

## RESEARCH ARTICLE

# Predicting the Transactivation Activity of p53 Missense Mutants Using a Four-Body Potential Score Derived From Delaunay Tessellations

Ewy Mathe,<sup>1,2</sup> Magali Olivier,<sup>2</sup> Shunsuke Kato,<sup>3</sup> Chikashi Ishioka,<sup>3</sup> Iosif Vaisman,<sup>1</sup> and Pierre Hainaut<sup>2\*</sup>

<sup>1</sup>Department of Bioinformatics and Computational Biology, George Mason University, Manassas, Virginia; <sup>2</sup>International Agency for Research on Cancer, Lyon, France; <sup>3</sup>Department of Clinical Oncology, Institute of Development Aging and Cancer, Tohoku University, Sendai, Japan

Communicated by Richard Wooster

We describe a novel statistical scoring method based on a computational geometry approach to predict the functional impact (transactivation activity) of missense mutations in the DNA-binding domain (DBD) of the tumor suppressor *TP53*, which is the most frequently mutated gene in human cancer. Residual scores (RS) for each residue were calculated to reflect differences in the compositional preferences of four nearest-neighbor residues between mutant and wild-type proteins. The RS were then combined into a residual score profile (RSP) representing the RS values for all 194 residues in the DBD. 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. While these functional categories showed significant differences in average RS, the latter lacked resolution power to predict the transactivation activities of individual mutants. In contrast, using decision tree models, we found that the RSP predicted transactivation with an accuracy varying between 64.2% and 78.5% depending on the promoter. Lastly, we used the best model to predict the functional outcome of all missense mutants in the DBD of p53 and compared the predictions with their frequency of occurrence in human cancers. We found that mutants predicted as functional (F) accounted for ~14% of all missense mutants found in cancers, while mutants predicted as nonfunctional (NF) represented ~86% of the mutants. These results show that this computational approach provides a fast and reliable method for predicting the functional impact of p53 mutants associated with cancer. *Hum Mutat* 0, 1–10, 2006. © 2006 Wiley-Liss, Inc.

KEY WORDS: Delaunay tessellations; *TP53* mutations; functional predictions; computational mutagenesis; p53

## INTRODUCTION

Although missense mutations are frequently observed in human diseases [Cooper et al., 1998], their impact on protein function and contribution to disease are often difficult to assess. Missense mutations can affect functionally important residues, such as those located in protein–DNA and protein–protein interactions or phosphorylation sites. They may also lead to structural modifications that increase or decrease protein stability. Measuring the stability and/or activity of missense mutants experimentally is often difficult and labor intensive. An increasing number of computational tools are being developed to analyze all possible point mutations in a given protein and determine which amino acids have key structural and functional features. These tools are based on different approaches, including sequence homology [Ferrer-Costa et al., 2002; Vitkup et al., 2003], structure homology [Dambosky et al., 2001; Prokop et al., 2000; Schwede et al., 2003], a combination of both [Kelley et al., 1999, 2000; Ramensky et al., 2002], and evolutionary conservation [Greenblatt et al., 2003; Ng and Henikoff, 2001, 2002; Walker et al., 1999]. Recently, machine learning methods have been applied to predict the stability and functional effects of single point mutations [Capriotti et al., 2004; Karchin et al., 2005; Krishnan and Westhead, 2003].

Computational geometry is also used to study the impact of missense mutations. Simplicial neighborhood analysis of protein packing (SNAPP) [Carter et al., 2001; Tropsha et al., 2003] uses scores 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. SNAPP scores have shown good correlations with experimental free energy values of hydrophobic core mutants

Received 29 July 2005; accepted revised manuscript 30 September 2005.

\*Correspondence to: Pierre Hainaut, International Agency for Research on Cancer, 150 cours Albert Thomas, 69372 Lyon Cedex 08, France. E-mail: hainaut@iarc.fr

Grant sponsor: International Agency for Research on Cancer; National Institute of Environmental Health Sciences; European Community (EC), FP6.

This publication reflects the authors' views and not necessarily those of the EC. The EC is not liable for any use that may be made of the information contained herein.

DOI 10.1002/humu.20284

Published online in Wiley InterScience (www.interscience.wiley.com).

for five proteins with available experimental data [Carter et al., 2001; Tropsha et al., 2003].

p53 is a transcription factor that is encoded by the tumor suppressor *TP53* (MIM# 191170), the most frequently mutated gene in human cancers. Mutations reported in the scientific literature are compiled in the IARC TP53 database ([www-p53.iarc.fr](http://www-p53.iarc.fr)) [Olivier et al., 2002], together with experimental data on various properties of mutant proteins. p53 exerts antiproliferative activities mainly by regulating target genes via DNA binding on specific sequences [Hofseth et al., 2004] located in either the promoter or intronic regulatory sites. The p53-response element is a degenerated and repeated sequence that matches the consensus 5'-PuPuPuC(A/T)(T/A)GPyPyPy-3' [el-Deiry et al., 1992], where both the number of repetitions and the sequence depend on the target gene. Most mutations in human cancers are located in the DNA-binding domain (DBD) [Olivier et al., 2004]. The structure of this domain (as protein : protein or protein : DNA complexes) has been determined by crystallography [Cho et al., 1994]. It consists of a globular domain made of two anti-parallel beta-sheets that provide a scaffold for flexible loops involved in contacting DNA. This DNA-binding surface is formed by two contiguous motifs: a loop-sheet-helix that binds in the major groove (LSH), and a large loop that makes contacts in the minor groove of DNA (L2). This latter structure is stabilized by a zinc atom bound to three cysteines and one histidine. The DBD is a well structured but flexible domain that can adopt different thermodynamic states [Bullock and Fersht, 2001].

The wild-type protein is rapidly degraded and is present at low concentrations in cells under normal conditions. After activation under stress conditions, such as DNA damage, oncogene activation, or hypoxia, it accumulates in the nucleus and regulates genes involved in cell cycle checkpoints (G1 and G2 phases), apoptosis, or DNA repair. In a systematic study using a yeast-based

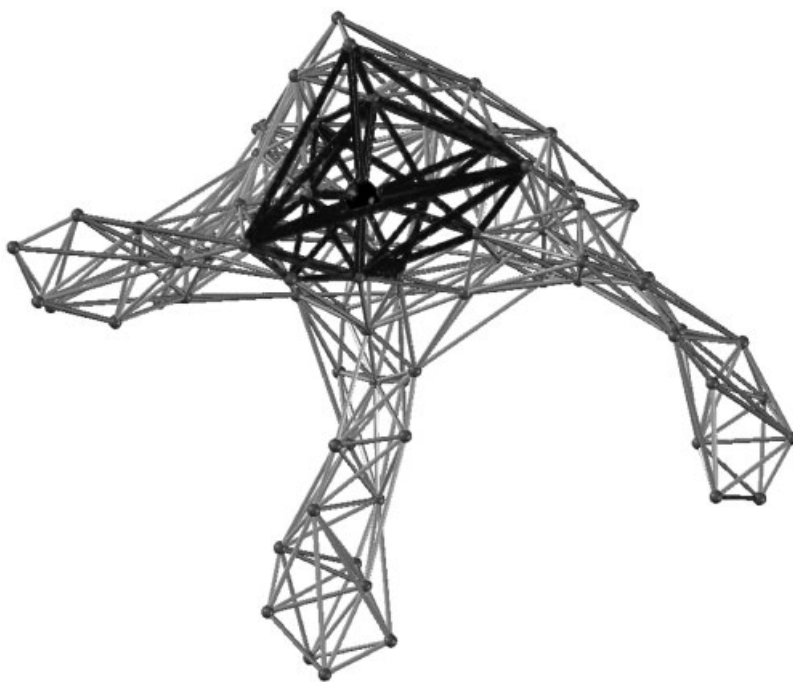
functional assay, Kato et al. [2003] measured on eight p53-response elements derived from p53-target genes (*WAF1*, *MDM2*, *BAX*, *h1433σ*, *AIP1*, *GADD45*, *NOXA*, and *P53R2*) the transactivation capacity of all missense mutants resulting from single nucleotide substitutions. They showed great variations in the degree of loss of function between mutant proteins and between the same mutants across different promoters.

In this study we calculated residual scores (RS), in a manner similar to that used for SNAPP scores, for all possible missense mutants resulting from a single nucleotide substitution in the DBD of p53. We also extended this method to calculate mutant RS profiles (RSP) based on RS for each amino acid position in the DBD. First, RS were compared with mutant p53 structural properties as classified by Martin et al. [2002]. Second, RS and RSP were used to predict transactivation activity based on the functional data of Kato et al. [2003].

## MATERIALS AND METHODS

### Potential Scores and Profiles

Delaunay tessellations were generated using the program Qhull [Barber et al., 1996]. Each point in the tessellation represents the position of the center of mass of a residue side chain in a given protein (Fig. 1). In 3D space, the simplices that result from the tessellation are tetrahedra, where the four vertices correspond to four nearest-neighbor residues. Based on this representation, a collection of all of the quadruplets (four nearest-neighbor residues) that make up a given protein is defined. A training set of 1,417 proteins with available structures in the Protein Data Bank (PDB) [Berman et al., 2000] were used to calculate the log likelihood scores of all the different quadruplets found in this training set. The inclusion criteria for selecting structures in the training set were as follows: 1) structure determined by X-ray crystallography,



**FIGURE 1. Representation of a Delaunay tessellation. Each point represents the center of mass of an amino acid side chain. The larger black dot represents a mutation site, with the black edges joining the mutation site and its nearest residue neighbors. Modifying the amino acid letter at that position will thus affect all the quadruplets in which that residue participates. The tessellation of a protein gives the list of all four nearest-neighbor residues that constitute this protein.**

2) maximum resolution of 2.2 Å, 3) maximum R-value of 0.23, and 4) maximum of 30% sequence identity. The maximum number of possible quadruplets using a 20-letter amino acid letter code and irrespective of sequence order is 8,855. The training set contained 8,843 quadruplets. Based on these quadruplet frequencies, the log likelihood score of a given quadruplet was calculated as follows [Vaisman et al., 1996]:

$$q_{ijkl} = \log \frac{f_{ijkl}}{p_{ijkl}}$$

where  $i, j, k,$  and  $l$  are the four amino acids that compose the quadruplet;  $f_{ijkl}$  is the frequency of the quadruplet in the training set; and  $p_{ijkl}$  is the frequency of random occurrence of the quadruplet. The latter is defined as  $c_{a_i a_j a_k a_l}$ , where  $c$  is a permutation factor that accounts for the permutability of replicated residues in a given quadruplet (i.e., AACF and ACAF are considered to be the same), and  $a_i, a_j, a_k,$  and  $a_l$  are the frequencies of amino acids  $i, j, k,$  and  $l$  in the training set.

The potential score for each residue in the DBD of p53 was computed as the sum of the scores of all quadruplets in which that residue participates, while the potential score for the entire protein is the sum of the log likelihood scores of all quadruplets in that protein. A potential score profile was then constructed as a vector  $\vec{V}$  of length  $n$ :

$$\vec{V} = (PSaa_1, PSaa_2, \dots, PSaa_n)$$

where  $n$  is the number of amino acids in the protein (194), and  $PSaa_1$  is the potential score of amino acid 1. All potential score calculations were based on the X-ray structure of the DBD of p53 (PDB [Berman et al., 2000] ID 1tsr, chain B), which includes 194 amino acids (residues 96–289) [Cho et al., 1994].

### RS and RSP

To simulate a single point mutation, the identity of an amino-acid residue at the desired position is changed in the tessellated structure of wild-type p53. This letter switch changes the amino acid composition of the quadruplets that include the substituted residue (Fig. 1). To evaluate the changes due to a single amino acid substitution, the RS are calculated by subtracting the mutant potential score from the wild-type potential score. Similarly, RSP are obtained by subtracting the residue potential scores for each amino acid in the mutant from their respective residue potential scores in the wild-type protein. Importantly, only the structural neighbors of the mutated residue will have nonzero values in the RSP. RS and RSP were calculated for all possible single point mutations in the DBD of p53 (1,147 mutants).

### P53 Mutants and Functional Data Sets

The data set on the functional activity of 426 missense mutant proteins in the IARC TP53 database (R9, July 2004, [www-p53.iarc.fr/index.html](http://www-p53.iarc.fr/index.html)) was used to classify mutants as functional or nonfunctional according to their capacity to transactivate. The functional tests considered were all those cited in the literature, and included assays performed in human or yeast cells. Mutants were categorized as functional for a given gene if it appeared in the database as “retained function,” and as non-functional if it appeared under “lost function.” In the case of conflicting data (i.e., one assay found that the mutant had lost function, while another found that the function was retained), the mutant remained unclassified and was not used in the analysis.

A transactivation data set, which was produced using a yeast functional assay by Kato et al. [2003] on eight different p53-binding sequences, was used to systematically classify all possible

missense mutants according to their transactivation. The eight p53-binding sequences were derived from the promoters of *WAF1*, *MDM2*, *BAX*, *h1433σ*, *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. These percent activity values were grouped into two categories: mutants with an activity below 40% were classified as NF (with significant loss of function), whereas mutants with an activity above 80% were classified as F (close to wild-type). Mutants with activity between 40% and 80% were excluded from further analysis. These percent cutoffs were chosen according to the distribution of the percent activity values.

To test the concordance between these categories and the transcriptional activities obtained in other experimental assays, these categories were compared with the data set of the IARC TP53 database. Of the 190 mutants that were found in the database and evaluated by Kato et al. [2003], 80.8% had the same function in both data sets, indicating that the F and NF categories are representative of mutant function independently of the experimental assay.

Additionally, a separate analysis was performed on a subset of 683 “consistent” mutants that had similar transactivation, according to the yeast functional assay [Kato et al., 2003], across all eight genes. For this subset, mutants with significant loss of activity across all genes (<40% activity) were classified as NF, and mutants with partial activity on all genes (≥40% activity) were classified as F.

### Biochemical and Structural Classifications

Mutants were allocated to different structural categories according to Martin et al. [2002], who evaluated the structural implications of p53 core domain mutants reported in the IARC database. They provided a systematic classification of p53 mutants based on the structural and functional properties of the mutated residue (available at [www.bioinf.org.uk](http://www.bioinf.org.uk)). In this approach, the Jones and Thornton [1997] method was used to determine the possible protein–protein interaction sites, and mutations classified as “in patch” correspond to those occurring at residues that may be involved in protein–protein interactions. Mutations classified as “clashes” are those that form bad contacts with nearest-neighbor residues. A detailed description of the classifiers is provided in Martin et al. [2002].

Other biochemical parameters were also calculated, including Grantham conservation scores [Grantham, 1974] and solvent accessibility of residues. Grantham conservation scoring is based on the composition (atomic weight ratio of hetero elements in the side chain), polarity, and volumes of the amino acids. These three parameters are combined into a Grantham conservation score that reflects the degree of changes that would be incurred by a specific substitution. Next, Naccess [Hubbard and Thornton, 1992–1996] was used to calculate the solvent accessibility of residues in the wild-type protein. All of the default parameters were applied, and the solvent accessible area for each residue relative to the entire protein was considered for our analysis.

### Machine Learning Algorithms for Predictions of Transactivation

Predictions of transactivation activities, based on the NF and F classifications described above, were made with the freely available machine learning software WEKA [Witten and Eibe, 2000]. The transactivation data of each promoter were treated independently such that one model was built for the functional predictions of one

promoter. The K-nearest-neighbor (KNN) classifier with  $K = 3$  (IBK) and decision tree (J48) algorithms, with default parameters and no normalization of the data, were used for RS and RSP, respectively. 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 is used as the test set while the remaining groups (all but the test set) are used for training. The stratification ensures that the proportion of F and NF mutants in the 10 groups is kept similar to that found in the entire data set. The predictions resulting from the model are directly compared with the transactivation activity categories, from the yeast-based assay, of a given promoter. From these predictions, the average accuracy score resulting from all test sets, which reflects the number of predictions that match the known activity measurements, is calculated. In addition, true-positive fractions (i.e., the number of correctly classified category NF mutations divided by the total number of category NF mutations) and false-positive fractions (the number of incorrectly classified category NF mutations divided by the total number of category NF mutations) were evaluated.

In the KNN classifier, the category of a test mutant is assigned according to the majority category of the three closest mutants determined by the Euclidean distance between them. For the decision tree models, receiver operating characteristic (ROC) curves (false-positive fraction vs. true-positive fraction) were drawn. Each point in the curve represents the 10-fold cross-validation results of one model using a user-specified, cost-sensitive matrix. The ROC curves describe how the false- and true-positive fractions fluctuate when a different cost-sensitive matrix is used to derive the models. From these ROC curves the most appropriate model can easily be found, which in our case corresponds to the one that yields a high true-positive fraction while maintaining a low false-positive fraction. In this sense, the best model predicts a nearly equal percentage of F and NF category mutants. The area under the curve (AUC) is then calculated for each ROC using the trapezoid rule. AUC values close to one indicate a good discrimination power between F and NF mutants. Cost-sensitive matrices that define different weights to be applied to each category were applied to the functional data prior to model building to artificially change the proportions in the number of F and NF mutants. For example, *MDM2* contains 99 F and 819 NF mutants. Because roughly 90% of the data are NF, a good resulting model may strive to predict only NF, which would result in 90% accuracy. Applying the cost-sensitive matrix will artificially increase the instances of F mutants, ensuring that the model will strive to predict both F and NF mutations.

## RESULTS

### Correlations Between Mutant Structural Features and Average RS

The RS for all missense mutants resulting from a single base substitution in the DBD of p53 (residues 96–289) ranged from  $-13.14$  to  $8.79$ , with an average of  $-0.85$ . A negative RS indicates that the log likelihoods of the mutant quadruplets are inferior to the log likelihoods of the wild-type quadruplets, whereas a positive RS implies that the mutant potential score is higher than that of the wild-type. Previous studies have shown that strongly negative scores are likely to reflect less stable (and possibly less active) mutants, and are generally found in the protein core [Carter et al., 2001; Masso and Vaisman, 2003; Tropsha et al., 2003]. Conversely, mutations with strongly positive RS may suggest that

the new amino acid introduces structural features that increase the stability and/or activity of the protein.

Our results show that 65 out of 66 mutants with RS below  $-5$  are in rigid secondary structures (beta and alpha strands, and beta turn, as determined by ProMotif [Hutchinson and Thornton, 1996] in PDBSum [Laskowski, 2001]). The lowest RS were observed for mutations at zinc-binding residues (Cys176, His179, Cys238, and Cys242). Given the structural and functional importance of cysteine motifs, quadruplets describing such motifs (i.e., CCCC, CCCH) occur frequently in the training data set, and RS for most mutants involving cysteine substitutions are expected to be far from zero. Furthermore, 26 of 31 (84%) mutants with RS  $> 3$  were also found at residues involved in rigid secondary structure. Additionally, when we compared the RS with the Grantham scores, which evaluate biochemical differences between mutant and wild-type residues, the highest absolute value of RS corresponded to higher Grantham scores, indicating that the high absolute value of RS reflects a substantial chemical/physical difference between the mutant and wild-type residues (Fig. 2).

Average RS values were then calculated for structural categories of p53 mutants as defined by Martin et al. [2002]. The average RS for mutations at residues that are non solvent-accessible, not in patch (i.e., not involved in potential protein–protein interaction sites), or non-DNA binding (i.e., making no direct contact with DNA) tended to be lower than for mutations at residues that are solvent-accessible, in patch, or DNA binding, respectively (Table 1). This observation is in agreement with the fact that residues in the hydrophobic core play stronger roles in protein stability than most solvent-accessible residues. The average RS for mutants classified as “clashes” (i.e., predicted to cause structural clashes based on differences in chemical properties or geometry of the mutant side chain [Martin et al., 2002]), was also well below zero.

### RS and Transactivation Activities

The transactivation activities of missense mutants were measured in yeast assays by Kato et al. [2003] on eight different promoters. Mutants in the DBD were classified according to their activity relative to the wild-type protein and classified as NF or F (see Materials and Methods). For each promoter, the average RS were calculated for NF and F categories. The average RS showed a significant difference between NF and F, with the most negative values corresponding to NF mutants (all P-values  $< 0.05$ ; Table 2). A set of “consistent” mutants was identified that contained mutants with similar activity on all promoters. This set consisted of 392 NF and 291 F mutants. For this subset, the average RS of the NF mutants was also lower than the average RS of the F mutants (Table 2). Thus, the average RS show a clear distinction between transcriptionally active and inactive p53 mutants.

### Predicting Transactivation Activity From RS

A KNN algorithm (with  $K = 3$ ) based on the Euclidean distances between RS was used to predict whether specific missense mutants would fall into the NF or F category. The predictions were made for each promoter, as well as for the “consistent” category. The proportion of correctly predicted mutants, assessed by a 10-fold cross-validation (see Materials and Methods), ranged from 54.8% to 87.6% depending upon the promoter, with true-negative fractions (TNF, percentage of F mutants correctly classified) ranging from 2.1% to 50.8%. In an attempt to improve the prediction power, the RS were combined with solvent accessibility (SA), Grantham scores (GS),

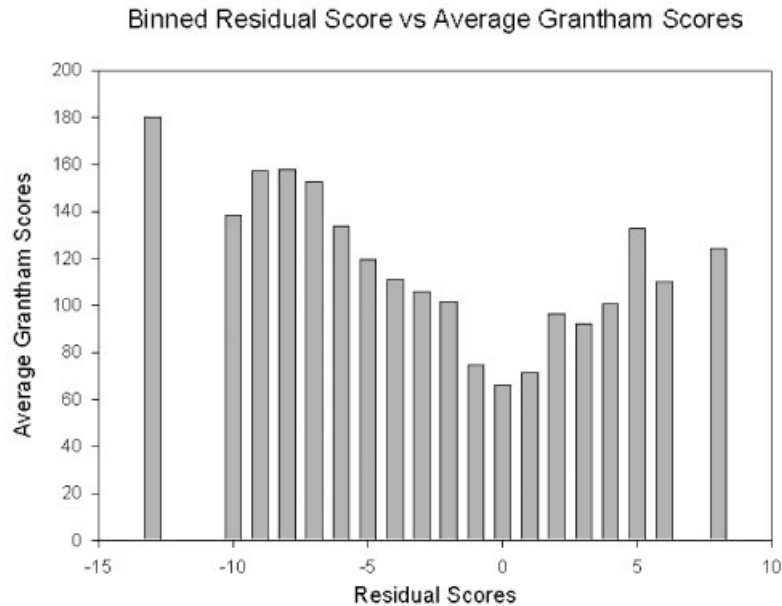


FIGURE 2. Histogram of RS binned by average Grantham score. The RS measure the difference in log likelihood of the quadruplets between the mutant and wild-type structures. Extreme values of RS indicate a large difference in log likelihood and suggest that the substitution substantially affects rearrangements around the mutated residue. Grantham scores measure the differences in chemical properties between wild-type and mutant residues. As expected, there is a clear trend between extreme values of RS and high Grantham scores.

TABLE 1. Comparison Between RS and Structural Classification of p53 Missense Mutants as Defined by Martin et al. [2002]

Categories	Number of codons	Number of missense mutations	Number of mutations in database <sup>d</sup>	Average residual score (range)
All mutations	189	883	12966	-1.1 (-13.1 to 8.8)
Zinc binding	4	25	860	-7.02 (-10.3 to -3.85)
DNA binding	14	67	3223	-0.54 (-5.9 to 6.08)
Non-DNA binding	175	816	9743	-1.12 (-13.1 to 8.8)
Not in patch <sup>a</sup>	96	476	10203	-1.39 (-13.1 to 8.8)
In patch <sup>b</sup>	93	407	2763	-0.71 (-8.83 to 5.6)
Solvent accessible	71	300	2783	-0.52 (-6.74 to 6.08)
Non-solvent accessible	118	583	10183	-1.37 (-13.1 to 8.8)
Clashes <sup>c</sup>	9	20	467	-3.59 (-13.1 to 1.48)

<sup>a</sup>Residues not in possible protein:protein interaction sites [Jones and Thornton, 1997].

<sup>b</sup>Residues in possible protein:protein interaction sites [Jones and Thornton, 1997].

<sup>c</sup>Residues involved in bad structural contacts with nearby residues.

<sup>d</sup>IARC TP53 Database, R9, July 2004.

TABLE 2. Comparing the Average Residual Score in Each Transactivation Category

Promoters	Average RS (range) NF mutants <sup>a</sup>	Average RS (range) F mutants <sup>a</sup>
WAF1	-1.2 (-13.1 to 8.8)	-0.29 (-6.7 to 5.7)
MDM2	-1.04 (-13.1 to 8.8)	-0.41 (-6.1 to 5.7)
BAX	-1.17 (-13.1 to 8.8)	-0.28 (-6.7 to 5.7)
h1433 $\sigma$	-1.27 (-13.1 to 8.8)	-0.39 (-6.7 to 5.7)
AIP1	-1.27 (-13.1 to 8.8)	-0.46 (-6.7 to 4.3)
GADD45	-1.42 (-13.1 to 8.8)	-0.42 (-6.7 to 5.2)
NOXA	-1.49 (-13.1 to 8.8)	-0.38 (-8.2 to 5.8)
P53R2	-1.59 (-13.1 to 8.8)	-0.28 (-8.2 to 5.8)
"Consistent"	-1.64 (-13.1 to 8.8)	-0.023 (-2.4 to 4.3)

<sup>a</sup>The P-value when testing the difference in average RS between NF and F mutants was below 0.05 for all promoters tested.

RS, residual score; NF, nonfunctional mutants; F, functional mutants; "Consistent", set of mutants with similar transcriptional activity across all eight promoters.

or SA+GS. While changes in prediction accuracies were observed (69.4–88.2%, 59.5–86.3%, and 67.1–86.7% for SA, GS, and SA+GS, respectively), the TNF were still low, ranging from 3.3–32.5%, 3.9–45%, and 3.4–35.5% for SA, GS, and SA+GS, respectively. The low TNF indicates that RS, alone or combined with SA or GS, cannot predict activity with sufficient accuracy.

### Predicting Transactivation Activity Using RSP

RSP were calculated for all missense mutants, where RSP are represented by a vector that comprises RS for all 194 amino acids in the DBD (see Materials and Methods). Because making predictions with RSP requires a model capable of handling 194 dimensions, a machine learning method was sought and decision trees were used. The predictions were made for each promoter and for the “consistent” mutants and assessed against NF and F categories using different models resulting from the decision trees. A cost-sensitive matrix, which applies different weights to the input data, was applied to the algorithm to artificially vary the proportion of NF and F mutants (see Materials and Methods) [Weiss and Provost, 2003; Witten and Eibe, 2000]. ROC curves were drawn and the AUCs were calculated to test the robustness of the predictions. The prediction accuracies ranged from 64.2% (MDM2 promoter) to 78.5% (“consistent” mutants) with corresponding AUCs of 0.673 and 0.795, respectively (Table 3). Figure 3 shows the ROC curves for “consistent mutants”, and for MDM2 and GADD45 promoters, that gave the lowest and highest promoter-specific predictions, respectively. Adding SA, GS, or SA+GS to RSP in the prediction models did not significantly change the prediction accuracies and AUC, except for GADD45 and the “consistent” mutants for which SA+GS slightly increased the prediction accuracy and AUC (Table 3).

As a negative control, the scores in the RSP of each mutant were shuffled to randomize their positions in the primary sequence, and ROC curves were drawn using these “shuffled” RSP. For all promoters, the AUC for the “shuffled” RSP was similar to random (dotted lines in Fig. 3), indicating that the position of the RS in the profiles is important.

### Comparison of Predicted Functional Categories With Observed Mutation Frequency in Cancers

To evaluate the biological significance of the functional categories predicted from our models, we calculated the mutation frequency observed in human cancers for each predicted category. The frequencies were extracted from the IARC TP53 database of somatic mutations (R9, July 2004). The database contains mutations reported in all types of human tumors. We applied the best prediction model, derived from using the “consistent” mutants as the training set, to each of the 1,147 possible missense mutants in the DBD of p53 (residues 96–289). Each mutant was predicted to be either F or NF. This model gave 596 (52%) NF mutants and 551 (48%) F mutants. When the sum of the frequencies of all mutants in these categories was calculated, NF mutants accounted for ~86% (12978/15076) of the missense mutants in the DBD reported in the IARC TP53 database.

## DISCUSSION

When applied to protein structures, Delaunay tessellations provide an objective definition of nearest neighbors for each residue [Poupon, 2004]. This method has been successfully used to discriminate between disease-causing and non-disease-causing SNPs [Stitzel et al., 2003, 2004]. It has been shown that four-

TABLE 3. Prediction Accuracy and AUC Results Using the Residual Score Profiles Alone, or in Combination with Solvent Accessibility, Grantham Scores, or Both, for Predicting Transactivation Activity

Results by promoter <sup>a</sup>	RSP	RSP+SA	RSP+GS	RSP+GS+SA
<b>WAF1</b>				
Prediction accuracy <sup>b</sup>	69.6	70.6	68	71.5
AUC	0.735	0.776	0.743	0.785
<b>MDM2</b>				
Prediction accuracy	64.2	65.1	67.4	65.7
AUC	0.673	0.685	0.669	0.719
<b>BAX</b>				
Prediction accuracy	70.7	69.9	70.2	69.2
AUC	0.738	0.637	0.739	0.76
<b>h1433σ</b>				
Prediction accuracy	72.6	74.9	70.6	73.1
AUC	0.753	0.772	0.747	0.774
<b>AIP1</b>				
Prediction accuracy	70.7	72.1	69.7	70.9
AUC	0.746	0.786	0.828	0.792
<b>GADD45</b>				
Prediction accuracy	74.2	76.7	74.0	77.8
AUC	0.783	0.829	0.781	0.832
<b>NOXA</b>				
Prediction accuracy	72.5	73.4	70.0	72.9
AUC	0.78	0.811	0.783	0.817
<b>P53R2</b>				
Prediction accuracy	73.0	73.4	71.9	74.6
AUC	0.795	0.813	0.791	0.799
<b>“Consistent”</b>				
Prediction accuracy	78.5	78.9	75.4	80.5
AUC	0.811	0.847	0.807	0.865

<sup>a</sup>The results are based on a 10-fold cross-validation using decision trees. The predictions are strong for all promoters except MDM2, and are strongest for the set of consistent mutants.

<sup>b</sup>The prediction accuracy reflects the average accuracy measured for each test set in the 10-fold cross-validation. RSP, residual score profiles; SA, solvent accessibility; GS, Grantham score.

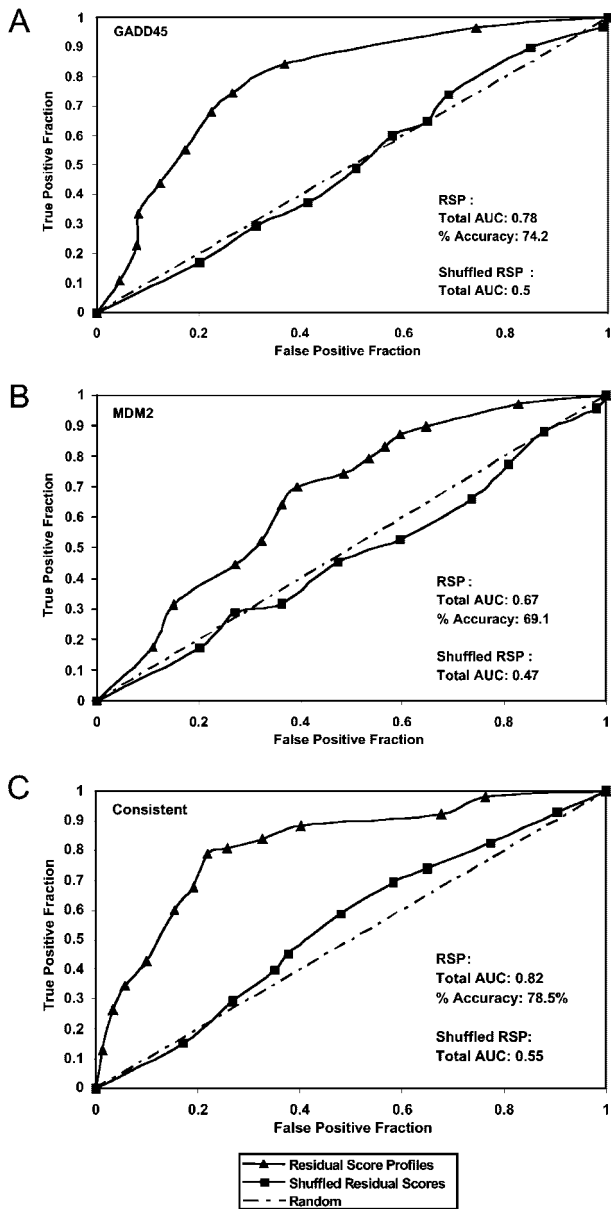


FIGURE 3. ROC curves for the best transactivation (TA) predictions across *GADD45* (A), worst predictions across *MDM2* (B), and predictions across “consistent mutants” (C). In each graph, the line with triangles reflects the results of the model using the RSP, the dotted line represents random classifications, and the line with squares reflects results obtained from random shuffling of the RSP. A total AUC closer to 0.5 indicates a discrimination power between both categories close to random, while a total AUC closer to 1 indicates a strong discrimination power between F and NF mutants.

body statistical potentials derived from the tessellations can distinguish native from non-native protein conformations [Krishnamoorthy and Tropsha, 2003]. Importantly, the amino acid composition of all nearest-neighbor quadruplets in a training set of representative protein structures is non-random [Singh et al., 1996; Vaisman et al., 1996], reflecting the frequent occurrence of structurally important motifs. In fact, the most common quadruplets found in the training set are dominated by cysteines, which often participate in metal or disulfide-linked structures, and are frequent structural neighbors. Based on these log likelihoods,

we calculated RS and RSP for all missense mutants in the DBD of p53, the most frequently mutated gene product in human cancer. To assess the reliability of these scores in predicting protein function, we used a data set established from a systematic experimental assessment of the transactivation activities of mutant forms of p53.

Although the structures of several domains of wild-type p53 have been described, there are currently no data available on the structure of the entire protein. Moreover, data on the structure of mutant proteins are still extremely limited. In the present study, we used the structure of the wild-type DBD of p53, bound to a consensus sequence, to construct the Delaunay tessellation. This structure was originally reported by Cho et al. [1994]. The original structure data file (PDB: 1TSR) contains three p53 monomers associated with one 10-mer DNA oligomer matching the p53-binding consensus. We used data for monomer B (residues 96–289), the only one that makes contacts in both major and minor grooves of target DNA, and is thus the best representation available of the conformational constraints involved in p53 binding to its target sequence.

The lowest RS were observed for mutations in the zinc-binding motif, the configuration of which is essential for the correct folding and stability of the DBD of p53 [Bullock et al., 1997; Hainaut and Mann, 2001]. Mutations at residues located in the hydrophobic core of the protein also had RS far from zero, consistent with previous results [Carter et al., 2001; Masso and Vaisman, 2003; Tropsha et al., 2003]. Because residues in the hydrophobic core play an important role in protein stability, mutating these residues is more likely to have a profound effect on local protein geometry. These observations are in agreement with previous findings that the stability of mutants at residues in the zinc binding regions or the beta-sandwich (hydrophobic core) was lower than that for mutants at residues in other structural domains of the protein [Bullock and Fersht, 2001; Bullock et al., 1997]. Thus, overall there was a good concordance between RS and structural features of the protein. Moreover, there was a trend for association between low RS values and NF groups of mutants. Depending on the promoter, the average RS varied between  $-1.64$  and  $-1.04$  for NF groups, and  $-0.46$  to  $-0.023$  for F groups ( $P < 0.05$ ).

Despite this general concordance among RS, structural effect, and loss of transcriptional activity, RS appeared to have limited accuracy for predicting the transactivation activity of individual mutants using a KNN algorithm. The accuracy of the prediction was not improved by using other parameters, such as solvent accessibility and/or Grantham scores in combination with RS. RS gives a global measure of changes in residue composition between wild-type and mutant proteins, but does not capture the impact of local changes that affect the mutant residue and its structural neighbors. To take these local changes into account, we first constructed, for each mutant, a vector  $\vec{V}$ , consisting of potential scores for each residue (sum of quadruplet likelihood scores in which a residue participates). Then we derived an RSP, calculated as the difference of  $\vec{V}(\text{mutant}) - \vec{V}(\text{wild-type})$ . In the RSP, only the mutant residue and its structural neighbors have nonzero values.

To handle the complexity generated by the high-dimensionality of RSP (194 dimensions, corresponding to the 194 residue positions in the protein structure), decision tree algorithms were applied to predict promoter-specific transactivation activities. This approach has the advantage of allowing the generation of different prediction models, depending on the promoter-specific input data sets. Decision tree algorithms (DTA) were preferred to other machine learning techniques, such as support vector machines

(SVM), mainly because decision trees result in interpretable rules for classifying mutants. We found that DTA yielded better promoter-specific predictions than the SVM models (data not shown). In a recent study Krishnan and Westhead [2003] compared the performance of SVM and DTA in predicting the functional effects of SNPs. They found that SVM models were more readily applicable to other proteins, but that DTA models worked better for protein-specific predictions.

Our predictions of transactivation activities using RSP correctly assigned mutants to either NF or F categories for seven of eight promoters, with accuracies ranging from 69.6% to 74.2%. An even better prediction accuracy was obtained with the group of “consistent” mutants (78.5%). This finding is in agreement with the notion that the “consistent” group includes mutants with either very limited (F on all promoters) or, in contrast, particularly severe (NF on all promoters) structural defects. In contrast, low prediction accuracy was found for the *MDM2* promoter (64.2%). Adding SA or GS to RSP did not change the prediction accuracies, which suggests that these variables do not add significant structural information in addition to what is captured by RSP. In addition, to verify that the increase in accuracy with RSP was indeed dependent on the use of RSP and not on the use of decision tree models, we derived decision tree models using RS. We found that the prediction accuracies resulting from decision tree models using RS were no higher than 56%. Thus the local structural impact taken into account in the RSP appears to be key for predicting transactivation.

Interestingly, there was a concordance between the accuracy of prediction for a given promoter and the degree of similarity between the p53 binding sequence in that promoter and the canonical p53 binding consensus. For example, the p53 binding sites in *GADD45* and *p53R2* promoters, which give the highest prediction accuracies, differed from the canonical consensus by only one base. In contrast, the p53 binding site of *MDM2*, which gave the worst prediction accuracy, differed from the consensus by three bases, including one of the invariant G in the central portion of the consensus. These results are compatible with the notion that wild-type p53 binds with different affinity to different promoters, and that *MDM2* may contain a promoter of lower affinity for p53 than most other p53-targeted genes.

Using the frequencies of mutations extracted from the IARC TP53 database as a separate data set, we found that 596 mutants in the DBD were predicted as NF, accounting for the majority (86%) of the missense mutants reported in the IARC TP53 database. The predictions were made on the 1,147 possible mutants using the model obtained with the group of 683 “consistent” mutants (the one yielding the highest prediction accuracy, 78%). Thus, using transactivation activities on 60% of the mutants we demonstrated a concordance between structural/functional properties and occurrence in cancer for all missense mutants in the DBD. These results show the biological significance of the predicted functional categories, since the mutants present in the database are those associated with human cancer and thus are more likely to be NF.

We analyzed the decision tree models in an attempt to identify residues with particular potential score difference values that would best classify the mutation as F or NF. Residues with scores that best discriminate between the two functional categories (that is, residues that fall at the root of the decision tree in the model), varied from model to model (depending on the promoter). These residues did not belong to a clear structural or functional category. Furthermore, there was no obvious structural rationale to account for misclassified mutants. In particular, these misclassified mutants

did not show obvious patterns with respect to solvent accessibility, type of amino-acid change (hydrophobic to hydrophobic, polar to polar, hydrophobic to polar, and polar to hydrophobic), or presence in patch/not in patch. These observations underline some of the limitations of our approach. First, calculations of RS and RSP are highly dependent on the input protein structure. In the present study all calculations were performed using a single structure of the core domain of p53 in complex with DNA. Using a complete p53 model structure, including the transactivation and oligomerization domains, should yield better prediction results. Unfortunately, such a model has not yet been established. On the other hand, since p53 binds DNA as tetramer [Anderson and Appella, 2003], tessellating structural models of p53 tetramers would undoubtedly reveal interactions that cannot be accounted for in this simplified DBD model. Second, using the same structure to calculate the RS makes the assumption that the mutant and wild-type structures are similar enough to yield the same overall Delaunay tessellation. The use of p53-specific antibodies that detect different conformations of p53 has shown that p53 has a relatively unstable structure [Legros et al., 1994]. Recent studies of mutant structures of the DBD [Joerger et al., 2004, 2005] suggested that some mutants cause substantial structural changes with respect to the wild-type structure (unbound to DNA) while others do not. Using the mutant structures to calculate RS and RSP may thus provide more accurate functional predictions, especially for mutations that cause larger structural changes. Third, it may be feasible to include information about bound DNA when constructing the tessellation. However, there may not be a large enough set of representative DNA-bound protein structures (that can be tessellated) in the PDB to constitute a training set. Despite these limitations, the Delaunay tessellation approach yields encouraging results, and the calculations of the RSP can be made rapidly. Furthermore, unlike most other structure-based methods (i.e., molecular dynamics and homology modeling), our approach has comparatively low input requirements and therefore allows the analysis of all possible missense substitutions for a given protein with known tertiary structure.

In conclusion, our results show that computational methods based on scores derived from Delaunay tessellations can predict with about 78% accuracy the transactivation activity, evaluated using a yeast functional assay, of p53 mutant proteins. The method is computationally inexpensive and can thus quickly generate the RSP of all possible mutants in a given protein whose tertiary structure is known. We expect our method to be applicable to other proteins for which a sufficient amount of functional data are available to initially train the machine learning models.

## ACKNOWLEDGMENTS

We thank Zhibin Lu for writing the JAVA programs that call Qhull application and calculate the potential scores, and Majid Masso for helpful discussions about the RSP. E. Mathe was supported by a Special Training Award from the International Agency for Research on Cancer (IARC), funded by the National Institute of Environmental Health Sciences (NIEHS).

## REFERENCES

- Anderson CW, Appella E. 2003. Signaling to the p53 tumor suppressor through pathways activated by genotoxic and non-genotoxic stresses. In: Bradshaw RA, Dennis E, editors. Handbook of cell signaling. New York: Academic Press. p 237–247.

- Barber CB, Dobkin DP, Huhdanpaa H. 1996. The Quickhull algorithm for convex hulls. *ACM Trans Mathematical Software* 22:469–483.
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The protein data bank. *Nucleic Acids Res* 28:235–242.
- Bullock AN, Henckel J, DeDecker BS, Johnson CM, Nikolova PV, Proctor MR, Lane DP, Fersht AR. 1997. Thermodynamic stability of wild-type and mutant p53 core domain. *Proc Natl Acad Sci USA* 94:14338–14342.
- Bullock AN, Fersht AR. 2001. Rescuing the function of mutant p53. *Nat Rev Cancer* 1:68–76.
- Capriotti E, Fariselli P, Casadio R. 2004. A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics* 20(Suppl 1):I63–I68.
- Carter CW Jr, LeFebvre BC, Cammer SA, Tropsha A, Edgell MH. 2001. Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. *J Mol Biol* 311:625–638.
- Cho Y, Gorina S, Jeffrey PD, Pavletich NP. 1994. Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. *Science* 265:346–355.
- Cooper DN, Ball EV, Krawczak M. 1998. The human gene mutation database. *Nucleic Acids Res* 26:285–287.
- Dambosky J, Prokop M, Koca J. 2001. TRITON: graphic software for rational engineering of enzymes. *Trends Biochem Sci* 26:71–73.
- el-Deiry WS, Kern SE, Pietenpol JA, Kinzler KW, Vogelstein B. 1992. Definition of a consensus binding site for p53. *Nat Genet* 1:45–49.
- Ferrer-Costa C, Orozco M, de la Cruz X. 2002. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *J Mol Biol* 315:771–786.
- Grantham R. 1974. Amino acid difference formula to help explain protein evolution. *Science* 185:862–864.
- Greenblatt MS, Beaudet JG, Gump JR, Godin KS, Trombley L, Koh J, Bond JP. 2003. Detailed computational study of p53 and p16: using evolutionary sequence analysis and disease-associated mutations to predict the functional consequences of allelic variants. *Oncogene* 22:1150–1163.
- Hainaut P, Mann K. 2001. Zinc binding and redox control of p53 structure and function. *Antioxid Redox Signal* 3:611–623.
- Hofseth LJ, Hussain SP, Harris CC. 2004. p53: 25 years after its discovery. *Trends Pharmacol Sci* 25:177–181.
- Hubbard S, Thornton J. 1992–1996. NACCESS V2.1.1: Solvent accessible area calculations. Version V2.1.1.
- Hutchinson EG, Thornton JM. 1996. PROMOTIF—a program to identify and analyze structural motifs in proteins. *Protein Sci* 5:212–220.
- Joerger AC, Allen MD, Fersht AR. 2004. Crystal structure of a superstable mutant of human p53 core domain. Insights into the mechanism of rescuing oncogenic mutations. *J Biol Chem* 279:1291–1296.
- Joerger AC, Ang HC, Veprintsev DB, Blair CM, Fersht AR. 2005. Structures of p53 cancer mutants and mechanism of rescue by second-site suppressor mutations. *J Biol Chem* 280:16030–16037.
- Jones S, Thornton JM. 1997. Analysis of protein–protein interaction sites using surface patches. *J Mol Biol* 272:121–132.
- Karchin R, Kelly L, Sali A. 2005. Improving functional annotation of non-synonymous SNPs with information theory. *Pac Symp Biocomput* 1:397–408.
- Kato S, Han SY, Liu W, Otsuka K, Shibata H, Kanamaru R, Ishioka C. 2003. Understanding the function–structure and function–mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. *Proc Natl Acad Sci USA* 100:8424–8429.
- Kelley LA, MacCallum R, Sternberg MJE. 1999. Recognition of remote protein homologies using three-dimensional information to generate a position specific scoring matrix in the program 3D-PSSM. In: Sorin Istrail PP, Michael Waterman, editors. *New York: Association for Computing Machinery*. p 218–225.
- Kelley LA, MacCallum RM, Sternberg MJ. 2000. Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J Mol Biol* 299:499–520.
- Krishnamoorthy B, Tropsha A. 2003. Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations. *Bioinformatics* 19:1540–1548.
- Krishnan VG, Westhead DR. 2003. A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. *Bioinformatics* 19:2199–2209.
- Laskowski RA. 2001. PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Res* 29:221–222.
- Legros Y, Meyer A, Ory K, Soussi T. 1994. Mutations in p53 produce a common conformational effect that can be detected with a panel of monoclonal antibodies directed toward the central part of the p53 protein. *Oncogene* 9:3689–3694.
- Martin AC, Facchiano AM, Cuff AL, Hernandez-Boussard T, Olivier M, Hainaut P, Thornton JM. 2002. Integrating mutation data and structural analysis of the TP53 tumor-suppressor protein. *Hum Mutat* 19:149–164.
- Masso M, Vaisman II. 2003. Comprehensive mutagenesis of HIV-1 protease: a computational geometry approach. *Biochem Biophys Res Commun* 305:322–326.
- Ng PC, Henikoff S. 2001. Predicting deleterious amino acid substitutions. *Genome Res* 11:863–874.
- Ng PC, Henikoff S. 2002. Accounting for human polymorphisms predicted to affect protein function. *Genome Res* 12:436–446.
- Olivier M, Eeles R, Hollstein M, Khan MA, Harris CC, Hainaut P. 2002. The IARC TP53 database: new online mutation analysis and recommendations to users. *Hum Mutat* 19:607–614.
- Olivier M, Hussain SP, Caron de Fromental C, Hainaut P, Harris CC. 2004. TP53 mutation spectra and load: a tool for generating hypotheses on the etiology of cancer. *IARC Sci Publ* 157:247–270.
- Poupon A. 2004. Voronoi and Voronoi-related tessellations in studies of protein structure and interaction. *Curr Opin Struct Biol* 14:233–241.
- Prokop M, Damborsky J, Koca J. 2000. TRITON: in silico construction of protein mutants and prediction of their activities. *Bioinformatics* 16:845–846.
- Ramensky V, Bork P, Sunyaev S. 2002. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res* 30:3894–3900.
- Schwede T, Kopp J, Guex N, Peitsch MC. 2003. SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res* 31:3381–3385.
- Singh RK, Tropsha A, Vaisman TA II. 1996. Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues. *J Comput Biol* 3:213–221.
- Stitzel NO, Tseng YY, Pervouchine D, Goddeau D, Kasif S, Liang J. 2003. Structural location of disease-associated single-nucleotide polymorphisms. *J Mol Biol* 327:1021–1030.
- Stitzel NO, Binkowski TA, Tseng YY, Kasif S, Liang J. 2004. topoSNP: a topographic database of non-synonymous single

- nucleotide polymorphisms with and without known disease association. *Nucleic Acids Res* 32:D520–D522.
- Tropsha A, Carter CW Jr, Cammer S, Vaisman TA II. 2003. Simplicial neighborhood analysis of protein packing (SNAPP): a computational geometry approach to studying proteins. *Methods Enzymol* 374:509–544.
- Vaisman II, Tropsha A, Zheng W. 1996. Compositional preferences in quadruplets of nearest neighbor residues in protein structures: statistical geometry analysis. *IEEE Symp Intell Syst* 163–168.
- Vitkup D, Sander C, Church GM. 2003. The amino-acid mutational spectrum of human genetic disease. *Genome Biol* 4:R72.
- Walker DR, Bond JP, Tarone RE, Harris CC, Makalowski W, Boguski MS, Greenblatt MS. 1999. Evolutionary conservation and somatic mutation hotspot maps of p53: correlation with p53 protein structural and functional features. *Oncogene* 18: 211–218.
- Weiss G, Provost F. 2003. Learning when training data are costly: the effect of class distribution on tree induction. *J Artif Intell Res* 10:315–354.
- Witten IH, Eibe F. 2000. *Data mining: practical machine learning tools with Java implementations*. San Francisco: Morgan Kaufmann 416p.