition protocols available. At the same time, we believe that a systematic approach needs to be undertaken for developing "optimal" DT-based scoring functions. We are currently investigating several methods, including optimization-based training methods, to fine tune specific combinations of our DT-based scoring functions for various purposes such as fold recognition and mutagenesis. We are also considering the extension of the hierarchy to alpha shapes of proteins, which is a natural generalization of the DT (as a first step, e.g., extending the scoring function of Li et al. (2003) to include buriedness levels). The main purpose of this paper is to lay down the fundamentals of the DT-based hierarchy of contacts, especially the various buriedness classes and efficient methods to assign the same.

ACKNOWLEDGMENT

Krishnamoorthy, Stratton, and Deutsch are thankful for the support provided by the NSF UBM Grant DEB 0531870 for working on the research presented in this paper.

REFERENCES

Banavar, J. R., Maritan, A., Micheletti, C., and Trovato, A. (2002) Geometry and physics of proteins. *Proteins: Structure, Function, and Genetics*, **47**(3), 315–322.

Ben-Naim, A. (1997) Statistical potentials extracted from protein structures: Are these meaningful potentials? *The Journal of Chemical Physics*, **107**(9), 3698–3706.

Bostick, D. L., Shen, M., and Vaisman, I. I. (2004) A simple topological representation of protein structure: Implications for new, fast, and robust structural classification. *Proteins: Structure, Function, and Bioinformatics*, **56**(3), 486–501.

Bryant, S. H. and Lawrence, C. E. (1991) The frequency of ion-pair substructures in proteins is quantitatively related to electrostatic potential: A statistical model for nonbonded interactions. *Proteins: Structure, Function, and Genetics*, **9**(2), 108–119.

Buchete, N.-V., Straub, J. E., and Thirumalai, D. (2004) Orientational potentials extracted from protein structures improve native fold recognition. *Protein Science*, **13**(4), 862–874.

Carter, Jr, C. W., LeFebvre, B., Cammer, S. A., Tropsha, A., and Edgell, M. H. (2001) Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. *Journal of Molecular Biology*, **311**, 625–638.

Cheng, J., Randall, A., and Baldi, P. (2006) Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins: Structure, Function, and Bioinformatics*, **62**(4), 1125–1132.

Delarue, M. and Koehl, P. (1995) Atomic environment energies in proteins defined from statistics of accessible and contact surface areas. *Journal of Molecular Biology*, **249**(3), 675–690.

Deutsch, C. and Krishnamoorthy, B. (2007) Four-body scoring function for mutagenesis. *Bioinformatics*, **23**(22), 3009–3015.

DeWitte, R. S. and Shakhnovich, E. I. (1994) Pseudodihedrals: Simplified protein backbone representation with knowledge-based energy. *Protein Science*, **3**(9), 1570–1581.

Edelsbrunner, H. (2001) *Geometry and Topology for Mesh Generation*. Cambridge University Press, England.

Edelsbrunner, H. and Koehl, P. (2005) The geometry of biomolecular solvation. In *Combinatorial and Computational Geometry*, volume 52 of *MSRI Publications*, 243–275.

Feng, Y., Kloczkowski, A., and Jernigan, R. L. (2007) Four-body contact potentials derived from two protein datasets to discriminate native structures from decoys. *Proteins: Structure, Function, and Bioinformatics*, **68**(1), 57–66.

Gan, H. H., Tropsha, A., and Schlick, T. (2001) Lattice protein folding with two and four-body statistical potentials. *PROTEINS:Structure, Function, and Genetics*, **43**, 161–174.

Gilis, D. and Rooman, M. (1997) Predicting protein stability changes upon mutation using database-derived potentials: Solvent accessibility determines the importance of local versus non-local interactions along the sequence. *Journal of Molecular Biology*, **272**, 276–290.

Halgren, T. A. (1995) Potential energy functions. *Current Opinion in Structural Biology*, **5**, 205–210(6).

Huan, J., Bandyopadhyay, D., Wang, W., Snoeyink, J., Prins, J., and Tropsha, A. (2005) Comparing graph representations of protein structure for mining family-specific residue-based packing motifs. *Journal of Computational Biology*, **12**(6), 657–671.

Idicula-Thomas, S. and Balaji, P. V. (2005) Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in Escherichia coli. *Protein Sci*, **14**(3), 582–592.

Idicula-Thomas, S., Kulkarni, A. J., Kulkarni, B. D., Jayaraman, V. K., and Balaji, P. V. (2006) A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on overexpression in Escherichia coli. *Bioinformatics*, **22**(3), 278–284.

Krishnamoorthy, B. and Tropsha, A. (2003) Development of a four-body statistical pseudo-potential for discriminating native from non-native protein conformations. *Bioinformatics*, **19**(12), 1540–1549.

Kumar, M. S., Bava, K. A., Gromiha, M. M., Prabakaran, P., Kitajima, K., Uedaira, H., and Sarai, A. (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Research*, **34**, D204–D206.

Lazaridis, T. and Karplus, M. (1999) Discrimination of native from misfolded protein models with an energy function including implicit solvation. *J. Mol. Biol.*, **288**, 477–487.

Li, X., Hu, C., and Liang, J. (2003) Simplicial edge representation of protein structures and alpha contact potential with confidence measure. *Proteins: Structure, Function, and Bioinformatics*, **53**(4), 792–805.

Li, X. and Liang, J. (2005) Geometric cooperativity and anticooperativity of three-body interactions in native proteins. *Proteins: Structure, Function, and Bioinformatics*, **60**(1), 46–65.

Masso, M., Lu, Z., and Vaisman, I. I. (2006) Computational mutagenesis studies of protein structure-function correlations. *Proteins: Structure, Function, and Bioinformatics*, **64**(1), 234–245.

Masso, M. and Vaisman, I. I. (2007) Accurate prediction of enzyme mutant activity based on a multibody statistical potential. *Bioinformatics*, **23**(23), 3155–3161.

Maxwell, K. L., Mittermaier, A. K., Forman-Kay, J. D., and Davidson, A. R. (1999) A simple in vivo assay for increased protein solubility. *Protein Science*, **8**(9), 1908–1911.

McConkey, B. J., Sobolev, V., and Edelman, M. (2003) Discrimination of native protein structures using atom-atom contact scoring. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(6), 3215–3220.

Melo, F. and Feytmans, E. (1997) Novel knowledge-based mean force potential at atomic level. *Journal of Molecular Biology*, **267**(1), 207–222.

Miyazawa, S. and Jernigan, R. L. (1985) Estimation of effective inter-residue contact energies from protein crystal structures: A quasi-chemical approximation. *Macromolecules*, **18**, 534–552.

Miyazawa, S. and Jernigan, R. L. (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal of Molecular Biology*, **256**(3), 623–644.

Park, B., Huang, E., and Levitt, M. (1997) Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J. Mol. Biol.*, **266**, 831–846.

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

Reck, G. M. and Vaisman, I. I. (2007) Decoy discrimination using contact potentials based on Delaunay tessellation of hydrated proteins. In *ISVD2007: Fourth IEEE International Symposium on Voronoi Diagrams in Science and Engineering*.

Samudrala, R. and Levitt, M. (2000) Decoys 'R' Us: A database of incorrect conformations to improve protein structure prediction. *Protein Science*, **9**, 1399–1401.

Samudrala, R. and Moult, J. (1998) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *Journal of Molecular Biology*, **275**(5), 895–916.

Seno, F., Maritan, A., and Banavar, J. R. (1998) Interaction potentials for protein folding. *Proteins: Structure, Function, and Genetics*, **30**(3), 244–248.

Sim, J. and Sim, T. (1999) Amino acid substitutions affecting protein solubility: high level expression of streptomyces clavuligerus isopenicillin N synthase in escherichia coli. *Journal of Molecular Catalysis B: Enzymatic*, **6**(3), 133–143.

Singh, R. K., Tropsha, A., and Vaisman, I. I. (1996) Delaunay tessellation of proteins: Four body nearest neighbor propensities of amino acid residues. *Journal of Computational Biology*, **3**(2), 213–222.

Sippl, M. J. (1990) Calculation of conformational ensembles from potentials of mean force. *Journal of Molecular Biology*, **213**, 859–883.

Sippl, M. J. (1995) Knowledge-based potentials for proteins. *Current Opinion in Structural Biology*, **5**, 229–235.

Smialowski, P., Martin-Galiano, A. J., Mikolajika, A., Girschick, T., Holak, T. A., and Frishman, D. (2007) Protein solubility: sequence based prediction and experimental verification. *Bioinformatics*, **23**(19), 2536–2542.

Summa, C. M., Levitt, M., and DeGrado, W. F. (2005) An atomic environment potential for use in protein structure prediction. *Journal of Molecular Biology*, **352**(4), 986–1001.

Taylor, T. J., Rivera, M., Wilson, G., and Vaisman, I. I. (2005) New method for protein secondary structure assignment based on a simple topological descriptor. *Proteins: Structure, Function, and Bioinformatics*, **60**(3), 513–524.

Taylor, T. J. and Vaisman, I. I. (2006) Graph theoretic properties of networks formed by the delaunay tessellation of protein structures. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, **73**(4), 041925.

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

Trevino, S. R., Scholtz, J., and Pace, C. (2007) Amino acid contribution to protein solubility: Asp, Glu, and Ser contribute more favorably than the other hydrophilic amino acids in RNase Sa. *Journal of Molecular Biology*, **366**(2), 449–460.

Wang, G. and Dunbrack, Jr., R. (2003) Pisces: a protein sequence culling server.

Watson, D. (1992) *CONTOURING: A guide to the analysis and display of spatial data*. Pergamon Press, http://members.iinet.net.au/∼watson/software.htm.

Wodak, S. and Rooman, M. (1993) Generating and testing protein folds. *Current Opinion in Structural Biology*, **3**, 247–259.

Zheng, W., Cho, S., Vaisman, I. I., and Tropsha, A. (1997) A new approach to protein fold recognition based on Delaunay tessellation of protein structure. In *Pacific Symposium on Biocomputing 1997*, 487–496.

Zomorodian, A., Guibas, L., and Koehl, P. (2006) Geometric filtering of pairwise atomic interactions applied to the design of efficient statistical potentials. *Computer Aided Geometric Design*, **23**(6), 531–544.