Beilstein-Institut

Bozen 2006, May 15th – 19th, 2006, Bozen, Italy

Estimating the Effects Of Single Nucleotide Polymorphisms On Protein Structure: How Good are we at Identifying Likely Disease Associated Mutations?

Catherine L. Worth, David Burke and Tom L. Blundell

Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, U.K.

E-Mail: *catherine@cryst.bioc.cam.ac.uk

Received: 27th July 2006 / Published: 5th November 2007

Abstract

Understanding the effects that non-synonymous single nucleotide polymorphisms have on the structures of the gene products, the proteins, is important in identifying the origins of complex diseases. A method based on amino acid substitutions observed within homologous protein families with known 3D structures was used to predict changes in stability caused by mutations. In the task of predicting only the sign of stability change, our method performs comparably or better to other published methods with an accuracy of 71%. The method was applied to a set of disease associated and non-disease associated mutations and was shown to distinguish the two sets in terms of protein stability. Our method may therefore have application in correlating SNPs with diseases caused by protein instability.

Introduction

Single nucleotide polymorphisms (SNPs) are the most common source of variation in the genome. Due to the redundant nature of the RNA triplet code that encodes proteins, many of these SNPs will not cause an amino acid change in the encoded protein (synonymous

mutations). However, where a SNP causes an amino acid change (a non-synonymous mutation) there may be an effect on the structure or function of the encoded protein. Where protein function is lost, this may lead to disease. It would be extremely useful to be able to predict which mutations are likely to cause disease. Identifying those SNPs that infer susceptibility or protection to complex diseases will aid early diagnosis, prevention and treatments to these diseases [1].

A SNP may affect the function of a protein in three main ways. Firstly, a SNP may affect the functional residues of a protein i.e. the active site or protein-protein interaction site, impairing the protein's ability to carry out its function and hence affecting the molecular pathway within which the protein functions. Secondly, a SNP may affect the stability of a protein by either destabilizing it (increasing the ratio of unfolded protein to folded protein) or stabilizing it (decreasing the ratio of unfolded protein to folded protein). A third effect of SNPs, related to protein stability, is that of causing protein aggregation.

In its native state a protein's 3D structure is folded into regions of secondary structure. However, under conditions of stress e.g. high temperature, the protein may denature to an unfolded state which is more flexible and highly hydrated. The stability of a protein reflects its ability to resist this conformational change under stress. Protein stability differences between wild-type and mutant proteins can be calculated using the thermodynamic cycle (Fig. 1).

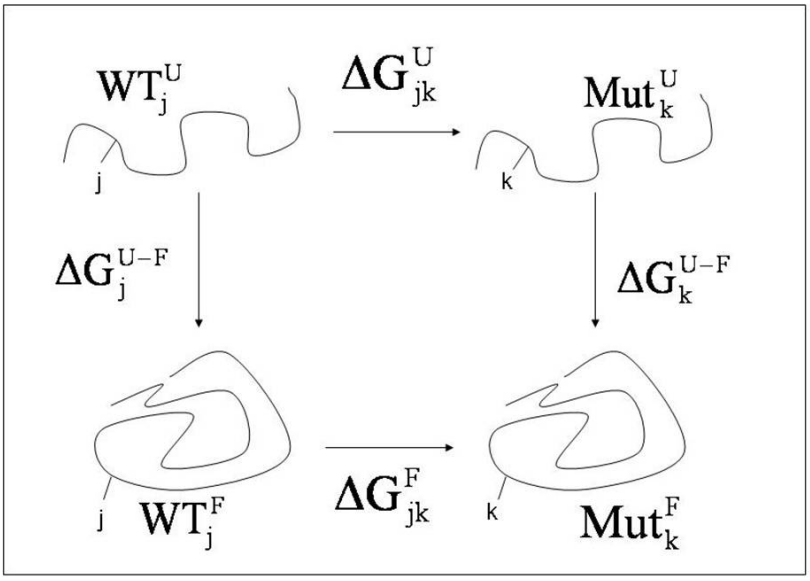

Figure 1. The thermodynamic cycle can be used to calculate protein stability changes between wild-type and mutant proteins.

The difference in free energy of unfolding of the wild type (*j*) and mutant (*k*), ΔΔG, is calculated by:

$$\Delta\Delta G = \Delta G_k^{U-F} - \Delta G_j^{U-F} = \Delta G_{jk}^{U} - \Delta G_{jk}^{F} \quad (1)$$

where ΔG_k^{U-F} and ΔG_j^{U-F} represent the free energy change going from the unfolded (U) to the folded (F) state for the mutant and wild type proteins respectively. Direct simulation of the unfolding process is not possible. As the total free energy in the full cycle is zero, the ΔΔG can instead be calculated using the free energy changes associated with the transformation of $j \rightarrow k$ in the unfolded and folded state (ΔG_{jk}^{U} and ΔG_{jk}^{F} respectively).

Various methods of predicting protein stability changes caused by mutation have been described and can be grouped into four main categories based on the method used in the calculation; (1) physical effective energy functions, (2) empirical potential energy functions, (3) machine learning methods and (4) statistical potential energy functions.

Physical effective energy functions (such as molecular mechanics approaches) are currently only useful for testing small sets of mutants due to the large amount of time required to compute calculated ΔΔG values [2 – 5]. The reliability of predictions are also questionable due to difficulties in sampling in the folded and unfolded states [6]. Empirical energy functions are fitted to experimental data using a set of weighted terms incorporating physical and statistical factors with structural knowledge [7, 8]. The empirical energy function is then tested on a second set of mutants in order to assess the accuracy of the method. Machine learning methods include neural networks and support vector machines (SVMs) and use information about mutations, protein sequence and structural information to fit a non-linear function to experimental data [9 – 12]. They are similar to empirical energy functions in their use of experimental data to fit their function and in both cases, care must be taken that the function is not over-fitted to the training data set. Statistical potential energy functions are derived using statistical analysis of information from protein databases such as substitution frequencies, distance potentials and amino acid environmental propensities [13 – 15].

Site Directed Mutator (SDM) is a statistical potential energy function developed by Topham *et al.*, [13] to predict the effect that SNPs will have on the stability of proteins. SDM uses amino acid substitution frequencies within homologous protein families to calculate a stability score which is analogous to the free energy difference between a wild-type and mutant protein. Blind testing on a set of 83 staphylococcal nuclease and 63 barnase mutants showed a correlation of 0.80 in the predicted stability changes with experimental data [13].

Here we apply SDM to a more extensive set of mutant proteins taken from the Protherm database [16] and obtain correlations of 0.60 and 0.68 for monomeric and crystallographic mutants. We also compare SDM's predictive power to other published methods and find that our method performs comparably or better to other methods in the task of predicting

whether a mutation will be stabilizing or destabilizing. We apply SDM to a set of disease-associated mutations and a set of non disease-associated mutations and find that our method is able to distinguish the two sets of mutations.

Methods

Homologous Structure Alignment Database (HOMSTRAD)

HOMSTRAD [17] clusters all known protein structures from the PDB [18] into homologous protein families. These families represent groups of proteins that have a common evolutionary origin. Most families have an average of 30% sequence identity and no pair has more than 90% sequence identity to each other. Representative structures for each protein family are chosen based on the quality of the X-ray analysis and resolution. Where a family contains two members or more a structural alignment is carried out using COMPARER [19]. The alignment is annotated using JOY to identify the local structural environment of each residue in the alignment [20].

Environment-specific substitution tables

A set of conformationally constrained environment-specific substitution tables (ESSTs) was constructed as described previously by Topham *et al.* [21]. The tables were derived from 371 protein families from the HOMSTRAD database, consisting of 1357 structures, and were built using the program Makesub (C. Topham, unpublished). The ESSTs hold the probability of each amino acid type existing in a particular environment being substituted by any other amino acid.

Definition of structural environment

The structural parameters that were used to define the local environment of amino acid residues were main chain conformation, solvent accessibility and hydrogen-bonding class.

1. *Main-chain conformation and secondary structure*
   Nine classes of main-chain conformation were defined: residues were identified as belonging to either a helix or β-sheet first and the remaining residues were classified as being *a, b, p, t, l, g* or *e* according to their main-chain φ-ψ torsion angles [21, 22]. The torsion angles and secondary structure assignments were calculated using the SSTRUC program (D. Smith, unpublished).
2. *Relative side-chain solvent accessibility*
   Three classes of relative side-chain solvent accessibility were defined based on the method of Lee and Richards [22]. Residues with side-chain relative accessibilities of:

I. < 7% were defined as inaccessible

II. 7 to 40% were defined as partially accessible

III. > 40% were defined as accessible

3. *Hydrogen bonding*
   Two classes of hydrogen bonding were defined: residues were classed as either forming a side-chain hydrogen bond or not. The program, Hbond (J. Overington, unpublished), was used to identify hydrogen bonds defined by the criterion that the distance between donor and acceptor was less than 3.5Å.

These structural parameters gave a total of 54 local environments (9 main-chain x 3 solvent accessibility x 2 hydrogen bonding terms).

Prediction of protein stability changes caused by mutation

The algorithm described by Topham *et al.* [13] was used to calculate a stability score difference between wild-type and mutant proteins. By analogy to the folding-unfolding cycle in Fig. 1, the algorithm uses ESSTs to calculate the difference in the stability scores for the folded and unfolded state for the wild-type and mutant protein structures:

$$\Delta\Delta S = \Delta s^{U}_{jk} - \Delta s^{F}_{jk} \tag{2}$$

ESSTs only take into account the environment of one of the two residues (wild-type or mutant), therefore it is necessary to consider not only the probability of replacement of the wild-type residue (R_j) in the wild-type environment (ε_{wt}) by a mutant residue type (r_k) in an undefined environment ($P(r_k/R_{j,}, \varepsilon_{wt})$) but also the probability of replacement of the mutant residue type (R_k) in the mutant environment (ε_{mut}) by the wild-type residue (r_j) in an undefined environment ($P(r_k/R_{j,}, \varepsilon_{mut})$).

In order to normalize the probabilities that are combined from different substitution tables, it is necessary to introduce a reference state. For the wild-type residue (R_j) in the wild-type environment a suitable reference state is the probability of it being conserved in that environment ($P(r_j/R_{j,}, \varepsilon_{wt})$). In an analogous way, for the mutant residue type (R_k) in the mutant environment, a suitable reference state is the probability of it being conserved in that environment ($P(r_k/R_{k,}, \varepsilon_{mut})$).

The difference in stability scores for a mutation in the folded state is therefore calculated by:

$$\Delta s^{F}_{jk} = \sum_{i} -\ln\left\{\frac{P(r_{ki}/R_{ji},\mu_{wt})}{P(r_{ji}/R_{ji},\mu_{wt})}\cdot\frac{P(r_{ki}/R_{ki},\mu_{mut})}{P(r_{ji}/R_{ki},\mu_{mut})}\right\} \tag{3}$$

The difference in stability scores in the unfolded state (Δs_{jk}^{U}) is also calculated using Equation 3 but uses an environmental substitution table derived from non-hydrogen bonded, surface exposed amino acid residues falling outside regions of regular secondary structure. The stability difference score for the folded and unfolded state for the wild-type and mutant protein structures is then calculated using Equation 2.

The definition used for accessible, partially buried and buried residues was different to that used to generate the ESSTs. Our earlier benchmarking had shown that the best results were obtained when residues with side-chain relative accessibilities of $< 17\%$ were defined as inaccessible, 17 to 59% as partially accessible and $> 59\%$ as accessible. The higher percentage solvent accessibilities used here are probably due to the fact that we are trying to predict the effect of single mutations on protein structure whereas the ESSTs occur in the context of a protein that may have accepted compensating substitutions elsewhere as a result of evolution.

Mutant thermodynamic datasets

A subset of the dataset used by Capriotti *et al.* [11] was used in this study. The mutant dataset was taken from the Protherm database which houses thermodynamic data for proteins and mutants [16]. Our method requires knowledge of the local structural environment of wild-type and mutant residues in order to predict the effect of mutation on the stability of a protein. If the local environment is incorrectly defined e.g. the protein functions as a trimer but is defined in the crystallographic asymmetric unit as the protomer, this may affect our calculation. To remove the effect of such errors we used the Protein Quaternary Structure (PQS) database to predict the oligomeric state of each of the proteins in the dataset [23]. Only those proteins that were predicted to be and solved as a monomer were used. For the same reason, proteins containing heteroatoms in their PDB file other than water or that were resolved at a resolution > 2 Å were also removed from the dataset. This dataset is hereafter referred to as the *monomeric set*.

A second set of mutants with crystal structures was taken from the Protherm database. These were all single mutants with monomeric structures. This dataset is hereafter referred to as the *crystallographic set*.

A third set of 388 mutants (*S388*) with thermodynamic measurements conducted at physiological conditions was also used to test our method. The *S388* dataset has been used to test other published methods and therefore allows us to perform a direct comparison of our method to them.

Building models of mutant proteins

The program, ANDANTE, was used to build models of the mutant proteins by building the mutant side-chains from a high-quality rotamer library. ANDANTE adds the lowest energy rotamer to the target and checks for clashes against the backbone [24].

Assessment of performance

To assess performance of our method in predicting the sign of stability change caused by mutation we calculate the accuracy:

$$Q = \frac{\left(t_n + t_p\right)}{\left(t_n + f_n + t_p + f_n\right)} \tag{4}$$

where TN, TP, FN and FP refer respectively to the number of true negatives, true positives, false negatives and false positives.

An alternative way of assessing the performance of our method in classifying correctly the sign of free energy change caused by mutation is to calculate the Matthew's Correlation Coefficient (MCC) [25]:

$$c = \frac{t_p t_n - f_p f_n}{\sqrt{\left(t_p + f_p\right)\left(t_p + f_n\right)\left(t_n + f_p\right)\left(t_n + f_n\right)}} \tag{5}$$

In order to assess how well our method predicts stabilizing and destabilizing mutations we calculate the sensitivity [TP/(TP+FN)] and specificity [TP/(TP+FP)] of stabilizing mutations and the sensitivity [TN/(TN+FP)] and specificity [TN/(TN+FN)] of destabilizing mutations.

We use a linear correlation coefficient measure (LCC) to assess the performance of our method in predicting the amount of free energy change caused by mutation

$$r = \frac{\left(n\sum XY - (\sum X \sum Y)\right)}{\sqrt{\left[n\sum X^2 - \left(\sum X\right)^2\right]\left[n\sum Y^2 - \left(\sum Y\right)^2\right]}} \tag{6}$$

Where r is the correlation coefficient, n is the number of data, and X and Y are the experimental and assigned stability respectively.

Disease-associated and non-disease-associated SNP data sets

A set of disease-associated mutations (da-SNPs) was compiled from the Online Mendelian Inheritance in Man (OMIM) and Catalogue of Somatic Mutations in Cancer (Cosmic) [26, 27]. An alternative set of non-disease-associated mutations (nd-SNPs) was compiled from dbSNP [28]. Where possible, structural homologues were identified using FUGUE [29] and comparative models built using the program Modeller [30]. Models of the mutant proteins were created using ANDANTE, as described previously. SDM was then used to predict the effect of the mutations on the stability of the proteins.

Results and Discussion

Monomeric dataset

A dataset of 223 mutants comprising 3 proteins was created after filtering for proteins that:

i. were predicted to function as monomers as well as being resolved as monomers

ii. were resolved at a resolution of < 2Å

iii. did not contain any HET atoms other than water in their PDB file.

Prediction of protein stability changes caused by mutation

The correlation of the predicted and observed ΔΔG values for the 223 mutants was 0.60 (Fig. 2). A breakdown of the prediction performance shows that our method has an accuracy of 0.74 with sensitivity of 0.76 and 0.72 for stabilizing and destabilizing mutations respectively and specificity of 0.7 and 0.77 for stabilizing and destabilizing mutations respectively (Table 1).

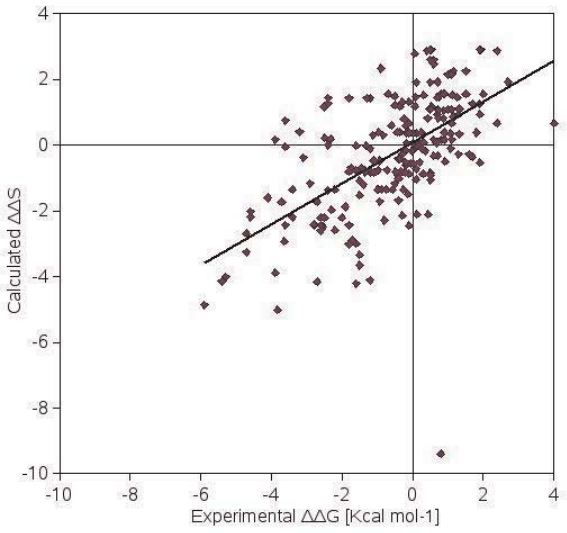

Figure 2. The experimentally measured energy changes versus the predicted energy changes using our method, SDM, on the monomeric dataset. The correlation is 0.60 and the standard error is 1.36 kcal mol-1. Removal of the outlying data point increases the correlation to 0.66.

There is one outlying data point (lower right hand corner of Fig. 2) which involves a mutation from alanine to cysteine in Ribonuclease H from *Escherichia coli* (PDB code: 2RN2). Our method predicts this mutation to be highly destabilizing when in fact it is mildly stabilizing. Looking at the substitution tables used to calculate the score, it is observed that mutating cysteine to alanine or conserving cysteine in the folded state has not been observed for that environment (buried, helix, hydrogen bonded). In the unfolded state mutating cysteine to alanine has not been observed but there is a high probability of cysteine being conserved. This results in a predicted value of $\Delta\Delta G$ that is hugely destabilizing. It appears that in this case there were insufficient data in the substitution table to be able to predict the effect of mutating alanine to cysteine on the stability of the protein. Increasing the number of families used to generate the ESSTs may help to tackle this. In view of these uncertainties we investigated removal of this one outlier and found that this increases the correlation to 0.66.

Table 1. Results (accuracy, sensitivity and specificity of stabilizing and destabilizing mutations and linear correlation coefficient) of SDM's stability predictions for the monomeric and crystallographic datasets.

Dataset	Accuracy	Specificity +ve	Sensitivity +ve	Specificity -ve	Sensitivity -ve	LCC
Monomeric	0.74	0.7	0.76	0.77	0.72	0.60
Crystallographic	0.72	0.67	0.66	0.75	0.76	0.68

Crystallographic dataset

A dataset of 252 mutants comprising 3 proteins (Ribonuclease H, lysozyme & trypsin) was created after filtering for proteins that were predicted to function as monomers as well as being resolved as monomers. All of these proteins were resolved at a resolution of < 2 Å.

The correlation of the predicted and observed $\Delta\Delta G$ values for the crystallographic mutants was 0.68 (Fig. 3). A breakdown of the prediction performance shows that our method has an accuracy of 0.72 with sensitivity of 0.66 and 0.76 for stabilizing and destabilizing mutations respectively and specificity of 0.67 and 0.75 for stabilizing and destabilizing mutations respectively (Table 1).

The results from the *monomeric* and *crystallographic set*s are extremely similar except that the *monomeric set* has a slightly higher accuracy (0.74 compared to 0.72) and the *crystallographic set* a higher LCC (0.68 compared to 0.60) (Table 1). The general trend is the same between the two sets, with the specificity of predicting destabilizing mutations being slightly higher than that for stabilizing (by ~8%). The results show that it is not necessary to have a crystal structure of a mutant protein in order to predict the effect of mutation on the stability of the protein.

The fact that the *crystallographic set* obtained a higher LCC than the *monomeric set* even though it had lower accuracy and sensitivity for predicting stabilizing and destabilizing mutations would appear to be inconsistent. However, the LCC obtained may have been improved by the presence of data points falling in the lower right hand part of the graph (Fig. 3).

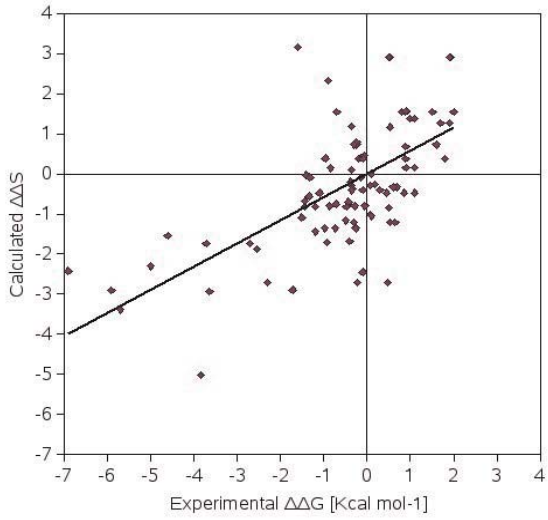

Figure 3. The experimentally measured energy changes versus the predicted energy changes using our method, SDM, on the crystallographic mutant dataset. The correlation is 0.68 and the standard error is 1.35 kcal mol-1.

S388 dataset

We compared our method's ability to classify mutations as stabilizing or destabilizing to other published methods by using the *S388* dataset. Predictions of the following methods were used: FOLDX [6], DFIRE [31], PoPMuSiC [14] (all energy-based methods), NeuralNet [11] and three SVM methods using sequence only (SO), structure only (TO) and sequence and structure (ST) information [10]. Results for the energy-based methods and NeuralNet were taken from Capriotti *et al.* [11] and the SVM methods from Cheng *et al.* [10].

Our method performs comparably or better than the other methods in the task of classifying mutations as stabilizing or destabilizing (Table 2). Our method has the 2nd highest correlation coefficient (0.27), equal to the ST SVM (0.27) and bettered only by the TO SVM (0.28). If we used only accuracy as a measure of performance then all methods perform comparably with NeuralNet having the highest accuracy (0.87) and DFIRE the lowest accuracy (0.68). Although the accuracy of predicting whether a mutation is stabilizing or destabilizing is above 68% for all the methods, the sensitivity of predicting stabilizing

mutations is poor. Five out of the seven methods incorrectly classify > 56% of the stabilizing mutations. Our method has a much improved sensitivity of predicting stabilizing mutations (0.67) compared to the others reported but still classifies 33% of stabilizing mutations incorrectly. The *S388* dataset was a more challenging test for our method considering that it was not filtered using the parameters described for the *monomeric set*. It is therefore very encouraging that our method performed comparably or better to other methods for this task.

Table 2. Comparison of SDM with other methods on *S388*.

Method	MCC	Accuracy	Sens. (+)	Spec. (+)	Sens. (-)	Spec. (-)
FOLDX	0.25	0.75	0.56	0.26	0.78	0.93
DFIRE	0.11	0.68	0.44	0.18	0.71	0.90
PoPMuSic	0.20	0.85	0.25	0.33	0.93	0.90
NeuralNet	0.25	0.87	0.21	0.44	0.96	0.90
SO	0.26	0.86	0.30	0.40	0.94	0.90
TO	0.28	0.86	0.31	0.42	0.94	0.91
ST	0.27	0.86	0.31	0.40	0.93	0.91
SDM	0.27	0.71	0.67	0.25	0.72	0.94

A general problem with current methods of predicting protein stability changes caused by SNPs is that they tend to be over-fitted to the mutant dataset they have been developed on. Most mutations are destabilizing and this is reflected in the mutant thermodynamic datasets used for developing and testing such methods. Methods that assign all of the samples to the majority class (destabilizing mutations) will have high accuracy even though the performance is poor for the minority class (stabilizing mutations). This trend is observed with the five methods with the lowest sensitivities for stabilizing mutations (Table 2).

Although our method performs comparably to other methods, it is currently not robust enough to be applied to all mutant proteins with confidence. The substitutions that the ESSTs hold are the result of evolution – they do not take into account single mutations. We are trying to predict the effect of mutating single residues and therefore hypothesize that the space occupied by the mutant amino acid will not change. Where a single mutation involves a size change there will be a cost associated. The ESSTs could therefore have limitations in this respect.

Disease and non-disease-associated SNP data sets

A total of 6182 nd-SNPs and 879 da-SNPs (797 from OMIM and 82 from Cosmic) had comparative models built of their encoding proteins. The stabilities of all of these modelled mutant structures were predicted using our method.

The stability predictions for the nd-SNP set of mutations have a normal distribution for both buried and accessible mutations (Fig. 4) with most mutations tending to have a neutral effect on protein stability. However, there is a slight shift in the distribution of buried and accessible mutations, with buried mutants tending to be more destabilizing. This is to be expected as mutations within the core of a protein are more likely to perturb the structure than accessible mutations, and hence are more likely to destabilize the protein.

The stability predictions for the da-SNP set of mutations have a somewhat different distribution to that observed with the nd-set. Accessible mutations in the da-SNP set have a largely normal distribution but with smaller peaks observed at -9 to -7 kcal mol^{-1} and 8 to 9 kcal mol^{-1} (Fig. 5). Buried mutations, however, have a skewed distribution, with a higher proportion of buried residues being destabilizing (70.5 %) at -5 to 0 kcal mol^{-1} compared to accessible mutations (52 %). This result is similar to that observed with the nd-SNP set where 72 % of buried mutations are destabilizing at the range of -5 to 0 kcal mol^{-1} compared to 52 % of accessible mutations.

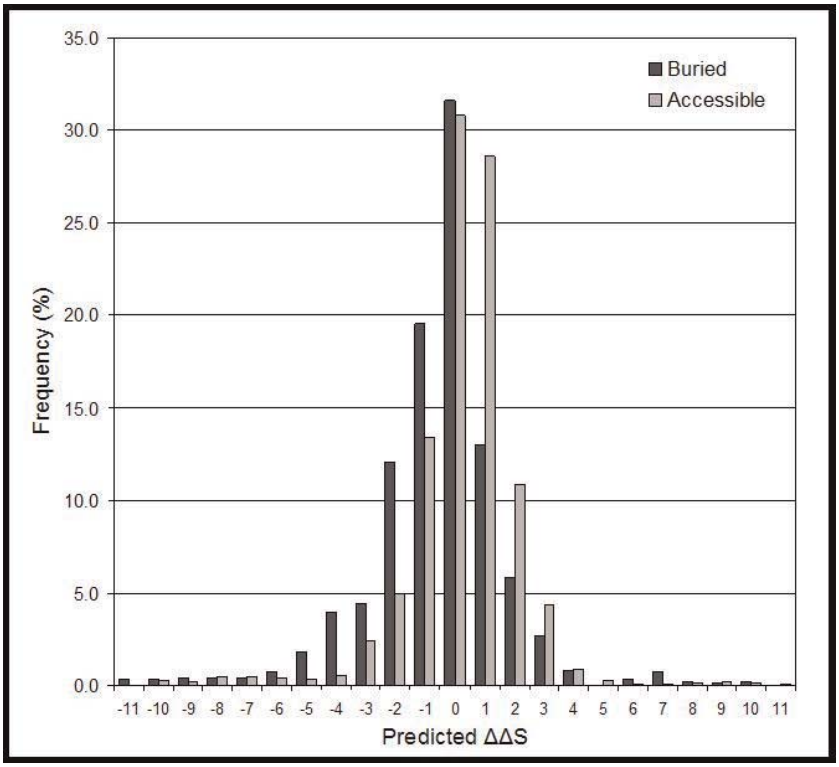

Figure 4. Distribution of accessible and buried residues relative to changes in stability (ΔΔS) in the nd-SNP set of mutations. Most mutations lie within the -4 to 3 kcal mol-1 range.

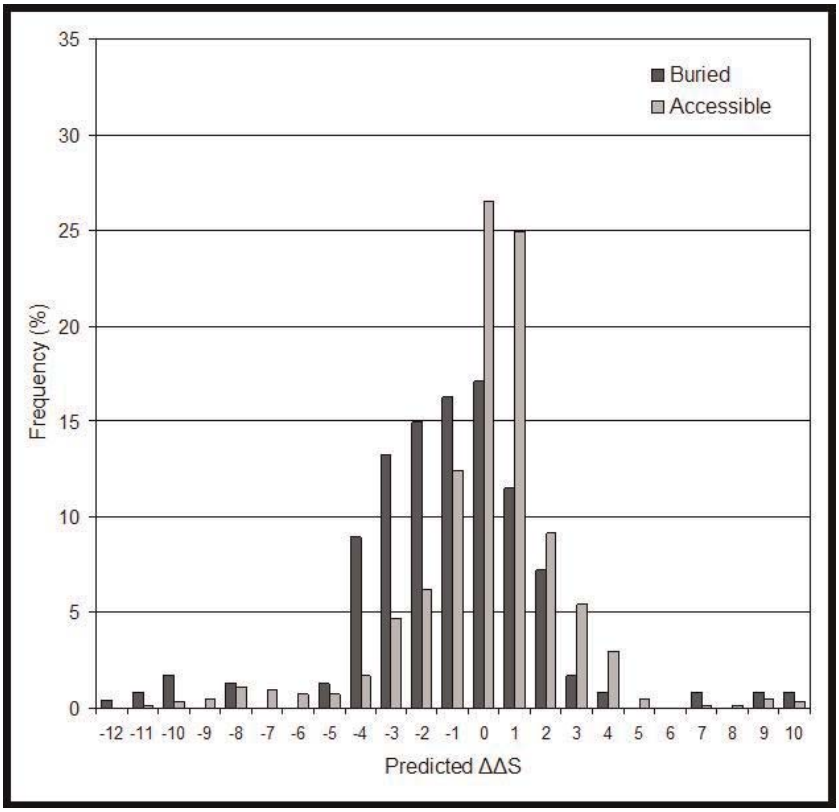

Figure 5. Distribution of accessible and buried residues relative to changes in stability (ΔΔS) in the da-SNP set of mutations. Most mutations lie within the -5 to 4 kcal mol-1 range.

Work by Randles *et al.* [32] has shown that mutations that decrease the stability of a single domain by > 2 kcal mol^{-1} result in severe disease. This trend has also been observed with Ig-like protein superoxide dismutase where mutations which lower the stability by more than 2 kcal mol^{-1} are associated with reduced survival times of patients [33]. Therefore, if we look at those mutations causing ΔΔG values < -2 kcal mol^{-1} it is observed that 25% of buried mutations in the nd-SNP set fall within this range, compared to 43% in the da-SNP set. This result is in agreement with previous work which has found that destabilization of proteins is associated with disease [34 – 36]. More importantly, it indicates that our method can distinguish disease associated SNPs from non-disease associated SNPs.

We find that 27% of the da-SNP set are located at buried sites compared to 16% in the nd-SNP set. This result is consistent with findings by Ferrer-Costa *et al.* [34] that 32% of da-SNPs are located at highly buried sites (< 5% relative accessibility) compared to 7% of nd-SNPs in these locations and estimates by Sunyaev *et al.* [37] that 35% of da-SNPs are

located at buried locations. Our method clearly distinguishes the da-SNP and nd-SNP sets in terms of protein stability and therefore may be of use in correlating SNPs with diseases caused by protein instability.

Conclusion

We have shown that our method performs comparably or better to other published methods in the task of predicting whether a mutation will be stabilizing or destabilizing. An advantage of our method is that it does not use *a priori* knowledge about mutants' thermodynamic measurements. Therefore, there is no bias caused by destabilizing mutations making up the majority class. This is also reflected in the sensitivity and specificity obtained when predicting whether a mutation is stabilizing or destabilizing (Table 1).

Substitutions that have been made within protein families during evolution should be concordant with the underlying protein structure. Although our method has been applied to a limited set of proteins and mutations, it has been shown that our ESSTs are able to reflect how mutations can affect the stability of a protein.

References

[1] Suh, Y., Vijg, J.(2005) SNP discovery in associating genetic variation with human disease phenotypes. *Mutat. Res.* **573**(1 – 2)**:**41**-**53.

[2] Bash, P.A., *et al*. (1987) Free energy calculations by computer simulation. *Science* **236**(4801)**:**564 – 568.

[3] Kollman, P. *et al*. (2000) Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc*. *Chem*. *Res*. **33**(12)**:**889**-**897.

[4] Funahashi, J. *et al*. (2003) How can free energy component analysis explain the difference in protein stability caused by amino acid substitutions? Effect of three hydrophobic mutations at the 56th residue on the stability of human lysozyme. *Protein Eng.* **16**(9)**:**665 – 671.

[5] Park, H., Lee, S. (2005) Prediction of the mutation-induced change in thermodynamic stabilities of membrane proteins from free energy simulations. *Biophys*. *Chem.* **114**(2 – 3)**:**191**-**197.

[6] Shi, Y.Y. *et al*. (1993) Can the stability of protein mutants be predicted by free energy calculations? *Protein Eng*. **6**(3)**:**289**-**295.

[7] Guerois, R., Nielsen, J.E., Serrano, L. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* **320**(2)**:**369 – 387.

[8] Bordner, A.J., Abagyan, R.A.(2004) Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations. *Proteins* **57**(2)**:**400 – 413.

[9] Capriotti, E. *et al*. (2005) Predicting protein stability changes from sequences using support vector machines. *Bioinformatics* **21 Suppl. 2:** ii54-ii58.

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

[11] Capriotti, E., Fariselli, P., Casadio, R.(2004) A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics* **20 Suppl 1:** I63**-**I68.

[12] Frenz, C.M. (2005) Neural network--based prediction of mutation-induced protein stability changes in Staphylococcal nuclease at 20 residue positions. *Proteins* **59**(2)**:**147 – 151.

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

[14] Gilis, D., 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. *J. Mol. Biol*. **272**(2)**:**276 – 290.

[15] Saraboji, K., Gromiha, M.M., Ponnuswamy, M.N. (2006) Average assignment method for predicting the stability of protein mutants. *Biopolymers* **82**(1)**:**80-92.

[16] Kumar, M.D. *et al.* (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res*. **34**(Database issue)**:** D 204 – 206.

[17] Mizuguchi, K. *et al*. (1998) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci*. **7**(11)**:**2469 – 2471.

[18] Berman, H.M. *et al*. (2000) The Protein Data Bank. *Nucleic Acids Res*. **28**(1)**:**235 – 242.

[19] Sali, A., Blundell, T.L. (1990) Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J. Mol. Biol*. **212**(2)**:**403 – 428.

[20] Mizuguchi, K. *et al*. (1998) JOY: protein sequence-structure representation and analysis. *Bioinformatics* **14**(7)**:**617 – 623.

[21] Topham, C.M. *et al*. (1993) Fragment ranking in modelling of protein structure. Conformationally constrained environmental amino acid substitution tables. *J. Mol. Biol*. **229**(1)**:**194 – 220.

[22] Lee, B., Richards, F.M. (1971) The interpretation of protein structures: estimation of static accessibility. *J. Mol*. *Biol*. **55**(3)**:**379-400.

[23] Henrick, K., Thornton, J.M. (1998) PQS: a protein quaternary structure file server. *Trends Biochem. Sci*. **23**(9)**:**358 – 361.

[24] Smith, R.E. *et al*. (2006) Andante: Reducing side-chain rotamer search space during comparative modeling using environment-specific substitution probabilities. *Bioinformatics* (in press).

[25] Matthews, B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **405**(2)**:**442 – 451.

[26] Hamosh, A. *et al*. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **30**(1)**:**52 – 55.

[27] Forbes, S. *et al*. (2001) Cosmic 2005. *Br. J. Cancer* **94**(2)**:**318**-**322.

[28] Sherry, S.T. *et al*. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*. **29**(1)**:**308 – 311.

[29] Shi, J., Blundell, T.L., Mizuguchi, K. (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.* **310**(1)**:**243 – 257.

[30] Sali, A., Blundell, T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**(3)**:**779 – 815.

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

[32] Randles, L.G., Lappalainen, I., Fowler, S.B., Moore, B., Hamill, S.J., Clarke, J. (2006) Using protein models to quantify the effects of pathogenic mutations in Ig-like proteins. *J. Biol. Chem.* **281**(34)**:**24216 – 24226.

[33] Lindberg, M.J. *et al*. (2005) Systematically perturbed folding patterns of amyotrophic lateral sclerosis (ALS)-associated SOD 1 mutants. *Proc. Natl Acad. Sci. U S A* **102**(28)**:**9754 – 9759.

[34] Ferrer-Costa, C.,Orozco, M., de la Cruz, X. (2002) Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *J. Mol. Biol.* **315**(4)**:**771 – 786.

[35] Yue, P., Li, Z., Moult, J. (2003) Loss of protein structure stability as a major causative factor in monogenic disease. *J. Mol. Biol.* **353**(2)**:**459 – 473.

[36] Stitziel, N.O. *et al*. (2003) Structural location of disease-associated single-nucleotide polymorphisms. *J. Mol. Biol.* **327**(5)**:**1021 – 1030.

[37] Sunyaev, S., Ramensky,V., Bork, P. (2000) Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet.* **16**(5)**:**198**-**200.