BIOINFORMATICS

Vol. 00 no. 00 2007
Pages 1–7

Four-Body Scoring Function for Mutagenesis

Chris Deutsch [a] and Bala Krishnamoorthy [a]

[a]Department of Mathematics, Washington State University

ABSTRACT

Motivation: There is a need for an efficient and accurate computational method to identify the effects of single and multiple residue mutations on the stability and reactivity of proteins. Such a method should ideally be consistent and yet applicable in a widespread manner, i.e., it should be applied to various proteins under the same parameter settings, and have good predictive power for all of them.

Results: We develop a Delaunay tessellation-based four-body scoring function to predict the effects of single and multiple residue mutations on the stability and reactivity of proteins. We test our scoring function on sets of single point mutations used by several previous studies. We also assemble a new, diverse set of 237 single and multiple residue mutations, from over twenty four different publications. The four-body scoring function correctly predicted the changes to the stability of 169 out of 210 mutants (80.5%), and the changes to the reactivity of 17 out of 27 mutants (63%). For the mutants that had the changes in stability/reactivity quantified (using reaction rates, temperatures etc.), an average Spearman rank correlation coefficient of 0.67 was achieved with the four-body scores. We also develop an efficient method for screening huge numbers of mutants of a protein, called *combinatorial mutagenesis*. In one study, 64 million mutants of a cold-shock nucleus binding domain protein 1CSQ, with six of its residues being changed to all possible (20) amino acids, were screened within a few hours on a PC, and all five stabilizing mutants reported were correctly identified as stabilizing by combinatorial mutagenesis.

Availability: All lists of mutants scored, and executables of programs developed as part of this study are available from this web page: http://www.wsu.edu/∼kbala/Mutate.html.

Contact: kbala@wsu.edu

1 INTRODUCTION AND PREVIOUS WORK

Mutagenesis is the process of replacing one or more amino acids in a wild-type (WT) protein by alternate amino acids to generate a mutant protein. The goal of the process is to create a protein with certain desirable biochemical properties that are lacking in the WT protein. For instance, a protein that is more reactive, or more stable, in a particular reaction than the WT can often be generated by altering the identity of a single key residue (to generate a single-point mutant). Mutagenesis finds applications naturally in protein design, drug discovery, and other similar areas (see the Supplementary Document for a listing of many such applications).

The experimental process of creating mutants can often be expensive and time-consuming. To start with, it is often not straightforward to identify the key residue(s) that need to be mutated in order to achieve the desired biochemical properties. Once the critical residue positions are identified, it can still be non-trivial to decide what the new amino acids should be. Hence biochemists often end up having to create and analyze a large number of mutants in order to identify a handful of desirable ones. In a typical example (1CSQ, one of the proteins included in our research), six amino acid positions were identified as desirable mutation sites. The experimentalists wanted to try all possible alternate amino acid combinations for these six residues, changing at least one amino acid in each case. The total number of single- and multiple-point mutants that they could have considered is 64 million ($20^6 - 1$ to be exact)! Of course, they only tried a few hundreds of the mutants, and reported five of them that had the desired properties. As illustrated by this example, it is quite desirable to use computational methods to reduce the number of mutants that need to be generated experimentally.

One of the key steps in most protein structure prediction methods is the screening of multiple candidate conformations to select the best one(s), and a scoring function is used for this purpose. Scoring functions have been used for protein fold recognition for several years – many of them are studied in the following references (Miyazawa and Jernigan, 1985; Sippl, 1990; Park and Levitt, 1996; Krishnamoorthy and Tropsha, 2003). We could study the use of any such scoring function for the purpose of virtual mutagenesis – score the original (WT) protein, and then score the original protein after making the proposed changes to the sequence while keeping the structure unchanged. We could then correlate the changes in score and the effects of the mutations on various properties of the original protein, thus developing a *predictor* for the effects of mutations. In spite of the apparent simplicity, only a few such studies have been undertaken so far. Carter et al. (2001) obtained high correlations between changes in the four-body scores and the free energy changes (measured as $\Delta\Delta G$ values) resulting from mutations to residues in the hydrophobic cores of five different proteins. More recently, Masso, Lu, and Vaisman (2006) used the same four-body scoring function to study the structure-function correlations of the mutants of HIV-1 protease and T4 lysozyme.

On the other hand, direct computational strategies (i.e., without any connections to fold recognition) have been used to predict the effects of mutations on the stability of proteins. Gilis and Rooman (1997) developed database-derived potentials based on solvent accessibility to predict the effects of single-point mutations on the stability of proteins. Topham et al. (1997) used tabulated structural propensities of amino acids to predict the changes to the stability of several T4-lysozyme structures. In addition to the 3D structures of the WT proteins, this study also used the 3D structures of mutants. Guerois et al. (2002) developed the energy function called FOLD-X for predicting the $\Delta\Delta G$ values due to mutations. More recently, Cheng et al. (2006) developed support vector machines-based (SVM) models that used both sequence and structure information to predict stability changes due to single-point mutations. The SVM-based method has the best accuracy

reported so far – 84%. The common feature of the above methods is that they all employ various sequence and structure interaction terms, and the best way to combine these terms is determined (i.e., various parameters are tuned) using a training set of mutations. The accuracy of these methods mainly stems from the training procedure, and hence is quite dependent on the training set of mutations used.

We believe that the accuracy of the underlying scoring function (or interaction terms) is most critical for predicting the effects of mutations. Our main goal is to develop, and test, an accurate underlying scoring function to predict the changes to the stability and reactivity of proteins due to mutations *from scratch*, i.e., without having to learn from any mutations. The scoring function we use is developed from the four-body scoring function that is based on the Delaunay tessellation of proteins. The latest (and most accurate) version of the four-body scoring function as used for protein decoy discrimination was proposed by the author previously (Krishnamoorthy and Tropsha, 2003), and has been tested extensively on many test sets of decoys. Though two previous studies (Carter et al., 2001; Masso et al., 2006) used the four-body scoring function to analyze mutagenesis, they both tested the scoring function only on limited sets of proteins. The first study considered mutations that are made only in the hydrophobic core of five proteins. The second study considers most possible mutations for two different proteins. Further, they both use an older version of the scoring function, and the second study uses different settings for scoring the mutants of the two proteins considered. We address these and other shortcomings when defining our four-body scoring function.

We first test our scoring function on 1558 mutants (all single-point mutations except three) considered by the previous studies for stability changes. The overall accuracy was 65.2% (see Section 3 for details). We also assemble a new, comprehensive list of proteins and their mutants (both single and multiple point), along with experimental data that quantifies the change in stability or reactivity of the WT protein. This test set of 237 mutants is collected from twenty four different experimental studies. We correctly predict the effects of the mutations on the stability of 169 out of 210 mutants (80.5%), and those on the reactivity of 17 out of 27 (63%) mutants. For the sets of mutants that had the effects on stability or reactivity quantified (reaction rates, free energy changes etc.), we obtain an average Spearman rank correlation coefficient of 0.67 between the quantifying data and the four-body scores. We also propose an efficient method to evaluate huge numbers of mutants by working with only the Delaunay tetrahedra that the mutated residues participate in (as opposed to scoring the WT protein repeatedly).

2 METHODS AND MATERIALS

We describe the details of the four-body scoring function used for mutagenesis, and the test set of single- and multiple-residue mutants that we assembled, and then introduce combinatorial mutagenesis.

2.1 Four-body scoring function

The idea of a scoring function for protein fold recognition that is built on the Delaunay tessellation of proteins was first proposed by Tropsha et al. (1996). Singh, Tropsha, and co-workers subsequently developed the scoring function (Munson and Singh, 1997; Tropsha et al., 1998), and also explored the possibility of using the same for *ab initio* protein folding (Gan et al., 2001). The formulation of the scoring function was improved by Krishnamoorthy and Tropsha (2003), and the applicability for decoy discrimination was tested on various decoy sets. Further extensive testing of the scoring function has been conducted recently by Krishnamoorthy et al. (Fowler et al., 2007; Krishnamoorthy and Stratton, 2007). This latest formulation defines the scoring function as the following log-likelihood ratio:

$$Q^{\alpha}_{ijkl} = \log\left[\frac{f^{\alpha}_{ijkl}}{p^{\alpha}_{ijkl}}\right]. \quad (1)$$

i, j, k, and l represent the residue identities of the four amino acids (20 possibilities) in a Delaunay tetrahedron from the tessellation of the protein. Each amino acid is represented by a single point located at the centroid of the atoms in its side-chain (including the C_α atom). α represents the *type* of the tetrahedron based on the back-bone chain connectivity of the four participating amino acids. There are five tetrahedron types possible, and α takes one of the values $0, 1, 2, 3$ or 4 corresponding to these types (Krishnamoorthy and Tropsha, 2003). The total score (or simply, the *score*) of a protein is then defined as the sum of the log-likelihood ratios of all tetrahedra in its Delaunay tessellation. A cut-off value of 10 Å (Angstroms) was used for the length of any edge of the tetrahedra that are scored, thus discarding biochemically irrelevant tetrahedra with huge edge lengths. Simply put, the score of a protein gives a measure of how *well-packed* its residues are (hence it was also called the Simplicial Neighborhood Analysis of Protein Packing, or SNAPP, score). Further, the correct way to interpret the score is in a *relative* sense, i.e., we can compare the scores of two otherwise similar conformations to quantify how one of them is packed better than the other.

The back-bone chain connectivity of the tetrahedra is not considered by Carter et al. (2001) or in the more recent study by Masso et al. (2006). Further, in both these studies, there is some ambiguity regarding the choice of side-chain centers of residues versus back-bone C_α atom coordinates that should be used to represent each amino acid. It is not desirable to change the settings and other parameters of the scoring function when scoring different proteins. In fact, Masso et al. use two different settings for the two proteins that they study. Their justification is the robustness of the four-body score under small perturbations of the points representing each amino acid. The authors claim that the total score of a protein does not change *by much* when the representation of the amino acids is changed from C_α to side-chain centers.

The question of robustness of the Delaunay tessellation of proteins (and point sets in general) was addressed by Bandyopadhyay and Snoeyink (2004) – they defined the concept of *almost Delaunay* simplices, where the positions of the points defining the simplex are allowed to vary in a controlled range (as opposed to being fixed). The four-body scoring function was tested under the almost-Delaunay setting to obtain decoy discrimination results that are roughly comparable to those obtained by the original scoring function. Still, the results obtained using the side-chain center representation (by Krishnamoorthy and Tropsha (2003)) are markedly better than those obtained using C_α representation, or using almost Delaunay tetrahedra. In fact, Krishnamoorthy and Tropsha did obtain the C_α results for the decoy sets reported in their paper; but they were uniformly inferior to those using side-chain centers, and hence not reported at all. They suggested that side-chain

centers be used always in order to obtain the most accurate results. This suggestion has been further validated by recent results obtained by Krishnamoorthy et al. (Fowler et al., 2007; Krishnamoorthy and Stratton, 2007).

Another aspect of the side-chain center versus C_α option is the change in representation between the WT and the *actual* mutant protein. We use the structure of the WT for the mutant as well (only the sequence is changed). To check the validity of this assumption, we need to compare the 3D structure of mutants (when available) with that of the WT, as represented by the set of Delaunay tetrahedra formed. Naturally, the tetrahedra set could see several more changes with the side-chain center representation as compared to that of C_α's; especially, when small residues in the WT are replaced by bulky ones in the mutant (e.g., a GLY replaced by TYR). Topham et al. (1997) provide a list of PDB codes for several WT-mutant pairs of T4 lysozyme. We calculate the *edit distance* between the tetrahedra sets participated by the mutation sites in the WT and in the mutant – how many residue number substitutions have to be made to get from the tetrahedra set of the WT to that of the mutant, given as a percentage of the total number of residues (counting repetitions) in the tetrahedra in the set of the WT. Under the side-chain center representation, the average edit distance is 35%, while under C_α representation, it is only 12%. At the same time, the above calculation completely ignores the sequence of the residues. Even though the C_α atoms of the WT are a lot closer to the C_α of the actual mutant, the sequence-structure correlations are far more accurate under the side-chain center representation (Krishnamoorthy and Tropsha, 2003). To make sure, we scored our set of mutants using C_α's, only to obtain an accuracy of less than 50%. Hence, we stick with the side-chain centers.

On a related note, the claim of Masso et al. that the score of the protein does not change much when C_α atoms are used in place of side-chain centers might hold only for the *total score* of the protein, and not for the case of *change* in the total score – especially when the change is small. When only a single residue is changed, only a small subset of the full set of Delaunay tetrahedra is affected. The change in the total score in this case might be sensitive to the way the residues are represented, and also to perturbations in the positions of these residues. The residues that are in the inner portions of the protein (buried) participate in many more tetrahedra than those that are on the outside (surface), and hence the robustness result might apply more for the case of the buried residues. All mutations studied by Carter et al. (2001) are performed on hydrophobic core residues. On the other hand, many mutations that we considered involve changes to surface residues. Hence we suggest the consistent use of side-chain centers when scoring mutations using the four-body scoring function. We also use the weights for the scores of different classes of tetrahedra as defined by Krishnamoorthy and Tropsha (2003).

In addition, some key long-range interactions between amino acids are missed out by the use of a 10 Å cut-off on the Delaunay edges, especially for the case of surface residues. Hence we use an increased cut-off of 12 Å when scoring mutations. Notice that the contacts made by most buried residues remain unchanged, as such contacts are well within the 10 Å range. At the same time, several key interactions of surface residues that are left out by the 10 Å cut-off are now included in the calculations, thus making the scoring function more accurate.

Under the settings described above, we calculate the change in total score between the WT and the mutant protein (mutant score - WT score). A positive change (i.e., the mutant score is more than the WT score) indicates that the mutant is more stable than the WT, while a negative change indicates lower stability. Instead of using the raw change in total score, we use the fraction (given as percentage) of change to the sum of the scores of the tetrahedra that see *any* change due to the mutations. Thus we exclude from the calculations those tetrahedra that are present both in the WT and the mutant. We use a cut-off value of 0.1% to determine if this percentage change is significant (i.e., if the percentage change is below 0.1% in absolute value, we assume there is *no* change).

We also correlate increased (decreased) activity with a negative (positive) change in the total score. The intuition behind this definition is that well-packed proteins are typically not highly active, and hence the high total score is correlated with less activity. We must mention, though, that as of now, we could only assemble a limited number of mutants with activity data to test this assumption (see Section 3).

2.2 Test sets of mutants

The ProTherm database (Kumar et al., 2006) lists a huge number of mutations, and some of the previous studies have created mutant data sets from there (Cheng et al., 2006). At the same time, ProTherm typically does not list multi-point mutants, and reactivity data is not listed in all the cases as well. Hence, we have searched the literature to identify a comprehensive list of single- and multiple-point mutations. Overall, there are 237 mutants taken from 24 different papers. 210 of the mutants are analyzed for changes in stability, while the remaining 27 are analyzed for changes in reactivity. After assembling the data set, we found that ProTherm in fact listed 80 of them. The whole data set, along with the performance of the four-body scoring function on the mutants, is presented in Table 1. We describe the various types of mutations assembled briefly in a supplementary document.

Apart from our mutant list, we also analyze the two lists of 1096 and 388 single point mutants considered by Cheng et al. (2006), the 50 single point mutants that were considered outliers in the study by Guerois et al. (2002), and 24 T4 lysozyme mutants (3 being multiple-point) considered by Topham et al. (1997).

2.3 Combinatorial Mutagenesis

After the potential mutation sites in the WT are identified, it is often not straightforward to decide the new residues to be put in these sites. Experimentalists might want to try several amino acids for each mutation site, and hence are faced with the task of generating a large number of mutants. For example, if three potential mutation sites have been identified, and we want to try the residues Ala, Val, Ile for the first site, Val and Leu for the second site, and Cys, His, Lys, and Arg for the third site. The total number of mutants we have to analyze is $3 \times 2 \times 4 = 24$. In one of the studies that we used to create our list of mutants, Martin et al. (2001) identified six amino acid sites of the protein 1CSQ to be mutated to *all other* amino acids. The result is a staggering 64,000,000 proteins to be analyzed (including the WT). Our scoring function is most valuable for such mutagenesis experiments – we could identify (computationally) a relatively small set of mutants that are potentially the most suitable ones, and experimentally generate them before considering others. Since we consider all possible combinations of mutations at the

Table 1. Test set of mutations studied. Each mutation (if given) is indicated by the residue number and the new amino acid to which it is mutated. If there are multiple changes defining a single mutant, these are separated by "/", and all the mutations enclosed inside braces. Pred gives the number of correctly predicted mutants, out of the total number given by TOT. The complete list of mutants and the corresponding four-body scores are provided in a supplementary document. They are also available from the author's web page: http://www.wsu.edu/∼kbala/Mutate.html.

#	Article	Study	Mutants scored	Pred	TOT
1	Bonander et al. (2000)	Disulfide bond-deficient azurin mutant	(3A/26I), (3A/26A), (3A/25R/26A/27R)	3	3
2	Martin et al. (2001)	In-vitro selection of highly stabilized mutants with optimized Surface	G-1 (66L/67P), G-2 (2I/3S/46Q/64L/66L/67P), G-3 (46L/66L/67H), G-4 (2Y/3R/46L), G-5 (2Y/3I/46Q/64L/66L/67P)	5	5
3	de Antonio et al. (2000)	Contribution of tryptophan to the properties of ribotoxin α-sarcin	4F, 51F, (4F/51F)	0	3
4	Chen and Gouaux (1997)	Reduction of hydrophobicity in bacteriorhodopsin	Q1 (113Q), Q2 (113Q/116Q), Q3 (113Q/116Q/120Q), Q4 (113Q/116Q/120Q/124Q), Q4D (113Q/116Q/117D/120Q/124Q)	5	5
5	Huang et al. (1996)	Mutagenesis of protein phosphatase 1 for catalysis and inhibitor binding	96A, 124D, 248N, 221S, 395A, 208A	2	6
6	Ge et al. (2003)	Antifungal activity of a rice lipid transfer protein	45A, 46A, 72L	2	3
7	Suresh et al. (2006)	C-reactive protein mutants and pneumococcal infection in mice	175A, 114A	2	2
8	Oppermann et al. (1997)	Mutagenesis of hydroxysteroid dehydrogenase	12A, 87A, 138A	2	3
9	Pathange et al. (2006)	Correlation between protein binding strength	29 single-point mutants from 82-93H, 136-155H	26	29
10	Funahashi et al. (2002)	Surface hydration and stability of lysozyme	2G, 2A, 2L, 2M, 2F, 2S, 2Y, 2D, 2N, 2R, 2I 74G, 74A, 74F, 74S, 74Y, 74D, 74N, 74R, 74M, 74L, 74I, 110F, 110Y, 110R, 110N, 110D, 110M, 110L, 110I, 110A, 110G (original residue in all cases is V)	24	32
11	Takano et al. (1999)	N-terminal residues and conformational stability of human lysozyme	1M, 1A	0	2
12	Hahn et al. (1995)	Mutagenesis of Glucanase	101Y, 103Q, 105N, 105K, 107D, 101F, 103D	5	7
13	Sun et al. (2001)	Mutations and stability of methylamine dehydrogenase	76N, 122A, 122C, 119F, 119E, 119K	4	6
14	Dvir et al. (2003)	Human acid-β-glucosidase and Gaucher disease	370S, 394L, 463C, 496H	2	4
15	Korkegian et al. (2005)	Thermostabilization of an enzyme	(23L/140L/108I), (23L/140L), 23L, 140L, 10T, 67E, 69L	5	7
16	Brownlie et al. (1994)	Structures of the mutants of porphobilinogen deaminase	26H, 149L, 173W, 31T, 34K, 116T, 116W, 116Q, 177R, 223K, 250K, 252T, 93F, 201W, 247F, 256N, 167W	12	17
17	Braun et al. (1997)	Alanine insertion in lactose permease transmembrane helices	83A, 87A, 90A, 91A, 93A, 79A, 96A	5	7
18	Siadat et al. (2006)	Disulfide bonds and stability of an acetylcholinesterase	M2 (327C/375C), M3 (354C/456C), M4 (369C/476C), M6 (452C/533C), M7 (464C/543C)	4	5
19	Almog et al. (2002)	Stabilizing mutations in subtilisin BPN	S63 (41A/50F/73L/206W/217K/218S/221C/271E), S88 (2K/3C/5S/43N/50F/73L/206C/217K/218S/271E)	2	2
20	Erwin et al. (1990)	Salt bridges and stability of subtilisin BPN	271E, 51K, 164R	3	3
21	Köditz et al. (2004)	Mutagenesis of the unfolding region of ribonuclease A	35S, 35A, 46Y, (31A/33S), (35S/46Y), (35A/46Y), (31A/33S/46Y)	5	7
22	Ormö et al. (1995)	Radical stability in ribonucleotide reductase from E-coli	212W, 234N	2	2
23	Kong et al. (1993)	Evolutionally conserved aspartic acid residues in human glutathione S-transferase P1-1	57A, 98A, 152A	0	3
24	Carter et al. (2001)	Hydrophobic core mutations and stability	73 mutants (five different proteins)	66	73

individual sites, we term the process of scoring all possible mutants using the four-body scoring function as *combinatorial mutagenesis*.

As seen by the example of 1CSQ, the number of combinations could be quite huge, and for such cases, the usual way of scoring the mutants turns out to be highly inefficient. By default, we would score the WT protein once, and then score the same again for each mutant, with the appropriate changes made in the amino acid sequence. Each call to the four-body scoring function involves the computation of the Delaunay tessellation of the protein, which proves to be the bottleneck as far as the overall running time of the algorithm is concerned. The most efficient algorithms for computing Delaunay tessellations have a worst-case running time

of $O(n \log n)$, where n is the number of points (see Edelsbrunner, 2001, Chapters 1,5). The average running times in practice also follow the same bounds. At the same time, we notice that the structure of the WT is not altered in any of the mutants, and hence the residue numbers of the four amino acids forming each tetrahedron remains unaltered, even though the identities of some of the amino acids are changed. Hence we calculate the Delaunay tessellation only *once* as part of combinatorial mutagenesis, when scoring the WT protein. We just need to change the amino acid identities corresponding to each mutant.

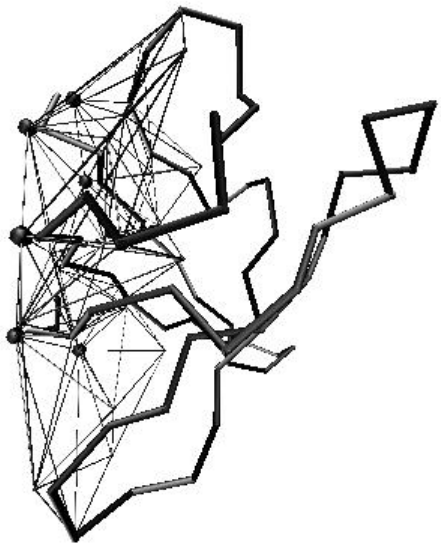

Fig. 1. 1CSQ (backbone trace shown as the thick line) along with the Delaunay tetrahedra that the six mutation sites participate in. The six residues considered for mutation are 2-Leu, 3-Glu, 46-Ala, 64-Thr, 66-Glu, and 67-Ala. The spheres are at the side-chain centers of these residues, and represent the six residues in the increasing order when viewed from the bottom part of the figure to the top. The thin lines are the edges of the 58 tetrahedra that contain at least one of these six residues. If we consider all 67 amino acids, we get a total of 256 Delaunay tetrahedra (at 12 Å distance cut-off). This image was created using VMD (Humphrey et al., 1996).

As a result, only those tetrahedra that involve one or more of the mutation sites see changes to the sequence identities of the participating amino acids. For instance, the six mutation sites of 1CSQ participate in (i.e., at least one of the six is in) 58 Delaunay tetrahedra, which is only a fraction of the total of 256 tetrahedra formed by the entire protein (see Figure 1). From the theoretical point of view, there are some bounds for the number of Delaunay triangles that each point (out of the total n points) participates in when we consider the two-dimensional case (Edelsbrunner, 2001), and some more conservative estimates could be derived in 3D. From among the 4,000-odd proteins that we analyzed, the maximum number of tetrahedra that a single amino acid participated in is 48 (43-Arg in 2BOQ), but the typical number of tetrahedra is much smaller (average is 17.65). This number is even smaller if the amino acid in question is on the surface of the protein. So we identify the smaller set of tetrahedra that the mutation sites participate in (by searching the Delaunay tessellation of the WT). For each mutant, we calculate the difference in the sum of the log-likelihood ratios for these tetrahedra alone in the WT and in the mutant, and we use this difference to score and rank all the mutants. This implementation of combinatorial mutagenesis proves to be far more efficient than repeated calls to the default four-body scoring function for each mutant (see Section 3.1).

3 RESULTS AND DISCUSSION

We say that the four-body scoring function predicts the effect of a mutation *correctly* if an increase (correspondingly, a decrease) in the four-body total score is observed for mutations that are experimentally observed to be stabilizing or decreasing the reactivity (destabilizing or increasing the reactivity) of the WT protein. Overall on our data set, 78% (186) of the mutants were identified correctly (see Table 1), with 169 out of 210 correct predictions for stability (80.5%) and 17 out of 27 for activity (63%). We are currently undertaking a detailed examination of how the scoring function performed for each of the twenty four mutant sets, and especially the cases of de Antonio et al. (2000), Takano et al. (1999), and Kong et al. (1993) (articles #3, #11, and #23 in Table 1), for which the scoring function failed on all the mutants considered from each set. For the human lysozyme mutants studied by Takano et al., only the N-terminal residue is mutated, which does not form enough Delaunay tetrahedra (being on one end of the chain, and on the surface of the protein).

As of now, we only have limited (27) number of mutants with activity data available. We need more such mutants to test out assumption that high total scores correspond to lower activity. The accuracy of 63% on this set of mutants with activity data is encouraging still.

The effects of the mutants are quantified for ten of the mutant sets, five of them being different proteins studied by Carter et al. (2001). Even though the authors calculated linear correlation coefficients between four-body scores and free energy changes of these mutants, there is no clear evidence to suggest that the four-body scores follow a linear relationship with the experimental quantities reported. (Masso et al. (2006) also report linear correlation coefficients, but see Section 3.2 for discussion on their work). The Spearman rank correlation coefficient between the four-body scores and the experimental values seems more appropriate, with no assumptions of linear relationships involved. We present the rank correlation coefficients for the ten mutant sets in Table 2. The overall average Spearman rank correlation coefficient is 0.67. The rank correlations for the set of mutants studied by Carter et al. are markedly high – the average for these five mutant sets is 0.77. This result is not surprising, as all these mutations are done on sites in the hydrophobic cores of the proteins in question. In general, the more tetrahedra the mutation sites participate in, the more accurate the prediction is.

The performance of our scoring function on mutant data sets compiled by others are as follows (% correct predictions): 66% for the 1096-set and 63% for the 388-set from Cheng et al. (2006), 60% for the 50 outliers from Guerois et al. (2002), and 79% of the 24 mutants from Topham et al. (1997). We also scored our 210 mutants with stability data using the FOLD-X program (Guerois et al., 2002), and 68% of these mutants were identified correctly by this program (compare to our accuracy of 80.5%). While the web interface for the FOLD-X program is handy when scoring a handful

Table 2. Spearman rank correlation coefficients for mutant sets whose change in stability/reactivity has been quantified. # Mut gives the number of mutants, and RC gives the rank correlation coefficient. *: indicates that the WT is also included in the rank calculations. †: We could use only five of the mutants reported by Hahn et al. out of a total of seven (see Entry 12 in Table 1), as the mutants 103Q and 105K were listed as unmeasurable. The last five sets of mutants were all reported by Carter et al. (2001).

Study	Experimental quantity	# Mut	RC
Martin et al. (2001)	melting temperature	5	0.90
Chen and Gouaux (1997)	enthalpy of activation	6*	1.00
Oppermann et al. (1997)	reaction rates (K)	6*	-0.37
Funahashi et al. (2002)	stability z-values	32	0.67
Hahn et al. (1995)	WT/mutant reactivity	5 †	-0.10
barnase	$\Delta\Delta G_{\text{unfold}}$	9	0.90
chymotripsin inhibitor	$\Delta\Delta G_{\text{unfold}}$	9	0.82
staphylococcal nuclease	$\Delta\Delta G_{\text{unfold}}$	19	0.87
calbindin	$\Delta\Delta G_{\text{unfold}}$	9	0.78
T4 lysozyme	$\Delta\Delta G_{\text{unfold}}$	27	0.63

of mutations, we found it quite tedious to score all the 210 mutants from our test set (took us several hours). We believe that researchers should provide executable file(s) for scoring functions that could handle large sets of mutations simultaneously.

The key point to note when comparing our scoring function to others is that unlike the previous methods, we have *not trained* our scoring function on a set of mutations. Thus, an SVM trained on our scoring function could well have the largest accuracy yet reported – we are currently trying to implement this idea.

3.1 Combinatorial mutagenesis: an example

Our implementation of combinatorial mutagenesis (Section 2.3) scored all 64,000,000 mutants of 1CSQ (including the WT) within 6 hours (on a typical PC). In comparison, calling the four-body scoring function separately for each mutant did not finish in 24 hours. The original authors reported only five stabilizing mutants. Combinatorial mutagenesis predicted all five of them correctly. Furthermore, they were in the top 17.7% (of 64 million mutants). Analysis of the top-scoring mutants shows that high scores are assigned for mutants with Cystines in the selected sites. As illustrated in Figure 1, the six mutation sites participate in several tetrahedra together, i.e., they are linked to each other. The occurrences of stabilizing disulfide bonds between cystines is scored among the highest by the four-body scoring function, and hence the mutants with two or more Cystines are naturally scored high (Tropsha et al., 1996, 1998).

3.2 Comments on the work of Masso et al. (2006)

Masso, Lu, and Vaisman analyzed mutants reported in three different papers (and some more mutants that were not reported in these papers) using an earlier version of the four-body scoring function. In the first of these papers, Loeb et al. (1989) studied mutants of HIV-1 Protease. They reported a western blot assay analysis as well as an enzyme activity analysis. Most mutations were reported to have ambiguous "WT-like" behavior. Furthermore, only mutations of the wild type western blot assay classification could be considered, as they were the only group for which the production of the protein was explicitly proven. This is an important factor, as the authors were mutating an operon, which leads to the production of multiple HIV proteins. Our results of this study were poorer than the other reported results. Out of 56 mutants, only 21 were correctly scored. This lower accuracy may be accounted for by the unreported global destabilization of the enzyme. In another paper considered, Wrobel et al. (1998), analyzed mutants of HIV-1 reverse transcriptase. The analysis as well as the results presented were similar to those presented by Loeb et al., and a subset of appropriate mutations was selected in a similar manner. From among the 105 selected mutants, the four-body scoring function correctly predicted 51 mutants. Once again, the incorrect scoring of the remaining mutants in this study could very well be due to the global destabilization of the protein, which the authors did not report explicitly.

4 CONCLUSIONS

The strengths of the four-body scoring function for predicting stability and reactivity effects of mutations are widespread applicability, consistency (one setting works for all cases), computational efficiency (combinatorial mutagenesis), and accuracy. The idea of combinatorial mutagenesis can in principle be used even for a single mutation, or a few of them, but the gain in computational efficiency might not be noticeable.

Our test set is comprehensive, but is not *complete* – we plan to further explore previous as well as forthcoming literature to add new sets of mutants to the current ones. Even though the same settings are recommended for applying our scoring function to all proteins, we could customize it with different settings specifically for certain classes of proteins, thus increasing its accuracy (of course, the scoring function will not perform as well under such customized settings for other classes of proteins). Another idea for increasing the accuracy of the scoring function is to use different weights for various quadruplets (rather than simply adding them all up). We could divide the set of mutants into training and test sets, determine the weights by learning from the training set and then validate them on the test set. We are currently investigating these and other ideas.

5 ACKNOWLEDGMENTS

Both authors are thankful for the support provided by the NSF UBM Grant DEB 0531870 while working on this research project.

REFERENCES

Almog, O., Gallagher, D. T., Ladner, J. E., Strausberg, S., Alexander, P., Bryan, P., and Gilliland, G. L. (2002) Structural basis of thermostability analysis of stabilizing mutations in subtilisin bpn. *The Journal of biological chemistry*, **277**(30), 27553–27558.

Bandyopadhyay, D. and Snoeyink, J. (2004) Almost-Delaunay simplices: nearest neighbor relations for imprecise points. In *SODA '04: Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms*, 410–419, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

Bonander, N., Leckner, J., Guo, H., Karlsson, B. G., and Sjölin, L. (2000) Crystal structure of the disulfide bond-deficient azurin mutant c3a/c26a. *European Journal of Biochemistry*, **267**, 4511–4519.

Braun, P., Persson, B., Kaback, H. R., and von Heijne, G. (1997) Alanine insertion scanning mutagenesis of lactose permease transmembrane helices. *Journal of Biological Chemistry*, **272**(47), 29566–29571.

Brownlie, P. D., Lambert, R., Louie, G. V., Jordan, P. M., Blundell, T. L., Warren, M. J., Cooper, J. B., and Wood, S. P. (1994) The three-dimensional structures of mutants of porphobilinogen deaminase: Toward an understanding of the structural basis of acute intermittent porphyria. *Protein Science*, **3**(10), 1644–1650.

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.

Chen, G.-Q. and Gouaux, E. (1997) Reduction of membrane protein hydrophobicity by site-directed mutagenesis: introduction of multiple polar residues in helix d of bacteriorhodopsin. *Protein Engineering*, **10**, 1061–1066.

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**, 1125–1132.

de Antonio, C., del Pozo, A. M., Mancheño, J. M., Oñaderra, M., Lacadena, J., Martínez-Ruiz, A., Pérez-Cañadillas, J. M., Bruix, M., and Gavilanes, J. G. (2000) Assignment of the contribution of the tryptophan residues to the spectroscopic and functional properties of the ribotoxin alpha-sarcin. *Proteins: Structure, Function, and Genetics*, **41**(3), 350–361.

Dvir, H., Harel, M., McCarthy, A. A., Toker, L., Silman, I., Futerman, A. H., and Sussman, J. L. (2003) X-ray structure of human acid-β-glucosidase, the defective enzyme in gaucher disease. *EMBO Reports*, **4**(7), 704–709.

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

Erwin, C., Barnett, B., Oliver, J., and Sullivan, J. (1990) Effects of engineered salt bridges on the stability of subtilisin bpn. *Protein Engineering*, **4**(1), 87–97.

Fowler, A., Krishnamoorthy, B., and Stratton, K. (2007) A hierarchy of scoring functions for protein decoy discrimination based on Delaunay tessellation of proteins. *Bioinformatics*, under review.

Funahashi, J., Takano, K., Yamagata, Y., and Yutani, K. (2002) Positive contribution of hydration structure on the surface of human lysozyme to the conformational stability. *The Journal of biological chemistry*, **277**(24), 21792–21800.

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.

Ge, X., Chen, J., Sun, C., and Cao, K. (2003) Preliminary study on the structural basis of the antifungal activity of a rice lipid transfer protein. *Protein engineering*, **16**(6), 387–390.

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.

Guerois, R., Nielsen, J. E., and Serrano, L. (2002) Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *Journal of Molecular Biology*, **320**, 369–387.

Hahn, M., Olsen, O., Politz, O., Boriss, R., and Heinemann, U. (1995) Crystal structure and site-directed mutagenesis of bacillus macerans endo-1, 3-1, 4 -beta glucanase. *Journal Biology and Chemistry*, **270**, 3081–3088.

Huang, H.-B., Horiuchi, A., Goldberg, J., Greengard, P., and Nairn, A. C. (1996) Site-directed mutagenesis of amino acid residues of protein phosphatase 1 involved in catalysis and inhibitor binding. *The American Society for Biochemistry and Molecular Biology, Inc*, **271**(5), 2574–2577.

Humphrey, W., Dalke, A., and Schulten, K. (1996) VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, **14**, 33–38.

Köditz, J., Ulbrich-Hofmann, R., and Arnold, U. (2004) Probing the unfolding region of ribonuclease a by site-directed mutagenesis. *European Journal of Biochemistry*, **271**, 4147–4156.

Kong, K.-H., Inoue, H., and Takahashi, K. (1993) Site-directed mutagenesis study on the roles of evolutionally conserved aspartic acid residues in human glutathione s-transferase p1-1. *Protein Engineering*, **6**(1), 93–99.

Korkegian, A., Black, M. E., Baker, D., and Stoddard, B. L. (2005) Computational thermostabilization of an enzyme. *Science*, **308**(5723), 857–860.

Krishnamoorthy, B. and Stratton, K. (2007) Ranking CASP predictions using a four-body scoring function. *Proteins: Structure, Function, and Bioinformatics*, under review.

Krishnamoorthy, B. and Tropsha, A. (2003) Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations. *Bioinformatics*, **19**(12), 1540–1548.

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, database issue; ProTherm link: http://gibk26.bse.kyutech.ac.jp/jouhou/Protherm/protherm.html.

Loeb, D. D., Swanstrom, R., Everitt, L., Manchester, M., Stamper, S. E., and Hutchison, C. A. (1989) Complete mutagenesis of hiv-1 protease. *Nature*, **340**(6232), 397–400.

Martin, A., Sieber, V., and Schmid, F. (2001) In-vitro selection of highly stabilized protein variants with optimized surface. *Journal of Molecular Biology*, **309**(3), 717–726.

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.

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.

Munson, P. J. and Singh, R. K. (1997) Statistical significance of hierarchical multi-body potentials based on Delaunay tessellation and their application in sequence-structure alignment. *Protein Science*, **6**, 1467–1481.

Oppermann, U. C. T., Filling, C., Berndt, K. D., Persson, B., Benach, J., Ladenstein, R., , and JÕrnvall, H. (1997) Active site directed mutagenesis of $3^{\beta}/17^{\beta}$-hydroxysteroid dehydrogenase establishes differential effects on short-chain dehydrogenase/reductase reactions. *Biochemistry*, **36**(1), 34–40.

Ormö, M., Regnström, K., Wang, Z., Jr., L. Q., Sahlin, M., and Sjöberg, B.-M. (1995) Residues important for radical stability in ribonucleotide reductase from escherichia coli. *Journal of Biological Chemistry*, **270**(12), 6570–6576.

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

Pathange, L. P., Bevan, D. R., Larson, T. J., and Zhang, C. (2006) Correlation between protein binding strength on immobilized metal affinity chromatography and the histidine-related protein surface structure. *Analytical Chemistry*, **78**(13), 4443–4449.

Siadat, O. R., Lougarre, A., Lamouroux, L., Ladurantie, C., and Fournier, D. (2006) The effect of engineered disulfide bonds on the stability of drosophila melanogaster acetylcholinesterase. *Journal of Biological Chemistry*, **7**(12).

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

Sun, D., Jones, L. H., Mathews, F. S., and Davidson, V. L. (2001) Active-site residues are critical for the folding and stability of methylamine dehydrogenase. *Protein Engineering*, **14**(9), 675–681.

Suresh, M. V., Singh, S. K., Jr., D. A. F., and Agarwal, A. (2006) Role of the property of c-reactive protein to activate the classical pathway of complement in protecting mice from pneumococcal infection. *The Journal of Immunology*, **176**, 4369–4374.

Takano, K., Tsuchimori, K., Yamagata, Y., and Yutani, K. (1999) Effect of foreign n-terminal residues on the conformational stability of human lysozyme. *European Journal of Biochemistry*, **266**, 675–682.

Topham, C. M., Srinivasan, N., and Blundell, T. L. (1997) Prediction of the stability of protein mutants based on structural enviornment-dependent amino acid substitution and propensity tables. *Protein Engineering*, **10**(1), 7–21.

Tropsha, A., Singh, R. K., 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.

Tropsha, A., Vaisman, I. I., and Zheng, W. (1998) Compositional preferences in quadruplets of nearest neighbor residues in protein structures: statistical geometry analysis. In *IEEE Symposia on Intelligence and Systems*, 163–168.

Wrobel, J. A., Chao, S.-F., Conrad, M. J., Merker, J. D., Swanstrom, R., Pielak, G. J., and Hutchison, C. A. (1998) A genetic approach for identifying critical residues in the fingers and palm subdomains of hiv-1 reverse transcriptase. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 638–645.