COMPUTATIONAL MUTAGENESIS MODELS FOR PROTEIN ACTIVITY AND STABILITY ANALYSIS

by

Bill Shili Zhan

A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Bioinformatics

Committee:

	Dr. Iosif Vaisman, Dissertation Director
	Dr. John Grefenstette, Committee Member
	Dr. Timothy Born, Committee Member
	Dr. James Willett, Committee Member
	Dr. Saleet Jafri, Department Chairperson
	Dr. Peter Becker, Associate Dean for Graduate Studies, College of Science
	Dr. Vikas Chandhoke, Dean, College of Science

Date: 12/07/2007

Fall Semester 2007
George Mason University
Fairfax, VA

Computational Mutagenesis Models for Protein Activity and Stability Analysis

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Bill Shili Zhan
Master of Science
Southeastern University, 2000

Director: Iosif I. Vaisman, Associate Professor
Department of Computational Biology and Bioinformatics

Fall Semester 2007
George Mason University
Fairfax, VA

ACKNOWLEDGEMENTS

I am very grateful to my dissertation director, Dr. Vaisman for providing me with a solid working foundation for my research, while inspiring me to explore new ideas. His support extended to many areas of my dissertation studies, including topic selection, detailed technical guidance, and critical review of experimental results.

I would like to express my deep gratitude to my advisory committee members, Dr. John Grefenstette, Dr. Timothy Born, and Dr. James Willett, for their insightful advice and valuable ideas, as well as their precious time for dissertation-related meetings and discussions.

I would like to thank members of the Structural Bioinformatics group at George Mason University. In particular, Dr. Majid Masso and Dr. Yan Ding have always been readily available for sharing and discussing programs and results. Their help throughout my dissertation was invaluable. I am grateful to Zhibin Lu and Rena Liang for writing some of the software our group uses to tessellate proteins.

My appreciation also goes to Mary Margaret Flannery and Glenda Wilson, for all their prompt and efficient administrative support and help and to Chris Ryan for his computer and lab assistance.

Last but not least, a special thanks to my wife, Jennifer, my daughter, Shirley, and my son, Richard, who have always supported me throughout my studies.

TABLE OF CONTENTS

List of Tables

List of Figures

List of Abbreviations

AUC	area under the ROC curve
CM	amino acid side chain center of mass coordinates
CMP	comprehensive mutational profile
C/NC	conservative/non-conservative amino acid substitutions
DT	decision tree
PDB	Protein Data Bank
RF	random forest
RT	reverse transcriptase
ROC	receiver operating characteristic curve
RS	Residual score
RSP	Residual score profile
SE	standard error
SERCA1	Sarcoendo plasmic reticulum calcium-ATPases
SNP	single nucleotide polymorphism
SVM	support vector machine
Weka	Waikato Environment for Knowledge Analysis
wt	wild-type
10 CV	tenfold cross-validation

Abstract

COMPUTATIONAL MUTAGENESIS MODELS FOR PROTEIN ACTIVITY AND STABILITY ANALYSIS

Bill Shili Zhan, Ph.D.

George Mason University, 2007

Dissertation Director: Iosif I. Vaisman

Missense mutations may cause structural alterations of a protein and lead to a loss/gain of activity and stability. Studies of missense mutations in proteins are important for understanding protein structure-function relationships, analyzing the function of gene variations, and designing new proteins. In this dissertation, we have developed computational mutagenesis models to predict the changes of stability and activity of protein mutants using the four-body statistical potential derived from Delaunay tessellations of protein structures. First, our results show that a strong correlation exists between the mean residual scores of mutants and the change of mutant stability in 18 proteins extracted from ProTherm database. Second, we developed robust and accurate machine-learning models based on the residual score profiles of protein mutants to predict the sign of mutant stability change. Third, we have demonstrated a correlation between changes of four-body statistical potential and activity alternation in human p53

and rabbit sarcoendo plasmic reticulum calcium-ATPases (SERCA1) mutants. The supervised machine-learning models based on the residual score profiles of protein mutants were also developed to predict the activity changes in p53 and SERCA1 mutants. Fourth, a highly significant correlation between changes in four-body statistical potential with conservation of amino-acid substitutions was observed. Finally, a novel statistical matrix based on the mean residual scores of all 380 types of mutations in 700 proteins was developed and a statistically significant correlation is revealed between the novel matrix and PAM/BLOSUM matrices. Overall, these conclusions support our hypothesis that computational mutagenesis models using four-body statistical potential present a powerful approach for predicting the changes of activity and stability in protein mutants.

1. Introduction

1.1 Activity and Stability Changes in Protein Mutants

Proteins are the final products of gene expression, but the biological process from DNA sequence to protein is complex. The first stage of the gene expression is to transcribe triplets of nucleotides (codons) in the coding regions of the DNA into RNA. Then RNA is translated into a linear sequence of amino acids in polypeptide chains. Finally, the synthesized amino-acid chain folds spontaneously and efficiently into the unique native conformation, the protein's three-dimensional shape. The precise folding pathways are far from being fully understood although it is thought that the main driving force is the tendency to bury hydrophobic amino-acids in the core of the protein, away from the mainly hydrophilic environment.

The three-dimensional structure of a protein defines its functional properties. Point mutations that result in amino-acid substitutions may disrupt the encoded protein's folding pathway, its three-dimensional structure, and ultimately its functionality. Missense mutations that alter the amino acid sequence of a gene product may potentially affect the cellular phenotype at different aspects. They may directly change the stability of the native protein structure and the folding rate, resulting in a decreased concentration of the protein (Karchin et al 2005). Point mutations located at ligand-binding and catalytic sites may further influence protein interactions and other biochemical activities

inside the cell (Sunyaev et al 2001). They may also have an effect at the level of transcription, translation as well as post-translational modification although these are relatively poorly understood (Wang and Moult 2001).

Analysis of the effects of missense mutation is important in a number of aspects. First, it gives insight into structural and functional features of the corresponding wild-type protein and the relationship of structure-function of proteins. Second, it helps to understand the biological effects of gene variations between individuals. Finally, it can guide for the design of new proteins.

One of the most reliable approaches for studying the effects of missense mutations on activity and stability relies on site-directed mutagenesis (Smith 1985), by which amino acid residue is replaced at key positions and their impacts are determined on activity and stability relative to wild-type (wt). However, such experimental mutagenesis studies by biologists are expensive and time consuming. When it is not feasible to conduct such experiments on all possible substitutions, theoretical prediction of the stability and activity of protein mutants would be useful for guiding site-directed mutagenesis and other protein-engineering techniques. As a result, an increasing number of computational approaches which are inexpensive and efficient have been developed to analyze all possible point mutations in a given protein and determine which amino acids have key structural and functional impacts.

With the steadily increasing availability of biological data extracted from mutagenesis experiments or data from a SNP database, there has been growing interest in developing various computational methods to predict structural and functional effects of

mutations (Nakken et al 2007). The ultimate goal is to discriminating between the neutral, non-functional amino acid mutations and the ones that are functional, leading to a damaging potential to the encoded protein. The computational methods for this problem mainly base on three categories of features for prediction, which are physicochemical features of the amino acids, structural features of the encoded protein, and evolutionary features derived from sequence alignments of homologous proteins (Mooney 2005).

The physicochemical features include molecular mass, polarity, acidity, basicity, aromaticity, conformational flexibility and ability to hydrogen bond. The classic Grantham matrix, measuring chemical distance between the different amino acids, was derived based on these features (Grantham, 1974). A more evolutionary, protein-specific Grantham score was developed by incorporation of sequence variation in a multiple sequence alignment (Tavtigian et al., 2006). The Align-GVGD (Align-Grantham Variation Grantham Deviation) was recently used to analyze missense substitutions in the BRCA1 and p53 (Tavtigian et al., 2006, Mathe et al., 2006a). In addition, a hydropathy scale, measuring the hydrophilicity and hydrophobicity along a protein chain, plays an important role in the protein folding process (Balasubramanian et al., 2005). Zhou et al have developed a hydrophobicity scale to accurately predict protein-protein binding free energies of 21 protein-protein complexes (Zhou and Zhou 2002b).

The structural features include the native amino acid's solvent accessibility, which measures its exposure to the surrounding environment, and the crystallographic B-factor that measures the atomic mobility of the wild-type amino acid and its ability to accommodate mutations (Bowie et al., 1990; Matthews, 1995). Structural information has

been used to predict deleterious human alleles (Chasman and Adams, 2001; Sunyaev et al 2001). Wang and Moult (2001) used a number of structural rules to analyze disease-causing missense SNPs (Wang and Moult 2001). Herrgard et al used structural motifs to predict the effects of amino acid mutations on enzyme catalytic activity (Herrgard et al 2003). Stitziel et al classified geometry location of disease associated nsSNPs using computational geometry method (Stitziel et al 2003).

The evolutionary-based method used evolutionary feature for prediction of deleterious mutations (Ng and Henikoff 2003; del Sol Mesa et al 2003). It is thought that evolutionary forces have selected the allowable set of amino-acid substitutions at a particular site in the protein (Cargill et al 1999). The allowable set of substitutions can be determined by aligning the query sequence with homologous sequences. If a mutated amino acid does not appear in the homology-derived set of substitutions at that site, a substitution by such an amino acid is likely to have a deleterious effect on the protein function. Ng et al developed an evolutionary-based tool, SIFT (Sorts Intolerant from Tolerant), which uses purely sequence homology data (Ng and Henikoff 2003).

Some proteins have similar sequences but substantially different structures and other proteins may have similar structures but divergent sequences. Therefore, the methods based on both sequence and structural information is likely to be complementary and improve predictions. Polyphen is an example of this method (Ramensky et al 2002). The program categorizes nsSNPs as probably damaging, possibly damaging, benign, or unknown. Other researchers have also used both structural and evolutionary information to achieve better prediction power (Saunders et al 2002; Bao et al 2005).

1.2 A Statistical Geometry Approach

A statistical geometry approach using the statistical potential derived from Delaunay tessellations of protein structures has been applied to study the impact of missense mutations. Delaunay tessellation naturally partitions a protein's tertiary structure into an aggregate of space-filling, nonoverlapping, irregular tetrahedras defined as Delaunay simplices whose edges represent all nearest neighbors and thereby reduce the complex three-dimensional interactions in a protein to explicit, elementary tertiary motifs (Singh et al 1996, Tropsha et al 1996, Vaisman et al 1998). Statistical analysis of the residue composition of Delaunay simplices reveals nonrandom preferences for certain quadruplets of amino acids to be clustered together. This nonrandom preference has been used to develop a four-body statistical potential, which is called simplicial neighborhood analysis of protein packing (SNAPP). Change of SNAPP scores has been shown good correlations with experimental change of free energy values of hydrophobic core mutants (Carter et al 2001, Tropsha et al 2003).

A statistical scoring method based on SNAPP scores has been developed to predict the functional effects of missense mutations in the DNA-binding domain (DBD) of the tumor suppressor protein p53, the gene of which is most frequently mutated in human cancer (Mathe et al 2006b). Residual scores (RS), which is calculated in a similar way as SNAAP scores and a residual score profile (RSP) representing the RS values for all residues were used to predict the transactivation activity of p53 mutants with an accuracy varying between 64.2% and 78.5% depending on the promoters (Mathe et al 2006b). Masso et al compared the activity of the experimentally synthesized HIV-1

protease mutants and their corresponding RS and revealed a strong structure-function correlation (Masso et al 2006). Such structure-function correlation has also been identified in two other proteins, bacteriophage T4 lysozyme and HIV-1 reverse transcriptase (Masso et al 2006).

The statistical geometry approach has been applied to a number of other areas of protein studies. These applications include inverted protein folding (Tropsha et al 1996), fold recognition (Zheng et al 1997), discrimination of native and decoy structures (Munson and Singh 1997, Krishnamoorthy and Tropsha 2003), structure similarity comparison (Bostick and Vaisman 2003), structure classification (Bostick et al 2004), protein design (Weberndorfer et al 1999), computational mutagenesis (Masso and Vaisman 2003), and secondary structure assignment (Taylor et al 2005).

1.3 Specific Aims and Dissertation Organization

The general goal of my dissertation are to developed fast and accurate computational mutagenesis models to predict the changes of stability and activity of protein mutants using the four-body statistical potential derived from Delaunay tessellations of protein structures.

The first aim of this dissertation was to correlate the mean residual scores with changes of midpoint temperature of the thermal unfolding (dTm) after single and double point mutations. 1650 single point mutants from 18 proteins and 76 double point mutants in T4 lysozyme from ProTherm database were chosen for the study. The correlation studies were also conducted by incorporating the information of pH, secondary structure, and accessibility of residues. The inferential models based on three machine learning

algorisms, decision tree (DT), random forest (RF), and support vector machine(SVM), were developed to predict mutant thermal stability using RSP in five proteins with sufficient numbers of experimentally synthesized mutants.

The second aim of this dissertation was to correlate the mean residual scores with change of free energy of unfolding (ddG) induced by single and double point mutations. 1856 experimentally synthesized single-point mutants of 17 proteins and 169 double-point mutants of three proteins from ProTherm database were used for the study. Such correlations were also studied in selected proteins by adding the information of pH, secondary structure, and accessibility of residues. Furthermore, the machine learning models using three algorithms, DT, RF, and SVM, and RSP were developed to infer the changes of thermodynamic stability of mutants in five proteins. The performance of thermodynamic stability models was compared to that of thermal stability models.

The third aim was to develop inferential models to predict the activity class of p53 and SERCA1 mutants. RS was compared with transactivation activity of eight different promoters for 932 experimentally synthesized missense mutants, extracted from the Universal Mutation Database (UMD) p53 database, in the DNA-binding domain (DBD) of the tumor suppressor TP53. Next, RS was also compared with the frequency of p53 mutations. The inferential models based on DT, RF, and SVM was developed to predict mutant activity in p53 using residual score profile and the model performance was evaluated. Furthermore, RS was compared with transport activity of 98 experimentally synthesized Sarcoendo plasmic reticulum calcium-ATPases (SERCA1) mutants and the inferential models to predicted SERCA1 transport activity were also derived.

The fourth aim was to investigate relationship between the residual score and conservatism of amino-acid substitution in 700 proteins with high resolution and low similarity. A novel scoring matrix was also developed based on the residual scores of 2,750,782 hypothetical mutations in 700 proteins. The correlation coefficients of this new matrix and PAM40, PAM80, PAM120, PAM250, BLOSUM62, and BLOSUM80 were computed using Simple Mantel test. In addition, RF inferential models based on RS matrix score, actual RS, and RSP were developed and compared in p53 single-point mutants.

This dissertation is composed of four manuscripts intended for publication. Specifically, chapters 2 through 5 reproduce the independent papers with minor formatting adjustment. Thus, some overlap of contents between chapters, especially for the materials and methods section, is inevitable. Two manuscripts comprising Chapters 4 and 5 have been submitted for publication. Chapter 6 describes future directions for the development of proposed methodologies.

2. Predicting Protein Thermal Stability Changes upon Point Mutations Using Machine Learning Methods and Four-Body Statistical Potential

2.1 Abstract

Analysis of protein stability change arising from single residue substitutions provides in depth information on protein structure and function and is also important for the design of new proteins. Residual scores (RS) is the difference of four-body statistical potential between the mutant and its wild-type protein. Residual score profile (RSP) quantifies the environmental change, relative to the wild-type protein, that occurs at every residue position following the point mutation. Here RS were compared with the sign of stability change, measured by changes of midpoint temperature of the thermal unfolding (dTm), for 1650 experimentally synthesized single-point mutants of 18 proteins and 76 double-point mutants of T4 lysozyme from ProTherm database, and a strong correlation is revealed. Furthermore, we observed accessibility of wild type residues has an effect on such correlation. The correlation is mainly driven by the amino-acid substitution in the buried sites of six proteins studied. Finally, stability inferential models based on three machine learning algorithms trained with RSP in five proteins were derived and the predicted accuracies of the models vary between 85% and 96% depending on the proteins and the machine learning algorithms used. These findings are consistent with our hypothesis that computational mutagenesis models based on four-body statistical potential can be used for thermal stability prediction of protein mutants.

2.2 Introduction

Understanding the mechanisms that govern protein stability is one of the long-term goals of protein structure analysis (Daggett and Fersht 2003). The prediction of the protein stability induced by point mutations is also important in protein design (Capriotti et al 2005). Experimental determination of folding free energy change between wild type and mutant protein is costly and time-consuming. Therefore, various computational methods based on different energy functions have been described to predict protein stability changes following single amino acid mutations. These methods can be classified into three major categories (Cheng et al., 2006): (1) physical potential approach; (2) statistical potential approach; (3) empirical potential approach. Physical potential approaches (Prevost et al., 1991; Pitera and Kollman, 2000), directly simulating the atomic force fields present in a given structure, are computationally intensive so that their usage is nearly impossible for applications on a large scale (Guerois et al., 2002). Statistical potential approaches use statistical analysis of the environmental propensities, substitution frequencies, and correlations of contacting residues in solved tertiary structures to derive potential functions (Gilis and Rooman, 1997; Kwasigroch et al., 2002; Carter et al., 2001). The empirical potential approach (Funahashi et al., 2001; Guerois et al., 2002; Zhou and Zhou, 2002a) obtain an energy function by fitting a linear combination of physical energy terms, statistical energy terms, and structural descriptors to the experimental energy data.

Recently, supervised classification techniques from machine learning have been applied to predict the stability of single point mutations (Capriotti et al., 2004; Frenz 2005). Two types of information are required for supervised classification. One is an

ordered data vector containing a set of attributes measured for each mutant (known as an instance in the machine learning language). The other is the experimentally determined stability level of mutants. Supervised classification algorithms make use of both the attribute vectors and the stability levels of the known mutants (the training set) in order to learn a model that can classify the stability levels of the unknown mutants (the testing set) based only on their attribute vectors. Model quality and predictive accuracy may be affected by various factors such as the choice of algorithm, the use of parameters, the size of training set, the cost of misclassification, the measurement accuracy of the experimental stability levels of the mutants.

Computational geometry techniques based on Delaunay tessellation of protein structure have also been used to study protein stability changes and functional effects following single amino acid mutations (Carter et al., 2001; Masso et al., 2006; Mathe et al., 2006b). Carter et al have shown correlations between simplicial neighborhood analysis of protein packing (SNAPP) scores and experimental free energy values of hydrophobic core mutants for five proteins with available experimental data (Carter et al., 2001; Tropsha et al., 2003). SNAPP scores are derived from the compositional likelihood of quadruplets in a protein. Each quadruplet is composed of four nearest-neighbor residues, as defined by the Delaunay tessellation of the tertiary structure of protein. Masso et al observed a significant correlation between the residual score (RS), which is defined as the difference between the mutant and wild-type topological scores, and the activity levels of mutants in human immunodeficiency virus (HIV)-1 protease, T4 lysozyme, and HIV-1 reverse transcriptase (Masso et al., 2006). In addition to the scalar

residual score, the four-body statistical potential can also be used to produce a vector characterization, the residual score profile (RSP), for every protein mutant. RSP is defined as the vector with the length of protein sequence containing the differences of residue potentials between mutant and wild type (wt) at each position. The components of the RSP of a mutant quantitatively measure the relative environmental changes from wt at each of the protein positions induced by the amino acid substitution that created the mutant protein. Mathe et al used RSP to predict the transactivation activity of missense mutations in the DNA-binding domain (DBD) of the tumor suppressor TP53 with an accuracy varying between 64.2% and 78.5% depending on the promoters (Mathe et al., 2006b).

In the current study, we first compare RS with sign of stability change, measured by changes of midpoint temperature of the thermal unfolding (dTm), for 1650 experimentally synthesized single-point mutants in 18 proteins from ProTherm database (Kumar et al 2006). Second, we studied such correlation by including information of secondary structure, pH, and accessibility of residues of mutants. Finally, we develop inferential models using supervised machine learning algorithms and RSP to predict whether a mutation will increase or decrease the thermal stability of protein structure.

2.3 Material and Methods

2.3.1 Delaunay Tessellation and Four-Body Statistical Potential.

Given a set of points, the Voronoi tessellation divides space into regions, called Voronoi cells, centered on these points (Poupon et al 2004). The cell is built by constructing the planes bisecting the lines drawn from the point to each of the other

points and selecting the smallest polyhedron formed by these planes. Voronoi cells fill space and define a tessellation. The related Delaunay tessellation is obtained by tracing vertices between all points that have a common face in their Voronoi cells. Each Voronoi diagram corresponds to one and only one Delaunay tessellation, the two tessellations being the duals of each other.

When Delaunay tessellation is applied to protein structure, each point in this new 3-D space represents one residue. These points may be represented by atomic coordinates of alpha-carbons, beta-carbons, or computed locations, such as the centroid of the atoms in each residue for a given protein in the protein database bank (PDB). For this study, alpha-carbons will be used and each protein structure is represented as a discrete set of points in 3-dimensional space, corresponding to the coordinates of the Cα atoms of the constituent amino acids in the protein. Delaunay tessellation of such a protein structure yields a set of non-overlapping, space-filling, irregular tetrahedra whose vertices are the points representing the constituent amino acids (Singh et al 1996, Vaisman et al 1998). Delaunay tessellations are then constructed using the program Quickhull (Barber et al 1996) that computes the convex-hull (smallest convex set that contains defined points) of the set of residue points.

The four-body statistical potential score is based on the tessellation of a protein, which objectively defines all quadruplets in a given structure that make up four nearest neighbor residues. There are total 8855 different possible types of quadruplets based on the 20 amino acid letter codes (Singh et al 1996, Vaisman et al 1998). A training set of 1417 proteins with known structures was used to calculate the propensities of different

types of quadruplets. Given a training set of proteins, the log-likelihood q for each quadruplet is defined as $q_{ijkl} = \log (f_{ijkl} / p_{ijkl})$. where i, j, k, and l are the four amino acids that compose the quadruplet, f_{ijkl} is the observed normalized frequency of quadruplet, and p_{ijkl} is expected normalized frequency of occurrence of the quadruplet. p_{ijkl} is defined as $p_{ijkl} = ca_ia_ja_ka_l$ where c is a permutation factor that accounts for the permutability of replicated residue types in a given quadruplet and a_r represents the normalized frequency of occurrence of the amino acid r among all of the training set proteins. The potential score for a given residue is then obtained by summing up all the log likelihoods q_{ijkl} for all the quadruplets in which that residue is found. The potential score for the entire protein is the sum of scores of all the quadruplets that define that protein (Singh et al 1996, Vaisman et al 1998). All programs used to call qhull and compute the statistical scores have been written in Java by Zhibin Lu.

2.3.2 Residual Scores and Residual Score Profiles

Each single point mutation causes a change in the log-likelihood scores of all simplices that the point participates in as a vertex. All the identity of the quadruplets that comprise that mutated amino acid is changed and thus their residue potentials are also changed. RS of a protein mutant is thus defined as the difference between the mutant and wt protein topological scores, and measures the relative change in sequence-structure compatibility caused by the amino acid replacement (Masso et al., 2006). RSP of a protein mutant is defined as the vector with the length of protein sequence containing the differences of residue potentials between mutant and wt at each position (Masso et al., 2006). Only the positions of the vector which are structural neighbors of the mutated

residue have non-zero values. All other positions are zero. RS and RSP were calculated for in 18 protein structures with single point mutations and in T4 lysozyme having double-point mutations from Protherm database.

2.3.3 Supervised Machine Learning Algorithms

The freely available machine learning software Weka (Frank et al 2004) was used to predict mutant stability in current study. Random Forest (RF), decision tree (DT), and support vector machine (SVM) algorithms, with default parameters and no normalization of the data, were used. To determine the overall prediction accuracy of a model, a stratified 10-fold cross-validation (10 CV) was applied to both prediction algorithms. The 10 CV splits the data sets into 10 nearly equal groups, and each group in turn was used as the test set while the remaining groups were used for training. The stability change of a mutant is classified as increased (I) and decreased (D), based on the sign of dTm. The stratification ensures that the proportion of I and D mutants in the 10 groups is kept similar in the entire data set. The predictions resulting from the model were directly compared with the class of stability change determined experimentally. From these predictions, the average accuracy score resulting from all test sets, which reflects the number of predictions that match the known stability measurements, was calculated.

The most common method for evaluating binary classification models that require cost weighting is examination of the receiver operating characteristic (ROC) curve (Witten et al 2005). A comparison of the actual and predicted activity classes for each of the mutants based on the outcome of 10 CV provides a simple accuracy measure for the model. Assuming generic class labels P (positive) and N (negative), accuracy = (TP +

TN) / (TP + FN + TN + FP) expressed as a percent true positive rate (TPR) = TP/(TP+FN) = sensitivity, and false positive rate (FPR) = FP/(FP+TN) = 1 – specificity where TP, TN, FP, and FN are true positive, true negative, false positive, and false negative respectively. The coordinates (FPR, TPR) in the unit square define a single point on the ROC curve and thus the ROC curve is a plot of the sensitivity versus the specificity as the parameter rules change (Witten et al 2005). The area under the curve (AUC) is commonly used to as a quantitative measure for quality of a model. The closer the AUC is to 1, the better the model. From these ROC curves, the most appropriate model is the one that yields a high true-positive fraction while maintaining a low false-positive fraction. Therefore, the best model predicts a nearly equal percentage of positive and negative category mutants. AUC values close to one indicate a good discrimination power between positive and negative mutants. A conservative estimate for the standard error (SE) of the AUC was computed using Hanley and McNeil methods (Hanley et al 1982).

2.3.4 Experimental Thermal Stability Dataset

ProTherm (http://gibk26.bse.kyutech.ac.jp/jouhou/propherm/protherm.html) is a collection of published experimental stability data (Kumar et al 2006). These data, which include Gibbs free energy change, enthalpy change, heat capacity change, and transition temperature for wild type and mutant proteins, are important for understanding the structure and stability of proteins. Midpoint transition temperature from folded to unfolded states following thermal denaturation, Tm, is practically used to represent

thermo-stabilization of a protein. Change in Tm, dTm, represents stabilizing or destabilizing mutations from wt proteins (dTm = Tm(mutant) - Tm(wt)).

In current study, ProTherm database was downloaded as of July 30, 2006 and the selection of mutant dataset is based on the following three criteria:

1. The protein mutants have known values for dTm.
2. The number of mutant entries for each protein is at least thirty for single-point mutation or double-point mutation.
3. The structure of wild-type protein has been solved and is tessellable.

Eighteen proteins met the above criteria for single point mutants are selected and one protein for double point mutants met the above criteria is Bacteriophage T4 lysozyme, which has 76 mutant entries. When the same mutation has different conditions for pH, the same mutation will be counted more than one times. If the same mutation has same pH and there is more than one entry, they are considered as duplicate records in ProTherm database and are only counted once. We extract tertiary structure files from the Protein Data Bank (PDB) for all proteins studied according to their PDB codes (Berman et al 2000).

2.4 Results and Discussion

2.4.1 Correlations between Sign of dTm and Average RS in Single-Point Mutants

Changes of midpoint transition temperature of the thermal unfolding is defined as the difference of Tm between wt protein and mutants (dTm = Tm (mutant)- Tm (wt)). A positive dTm means an increased thermal stability of a mutant while a negative dTm means a decreased stability of a mutant. According to the sign of dTm, 1650 mutants are

thus separated into 522 stability-increased mutants and 1128 stability-decreased mutants (Table 2.1). We were interested in understanding the relationship between the sign of stability changes of the mutants and their residual sequence-structure compatibility scores. A mean residual score was calculated by averaging the RS of all the mutants in the given class (positive dTm, negative dTm) in each of 18 proteins.

The results are summarized in Table 2.1 and several facts are observed. First, mean residual scores of positive dTm mutants are larger than that of negative dTm mutants in 17 of 18 proteins. Next, only ten proteins have more than 50 experimentally synthesized mutants available and the mean residual scores of positive dTm mutants are consistently larger than that of negative dTm mutants in each of these ten proteins. Finally, two sample t-tests show a statistically significant was observed in eight out of these ten proteins with p values ranging from 0.05-1.92E-12. The similarity of the results for these ten protein systems, especially given the limitations on number of single point mutants with known stability in each system available for the analysis, suggests that mutant residual scores encapsulate structural information about proteins that can be used to illuminate the strong impact of protein structure on stability. These results demonstrate that a strong correlation exists between the change of mutant stability and the mean residual scores of single-point mutants.

Table 2.1 Correlation of mean residual scores with the sign of dTm in single-point mutants. Mean residual scores of 1650 experimentally synthesized single-point mutants in 18 proteins were computed for positive dTm and negative dTm. The RS of a mutant is defined as the difference between mutant and wild-type topological scores. MRS POS is mean residual scores of positive dTm mutants (dTm>=0). MRS NEG is mean residual scores of negative dTm mutants (dTm<0). P Value is the indicator of statistically significant difference using two-sample t-test for independent samples with unequal variances. SD POS is the standard deviation of the mean residual scores of positive dTm mutants. SD NEG is the standard deviation of the mean residual scores of negative dTm mutants. Count is the total number of mutant entry for each protein.

Source	Protein	PDB ID	MRS POS	MRS NEG	P Value	SD POS	SD NEG	Count
Bacteriophage T4	Lysozyme	2LZM	0.45	-0.7	1.92E-12	1.43	1.94	509
Human	Lysozyme	1LZ1	0.76	-0.18	0.0053	1.8	1.41	157
Escherichia coli	Ribonuclease HI	2RN2	-0.11	-1.79	0.0017	1.51	3.36	140
Saccharomyces cerevisiae	Iso-1 cytochrome c	1YCC	-0.47	-1.25	0.022	1.61	1.3	77
Staphylococcus aureus	Staphylococcal nuclease	1STN	-0.08	-1.13	0.057	1.59	1.46	72
Streptomyces	Subtilisin inhibitor	3SSI	-0.02	-0.79	0.23	0.96	0.96	72
Bovine	Trypsin inhibitor	1BPI	0.08	-0.37	0.45	1.05	0.95	66
Anabaena	Apoflavodoxin	1FTG	0.89	-1.08	6.69E-10	1.27	1.21	64
Bacteriophage P22	Arc repressor	1ARR	0.33	-0.6	0.023	1.24	1.1	52
Streptomyces aureofaciens	Ribonuclease Sa	1RGG	1.65	0.38	0.017	2.12	1.31	51
Bovine	Ribonuclease A	1RTB	-1.31	0.37	0.036	1.63	1.74	44
Chicken	Lysozyme	4LYZ	0.24	-0.02	0.61	0.87	2.32	44
Saccharomyces cerevisiae	Ubiquitin	1OTR	0.59	-1.25	3.69E-06	0.52	1.67	41
Escherichia coli	Maltose-binding protein	3MBP	0.5	-0.82	NA	0.38	0.89	40
Sperm whale	Myoglobin	1BVC	-0.03	0.23	0.38	0.89	1.24	40
Aspergillus oryzae	Ribonuclease T1	1RN1	-0.37	-0.83	0.25	1.85	0.87	38
Bacillus amyloliquefaciens	Barnase	1BNI	-0.19	-0.57	0.75	0	1.48	37
Barley	Chymotrypsin inhibitor	2CI2	-0.09	-0.26	0.56	0.49	1.01	34

Twenty amino acids can be classified according to Dayhoff's substitution matrix, derived from the log odds ratio of the 250 PAM matrix, into six groups: (V,L,I,M), (R,K,H), (D,E,N,Q), (F,Y,W), (C), (A,S,T,G,P) (Dayhoff et al 1978). The amino acids

with common chemical and physical properties tend to fall into the same group. The substitution of amino-acid can be classified as conservative or non-conservative based on such grouping. When the mutant and wild-type amino acids fall within same groups, the substitution is considered conservative. Otherwise, the mutation is classified as non-conservative. We have demonstrated the strong correlation between mean residual scores and dTm in protein single-point mutants. In order to investigate whether or not the strength of correlation is different between conservative (C) and non-conservative (NC) substitutions, we further subdivided the mutants into C and NC substitutions in each stability class and a mean residual score was also calculated for each subgroup. Our data indicate that the strength of correlation between the level of stability and the mean residual score for the NC substitutions in T4 lysozyme is much stronger than that for the C substitutions and the correlation of mean residual score with stability is contributed mainly by the NC substitutions (Figure 2.1). Similar results are observed for human lysozyme and E. coli. Ribonuclease HI (data not shown). The results are consistent with previous results that a strong correlation were observed between the level of activity and the mean residual score for the NC substitutions while weak or no measurable correlation for C substitutions in T4 lysozyme (Masso et al., 2006).

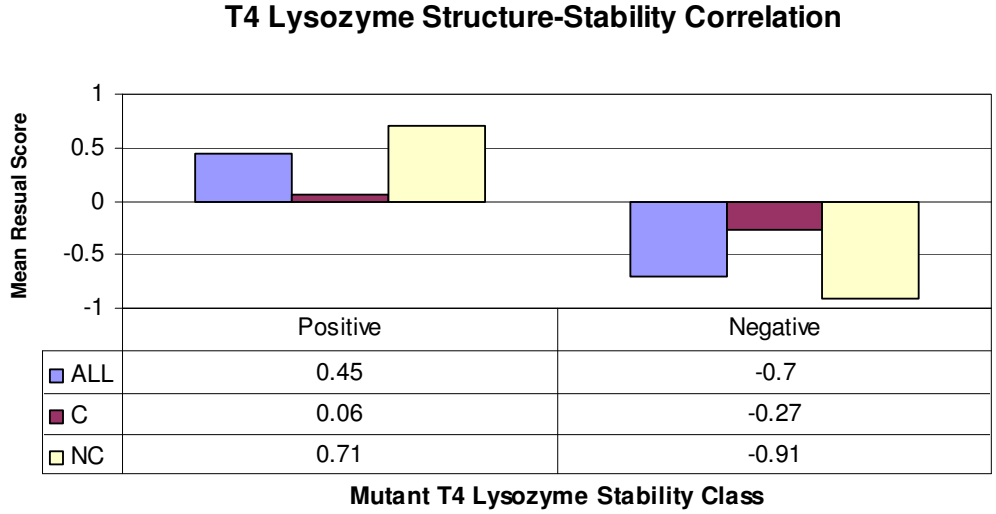


Figure 2.1 T4 lysozyme structure-stability correlation. Comparison of the stability of 509 experimentally synthesized T4 lysozyme mutants with the mean residual of the mutants within each stability class. A significant structure-stability correlation is demonstrated, driven specifically by the NC mutants within the stability classes. The C mutants in each of the stability classes generally have residuals that are small in magnitude due to the minimal change in sequence-structure compatibility from wt. The mean of the residual scores of the C mutants in each stability class remains relatively constant between positive (stability-increased) and negative (stability-decreased) classes.

2.4.2 Correlations between Sign of dTm and Average RS in T4 Lysozyme Double-Point Mutants.

Unlike single-point mutants, there are only a limited number of double-point mutants available in ProTherm database. Our interest is to understand the relationship between the level of stability changes of double-mutants and their residual sequence-structure compatibility scores. 76 double-point experimentally synthesized mutants in T4 lysozyme was used for the study. First, according to the sign of dTm, the mutants were divided into 29 positive and 47 negative mutants. Second, RS were calculated for 76 double-point mutants in T4 lysozyme. Finally, a mean residual score was calculated by

averaging the residual scores of all mutants in the positive class and the negative class. The mean residual score of positive mutants (MRS POS=2.1) are larger than that of negative mutants (MRS NEG= -0.92) and the difference is highly statistically significant with p-value =5.73e-07 using two sample t-tests for independent samples with unequal variances. These results indicate a strong correlation exists between the changes of stability and mean residual scores of the double mutants in T4 lysozyme.

2.4.3 Correlations between Sign of dTm and Average RS by pH.

We have demonstrated a strong correlation exists between mean residual scores and the sign of dTm in T4 lysozyme. The information for one of experimental conditions, pH, is also available in ProTherm database. Our interest is to know whether or not the strength of such correlation is related the range of pH values. Therefore, 507 single-point T4 lysozyme mutants were divided into three groups based on the range of pH values or randomly. Randomized groups serve as control and the control was repeated once (A1-3 and B1-3). The correlation studies were performed in each of the experimental groups and each of control groups and the results are shown in Table 2.2. Our data indicate that no significant difference of correlation is observed from experimental groups and control groups.

Table 2.2 Correlation between mean residual scores and dTm in different pH ranges. 507 single-point T4 lysozyme mutants were divided into three groups based on the range of pH values or randomly (Control A1-3 and Control B1-3). MRS POS is mean residual scores of positive dTm mutants (dTm>=0). MRS NEG is mean residual scores of negative dTm mutants (dTm<0). CNT POS is the count of positive mutations. CNT NEG is the count of negative mutations. P Value is the indicator of statistically significant difference using two-sample t-test for independent samples with unequal variances. Count is the total number of mutant entry.

pH Range	MRS POS	CNT POS	MRS NEG	CNT NEG	P Value	Count
1.99-3.00	0.36	31	-0.62	138	8.30E-04	169
3.00-5.40	0.21	47	-0.67	122	0.0036	169
5.40-7.00	0.65	64	-0.75	105	2.20E-07	169
Conrol A1	0.31	51	-0.51	118	0.00086	169
Conrol A2	0.53	49	-0.75	120	5.70E-05	169
Conrol A3	0.52	42	-0.75	127	3.90E-06	169
Conrol B1	0.45	52	-0.65	117	3.30E-05	169
Conrol B2	0.42	47	-0.63	122	1.80E-04	169
Conrol B3	0.48	43	-0.7	126	8.30E-05	169

2.4.4 Correlations between Sign of dTm and Average RS by Accessibility.

We have demonstrated the strong correlation between mean residual scores and the sign of dTm of single-point mutants. ProTherm database also contains information about accessibility of wild type residues. Our interest is to understand whether or not the strength of such correlation is related to the accessibility of the mutation sites studied.

Accessibility of a wild type residue (%) is defined as the accessible surface area (ASA) of the residue at the mutation site (X) in its parent protein, computed using the program Analytical Surface Calculation (ASC) divided by the ASA of the residue in an extended tripeptide Ala-X-Ala conformation. The extended state ASA was calculated using ECEPP/2 algorithm with dihedral angles given by Oobatake and Ooi (Oobatake and Ooi, 1993) and the van der Waals radius of atoms from Ooi et al. (Ooi et al 1987). The values are Ala-110.2; Asp-144.1; Cys-140.4; Glu-174.7; Phe-200.7; Gly-78.7; His-

181.9; Ile-185.0; Lys-205.7; Leu-183.1; Met-200.1; Asn-146.4; Pro-141.9; Gln-178.6; Arg-229.0; Ser-117.2; Thr-138.7; Val-153.7; Trp-240.5; Tyr-213.7.

ProTherm database classifies the residues with less than 20% accessibility as buried, between 20% and 50% as partially buried and more than 50% as exposed. In this study, we classifies the residues with less than 50% accessibility as buried and more than 50% as exposed due to limited number of mutants available. Six proteins with more than 70 single-point mutants were analyzed. The mutants in each of these six proteins are divided into two groups, buried (B) or exposed (E), based on the cut-off value of 50% accessibility. The correlation studies are performed in the B group or the E group in each of six proteins and the results are shown in Table 2.3. The correlation between mean residual scores and the sign of dTm is observed for the buried group in each of six proteins and such correlations in five of six proteins (except for 1LZ1) are statistically significant. However, the correlation between mean residual scores and the sign of dTm is not evident for the exposed group. The p-values in five of six proteins are consistently larger than 0.05 in the exposed groups. The data have demonstrated that the correlation between mean residual scores and the sign of dTm may be only applicable to the buried mutation sites, but not to the exposed sites. Because residues in the buried site or hydrophobic core play an important role in protein stability, substitution of these residues is more likely to have a profound effect on the local protein geometry. These results are also in agreement with previous findings that the simplicial neighborhood analysis of protein packing (SNAPP) scores correlate with experimental free energy values of hydrophobic core mutants for five proteins [Carter et al., 2001; Tropsha et al., 2003].

Our data also support previous observations that the stability of mutants at residues hydrophobic core was lower than that for mutants at residues in other structural domains of p53 protein [Bullock and Fersht, 2001; Bullock et al., 1997].

Table 2.3 Comparison of correlation between mean residual scores and dTm in the buried group and the exposed group. Mean residual scores in the buried (B) group or the exposed (E) group were computed in each of six proteins with more than 65 single-point mutants. MRS POS is mean residual scores of positive dTm mutants (dTm>=0). MRS NEG is mean residual scores of negative dTm mutants (dTm<0). CNT POS is the count of positive mutations. CNT NEG is the count of negative mutations. P Value is the indicator of statistically significant difference using two-sample t-test for independent samples with unequal variances.

PDB ID	ASA	MRS POS	CNT POS	MRS NEG	CNT NEG	P Value
2LZM	B	0.65	68	-0.91	295	1.56E-09
2LZM	E	0.26	74	0.14	72	0.48
1LZ1	B	0.34	11	-0.58	86	0.19
1LZ1	E	0.91	30	0.95	30	0.84
2RN2	B	0.06	67	-2.12	33	0.0028
2RN2	E	-0.48	30	-0.71	10	0.54
1YCC	B	0.5	25	-1.01	18	0.00026
1YCC	E	-2.48	12	-1.45	22	0.002
1STN	B	0.62	12	-0.99	42	0.012
1STN	E	-1.28	7	-1.68	11	0.42
3SSI	B	-0.81	2	-1.22	24	0.024
3SSI	E	0.04	25	-0.3	21	0.21

2.4.5 Correlations between Sign of dTm and Average RS by Secondary Structure.

ProTherm database also contains information about secondary structure of wild type residues. Our interest is to understand whether or not the strength of the correlation between mean residual scores and the sign of dTm has any relationship with the secondary structure of the mutation sites studied. Secondary structural data for the mutation sites are obtained from PDB in ProTherm database. Three proteins with more

than 140 single-point mutants in the dTm dataset were used for the analysis. The mutants in each of these three proteins are further divided into four subgroups, helix (H), strand (S), turn (T) and coil (C). The mean residual scores were computed in each of subgroups and the results are shown in Table 2.4. The mean residual scores of positive dTm mutants are larger than that of negative dTm mutants in each of four groups in all three proteins (Table 2.4). Since only the helix group has sufficient number of mutants and all other groups has limited number of mutants in each of three proteins studied, the data was unable to conclude whether or not the strength of such correlation in one secondary structure group is statistically higher than that in other secondary structure groups.

Table 2.4 Comparison of correlation between mean residual scores and dTm in different secondary structures. Mean residual scores and dTm in the helix (H) group, strand (S) group, turn (T) group, and coil (C) group were computed in each of three proteins with more than 140 single-point mutants. MRS POS is mean residual scores of positive dTm mutants (dTm>=0). MRS NEG is mean residual scores of negative dTm mutants (dTm<0). CNT POS is the count of positive mutations. CNT NEG is the count of negative mutations. P Value is the indicator of statistically significant difference using two-sample t-test for independent samples with unequal variances.

PDB ID	SECSTR	MRS POS	CNT POS	MRS NEG	CNT NEG	P Value
2LZM	H	0.52	126	-0.78	300	1.45E-12
2LZM	T	0.29	5	0.8	7	0.5849
2LZM	C	0.01	3	-0.37	48	0.5756
2LZM	S	-0.42	8	-1.13	12	0.1592
1LZ1	H	1.67	10	-0.89	18	0.02056
1LZ1	T	0.6	18	-0.36	40	0.02993
1LZ1	C	0.36	10	0.26	27	0.8126
1LZ1	S	0.02	3	0.06	31	0.888
2RN2	H	0.09	44	-2.54	27	0.002846
2RN2	T	0.16	9	-2.68	2	1.63E-07
2RN2	C	-0.82	24	-0.34	7	0.2518
2RN2	S	0.18	20	-0.1	7	0.5904

2.4.6 Predicting Stability Alternation of Single-Point Mutants Using RSP and Machine Learning Schemes

We have established strong correlation between mean residual scores and thermal stability alternations of single-point protein mutants. In addition, the RSP contains significantly more residue position-specific information than the residual scores (Masso et al., 2006; Mathe et al., 2006b). Mathe et al observed that prediction accuracies resulting from machine learning models using RSP is much higher than that using RS (Mathe et al., 2006b). As a result, we decided to investigate how well the information encoded in the RSP is able to distinguish between the mutants of two stability classes: the stability-increased class and the stability-decreased class.

RSP was calculated for all mutants in each of five proteins with greatest number of mutants from ProTherm database. To handle the complexity generated by the high-dimensionality of RSP, machine learning methods were applied to predict stability alternation of protein mutants. In this study, three machine learning algorithms, DT, RF, and SVM, available in the Weka (Witten et al 2005) were used for supervised classification. A ROC analysis was conducted in order to measure the robustness of the predictions (Witten et al 2005). An ROC curve is a plot of sensitivity versus 1 – specificity or True Positive Rate versus False Positive Rate. The AUC is 1.0 for a perfect classifier while the AUC of a random guessing model is 0.5. The RF, SVM, and DT models for five proteins all perform well at distinguishing between mutants of differing stability classes (Table 2.5). The prediction accuracies based DT model range from 85% to 92% with corresponding AUCs from 0.74 to 0.93 depending on the protein systems

analyzed (Table 2.5). The prediction accuracies based RF model range from 89% to 96% with corresponding AUCs from 0.94 to 0.98 depending on the protein systems analyzed (Table 2.5). Figure 2.2 shows the ROC curves obtained by applying 10 CV in conjunction with random forest learning on a training set of the residual profile vectors of T4 lysozyme mutants and Human lysozyme mutants. Figure 2.3 shows the ROC curves obtained by applying 10 CV in conjunction with random forest learning on a training set of the residual profile vectors of E. coli Ribonuclease HI and Anabaena apoflavodoxin mutants. The prediction accuracies based SVM model range from 86% to 95% with corresponding AUCs from 0.80 to 0.94 depending on the protein systems analyzed (Table 2.5). Inferential models based on the RF approach consistently outperform that using the DT and SVM in each of five proteins (Table 2.5). In order to build the control models, the class labels (the sign of dTm) were randomly shuffled among the training vectors. The AUC values for models generated with these controls have a value of about 0.5, which suggest that these control models are all equivalent to random guessing. These data indicates the models can be used to accurately predict the level of stability change of protein mutants.

Table 2.5 Comparison of prediction accuracy and AUC among DT, RF, and SVM. Prediction accuracy and AUC using the RSP in five proteins with largest number of mutants based on dTm were computed using Weka software. The results are based on a 10-fold cross-validation using DT, RF, and SVM with default parameters. The values of the control AUC and control accuracy are obtained from the models generated from the control data set vectors among which the class label (the sign of dTm) was randomly shuffled. SE is the conservative estimate for the standard error of the AUC. CTL AUC is the control AUC.

Protein	Model	AUC	SE	CTL AUC	AC	CTL AC
Bacteriophage T4 Lysozyme	DT	0.92	0.0162	0.5	92%	74%
	RF	0.97	0.0102	0.45	94%	66%
	SVM	0.8	0.024	0.5	86%	74%
Human Lysozyme	DT	0.74	0.0485	0.55	85%	69%
	RF	0.95	0.0243	0.54	89%	69%
	SVM	0.83	0.0418	0.54	87%	66%
E. coli Ribonuclease HI	DT	0.76	0.0406	0.5	85%	69%
	RF	0.97	0.0131	0.56	91%	71%
	SVM	0.89	0.0267	0.53	93%	69%
Anabaena Apoflavodoxin	DT	0.92	0.0333	0.6	86%	60%
	RF	0.94	0.0289	0.56	92%	58%
	SVM	0.94	0.0289	0.55	95%	55%
Iso-1 cytochrome c	DT	0.89	0.0515	0.46	88%	64%
	RF	0.98	0.023	0.46	96%	55%
	SVM	0.92	0.0446	0.48	94%	53%

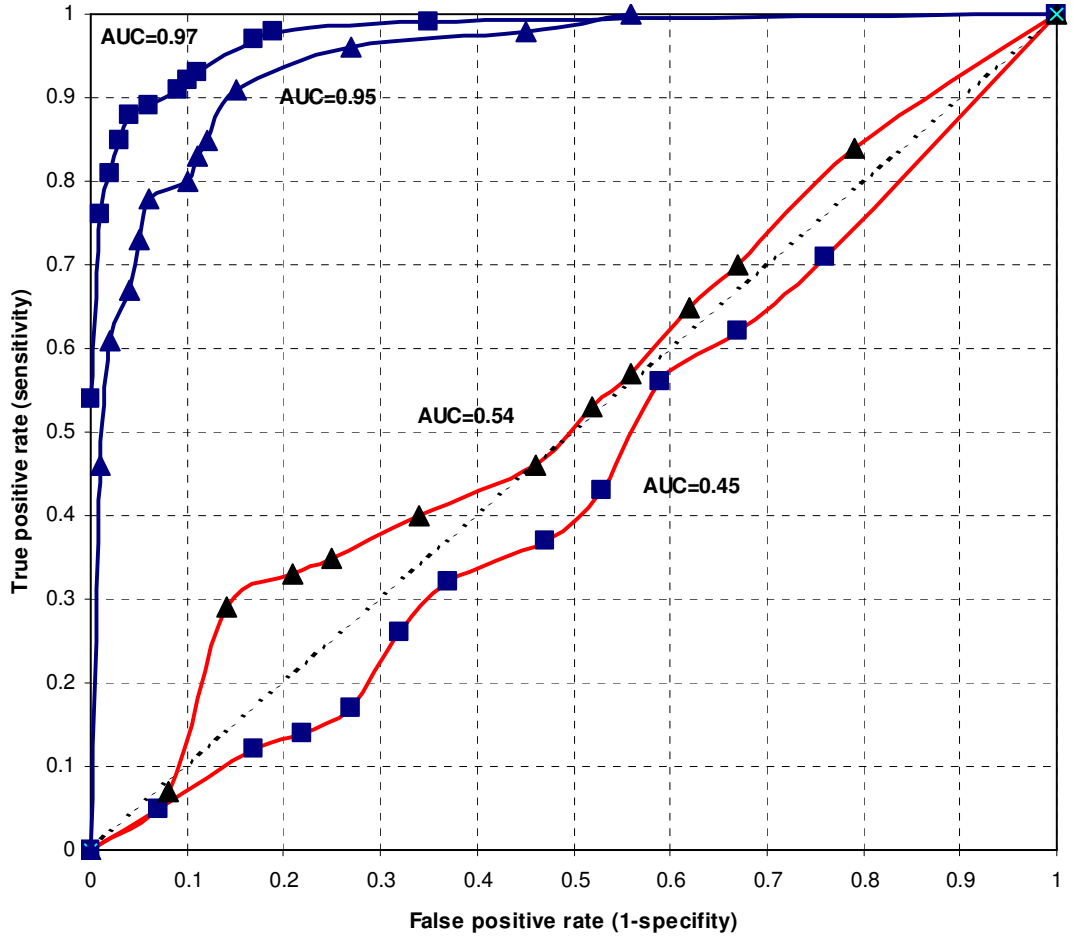


Figure 2.2 T4 lysozyme and Human lysozyme ROC curve. The curve is obtained by applying 10 CV in conjunction with random forest learning on a training set of the residual profile vectors of T4 lysozyme mutants (squares) and Human lysozyme mutants (triangles) that either increase stability or decrease stability by their respective amino acid replacements. Control ROC curves (red curves) obtained by initially performing a random shuffling of the classes (stability-increased class and stability-decreased class) among the mutants prior to training. The control AUC values (0.45 and 0.54) are close to 0.5, reflecting models that perform no better than random guessing.

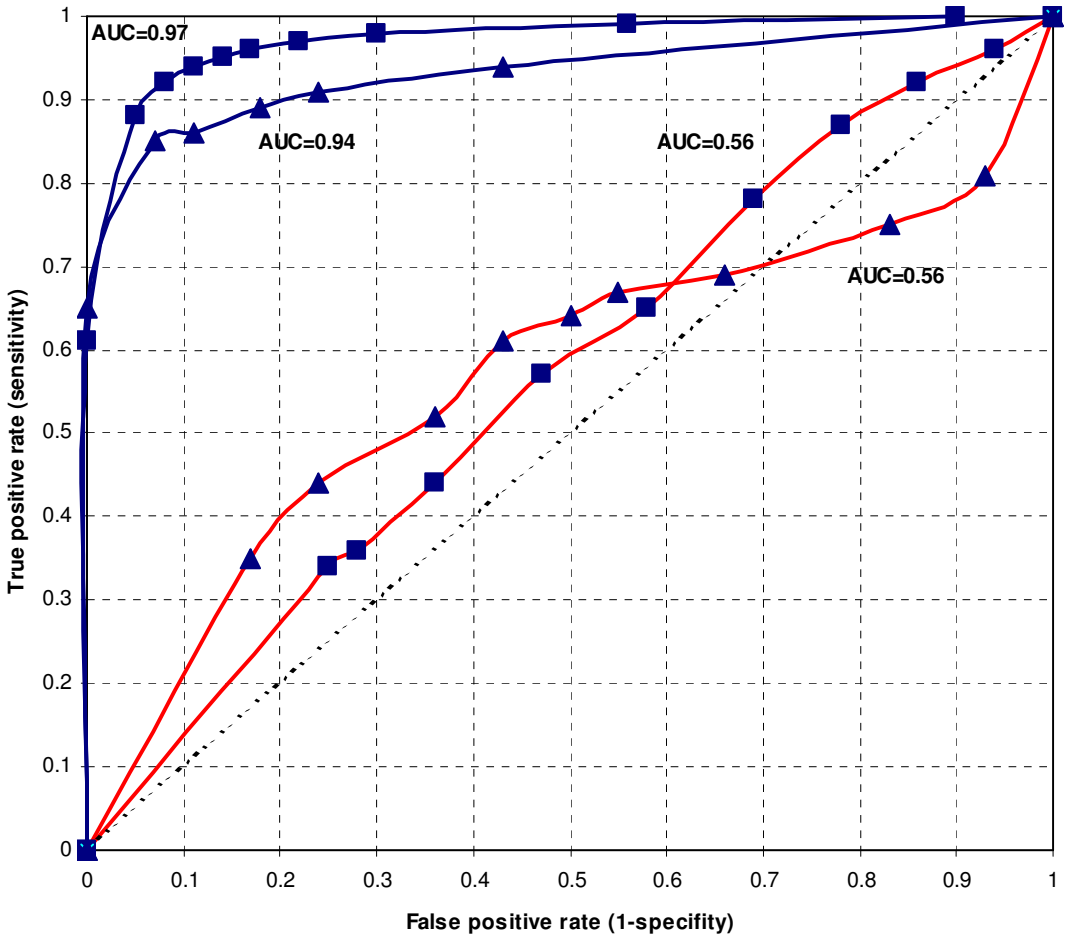


Figure 2.3 E. coli Ribonuclease HI and Anabaena apoflavodoxin ROC curve. The curve is obtained by applying 10 CV in conjunction with random forest learning on a training set of the residual profile vectors of E. coli Ribonuclease HI (squares) and Anabaena apoflavodoxin mutants (triangles) that either increase stability or decrease stability by their respective amino acid replacements. Control ROC curves (red curves) obtained by initially performing a random shuffling of the classes (stability-increased class and stability-decreased class) among the mutants prior to training. The control AUC values (0.56 and 0.56) are close to 0.5, reflecting models that perform no better than random guessing.

2.4.7 Predicting Stability Alternation of T4 Lysozyme Double-Point Mutants Using RSP and Machine Learning Schemes

We have previously demonstrated a strong correlation between stability alternations and mean residual scores of T4 lysozyme double-point mutants. In addition, the RSP contain significantly more residue position-specific information than the scalar residual scores. For the above reasons, we decided to investigate how well the information encoded in the RSP is able to distinguish between the mutants of differing stability classes in T4 lysozyme double-point mutants.

RSP were calculated for 76 double-point mutants of T4 lysozyme and DT, RF, and SVM were used for supervised classification. The AUCs were calculated to test the robustness of the predictions. The prediction accuracies are 83%, 89%, and 81% with corresponding AUC of 0.83, 0.93, and 0.80 for DT, RF, and SVM respectively (Table 2.6). The results demonstrate that inferential models based on the RF approach outperform those based on DT and SVM. These data indicates the models can be used to accurately predict the level of stability change of double mutants in T4 lysozyme.

Table 2.6 Prediction accuracy and AUC of DT and RF models in T4 Lysozyme double-point mutants. Prediction accuracy and AUC were computed using the RSP in 76 double mutants of Bacteriophage T4 lysozyme based on dTm. The results are based on a 10-fold cross-validation using DT, RF, and SVM. SE is the conservative estimate for the standard error of the AUC.

Protein	Model	AUC	SE	Accuracy
Bacteriophage T4 Lysozyme	DT	0.83	0.0518	83%
	RF	0.93	0.0346	89%
	SVM	0.8	0.0554	81%

2.4.8 Random Forest Learning Curves for the T4 Lysozyme Single-Point Mutants.

In order to gain insight into the effect of the training set size on prediction accuracy, we generated training sets of increasing size from the T4 lysozyme training set by sampling with replacement. First, the smallest training set contains 25 mutants were chosen randomly from 509 mutants 20 times and each of these training sets were used for generating models using RF with 10 CV option. The mean accuracy and standard deviation are computed after 20 runs of 10 CV. Next, we increment the number of mutants chosen by 25, and a mean accuracy and standard deviation are reported for 20 runs of 10 CV on the 50 mutants. We continue the procedure by increasing the randomly chosen training set size by 25 mutants at each increment until we reach a training set of size 500. By using these twenty different training sizes, a learning curve was generated and shown in Figure 2.2. The learning curve indicates that it would suffice to use only 275 training set mutants in order to develop an RF model with an accuracy of over 90%.

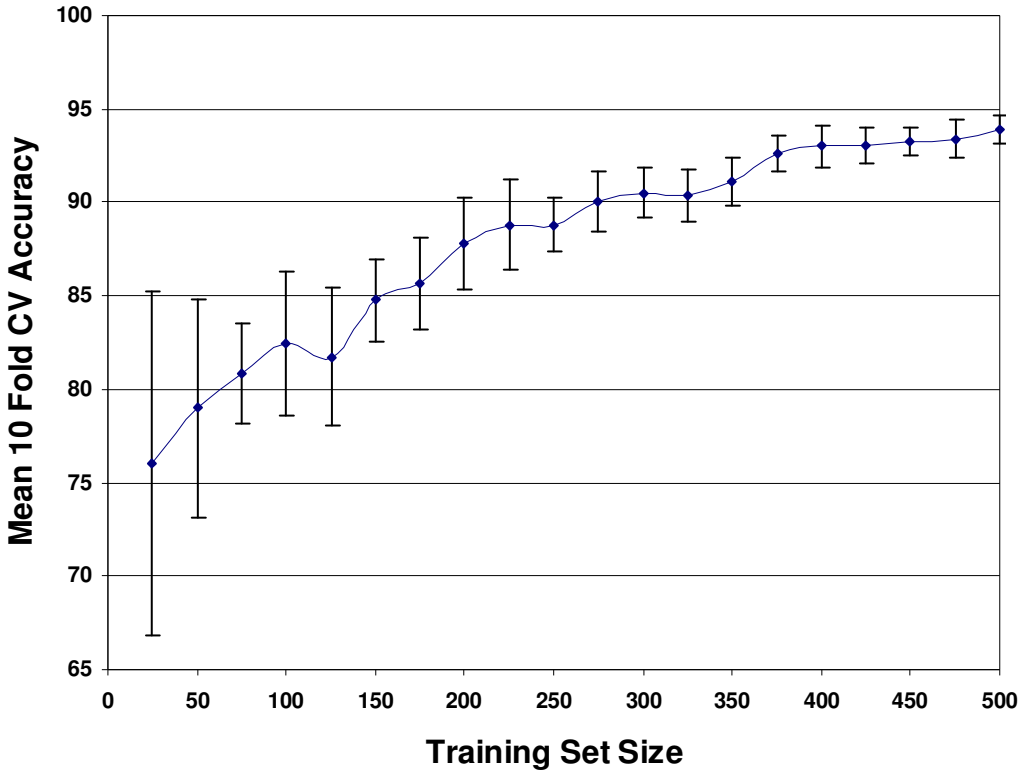


Figure 2.4 Random forest learning curve for the T4 lysozyme mutants. Training sets are randomly chosen with replacement in increments of 25 mutants. For each training set size, the mean 10 CV accuracy is obtained by averaging the accuracy over twenty 10 CV runs using random forest learning. Error bars represent ±1 standard deviation from the mean.

A limitation of our computational mutagenesis approach based on Delaunay tessellation of protein structures is that it requires the available tertiary structure of the protein. However, this method also has several advantages. First, the computation of the RS and RSP can be very fast. Second, unlike most other structure-based methods such as molecular dynamics and homology modeling, our method allows the analysis of all possible missense substitutions for a given protein due to relatively low input

requirements. Finally and most importantly, our best models derived from RF using RSP can achieve a prediction accuracy of over 90% for the stability classes of five proteins for which a certain amount of experimentally mutant stability data are available. We expect that our method is applicable to other proteins for which a sufficient amount of experimental stability data is available to initially train the machine learning models.

2.5 Conclusions

Computational mutagenesis technique based on a four-body statistical potential, generated via Delaunay tessellation of protein structures have been used to analyze the thermal stability alternation following both single-point mutation of 18 proteins and double-point mutation of T4 lysozyme. The computational mutagenesis results in a scalar value and a vector representation of every single amino acid replacement in these proteins. First, a comparison between the measured stability using dTm of the experimentally synthesized single-point mutants and their corresponding residual scores reveals a strong structure-stability correlation in 17 of 18 proteins with a minimum of 30 experimentally synthesized mutants available in ProThem database. Specifically, mutants are divided into two stability classes based on the sign of dTm in each protein. At each stability class, a mean residual score is calculated for the set of mutants in that class. We observe that a decrease in stability level is associated with a drop in mean residual score. Second, this correlation is contributed primarily by non-conservative amino acid substitutions. Third, we demonstrated the correlation is driven mainly by the residue mutations in the buried sites. Within each stability class, the mean residual of the substitutions in exposed sites is minimal in magnitude, and they do not contribute to the

correlation. Fourth, a strong correlation is observed between mean residual score and the stability levels of double-point mutants of T4 lysozyme. Fifth, RSP are used in conjunction with three supervised learning schemes, DT, RF, and SVM, to generate accurate inferential models of single-point mutant stability in five proteins, T4 lysozyme, Human lysozyme, E. coli Ribonuclease HI, Iso-1 cytochrome c, and Staphylococcal nuclease. Finally, the construction of a learning curve reveals that only about 275 single point mutants in T4 lysozyme are necessary as a training set for generating an inferential model with prediction accuracy over 90%.

3. Inferential Models of Mutant Thermodynamic Stability Alternation Using Four-Body Statistical Potential and Machine Learning Methods

3.1 Abstract

The four-body statistical potential can be used to generate a scalar and a vector representation for every point mutant of a protein. The scalar value called residual score (RS) is the difference of four-body statistical potential between the mutant and the wild-type protein. Vector components of a mutant or residual score profile (RSP) quantify the environmental change, relative to the wild-type protein, that occurs at every residue position following the point mutation. In current study, RS are first compared with stability change, measured by the change of thermodynamic stability following point mutations, ddG, for 1856 experimentally synthesized single-point mutants of 17 proteins and 169 double-point mutants of 3 proteins from ProTherm database, and a strong correlation is revealed. Second, we observed such correlation is only applied to the buried sites of seven proteins studied. Third, we developed thermodynamic stability inferential models using three machine learning algorithms and RSP for four protein and the predicted accuracies vary between 87% and 96% depending on the proteins and the machine learning algorithms used. Finally, we compared the performance of thermodynamic stability models with that of thermal stability models and no significant difference of prediction accuracy is observed. These results suggest that inferential

models either based on thermodynamic stability or based on thermal stability can be used for stability prediction of protein mutants.

3.2 Introduction

The stability of a protein structure can be significantly changed following single amino acid mutations. It is very important to accurately predict how single amino acid mutations will affect the stability of a protein structure in protein design (Capriotti et al 2005). Experimental determination of folding free energy change between wild type and mutant protein is costly and time-consuming. As a result, various computational methods based on different energy functions have been described to predict protein stability changes following single amino acid mutations in an attempt to reduce the need for experimental determination of stability effects. These methods can be classified into three major categories (Cheng et al., 2006): (1) physical potential approach; (2) statistical potential approach; (3) empirical potential approach; and (4) supervised machine learning approach. The first three methods all rely on energy functions. Physical potential approaches (Prevost et al., 1991; Pitera and Kollman, 2000), directly simulating the atomic force fields present in a given structure, are computationally intensive so that their usage is nearly impossible for applications on a large scale (Guerois et al., 2002). Statistical potential approaches use statistical analysis of the environmental propensities, substitution frequencies, and correlations of contacting residues in solved tertiary structures to derive potential functions (Gilis and Rooman, 1997; Kwasigroch et al., 2002; Carter et al., 2001). The empirical potential approach (Funahashi et al., 2001; Guerois et al., 2002; Zhou and Zhou, 2002a) obtain an energy function by fitting a linear

combination of physical energy terms, statistical energy terms, and structural descriptors to the experimental energy data. Supervised machine learning approaches make use of both the attribute vectors and the stability levels of the known mutants in order to learn a model that can classify the stability levels of the unknown mutants based only on their attribute vectors (Capriotti et al., 2004; Frenz 2005). The information in the attribute vectors includes physical and chemical features of the amino acids, structural features of the encoded protein, or evolutionary features derived from sequence alignments of homologous proteins.

Computational geometry technique based on Delaunay tessellation of protein structure has also been used to analyze the effects of single amino acid mutations (Carter et al., 2001; Masso et al., 2006; Mathe et al., 2006b). Carter et al have shown that a strong correlation between four-body statistical potential and experimental free energy values of hydrophobic core mutants for five proteins [Carter et al., 2001; Tropsha et al., 2003]. The four-body statistical potential is defined as a function that assigns to each quadruplet a log-likelihood score that compares the observed normalized frequency to an expected rate of occurrence (Singh et al 1996, Vaisman et al 1998). The residual score (RS), which is defined as the difference between the mutant and wild-type topological scores, has been shown to correlate with the activity levels of mutants in human immunodeficiency virus (HIV)-1 protease, T4 lysozyme, and HIV-1 reverse transcriptase (Masso et al., 2006). In addition to the scalar residual score, the four-body statistical potential can also be used to produce a vector characterization, the residual score profile, for every protein mutant. The components of the RSP of a mutant quantitatively measure

the relative environmental changes from wild type (wt) at each of the protein positions induced by the amino acid substitution that created the mutant protein. RSP has been used to predict the transactivation activity of missense mutations in the DNA-binding domain (DBD) of the tumor suppressor TP53 (Mathe et al., 2006b).

The stability of a protein can often be measured in two methods. One approach is to use midpoint transition temperature from folded to unfolded states following thermal denaturation, Tm, which represent thermo-stabilization of a protein. Change in Tm, ΔTm, represents stabilizing or destabilizing mutations from wild-type proteins. The other approach is to use the differences in free energy between the folded and unfolded states of proteins, dG, which represents protein thermodynamic stability. The use of dG to measure protein stability is more general since it can be associated with any kind (thermal, chemical, pH, etc) of destabilization process. It should be noted that thermodynamic stability is not the only problem at temperatures more than 80 since high temperatures may result in the irreversible inactivation of proteins due to possible chemical damages. Change in dG, ddG, can be used to measure increased or decreased stability of mutants from wild-type proteins.

In many cases, the correct prediction of the direction or sign of the stability change is more relevant than its magnitude (Capriotti et al., 2004). As a result, in the current study we first compare RS with sign of stability change, measured by differences in free energy between the folded and unfolded states of proteins (ddG), for 1856 experimentally synthesized single-point mutants in 17 proteins and 169 double-point mutants in 3 protein from ProTherm database (Kumar et al 2006). Second, we further

conduction such correlation studies by incorporating information of pH, accessibility of residues, and the secondary structure of mutants. Third, we develop inferential models using supervised machine learning algorithms and RSP to predict the sign of the relative thermodynamic stability change, ddG, to determine whether a mutation will increase or decrease the stability of protein structure. Finally, we compared the performance between thermodynamic stability models and thermal stability models.

3.3 Material and Methods

3.3.1 Experimental Thermodynamic Stability Dataset

ProTherm database was downloaded as of Feb 4, 2007 (Kumar et al 2006). A dataset was generated to meet three criteria: 1) The protein mutants have known values for ddG. 2) The number of mutant entries for each protein is at least thirty for single-point mutation or double-point mutation. 3) The tertiary structure of wild-type protein has been solved and is also tessellable. Each mutation in the dataset has six attributes: PDB codes, mutation, solvent accessibility, pH value, secondary structure, and energy change (ddG). The dataset includes 1856 single site mutations obtained from 17 different proteins and 169 double site mutations from 3 different proteins. When the same mutation has different conditions for pH, the same mutation will be counted more than one times. If the same mutation has same pH and there is more than one record, they are considered as duplicate records in ProTherm database and are only counted once. We extract tertiary structure files from the Protein Data Bank (PDB) for all proteins studied according to their PDB codes (Berman et al 2000).

3.3.2 Delaunay Tessellation and Four-Body Statistical Potential.

A non-homologous training set of 1417 high-resolution crystallographic protein structures is selected from the PDB (Berman et al 2000). Utilizing atomic coordinates of alpha-carbons, each structure is represented as a discrete set of points in 3-dimensional space, corresponding to alpha-carbon (Cα) coordinates of the constituent amino acid residues in the protein. A geometrical construct known as Delaunay tessellation uses Cα coordinates to create a partitioning of the space occupied by a protein structure. Delaunay tessellation of each protein structure yields an aggregate of non-overlapping, space-filling, irregular tetrahedra, referred to as Delaunay simplices, whose vertices are the amino acid point representations (Singh et al 1996, Vaisman et al 1998). Each Delaunay simplex are constructed using the program Quickhull (Barber et al 1996) that computes the convex-hull (smallest convex set that contains defined points) of the set of residue points. Assuming order independence, there are total 8855 different possible types of quadruplets based on the 20 amino acid letter codes (Singh et al 1996, Vaisman et al 1998). The four-body statistical potential is defined as a function that assigns to each quadruplet a log-likelihood score that compares the observed normalized frequency to an expected rate of occurrence (Singh et al 1996, Vaisman et al 1998). All programs used to call qhull and compute the statistical scores have been written in Java by Zhibin Lu.

3.3.3 Residual Scores and Residual Score Profiles

Individual residue potential of each constituent amino acid position in a protein is sum of log-likelihood scores of only those simplices in the Delaunay tessellation for which the point representing the residue position participates as a vertex (Masso et al., 2003). Potential profile vector of a protein is a vector of individual residue potentials for

all residues in the protein V = (Paa1, Paa2, … , Paan) where n is the position of residues in a given protein. The residual score of a mutant is the difference between the mutant and wt protein total potential (RS = total potential (mutant) – total potential (wild-type)). The residual profile vector or residual score profile of a mutant is the difference between the mutant and wt protein potential profile, and the value of each component is an environmental change (EC) score (Masso et al., 2006). The EC scores explicitly measure the change in the local environment at every residue position relative to wt due to the specific mutations, and implicitly indicate the topological connections between the mutated residue positions and their nearest-neighbors via the presence of non-zero EC scores as components in the residual profile vector. RS and RSP were calculated for in 17 protein structures with single point mutations and in 3 proteins having double-point mutations from ProTherm database (Kumar et al 2006).

3.3.4 Supervised Machine Learning Methods

The Weka suite of machine learning tools is used to generate and evaluate the performance of thermodynamic stability inference models, with a specific focus on supervised classification implementations of DT, RF, and SVM (Frank et al 2004, Witten et al 2005). To determine the overall prediction accuracy of a model, a stratified 10-fold cross-validation was applied to all three prediction algorithms. The 10-fold cross-validation splits the data sets into 10 nearly equal groups, and each group in turn was used as the test set while the remaining groups were used for training. The stability change of a mutant is classified as positive and negative, based on the sign of ddG. If the energy change dd*G* is positive, the mutation increases stability and is classified as a

positive example. If dd*G* is negative, the mutation is destabilizing and is classified as a negative example. The stratification ensures that the proportion of positive and negative mutants in the 10 groups is kept similar in the entire data set. The predictions resulting from the model were directly compared with the class of stability change determined experimentally. From these predictions, the average accuracy score resulting from all test sets, which reflects the number of predictions that match the known stability measurements, was calculated.

An approach based on the receiver operating characteristic (ROC) curve was used to evaluate model performance. The ROC curve is a plot of the sensitivity (true positive rate – TPR) versus 1 – specificity (false positive rate – FPR) in the unit square. ROC curves are a common method for comparing the tradeoff between the models sensitivity and specificity. The area under the curve (AUC) is commonly used to as a quantitative measure for quality of a model. A ROC curve with AUC = 0.5 represents a model of random guessing between two classes while a ROC curve with AUC = 1.0 represents a perfect model. A conservative estimate for the standard error (SE) of the AUC was computed using Hanley and McNeil methods (Hanley et al 1982).

3.4 Results and Discussion

3.4.1 Correlations between Sign of ddG and Average RS in Single-Point Mutants

The difference in free energy between the folded and unfolded states of proteins (dG) measures protein thermodynamic stability. Change in dG, ddG, can be used to measure increased or decreased stability of mutants from wild-type proteins. The thermodynamic stability change of a mutant is classified as either positive or negative,

according to the sign of ddG. A positive ddG means an increased stability of a mutant while a negative ddG represents a decreased stability of a mutant. Based on the sign of ddG, 1856 mutants are thus separated into 518 stability-increased mutants and 1338 stability-decreased mutants (Table 3.1). We were interested in understanding the relationship between the sign of ddG of the mutants and their residual sequence-structure compatibility scores. A mean residual score was calculated by averaging the residuals of all the mutants in each of given class (positive ddG and negative ddG) in 17 proteins.

The results are summarized in Table 3.1. The results show that mean residual scores of positive ddG mutants are larger than that of negative ddG mutants in 16 of 17 proteins. The only exception is Aspergillus oryzae Ribonuclease T1, in which the mean residual score (-0.55) of positive ddG mutants is less than that (-0.25) of negative ddG mutants. However, the difference in Aspergillus oryzae Ribonuclease T1 is not statistically significant. Furthermore, two sample t-tests for independent samples with unequal variances show that there is a highly statistically significant difference between the mean residual scores of the positive and negative classes in 9 of 16 proteins. The similarity of the results for several protein systems, especially given the limitations on number of single point mutants with known stability in each system available for the analysis, suggests that mutant residual scores encapsulate structural information about proteins that can be used to predict the strong impact of protein structure on thermodynamic stability. These results demonstrate that a strong correlation exists between the change of mutant thermodynamic stability and the mean residual scores of the mutants.

Table 3.1 Correlation of mean residual scores with the sign of ddG in single-point mutants. Mean residual scores of 1856 experimentally synthesized single-point mutants in 17 proteins were computed for positive ddG and negative ddG. The residual score of a mutant is defined as the difference between mutant and wild-type topological scores. MRS POS is mean residual scores of positive ddG mutants (ddG>=0). MRS NEG is mean residual scores of negative ddG mutants (ddG<0). P Value is the indicator of statistically significant difference using two-sample t-test for independent samples with unequal variances. SD POS is the standard deviation of the mean residual scores of positive ddG mutants. SD NEG is the standard deviation of the mean residual scores of negative ddG mutants. Count is the total number of mutant entry for each protein.

Source	Protein	PDB ID	MRS POS	MRS NEG	P Value	SD POS	SD NEG	Count
Bacteriophage T4	Lysozyme	2LZM	0.43	-0.89	2.34E-14	1.43	1.93	443
Bacillus amyloliquefaciens	Barnase	1BNI	0.19	-0.49	0.0026	0.9	1.06	236
Human	Lysozyme	1LZ1	0.73	-0.15	0.013	1.75	1.51	145
Escherichia coli	Ribonuclease HI	2RN2	-0.06	-2.09	0.0012	1.54	3.57	122
Barley	Chymotrypsin inhibitor	2CI2	0	-0.31	0.054	0.48	1.03	117
Bacteriophage f1	Gene V	1VQB	-0.04	-0.69	0.12	1.45	2.13	114
Anabaena	Apoflavodoxin	1FTG	0.32	-0.21	0.047	1.39	1.58	92
Sperm whale	Myoglobin	1BVC	-0.12	-0.55	0.11	0.72	1.69	84
Aspergillus oryzae	Ribonuclease T1	1RN1	-0.55	-0.25	0.63	2.24	1.09	67
Streptomyces aureofaciens	Ribonuclease Sa	1RGG	1.91	0.53	0.0028	2.17	1.21	60
Chicken	Lysozyme	4LYZ	0.17	-0.21	0.41	0.86	2.29	51
Bovine	Trypsin inhibitor	1BPI	0.28	-0.42	0.16	1.03	0.93	51
Streptomyces	Subtilisin inhibitor	3SSI	-0.87	-1.15	0.21	0.47	0.72	50
Bacillus caldolyticus	Cold shock protein	1C9O	0.39	-0.24	0.12	1.33	1.39	50
Staphylococcus aureus	Staphylococcal nuclease	1STN	-0.13	-0.78	0.21	1.23	1.27	49
Saccharomyces cerevisiae	Iso-1 cytochrome c	1YCC	0.47	-0.33	0.046	1.01	0.82	40
Streptococcus	Protein G	1PGA	-0.02	-1.56	0.011	1.22	1.01	34

According to Dayhoff's substitution matrix (Dayhoff et al 1978), twenty amino acids can be classified into six groups: (V,L,I,M), (R,K,H), (D,E,N,Q), (F,Y,W), (C),

(A,S,T,G,P). The amino acids with common chemical and physical properties tend to fall into the same group. When the mutant and wild-type amino acids fall within same groups, the substitution is considered conservative. Otherwise, it is non-conservative. We have demonstrated the strong correlation between mean residual scores and the sign of ddG in protein single-point mutants. In order to investigate whether or not the strength of correlation is different between conservative (C) and non-conservative (NC) substitutions, we further subdivided the mutants into C and NC substitutions in each stability class and a mean residual score was also calculated for each subgroup. The results demonstrated that the strength of correlation between the level of thermodynamic stability and the mean residual score for the NC substitutions in T4 lysozyme is much stronger than that for the C substitutions (Figure 3.1). The result indicates the correlation of mean residual score with stability is contributed mainly by the NC substitutions. Similar results are observed in two other proteins, Bacillus amyloliquefaciens Barnase and human lysozyme (data not shown). These data are consistent with previous results that a strong correlation were observed between the level of activity and the mean residual score for the NC substitutions while weak or no measurable correlation for C substitutions in T4 lysozyme (Masso et al., 2006).

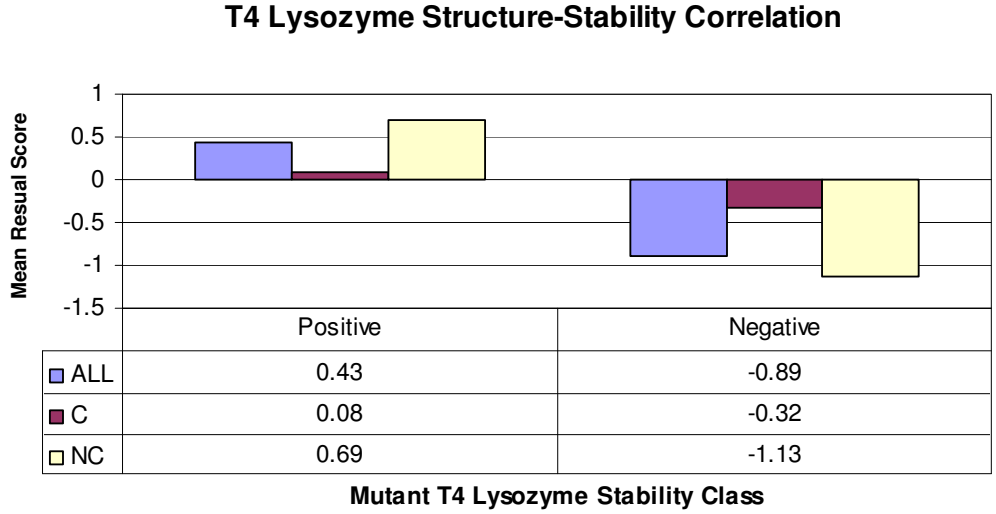


Figure 3.1 T4 lysozyme structure-stability correlation. Comparison of the stability of 443 experimentally synthesized T4 lysozyme mutants with the mean residual of the mutants within each stability class using ddG. A significant structure-stability correlation is demonstrated, driven specifically by the NC mutants within the stability classes. The C mutants in each of the stability classes generally have residuals that are small in magnitude due to the minimal change in sequence-structure compatibility from wt. The mean of the residual scores of the C mutants in each stability class remains relatively constant between positive (stability-increased) and negative (stability-decreased) classes.

3.4.2 Correlations between Sign of ddG and Average RS in T4 Lysozyme Double-Point Mutants.

Unlike single-point mutants, there are only a limited number of double-point mutants available in ProTherm database. Our interest is to understand the relationship between the level of thermodynamic stability changes of the T4 lysozyme double-mutants and their residual sequence-structure compatibility scores. Three protein systems, T4 lysozyme, Bacteriophage f1 Gene V, and Streptococcus Protein G, which have the highest number of double-point mutants in ProTherm database, are used for

correlation study. The results demonstrated that the mean residual scores of positive ddG mutants are consistently larger than that of negative ddG mutants in all three proteins (Table 3.2). Furthermore, two sample t-tests for independent samples with unequal variances show that there is a highly statistically significant difference between the mean residual scores of the positive and negative classes in all three proteins studied (Table 3.2). These results indicate a strong correlation exists between the changes of thermodynamic stability and mean residual scores of the double mutants in these three proteins.

Table 3.2 Correlation of mean residual scores with the sign of ddG in double-point mutants. Mean residual scores of 169 experimentally synthesized double-point mutants in 3 proteins were computed for positive ddG and negative ddG. The residual score of a mutant is defined as the difference between mutant and wild-type topological scores. MRS POS is mean residual scores of positive ddG mutants (ddG>=0). MRS NEG is mean residual scores of negative ddG mutants (ddG<0). P Value is the indicator of statistically significant difference using two-sample t-test for independent samples with unequal variances. SD POS is the standard deviation of the mean residual scores of positive ddG mutants. SD NEG is the standard deviation of the mean residual scores of negative ddG mutants. Count is the total number of mutant entry for each protein.

Source	Protein	PDB ID	MRS POS	MRS NEG	P Value	Count
Bacteriophage T4	Lysozyme	2LZM	2.05	-0.57	1.64E-05	74
Bacteriophage f1	Gene V	1VQB	0.31	-1.09	0.0016	49
Streptococcus	Protein G	1PGA	-0.25	-3.03	2.20E-16	46

3.4.3 Correlations between Sign of ddG and Average RS by pH.

A strong correlation has been established between mean residual scores and ddG in T4 lysozyme. The information for one of experimental conditions, pH, is also available in ProTherm database. Our interest is to know whether or not the strength of

such correlation is related the range of pH values. Therefore, 443 single-point T4 lysozyme mutants were divided into three groups based on the range of pH values or randomly. Randomized groups serve as control and the control was repeated once (A1-3 and B1-3). The correlation studies were performed in each of the experimental groups and each of control groups. Our results indicate that no significant difference of correlation strength is observed between experimental groups and control groups (Table 3.3).

Table 3.3 Correlation between mean residual scores and ddG in different pH ranges. 443 single-point T4 lysozyme mutants were divided into three groups based on the range of pH values or randomly (Control A1-3 and Control B1-3). MRS POS is mean residual scores of positive ddG mutants (ddG>=0). MRS NEG is mean residual scores of negative ddG mutants (ddG<0). CNT POS is the count of positive mutations. CNT NEG is the count of negative mutations. P Value is the indicator of statistically significant difference using two-sample t-test for independent samples with unequal variances. Count is the total number of mutant entry.

pH Range	MRS POS	CNT POS	MRS NEG	CNT NEG	P Value	Total
1.99-3.00	0.69	35	-0.8	112	1.20E-04	147
3.00-5.40	0.63	50	-0.94	97	0.00014	147
5.40-6.90	0.6	51	-0.85	96	7.33E-06	147
Conrol A1	0.58	45	-1.17	102	7.23E-06	147
Conrol A2	0.4	45	-0.9	102	1.36E-04	147
Conrol A3	0.94	46	-0.6	101	6.80E-05	147
Conrol B1	0.84	43	-0.83	104	2.54E-06	147
Conrol B2	0.53	47	-0.74	100	5.90E-04	147
Conrol B3	-0.55	46	-1.1	101	2.26E-05	147

3.4.4 Correlations between Sign of ddG and Average RS by Accessibility.

A strong correlation has been established between mean residual scores and ddG in protein single-point mutants. ProTherm database also contains information about accessibility of wild type residues. Our interest is to understand whether or not the strength of such correlation is related to the accessibility of the mutation sites studied.

ProTherm database classifies the residues with less than 20% accessibility as buried, between 20% and 50% as partially buried and more than 50% as exposed. In this study, we classifies the residues with less than 50% accessibility as buried and more than 50% as exposed due to limited number mutants available. We used seven proteins with more than 90 single-point mutants for this study. The mutants in each of these seven proteins are divided into two groups, buried (B) or exposed (E), based on the cut-off value of 50% accessibility. The correlation studies are performed in the B group or the E group in each of seven proteins. The correlation between mean residual scores and the sign of ddG is observed for the buried group in each of seven proteins and such correlations in six of seven proteins (except for human lysozyme) are statistically significant with p value <0.05 (Table 3.4). However, the correlation between mean residual scores and the sign of ddG is not observed for the exposed group. The p values in all seven proteins are consistently more than 0.05 in the exposed groups using two-sample t-test for independent samples with unequal variances. The data have demonstrated that the correlation between mean residual scores and the sign of ddG may be only applicable to the buried mutation sites, but not to the exposed sites. Since residues in the buried site or hydrophobic core play an important role in protein stability,

replacement of these residues is more likely to have a profound effect on local protein geometry. Our data are in agreement with previous findings that the four-body statistical potential scores correlate with experimental free energy values of hydrophobic core mutants [Carter et al., 2001; Tropsha et al., 2003].

Table 3.4 Comparison of correlation between mean residual scores and ddG in the buried group and the exposed group. Mean residual scores in the buried (B) group or the exposed (E) group were computed in each of seven proteins with more than 65 single-point mutants. MRS POS is mean residual scores of positive ddG mutants (ddG>=0). MRS NEG is mean residual scores of negative ddG mutants (ddG<0). CNT POS is the count of positive mutations. CNT NEG is the count of negative mutations. P Value is the indicator of statistically significant difference using two-sample t-test for independent samples with unequal variances.

PDB	ASA	MRS POS	CNT POS	MRS NEG	CNT NEG	P Value
2LZM	B	0.68	60	-1.17	246	1.367E-10
2LZM	E	0.24	77	0.29	60	0.7938
1BNI	B	0.24	6	-0.71	138	0.03398
1BNI	E	0.17	16	-0.1	76	0.3053
1LZ1	B	-0.07	10	-0.59	77	0.3976
1LZ1	E	1.02	27	0.97	31	0.9453
2RN2	B	0.08	64	-2.53	27	0.002666
2RN2	E	-0.47	22	-0.77	9	0.4899
2CI2	B	0.23	10	-0.29	75	0.003612
2CI2	E	-0.4	6	-0.37	26	0.8851
1VQB	B	0.09	15	-0.78	78	0.06041
1VQB	E	-0.68	3	-0.3	18	0.7579
1FTG	B	0.19	17	-0.84	41	0.02127
1FTG	E	0.5	13	1.02	21	0.3262

3.4.5 Correlations between Sign of ddG and Average RS by Secondary Structure.

The information about secondary structure of wild type residues is also available in ProTherm database. Our interest is to understand whether or not the strength of such correlation has any relationship with the secondary structure of the mutation sites studied. Secondary structural data for the mutation sites are extracted from ProTherm database.

Three proteins have more than 140 single-point mutants in the ddG dataset. The mutants in each of these three proteins are divided into four subgroups, helix (H), strand (S), turn (T) and coil (C). The correlation studies are performed in all four subgroups in each of three proteins. Our result indicates that the mean residual scores of positive ddG mutants are larger than that of negative ddG mutants in the helix group of all three proteins (Table 3.5). Since only the helix group has sufficient number of mutants and all other groups has limited number of mutants in each of three proteins studied, the data was unable to conclude whether or not the strength of such correlation in one secondary structure group is higher than that in other secondary structure groups.

Table 3.5 Comparison of correlation between mean residual scores and ddG in different secondary structures. Mean residual scores and ddG in the helix (H) group, strand (S) group, turn (T) group, and coil (C) group were computed in each of three proteins with more than 140 single-point mutants. MRS POS is mean residual scores of positive ddG mutants (ddG>=0). MRS NEG is mean residual scores of negative ddG mutants (ddG<0). CNT POS is the count of positive mutations. CNT NEG is the count of negative mutations. P Value is the indicator of statistically significant difference using two-sample t-test for independent samples with unequal variances.

PDB	SECSTR	MRS POS	CNT POS	MRS NEG	CNT NEG	P Value
2LZM	H	0.49	122	-1.03	229	5.55E-14
2LZM	T	0.69	4	0.64	7	0.96
2LZM	C	-0.21	5	-0.49	60	0.47
2LZM	S	-0.4	6	-1.18	10	0.21
1BNI	H	0.66	12	-0.21	73	0.000014
1BNI	T	-0.32	3	-0.97	5	0.21
1BNI	C	-0.55	6	-0.31	74	0.66
1BNI	S	0.48	1	-1	62	N/A
1LZ1	H	2.29	8	-0.88	17	0.012
1LZ1	T	0.46	18	-0.51	35	0.035
1LZ1	C	0.05	8	0.29	26	0.52
1LZ1	S	0.02	3	0.32	30	0.25

3.4.6 Predicting Stability Alternation of Single-Point Mutants Using RSP and Machine Learning Schemes

A strong correlation has been established between stability alternations and mean residual scores of protein single-point mutants. In addition, the RSP contains significantly more residue position-specific information than the scalar residual scores (Masso et al., 2006; Mathe et al., 2006b). Mathe et al has observed that prediction accuracies resulting from machine learning models using RSP is much higher than that using RS (Mathe et al., 2006b). As a result, we decided to investigate how well the information encoded in the RSP is able to distinguish between the mutants of two thermodynamic stability classes according to the sign of ddG.

RSP was calculated for all mutants in each of four proteins with greatest number of mutants based on ddG. In this study, three machine learning algorithms, DT, RF, and SVM, available in the Weka (Witten et al 2005) were used for supervised classification. A test option known as stratified tenfold cross-validation (10 Fold CV) was used for analysis.

A ROC analysis was conducted in order to measure the robustness of the predictions (Witten et al 2005) for each protein. The RF, SVM, and DT models for four proteins all perform well at distinguishing between mutants of differing stability classes (Table 3.6). The prediction accuracies based DT model range from 87% to 94% with corresponding AUCs from 0.62 to 0.89 depending on the protein systems analyzed (Table 3.6). The prediction accuracies based RF model range from 91% to 93% with corresponding AUCs from 0.76 to 0.99 depending on the protein systems analyzed (Table

3.6). The prediction accuracies based SVM model range from 87% to 94% with corresponding AUCs from 0.63 to 0.92 depending on the protein systems analyzed (Table 3.6). Inferential models based on the RF approach consistently outperform that using the DT and SVM in each of four proteins (Table 3.6). In order to build the control models, the class labels (the sign of ddG) were randomly shuffled among the data set vectors. The AUC values for models generated with these controls have an average value of about 0.5, which suggest that these control models are all equivalent to random guessing. These data indicates the models can be used to accurately predict the level of thermodynamic stability change of protein mutants. These models may be used to infer stability classes of unknown protein mutants.

Table 3.6 Comparison of prediction accuracy and AUC among DT, RF, and SVM. Prediction accuracy and AUC using the residual score profile in four proteins with largest number of mutants based on ddG were computed using Weka software. The results are based on a 10-fold cross-validation using DT, RF, and SVM with default parameters. The values of the control AUC and control accuracy are obtained from the models generated from the control data set vectors among which the class label (the sign of ddG) was randomly shuffled. SE is the conservative estimate for the standard error of the AUC. CTL AUC is the control AUC.

Protein	Model	AUC	SE	CTL AUC	Accuracy
Bacteriophage T4 Lysozyme	DT	0.85	0.022	0.48	88%
	RF	0.96	0.0119	0.44	91%
	SVM	0.82	0.0237	0.49	87%
Bacillus amyloliquefaciens Barnase	DT	0.62	0.0665	0.49	94%
	RF	0.76	0.0613	0.5	92%
	SVM	0.63	0.0664	0.49	90%
Human Lysozyme	DT	0.89	0.0367	0.42	90%
	RF	0.95	0.0255	0.4	92%
	SVM	0.88	0.0381	0.46	90%
E. coli Ribonuclease HI	DT	0.78	0.0419	0.54	87%
	RF	0.99	0.0078	0.49	93%
	SVM	0.92	0.0239	0.55	94%

3.4.7 Predicting Stability Alternation of T4 Lysozyme Double-Point Mutants Using RSP and Machine Learning Schemes

A strong correlation has been established between thermodynamic stability alternations and mean residual scores of T4 lysozyme double-point mutants. In addition, the RSP contain significantly more residue position-specific information than the scalar residual scores. As a result, we decided to investigate how well the information encoded in the residual profiles is able to differ between the mutants of differing thermodynamic stability classes in T4 lysozyme double-point mutants.

RSP were calculated for 74 double-point mutants of T4 lysozyme and DT, RF, and SVM algorithms were used for supervised classification. The AUCs were calculated to test the robustness of the predictions. The prediction accuracies are 85%, 91%, and 82% with corresponding AUC of 0.82, 0.94, and 0.82 for DT, RF, and SVM respectively (Table 3.7). The results demonstrate that inferential models based on the RF approach outperform those based on DT and SVM. These data indicates the models can be used to accurately predict the level of thermodynamic stability change of protein double mutants in T4 lysozyme.

Table 3.7 Prediction accuracy and AUC of DT and RF models in T4 Lysozyme double-point mutants. Prediction accuracy and AUC were computed using the residual score profile in 74 double mutants of Bacteriophage T4 lysozyme based on ddG. The results are based on a 10-fold cross-validation using DT, RF, and SVM. SE is the conservative estimate for the standard error of the AUC.

Protein	Model	AUC	SE	Accuracy
Bacteriophage T4 Lysozyme	DT	0.82	0.0539	85%
	RF	0.94	0.0326	91%
	SVM	0.82	0.0539	82%

3.4.8 Comparison of Inferential Models Using ddG and Using dTm

The experimental thermal stability data measured by dTm for single-point mutations of three proteins, Bacteriophage T4 lysozyme, Human lysozyme, and E. coli Ribonuclease HI, are also available in ProTherm database. We are interested in comparing the AUC and prediction accuracy of DT, RF, and SVM models using dTm and ddG. RSP was calculated for all mutants in each of three proteins which have a known dTm value and DT, RF, and SVM models were developed using the RSP and the

sign of dTm. The AUC and prediction accuracy of DT, RF, and SVM models using both ddG and dTm are shown in Table 3.8. Our data indicate that the models using either dTm or ddG make a similar prediction of change of stability in single point mutants (Table 3.8). We did not observe significant difference of AUC and prediction accuracy between models using dTm and those using ddG (Table 3.8).

Table 3.8 Comparison of AUC and prediction accuracy using dTm and uisng ddG in single-point mutants. DT, RF, and SVM models were generated to predict both thermal stability change (dTm) and thermodynamic stability change (ddG) for single-point mutations in Bacteriophage T4 lysozyme, Human lysozyme, and E. coli Ribonuclease HI.

Protein	Model	AUC (dTm)	AUC (ddG)	Accuracy (dTm)	Accuracy (ddG)
Bacteriophage T4 Lysozyme	DT	0.92	0.85	92%	88%
	RF	0.97	0.96	94%	91%
	SVM	0.8	0.82	86%	87%
Human Lysozyme	DT	0.74	0.89	85%	90%
	RF	0.95	0.95	89%	92%
	SVM	0.83	0.88	87%	90%
E. coli Ribonuclease HI	DT	0.76	0.78	85%	87%
	RF	0.97	0.99	91%	93%
	SVM	0.89	0.92	93%	94%

3.4.9 Comparison of Analysis of Mutants Using Coordinates of the Cα Atoms and Using the Weighted Center of Mass (CM) of Atoms

Utilizing the atomic coordinates of PDB files, we represent each protein structure as a discrete set of points in 3-dimensional space, corresponding to the coordinates of the Cα atoms of the constituent amino acids in the protein. All the results presented so far use the coordinates of the Cα atoms. Alternatively, the coordinates of the weighted center of mass (CM) of atoms comprising each amino acid side chain may be used. We would

like to know whether or not there is difference between these two approaches. Delaunay Tessellation of three proteins with highest number of mutants using Cα and CM are conducted. RSP were built and models of DT, RF, and SVM using ddG were generated. The AUC and prediction accuracy are shown in Table 3.9. We did not observe significant difference of AUC and prediction accuracy between models using Cα approach and those using CM approach (Table 3.9). The data indicate the models using either Cα approach or CM approach make a similar prediction of change of stability following single point mutation.

Table 3.9 Comparison of AUC and prediction accuracy using Cα and CM. Machine learning models using DT, RF, and SVM were generated using Cα approach and CM approach in Bacteriophage T4 lysozyme, Human lysozyme, and E. coli Ribonuclease HI to predict the class of ddG.

Protein	AUC (Cα)	AUC (CM)	Accuracy (Cα)	Accuracy(CM)
Bacteriophage T4 Lysozyme	0.85(DT) 0.96(RF) 0.82(SVM)	0.85(DT) 0.96(RF) 0.84(SVM)	88%(DT) 91%(RF) 87%(SVM)	90%(DT) 92%(RF) 88%(SVM)
Human Lysozyme	0.89(DT) 0.95(RF) 0.88(SVM)	0.85(DT) 0.93(RF) 0.88(SVM)	90%(DT) 92%(RF) 90%(SVM)	88%(DT) 92%(RF) 91%(SVM)
E. coli Ribonuclease HI	0.78(DT) 0.99(RF) 0.92(SVM)	0.84(DT) 0.98(RF) 0.94(SVM)	87%(DT) 93%(RF) 94%(SVM)	88%(DT) 93%(RF) 96%(SVM)

A limitation of our computational mutagenesis approach based on Delaunay tessellation of protein structures is that it requires the available tertiary structure of the protein. However, this method also has several advantages. First, the computation of the RS and RSP can be very fast. Second, unlike most other structure-based methods such as molecular dynamics and homology modeling, our method allows the analysis of all

possible missense substitutions for a given protein due to relatively low input requirements. Finally and most importantly, our best models derived from RF using RSP can achieve a prediction accuracy of over 90% for the stability classes of four proteins for which a certain amount of experimentally mutant stability data are available. We expect that our method is applicable to other proteins for which a sufficient amount of experimental stability data is available to initially train the machine learning models.

3.5 Conclusions

Computational mutagenesis technique based on a four-body statistical potential, generated via Delaunay tessellation of protein structures have been used to analyze the thermodynamic stability alternation following single-point mutations of 17 proteins and double-point mutants of three proteins. The computational mutagenesis results in a scalar value and a vector representation of every single amino acid replacement in these proteins. First, a comparison between the measured stability using ddG of the experimentally synthesized single-point mutants and their corresponding residual scores reveals a strong structure-stability correlation in 16 of 17 proteins with a minimum of 30 experimentally synthesized mutants available in ProThem database. Specifically, mutants are divided into two stability classes (positive and negative) based on the sign of ddG in each protein. At each stability class, a mean residual is calculated for the set of mutants in that class. We observe that a decrease in stability level is associated with a drop in mean residual score. Second, this correlation is contributed primarily by non-conservative amino acid substitutions. Third, we demonstrated the correlation is driven mainly by the residue mutations in the buried sites. Within each stability class, the mean residual of the

substitutions in exposed sites is minimal in magnitude, and they do not contribute to the correlation. Fourth, a strong correlation is observed between mean residual score and the thermodynamic stability changes following double-point mutations in T4 lysozyme, Bacteriophage f1 Gene V, and Streptococcus Protein G. Fifth, the residual profile vectors are used in conjunction with three supervised learning schemes, DT, RF, and SVM, to generate accurate inferential models of stability alternation induced by single-point mutations in four protein systems as well as by double-point mutation in three protein systems studied. Finally, the models using either Cα approach or CM approach make a similar prediction of change of thermodynamic stability following single point mutation.

4. Structure-Function Correlations and Accurate Inferential Models of p53 and SERCA1 Mutant Activity

4.1 Abstract

The four-body statistical potential function was applied to every point mutant of p53 to generate a scalar score, residual score (RS), and a vector characterization, residual score profile (RSP). RS is the difference of four-body statistical potential between the mutant and its wild-type protein. RSP quantifies the environmental change, relative to the wild-type protein, that occurs at every residue position following the point mutation. In current study, RS are first compared with transactivation activity of eight different promoters for 932 experimentally synthesized missense p53 mutants, extracted from the Universal Mutation Database (UMD) p53 database, in the DNA-binding domain (DBD) of the tumor suppressor TP53, and a strong correlation is revealed for each of eight promoters. Second, we compared RS with the frequency of p53 mutations and an inverse correlation is observed. Third, we developed p53 inferential models of p53 transactivation activity using three machine learning algorithms and RSP, and the predicted accuracies of the models vary between 63% and 80% depending on the promoters and the machine learning algorithms used. Finally, the correlation of RS with transport activity of 98 experimentally synthesized sarcoendo plasmic reticulum calcium

ATPases (SERCA1) mutants is similarly identified and inferential models to predicted SERCA1 transport activity were also derived.

4.2 Introduction

p53 is a transcription factor encoded by the tumor suppressor gene TP53, which is the most frequently mutated gene in human cancers, and plays a significant role in the development of cancer. The IARC TP53 Mutation Database compiles all TP53 gene mutations identified in human cancers and cell lines that have been reported in the scientific literature (www-p53.iarc.fr) (Olivier et al 2002). This database includes data set for somatic mutations in sporadic cancers, germline mutation in familial cancers, polymorphisms identified in human populations, functional properties of P53 mutant proteins, and gene status in human cell-lines**.** The p53 protein is activated by various stress conditions, including DNA damage, oncogene activation or hypoxia, and regulates the transcription of several genes involved in DNA repair, cell cycle checkpoints or apoptosis (Vogelstein et al 2000). In a systematic study, Kato et al used a yeast-based expression assay to measure the transactivation activity of all possible p53 missense mutations (Kato et al 2003).

Most mutations occur in the sequence-specific DNA binding domain (DBD) of the p53 protein (residues 102-292) (Kato et al 2003). Although the entire structure of p53 has not been solved, a high resolution (2.20 Å) x-ray structure of the DNA binding domain of p53 is determined (Cho et al 1994) and is available in Protein Data Bank (PDB) under the entry 1tsr (Berman et al 2000). The DBD domain consists of a beta sandwich that provides a scaffold for two large loops and a loop-sheet-helix motif (Cho et

al 1994). The two large loops are held together in part by a tetrahedrally coordinated zinc atom. DNA binding surface of p53 comprise the two large loops and the loop-sheet-helix motif. Residues from the loop-sheet-helix motif interact in the major groove of the DNA, while an arginine from one of the two large loops interacts in the minor groove. The loops and the loop-sheet-helix motif consist of the conserved regions of the core domain and account for the majority of the p53 mutations identified in tumors (Cho et al 1994).

Analysis of activity of protein mutants is important for understanding protein structure and function, studying the function of gene variations, and designing new proteins. Experimental mutagenesis studies of protein mutation are expensive and time-consuming, and thus the number of mutants that can be included in such studies is limited. When it is not feasible to conduct such experiments on all substitutions, theoretical prediction of the activity of protein mutants would be useful for guiding site-directed mutagenesis and other protein-engineering techniques. Various computational methods have been described to predict the effect of genetic variations. Based on the type of features used, the computational prediction methods can be classified as structure-based, evolutionary-based, or a combination of both. The structure-based approach utilizes the protein's structural information alone to predict the functional consequences of amino-acid substitutions (Wang et al 2001, Stitziel et al 2003). The structural features used include the native amino acid's solvent accessibility that measures its exposure to the surrounding environment and the crystallographic B-factor that measures the atomic mobility of the wild-type amino acid and its ability to accommodate mutations. The evolutionary-based method used evolutionary feature for prediction of deleterious

mutations (Ng et al 2003, del Sol Mesa et al 2003). It is thought that evolutionary forces have selected the allowable set of amino-acid substitutions at a particular site in the protein (Cargill et al 1999). The allowable set of substitutions can be determined by aligning the query sequence with homologous sequences. If a mutated amino acid does not appear in the homology-derived set of substitutions at that site, a substitution by such an amino acid is likely to have a deleterious effect on the protein function. The third approach uses both the structural and evolutionary information and appears to achieve better predictive power (Chasman et al 2001, Sunyaev et al 2001, Saunders et al 2002, Krishnan et al 2003, Karchin et al 2005, Verzilli et al 2005).

Computational geometry technique based on Delaunay tessellation of protein structure has recently been used to analyze the functional effects of single amino acid mutations (Masso et al., 2006). The four-body statistical potential is defined as a function that assigns to each quadruplet a log-likelihood score that compares the observed normalized frequency to an expected rate of occurrence (Singh et al 1996, Vaisman et al 1998). The four-body statistical potential can be used to generate a scalar and a vector representation for every point mutant of proteins. The scalar value called residual score (RS) is the difference of four-body statistical potential between the mutant and its wild-type (wt) protein. Vector components of a mutant or residual score profile (RSP) quantify the relative environmental changes from wt at each of the protein positions induced by the amino acid substitution that created the mutant protein. RS has been shown to correlate with the activity levels of mutants in human immunodeficiency virus (HIV)-1 protease, T4 lysozyme, and HIV-1 reverse transcriptase (Masso et al., 2006).

In a pervious study, p53 mutants were grouped into functional categories based on their transactivation activities experimentally measured in yeast functional assays using p53-response elements from eight different promoters and these functional categories was shown significant differences in average RS (Mathe et al., 2006b). Mathe et al used decision tree (DT) models and the RSP to predict p53 transactivation activity with an accuracy varying between 64.2% and 78.5% depending on the promoters (Mathe et al., 2006b). Present studies are different from Mathe's work in several ways. First, Mathe's work excluded the mutants with activity between 40% and 80% while present study use all available experimentally synthesized p53 mutants in the DNA binding domain. Exclusion of some mutants may introduce some bias of the data set. Usage of all available mutants will provide an unbiased result. Second, Mathe's analysis only used the two-class system: functional or non-functional. Current study used two, three, and four class systems to demonstrate structure-function correlation. Correlation demonstrated by three and four class systems will lead to a much more convincing conclusion than that used only by two class system. Third, only Decision Tree (DT) was used to develop the inferential models in Mathe's study. Present study used three machine learning algorithms, DT, RandomForest (RF), and support vector machine (SVM). Comparison of model performance among three different algorithms may identify the best algorithms. Finally, Mathe's p53 inferential models use only 'consistent' mutants that had similar transactivation activity according to the yeast functional assay across all eight promoters as the training data set while our inferential models are based

on all available mutants. Usage of all available mutants may eliminate any bias consequence.

In addition to the above differences, several new studies are also explored in the current work. First, since the more frequent p53 mutants is associated with more likely loss of transactivation activity (Soussi et al 2005), we compared RS and the frequency of p53 mutations in order to investigate whether or not there is any kind of correlation between RS and the frequency of p53 mutations. Such study may demonstrate an interesting association between RS and the biological effect of mutation. Second, it has been shown that the transcription activity of a p53 mutant is different at two temperatures: 30 and 37 degrees (Shiraishi et al 2004). If the transcription activity of a p53 mutant is different at these two temperatures, it is defined as a temperature sensitivity (ts) mutant. Otherwise, it is a non-ts mutant (Shiraishi et al 2004). In order to investigate whether RS are correlated with the ts of p53 mutants, we compared the average RS in two categories of p53 mutants: ts and non-ts mutants. Finally, we developed random forest learning curve in order to study the effect of the training set size on prediction accuracy of p53 mutant transactivation activity levels on one of the promoter, NOXA.

In the current study, we first compared RS with transactivation activity of eight different promoters for 932 experimentally synthesized missense mutants from the Universal Mutation Database (UMD) p53 database (http://p53.free.fr) in the DNA-binding domain (DBD) of the tumor suppressor TP53 (Hamroun et al 2006). Second, we analyzed the relationship between RS and the frequency of p53 mutations. Third, we develop inferential models using three supervised machine learning algorithms and RSP

to predict whether a mutation in the DBD of p53 will increase or decrease transactivation activity of p53. Finally, the relation of RS with transport activity of 98 experimentally synthesized sarcoendo plasmic reticulum calcium-ATPases (SERCA1) mutants was investigated and the inferential models to predicted SERCA1 transport activity were also derived.

4.3 Material and Methods

4.3.1 Experimental Data

The data set for transactivation activities of 932 missense mutants in the DBD of p53 on eight different promoters was extracted from the UMD p53 database (Hamroun et al 2006). The eight p53-binding sequences were derived from the promoters of WAF1, MDM2, BAX, h1433s, AIP1, GADD45, NOXA, and P53R2 genes. For each mutant on each target sequence, the transactivation activity was expressed as the percentage of activity relative to the wild-type protein. The mutants were classified using three approaches: two classes, three classes and four classes. The two class approach categorizes the phenotypes as positive (more than 50% of wild type activity) and negative (less than 50% of wild type activity). The three class approach divides the phenotypes as positive (more than 50% of wild type activity), intermediate (10 -50% of wild type activity), and negative (less than 10% of wild type activity). The four class approach groups the phenotypes as fully active (more than 100% of wild type activity), moderate (50-100% of wild type activity), low (10-50% of wild type activity), and inactive (less than10% of wild type activity). The structure of the entire p53 protein has not been solved. However, the structures of several domains of wild-type p53 have been described.

In the present study, the structure of the wild-type DBD of p53 bound to a consensus DNA sequence is used for Delaunay tessellation (Cho et al 1994). The original structure data file in PDB under the entry 1tsr contains three p53 monomers associated with one 10-mer DNA oligomer matching the p53-binding consensus. The monomer B (residues 96–289) is used for the tessellation because it is the only one that makes contacts in both major and minor grooves of target DNA and is the best representation of the conformational constraints involved in p53 binding to its responsive DNA sequence. The data for p53 mutation frequency are also obtained from the UMD p53 database.

p53 mutants can be categorized as temperature-sensitive (ts) or non -temperature-sensitive (non-ts) because they exhibit different transcription activities under different temperatures. The temperature-sensitivity data of p53 mutants were extracted from the published results reported by Shiraishi et al (Shiraishi et al 2004).

SERCA1 functional data were collected from Scanning Mutagenesis of 98 single-point mutations located on the transmembrane sequences, M4, M5, M6, and M8 (Rice et al 1996). The original structure data file in PDB under the entry 1su4 with two bound calcium ions are used for Delaunay tessellation (Toyoshima et al 2000).

4.3.2 Four-Body Statistical Potential.

A non-homologous training set of 1417 high-resolution crystallographic protein structures is selected from the PDB (Berman et al 2000). Utilizing atomic coordinates of alpha-carbons, each structure is represented as a discrete set of points in 3-dimensional space, corresponding to alpha-carbon (Cα) coordinates of the constituent amino acid residues in the protein. A geometrical construct known as Delaunay tessellation uses Cα

coordinates to create a partitioning of the space occupied by a protein structure. Delaunay tessellation of each protein structure yields an aggregate of non-overlapping, space-filling, irregular tetrahedra, referred to as Delaunay simplices, whose vertices are the amino acid point representations (Singh et al 1996, Vaisman et al 1998). Each Delaunay simplex are constructed using the program Quickhull (Barber et al 1996) that computes the convex-hull (smallest convex set that contains defined points) of the set of residue points. Assuming order independence, there are total 8855 different possible types of quadruplets based on the 20 amino acid letter codes (Singh et al 1996, Vaisman et al 1998). The four-body statistical potential is defined as a function that assigns to each quadruplet a log-likelihood score that compares the observed normalized frequency to an expected rate of occurrence (Singh et al 1996, Vaisman et al 1998). All programs used to call qhull and compute the statistical scores have been written in Java by Zhibin Lu.

4.3.3 Residual Scores and Residual Score Profiles

Total potential of the protein is the sum of log-likelihood scores of all simplices that form the Delaunay tessellation of the protein structure (Masso et al., 2003). The residual score of a mutant is the difference between the mutant and wt protein total potential (RS = total potential (mutant) – total potential (wt)). The RSP of a mutant is the difference between the mutant and wt protein potential profile, and the value of each component is an environmental change (EC) score (Masso et al., 2006). The dimensionality of RSP is equivalent to the number of residues in the primary sequence of the protein, and each vector component is a measure of the change in the local environment at every residue position relative to wt due to the specific mutations, and

implicitly indicate the topological connections between the mutated residue positions and their nearest-neighbors via the presence of non-zero EC scores as components in the RSP. RS and RSP were calculated for 932 single site p53 mutants using the structure of the DBD of wt p53 in PDB under the PDB entry 1tsr bound to a consensus DNA sequence (Cho et al 1994) and for 98 single site SERCA1 mutants using the original structure data file in PDB under the entry 1su4 with two bound calcium ions (Toyoshima et al 2000).

4.3.4 Comprehensive Mutational Profile

For each mutant protein, a residual score is calculated as the difference in topological scores between the mutant and the wt protein. When each residue position is mutated to other 19 amino acids, 19 residual scores were individually computed. The comprehensive mutational profile (CMP) of a protein is a vector of average of 19 residual scores for all residue positions in the protein. Mathematically, CMP Vector = (avg(ECSaa1), avg(ECSaa2), … , avg(ECSaan)) where n is the number of residues in a given protein, avg(ECSaa1) is the average of 19 EC scores of the first amino acid when it is mutated to other 19 amino acids. At each residue position in the protein, a CMP component or a CMP score is thus the mean of the residual scores associated with all possible 20 mutations. The zero residual score associated with a substitution of a wt residue with itself is included at each position for completeness.

4.3.5 Supervised Machine Learning Methods

The Weka suite of machine learning tools is used to generate and evaluate the performance of inference models for p53 transactivation activity and SERCA1 transport activity, with a specific focus on supervised classification implementations of three

algorisms, decision tree (DT), RandomForest (RF), and support vector machine (SVM) (Frank et al 2004, Witten et al 2005). The training set employed for DT, RF, and SVM model building consists of a slight modification of the RSP of the p53 and SERCA1 mutants. In particular, three additional components (wt residue, position number, replacement residue) are inserted at the beginning of the vector, and the activity class of the mutant is added to the end of the vector. To determine the overall prediction accuracy of a model, a stratified 10-fold cross-validation (10 CV) was applied to both prediction algorithms. The 10-fold cross-validation splits the data sets into 10 nearly equal groups, and each group in turn was used as the test set while the remaining groups were used for training. Mutants are classified as positive and negative according to their activity levels. In p53, if the mutant has greater than 50% of wt activity, it belongs to the positive class. Otherwise, it is the negative class. In SERCA1, if the mutant has greater than 75% of wt activity, it belongs to the positive class. Otherwise, it is the negative class.

An approach based on the receiver operating characteristic (ROC) curve was used to evaluate model performance for both p53 and SERCA1. The ROC curve is a plot of the sensitivity (true positive rate – TPR) versus 1 – specificity (false positive rate – FPR) in the unit square. ROC curves are a common method for comparing the tradeoff between the model sensitivity and specificity. The area under the curve (AUC) is commonly used to as a quantitative measure for quality of a model. A ROC curve with AUC = 0.5 represents a model of random guessing between two classes while a ROC curve with

AUC = 1.0 represents a perfect model. A conservative estimate for the standard error (SE) of the AUC was computed using Hanley and McNeil methods (Hanley et al 1982).

4.4 Results and Discussion

4.4.1 Structure-Function Correlation in p53 Based on Mutant Phenotypes

Theoretically, there are 3686 (194 x 19) possible single point mutants that can be generated for the 194 amino acids (positions 96 to 289) in the structure of the p53 DBD (PDB id is1tsr). The transactivation activities of 932 missense mutants in the p53 DBD on eight different promoters were extracted from the UMD p53 database (Hamroun et al 2006) and were classified according to their activity relative to the wild-type protein using three approaches: two classes, three classes, and four classes (see Materials and Methods).

In the case of two class approach, we calculated mean residual score of all the mutants in the positive and negative classes for each of eight promoters. The results are shown in Table 4.1. Mean residual scores of positive mutants are consistently larger than that of negative mutants in each of eight promoters. Furthermore, two sample t-tests for independent samples with unequal variances show that a highly statistically significant difference between the mean residual scores of the positive and negative classes was observed in each of eight promoters (p values range from 0.0012 and 3.75E-10). The similarity of the results for eight promoters suggests that mutant residual scores encapsulate structural information about p53 protein that can be used to predict the strong impact of protein structure on activity. These results demonstrate that a strong correlation

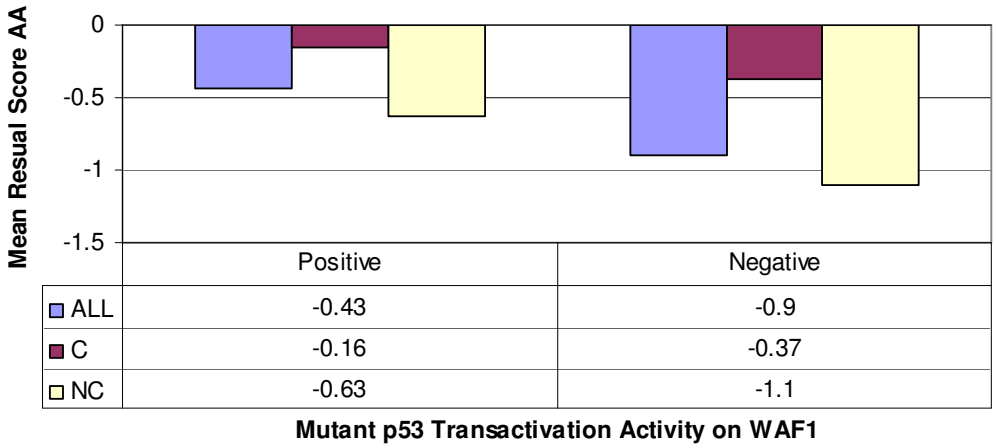
exists between the transactivation activities and the mean residual scores of the p53 mutants.

Table 4.1 Correlation of mean residual scores with p53 transactivation activity based on the approach of two mutant activity classes. RE is the p53-binding sequences from eight promoters. Mean residual scores of 932 experimentally synthesized p53 single-point mutants were computed for positive and negative classes. The residual score of a mutant is defined as the difference between mutant and wild-type topological scores. MRS POS is mean residual scores of positive mutants. CNT POS is the number of positive mutants. MRS NEG is mean residual scores of negative mutants. CNT NEG is the number of negative mutants. P Value is the indicator of statistically significant difference using two-sample t-test for independent samples with unequal variances. SD POS is the standard deviation of the mean residual scores of positive mutants. SD NEG is the standard deviation of the mean residual scores of negative mutants.

RE	MRS POS	CNT POS	MRS NEG	CNT NEG	P Value	SD POS	SD NEG
WAF1	-0.43	199	-0.9	733	0.0012	1.63	2.36
NOXA	-0.36	475	-1.25	457	1.10E-09	1.76	2.56
MDM2	-0.38	213	-0.92	719	1.00E-04	1.56	2.38
BAX	-0.43	300	-0.97	632	8.64E-05	1.69	2.43
v14_3_3_s	-0.35	339	-1.06	593	1.58E-07	1.63	2.48
AIP	-0.37	343	-1.05	589	7.09E-07	1.64	2.48
GADD45	-0.3	396	-1.17	536	3.75E-10	1.74	2.47
p53R2	-0.38	507	-1.3	425	8.82E-10	1.79	2.58

Based on Dayhoff's substitution matrix (Dayhoff et al 1978), twenty amino acids can be classified into six groups: (V,L,I,M), (R,K,H), (D,E,N,Q), (F,Y,W), (C), (A,S,T,G,P). The amino acids with common chemical and physical properties tend to fall into the same group. The substitution of amino-acid can be categorized as conservative or non-conservative based on such grouping. When the mutant and wild-type amino acids fall within same groups, the substitution is considered conservative. Otherwise, the mutation is classified as non-conservative. The p53 mutants in each class were further

A.



B.

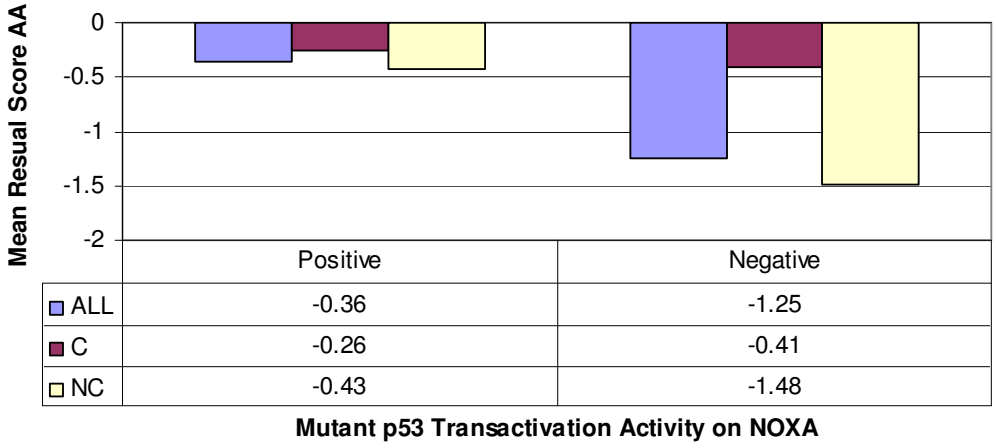
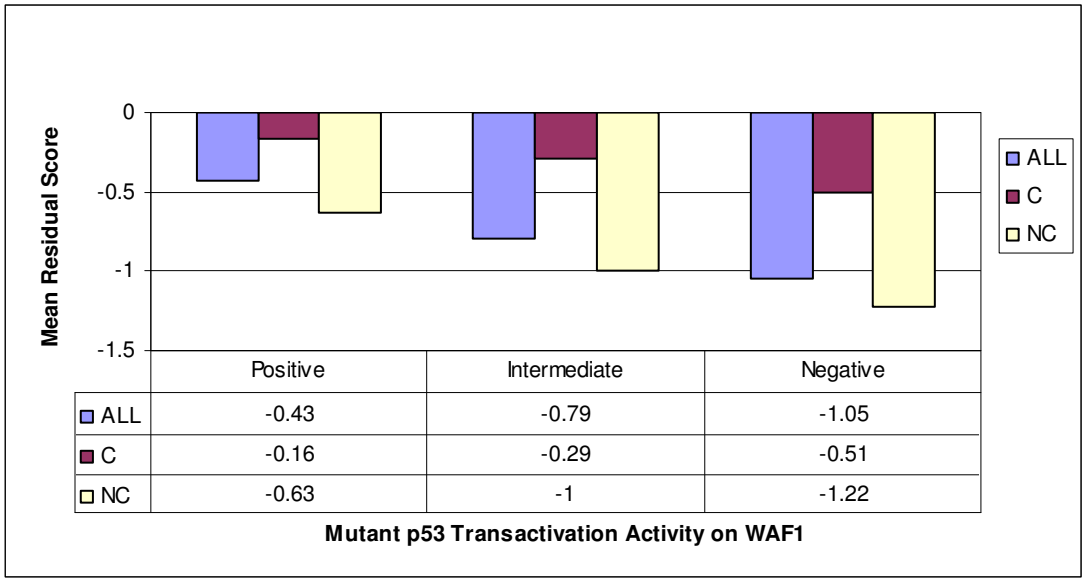


Figure 4.1 Two-class p53 structure-function correlation. Comparison of the p53 transactivation activity on the promoters WAF1 (A) and NOXA (B) with the mean residual scores of positive and negative mutants. The residual score of a mutant is defined as the difference between mutant and wild-type topological scores. Mutants in each activity class are further subdivided based on whether they are conservative or non-conservative substitutions of the wild-type residue.

divided based on whether a mutation represented a conservative (C) or non-conservative (NC) substitution of the wt residue. For each class, a mean residual score for each of these subgroups were computed for the p53-binding sequences from two promoters, WAF1 and NOXA and the results are shown in Figure 4.1A (WAF1) and Figure 4.1B (NOXA). Our data demonstrate that the correlation of mean residual score with transactivation activity on these two promoters is driven primarily by the NC substitutions with little or no contribution from the C substitutions.

In order to study whether or not there is a trend for this correlation, we further subdivided the negative mutants to generate a three-class system, and then further subdivided the positive mutants to yield a four-class system (see Materials and Methods). For each class of the three-class system, the mean residual score for each of C and NC subgroups were computed for the p53-binding sequences from two promoters, WAF1 and NOXA and the results are shown in Figure 4.2A (WAF1) and Figure 4.2B (NOXA) respectively. A trend that more active mutants have higher mean residual score is clearly observed in both promoters. We obtained similar results in the four-class systems (Figure 4.3A and Figure 4.3B). In all cases, the trend is driven predominantly by the NC substitutions while C substitutions generally have a minimal impact on sequence-structure compatibility regardless of the observed phenotype. The data shown are from the analysis using the activity of theWAF1 and NOXA promoters, but similar results were obtained with the other six promoters (data not shown). Since such structure-function correlations based on mutant activity have been observed for other proteins

A.



B.

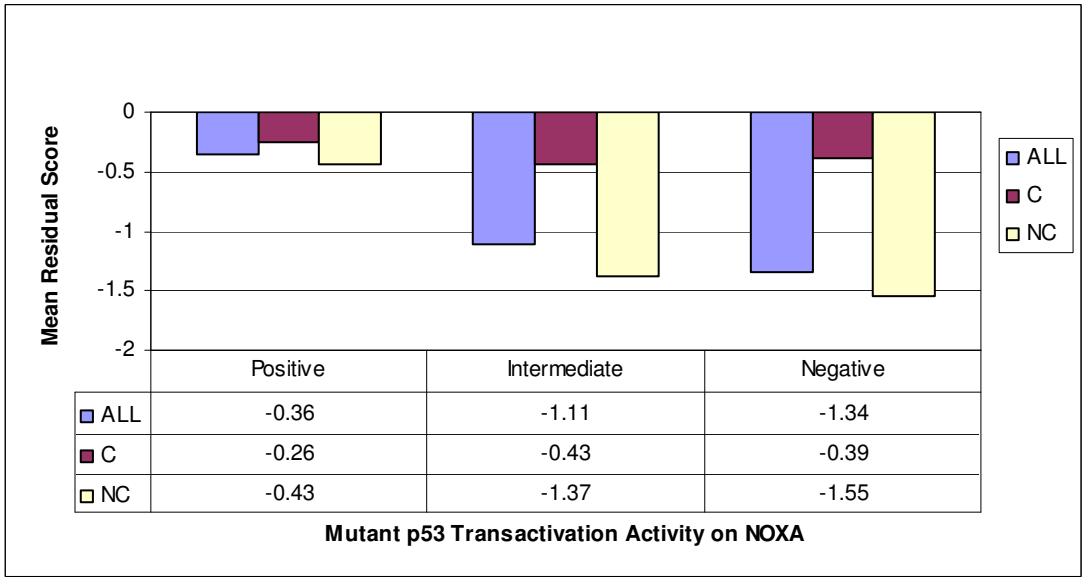
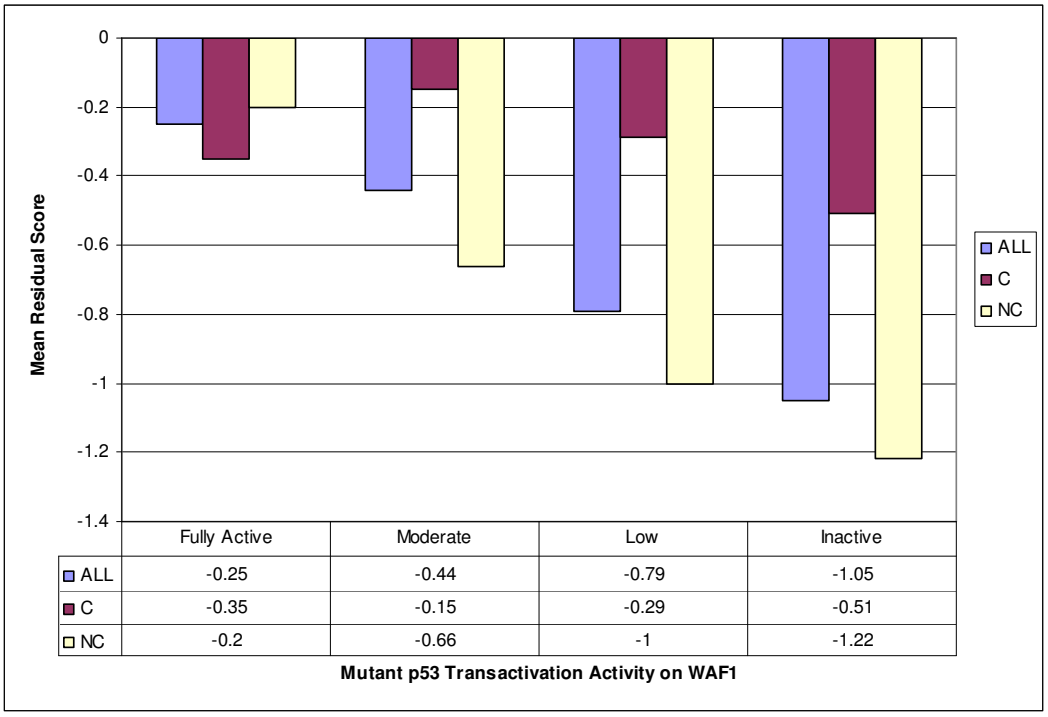


Figure 4.2 Three-class p53 structure-function correlation. Comparison of the p53 transactivation activity on the promoters WAF1 (A) and NOXA (B) with the mean residual scores of positive and negative mutants. The residual of a mutant is defined as the difference between mutant and wild-type topological scores. Mutants in each activity class are further subdivided based on whether they are conservative or non-conservative substitutions of the wild-type residue.

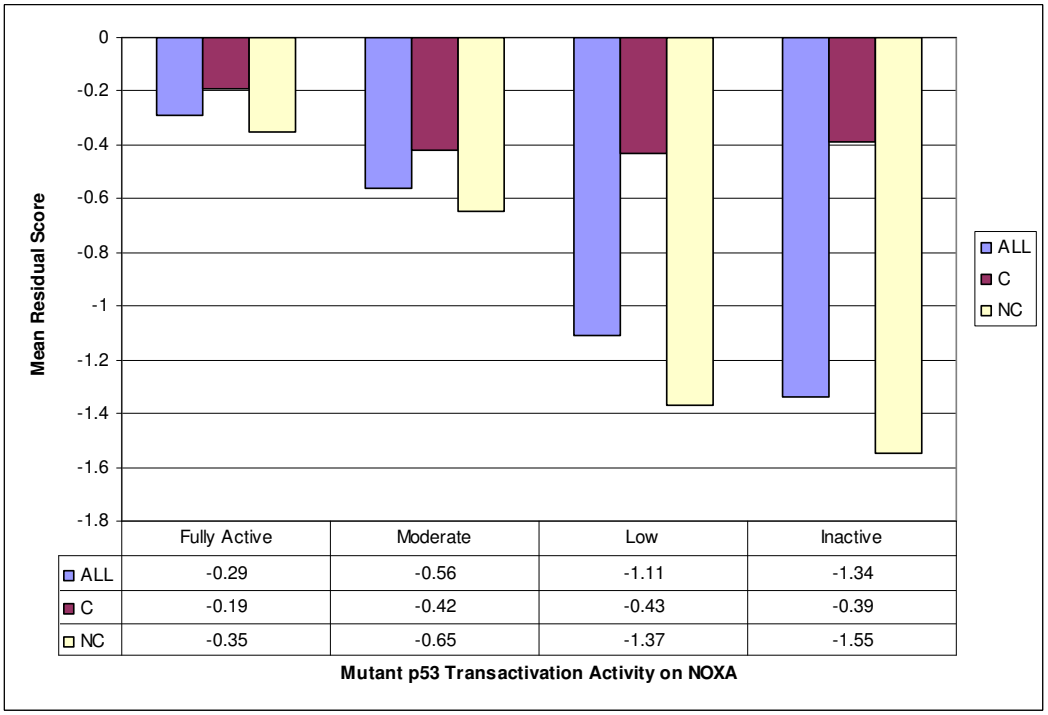
A.



	Fully Active	Moderate	Low	Inactive
ALL	-0.25	-0.44	-0.79	-1.05
C	-0.35	-0.15	-0.29	-0.51
NC	-0.2	-0.66	-1	-1.22

Mutant p53 Transactivation Activity on WAF1

B.



	Fully Active	Moderate	Low	Inactive
ALL	-0.29	-0.56	-1.11	-1.34
C	-0.19	-0.42	-0.43	-0.39
NC	-0.35	-0.65	-1.37	-1.55

Mutant p53 Transactivation Activity on NOXA

Figure 4.3 Four-class p53 structure-function correlation. Comparison of the p53 transactivation activity on the promoters WAF1 (A) and NOXA (B) with the mean residual scores of positive and negative mutants. The residual of a mutant is defined as the difference between mutant and wild-type topological scores. Mutants in each activity class are further subdivided based on whether they are conservative or non-conservative substitutions of the wild-type residue.

(Masso et al., 2006), the findings for p53 further support our hypothesis that structure-function correlation may be applicable to all proteins.

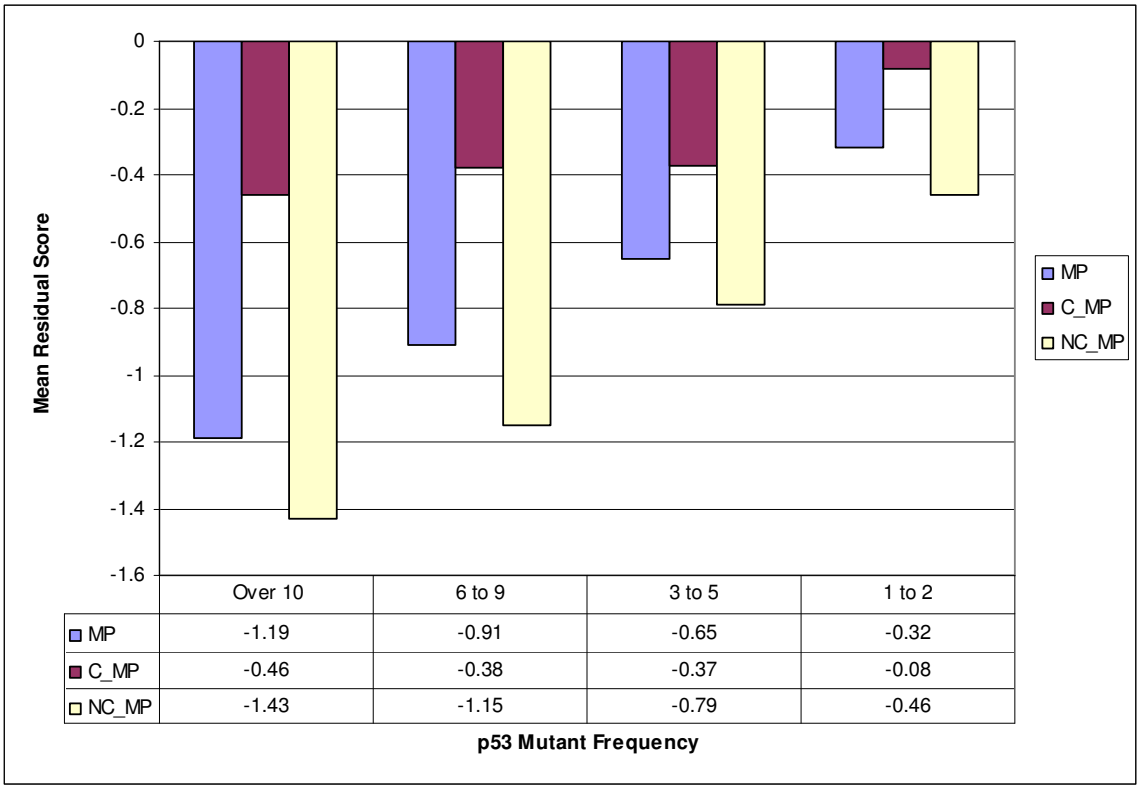
4.4.2 Inverse Correlation RS with the Frequency of p53 Mutations.

TP53 mutation is the most frequent genetic alteration found in human cancers. Over 15,000 tumors with TP53 mutations have been published, leading to the description of more than 1,500 different TP53 mutants (Soussi and Béroud, 2003). The frequency of these mutants is highly heterogeneous, with 11 hotspot mutants found more than 100 times, whereas 306 mutants have been reported only once (Soussi and Béroud, 2003). It has been shown that there is an inverse correlation between the transactivation activity of p53 mutants and their frequency of mutations (Soussi et al 2005, Soussi et al 2006). The most frequent TP53 mutants sustain a clear loss of transactivation activity while more than 50% of the rare TP53 mutants display significant activity (Soussi et al 2005). It would be interesting to know whether or not there is any kind of correlation between the mean residual score and the frequency of p53 mutations.

According to their frequency reported in the UMD p53 database, 932 missense mutants in the DBD of p53 were classified into four categories: 1-2, 3- 5, 7-9, and over 10. Mutants in each frequency category are further subdivided based on whether they are C or NC substitutions of the wt residue. The mean residual scores were computed for

each of four frequency categories and each of C and NC sub-groups. The results reveal an inverse correlation between p53 mutant frequencies and mean residual scores (Figure 4.4). The higher frequency is related to the lower mean residual score (Figure 4.4). Our data demonstrate that such inverse correlation of mean residual score with p53 mutation frequency is primarily driven by the NC substitutions.

Statistical analysis using two sample t-tests for independent samples with unequal



	Over 10	6 to 9	3 to 5	1 to 2
MP	-1.19	-0.91	-0.65	-0.32
C_MP	-0.46	-0.38	-0.37	-0.08
NC_MP	-1.43	-1.15	-0.79	-0.46

Figure 4.4 Inversed correlation of mean residual score with p53 mutant frequency. 932 p53 missense mutants are classified into four categories: 1 to 2, 3 to 5, 7 to 9, and over 10 according to the number of occurrences reported the UMD p53 database. Mutants in each frequency category are further subdivided based on whether they are conservative (C) or non-conservative (NC) substitutions of the wt residue. The mean residual scores were computed for each of four frequency categories and each of C and NC sub-groups.

variances did not reveal any statistical difference between categories 1-2 and 3-5, between categories 3-5 and 7-9, or between categories 7- 9 and over 10. However, The analysis of the remaining three pairs, 1-2 and 7-9, 1-2 and over 10, and 3-5 and over 10, showed a highly significant difference ($P<0.01$).

4.4.3 Comparison of RS with Temperature Sensitivity of p53 Mutants

P53 mutants can be categorized as temperature-sensitive (ts) or non -temperature-sensitive (non-ts) based on their different transcription activities at two temperatures: 30 and 37 degrees (Shiraishi et al 2004). If the transcription activity of a mutant is different at 30 and 37 degrees, it is ts mutant, otherwise, it is non-ts mutant.

In order to investigate whether RS are correlated with the temperature sensitivity(ts) of p53 mutants, we first extracted the 133 ts mutants and 1014 non-ts mutants in the DNA binding domain of p53 reported by Shiraishi et al (Shiraishi et al 2004). Then we computed the mean residual scores for ts mutants and non-ts mutants and the results are shown in Table 4.2. Our data indicate that the mean residual score between ts-mutants and that of non-ts mutants are not statistically different (p value >0.05) (Table 4.2).

Table 4.2 Comparison of RS with temperature sensitivity of p53 mutants. MRS is mean residual scores. Count is the total number of each type of mutants.

Temperature Sensitivity Type	MRS	Count
Non-ts	-0.7	1014
ts	-0.5	133

4.4.4 Inferential Models of p53 Mutant Activity using RSP

Although the mean of the mutant residual scores utilized thus far have been demonstrated to correlate well with the mutant activity classes, the intraclass variance of these scores is clearly too large to make them useful as an accurate predictive measure of mutant activity. On the other hand, RSP can be used to represent the p53 single point mutants and the RSP contains additional topological information as well as EC scores for every residue position affected by the single point mutation, and thus it may be better to be used for developing inferential models. As a result, we decided to investigate how well the information encoded in the RSP is able to differ between the mutants of different activity classes in p53 single-point mutants.

RSP were calculated for 932 single-point mutants of p53 and DT, RF, and SVM algorithms were used for supervised classification of two-class system in each of eight promoters. The AUCs were calculated to test the robustness of the predictions. For the promoter WAF1, the prediction accuracies are 80%, 79%, and 78% with corresponding AUC of 0.737, 0.68, and 0.51 for DT, RF, and SVM respectively (Table 4.3). Similar results are obtained in other seven promoters (Table 4.3). These data indicate the models can be used to predict the activity classes of p53 single-point mutants in each of eight promoters. In addition, our results demonstrate that inferential models using RF outperform those using DT and SVM.

Table 4.3 Prediction accuracy and AUC of DT and RF models in p53 single-point mutants. RE is the p53-binding sequences from eight promoters. The results are based on a 10-fold cross-validation using DT, RF, and SVM. Prediction accuracy and AUC were computed using the RSP in 932 p53 mutants for eight p53-binding sequences the two-class system. SE is the conservative estimate for the standard error of the AUC.

RE	Model	Accuracy	AUC	SE
WAF1	RF	80	0.737	0.0178
WAF1	DT	79	0.681	0.0196
WAF1	SVM	78	0.507	0.023
NOXA	RF	71	0.768	0.0154
NOXA	DT	71	0.725	0.0165
NOXA	SVM	65	0.652	0.0178
AIP	RF	72	0.761	0.0154
AIP	DT	70	0.711	0.0167
AIP	SVM	65	0.582	0.019
BAX	RF	75	0.795	0.0144
BAX	DT	75	0.729	0.0165
BAX	SVM	69	0.584	0.0195
GADD45	RF	69	0.75	0.0165
GADD45	DT	66	0.69	0.0178
GADD45	SVM	63	0.617	0.0187
MDM2	RF	78	0.729	0.0177
MDM2	DT	78	0.67	0.0195
MDM2	SVM	77	0.511	0.0224
P53R2	RF	73	0.802	0.0142
P53R2	DT	69	0.698	0.0169
P53R2	SVM	66	0.653	0.0177
V14_3_3_S	RF	73	0.775	0.0149
V14_3_3_S	DT	71	0.723	0.0164
V14_3_3_S	SVM	67	0.598	0.0189

4.4.5 Learning Curves for the Random Forest Models of p53 Single-Point Mutants

In order to gain insight into the effect of the training set size on prediction accuracy of p53 mutant transactivation activity levels on the promoter NOXA, we generated training sets of increasing size from the training set by sampling with replacement. First, the smallest training set contains 45 mutants were chosen randomly from 932 mutants ten times and each of thes training sets were used for generating models using RF with 10 CV option. The mean accuracy and standard deviation are

computed after ten runs of 10 CV. Next, we increment the number of mutants chosen by 45, and a mean accuracy and standard deviation are reported for 20 runs of 10 CV on the 90 mutants. We continue the procedure by increasing the randomly chosen training set size by 45 mutants at each increment until we reach a training set of size 900. By using these twenty different training sizes, a learning curve was generated and shown in Figure 4.5. According to the learning curve shown in Figure 4.5, it would require to use about 800 training set mutants in order to develop an RF model with an accuracy of over 70%.

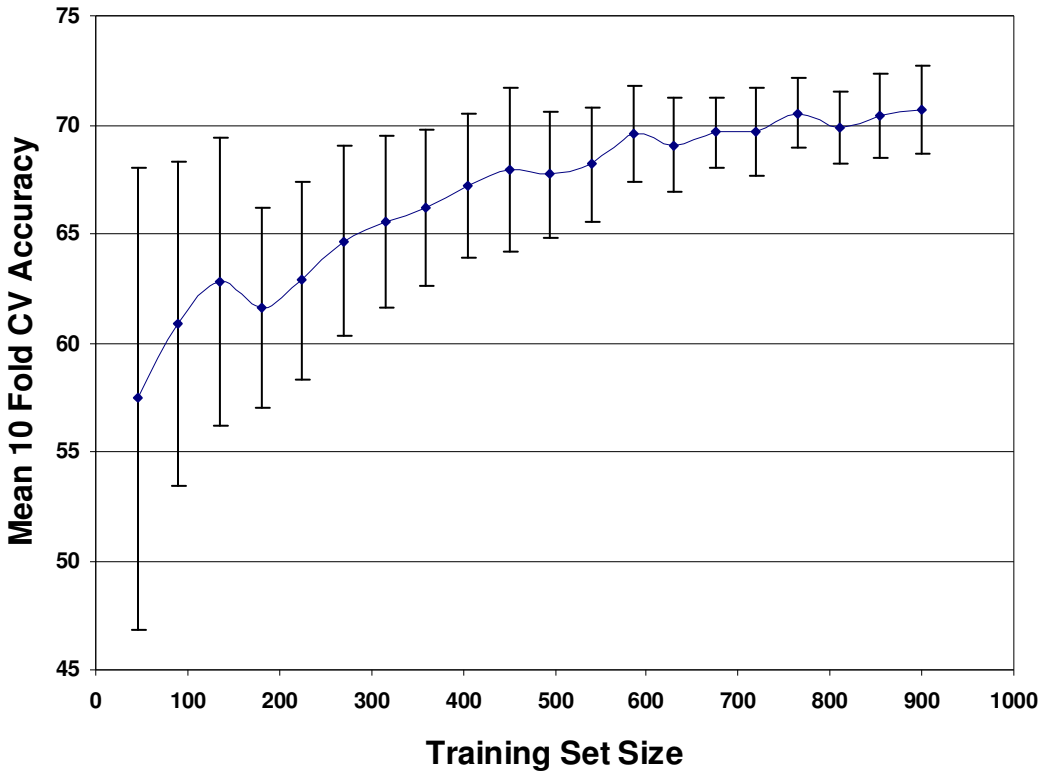


Figure 4.5 Random forest learning curve for the two-class labeling the p53 mutants using the promoter NOXA. Training sets are randomly chosen with replacement in increments of 45 mutants. For each training set size, the mean 10 CV accuracy is obtained by averaging the accuracy over twenty 10 CV runs using random forest learning. Error bars represent ±1 standard deviation from the mean.

4.4.6 SERCA1 Potential Profile and Comprehensive Mutational Profile

Total potential of a protein is calculated by summing the log-likelihoods of all Delaunay simplices formed by the tessellation. By summing the log-likelihoods of all simplices in which a particular residue participates, individual residues are each assigned a score, yielding a 3D–1D potential profile. The total statistical potential of SERCA1 with two bound calcium ions (PDB ID: 1su4) is calculated to be 181.91 and its potential profile is shown in Figure 4.6. Highest scoring residues tend to be located in the

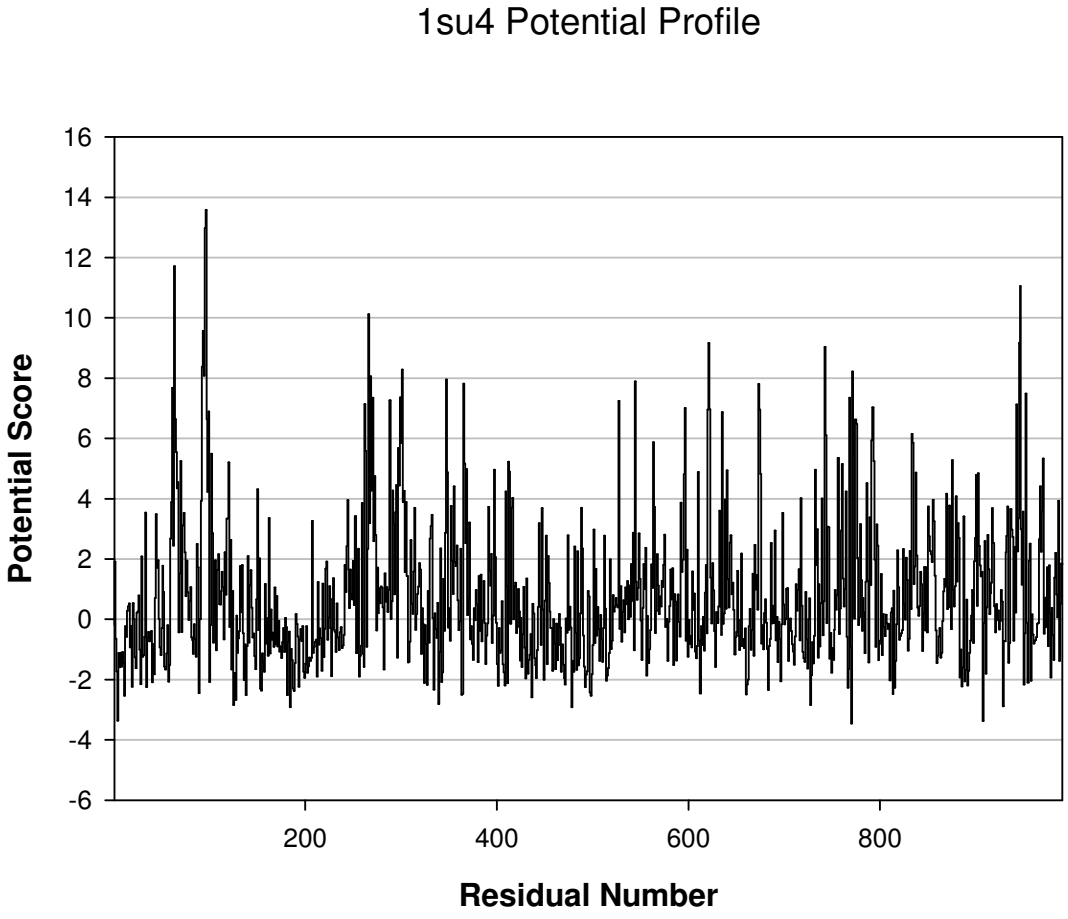


Figure 4.6 3D–1D potential profile of a wild-type SERCA1 with two bound calcium ions (PDB ID: 1su4).

hydrophobic core, while lower scoring residues are exposed amino acids.

In order to develop the comprehensive mutational profile (CMP) of SERCA1, each position the 994 amino acids is mutated 20 times (including the mutation to itself). As a result, a 20*994 matrix of total statistical potentials for all possible single residue mutants is produced. Each of the 994 columns is labeled with the corresponding residue present in the primary sequence of wt SERCA1, and each of the 20 rows is labeled with an amino acid chosen to replace the given residue in the primary sequence. Subtracting the wt potential from each cell yields the difference between mutant and wt total potentials. Finally, by averaging the values in each column, the CMP is obtained (Figure 4.7). Comparison of the SERCA1 potential profile in Figure 4.6 and the corresponding CMP in Figure 4.7 reveals a strong inverse correlation (with correlation coefficient R^2 = 0.79) as shown in Figure 4.8.

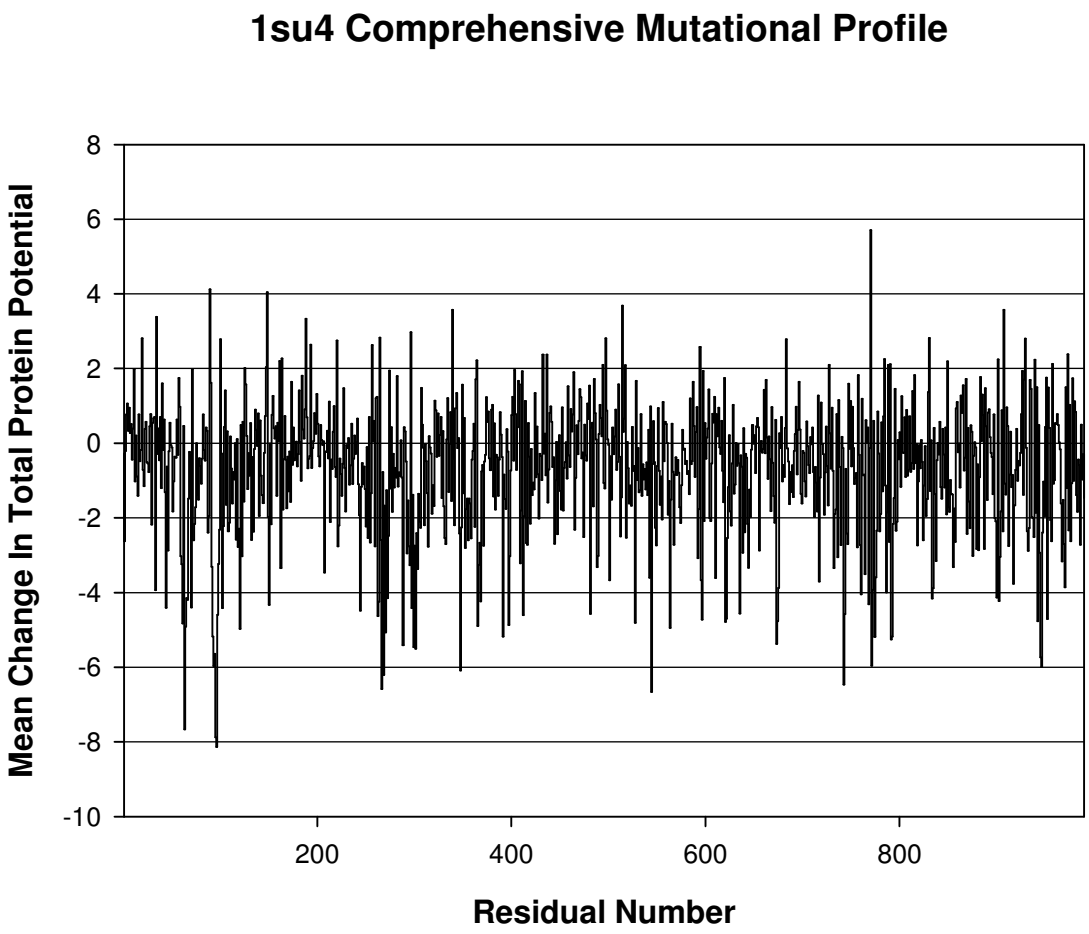


Figure 4.7 Comprehensive mutational profile of SERCA1. For each residue, the graph reflects the mean difference in total potential from wild-type protease resulting from all possible substitutions at the given position.

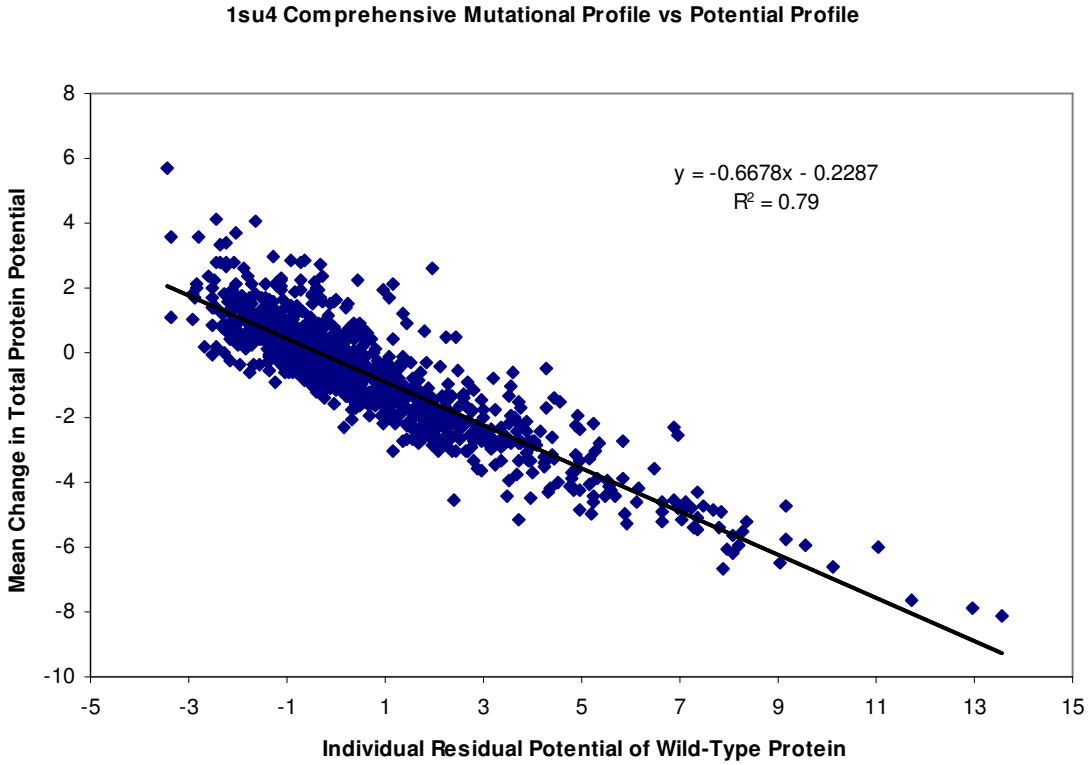


Figure 4.8 Correlation of the CMP with the individual residue potential profile of SERCA1. The graph reveals a strong inverse relationship, with correlation coefficient $R^2 = 0.79$.

4.4.7 Structure-Function Correlation of SERCA1

In order to understand the relationship between the transport activity of the single point mutants of SERCA1 and their residual scores, we separated the 98 experimentally synthesized mutants into two classes, 65 positive and 33 negative, based on transport activity of SERCA1 (Rice et al 1996) and calculated mean residual scores of the mutants in both classes. The results are shown in Figure 4.9. Mean residual scores of active mutants are larger than that of inactive mutants and the difference is statistically significant (with p-value < 0.05) using two sample t-tests for independent samples with

unequal variances. These results indicate a correlation exists between the transport activities and mean residual scores of SERCA1 mutants. Since such structure-function correlations based on mutant activity have been observed in the p53 protein shown our previous results as well as in HIV-1 protease, T4 lysozyme, and HIV-1 reverse transcriptase (Masso et al., 2006), the findings for SERCA1 further support our hypothesis that structure-function correlation may be applicable to all proteins.

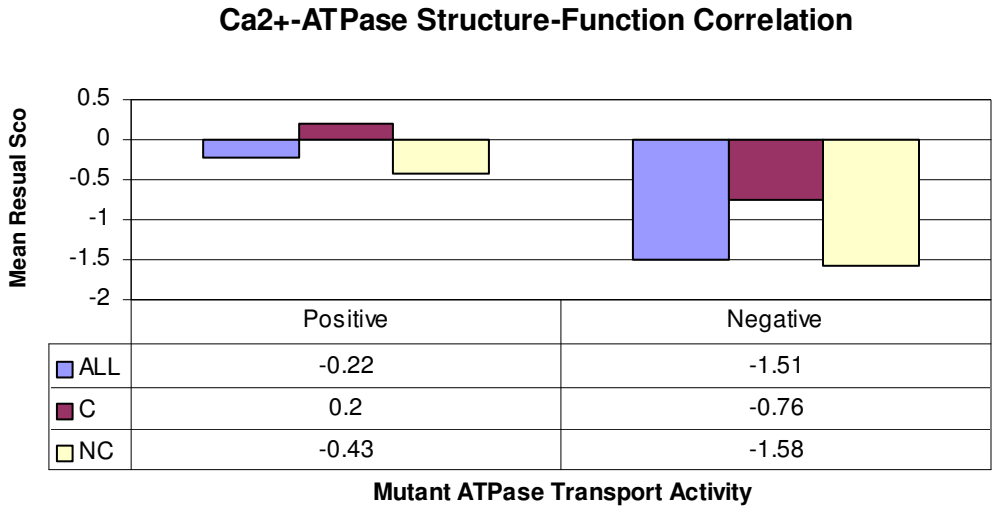


Figure 4.9 Sarcoendo plasmic reticulum calcium-ATPases (SERCA1) structure function correlation. Comparison of the activity of 98 experimentally synthesized SERCA1 mutants with the mean residual score of the mutants within each functional class. The residual score of a mutant is defined as the difference between mutant and wild-type topological scores. Two sample t-tests for independent samples with unequal variances show that there is a statistically significant difference between the mean residual scores of the positive and negative classes (p value < 0.05). Mutants in each activity class are further subdivided based on whether they are conservative or non-conservative substitutions of the wild-type residue.

4.4.8 Inferential Models of SERCA1 Mutant Activity using RSP

RSP were calculated for all 98 SERCA1 mutants. Three supervised learning schemes, DT, RF, and SVM, were used to discriminate the positive and negative activity classes. The AUCs were calculated to test the robustness of the predictions. The prediction accuracies are 82%, 82% and 75% with corresponding AUC of 0.70, 0.83 and 0.57 for DT, RF and SVM respectively (Table 4.4). These data indicates the models based on RSP may be used to accurately predict the activity classes of SERCA1 single-point mutants.

Table 4.4 Comparison of prediction accuracy and AUC for residual score of SERCA1 mutants. RSP were calculated for SERCA1 single-site mutants respectively Models were generated using DT, RF and SVM via 10 fold cross-validation method.

MODEL	AUC	ACCURACY
DT	0.7	82%
RF	0.83	82%
SVM	0.57	75%

4.5 Conclusions

Computational mutagenesis technique based on a four-body statistical potential, derived from Delaunay tessellation of protein structures, yields a scalar (the residual score) and a vector (the residual profile) representation of every single amino acid replacement in p53 at positions 96 to 289, which forms the DNA binding domain. The residual score associated with each mutant provides a quantitative measure of the relative

change in sequence-structure compatibility of the mutant from the wt p53. The RS for 932 p53 mutants belonging to discrete activity classes become increasingly negative as the classes diminish in level of activity, indicating a significant structure-function correlation. In addition, a comparison was made between RS and the frequency of p53 mutations, yielding a statistically significant inverse correlation. Furthermore, the residual profile vectors are used in conjunction with three supervised machine learning schemes (RF, DT and SVM) to develop accurate inferential models of p53 mutant activity. Finally, the correlation of RS with transport activity of 98 experimentally synthesized SERCA1mutants is similarly observed and inferential models to predicted SERCA1 transport activity were also developed.

5. Correlation of Four-Body Potential Score Derived from Delaunay Tessellations and Conservatism of Residue Substitution in Proteins

5.1 Abstract

The residual score of a mutant, which measures the relative change in overall sequence-structure compatibility from wild-type, is the difference of four-body statistical potential between the mutant and its wild-type protein. The point-mutation can be categorized as conservative or non-conservative based on the amino-acid groups classified according Dayhoff's substitution matrix. In the present study, the mean of residual scores were computed in these two categories of mutations in each of 700 proteins with low sequence similarity. The mean of residual score in conservative class is higher than that in non-conservative class in 685 of 700 protein chains (97.9%). Such difference is statistically significant in 664 protein chains using two sample t test (p value <0.05). These data indicates a strong correlation between the mean of residual score and conservatism of substitutions in proteins. Furthermore, we developed a novel statistical matrix based on the mean residual scores of all 380 types of mutations in 700 proteins studied. A statistically significant correlation is observed between this novel matrix and PAM40, PAM 80, PAM 120, PAM 250, BLOSUM62, or BLOSUM80. Finally, the inferential models of p53 activity class using this novel matrix scores are comparable to those models using residual scores. Our results suggest that this novel matrix may be served as

one of attributes for predicting functional effects of mutation in any proteins and may be applied to other areas of statistical geometry studies on the proteins with unknown three-dimensional structures.

5.2 Introduction

Twenty amino acids can be classified according to Dayhoff's substitution matrix, derived from the log odds ratio of the 250 PAM matrix, into six groups: (V,L,I,M), (R,K,H), (D,E,N,Q), (F,Y,W), (C), (A,S,T,G,P) (Dayhoff et al 1978). The amino acids with common chemical and physical properties tend to fall into the same group. The mutations can be classified as conservative or non-conservative based on such grouping. When the mutant and wild-type amino acids fall within same groups, the mutation is considered conservative. Otherwise, the mutation is classified as non-conservative.

The type of amino acid changes producing disease are not the same as the type of amino acid substitutions commonly observed among species (Miller et al 2001). The amino acid changes observed in disease patients are far more radical than the variation found among species and in non-diseased humans, implying non-conservative substitutions are more likely observed in disease-associated mutations (Miller et al 2001).

Computational geometry technique based on Delaunay tessellation of protein structure has also been used to study the functional effects of single amino acid mutations (Masso et al., 2006; Mathe et al., 2006b). The four-body statistical potential is defined as a function that assigns to each quadruplet a log-likelihood score that compares the observed normalized frequency to an expected rate of occurrence (Singh et al 1996, Vaisman et al 1998). The residual score (RS) of a mutant, which measures the relative

change in overall sequence-structure compatibility from wild-type, is the difference of four-body statistical potential between the mutant and its wild-type protein. It has been shown that the average RS of proteins with disease-associated nsSNPs was significantly lower than that of the proteins with neutral SNPs (Barenboim et al 2005). RS has also been shown to correlate with the activity levels of mutants in human immunodeficiency virus (HIV)-1 protease, T4 lysozyme, and HIV-1 reverse transcriptase (Masso et al., 2006). All these studies required the 3-dimetion structure of proteins studied have been solved.

In current study, a comparison of the average residual score was first made between conservative substitutions and non-conservative substitutions in each of 700 proteins with low sequence similarity. Second, each amino may be mutated to other nineteen amino acids and thus there are total of 380 different types of single-point mutations for all twenty amino acids based on a difference in the combination of wild-type amino acid and mutated amino acid. The mean residual scores by 380 types of mutations were calculated based on the 700 proteins studied and a novel substitution matrix was developed. Third, this novel statistical matrix was compared to PAM40, PAM80, PAM120, PAM250, BLOSUM62, and BLOSUM80 and correlation values were calculated. Finally, we developed inferential models of p53 activity based the novel matrix scores and compared the performance of models using the matrix scores and RS.

5.3 Materials and Methods

5.3.1 Four-Body Statistical Potential Function, RS and RSP

A non-homologous training set of 1417 high-resolution crystallographic protein structures is selected from the protein database bank (PDB) (Berman et al 2000). Utilizing atomic coordinates of alpha-carbons, each structure is represented as a discrete set of points in 3-dimensional space, corresponding to alpha-carbon (Cα) coordinates of the constituent amino acid residues in the protein. Delaunay tessellation of each protein structure yields an aggregate of non-overlapping, space-filling, irregular tetrahedra, referred to as Delaunay simplices, whose vertices are the amino acid point representations (Singh et al 1996, Vaisman et al 1998). Each Delaunay simplex are constructed using the program Quickhull (Barber et al 1996) that computes the convex-hull (smallest convex set that contains defined points) of the set of residue points. Assuming order independence, there are total 8855 different possible types of quadruplets based on the 20 amino acid letter codes (Singh et al 1996, Vaisman et al 1998). The four-body statistical potential is defined as a function that assigns to each quadruplet a log-likelihood score that compares the observed normalized frequency to an expected rate of occurrence (Singh et al 1996, Vaisman et al 1998).

Total potential of the protein is the sum of log-likelihood scores of all simplices that form the Delaunay tessellation of the protein structure (Masso et al., 2003). The RS of a mutant is the difference between the mutant and wild-type protein total potential (RS = total potential (mutant) – total potential (wild-type)). The residual score profile (RSP) of a mutant is the difference between the mutant and wt protein potential profile, and the value of each component is an environmental change (EC) scores (Masso et al., 2006).

5.3.2 Protein Data Set With Low Sequence Similarity

A pre-compiled culled PDB list of 1540 protein chains (cullpdb_pc30_res1.6_R0.25_d070310_chains1540) was obtained from PISCES: a protein sequence culling server (Wang et al 2003). This list has the percentage identity cutoff value of less than 30%, the resolution cutoff value of less than1.6 angstroms, and the R-factor cutoff value of 0.25. The list was generated on March 10, 2007. All 1540 PDB files were downloaded from PDB based on the PDB id of the list. 700 of 1540 chains are tessellable and were used for current study. The total number of hypothetical substitutions in 700 proteins is 2,750,782 and RS was calculated for all of them.

5.3.3 Measurement of the Correlation between Matrices

Mantel's test is a widely used method for assessing the relationships between two distance matrices or, more generally, two resemblance or proximity matrices (Mantel 1967, Mantel and Valand 1970). A free software, zt program, was used to perform the simple Mantel test between matrixes (Bonnet and Van de Peer 2002). The Mantel test involves measuring the association between the elements in two matrices by a suitable statistic, and then assessing the significance of this statistic by comparison with the distribution found by randomly reallocating the order of the elements in one of the matrices. The statistic used for the measure of the correlation between the matrices is the classical Pearson correlation coefficient.

The PAM matrices, PAM40, PAM80, PAM120, and PAM250 and BLOSUM matrices, BLOSUM62 and BLOSUM 80, were downloaded from a website of Pittsburgh Supercomputing Center

(http://www.psc.edu/general/software/packages/nwgap/manual/manual.html). Wolfson-Nussinov matrix is extracted from the public paper by Azarya-Sprinzak (Azarya-Sprinzak et al 1997). The novel matrix was created by using the mean residual scores of 380 substitution types from 2,750,782 hypothetical substitutions in 700 proteins. This matrix was called RS matrix. The lower-right half of elements in the RS matrix are used to be compared with that of PAM matrices and BLOSUM matrices. The correlation coefficient and its associated probability between the described matrices were computed using the zt program (Bonnet and Van de Peer 2002).

5.3.4. p53 Functional Data Set

The data set for transactivation activities of 932 missense mutants in the DBD of p53 on eight different promoters was extracted from the UMD p53 database (Hamroun et al 2006). The eight p53-binding sequences were derived from the promoters of WAF1, MDM2, BAX, h1433s, AIP1, GADD45, NOXA, and P53R2 genes. For each mutant on each target sequence, the transactivation was expressed as the percentage of activity relative to the wild-type protein. The mutants were classified the phenotypes as positive (more than 50% of wild type activity) and negative (less than 50% of wild type activity). The structure of the wild-type DBD of p53 bound to a consensus DNA sequence is used for Delaunay tessellation (Cho et al 1994). The original structure data file in PDB under the entry 1tsr contains three p53 monomers associated with one 10-mer DNA oligomer matching the p53-binding consensus. The monomer B (residues 96–289) is used for the tessellation because it is the only one that makes contacts in both major and minor

grooves of target DNA and is the best representation of the conformational constraints involved in p53 binding to its responsive DNA sequence.

5.3.5. Supervised Machine Learning Methods

The Weka suite of machine learning tools is used to generate and evaluate the performance of inference models for p53 transactivation activity, with a specific focus on supervised classification implementations of random forest (RF) (Frank et al 2004, Witten et al 2005). The training set vectors employed for RF model building consist of a slight modification of the matrix score, RS, and RSP of the p53 mutants. Specifically, three additional components (wt residue, position number, replacement residue) are inserted at the beginning of the vector, and the activity class of the mutant is added to the end of the vector. To determine the overall prediction accuracy of a model, a stratified 10-fold cross-validation was applied to both prediction algorithms. The 10-fold cross-validation splits the data sets into 10 nearly equal groups, and each group in turn was used as the test set while the remaining groups were used for training. Mutants are classified as positive and negative according to their activity levels. If a mutant has greater than 50% of wt activity, it belongs to the positive class. Otherwise it falls to the negative class.

An approach based on the receiver operating characteristic (ROC) curve was used to evaluate model performance for p53. The ROC curve is a plot of the sensitivity (true positive rate – TPR) versus 1 – specificity (false positive rate – FPR) in the unit square. ROC curves are a common method for comparing the tradeoff between the models sensitivity and specificity. The area under the curve (AUC) is commonly used to as a

quantitative measure for quality of a model. A ROC curve with AUC = 0.5 represents a model of random guessing between two classes while a ROC curve with AUC = 1.0 represents a perfect model. A conservative estimate for the standard error (SE) of the AUC was computed using Hanley and McNeil methods (Hanley et al 1982).

5.4 Results and Discussion

5.4.1 Correlation of RS with Conservatism of Substitutions

According to Dayhoff's substitution matrix, twenty amino acids can be classified into six groups: (V,L,I,M), (R,K,H), (D,E,N,Q), (F,Y,W), (C), (A,S,T,G,P) (Dayhoff et al 1978). Conservative substitution is defined as the replacement of amino acids within the same group which non-conservative substitution is the replacement of amino acids between groups. Each residue in a protein chain is possibly substituted by any other 19 amino acids and each substitution will be either conservative or non-conservative. In order to investigate the possible correlation between RS and the conservatism of substitutions, we calculate RS for all possible hypothetical mutations (2,750,782) in each of 700 proteins and then compute the average RS in both conservative and non-conservative groups in each protein. Our results indicate that conserved mean residual score is higher than non-conserved mean residual score in 685 of 700 protein chains (97.9%). The difference is statistically significant in 664 protein chains (p value <0.05) (Table 5.1) and only 21 protein chains do not show a statistically significant (Table 5.2). 15 protein chains (2.1%) have higher non-conserved mean residual scores and the difference is statistically insignificant in all of these 15 proteins chains (Table 5.3). The reason that the 36 (21+15) proteins, which represent only 5 percent of 700 proteins

studied, do not show such correlation is not clear. These data indicates a strong correlation between mean residual topological scores and conservation of substitution and support the notions that structural disruptions in conservative mutations are smaller that those in non-conservative mutations.

Table 5.1 Correlation of RS with Conservatism of Substitutions. 664 protein chains have higher conserved mean residual scores than non-conserved ones and their difference is statistically significant (p value <0.05). Only the first 12 entries are shown. Appendix A.2 shows the complete list. MRS C is conserved mean residual score. CNT C is the count of conserved mutations. MRS NC is non-conserved mean residual score. CNT NC is the count of non-conserved mutations. P value is the result of two-sample t test for independent samples with unequal variances.

PDB ID	MRS C	CNT C	MRS NC	CNT NC	P value	Length
1M2DA	-0.27	303	-1.35	1616	<2.2e-16	101
2I4AA	-0.03	331	-0.83	1702	<2.2e-16	107
1M55A	-0.27	581	-0.76	3086	<2.2e-16	193
1TUAA	-0.51	569	-1.15	3003	<2.2e-16	188
1H5QA	-0.29	822	-0.7	4118	1.44E-14	260
1UF5A	-0.25	918	-0.64	4839	3.75E-14	303
1O9RA	-0.16	507	-0.67	2571	<2.2e-16	162
2GMYA	-0.1	445	-0.62	2348	5.19E-15	147
1VKIA	-0.08	515	-0.92	2620	<2.2e-16	165
1ZHVA	-0.27	414	-0.97	2132	<2.2e-16	134
2AGYA	-0.04	1110	-0.82	5730	<2.2e-16	360
1QLWA	-0.13	1003	-0.71	5039	<2.2e-16	318

Table 5.2 Twenty-One Protein Chains with Higher Conserved Mean RS Than Non-Conserved Ones. The differences of mean RS in the classes of conserved and non-conserved categories are not statistically significant in all 21 proteins (p value >0.05).

SPECIES	PROTEIN	PDB ID
AGROBACTERIUM TUMEFACIENS	Transcriptional repressor traM	1RFYA
AGROBACTERIUM TUMEFACIENS	Hypothetical protein Atu1052	2GZ4A
ARABIDOPSIS THALIANA	expressed protein	1VK5A
ARABIDOPSIS THALIANA	WRKY transcription factor 1	2AYDA
ASPERGILLUS NIGER	Aspergillopepsin II light chain	1Y43A
BOS TAURUS	FIBRINOGEN GAMMA-B CHAIN	1JY2P
HOMO SAPIENS	TPR1-DOMAIN OF HOP	1ELWA
HOMO SAPIENS	VIMENTIN	1GK7A
HOMO SAPIENS	Myotonin-protein kinase	1WT6A
HOMO SAPIENS	FYVE-finger-containing Rab5 effector protein rabenosyn-5	1Z0JB
HUMAN PAPILLOMAVIRUS TYPE 16	Regulatory protein E2	2NNUA
MELEAGRIS GALLOPAVO	PANCREATIC HORMONE	2BF9A
MOLONEY MURINE LEUKAEMIA VIRUS	Core protein p15	1MN8A
MUS MUSCULUS	RADR ZIF268 ZINC FINGER PEPTIDE	1A1IA
NA	DOMAIN SWAPPED DIMER	1G6UA
NA	ENV POLYPROTEIN	1JEKA
NA	FxxYF motif peptide	1T7MB
NA	Voltage-dependent L-type calcium channel alpha-1C subunit	2F3YB
SACCHAROMYCES CEREVISIAE	Transcription regulatory protein SWI3	2FQ3A
SACCHAROMYCES CEREVISIAE	General control protein GCN4	2HY6A
XYLELLA FASTIDIOSA	Periplasmic divalent cation tolerance protein	2NUHA

Table 5.3 Fifteen Protein Chains with Higher Non-Conserved Mean RS Than Conserved Mean RS.

SPECIES	PROTEIN	PDB ID
ARTOCARPUS INTEGRIFOLIA	Agglutinin beta-3 chain	1UGXB
BOS TAURUS	FIBRINOGEN ALPHA CHAIN	1JY2N
BOS TAURUS	FIBRINOGEN BETA CHAIN	1JY2O
HELICOBACTER PYLORI	Aspartate 1-decarboxylase beta chain	1UHEB
HOMO SAPIENS	PARATHYROID HORMONE	1ET1A
HOMO SAPIENS	UROKINASE-TYPE PLASMINOGEN ACTIVATOR	1GJ7A
HOMO SAPIENS	NUCLEAR RNA EXPORT FACTOR	1OAIA
HOMO SAPIENS	Cellular tumor antigen p53	2B3GB
HOMO SAPIENS	Bromodomain-containing protein 4	2NNUB
NA	ENV POLYPROTEIN	1JEKB
NA	cAMP-dependent protein kinase inhibitor, alpha form	1RDQI
NA	LxxLL motif peptide	1T7FB
NA	ANTIFREEZE PROTEIN ISOFORM HPLC6	1WFBA
NA	Voltage-dependent L-type calcium channel alpha-1C subunit	2F3ZB
NA	A Kinase binding peptide	2HWNE

The Structural Classification of Proteins (SCOP) database is a comprehensive ordering of all proteins of known structure, according to their evolutionary and structural relationships (Andreeva et al 2004). Protein domains in SCOP are hierarchically classified into families, superfamilies, folds and classes. SCOP classifies proteins into ten categories:

a. All alpha proteins

b. All beta proteins

c. Alpha and beta proteins (a/b): mainly parallel beta sheets or beta-alpha-beta units

d. Alpha and beta proteins (a+b): mainly antiparallel beta sheets or segregated alpha and beta regions

e. Multi-domain proteins (alpha and beta)

f. Membrane and cell surface proteins and peptides

g. Small proteins

h. Coiled coil proteins

j. Low resolution protein structures

h. Peptides

k. Designed proteins

In order to investigate whether or not there is any difference of mean residual scores among the protein categories classified in SCOP, the class labels for the matched proteins studies were downloaded and the mean residual scores were calculated for each category. The results are shown in Table 5.4. Only the top four categories have enough number of proteins to produce a statistically meaningful result. The conserved mean residual scores between all alpha proteins and all beta proteins are very similar (-0.19 vs -0.18). Similarly, the non-conserved mean residual scores between all alpha proteins and all beta proteins are very similar (-0.75 vs -0.77). In addition, the conserved and non-conserved mean residual scores between alpha and beta proteins (a/b) and alpha and beta proteins (a+b) are also very closed to each other, which are -0.22 vs –0.19 and –0.81 vs –0.71 respectively.

Table 5.4 Conserved and Non-Conserved Mean RS in Different Protein Categories. 472 protein chains a match with the proteins in SCOP. MRS C is conserved mean residual score. CNT C is the count of conserved mutations. MRS NC is non-conserved mean residual score. CNT NC is the count of non-conserved mutations. CNT PDB is the number of matched proteins in each category.

Class Name	MRS C	CNT C	MRS NC	CNT NC	CNT PDB
All alpha proteins	-0.19	52469	-0.75	277257	84
All beta proteins	-0.18	102388	-0.77	526968	100
Alpha and beta proteins (a/b)	-0.22	134993	-0.81	696561	130
Alpha and beta proteins (a+b)	-0.19	75794	-0.71	396698	116
Multi-domain proteins (alpha and beta)	-0.21	9222	-0.73	47436	7
Membrane and cell surface proteins and peptides	-0.39	626	-1.49	3478	1
Small proteins	-0.16	3903	-0.96	22830	15
Coiled coil proteins	-0.19	2046	-0.54	10703	13
Low resolution protein structures	-0.75	340	-0.59	1674	3
Designed proteins	-0.35	693	-0.81	3582	3

5.4.2 A Novel Statistical Matrix Based on RS

Each amino acid may be mutated to other nineteen amino acids and thus there are total of 380 different types of single-point mutations for all twenty amino acids based on a difference in the combination of wild-type amino acid and mutated amino acid. The mean residual scores by 380 types of mutations were calculated and sorted in all 700 proteins and a graph was plotted based on this sorted data (Figure 5.1). Table 5.5 shows 15 mutations with lowest mean residual scores and 15 mutations with highest mean residual scores. Of 15 mutations with lowest mean residual scores, 5 are associated with a wild type amino acid of cysteine. It is expected that the structural disruption could be vigorous when cysteine is substituted by other amino acids because the disulfide bonds

may be impaired. Since an amino acid may be mutated to itself and the RS of such mutant is 0.

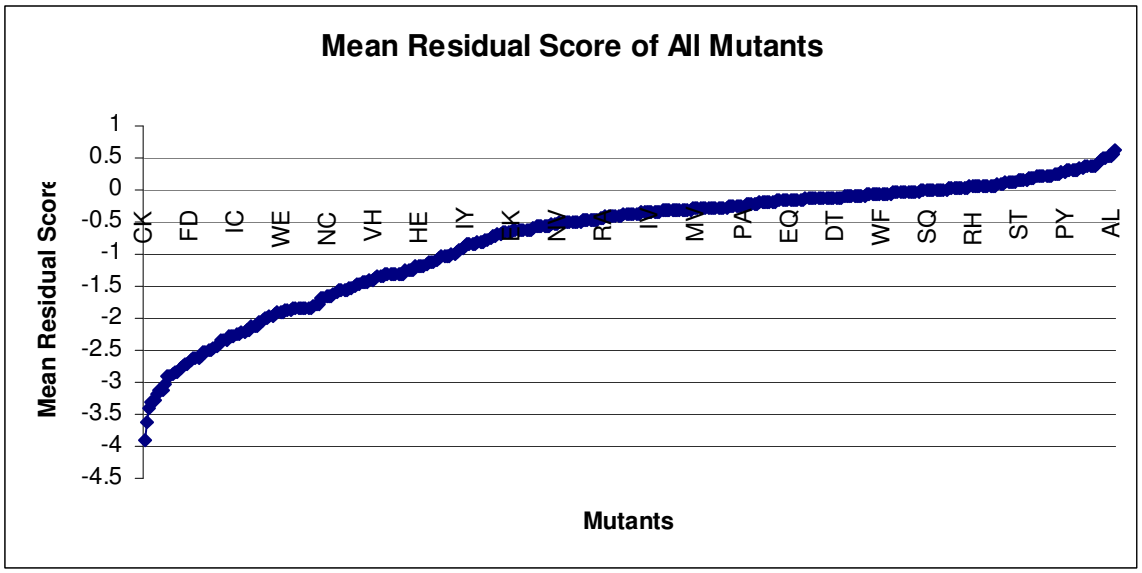


Figure 5.1 Sorted mean residual score of all possible 380 types of mutations. From the distribution char, residual scores are statistically more likely negative following mutation.

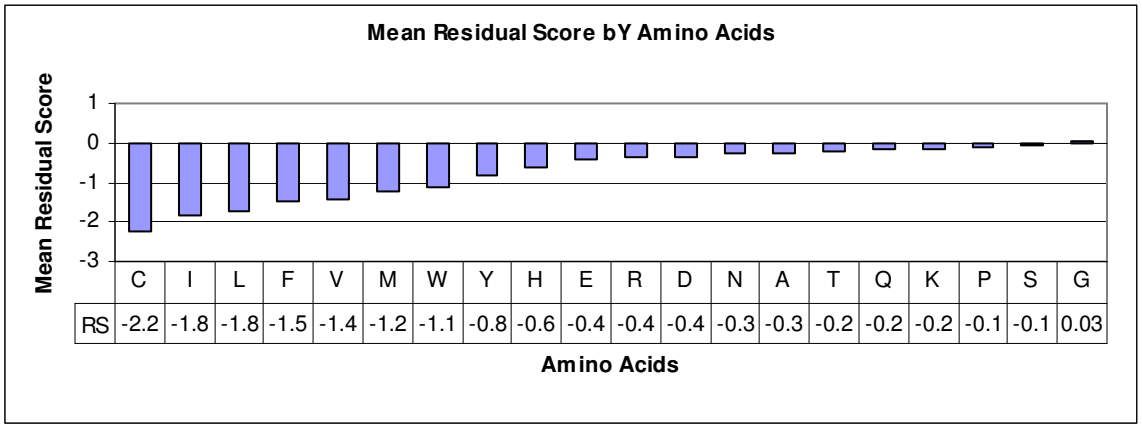
Table 5.5 Substitutions Types with Lowest and Highest Mean RS. MRS is mean residual scores for each substitution type. Count is the total number of hypothetical substitutions for each substitution type.

Level of RS	Substitution Type	MRS	CNT
Lowest 15	CK	-3.9	1971
	CE	-3.63	1971
	CD	-3.4	1971
	IK	-3.3	7791
	ID	-3.27	7791
	LK	-3.2	12870
	LD	-3.13	12870
	IE	-3.12	7791
	CN	-3.02	1971
	LE	-2.9	12870
	IN	-2.9	7791
	FK	-2.89	5798
	VK	-2.84	10199
	LN	-2.83	12870
	CQ	-2.78	1971
Highest 15	GM	0.34	11179
	GH	0.35	11179
	TW	0.36	8160
	AV	0.38	12237
	QW	0.38	5670
	GY	0.39	11179
	GF	0.39	11179
	AW	0.4	12237
	PW	0.47	6737
	SW	0.49	8623
	AI	0.51	12237
	AM	0.53	12237
	GW	0.54	11179
	AL	0.56	12237
	AF	0.61	12237

Each of twenty amino acids may be mutated to other amino acids and thus there are 20 different types of single-point mutations based on a difference of the wild-type amino acids. The mean residual scores by these 20 types of mutations were calculated

and sorted and a graph was plotted based on this sorted data (Figure 5.2A) and frequency of amino acids that appears in these 700 proteins are also calculated and depicted in Figure 5.2B. Our data indicates that cysteine has a lowest score of -2.2 while proline has a biggest score of 0.03. It is expected because the disulfide bonds may be impaired when cysteine is substituted by other amino acids.

A.



B.

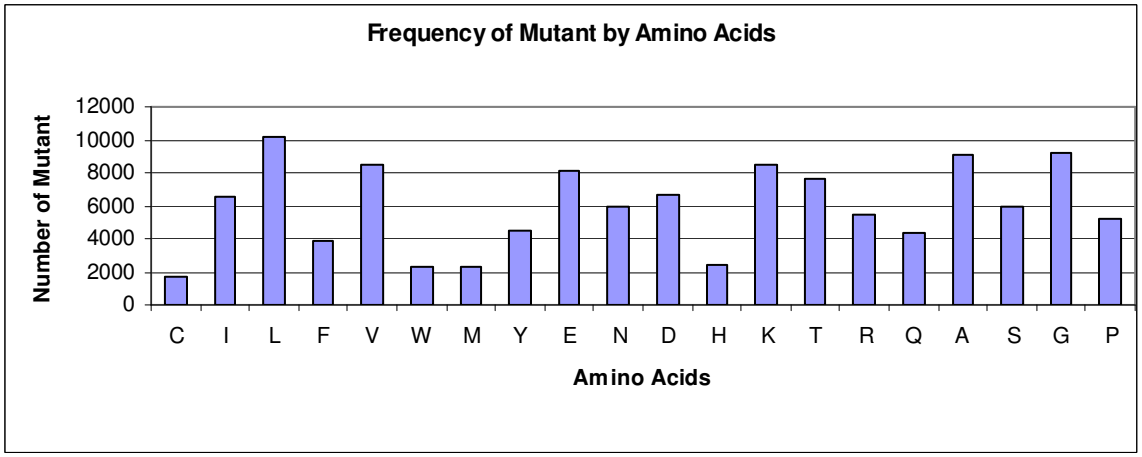


Figure 5.2 Mean residual scores and frequency of amino acids. Total number of mutants is 2,750,782 from 700 proteins.

5.4.3 Correlation RS Matrix with PAM, BLOSUM, and Wolfson-Nussinov Matrices.

Motivated by previous previously established strong correlation between the RS and conservatism of substitutions, coupled with the fact that PAM matrix and BLOSUM matrix are related to evolutionary distance, we decided to investigate any possible correlation between our newly developed RS matrix and PAM matrices, BLOSUM matrices, or Wolfson-Nussinov matrix. The correlation coefficient and its associated probability between the above described matrices were computed using simple Mantel test implemented by the zt program (Bonnet and Van de Peer 2002, Mantel 1967, Mantel and Valand 1970). Two totally unrelated matrices will have a correlation coefficient value of 0 while two highly correlated matrices have a correlation value close to 1. The comparison results are shown in Table 5.6. The observed correlation between the new matrix and PAM matrices, BLOSUM matrices, or Wolfson-Nussinov matrix ranges from 0.50 to 0.56 with an associated p value of less than 0.0001 (Table 5.6). The observed correlation value between the new matrix and Wolfson-Nussinov matrix is 0.61, which is better than the correlation between the new matrix and PAM matrices or BLOSUM matrices. Such result is expected since Wolfson-Nussinov matrix is a structural-based substitution matrix derived from the frequencies of interchanges of spatially adjacent residue pairs in conserved 3D environments in globally dissimilar protein structures (Azarya-Sprinzak et al 1997). In order to build the negative control matrix, we randomly shuffled all elements in the RS matrix. The correlation coefficients between the negative control matrix and BLOSUM/PAM/RS/Wolfson-Nussinov matrices range between 0.019 and 0.072 with associated probabilities of 0.15 to 0.38, demonstrating they are not related

to each other. PAM40 were also compared with PAM matrices and BLOSUM matrices to serve as positive control and their correlation coefficients range from 0.85 to 0.95 with an associated p value of less than 0.0001. Therefore, our data demonstrate that there is good correlation between RS matrix and PAM matrices or BLOSUM matrices.

Table 5.6 Correlation Coefficients and Associated Probabilities between RS-Based Matrix, PAM Matrices, and BLOSUM Matrices. RS is the new matrix based on RS. RS Control is the control matrix created by randomly shuffling all elements of the matrix RS. P value is the associated probability for the observed correlation coefficient.

	RS (p value)	RS Control (p value)	PAM40 (p value)
RS	1 (<0.0001)	0.019(0.38)	
PAM40	0.5(<0.0001)	0.050(0.23)	1(<0.0001)
PAM80	0.52(<0.0001)	0.072(0.15)	0.95(<0.0001)
PAM120	0.53(<0.0001)	0.020(0.39)	0.98(<0.0001)
PAM250	0.54(<0.0001)	0.039(0.29)	0.89(<0.0001)
BLOSUM62	0.54(<0.0001)	0.056(0.22)	0.85(<0.0001)
BLOSUM80	0.56(<0.0001)	0.056(0.22)	0.86(<0.0001)
Wolfson-Nussinov	0.61(<0.0001)	0.035(0.31)	0.59(<0.0001)

5.4.4 Comparison of p53 Activity RF Models using Matrix Score, RS and RSP.

The vectors containing RS matrix score, RS, RSP, PAM120 matrix score, BLOSUM62 matrix score, and Wolfson-Nussinov matrix score were created for 932 single-point p53 mutants. The RF algorithm was used for supervised classification. The AUCs were calculated to test the robustness of the predictions. For the promoter WAF1, the prediction accuracies are 74%, 75%, 74%, 73%, 75%, and 80% with corresponding AUC of 0.692, 0.697, 0.692, 0.682, 0.704, and 0.737 using RS matrix score, PAM120 matrix score, BLOSUM62 matrix score, Wolfson-Nussinov matrix score, actual RS, and RSP respectively (Table 5.7). The data indicate that the performance of inferential model

based on the RS matrix score, PAM120 matrix score, BLOSUM62 matrix score, and Wolfson-Nussinov matrix score are very similar. Furthermore, the inferential model using RSP is better than those using scores of the new RS matrix, PAM120, BLOSUM62, and Wolfson-Nussinov matrix, actual RS. The models based on actual RS are only slightly better than that of scores of the new RS matrix, PAM120, BLOSUM62, and Wolfson-Nussinov matrix. Similar results are obtained in other three promoters (Table 5.7).

Table 5.7 Comparison of Prediction Accuracy and AUC among p53 Activity Models using Matrix Score, RS and RSP. The results are based on a 10-fold cross-validation using RF with default parameters. Prediction accuracy and AUC were computed using scores of the new matrix, PAM120, BLOSUM62, and Wolfson-Nussinov matrix, actual RS, and RSP in 932 p53 mutants for four p53-binding sequences in the two-class system. RE is the p53-binding sequences from four promoters. SE is the conservative estimate for the standard error of the AUC.

RE	Model Basis	AUC	SE	Accuracy
WAF1	Matrix Score	0.692	0.0225	74%
	PAM120	0.697	0.0224	75%
	BLOSUM62	0.692	0.0225	74%
	Wolfson-Nussinov	0.682	0.0227	73%
	RS	0.704	0.0223	75%
	RSP	0.737	0.0217	80%
NOXA	Matrix Score	0.694	0.0171	63%
	PAM120	0.698	0.017	64%
	BLOSUM62	0.694	0.0171	64%
	Wolfson-Nussinov	0.701	0.017	63%
	RS	0.726	0.0164	67%
	RSP	0.768	0.0154	71%
AIP	Matrix Score	0.653	0.019	63%
	PAM120	0.668	0.0188	63%
	BLOSUM62	0.661	0.0189	63%
	Wolfson-Nussinov	0.666	0.0188	62%
	RS	0.685	0.0185	64%
	RSP	0.761	0.017	72%
BAX	Matrix Score	0.685	0.0194	65%
	PAM120	0.67	0.0196	65%
	BLOSUM62	0.665	0.0196	65%
	Wolfson-Nussinov	0.655	0.0198	65%
	RS	0.713	0.0189	67%
	RSP	0.795	0.0169	75%

5.5 Conclusions

The point-mutation can be categorized as conservative or non-conservative based on the amino-acid groups classified according Dayhoff's substitution matrix. We first demonstrated that the mean of residual score in conservative class is higher than that in non-conservative class in 685 of 700 protein chains (97.9%) studied. Such difference is statistically significant in 664 protein chains using two sample student t test (p value <0.05) and thus a strong correlation is revealed between the residual scores and conservatism of substitutions in proteins. Second, we have developed a novel statistical substitution matrix using the mean residual scores from 2,750,782 hypothetical mutations in 700 proteins. Third, the correlation coefficients of this novel matrix and PAM40, PAM 80, PAM 120, PAM 250, BLOSUM62, and BLOSUM80 range from 5.0 to 5.6 using simple Mantel test and thus a good correlation is observed. Finally, we have shown that the RF models using RS matrix score are comparable to that of using actual RS in single-point p53 mutants. Our data suggest that this novel matrix may be served as one of attributes for predicting functional effects of mutations and may be applied to other areas of statistical geometry studies on the proteins with unknown three-dimensional structures.

6. Future Directions

In this dissertation, the supervised machine learning models for stability and activity of protein mutants are trained with only residual score profiles. One of the ways to improve the results would be to incorporate other features into the training vectors in order to develop more accurate inferential models. The other features may include physicochemical properties of the amino acids and evolutionary properties derived from sequence alignments of homologous proteins.

Studies from this dissertation have demonstrated the structure-function correlation in p53 and SERCA1. Other researchers have previous shown such correlation in a limited number of other proteins. It would be desirable to expand such study to more proteins with known three-dimensional structures and available experimentally measured activities from a large number of single point mutants. Such expanded studies will further test our hypothesis that the structure-function correlation is applicable to all proteins.

The statistical geometry approach based on Delaunay tessellation of protein structure has been applied to a number of areas of protein studies. However, this approach requires the tertiary structure of the protein studied has been solved, limiting the number of proteins that can be analyzed. The newly developed substitution matrix scores or 380 general mean residual scores computed from the residual scores of 2,750,782

hypothetical mutations makes it possible to study the mutants of any protein with known primary sequence information using this approach. It enables future tests of applicability of general mean residual score for the analysis of mutants of proteins with unknown tertiary structure. Furthermore, this new substitution matrix may serve as a substitution matrix for the BLAST program or any other programs where substitution matrix is needed.

7. Conclusions

By analyzing the stability change of protein mutants, we demonstrate that a strong correlation was observed between RS and the sign of dTm/ddG in the single or double-point mutants. Such correlation between RS and dTm/ddG is only applicable to the buried mutation sites, but not to the exposed sites. Furthermore, our data indicate that the machine learning models using RSP and three supervised machine learning schems (DT,RF, and SVM) can be used to accurately predict the level of stability change of single or double-point mutants. The performance of RF models is better than those of DT and SVM models. Finally, the inferential models using either dTm or ddG can make a similar prediction of change of stability in single-point mutants.

By studying the activity change of p53 and SERCA1 proteins mutants, we show that a significant structure-function correlation is observed in human p53 and rabbit SERCA1, which suggests that such structure-function correlation may be a general property of proteins. Furthermore, we demonstrate that the inferential models based on RSP in conjunction with three supervised machine learning schemes (RF, DT and SVM) can be used to accurately predict the activity change of p53 and SERCA1 mutants. This result suggests that the method is applicable to other proteins with a sufficient amount of experimentally functional data in order to derive machine learning models. Finally, an inverse correlation was observed between the sequence-structure compatibility scores and

the frequency of p53 mutants, suggesting p53 mutants with lower sequence-structure compatibility scores are more likely to occur in human tumors.

In the conservation analysis of protein mutants, a strong correlation was observed between mean residual scores and conservation of substitutions in proteins. A novel statistical substitution matrix was developed using mean residual scores of 380 types of mutations from 700 proteins. A good correlation exists between this novel matrix and PAM/BLOSUM matrices. The inferential models using the new matrix scores are comparable to that of using actual residual in predicting the activity of p53 mutants.

Overall, these above conclusions support our hypothesis that computational mutagenesis models using four-body statistical potential present a powerful approach for predicting the changes of activity and stability in protein mutants.

Appendix A

A.1 A Novel Conservation Matrix Based on Delaunay Tessellations The novel matrix score are computed based on RS of 2,750,782 hypothetical substitutions in 700 proteins. Part A is the first half of the matrix where the replaced amino acids are A, R, N, D, C, Q, E, G, H, and I. Part B is the second half of the matrix where the replaced amino acids are L,K,M,F,P,S,T,W,Y, and V.

Part A.

	A	R	N	D	C	Q	E	G	H	I
A	0	-0.41	-0.82	-1	-1.15	-0.48	-0.94	-0.69	-0.13	0.5
R	-0.42	0	-0.49	-0.67	-1.67	-0.31	-0.65	-0.62	0.05	-0.2
N	-0.51	-0.27	0	-0.2	-1.66	-0.11	-0.46	-0.3	0.04	-0.4
D	-0.55	-0.32	-0.1	0	-1.83	-0.1	-0.17	-0.31	-0.04	-0.5
C	-2.25	-2.72	-3.02	-3.4	0	-2.78	-3.63	-2.71	-1.7	-1.3
Q	-0.28	-0.13	-0.19	-0.3	-1.47	0	-0.38	-0.36	0.12	-0.1
E	-0.49	-0.29	-0.33	-0.13	-2.01	-0.16	0	-0.5	-0.17	-0.5
G	-0.1	-0.01	0.07	-0.05	-1.12	0.07	-0.26	0	0.35	0.1
H	-0.76	-0.63	-0.87	-1.03	-1.42	-0.72	-1.18	-0.9	0	-0.4
I	-1.42	-2.34	-2.9	-3.27	-2.26	-2.46	-3.12	-2.63	-1.83	0
L	-1.3	-2.14	-2.83	-3.13	-2.2	-2.27	-2.9	-2.53	-1.69	-0.1
K	-0.34	0.15	0	-0.15	-1.78	0.01	-0.15	-0.28	0.14	-0.3
M	-1.04	-1.54	-2.01	-2.29	-1.84	-1.62	-2.22	-1.85	-1.02	-0.1
F	-1.26	-1.88	-2.33	-2.66	-1.98	-1.95	-2.63	-2.14	-1.31	-0.1
P	-0.24	-0.06	-0.15	-0.29	-1.3	-0.05	-0.4	-0.27	0.23	-0
S	-0.23	-0.09	0	-0.17	-1.19	0	-0.41	-0.14	0.27	-0
T	-0.27	-0.3	-0.32	-0.52	-1.3	-0.24	-0.68	-0.4	0.06	0.1
W	-1.11	-1.31	-1.57	-1.83	-1.85	-1.33	-1.92	-1.57	-0.78	-0.4
Y	-0.81	-1.07	-1.31	-1.58	-1.54	-1.09	-1.65	-1.3	-0.56	0
V	-1.02	-1.87	-2.42	-2.76	-1.84	-1.98	-2.62	-2.14	-1.38	0.2

Part B.

	L	K	M	F	P	S	T	W	Y	V
A	0.56	-1.18	0.53	0.61	-0.47	-0.63	-0.24	0.4	0.27	0.4
R	-0.14	-0.66	0.05	0.07	-0.31	-0.49	-0.28	0.19	0.02	-0.3
N	-0.43	-0.59	-0.08	-0.03	-0.15	-0.17	-0.02	0.23	0.08	-0.5
D	-0.51	-0.63	-0.19	-0.18	-0.19	-0.25	-0.12	0.11	-0.03	-0.6
C	-1.33	-3.9	-1.25	-0.96	-2.51	-2.58	-2.28	-1.26	-1.43	-1.5
Q	-0.03	-0.63	0.18	0.21	-0.12	-0.25	-0.04	0.38	0.2	-0.2
E	-0.34	-0.64	-0.14	-0.18	-0.31	-0.47	-0.29	0.01	-0.08	-0.6
G	0.04	-0.48	0.34	0.39	0.09	0.04	0.23	0.54	0.39	-0
H	-0.31	-1.43	-0.04	0.04	-0.65	-0.79	-0.56	0.15	-0.09	-0.5
I	-0.13	-3.3	-0.4	-0.17	-2.33	-2.54	-1.92	-0.84	-0.85	-0.3
L	0	-3.2	-0.37	-0.14	-2.19	-2.45	-1.87	-0.73	-0.81	-0.3

K	-0.28	0	0	-0.03	-0.06	-0.14	0.01	0.13	0.07	-0.4
M	-0.07	-2.5	0	0.05	-1.55	-1.77	-1.32	-0.28	-0.41	-0.3
F	-0.16	-2.89	-0.27	0	-1.82	-2.04	-1.55	-0.45	-0.56	-0.4
P	0.02	-0.61	0.25	0.31	0	-0.18	0.02	0.47	0.3	-0.1
S	-0.05	-0.56	0.23	0.31	0	0	0.15	0.49	0.32	-0.1
T	0.04	-0.83	0.24	0.32	-0.22	-0.3	0	0.36	0.22	-0.1
W	-0.33	-2.22	-0.25	-0.05	-1.24	-1.44	-1.12	0	-0.31	-0.5
Y	0	-1.92	0.06	0.24	-1.01	-1.18	-0.84	0.1	0	-0.2
V	0.15	-2.84	-0.09	0.12	-1.85	-2.06	-1.47	-0.46	0.51	0

A.2 A Full List of Correlation of RS with Conservatism of Substitutions in 664 Proteins. 664 protein chains have higher conserved mean residual scores than non-conserved ones and their difference is statistically significant (p value <0.05). MRS_C is conserved mean residual score. CNT_C is the count of conserved mutations. MRS_NC is non-conserved mean residual score. CNT_NC is the count of non-conserved mutations. P_VALUE is the result of two-sample t test for independent samples with unequal variances.

PDB	MRS_C	CNT_C	MRS_NC	CNT_NC	P_VALUE	LENGTH
1M2DA	-0.27	303	-1.35	1616	<2.2e-16	101
2I4AA	-0.03	331	-0.83	1702	<2.2e-16	107
1M55A	-0.27	581	-0.76	3086	<2.2e-16	193
1TUAA	-0.51	569	-1.15	3003	<2.2e-16	188
1H5QA	-0.29	822	-0.7	4118	1.44E-14	260
1UF5A	-0.25	918	-0.64	4839	3.75E-14	303
1O9RA	-0.16	507	-0.67	2571	<2.2e-16	162
2GMYA	-0.1	445	-0.62	2348	5.19E-15	147
1VKIA	-0.08	515	-0.92	2620	<2.2e-16	165
1ZHVA	-0.27	414	-0.97	2132	<2.2e-16	134
2AGYA	-0.04	1110	-0.82	5730	<2.2e-16	360
1QLWA	-0.13	1003	-0.71	5039	<2.2e-16	318
2COVD	-0.09	262	-0.55	1486	7.20E-10	92
1E85A	-0.05	387	-0.59	1969	<2.2e-16	124
1URSA	-0.24	1162	-0.87	5792	<2.2e-16	366
1OI6A	-0.12	635	-0.43	3203	2.55E-08	202
1OBOA	-0.13	519	-0.85	2692	<2.2e-16	169
2BMWA	-0.24	902	-0.9	4703	<2.2e-16	295
2ERBA	-0.26	344	-0.93	1993	<2.2e-16	123
1R5RA	-0.13	347	-1.19	1876	<2.2e-16	117
1T1DA	-0.12	289	-0.67	1611	<2.2e-16	100
1T6CA	-0.35	913	-1.11	4901	<2.2e-16	306
1RP0A	-0.16	884	-0.77	4398	<2.2e-16	278
1XMTA	-0.16	281	-0.38	1524	0.001647	95
2H8GA	-0.14	776	-0.72	3898	<2.2e-16	246
1I0RA	-0.26	486	-0.84	2573	<2.2e-16	161
1JNRA	-0.35	1928	-0.81	10270	<2.2e-16	642

1Q4UA	-0.35	440	-0.58	2220	0.0009146	140
1UGXA	-0.06	415	-0.67	2112	<2.2e-16	133
1K7CA	-0.37	748	-0.86	3679	<2.2e-16	233
1NKGA	-0.26	1649	-0.89	8003	<2.2e-16	508
1Y4WA	-0.09	1643	-0.59	8180	<2.2e-16	517
2IBAA	-0.03	893	-0.47	4712	<2.2e-16	295
1UWCA	-0.1	820	-0.55	4139	<2.2e-16	261
1Y43B	-0.09	553	-0.39	2696	2.42E-06	171
2GUYA	-0.13	1480	-0.84	7564	<2.2e-16	476
1EB6A	-0.24	558	-0.52	2805	1.06E-05	177
2GECA	-0.04	425	-0.51	2197	2.21E-14	138
1M1NA	-0.21	1421	-0.55	7642	<2.2e-16	477
1M1NB	-0.15	1561	-0.79	8357	<2.2e-16	522
1H4GA	-0.24	635	-0.69	3260	<2.2e-16	205
1W23A	-0.16	1111	-0.86	5729	<2.2e-16	360
1A2PA	-0.23	332	-0.39	1720	0.004726	108
1WCKA	-0.16	463	-0.93	2121	<2.2e-16	136
1YKUA	-0.2	369	-0.89	2139	<2.2e-16	132
1C9OA	-0.22	200	-0.78	1054	8.70E-11	66
2B5AA	-0.2	226	-0.53	1237	0.0001485	77
1QGIA	-0.12	792	-0.75	4129	<2.2e-16	259
4UBPA	-0.21	310	-0.47	1590	0.0001707	100
4UBPB	-0.25	378	-0.66	1940	5.83E-11	122
1V5DA	-0.15	1190	-0.57	6144	<2.2e-16	386
2AHFA	-0.28	1145	-0.64	6018	<2.2e-16	377
1QW9A	-0.18	1485	-0.74	7958	<2.2e-16	497
1U84A	-0.36	247	-0.6	1292	0.01567	81
2HHVA	-0.36	1751	-0.97	9269	<2.2e-16	580
1F7LA	-0.24	355	-0.59	1887	8.65E-07	118
1ISPA	-0.17	1126	-0.95	5676	<2.2e-16	358
1KQPA	-0.15	837	-0.72	4312	<2.2e-16	271
1L7AA	-0.21	976	-0.64	5066	<2.2e-16	318
1NC5A	-0.15	1078	-0.67	5819	<2.2e-16	363
1NG6A	-0.22	449	-0.91	2363	<2.2e-16	148
1OYGA	-0.21	1351	-1.01	7009	<2.2e-16	440
1UV4A	-0.24	918	-0.82	4611	<2.2e-16	291
1WPUA	-0.24	456	-0.9	2337	<2.2e-16	147
1Z3EA	-0.16	346	-0.99	1896	<2.2e-16	118
1Z3EB	-0.15	197	-1.01	1076	<2.2e-16	67
2BKXA	-0.33	749	-1.02	3849	<2.2e-16	242
2OHWA	-0.19	385	-0.8	2047	<2.2e-16	128
2TPSA	-0.29	717	-1.05	3577	<2.2e-16	226
1IQZA	-0.04	490	-0.79	2588	<2.2e-16	162
2IBLA	0.05	343	-0.22	1709	2.24E-06	108
1C5EA	-0.14	320	-0.88	1485	<2.2e-16	95
1B5EA	-0.14	713	-0.59	3866	<2.2e-16	241
1C1KA	-0.23	609	-0.98	3514	<2.2e-16	217
1KAFA	-0.24	315	-0.76	1737	2.29E-13	108
1MK0A	-0.11	273	-0.28	1570	0.008657	97

1OCYA	0.01	612	-0.26	3150	3.44E-08	198
2D73A	-0.14	2162	-0.75	11461	<2.2e-16	717
2H7ZB	-0.01	205	-1.61	1258	<2.2e-16	77
1YU0A	-0.19	1228	-0.65	5916	<2.2e-16	376
1ZZ1A	-0.29	1163	-0.74	5810	<2.2e-16	367
1B8OA	-0.15	855	-0.81	4465	<2.2e-16	280
1D4OA	-0.2	553	-0.95	2810	<2.2e-16	177
1ES9A	-0.28	638	-1.09	3390	<2.2e-16	212
1KT6A	-0.25	506	-0.72	2819	<2.2e-16	175
1O7QA	-0.09	830	-0.61	4623	<2.2e-16	287
1OMRA	-0.31	600	-1.15	3219	<2.2e-16	201
1PIDA	-0.06	53	-0.77	346	0.000218	21
1T61A	-0.13	673	-0.58	3564	<2.2e-16	223
2BJIA	-0.22	847	-0.81	4359	<2.2e-16	274
1JFUA	-0.25	551	-1	2793	<2.2e-16	176
1KNGA	-0.13	447	-0.63	2289	2.35E-14	144
1XT5A	0.02	421	-0.4	2144	1.88E-13	135
1F94A	-0.15	160	-1.46	1037	<2.2e-16	63
1Y37A	-0.25	887	-0.79	4699	<2.2e-16	294
1SZHA	-0.05	396	-1.03	2397	<2.2e-16	147
1LLFA	-0.11	1697	-0.82	8449	<2.2e-16	534
2J2JA	-0.1	568	-0.59	2890	<2.2e-16	182
1SU8A	-0.32	1977	-1.12	10050	<2.2e-16	633
2IWAA	-0.06	743	-0.61	4083	<2.2e-16	254
1GXMA	-0.28	1013	-0.79	5143	<2.2e-16	324
1UXZA	-0.14	433	-0.77	2056	<2.2e-16	131
1KR7A	0.02	337	-0.2	1734	0.0007402	109
1E2WA	-0.21	781	-0.91	3988	<2.2e-16	251
1T9IA	-0.25	456	-1.02	2451	<2.2e-16	153
1VMHA	-0.1	392	-0.7	2059	4.89E-14	129
1Y7BA	-0.17	1583	-0.65	8563	<2.2e-16	534
1U8VA	-0.11	1470	-0.5	7840	<2.2e-16	490
1YBIA	-0.26	856	-0.87	4540	<2.2e-16	284
1R45A	-0.04	610	-0.72	3209	<2.2e-16	201
1CCWA	-0.2	414	-0.69	2189	1.21E-11	137
1CCWB	-0.23	1472	-0.51	7705	2.23E-15	483
1GUTA	-0.07	216	-0.65	1057	1.15E-09	67
1OD3A	-0.27	432	-0.83	2057	<2.2e-16	131
2FGQX	-0.04	1058	-0.47	5212	<2.2e-16	330
1IS3A	-0.12	814	-0.81	4278	<2.2e-16	268
1WZDA	-0.13	635	-0.54	3336	3.68E-15	209
1O8XA	-0.11	430	-0.99	2287	<2.2e-16	143
2CZQA	-0.04	666	-0.42	3229	1.21E-12	205
1WS8A	-0.11	314	-0.61	1662	1.09E-11	104
2AQPA	-0.31	490	-1.1	2455	<2.2e-16	155
1KQWA	-0.1	379	-0.35	2167	7.38E-06	134
2C2UA	0.01	544	-0.39	2838	2.98E-14	178
1VZIA	-0.31	358	-0.66	2017	1.38E-05	125
2DSXA	0.11	151	-0.88	837	<2.2e-16	52

1FLMA	-0.22	386	-0.92	1932	<2.2e-16	122
1J0PA	-0.33	301	-1.3	1732	3.95E-15	107
1UCRA	-0.08	225	-0.8	1181	<2.2e-16	74
2AVKA	-0.23	383	-0.81	2144	<2.2e-16	133
1RL0A	-0.19	776	-0.84	4069	<2.2e-16	255
1NLQA	-0.08	323	-0.94	1672	<2.2e-16	105
1PFBA	-0.12	153	-0.42	892	0.0007458	55
1SXRA	-0.24	523	-0.65	2764	8.31E-09	173
1URRA	0.08	282	-0.08	1561	0.02432	97
1ZV1A	0.04	166	-0.56	955	2.12E-10	59
2BK9A	-0.1	473	-0.77	2434	<2.2e-16	153
1H2CA	-0.13	395	-0.87	1961	<2.2e-16	124
1Z2NX	-0.1	915	-0.84	4994	<2.2e-16	311
1O82A	-0.46	226	-1.25	1104	2.65E-11	70
1IT2A	-0.21	429	-0.91	2345	<2.2e-16	146
1N8KA	-0.13	1144	-0.64	5962	<2.2e-16	374
1K7IA	-0.17	1477	-0.77	7301	<2.2e-16	462
1NOFA	-0.16	1213	-0.62	6064	<2.2e-16	383
1PE9A	-0.18	1148	-0.91	5711	<2.2e-16	361
1RU4A	-0.22	1269	-0.85	6331	<2.2e-16	400
1C4QA	-0.37	212	-1.06	1099	3.45E-15	69
1DJ0A	-0.15	800	-0.64	4216	<2.2e-16	264
1E5KA	-0.21	577	-0.93	2995	<2.2e-16	188
1E6UA	-0.13	960	-0.59	5025	<2.2e-16	315
1EJ0A	-0.19	550	-0.92	2870	<2.2e-16	180
1EVLA	-0.17	1171	-0.79	6448	<2.2e-16	401
1EW4A	-0.18	317	-0.72	1697	4.78E-12	106
1F46A	-0.17	430	-0.77	2211	<2.2e-16	139
1FM0D	-0.43	258	-1.04	1281	6.02E-12	81
1FR2A	-0.11	258	-0.7	1319	2.45E-14	83
1FR2B	-0.26	394	-0.9	2095	<2.2e-16	131
1G6HA	-0.14	768	-1.13	4058	<2.2e-16	254
1G6SA	-0.41	1346	-1.09	6767	<2.2e-16	427
1GS5A	-0.38	827	-1.02	4075	<2.2e-16	258
1HNJA	-0.24	1011	-0.79	5012	<2.2e-16	317
1HW1A	-0.14	693	-0.7	3601	<2.2e-16	226
1HZTA	-0.08	462	-0.24	2445	0.005071	153
1I52A	-0.28	694	-0.74	3581	4.43E-14	225
1J2RA	-0.19	590	-0.79	2982	<2.2e-16	188
1JCDA	-0.55	169	-1.48	781	<2.2e-16	50
1JHGA	-0.26	310	-0.89	1609	9.81E-16	101
1JKEA	-0.16	449	-1.07	2306	<2.2e-16	145
1JKXA	-0.17	650	-0.86	3321	<2.2e-16	209
1JL1A	-0.31	468	-0.87	2420	2.83E-15	152
1JZ8A	-0.2	3079	-0.64	16130	<2.2e-16	1011
1KQFC	-0.39	626	-1.49	3478	<2.2e-16	216
1MXRA	-0.23	1008	-0.64	5433	<2.2e-16	339
1NYTA	-0.29	850	-0.89	4299	<2.2e-16	271
1P1XA	-0.3	779	-0.81	3971	<2.2e-16	250

1Q5YA	-0.49	243	-0.73	1353	0.04251	84
1Q6OA	-0.29	655	-0.73	3392	2.48E-14	213
1QTWA	-0.26	869	-0.58	4546	5.01E-10	285
1R9LA	-0.24	966	-0.84	4905	<2.2e-16	309
1RA0A	-0.23	1307	-0.68	6730	<2.2e-16	423
1RKQA	-0.24	839	-0.83	4310	<2.2e-16	271
1RYAA	-0.06	486	-0.54	2554	<2.2e-16	160
1S3CA	-0.08	426	-1	2196	<2.2e-16	138
1SX5A	-0.04	720	-0.67	3916	<2.2e-16	244
1T4BA	-0.21	1154	-0.76	5819	<2.2e-16	367
1TKEA	-0.1	661	-0.44	3595	3.30E-10	224
1TT8A	-0.35	513	-0.83	2603	3.28E-15	164
1U07A	-0.1	278	-0.56	1432	2.14E-12	90
1USGA	-0.14	1089	-0.84	5466	<2.2e-16	345
1VH5A	0	413	-0.34	2190	5.02E-08	137
1VHTA	-0.29	650	-0.81	3283	<2.2e-16	207
1WBHA	-0.29	683	-0.85	3364	<2.2e-16	213
1XEOA	-0.17	489	-0.92	2646	<2.2e-16	165
1XG4A	-0.21	909	-0.63	4544	<2.2e-16	287
1XS0A	-0.11	403	-0.78	2029	<2.2e-16	128
1YT3A	-0.2	1146	-0.61	5979	<2.2e-16	375
256BA	-0.33	326	-0.76	1688	5.14E-08	106
2AXWA	-0.16	406	-0.77	2140	7.88E-16	134
2B82A	-0.2	662	-0.77	3347	<2.2e-16	211
2DQ6A	-0.21	2628	-0.85	13807	<2.2e-16	865
2DS5A	0.1	113	-0.8	704	8.45E-12	43
2DY0A	-0.13	572	-0.8	2886	<2.2e-16	182
2EX2A	-0.14	1437	-0.77	7227	<2.2e-16	456
2FHZA	-0.12	321	-0.37	1693	0.0006668	106
2FHZB	-0.2	292	-0.91	1475	<2.2e-16	93
2G7OA	-0.03	203	-0.34	1089	8.86E-06	68
2IW1A	-0.23	1112	-0.82	5918	<2.2e-16	370
1F7DA	0	362	-0.94	1880	<2.2e-16	118
1IO0A	-0.51	514	-1.1	2640	<2.2e-16	166
1JB3A	-0.23	372	-0.77	2041	4.42E-13	127
1W4SA	0.01	407	-0.55	2367	8.18E-16	146
1WBIA	0.07	379	-0.14	1958	0.0004153	123
1YIIA	-0.11	962	-0.86	5042	<2.2e-16	316
2BKMA	-0.2	387	-0.53	2045	9.06E-08	128
1OS6A	-0.5	190	-1.4	1159	2.10E-12	71
2CZSA	-0.25	191	-0.58	1139	0.008981	70
1JF3A	0.03	474	-0.48	2319	<2.2e-16	147
1WDPA	-0.13	1489	-0.7	7878	<2.2e-16	493
1G12A	-0.2	532	-0.73	2641	6.23E-16	167
1JO0A	-0.17	300	-1.02	1543	<2.2e-16	97
1NNFA	-0.34	963	-1.06	4889	<2.2e-16	308
1OU8A	-0.17	327	-0.87	1687	2.28E-16	106
2GKEA	-0.12	827	-0.68	4379	<2.2e-16	274
2LISA	-0.08	376	-0.26	2113	0.003933	131

1TJOA	-0.09	561	-0.33	2859	1.28E-05	180
2CC6A	-0.21	202	-0.49	1014	0.0007513	64
1WZAA	-0.2	1451	-0.84	7821	<2.2e-16	488
2EWHA	-0.36	297	-0.63	1451	0.00742	92
1ZKEA	-0.21	242	-0.78	1297	3.37E-13	81
2FNUA	-0.21	1112	-0.56	5975	6.52E-15	373
2HALA	-0.2	655	-0.77	3373	<2.2e-16	212
1QJ4A	-0.34	760	-0.89	4104	<2.2e-16	256
1BKRA	-0.15	316	-0.57	1736	1.06E-08	108
1CY5A	-0.37	272	-0.96	1476	3.85E-13	92
1D4TA	0.04	313	-0.78	1663	<2.2e-16	104
1D7PM	-0.07	484	-0.44	2537	1.91E-09	159
1DK8A	-0.09	440	-0.46	2353	1.24E-09	147
1EAZA	-0.01	300	-0.32	1657	1.71E-05	103
1EK6A	-0.11	1054	-0.64	5520	<2.2e-16	346
1ELKA	-0.25	461	-1.09	2446	<2.2e-16	153
1F86A	-0.06	360	-0.62	1825	4.48E-15	115
1FL0A	-0.2	497	-0.79	2619	<2.2e-16	164
1G1TA	-0.08	448	-1.04	2535	<2.2e-16	157
1G8QA	-0.3	264	-1.03	1446	1.48E-15	90
1GP0A	-0.16	399	-0.71	2128	5.60E-16	133
1GVJA	-0.36	409	-0.78	2270	9.33E-10	141
1HD2A	-0.1	502	-0.66	2557	<2.2e-16	161
1I27A	-0.25	220	-0.63	1167	1.56E-06	73
1I2TA	-0.6	193	-1.1	966	3.62E-05	61
1I71A	-0.38	236	-0.93	1322	3.36E-07	82
1IFRA	-0.06	352	-0.56	1795	8.07E-14	113
1IRDB	-0.1	886	-0.66	4662	<2.2e-16	292
1J2JB	-0.35	119	-0.97	660	3.43E-08	41
1JHJA	-0.17	490	-0.89	2569	<2.2e-16	161
1K0MA	-0.13	709	-0.63	3756	<2.2e-16	235
1K3YA	-0.28	654	-1.02	3545	<2.2e-16	221
1KMTA	-0.15	417	-0.69	2205	<2.2e-16	138
1L9LA	-0.24	208	-0.74	1198	2.17E-08	74
1LKKA	-0.09	315	-0.56	1680	4.98E-11	105
1LQVA	-0.26	517	-0.57	2751	1.43E-07	172
1LUGA	-0.26	786	-1.02	4135	<2.2e-16	259
1M9ZA	-0.22	283	-1.42	1712	<2.2e-16	105
1MF7A	-0.14	578	-1.01	3108	<2.2e-16	194
1MFMA	-0.25	485	-1.21	2422	<2.2e-16	153
1MG4A	-0.21	298	-0.6	1621	2.88E-08	101
1MKKA	0.08	262	-0.38	1505	1.05E-08	93
1OQJA	0.07	250	-0.37	1460	1.11E-06	90
1OZ2A	-0.25	967	-0.5	5189	6.72E-09	324
1P4OA	-0.09	932	-0.51	4920	<2.2e-16	308
1PA7A	-0.04	376	-0.35	2094	3.60E-06	130
1PSRA	-0.18	296	-1.38	1604	<2.2e-16	100
1Q7LA	-0.21	580	-0.5	3068	1.88E-07	192
1Q7LB	-0.19	274	-0.39	1398	0.00288	88

1Q92A	-0.05	576	-0.63	3129	<2.2e-16	195
1QDDA	-0.07	428	-0.64	2308	1.86E-15	144
1R29A	-0.13	361	-0.86	1957	<2.2e-16	122
1S1DA	-0.15	976	-0.95	5047	<2.2e-16	317
1S95A	-0.14	960	-0.79	5196	<2.2e-16	324
1SHUX	-0.12	554	-0.94	2885	<2.2e-16	181
1SZ7A	-0.16	475	-0.6	2546	2.25E-11	159
1T3YA	-0.13	390	-0.58	2099	1.58E-11	131
1U2HA	-0.02	285	-0.33	1539	3.13E-05	96
1UNQA	-0.18	339	-0.76	1865	3.58E-16	116
1UPQA	-0.21	318	-0.59	1715	4.81E-08	107
1UYLA	-0.21	642	-0.88	3291	<2.2e-16	207
1UZ3A	-0.35	313	-0.65	1625	0.0001425	102
1W1HA	-0.09	427	-0.69	2366	<2.2e-16	147
1WLZA	0	247	-0.38	1368	5.86E-08	85
1WM3A	0.12	213	-0.28	1155	9.09E-08	72
1WMHA	-0.04	238	-0.33	1339	0.0005281	83
1WMHB	-0.09	253	-0.52	1305	2.61E-06	82
1WPAA	-0.25	305	-0.57	1728	1.21E-05	107
1WQJB	0.01	221	-1.04	1299	<2.2e-16	80
1WU9A	-0.17	172	-0.6	949	4.19E-07	59
1XPCA	-0.13	732	-0.62	3923	6.06E-16	245
1XU9A	-0.16	829	-0.75	4282	<2.2e-16	269
1Y93A	-0.2	487	-0.58	2515	3.93E-08	158
1YPQA	-0.2	385	-0.78	2104	3.72E-16	131
1ZEDA	-0.25	1508	-0.8	7631	<2.2e-16	481
1ZKKA	-0.07	469	-0.34	2571	3.51E-06	160
1ZMMA	0.06	76	-1.48	513	6.04E-15	31
1ZZWA	-0.02	437	-0.73	2356	<2.2e-16	147
2AEBA	-0.2	983	-0.79	4983	<2.2e-16	314
2AKZA	-0.06	1342	-0.72	6923	<2.2e-16	435
2ASKA	-0.16	301	-1.26	1618	<2.2e-16	101
2AWGA	-0.06	373	-0.26	1869	0.003161	118
2AXIA	-0.12	265	-1.16	1483	<2.2e-16	92
2BKFA	-0.05	262	-0.64	1353	4.42E-15	85
2BNUA	-0.12	636	-0.5	3221	1.17E-13	203
2BNUB	-0.05	742	-0.5	3837	<2.2e-16	241
2BO9B	-0.11	652	-0.58	3471	<2.2e-16	217
2C3NA	-0.14	718	-0.64	3823	<2.2e-16	239
2C4JA	-0.06	632	-0.42	3491	1.09E-12	217
2CB8A	-0.24	256	-0.49	1378	0.0003015	86
2CBZA	-0.15	719	-1.08	3651	<2.2e-16	230
2CCQA	-0.13	302	-0.64	1579	3.27E-11	99
2CKKA	-0.3	364	-0.78	1916	1.59E-10	120
2CVDA	-0.06	584	-0.53	3178	<2.2e-16	198
2DKOA	-0.16	428	-0.8	2346	<2.2e-16	146
2DKOB	0.15	298	-0.16	1659	6.26E-05	103
2ERFA	-0.2	645	-1.08	3326	<2.2e-16	209
2FCWA	-0.13	309	-0.5	1705	7.72E-07	106

2FCWB	-0.04	211	-0.75	1271	1.54E-09	78
2FRGP	0.04	324	-0.32	1690	3.25E-07	106
2G30A	-0.1	705	-0.66	3722	<2.2e-16	233
2GRRB	-0.24	490	-1.01	2493	<2.2e-16	157
2H14A	-0.1	908	-0.53	4849	<2.2e-16	303
2H2BA	-0.13	333	-0.73	1700	<2.2e-16	107
2H3LA	-0.32	323	-0.99	1634	<2.2e-16	103
2H6FA	-0.24	910	-0.73	5075	<2.2e-16	315
2H6FB	-0.25	1224	-0.54	6566	9.46E-10	410
2HQXA	0.1	274	-0.12	1436	0.0005887	90
2HXMA	-0.12	660	-0.82	3577	<2.2e-16	223
2I53A	-0.17	727	-0.6	4099	6.89E-16	254
2ICCA	-0.1	357	-0.37	1885	3.59E-06	118
2IZXA	-0.27	120	-0.55	621	0.00833	39
2NLSA	-0.01	93	-1.41	591	2.79E-12	36
2NMLA	-0.06	289	-0.73	1611	<2.2e-16	100
2NVHA	-0.17	457	-0.81	2431	<2.2e-16	152
2O3SA	-0.16	723	-0.82	4065	<2.2e-16	252
2ET1A	0	644	-0.51	3175	<2.2e-16	201
2FJ8A	-0.28	316	-1.64	1964	<2.2e-16	120
1NLNA	-0.16	605	-0.62	3252	2.90E-15	203
2B9DA	-0.18	147	-1.15	841	6.97E-13	52
1GMUA	-0.18	425	-0.43	2197	0.0002949	138
1T92A	0.05	343	-0.44	1709	1.36E-12	108
1O9IA	-0.15	829	-0.65	4225	<2.2e-16	266
2EZ9A	-0.28	1833	-0.83	9282	<2.2e-16	585
1O08A	-0.25	686	-1.05	3513	<2.2e-16	221
2IU5A	-0.32	510	-0.84	2891	3.71E-15	179
1M15A	-0.16	1073	-0.88	5691	<2.2e-16	356
1UCSA	-0.12	207	-1.04	1009	<2.2e-16	64
1IDPA	-0.06	431	-0.44	2362	5.40E-12	147
1K4IA	-0.27	673	-0.82	3431	<2.2e-16	216
2CFEA	-0.04	500	-0.62	2578	<2.2e-16	162
1N8VA	-0.19	279	-1.03	1640	<2.2e-16	101
1FP2A	-0.17	1024	-0.79	5531	<2.2e-16	345
2OPCA	0.07	347	-0.35	1838	3.91E-09	115
1R0RI	-0.03	141	-0.76	828	1.40E-10	51
1HBNB	-0.31	1403	-0.72	6995	<2.2e-16	442
1HBNC	-0.19	742	-0.56	3951	1.09E-14	247
1G61A	-0.37	703	-1.35	3572	<2.2e-16	225
1PKHA	-0.24	538	-1.07	2920	<2.2e-16	182
1QWGA	-0.27	739	-0.95	4030	<2.2e-16	251
1NTHA	-0.18	1404	-0.48	7279	2.24E-16	457
1QREA	-0.21	673	-0.84	3317	<2.2e-16	210
1W6SA	-0.16	1843	-0.7	9462	<2.2e-16	595
1W6SB	-0.24	205	-0.83	1163	1.28E-10	72
2C8SA	-0.02	456	-0.43	2375	1.90E-11	149
1GU2A	-0.18	374	-0.71	1982	1.47E-13	124
1MCTI	0.06	67	-1.82	465	3.75E-13	28

1UCDA	-0.26	568	-0.79	3042	<2.2e-16	190
1HYOA	-0.24	1293	-0.73	6611	<2.2e-16	416
1IJYA	-0.07	340	-0.59	1978	7.98E-13	122
1KYFA	-0.13	749	-0.56	3944	<2.2e-16	247
1M4JA	-0.14	391	-0.84	2136	<2.2e-16	133
1MD6A	-0.05	471	-0.64	2455	<2.2e-16	154
1MQKH	-0.04	378	-0.64	1959	<2.2e-16	123
1MQKL	-0.12	340	-0.51	1712	9.13E-09	108
1MY7A	-0.04	318	-0.63	1715	<2.2e-16	107
1NTVA	-0.2	446	-0.86	2442	<2.2e-16	152
1O3YA	-0.2	502	-1.2	2652	<2.2e-16	166
1Q1FA	-0.14	464	-0.75	2348	<2.2e-16	148
1T1VA	-0.08	286	-0.42	1481	1.35E-06	93
1XTEA	0.02	331	-0.57	1873	8.53E-16	116
1Z0JA	-0.2	510	-0.69	2682	5.98E-11	168
1ZNDA	-0.25	462	-1.04	2521	<2.2e-16	157
2CXNA	-0.23	1697	-0.79	8886	<2.2e-16	557
2FR5A	-0.1	408	-0.46	2176	2.27E-07	136
2GDGA	-0.31	354	-0.84	1812	5.81E-12	114
2HEWF	-0.11	376	-1.21	2056	<2.2e-16	128
2IOIA	-0.15	552	-0.7	3001	<2.2e-16	187
2CYGA	-0.19	1002	-1.12	4926	<2.2e-16	312
1W5RA	-0.13	856	-0.54	4331	<2.2e-16	273
1DQZA	-0.2	892	-0.63	4428	<2.2e-16	280
1LMIA	0.05	436	-0.54	2053	<2.2e-16	131
1M4IA	-0.31	563	-0.72	2876	3.26E-10	181
1N40A	-0.18	1240	-0.8	6246	<2.2e-16	394
1NBUA	-0.26	363	-0.9	1879	7.18E-15	118
1Y0HA	-0.14	317	-0.56	1602	5.73E-08	101
2ASBA	-0.21	717	-0.69	3577	<2.2e-16	226
2C92A	-0.47	474	-0.93	2319	4.69E-09	147
2EV1A	-0.24	586	-0.77	2929	<2.2e-16	185
2JEKA	-0.17	425	-0.59	2235	1.62E-11	140
1S9RA	-0.19	1236	-0.97	6535	<2.2e-16	409
2C0HA	-0.03	1083	-0.42	5624	<2.2e-16	353
1B3AA	-0.13	191	-0.8	1082	2.04E-09	67
1DP7P	-0.01	228	-0.37	1216	5.35E-06	76
1FD3A	0.01	111	-1.45	668	1.55E-13	41
1H2WA	-0.1	2098	-0.72	11392	<2.2e-16	710
1I4UA	-0.12	530	-0.61	2909	<2.2e-16	181
1ISUA	0.06	189	-0.56	989	3.18E-13	62
1MSOA	-0.09	53	-0.85	346	3.47E-05	21
1N0QA	-0.53	281	-1.28	1448	3.16E-15	91
1NA3A	0.01	260	-0.38	1374	1.22E-09	86
1P9IA	-0.46	90	-1.1	461	1.11E-06	29
1RJUV	-0.22	84	-5.1	600	<2.2e-16	36
1T6FA	-0.56	106	-1.09	597	8.37E-05	37
1W0PA	-0.1	2398	-0.71	11909	<2.2e-16	753
2AKFA	-0.05	94	-0.51	514	4.92E-05	32

2F91B	-0.08	90	-0.85	537	7.28E-11	33
1KS8A	-0.33	1349	-0.54	6878	8.06E-07	433
1FIUA	-0.18	882	-0.59	4552	<2.2e-16	286
1J77A	-0.12	601	-0.46	3180	1.03E-09	199
1RV9A	-0.25	770	-0.53	3828	6.47E-07	242
1G5AA	-0.23	1905	-0.82	10027	<2.2e-16	628
1IOOA	-0.2	554	-0.89	3170	<2.2e-16	196
1FT5A	-0.28	608	-1.05	3401	<2.2e-16	211
1WUIS	-0.18	818	-0.84	4255	<2.2e-16	267
1N62A	-0.25	491	-0.76	2568	1.23E-12	161
1N62B	-0.14	2480	-0.49	12796	<2.2e-16	804
1PO5A	-0.24	1405	-0.88	7430	<2.2e-16	465
1WDDS	-0.11	349	-0.29	1950	0.01181	121
1UASA	-0.12	1118	-0.62	5760	<2.2e-16	362
1T7RA	-0.1	725	-0.5	4025	2.27E-12	250
2GBAA	-0.33	326	-0.75	1669	3.15E-08	105
1QKSA	-0.17	1755	-0.94	8866	<2.2e-16	559
2C1VA	-0.12	1071	-0.52	5294	<2.2e-16	335
1DLWA	-0.38	380	-0.85	1824	1.68E-10	116
1H97A	-0.17	445	-0.83	2348	<2.2e-16	147
1UOYA	-0.29	183	-1.07	1033	1.74E-12	64
1M7GA	-0.08	640	-0.6	3312	<2.2e-16	208
1GPIA	-0.26	1346	-1.12	6824	<2.2e-16	430
2AIBA	-0.1	308	-0.97	1554	<2.2e-16	98
1T2DA	-0.27	970	-1.1	5015	<2.2e-16	315
1V5IB	0.01	237	-0.42	1207	6.47E-12	76
1WKRA	-0.2	1149	-0.97	5311	<2.2e-16	340
3PVIA	-0.3	460	-0.82	2504	1.78E-15	156
1JU2A	-0.17	1642	-0.7	8257	<2.2e-16	521
2BWRA	-0.15	1240	-0.79	6379	<2.2e-16	401
1L7LA	-0.06	386	-0.71	1913	<2.2e-16	121
1NF9A	-0.39	639	-0.9	3294	1.46E-15	207
1QVEA	-0.22	414	-0.92	1980	<2.2e-16	126
1SH8A	-0.28	476	-0.73	2412	5.66E-11	152
1TP6A	-0.15	384	-0.5	2010	1.03E-07	126
1U4GA	-0.18	918	-0.53	4744	5.01E-15	298
1X6ZA	-0.09	388	-0.82	1873	<2.2e-16	119
1Z6NA	-0.34	511	-0.98	2643	<2.2e-16	166
2FAOA	-0.27	918	-0.69	4687	<2.2e-16	295
1RKUA	-0.14	613	-0.83	3282	<2.2e-16	205
1RTTA	-0.34	556	-0.83	2750	8.40E-13	174
1GQIA	-0.16	2173	-0.77	11279	<2.2e-16	708
1UK8A	-0.14	838	-0.81	4311	<2.2e-16	271
1VM9A	-0.13	335	-0.41	1736	7.72E-05	109
1OH0A	-0.05	383	-0.37	1992	1.22E-07	125
1Q6ZA	-0.19	1677	-0.91	8279	<2.2e-16	524
1UWKA	-0.13	1731	-0.47	8795	<2.2e-16	554
1XLQA	-0.08	322	-0.97	1692	<2.2e-16	106
1YRCA	-0.14	1233	-0.65	6462	<2.2e-16	405

1ZI8A	-0.27	736	-0.81	3691	<2.2e-16	233
2H8ZA	-0.29	1128	-0.72	5693	<2.2e-16	359
1GA6A	-0.19	1227	-0.72	5784	<2.2e-16	369
1LO7A	0	411	-0.25	2249	1.95E-05	140
1V7ZA	-0.2	791	-0.99	4092	<2.2e-16	257
1M70A	0.04	603	-0.42	3007	<2.2e-16	190
1UKFA	-0.23	582	-0.56	2990	2.31E-09	188
1WTJA	-0.25	1059	-0.84	5249	<2.2e-16	332
1RKIA	-0.21	274	-0.95	1645	<2.2e-16	101
1E19A	-0.4	972	-1.33	4975	<2.2e-16	313
1F2TB	-0.34	436	-1.41	2281	<2.2e-16	143
1JG1A	-0.22	656	-0.84	3429	<2.2e-16	215
1X2IA	-0.23	209	-0.92	1083	1.20E-11	68
2DSKA	-0.13	925	-0.76	4775	<2.2e-16	300
1G8AA	-0.18	684	-0.81	3629	<2.2e-16	227
1IU8A	-0.27	645	-0.76	3269	<2.2e-16	206
1V30A	-0.32	336	-0.78	1906	2.11E-07	118
1V4PA	-0.31	440	-0.66	2429	3.23E-07	151
1WN2A	-0.51	360	-1.11	1882	1.92E-13	118
1X0TA	-0.36	284	-0.57	1730	0.03861	106
1X54A	-0.22	1298	-0.83	6948	<2.2e-16	434
2CVIA	-0.41	247	-1.5	1330	1.64E-15	83
2DF8A	-0.17	984	-0.87	5191	<2.2e-16	325
2EVBA	-0.15	232	-0.97	1174	<2.2e-16	74
2HD9A	-0.1	407	-0.74	2348	<2.2e-16	145
2CYJA	-0.6	341	-1.63	1901	<2.2e-16	118
2BT9A	-0.25	289	-0.67	1421	3.16E-08	90
1YV4A	0.05	285	-0.11	1672	0.02647	103
1CIPA	-0.33	924	-0.94	5080	<2.2e-16	316
1DCIA	-0.18	835	-0.66	4390	<2.2e-16	275
1EYHA	-0.1	427	-0.79	2309	<2.2e-16	144
1GVEA	-0.21	976	-0.68	5180	<2.2e-16	324
1LC0A	-0.24	873	-0.8	4637	<2.2e-16	290
1N7SA	-0.08	190	-0.27	1007	0.00643	63
1N7SB	-0.15	205	-0.67	1087	6.18E-10	68
1N7SC	-0.08	239	-0.64	1262	1.74E-11	79
1N7SD	0.1	204	-0.38	1050	8.41E-11	66
1NQ7A	-0.21	713	-0.57	3923	3.70E-10	244
1QAUA	-0.13	356	-0.96	1772	<2.2e-16	112
1QQFA	-0.21	849	-0.77	4395	<2.2e-16	276
1T1UA	-0.18	1786	-0.68	9557	<2.2e-16	597
1TJXA	-0.12	470	-0.58	2494	3.16E-12	156
1YD9A	-0.24	584	-0.95	2988	<2.2e-16	188
2B5HA	-0.15	545	-0.61	2989	6.79E-11	186
2CL5A	-0.07	649	-0.69	3417	<2.2e-16	214
2G6FX	-0.1	180	-0.54	941	1.25E-08	59
2B4HA	-0.02	699	-0.63	3614	<2.2e-16	227
1QFTA	0.01	526	-0.71	2799	<2.2e-16	175
1EW0A	-0.01	403	-0.47	2067	8.77E-16	130

2NVGA	0.06	437	-0.62	2242	<2.2e-16	141
2CZ1B	-0.16	657	-0.66	3352	<2.2e-16	211
1LZLA	-0.2	1027	-0.62	4996	<2.2e-16	317
2DE3A	-0.16	1113	-0.7	5423	<2.2e-16	344
1C1DA	-0.29	1144	-0.74	5487	<2.2e-16	349
1JU3A	-0.12	1812	-0.64	9018	<2.2e-16	570
1W2LA	-0.17	301	-0.63	1542	1.76E-09	97
1UQ5A	0.03	821	-0.73	4176	<2.2e-16	263
1CC8A	-0.23	211	-0.79	1157	7.93E-12	72
1EUVA	-0.24	653	-0.9	3546	<2.2e-16	221
1G6GA	-0.21	377	-0.86	2036	<2.2e-16	127
1GY7A	-0.17	366	-0.71	1933	<2.2e-16	121
1JR8A	-0.3	291	-0.85	1704	3.69E-10	105
1KA1A	-0.21	1079	-0.59	5647	<2.2e-16	354
1P6OA	-0.18	461	-0.38	2503	0.0001689	156
1R6XA	-0.34	1179	-0.74	6155	<2.2e-16	386
1ROCA	-0.02	465	-0.51	2461	<2.2e-16	154
1VFYA	-0.1	180	-0.68	1093	1.19E-05	67
1YFQA	-0.14	1014	-0.94	5484	<2.2e-16	342
2A6ZA	-0.06	683	-0.85	3535	<2.2e-16	222
2BAYA	-0.03	174	-0.31	890	0.0001727	56
2CIUA	-0.12	355	-0.5	1982	8.77E-08	123
2EUTA	-0.11	879	-0.67	4650	<2.2e-16	291
2FBAA	-0.25	1547	-0.72	7801	<2.2e-16	492
1L5OA	-0.32	1113	-0.79	5461	<2.2e-16	346
1LC5A	-0.31	1078	-0.77	5667	<2.2e-16	355
1ORRA	-0.19	1012	-0.62	5410	<2.2e-16	338
1JETA	-0.13	1568	-0.73	8255	<2.2e-16	517
2GWMA	-0.06	610	-0.68	3190	<2.2e-16	200
1TJYA	-0.27	992	-0.89	5012	<2.2e-16	316
1ZVAA	-0.11	240	-0.66	1185	3.20E-15	75
3SDHA	-0.19	442	-0.88	2313	<2.2e-16	145
1N08A	-0.09	458	-0.61	2468	<2.2e-16	154
1QL0A	-0.13	753	-0.59	3826	<2.2e-16	241
2BEMA	-0.11	527	-0.43	2703	1.50E-08	170
1SVFA	-0.55	214	-0.85	1002	0.003584	64
1SVFB	-0.07	124	-0.55	598	4.31E-06	38
2IC6A	-0.18	220	-0.42	1129	0.000752	71
1J1NA	-0.11	1464	-0.88	7884	<2.2e-16	492
1QMGA	-0.16	1598	-0.72	8168	<2.2e-16	514
1Y1PA	-0.1	1084	-0.66	5414	<2.2e-16	342
1EY4A	-0.29	404	-0.71	2180	2.31E-12	136
1JF8A	-0.13	391	-0.55	2079	6.00E-12	130
1KQ1A	0.02	177	-0.7	963	<2.2e-16	60
1YN3A	-0.28	304	-0.74	1558	8.94E-12	98
1YQZA	-0.2	1327	-1	6976	<2.2e-16	437
2O8LA	-0.08	685	-0.75	3419	<2.2e-16	216
1YBKA	-0.21	161	-0.58	827	7.53E-06	52
1XRKA	-0.17	371	-0.69	1909	4.06E-13	120

2CPGA	-0.21	132	-0.73	685	2.45E-07	43
1N7OA	-0.1	2203	-0.94	11496	<2.2e-16	721
2CS7A	-0.32	171	-0.61	855	0.0006365	54
2FI1A	-0.06	576	-0.65	2977	<2.2e-16	187
1ZOSA	-0.3	729	-0.94	3641	<2.2e-16	230
1YXYA	-0.27	724	-0.84	3646	<2.2e-16	230
1C7KA	-0.17	419	-0.56	2089	6.07E-11	132
2A4XA	-0.13	409	-0.5	2080	5.17E-09	131
1WXCA	-0.17	846	-0.49	4341	1.63E-09	273
1DS1A	-0.29	1018	-0.83	5119	<2.2e-16	323
2G2UB	-0.11	523	-0.44	2612	3.71E-10	165
1LQ9A	-0.07	364	-0.43	1764	2.95E-08	112
2CJLA	-0.06	643	-0.37	3233	3.70E-09	204
1KNMA	-0.13	405	-0.68	2046	<2.2e-16	129
1SJWA	-0.08	427	-0.39	2271	5.85E-09	142
1Q0RA	-0.28	958	-0.71	4685	<2.2e-16	297
1ES5A	-0.25	840	-0.59	4100	9.80E-13	260
1YQSA	-0.14	1110	-0.73	5445	<2.2e-16	345
1ZDYA	-0.13	935	-0.8	4784	<2.2e-16	301
1LWBA	-0.07	373	-0.53	1945	4.99E-14	122
1XYIA	-0.67	193	-1.29	1061	1.95E-10	66
1IO7A	-0.36	1088	-1.14	5866	<2.2e-16	366
1O7IA	-0.17	369	-0.83	1816	<2.2e-16	115
1OXXK	-0.31	1075	-1.3	5613	<2.2e-16	352
1R7JA	-0.27	263	-1.16	1447	<2.2e-16	90
1AJSA	-0.14	1259	-0.81	6569	<2.2e-16	412
1DZKA	-0.15	448	-0.87	2364	<2.2e-16	148
1NUYA	-0.19	1008	-0.88	5224	<2.2e-16	328
1UTEA	-0.19	917	-0.78	4821	<2.2e-16	302
1ELUA	-0.27	1173	-0.91	6066	<2.2e-16	381
1E29A	-0.24	417	-0.52	2148	1.91E-06	135
2D1EA	-0.34	732	-0.66	3885	1.56E-08	243
1S2OA	-0.26	747	-0.95	3889	<2.2e-16	244
1EZGA	-0.46	233	-3.53	1325	<2.2e-16	82
1JI1A	-0.24	2005	-0.91	10098	<2.2e-16	637
1H1NA	-0.2	967	-0.78	4828	<2.2e-16	305
1MTPA	-0.32	1031	-1.02	5049	<2.2e-16	320
1MTPB	-0.09	108	-0.44	557	0.0001667	35
1BQCA	-0.12	947	-0.88	4791	<2.2e-16	302
1PVMA	-0.17	517	-1.03	2865	<2.2e-16	178
1I58A	-0.41	577	-1.43	3014	<2.2e-16	189
1KQ3A	-0.29	1117	-0.84	5780	<2.2e-16	363
1OH4A	-0.2	519	-0.76	2787	<2.2e-16	174
1P1MA	-0.21	1217	-0.92	6459	<2.2e-16	404
1THFD	-0.33	780	-1.04	4027	<2.2e-16	253
1YD0A	-0.12	258	-0.53	1433	1.50E-07	89
2FN9A	-0.37	862	-0.86	4458	<2.2e-16	280
2GQTA	-0.27	820	-0.84	4253	<2.2e-16	267
1IUJA	-0.15	310	-0.52	1628	2.70E-08	102

1IV3A	-0.51	469	-1.05	2381	9.09E-13	150
1J3WA	-0.6	422	-1.2	2124	9.78E-12	134
1KWGA	-0.43	1974	-0.94	10262	<2.2e-16	644
1NYKA	-0.38	491	-1.32	2473	<2.2e-16	156
1UAYA	-0.71	774	-1.25	3805	2.02E-15	241
1UFYA	-0.32	381	-0.87	1918	1.50E-11	121
1UG6A	-0.4	1311	-0.89	6783	<2.2e-16	426
1UI0A	-0.12	581	-0.82	3067	<2.2e-16	192
1V2XA	-0.46	578	-0.85	3051	1.79E-08	191
1V37A	-0.32	532	-0.81	2717	1.05E-12	171
1V6SA	-0.48	1235	-1.31	6175	<2.2e-16	390
1V70A	-0.34	329	-0.98	1666	1.37E-13	105
1V8CA	-0.32	516	-0.94	2619	<2.2e-16	165
1VE1A	-0.37	952	-0.99	4786	<2.2e-16	302
1VEFA	-0.46	1211	-1.13	6142	<2.2e-16	387
1WCV1	-0.36	779	-1.18	3838	<2.2e-16	243
1WCWA	-0.67	796	-1.03	4030	1.63E-08	254
1WMWA	-0.57	1024	-0.98	5208	6.32E-15	328
1WNYA	-0.29	563	-0.81	2857	8.30E-15	180
2C78A	-0.27	1232	-0.96	6311	<2.2e-16	397
2CUAA	-0.12	375	-0.83	1943	<2.2e-16	122
2CVEA	-0.53	597	-0.93	3013	1.73E-08	190
2D3YA	-0.25	670	-0.72	3491	3.20E-12	219
2D5WA	-0.17	1826	-0.71	9612	<2.2e-16	602
2D8DA	-0.29	237	-0.69	1283	1.59E-07	80
2F23A	-0.16	482	-0.79	2444	<2.2e-16	154
2B3FA	-0.12	1227	-0.7	6221	<2.2e-16	392
1V8HA	-0.17	331	-0.76	1683	<2.2e-16	106
1WLUA	-0.47	367	-0.69	1856	0.006506	117
1QNRA	-0.22	1071	-0.77	5465	<2.2e-16	344
1KUGA	-0.24	581	-0.84	3238	<2.2e-16	201
1TUKA	-0.17	195	-1.13	1078	<2.2e-16	67
1YPYA	-0.13	573	-0.62	2885	<2.2e-16	182
1OQVA	-0.18	560	-0.8	2689	<2.2e-16	171
1VHWA	-0.32	723	-0.83	3780	<2.2e-16	237
3CHBD	-0.21	312	-0.92	1645	<2.2e-16	103
1JX6A	-0.2	1026	-0.79	5396	<2.2e-16	338
1RTQA	-0.07	941	-0.69	4588	<2.2e-16	291
2DDXA	0.06	1016	-0.64	5140	<2.2e-16	324
1FS7A	-0.28	1389	-0.87	7560	<2.2e-16	471
2FUKA	-0.33	695	-0.83	3447	2.60E-15	218
2GU9A	-0.12	354	-0.52	1755	2.21E-10	111
2NW8A	-0.17	793	-0.52	4261	2.64E-11	266
1LYVA	-0.15	888	-0.72	4489	<2.2e-16	283
2JDAA	-0.21	441	-1.03	2200	<2.2e-16	139
1FK5A	-0.28	295	-1.2	1472	<2.2e-16	93
1O4YA	-0.15	837	-0.65	4293	<2.2e-16	270
1S3EA	-0.1	1505	-0.66	7976	<2.2e-16	499
1SG0A	-0.14	698	-0.8	3672	<2.2e-16	230

Reference List

Reference List

Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2004: refinements integrate structure and sequence family data.
Nucleic Acids Res. 2004; 32(Database issue):D226-9.

Azarya-Sprinzak E, Naor D, Wolfson HJ, Nussinov R. Interchanges of spatially neighbouring residues in structurally conserved environments. Protein Engineering. 1997; 10: 1109–1122

Balasubramanian S, Xia Y, Freinkman E, Gerstein M. Sequence variation in G-protein-coupled receptors: analysis of single nucleotide polymorphisms. Nucleic Acids Res. 2005; 33:1710–1721.

Bao L, Cui Y. Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. Bioinformatics. 2005; 21(10):2185-2190.

Barber CB, Dobkin DP, Huhdanpaa H. The quickhull algorithm for convex hulls. ACM Transactions on Mathematical Software. 1996; 22(4):469-483.

Barenboim M, Jamison DC, Vaisman II. Statistical Geometry Approach to the Study of Functional Effects of Human Nonsynonymous SNPs. Hum Mutat. 2005; 26(5),471-476

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000; 28(1):235-242.

Bonnet E, Van de Peer Y. zt: a software tool for simple and partial Mantel tests. Journal of Statistical software. 2002; 7(10): 1-12

Bostick D, Vaisman II. A new topological method to measure protein structure similarity. Biochem Biophys Res Commun. 2003; 304(2):320-5.

Bostick D, Shen M, Vaisman II. A simple topological representation of protein structure: Implications for new, fast, and robust structural classification. Proteins, 2004; 56(3):487-501.

Bowie JU, Reidhaar-Olson JF, Lim WA, Sauer RT. Deciphering the message in protein sequences: tolerance to amino acid substitutions. Science. 1990; 247:1306–1310.

Bullock AN, Henckel J, DeDecker BS, Johnson CM, Nikolova PV, Proctor MR, Lane DP, Fersht AR. Thermodynamic stability of wild-type and mutant p53 core domain. Proc Natl Acad Sci USA. 1997; 94:14338–14342.

Bullock AN, Fersht AR. Rescuing the function of mutant p53. Nat Rev Cancer. 2001; 1:68–76.

Capriotti E, Fariselli P, Calabrese R, Casadio R. Predicting protein stability changes from sequences using support vector machines. Bioinformatics. 2005; 21 Suppl 2:ii54-ii58.

Capriotti E, Fariselli P, Casadio R. A neural-network-based method for predicting protein stability changes upon single point mutations. Bioinformatics. 2004; 20 Suppl 1:i63-i68.

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat. Genet. 1999; 22:231–238.

Carter C Jr, LeFebvre B, Cammer S, Tropsha A, Edgell M. Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. J Mol Biol. 2001; 311:625-638.

Chasman D, Adams RM. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure based assessment of amino acid variation. J Mol Biol. 2001; 307(2):683–706.

Cheng J, Randall A, Baldi P. Prediction of protein stability changes for single-site mutations using support vector machines. Proteins. 2006; 62(4):1125-32.

Cho Y, Gorina S, Jeffrey PD, Pavletich NP Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. Science. 1994; 265:346-355.

Daggett V, Fersht AR. Is there a unifying mechanism for protein folding? Trends Biochem. Sci. 2003; 28, 18–25

Dayhoff MO, Schwartz RM, Orcut BC. A model for evolutionary change in proteins. In: Dayhoff MO, editor. Atlas of protein sequence and structure. Washington, DC: National Biomedical Research Foundation. 1978; 5:345–352.

Del Sol Mesa A, Pazos F, Valencia A. Automatic methods for predicting functionally important residues. J. Molec. Biol. 2003; 326:1289–1302.

Frank E, Hall M, Trigg L, Holmes G, Witten IH. Data mining in bioinformatics using Weka. Bioinformatics. 2004; 20:2479-2481

Frenz CM. Neural network-based prediction of mutation-induced protein stability changes in Staphylococcal nuclease at 20 residue positions. Proteins. 2005; 59(2):147-51.

Funahashi J, Takano K, Yutani K. Are the parameters of various stabilization factors estimated from mutant human lysozymes compatible with other proteins? Protein Eng 2001; 14:127–134

Gillis D, Rooman M. 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 1997; 272:276–290.

Grantham R. Amino acid difference formula to help explain protein evolution. Science. 1974; 185:862–864.

Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 2002; 320:369–387.

Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982; 143:29-36.

Hamroun D, Kato S, Ishioka C, Claustres M, Beroud C, Soussi T. The UMD TP53 database and website: update and revisions. Hum Mutat. 2006; 27:14-20.

Herrgard S, Cammer SA, Hoffman BT, Knutson S, Gallina M, Speir JA, Fetrow JS, Baxter SM. Prediction of deleterious functional effects of amino acid mutations using a library of structurebased function descriptors. Proteins: Structure, Function, and Genetics 2003; 53(4):806-816.

Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, Eswar N, Haussler D, Sali A. LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics. 2005; 21(12):2814–2820.

Kato S, Han SY, Liu W, Otsuka K, Shibata H, Kanamaru R, Ishioka C. Understanding the function–structure and function–mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. Proc Natl Acad Sci USA. 2003; 100:8424–8429.

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

Krishnan VG, Westhead DR. A comparative study of machine learning methods to predict the effects of single nucleotide polymorphisms on protein function. Bioinformatics. 2003; 19(17):2199–2209.

Kumar MD, Bava KA, Gromiha MM, Parabakaran P, Kitajima K, Uedaira H, Sarai A. ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. Nuleic Acids Res. 2006; 34:D204-6, Database issue

Kwasigroch JM, Gillis D, Dehouck Y, Rooman M. Popmusic, rationally designing point mutations in protein structures. Bioinformatics. 2002; 18:1701–1702.

Mantel N. The detection of disease clustering and a generalized regression approach. Cancer Res. 1967; 27: 209-220.

Mantel N, Valand RS. A technique of nonparametric multivariate analysis.
Biometrics. 1970; 26: 547-558.

Masso M, Lu Z, Vaisman II. Computational mutagenesis studies of protein structure-function correlations. Proteins. 2006; 64(1):234-45.

Masso M, Vaisman II. Comprehensive mutagenesis of HIV-1 protease: a computational geometry approach. Biochem Biophys Res Commun. 2003; 305(2):322-6.

Mathe E, Olivier M, Kato S, Ishioka C, Hainaut P, Tavtigian SV. Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods.
Nucleic Acids Res. 2006a; 34(5):1317-25.

Mathe E, Olivier M, Kato S, Ishioka C, Vaisman II. Hainaut P. Predicting the transactivation activity of p53 missense mutants using a four-body potential score derived from Delaunay tessellations. Hum Mutat. 2006b; 27(2):163-72.

Matthews BW. Studies on protein stability with T4 lysozyme. Adv Protein Chem. 1995; 46:249–278.

Miller MP, Kumar S. Understanding human disease mutations through the use of interspecific genetic variation Hum. Mol. Genet. 2001; 10:2319–2328

Mooney S. Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. Brief Bioinform. 2005; 6:44-56.

Munson PJ, Singh RK. Statistical significance of hierarchical multi-body potentials based on Delaunay tessellation and their application in sequence-structure alignment. Protein Sci. 1997; 6(7):1467-1481.

Nakken S, Alseth I, Rognes T. Computational prediction of the effects of non-synonymous single nucleotide polymorphisms in human DNA repair genes. Neuroscience. 2007; 145(4):1273-9.

Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003; 31(13):3812–3814.

Olivier M, Eeles R, Hollstein M, Khan MA, Harris CC, Hainaut P. The IARC TP53 database: new online mutation analysis and recommendations to users. Hum Mutat. 2002; 19:607–614.

Oobatake M, Ooi T. Hydration and heat stability effects on protein unfolding. Prog Biophys Mol Biol. 1993; 59(3):237-84

Ooi T, Oobatake M, Nemethy G, Scheraga HA. Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. Proc Natl Acad Sci U S A. 1987; 84(10):3086-90.

Pitera JW, Kollman PA. Exhaustive mutagenesis in silico: multicoordinate free energy calculations on proteins and peptides. Proteins. 2000; 41:385–397.

Poupon A. Voronoi and Voronoi-related tessellations in studies of protein structure and interaction. Curr Opin Struct Biol. 2004; 14(2):233-41.

Prevost M, Wodak SJ, Tidor B, Karplus M. Contribution of the hydrophobic effect to protein stability: analysis based on simulations of the ile-96-ala mutation in barnsase. Proc Natl Acad Sci USA. 1991; 88:10880–10884.

Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. Nucleic Acids Res. 2002; 30:3894–3900

Rice WJ, MacLennan DH. Scanning mutagenesis reveals a similar pattern of mutation sensitivity in transmembrane sequences M4, M5, and M6, but not in M8, of the Ca2+-ATPase of sarcoplasmic reticulum (SERCA1a). J Biol Chem. 1996; 271(49):31412-9.

Saunders CT, Baker D. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. J. Molec. Biol. 2002; 322:891–901.

Singh R, Tropsha A, Vaisman II. Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues. J Comput Biol. 1996; 3:213-221.

Shiraishi K, Kato S, Han SY, Liu W, Otsuka K, Sakayori M, Ishida T, Takeda M, Kanamaru R, Ohuchi N, Ishioka C. Isolation of temperature-sensitive p53 mutations from a comprehensive missense mutation library. J Biol Chem. 2004; 279(1):348-55.

Smith M. In vitro mutagenesis. Annu Rev Genet. 1985; 19:423-62

Stitziel NO, Tseng YY, Pervouchine D, Goddeau D, Kasif S, Liang J. Structural location of disease-associated single-nucleotide polymorphisms. J. Molec. Biol. 2003; 327:1021–1030.

Soussi T, Béroud C. Significance of TP53 mutations in human cancer: A critical analysis of mutations at CpG dinucleotides. Hum Mutat. 2003; 21:192-200

Soussi T, Ishioka C, Claustres M, Beroud C. Locus-specific mutation databases: pitfalls and good practice based on the p53 experience. Nat Rev Cancer. 2006; 6, 83-90.

Soussi T, Kato S, Levy PP, Ishioka C. Reassessment of the TP53 mutation database in human disease by data mining with a library of TP53 missense mutations.
Hum Mutat. 2005; 25(1):6-17.

Sunyaev S, Ramensky V, Koch I, Lathe IIIW, Kondrashov AS, Bork P. Prediction of deleterious human alleles. Hum. Molec. Genet. 2001; 10:591–597.

Tavtigian SV, Deffenbaugh AM, Yin L, Judkins T, Scholl T, Samollow PB, de Silva D, Zharkikh A, Thomas A. Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. J Med Genet. 2006; 43:295–305.

Taylor T, Rivera M, Wilson G, Vaisman II. New method for protein secondary structure assignment based on a simple topological descriptor. Proteins. 2005; 60(3):513-24.

Toyoshima C, Nakasako M, Nomura H, Ogawa H. Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 A resolution. Nature. 2000; 405:647–55

Tropsha A, Carter CJr, Cammer S, Vaisman II. Simplicial neighborhood analysis of protein packing (SNAPP): a computational geometry approach to studying proteins. Methods Enzymol. 2003; 374:509–544.

Tropsha A, Singh R, Vaisman II, Zheng W. Statistical geometry analysis of proteins: implications for inverted structure prediction. Pac Symp Biocomput. 1996; 614-623.

Vaisman II, Tropsha A, Zheng W. Compositional preferences in quadruplets of nearest neighbor residues in protein structures: statistical geometry analysis. Proc IEEE Symp Intell Sys. 1998; 163-168.

Verzilli CJ, Whittaker JC, Stallard N, Chasman D. A hierarchical Bayesian model for predicting the functional consequences of amino acid polymorphisms. Appl Statistics. 2005; 54:191–206.

Vogelstein B, Lane D, Levine AJ. Surfing the p53 network. Nature. 2000; 408:307–310.

Wang G, Dunbrack RL. Jr. PISCES: a protein sequence culling server. Bioinformatics. 2003; 19:1589-1591

Wang Z, Moult J. SNPs, protein structure, and disease. Hum Mutat. 2001; 17(4):263–270.

Weberndorfer G, Hofacker IL, Stadler PF. An efficient potential for protein sequence design. Proc German Conf Bioinformatics. 1999; 107-112.

Witten IH, Frank E. Data Mining: Practicle Machine Learning Tools and Techniques. 2nd ed. 2005; San Fransisco. Morgan Kauffman. 524.

Zheng W, Cho SJ, Vaisman II, Tropsha A. A new approach to protein fold recognition based on Delaunay tessellation of protein structure. Pac Symp Biocomput. 1997; 486-497.

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

Zhou H, Zhou Y. Stability scale and atomic solvation parameters extracted from 1023 mutation experiments. Proteins. 2002b; 49(4):483-92

CURRICULUM VITAE

Bill Shili Zhan was born in Fuzhou, China and is an American citizen. He received his Bachelors of Science degree in Biology from Xiamen University in Xiamen, China. After graduation, he continued pursuing his graduate study in Biochemistry at the same school, and was awarded Master of Science in Biochemistry in 1992. In 1991, he received 'Guan Hua Awards', which is given to the ten most outstanding graduate students in Xiamen University. He worked as a research assistant at National Institutes of Health from 1992 to 1998. In 2000, he received Master of Science in computer science with honors from Southeastern University, Washington, DC. From 1999 to present, he worked as a Senior Database Administrator in CALIBRE Systems Inc. He joined the Ph.D. program in Bioinformatics at George Mason University in the spring of 2001, and completed his Ph.D. in the fall of 2007.