

# Functional Impact of Missense Variants in BRCA1 Predicted by Supervised Learning

Rachel Karchin<sup>1,2\*</sup>, Alvaro N. A. Monteiro<sup>3</sup>, Sean V. Tavtigian<sup>4</sup>, Marcelo A. Carvalho<sup>3</sup>, Andrej Sali<sup>5,6\*</sup>

**1** Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland, United States of America, **2** Institute of Computational Medicine, Johns Hopkins University, Baltimore, Maryland, United States of America, **3** Risk Assessment, Detection, and Intervention Program, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida, United States of America, **4** International Agency for Research on Cancer, Lyon, France, **5** Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, California, United States of America, **6** California Institute for Quantitative Biomedical Research, University of California San Francisco, San Francisco, California, United States of America

**Many individuals tested for inherited cancer susceptibility at the *BRCA1* gene locus are discovered to have variants of unknown clinical significance (UCVs). Most UCVs cause a single amino acid residue (missense) change in the *BRCA1* protein. They can be biochemically assayed, but such evaluations are time-consuming and labor-intensive. Computational methods that classify and suggest explanations for UCV impact on protein function can complement functional tests. Here we describe a supervised learning approach to classification of *BRCA1* UCVs. Using a novel combination of 16 predictive features, the algorithms were applied to retrospectively classify the impact of 36 *BRCA1* C-terminal (*BRCT*) domain UCVs biochemically assayed to measure transactivation function and to blindly classify 54 documented UCVs. Majority vote of three supervised learning algorithms is in agreement with the assay for more than 94% of the UCVs. Two UCVs found deleterious by both the assay and the classifiers reveal a previously uncharacterized putative binding site. Clinicians may soon be able to use computational classifiers such as those described here to better inform patients. These classifiers can be adapted to other cancer susceptibility genes and systematically applied to prioritize the growing number of potential causative loci and variants found by large-scale disease association studies.**

Citation: Karchin R, Monteiro ANA, Tavtigian SV, Carvalho MA, Sali A (2007) Functional impact of missense variants in *BRCA1* predicted by supervised learning. *PLoS Comput Biol* 3(2): e26. doi:10.1371/journal.pcbi.0030026

## Introduction

The *BRCA1* gene encodes a large multifunction protein involved in cell-cycle and centrosome control, transcriptional regulation, and in the DNA damage response [1–3]. Inherited mutations in this gene have been associated with an increased lifetime risk of breast and ovarian cancer (6–8 times that of the general population) [4]. There are several thousand known deleterious *BRCA1* mutations that result in frame-shifts and/or premature stop codons, producing a truncated protein product [5]. In contrast, the functional impact of most missense variants that result in a single amino acid residue change in *BRCA1* protein is not known. The Breast Cancer Information Core database (<http://research.nhgri.nih.gov/bic/>), a central repository of *BRCA1* and *BRCA2* mutations identified in genetic tests, currently contains 487 unique missense *BRCA1* variants (April 2006), of which only 17 have sufficient genetic/epidemiological evidence to be classified as deleterious (Clinically Important) and 33 as neutral or of little clinical importance (Not Clinically Important). As genetic testing for inherited disease predispositions becomes more commonplace, predicting the clinical significance of missense variants and other UCVs will be increasingly important for risk assessment.

Because most UCVs in *BRCA1* and *BRCA2* occur at very low population frequencies (<0.0001) [6], direct epidemiological measures, such as familial cosegregation with disease, are often not sufficiently powerful to identify the variants associated with cancer predisposition. A promising approach is to supplement epidemiological and clinical analysis of UCVs with indirect approaches such as biochemical studies of

protein function and bioinformatics analysis [6–8]. In the future, physicians and genetic counselors may be able to rely on all these sources of information about UCVs when counseling their patients.

Previous bioinformatics analysis of *BRCA1* UCVs has depended primarily on measures of evolutionary conservation in multiple sequence alignments of human *BRCA1* and related proteins from other organisms [9–11]. Two groups have attempted to include information about *BRCA1* protein structure. Williams et al. predicted the impact of 25 missense variants in *BRCA1*'s C-terminal *BRCT* domains by considering both conservation and location of variant amino acid residues in an X-ray crystal structure [12]. Variants were predicted deleterious if their properties were similar to

**Editor:** Greg Tucker-Kellogg, Lilly Systems Biology, Singapore

**Received** September 5, 2006; **Accepted** December 27, 2006; **Published** February 16, 2007

A previous version of this article appeared as an Early Online Release on December 28, 2006 (doi:10.1371/journal.pcbi.0030026.eor).

**Copyright:** © 2007 Karchin et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Abbreviations:** Align-GVGD, Align Grantham Variation Grantham Deviation; AUC, area under the ROC curve; BIC, breast information core database; BRCT, *BRCA1* C-terminal domains; BRCT-C, BRCT C-terminal domain; BRCT-N, BRCT N-terminal domain; GD, Grantham Deviation; GV, Grantham Variation; ROC, receiver operating characteristic; Rule-based decision tree, empirically derived rules encoded in a decision tree; SIFT, Sorting Intolerant from Tolerant; UCV, variant of unknown clinical significance

\* To whom correspondence should be addressed. E-mail: karchin@karchinlab.org (RK); sali@salilab.org (AS)

## Author Summary

A significant number of breast and ovarian cancers are due to inherited mutations in the BRCA1 and BRCA2 genes. Many women who receive genetic testing for these mutations are found to have variants of the genes that result in changed amino acids in the BRCA1 or BRCA2 proteins. The effect of these variants on cancer risk is not well-understood, posing a problem for patients and their health providers. We describe computational biology methods that predict and analyze the impact of 36 BRCA1 variants on protein function. The predictions are validated by biochemical assays of BRCA1 in yeast and mammalian cell cultures. The speed and accuracy of the computational methods is well-suited to rapid evaluation of large numbers of variants in genes that predispose to inherited diseases.

properties of biochemically characterized deleterious variants in *Escherichia coli* Lac Repressor and bacteriophage T4 lysozyme. Mirkovic et al. developed a set of hierarchical rules (Rule-based decision tree) based on the conservation, variant structural location, and amino acid residue physiochemical properties of 30 deleterious and seven neutral biochemically characterized BRCA1 missense variants [7].

We have developed a novel combination of 16 predictive features that describe conservation, impact of mutation on protein structure, and amino acid residue properties, and used them as input to computational supervised learning algorithms. These algorithms are trained to learn a generic classification of amino acid residue substitutions and positional contexts. The training set is composed of 618 missense variants in the transcription factor TP53 biochemically characterized as functional or nonfunctional in a transactivation assay [13]. TP53 is a tumor suppressor gene that is inactivated in the majority of human cancers.

Our validation set is composed of 36 missense variants in BRCA1's BRCT domains that were biochemically characterized with a transactivation assay [14]. These 36 variants were selected because they occur in individuals from families with breast or ovarian cancer in which no other deleterious mutation in *BRCA1* or *BRCA2* was found and were functionally tested under the same protocols and conditions, yielding standardized measurements of each variant's transactivation activity with respect to wild-type. We use the validation set to assess the supervised learners and compare them with algorithms based on evolutionarily allowed amino acid residues or empirically derived rules. The algorithms with greatest correlation between assay and computational predictions are the supervised learners Naïve Bayes [15], Support Vector Machine [16], and Random Forest [17].

Given a protein X-ray crystal structure, the supervised learning approach can quickly and accurately predict the outcome of our BRCA1 transactivation assay with greater than 94% accuracy on tested missense variants in the BRCT domains. We have applied the best performing supervised learners to blind prediction of the functional impact of 54 UCVs found in BIC and occurring in the BRCA1 BRCT domains. For each of these UCVs, we produce a consensus prediction and, where possible, a molecular explanation for the impact of the variant.

Next, we describe the protocol used to train and validate the supervised learning algorithms, the selection of 16

features used to represent each missense variant to the algorithms, implementation details of each algorithm, and performance assessment criteria (Methods). We then show how a combination of sequence- and structure-based features in a supervised learning setting obviates some of the problems with evolutionary analysis and empirically derived rules, providing specific examples of the strengths and weaknesses of each approach (Results, Discussion). We show that two of the variants found to be deleterious by both the assay and the classifiers may be at a previously uncharacterized protein binding site and that electrostatic changes at the site may weaken the interactions of BRCA1 and protein partners that are important for its functions (Discussion). Finally, we discuss the generalizability of our methods to other cancer susceptibility genes and to large-scale disease association studies (Discussion).

## Methods

### Training Set

We trained four supervised learning algorithms to discriminate between a set of 398 deleterious/nonfunctional and 220 neutral/functional TP53 missense variants, biochemically characterized in a transactivation assay [13]. The variants were downloaded from the IARC TP53 website (<http://www-p53.iarc.fr>). We only used variants capable or incapable of activating transcription for all eight of the TP53 promoters tested in the transactivation assay and located in the core DNA binding domain of TP53.

### Validation Set

The 36 BRCA1 BRCT missense variants described in our companion paper [14] were used as an independent validation set for the supervised learners. These variants were also classified by sequence-analysis methods based on evolutionarily allowed amino acid residues: Align Grantham Variation Grantham Deviation (Align-GVGD) [18], Sorting Intolerant from Tolerant (SIFT) [19], Ancestral Sequence [9,11], and empirically derived rules encoded in a decision tree (Rule-based decision tree) [7]. Each method was evaluated by its agreement with the BRCA1 transactivation assay on the validation set, according to accuracy (fraction of all variants correctly classified), sensitivity or true positive rate (fraction of all nonfunctional variants correctly classified), specificity or true negative rate (fraction of all functional variants correctly classified), Matthews correlation coefficient [20], and coverage (fraction of variants for which a prediction was made) (Table 1). Matthews correlation coefficient is defined as

$$MCC = \frac{(TP \quad TN) - (FP \quad FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (1)$$

and ranges from  $-1.0$  (worst) to  $1.0$  (best). A coefficient of  $0$  is equivalent to a random prediction, and less than  $0$  indicates a worse than random prediction.  $TP$  is the number of correctly classified nonfunctional variants,  $TN$  the number of correctly classified functional variants,  $FP$  the number of incorrectly classified nonfunctional variants, and  $FN$  the number of incorrectly classified functional variants.

For the Naïve Bayes, Support Vector Machine, Random Forest, Decision Tree, Align-GVGD, and SIFT classifiers, we computed a receiver operating characteristic (ROC) curve

**Table 1.** Classification Performance of Nine Computational Methods for UCV Classification, According to Agreement with the BRCA1 Transactivation Assay of 36 Missense Variants

Computational Method	Information	Method	Accuracy	True Positive Rate	True Negative Rate	Matthews Correlation Coefficient	Coverage
Naïve Bayes	Structure/sequence	Supervised learning	0.97	1.00	0.89	<b>0.93</b>	1.00
Support Vector Machine	Structure/sequence	Supervised learning	0.94	1.00	0.78	<b>0.85</b>	1.00
Random Forest	Structure/sequence	Supervised learning	0.94	1.00	0.78	<b>0.85</b>	1.00
Ancestral Sequence	Sequence	Sequence analysis	0.92	1.00	0.67	<b>0.77</b>	1.00
Align-GVGD Spur	Sequence	Sequence analysis	0.83	0.85	0.78	<b>0.59</b>	0.92
Align-GVGD Tnig	Sequence	Sequence analysis	0.83	0.89	0.67	<b>0.56</b>	0.92
SIFT	Sequence	Sequence analysis	0.78	0.78	0.78	<b>0.50</b>	1.00
Rule-based decision tree	Structure/sequence	Human rules	0.81	0.93	0.44	<b>0.43</b>	1.00
Decision Tree	Structure/sequence	Supervised learning	0.72	0.89	0.22	<b>0.14</b>	1.00

Results are sorted by Matthews correlation coefficient (bold). Shown are four supervised machine learning methods, trained on 618 biochemically characterized missense variants in the human transcription factor TP53. *Accuracy*, fraction of all variants correctly classified. *True Positive Rate*, fraction of correctly classified deleterious variants. *True Negative Rate*, fraction of correctly classified neutral variants. *Coverage*, fraction of the 36 mutants classified. doi:10.1371/journal.pcbi.0030026.t001

that quantifies the tradeoff between coverage of detected nonfunctional variants (*true positive rate*) and misclassified functional variants (*false positive rate* =  $1 - \text{specificity}$ ). ROC analysis was not possible for the Rule-based decision tree and Ancestral Sequence algorithms, which predict the class of a missense variant but do not provide an associated score.

### Feature Selection

The supervised learning algorithms (Naïve Bayes, Support Vector Machine, Random Forest, Decision Tree) were trained by associating each amino acid residue substitution in the TP53 training set with 16 carefully selected predictive features (Table 2). A vector of features for a single substitution is denoted as  $\vec{X}$ . The features describe properties of variant and wild-type residues: local structural environment; physiochemical attributes; and evolutionary conservation. To compute the features, we used DSSP (a program that calculates a variety of geometrical properties for each amino acid residue in a protein structure) [21], MODELLER for comparative protein structure modeling [22], SAM-T2K for protein sequence alignments and hidden Markov models [23], and in-house PERL code.

We began with a core set of 13 features selected by a correlation analysis between features and classes (functional or nonfunctional) of the TP53 variants, as described previously [24] (Table S3). An additional 18 candidate features were evaluated by adding them to the core set and doing 10-fold cross-validation tests of Support Vector Machine performance. Three features were found to improve performance and were added to the optimal feature set; the others were rejected (Tables S4 and S5). Next we evaluated Support Vector Machine performance with each of the best 16 features held out. In each case, the 10-fold cross-validation test yielded decreased performance.

### Protein Modeling

We used X-ray crystal structures from the Protein Data Bank [25] for the BRCA1 BRCT domains (1t29 chain A in complex with BACH1 peptide) [26] and the DNA binding domain of TP53 (1kzy) [27]. We performed *in silico* mutations on the structures with the MUTATE\_MODEL routine of

MODELLER (available as a Python script at [http://salilab.org/modeller/wiki/Mutate\\_model](http://salilab.org/modeller/wiki/Mutate_model)). MUTATE\_MODEL substitutes the wild-type amino acid residue at a position of interest with a variant amino acid residue, and optimizes the coordinates of the variant's backbone and sidechain atoms with an initial conjugate gradient minimization, molecular dynamics optimization with simulated annealing, and a final conjugate gradient minimization (E. Feyfant, 2004, private communication).

### Protein Sequence Alignments

The amino acid residue sequences of human TP53 (P04637) and BRCA1 (P38398) were downloaded from UNIPROT [28], and each was used as a seed sequence for the SAM-T2K iterative alignment-building algorithm [23]. For BRCA1, only amino acid residues in the BRCT domains (1649–1859) were aligned. We used the SAM w0.5 program to apply sequence weighting and regularization with Dirichlet mixtures [29] to each resulting alignment and to produce a profile hidden Markov model [30]. The TP53 and BRCA1 alignments and hidden Markov models are available upon request.

### Support Vector Machine

We trained a soft margin Support Vector Machine classifier with a radial basis kernel using the e1071 package in R [31]. The Support Vector Machine algorithm optimizes a vector of weights  $\vec{\alpha}$  (one weight for each training example) and a bias parameter  $b$ . The parameters  $g$  (radial basis kernel width) and  $C$  (penalty for violating the soft margin) were optimized on the training set with grid search using default parameters. Each of the 28 missense variants  $\vec{X}$  was then scored with the discriminant function

$$g(\vec{X}, \vec{X}_i) = \sum_{i=1}^l y_i \alpha_i k(\vec{X}, \vec{X}_i) + b \quad (2)$$

where  $l$  is the number of examples in the training set,  $y_i$  is the class label of each example in the training set (for deleterious/nonfunctional variants  $y_i = -1$  and for neutral/functional variants  $y_i = 1$ ), and  $K(\vec{X}, \vec{X}_i)$  is the value of the radial basis kernel function given  $\vec{X}$  and training example  $\vec{X}_i$ . Variants are

**Table 2.** Predictive Features Describing Evolutionary Conservation, Impact of Mutation on Protein Structure, and Amino Acid Residue Properties Used as Input to the Computational Supervised Learning Algorithms

Feature Category	Feature Description
<b>Structural</b>	Solvent Accessibility of wild-type amino acid residue ( $\text{\AA}^2$ )
	Solvent Accessibility of wild-type residue normalized by maximum exposed Solvent Accessibility of that residue type in a GLY-X-GLY tripeptide, using values given by Rose et al. [80]
	Solvent Accessibility of variant residue
	Normalized Solvent Accessibility of variant residue
	Number of methyl(ene) groups within 6 $\text{\AA}$ of the variant sidechain [81]
	Number of unsatisfied spatial restraints in the MODELLER objective function after in silico mutation and simulated annealing refinement of the variant <sup>a</sup>
	$\Phi$ and $\Psi$ backbone dihedral angles at the mutated position
<b>Physiochemical differences between wild-type and variant amino acid residues</b>	Whether the mutation results in buried charge
	Change in formal charge
	Change in volume ( $\text{\AA}^3$ ) [82]
<b>Evolutionary conservation of amino acid residues in protein orthologs</b>	Change in polarity [83]
	Grantham difference [37]
	Relative entropy estimated by amino acids in the variant's alignment column [84]
	Positional hidden Markov model conservation score based on the probabilities of the wild-type, variant, and most probable amino acid residue in the variant's alignment column <sup>b</sup> [24]

<sup>a</sup>Violated restraints suggest that the mutated sidechain introduced steric clashes or unusual geometries into the protein model. Examples of violated restraints include extreme values of the Lennard-Jones 6–12 potential [85], bond angle potential, bond length potential, sidechain dihedral angle restraints, and nonbonded restraints. Two thresholds are used to identify violated restraints yielding two features.

<sup>b</sup>The probabilities are estimated by a hidden Markov model built with SAM-T2K and the w0.5 script [23].

PHC =  $\log(p(\text{Wild-type}) - p(\text{Variant})) + \log(p(\text{Wild-type})) + \log(P(\text{Most Probable})) - \log(p(\text{Variant}))$

The features were computed for 618 TP53 missense variants, 36 BRCA1 BRCT missense variants biochemically characterized in our companion paper [14], and 54 BRCA1 BRCT UCVs found in BIC.

doi:10.1371/journal.pcbi.0030026.t002

classified as deleterious/nonfunctional if  $g(\vec{X}, \vec{X}_i) < 0$  and neutral/functional if  $g(\vec{X}, \vec{X}_i) > 0$ .

### Naïve Bayes

The Naïve Bayes algorithm estimates the probability that each variant belongs to deleterious or neutral classes  $C \in \{D, N\}$  by applying the Bayes rule:

$$P(C|\vec{X}) \propto P(\vec{X}|C)P(C) \quad (3)$$

where the prior class probability  $P(C)$  is the fraction of deleterious (or neutral) missense variants in the training set and each feature  $X_i$  is assumed to be conditionally independent of the  $k - 1$  other features, given its class membership, so that

$$P(\vec{X}|C) = \prod_{i=1}^k P(X_i|C) \quad (4)$$

where  $P(X_i|C)$  is estimated from the training set. We used the Naïve Bayes method in R's e1071 package. Each feature was approximated to be normally distributed and no smoothing was applied to the feature distributions.

### Decision Tree

We used the rpart package in R [32] to train a Decision Tree with the following parameters: *minsplit* = 20 (minimum number of observations required at a tree node before a split is attempted) and *cp* = 0 (no pruning of tree regardless of whether a split will improve model fit). To reduce overfitting, we pruned the resulting tree using the standard heuristic “1 Standard Error rule” [33] and 10-fold cross-validation. According to the 1 Standard Error rule, the pruned tree

with best generalization properties has a cross-validation error on the training set 1 Standard Error worse than the tree with the lowest cross-validation error. The pruning process yielded a reduced set of features:  $\Phi$  and  $\Psi$  mainchain dihedral angles, normalized solvent accessibility of wild-type, Grantham difference, volume change, relative entropy, and positional hidden Markov model conservation score.

### Random Forest

We used the randomForest package in R [34] to train a Random Forest, an algorithm based on a majority vote of a large number of decision trees, in which the candidate features at each tree node are randomly sampled [17]. The user-defined input parameters to randomForest are total number of trees in the forest and *mtry* (number of randomly sampled features considered as candidates for a split at each tree node). Both were selected with grid-search optimization as described for the Support Vector Machine [31].

### Log Likelihood Ratios

Predictions of Naïve Bayes, Decision Tree, and Random Forest are in the form of class conditional probabilities, where the two classes are  $D$  (deleterious/nonfunctional) and  $N$  (neutral/functional). For each example, the classifiers report  $P(D|\vec{X})$  (probability that the variant is deleterious, given feature vector  $\vec{X}$ ) and  $P(N|\vec{X})$  (probability that the variant is neutral, given feature vector  $\vec{X}$ ). To evaluate accuracy, true positive rate, true negative rate, and Matthews correlation coefficient, we classified variants as deleterious if  $P(D|\vec{X}) >$

0.5 and neutral otherwise. To compute ROC curves, we used the log likelihood ratio

$$LLR = \text{Log} \left( \frac{P(N|\vec{X})}{P(D|\vec{X})} \right) \quad (5)$$

as the output score of Naïve Bayes, Decision Tree, and Random Forest.

### SIFT

The SIFT algorithm [19] predicts the probability that a missense mutation occurs at a given alignment position. Variants that occur at conserved alignment positions are expected to be tolerated less than those that occur at diverse positions. The algorithm uses a modified version of PSI-BLAST [35] and Dirichlet mixture regularization [29] to construct a multiple sequence alignment of proteins that can be globally aligned to the query sequence and belong to the same clade. We used the SIFT server (<http://blocks.fhcrc.org/sift/SIFT.html>), with PSI-BLAST search set to the Swissprot-TrEMBL protein sequence database [36]. Both the full-length human BRCA1 sequence (amino acid residues 1–1863) and the BRCT C-terminal domain sequence only (amino acid residues 1649–1859) were submitted to the server. For the full-length sequence, SIFT reported low confidence predictions for 34 out of 36 missense variants. Consequently, we based our SIFT predictions on the C-terminal domain sequence. To compute accuracy, true positive rate, true negative rate, and Matthews correlation coefficient, we used the binary class predictions of the SIFT server (deleterious or neutral), based on the default SIFT threshold (*tolerated mutation probability* > 0.05). For ROC analysis, we used the raw SIFT probabilities, which range from 0 to 1.

### Ancestral Sequence

The Ancestral Sequence classifications were computed as described [9,11]. Each position in an alignment of eight mammalian *BRCA1* orthologs identified as giving best results by Pavlicek et al. was categorized as fixed (completely conserved), conserved (substitution of similar amino acid residues), or nonconserved (dissimilar amino acid residues or gaps). Any substitution at a fixed position and any non-conservative substitution at a conserved position is classified as deleterious. Amino acid residue similarity is based on the Gonnet PAM250 score (i.e., the likelihood that amino acid residue A has mutated into amino acid residue B in a pair of sequences that have diverged by 250 mutations per 100 amino acid residues of sequence) [37].

### Align-GVGD

The Align-GVGD method calculates two scores for each amino acid residue substitution, Grantham Deviation (GD) and Grantham Variation (GV), based on a modified Grantham distance measure [18,38]. The scores define four categories of missense variants: “Enriched deleterious 1” variants occur at invariant alignment positions for which the substitution is outside the range of variation observed at the position ( $GV = 0$ ,  $GD > 0$ ); “Enriched deleterious 2” occur at variable alignment positions containing physiochemically similar amino acid residues where the substitution is outside the range of observed variation ( $0 < GV < 61.3$ ,  $GD = 0$ ); “Enriched neutral 1” occur at variable positions containing

physiochemically similar amino acid residues where the substitution is inside the range of variation ( $GV > 0$ ,  $GD = 0$ ); and “Enriched neutral 2” occur at variable positions containing dissimilar amino acid residues where the substitution is slightly outside the range of variation ( $GV > 61.3$ ,  $0 < GD < 61.3$ ). We classified variants using first an alignment of placental mammals, a marsupial (gray short-tailed opossum), chicken, frog, and the pufferfish *Tetraodon* (“Align-GVGD Tnig”), and second an alignment that also includes the sea urchin *Strongylocentrotus purpuratus* (“Align-GVGD Spur”). Accuracy, true positive rate, true negative rate, and Matthews correlation coefficient were evaluated by reducing the four categories to deleterious/nonfunctional or neutral/functional. Variants may have GV and GD values that do not match any of the four categories (e.g., a variant with  $GV = 80$  and  $GD = 80$ ), which lowers coverage, Matthews correlation coefficient, true positive, and false positive rates. For ROC analysis, rather than fixing thresholds on GV and GD at 61.3 for each substitution, we considered the number of true positives and false positives over a range of thresholds, from the smallest to largest values of GV and GD in our dataset (0 to 215).

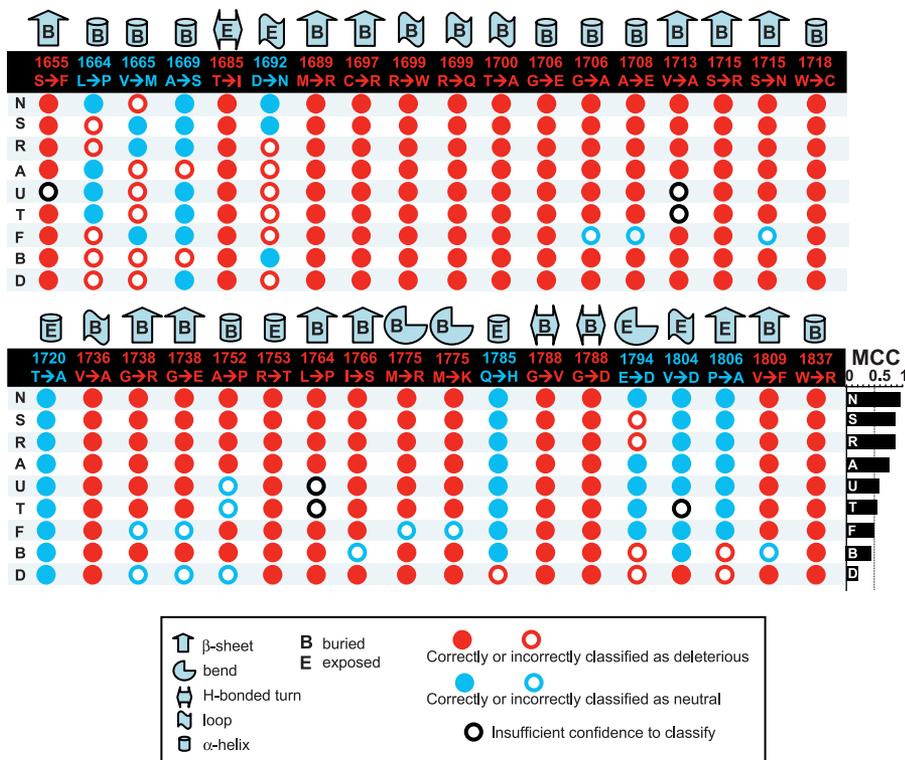
### Rule-Based Decision Tree

A Rule-based decision tree is a classification tree with human-designed rules that uses both structure- and sequence-based information, implemented in PERL [7]. Rule-based Decision Tree classifies a missense variant as either deleterious/nonfunctional or neutral/functional, but does not compute numerical scores.

### Structural Variant Analysis

The structural models of all BRCA1 BRCT missense variants were visually compared with the wild-type structure (1t29) using the molecular graphics program Chimera [39]. We explored changes in hydrogen bonding patterns and geometric properties of the molecular surface with Chimera’s FindHBond and MSMS routines. To visualize the distribution of amino acid residue conservation on the protein surface, the RenderByAttribute routine was used, with coloration defined by percent conserved in a hand-edited SAM-T2K alignment of BRCA1 orthologs. Species used in this alignment were *Homo sapiens* (AAA 73985), *Pan troglodytes* (AAG43492), *Gorilla gorilla* (AAT44835), *Pongo pygmaeus* (AAT44834), *Macaca mullata* (AAT44833), *Canis familiaris* (AAC48663), *Bos taurus* (AAL76094), *Monodelphis domestica* (AAX92675), *Mus musculus* (AAD00168), *Rattus norvegicus* (AAC36493), *Gallus gallus* (AAK83825), *Xenopus laevis* (AAL13037), and *Tetraodon nigroviridis* (AAR89523).

A highly conserved surface patch was identified as a possible binding site and subjected to further analysis. We used DELPHI [40] to compute the electrostatic surface potential at the putative binding site for the wild-type structure and for models of two solvent-exposed variants characterized as deleterious in our functional assays (T1685A and R1753T) [14]. The solvent relative dielectric constant was set to 4.0, the protein relative dielectric constant to 20.0, and ionic strength to the physiological value of 0.2 mM. Charges were estimated with the united atom AMBER model [41]. The proteins were prepared for DELPHI by adding heavy atoms missing from the 1t29 crystal structure with MODELLER’s COMPLETE\_PDB routine and adding hydrogens with



**Figure 1.** Computational Classifications of 36 BRCA1 BRCT Variants Functionally Characterized by the Transactivation Assay  
 For each variant, the local protein structure environment is represented by secondary structure type and whether the amino acid residue is buried (*normalized solvent accessibility* < 0.2) or exposed (*normalized solvent accessibility* ≥ 0.2). Labels (“1655 S->F”) are colored according to whether the variant was functional in the assays (blue) or nonfunctional (red). Computational classifications in agreement with the assay are indicated by filled circles. Computational classifications not in agreement with the assay are indicated by outlined circles. Computational classifications yielding “unclassified” are indicated by an outlined black circle. The variant D1692N is fully functional as a transcriptional activator but results in incorrect splicing in vivo. Results from variant M1775K are unpublished (Foulkes et al.).  
 A, Ancestral Sequence; B, Rule-based decision tree; D, Decision Tree; F, SIFT; MCC, Matthews correlation coefficient; N, Naïve Bayes; R, Random Forest; S, Support Vector Machine; T, Align-GVGD Tnig; U, Align-GVGD Spur.  
 doi:10.1371/journal.pcbi.0030026.g001

REDUCE [42], then visualized in Chimera with a GRASP surface representation [43].

**Results**

We compared the transactivation activity of the wild-type LexA DBD-BRCA1 or GAL4DBD-BRCA1 fusion construct in both yeast and mammalian cells with the activity of constructs containing 36 single missense variants in the BRCT domains [14]. Variant constructs presenting 50% or more of wild-type activity are characterized as neutral and those with 45% or less are characterized as deleterious, thresholds that are in agreement with available genetic evidence. These functional characterizations were used as a standard to evaluate the reliability of nine computational classifiers (Figure 1). We also provide a post-prediction analysis of these variants (Table S1). Three classifiers with the highest correlation to the functional assay were applied to predict the impact of 54 UCVs in the BRCT domains currently listed in the Breast Information Core database.

**Algorithm Evaluation**

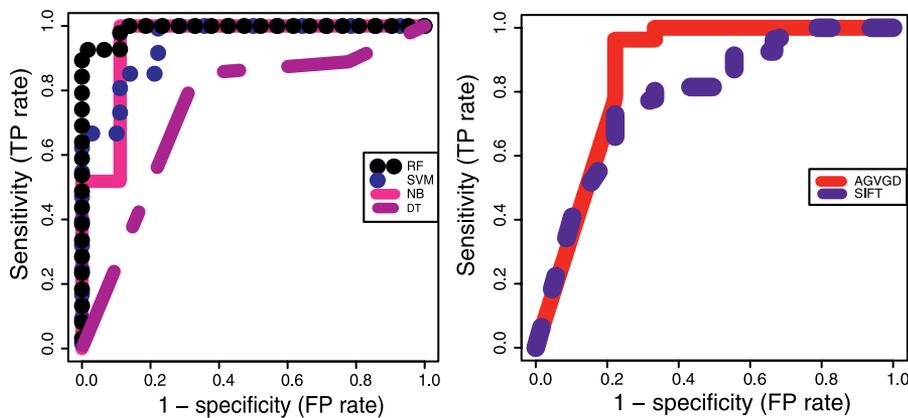
Based on ROC analysis, the supervised learners Random Forest, Support Vector Machine, and Naïve Bayes yield the most reliable computational classifications of the 36 variants (Figure 2). The area under the ROC curve (AUC) quantifies

the probability that a classifier will give a randomly drawn deleterious example a lower score than a randomly drawn neutral example. AUC is 0.992 for Random Forest, 0.947 for Support Vector Machine and Naïve Bayes, 0.86 for Align-GVGD Tnig, 0.852 for Align-GVGD Spur, 0.783 for SIFT, and 0.738 for Decision Tree (Figure 2). The Decision Tree algorithm appears to overfit the training set and generalizes less well than the other supervised learners.

Three of the supervised learning algorithms (Naïve Bayes, Support Vector Machine, and Random Forest) produce the best classifications of the 36 variants, as measured by accuracy, true positive rate, true negative rate, Matthews correlation coefficient, and coverage, using default thresholds (Table 1). According to these statistical measures, the best sequence analysis methods are Ancestral Sequence and Align-GVGD. Random Forest, Naïve Bayes, and Support Vector Machine are the most accurate scoring predictors, according to the AUC. The methods rankings are slightly different when evaluated by threshold-dependent statistics that reduce predictive scores to deleterious/neutral classes or by the score-based and threshold-independent ROC statistic of AUC.

**BIC BRCA1 UCVs in BRCT Domains**

We applied the top performing algorithms (Naïve Bayes, Support Vector Machine, and Random Forest) to predict the



**Figure 2.** Sensitivity versus 1-Specificity of Classifiers That Use a Numerical Score to Predict the Functional Impact of 34 BRCA1 BRCT UCVs

Comparison of four supervised machine learning methods, trained on 618 biochemically characterized missense variants in the human transcription factor TP53 and two sequence analysis methods that consider evolutionary conservation and physicochemical properties of amino acids (SIFT and Align-GVGD Tnig based on alignment of eight placental mammals, marsupial, chicken, frog, and pufferfish). Align-GVGD Spur, using an alignment that includes these species plus sea urchin, performs slightly worse than Align-GVGD Tnig in terms of ROC analysis and is not shown. Plot created with ROCR [86].

DT, decision tree; NB, Naive Bayes; RF, random forest; SVM, support vector machine.  
doi:10.1371/journal.pcbi.0030026.g002

impact of 54 BRCA1 UCVs listed in BIC that (a) are located in the BRCT domains, and (b) have not been functionally characterized by our transactivation assays. Based on a majority vote of the computational predictors, we computed a “consensus prediction” for each UCV (Figure 3). We provide structural explanations for the impact of as many variants as possible, and indicate where the predictions are supported by biochemical experiments found in the literature (Table S2).

The predicted deleterious UCVs are predominantly in the core secondary structure elements of the BRCT domains, rather than in loops, particularly in the  $\beta$  sheet of BRCT-N ( $\beta$ 1,  $\beta$ 3, and  $\beta$ 4), and in helix  $\alpha$ '3 and the turn connecting helix  $\alpha$ '1 and strand  $\beta$ '2 in BRCT-C (Figure 4).

### Binding-Site Prediction

We observed a patch of highly conserved amino acid residues that form a groove on the BRCA1 surface, on the opposite face from the known phosphopeptide binding cleft (Figure 5A–5C). These residues are T1684, T1685, H1686, K1711, W1712, and R1753. Both T1685 and H1686 have been shown to be highly sensitive to mutation, and our companion paper [14] contains new experimental evidence that R1753T has markedly reduced transactivation activity in both yeast and mammalian cells. The groove residues form hydrogen bonds with each other and several other conserved residues, including S1651, V1687, T1681, G1706, E1731, E1735, and P1749, producing two hydrogen bonding networks. The first network is found in BRCT-N (S1651, T1684, T1685, H1686, V1687, G1706, K1711, W1712, E1731) and the second network connects BRCT-N residues with the linker region that connects BRCT-N and BRCT-C (E1735, P1749, R1753) (Figure 5A–5C). All the residues lining this groove are completely conserved in our alignment of *BRCA1* orthologs, except for T1684, which is conserved in all orthologs except for *Tetraodon* (pufferfish), the organism most distant from human in our alignment (Figure 5A). Previous studies have shown that G1706 and P1749 are also sensitive to mutation [44,45] (S. Marsillac, 2006, private communication). The proposed

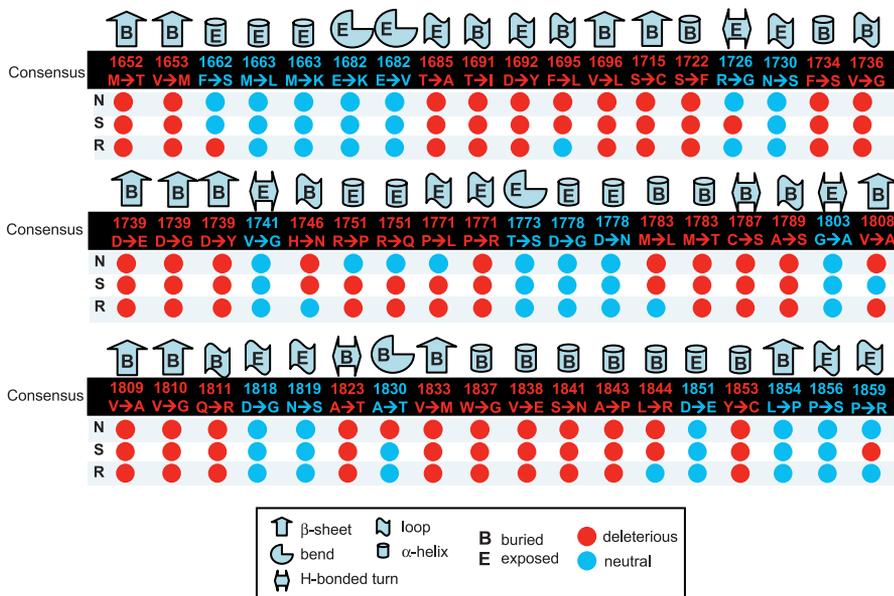
binding site would be specific to BRCA1, as most of these positions (except for H1686) are not highly conserved across tandem BRCT repeats in MDC1, PTIP, BARD1, and 53BP1.

The solvent-exposed missense variant R1753T found at the proposed binding site has <20% of the wild-type transactivation activity in yeast and <5% in mammalian cells [14], suggesting that the wild-type arginine amino acid residue might be important for binding of BRCA1 to a protein partner (or nucleic acid ligand). Although the mechanism of the BRCA1 BRCT domains in transactivation is not known, it is believed to depend on interactions with a variety of partners [46,47]. The mutation of R1753 to a threonine changes the local electrostatic surface potential from primarily positive and neutral (depicted as blue and white) to negative (red) (Figure 5D). This change may weaken the binding of protein partner(s) or nucleic acid ligand(s) necessary for transactivation.

### Discussion

We have developed an approach to rapid characterization of inherited missense variants in the BRCT domains of *BRCA1* that is able to retrospectively predict the outcome of a functional transactivation assay with greater than 94% accuracy. Our method makes no a priori assumptions about which predictive features might best inform a classification algorithm. Rather, we hypothesize that given a large sample of deleterious and neutral variants, we can quantitatively measure the most informative predictors and learn to distinguish between these two functional classes with a supervised learning algorithm. We discuss (a) how our approach compares with similar work [7,12,18,19,48–50], (b) how prediction of deleterious variants can identify putative binding sites, (c) how computational classifiers can save time and money required for biochemical assays of many candidate variants, and (d) the possibilities of generalizing the methods to large numbers of disease-associated genes.

Supervised learning algorithms have previously been applied to predicting the functional impact of missense



**Figure 3.** Computational Classifications of 54 Uncharacterized Variants Found in BIC

For each variant, the local protein structure environment is represented by secondary structure type and whether the amino acid residue is buried (*normalized solvent accessibility* < 0.2) or exposed (*normalized solvent accessibility* ≥ 0.2). For the 54 uncharacterized variants, labels (“1652 M->T”) are colored according to consensus prediction from Naïve Bayes, Support Vector Machine, and Random Forest. Predictions of each method are indicated by filled circles (blue, neutral; red, deleterious).

N, Naïve Bayes. R, Random Forest; S, Support Vector Machine.  
doi:10.1371/journal.pcbi.0030026.g003

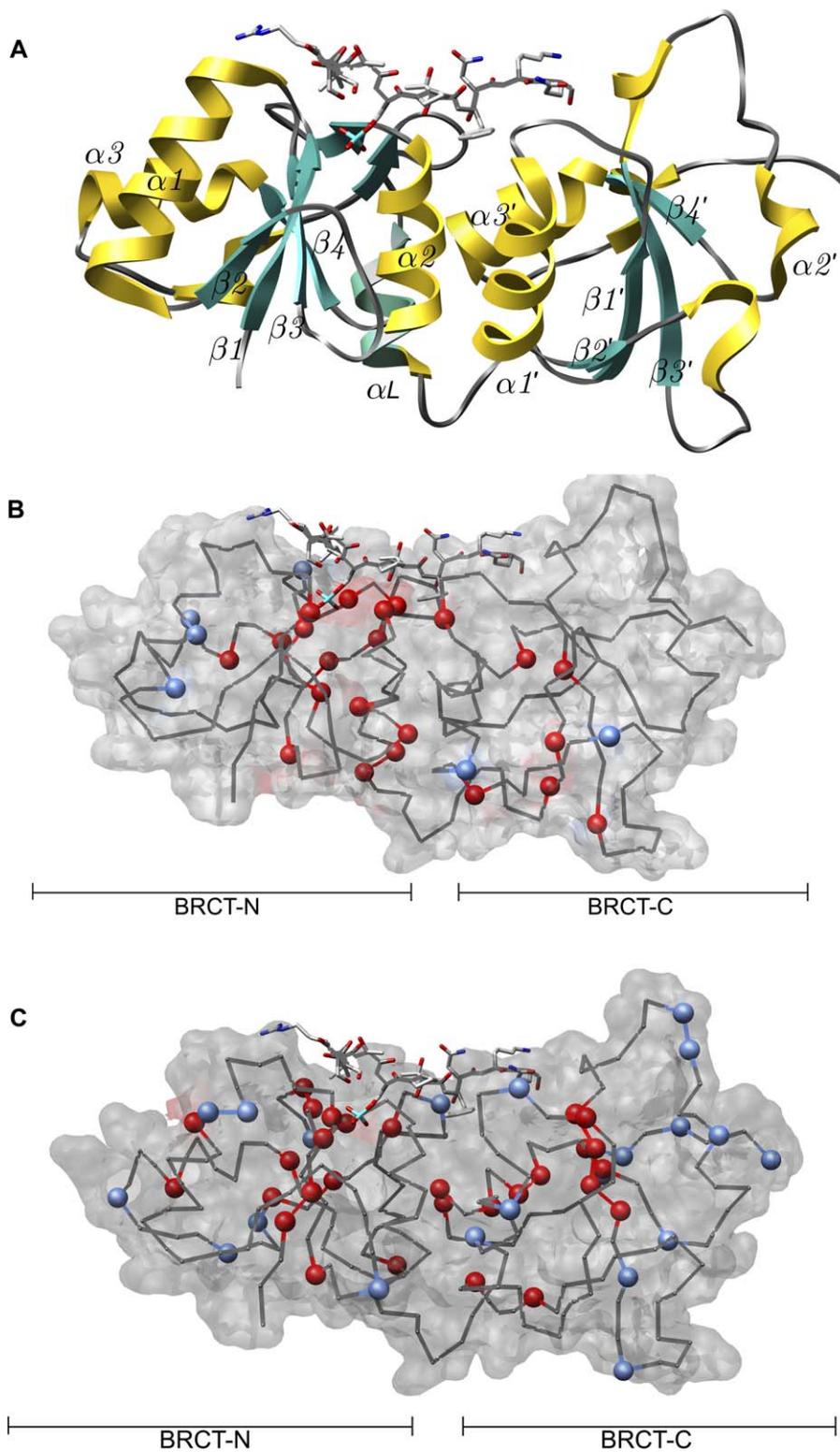
variants in TP53 with a four-body “potential” based on Delaunay tessellation [51], to engineered variants in *E. Coli* Lac Repressor, HIV protease, and T4 bacteriophage lysozyme [24,52,53], and to large sets of single nucleotide polymorphisms [54–56]. Much of this work has been limited by overfitting problems. Benchmarking of Support Vector Machines and Decision Trees in several studies has shown that high numbers of false positive and false negative classification errors (~0.30) are generated when the learners are applied to proteins other than those in their training sets [24,52]. Supervised learning using four-body potentials is further limited in application, because each missense variant is represented by a profile of *n* features (amino-acid residue potential scores), where *n* is the number of amino acid residues in the protein. Supervised learning algorithms require fixed-length feature vectors; thus, an algorithm trained on missense variants represented by *n* features can only classify missense variants that are represented by *n* features. For example, if the training set is composed of missense variants in Lac Repressor (327 amino acid residues), the algorithm cannot be used to classify mutants in Lysozyme (164 amino acid residues).

Here we identify a set of 16 predictive features that, in combination with Support Vector Machine, Random Forest, and Naïve Bayes supervised learning algorithms, avoids the overfitting problem when the training set is composed of TP53 variants and the validation set is composed of BRCA1 missense variants. Initially, we used 31 deleterious and eight neutral BRCA1 BRCT variants that had been functionally tested as our training set. However, this approach yielded poor classification performance in a cross-validation test, presumably because of small sample size (unpublished data). As an alternative, we selected our features and performed

supervised learning with a training set of 600+ artificially engineered TP53 missense variants. The ability of computational learning algorithms trained on TP53 variants to classify BRCA1 missense variants in agreement with the BRCA1 functional assay (94%+) suggests that mechanisms underlying structural and functional defects may be similar in TP53 and BRCA1.

In comparison, an approach based on sequence analysis and expected frequencies of structural features inferred from mutagenesis studies of *E. Coli* lac repressor and T4 lysozyme resulted in only 75% agreement with a BRCA1 BRCT trypsin sensitivity assay of 22 variants [12,50]. We find that the best supervised learners are in greater agreement with the BRCA1 transactivation assay than several sequence analysis methods and an empirically designed set of rules and thresholds [7].

The sequence analysis methods that incorporate physicochemical properties of amino acid residues as well as evolutionary conservation (Ancestral Sequence and Align-GVGD) are more accurate than SIFT, which only considers evolutionary conservation. A weakness of these methods is that, for purposes of classifying deleterious variants, there is no principled way to choose the optimal set of evolutionarily related sequences to align and analyze. In this work, we used sets of aligned sequences taken from published work (Ancestral Sequence) [9], the SIFT and Align GVGD web-servers [18,57], and a deep alignment (out to the sea urchin *Strongylocentrotus purpuratus*) generated by the creators of Align GVGD. Different sequence sets produce different classifications of the variants, and choice is biased by available genomes and decisions about appropriate thresholds of relatedness. The problem is illustrated with classifications of the BRCA1 BRCT missense variant V1665M. Align-GVGD Tnig, Align-GVGD Spur, and Ancestral Sequence incorrectly



**Figure 4.** Spatial Distribution of Predicted Deleterious and Neutral Missense Variants in the BRCA1 BRCT Domains

(A) Ribbon representation of the two domains with labeled helices ( $\alpha 1$ ,  $\alpha 2$ , etc.) and strands ( $\beta 1$ ,  $\beta 2$ , etc.). Recreation of Figure 1A [64].

(B) BRCA1 BRCT missense variants reported as neutral (blue) and deleterious (red) in the mammalian transactivation assay shown mapped onto the BRCA1 BRCT X-ray crystal structure (1t29).

(C) Consensus predictions of Random Forest, Naïve Bayes, and Support Vector Machine for 54 BRCA1 BRCT VUS in the Breast Information Core database (<http://research.nhgri.nih.gov/bic/BIC/>) mapped onto the same structure, with predicted neutral shown in blue and predicted deleterious in red. doi:10.1371/journal.pcbi.0030026.g004



classify V1665M as deleterious, because valine is completely conserved in their multiple sequence alignments. In contrast, SIFT constructs an alignment that includes two *Arabidopsis* proteins containing BRCT domains (UNIPROT Q9ZWC2, Q3E7F4) with a methionine aligned at this position, and thus correctly classifies the variant as neutral.

SIFT uses *Dirichlet mixture pseudocounts* to estimate allowed amino acid residues at each alignment position [29]. Pseudocount approaches compensate for incomplete sequence sampling in a multiple sequence alignment by adding counts for imaginary amino acid residues that are statistically likely to occur at each position. Our analysis indicates that several SIFT errors are the result of poorly estimated pseudocounts (Table S1). Importantly, SIFT is the only sequence analysis method that includes an automated alignment algorithm. The accuracy of Ancestral Sequence and Align-GVGD depends on manual sequence selection, so these methods cannot automatically be applied to other cancer susceptibility genes or to whole genome analysis.

Despite their limitations, sequence analysis methods have the advantage that they can be applied to any position in BRCA1, whereas our supervised learners require protein structure information and are thus limited to regions for which accurate protein structure or structural models are available. For example, the BRCA1 BRCT missense variant E1794D is misclassified by Support Vector Machine, Random Forest, Rule-based decision tree, and Decision Tree. Analysis of X-ray crystal protein structure quality with MolProbity and PROCHECK [58,59] indicates that in the 1t29 structure, the atomic coordinates at this position, may be incorrect. The corresponding backbone dihedral angles  $\Phi$  and  $\Psi$  of 143.8° and 112.9°, respectively, are statistical outliers. Using sequence analysis alone, with multiple sequence alignments in which *Tetraodon* BRCA1 contains an aspartic acid residue at this position, Align-GVGD, Ancestral Sequence, and SIFT correctly classify the variant as neutral.

The empirically designed rules (Rule-based decision tree) are more accurate than a Decision Tree of rules learned by a supervised algorithm but not as accurate as Support Vector Machine, Random Forest, Naïve Bayes, Ancestral Sequence, or Align-GVGD (Table 1 and Figure 1). It appears that it is difficult to correctly set thresholds and weigh the relative importance of empirically designed rules. For example, Rule-based decision tree classifies A1669S as deleterious, in disagreement with the functional assay (Figure 1 and Table S1). Rule-based decision tree uses a “mutation likelihood” rule that would classify A1669S as neutral because there is a serine in frog BRCA1 at this position. It also uses a rule that would classify A1669S as deleterious because the amino acid residue position is buried in the protein core and close to a buried charged residue. The order is such that the latter rule dominates the final decision. In contrast, the supervised learning approach makes no a priori decisions about thresholds or predictor ordering, but learns this information implicitly during the training process. Given an informative training set, such as the TP53 variants used in the present work, it is able to make highly accurate decisions about the BRCA1 variants.

The supervised learning algorithms described here make classification decisions based on nonlinear combinations of predictive features and fail to provide rationalizations that can be understood by humans. To address this issue, we apply a post-predictive step in which we analyze protein structure

models and alignments. For example, the wild-type arginine amino-acid residue at position 1753 of BRCA1 forms a salt bridge with the glutamate at position 1735. The arginine is found in the linker region that connects the BRCT-N and BRCT-C domains and the glutamate is found in BRCT-N (Figure 4). This charge-charge interaction may be important for the stability of the BRCT homodimer. In our structural model of the threonine variant, the salt bridge is broken, potentially destabilizing the homodimer. Both R1753 and E1735 are completely conserved in our alignment of 13 BRCA1 orthologs (Figure 5), suggesting possible selective pressure to preserve their pairwise interaction. Such rationalizations increase our confidence in the predictions and suggest ways to test them experimentally—for example, by site-directed mutagenesis. Fifty-four percent of the variants can be explained by combining structural and evolutionary analysis in post-prediction (Tables S1 and S2).

### Binding Site

Several studies of the BRCA1 BRCT domains have suggested that there may be surface patches that interact with protein partners [60–63]. In previous work, we predicted that a groove formed by both BRCT repeats (near nonfunctional variants L1657 and K1702) and the ridge that delimits the groove (near nonfunctional variant E1660) constitutes such a surface patch [7]. Our prediction was subsequently confirmed through X-ray crystallographic studies, which revealed a site where the phosphorylated peptides of BACH1 and CtIP have been found to bind [64,65]. Importantly, several missense variants were found to disrupt this interaction [64,65], suggesting that clustering of deleterious variants at solvent-exposed amino acid residue positions is indeed a useful indicator of binding site location. There is a large literature on the general topic of the relationship between deleterious mutations and binding sites [66–74].

Two surface variants found to be deleterious to BRCA1 transactivation activity in our companion paper (R1753T and T1685I) [14] lie on a highly conserved patch of amino acid residues, forming an exposed groove. The R1753T variant yields a changed electrostatic surface potential, which may be sufficient to disrupt the binding of BRCA1 to a protein partner or nucleic acid ligand important for transactivation. Following the logic that predicted the BACH1/ CtIP binding site, we suggest that this groove may be a previously uncharacterized binding site, whose disruption inactivates BRCA1 transactivation function. Accordingly, we are currently testing the binding of several candidate protein partners to the predicted binding site using site-directed mutagenesis and a yeast two-hybrid assay.

### Prioritizing Variants for Biochemical Testing

Biochemical assays can play an important role in identifying deleterious UCVs in cancer susceptibility genes [75], but the work is labor-intensive and time-consuming. We estimate that, on average, an assay for one BRCA1 UCV in a mammalian cell system costs US\$125–US\$150 and requires three weeks of personnel time, from ordering primers to final results. The time can be reduced by processing the variants in batches. To assay the 54 uncharacterized BRCT BRCA1 missense variants found in the BIC database (April 2006) would take approximately 18 months of personnel time. Accurate computational classification of UCVs can signifi-

cantly reduce the required time by prioritizing UCVs most likely to be deleterious. Importantly, while computational classification and functional assays can contribute to medical decision making, other factors such as family history, co-occurrence with known deleterious mutations, and studies of patient tumor tissue will continue to be important in a clinical setting.

### Generalizability of the Methods

We have applied supervised learners trained on the TP53 variant set to prediction of UCVs in BRCA2, with promising results (unpublished data). We are currently exploring whether this training set and our current set of features can be used to evaluate UCVs in other genes associated with familial cancer syndromes: MLH1, MSH2, MSH6 (hereditary nonpolyposis colon cancer), APC (familial adenomatous polyposis), MYH (MYH adenomatous polyposis), and P16 (melanoma). We have applied a modified version of this method to classify all human amino-acid changing SNPs found in the dbSNP database [76] as deleterious or neutral [56]. The SNPs were classified with a support vector machine trained on amino acid residue substitutions from more than 1,500 human proteins. Because X-ray crystal structures are not available for most human proteins [77], we built homology models with an automated modeling pipeline MODPIPE that relies on the MODELLER package for fold assignment, sequence-structure alignment, model building, and model assessment [78]. A small number of these predictions have been validated by biochemical and epidemiological studies found in the literature.

We are exploring the extent to which a decision rule learned with a training set of variants from one protein, such as TP53, can be generalized to variants from other proteins. One possibility is that most deleterious missense mutants do not affect specific binding interactions, but are instead slightly destabilizing [79]. If this is true, a training set of missense variants from a protein with similar stability to the protein of interest may be the best choice. Other possibilities include training on a protein sharing GO terms, or from the same fold family (all-alpha, all-beta, alpha-beta, etc.) as the protein of interest. We are working on generating large variant datasets from selected proteins to test these hypotheses.

In summary, we have systematically and comprehensively evaluated structure- and sequence-based computational prediction methods applied to variants in the BRCA1 BRCT domains and developed detailed structural explanations for the measured and predicted impact of 49 BRCA1 variants. When combined with 16 carefully selected predictive features, the best-supervised learning algorithms are in greater agreement with experimental results than has been reported previously.

The increased use of sequencing methods to genotype individuals at risk for inherited cancers and the observation that sequence variation is greater in ethnic minorities than in Caucasians highlight the need for improved methods of UCV risk assessment. Bioinformatics approaches including supervised learning algorithms, protein structure modeling, and evolutionary sequence analysis can contribute to an integrated approach to risk assessment by increasing coverage of classified UCVs more rapidly than is possible by functional assays. In the future, when clinicians counsel patients about their cancer risk, they will be able to take advantage of these bioinformatics prediction methods. Finally, successful generalization of these methods to a large

number of disease-associated genes will play an important role in reducing the growing number of loci, variants, and phenotypes that confound modern whole genome disease-association studies.

### Supporting Information

**Table S1.** Post-Prediction Analysis of BRCA1 BRCT Missense Variants Characterized by Transactivation Assays

Transcription assay in yeast (Y) or mammalian (M) cells. We used a 50% activity cutoff value. ○, neutral/low clinical relevance; ●, deleterious/high-risk variant; red, variant names are colored red if classification considering both yeast and mammalian assays is deleterious/high-risk, blue, variant names are colored blue if classification considering both yeast and mammalian assays are neutral/low clinical relevance. There is no available post-prediction analysis for S1715R, T1720A, and P1806A.

Found at doi:10.1371/journal.pcbi.0030026.st001 (135 KB DOC).

**Table S2.** Structural Explanations for the Impact of BRCA1 BRCT UCVs Found in BIC (Excluding Those in Table S1 and Those for Which Definitive Genetic Evidence Exists)

Consensus Classification is majority vote of Random Forest, Naïve Bayes, and Support Vector Machine classifiers. Red, deleterious; blue, neutral.

Found at doi:10.1371/journal.pcbi.0030026.st002 (101 KB DOC).

**Table S3.** Core Set of 13 Predictive Features Identified Using a Mutual Information Analysis

The mutual information between each feature and the functional/nonfunctional class of each TP53 missense mutant was computed as described previously [24].

Found at doi:10.1371/journal.pcbi.0030026.st003 (50 KB DOC).

**Table S4.** Predictive Features of Missense Mutants That Were Tried but Failed To Improve Support Vector Machine Performance on the TP53 Training Set

These were not included in the optimal 16 features selected for use with the supervised learning algorithms.

Found at doi:10.1371/journal.pcbi.0030026.st004 (40 KB DOC).

**Table 3.** Accession Numbers

Protein	Organism	NCBI Entrez ID
BRCA1	<i>H. sapiens</i>	AAA73985
BRCA2	<i>H. sapiens</i>	NP_000050
BRCA1	<i>P. troglodytes</i>	AAG43492
BRCA1	<i>G. gorilla</i>	AAT44835
BRCA1	<i>P. pygmaeus</i>	AAT44834
BRCA1	<i>Mac. mullata</i>	AAT44833
BRCA1	<i>C. familiaris</i>	AAC48663
BRCA1	<i>B. taurus</i>	AAL76094
BRCA1	<i>Mono. domestica</i>	AAX92675
BRCA1	<i>Mus musculus</i>	AAD00168
BRCA1	<i>R. norvegicus</i>	AAC36493
BRCA1	<i>G. gallus</i>	AAK83825
BRCA1	<i>X. laevis</i>	AAL13037
BRCA1	<i>T. nigroviridis</i>	AAR89523
BRCA1	<i>S. purpuratus</i>	EF152287
Lac Repressor	<i>E. coli</i>	AAM86784
Lysozyme	<i>bacteriophage T4</i>	NP_839873
TP53	<i>H. sapiens</i>	NP_000537
MDC1	<i>H. sapiens</i>	NP_055456
PTIP	<i>H. sapiens</i>	NP_031375
BARD1	<i>H. sapiens</i>	NP_000456
53BP1	<i>H. sapiens</i>	NP_005648
F21M11.4	<i>A. thaliana</i>	NP_005648
At1g04020	<i>A. thaliana</i>	NP_973758

doi:10.1371/journal.pcbi.0030026.t003

**Table S5.** Performance of Optimal 16 Features Evaluated with 10-Fold Cross-Validation Using the TP53 Missense Variant Training Set Performance was computed with all 16 features, then with each of the 16 features held out. MCC, Matthews correlation coefficient; TNR, true negative rate; TPR, true positive rate.

Found at doi:10.1371/journal.pcbi.0030026.st005 (63 KB DOC).

#### Accession Numbers

Accession numbers from the National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov>) are shown in Table 3.

#### References

- Gudmundsdottir K, Ashworth A (2006) The roles of BRCA1 and BRCA2 and associated proteins in the maintenance of genomic stability. *Oncogene* 25: 5864–5874.
- Starita LM, Parvin JD (2006) Substrates of the BRCA1-dependent ubiquitin ligase. *Cancer Biol Ther* 5: 137–141.
- Venkitaraman AR (2002) Cancer susceptibility and the functions of BRCA1 and BRCA2. *Cell* 108: 171–182.
- Nathanson KN, Wooster K, Weber BL (2001) Breast cancer genetics: What we know and what we need. *Nat Med* 7: 552–556.
- Szabo CI, Worley T, Monteiro AN (2004) Understanding germ-line mutations in BRCA1. *Cancer Biol Ther* 3: 515–520.
- Goldgar DE, Easton DF, Deffenbaugh AM, Monteiro AN, Tavtigian SV, et al. (2004) Integrated evaluation of DNA sequence variants of unknown clinical significance: Application to BRCA1 and BRCA2. *Am J Hum Genet* 75: 535–544.
- Mirkovic N, Marti-Renom MA, Weber BL, Sali A, Monteiro AN (2004) Structure-based assessment of missense mutations in human BRCA1: Implications for breast and ovarian cancer predisposition. *Cancer Res* 64: 3790–3797.
- Monteiro AN, Couch FJ (2006) Cancer risk assessment at the atomic level. *Cancer Res* 66: 1897–1899.
- Pavlicek A, Noskov VN, Kouprina N, Barrett JC, Jurka J, et al. (2004) Evolution of the tumor suppressor BRCA1 locus in primates: Implications for cancer predisposition. *Hum Mol Genet* 13: 2737–2751.
- Tavtigian SV, Samollow PB, de Silva D, Thomas A (2006) An analysis of unclassified missense substitutions in human BRCA1. *Fam Cancer* 5: 77–88.
- Fleming MA, Potter JD, Ramirez CJ, Ostrander GK, Ostrander EA (2003) Understanding missense mutations in the BRCA1 gene: An evolutionary approach. *Proc Natl Acad Sci U S A* 100: 1151–1156.
- Williams RS, Chasman DI, Hau DD, Hui B, Lau AY, et al. (2003) Detection of protein folding defects caused by BRCA1-BRCT truncation and missense mutations. *J Biol Chem* 278: 53007–53016.
- Kato S, Han SY, Liu W, Otsuka K, Shibata H, et al. (2003) Understanding the function–structure and function–mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. *Proc Natl Acad Sci U S A* 100: 8424–8429.
- Carvalho MA, Marsillac S, Karchin R, Manoukian S, Grist S, et al. (2007) Determination of cancer risk associated with germline BRCA1 missense variants by functional analysis. *Cancer Res*. In press.
- Domingos P, Pazzani M (1997) On the optimality of the simple Bayesian classifier under zero-one loss. *Mach Learn* 29: 103–130.
- Vapnik V (1995) The nature of statistical learning theory. New York: Springer-Verlag. 188 p.
- Breiman L (2001) Random forest. *Mach Learn* 45: 5–32.
- Tavtigian SV, Deffenbaugh AM, Yin L, Judkins T, Scholl T, et al. (2006) Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *J Med Genet* 43: 295–305.
- Ng PC, Henikoff S (2001) Predicting deleterious amino acid substitutions. *Genome Res* 11: 863–874.
- Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* 405: 442–451.
- Kabsch W, Sander C (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22: 2577–2637.
- Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234: 779–815.
- Karplus K, Karchin R, Barrett C, Tu S, Cline M, et al. (2001) What is the value added by human intervention in protein structure prediction? *Proteins (Supplement 5)*: 86–91.
- Karchin R, Kelly L, Sali A (2005) Improving functional annotation of non-synonymous SNPs with information theory. *Pac Symp Biocomput*: 397–408.
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. *Nucleic Acids Res* 28: 235–242.
- Shiozaki EN, Gu L, Yan N, Shi Y (2004) Structure of the BRCT repeats of

## Acknowledgments

**Author contributions.** RK, ANAM, MAC, and AS conceived and designed the experiments. RK and MAC performed the experiments. RK and ANAM analyzed the data. RK, ANAM, and SVT contributed reagents/materials/analysis tools. RK, ANAM, and AS wrote the paper.

**Funding.** The project has been supported by US National Institutes of Health grants F32 GM-072403–02, U01 GM-61390–04, R01 CA92309; the Sandler Family Supporting Foundation; an IBM SUR grant; and computer hardware gifts from Intel and Network Appliances.

**Competing interests.** The authors have declared that no competing interests exist.

- BRCA1 bound to a BACH1 phosphopeptide: Implications for signaling. *Mol Cell* 14: 405–412.
- Joo WS, Jeffrey PD, Cantor SB, Finnin MS, Livingston DM, et al. (2002) Structure of the 53BP1 BRCT region bound to p53 and its comparison to the Brcal BRCT structure. *Genes Dev* 16: 583–593.
  - Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, et al. (2006) The Universal Protein Resource (UniProt): An expanding universe of protein information. *Nucleic Acids Res* 34: D187–D191.
  - Sjolander K, Karplus K, Brown M, Hughey R, Krogh A, et al. (1996) Dirichlet mixtures: A method for improved detection of weak but significant protein sequence homology. *Comput Appl Biosci* 12: 327–345.
  - Krogh A, Brown M, Mian IS, Sjolander K, Haussler D (1994) Hidden Markov models in computational biology: Applications to protein modeling. *J Mol Biol* 235: 1501–1531.
  - Dimitriadou E, Hornik K, Liesch F, Meyer D, Weingessel A (2006) e1071: Misc functions of the Department of Statistics, Technical University of Vienna. R package version 15–13.
  - Therneau TA, Ripley B (2005) rpart: Recursive partitioning. R package version 31–23. Available: <http://www.r-project.org>. Accessed 22 January 2007.
  - Breiman LF, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Belmont (California): Wadsworth. 385 p.
  - Liaw A, Wiener M (2002) Classification and regression by randomForest. *R News* 2: 18–22.
  - Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research* 25: 3389–3402.
  - Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 31: 365–370.
  - Benner SA, Cohen MA, Gonnet GH (1993) Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *J Mol Biol* 229: 1065–1082.
  - Grantham R (1974) Amino acid difference formula to help explain protein evolution. *Science* 185: 862–864.
  - Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, et al. (2004) UCSF Chimera—A visualization system for exploratory research and analysis. *J Comput Chem* 25: 1605–1612.
  - Nicholls A, Honig B (1991) A rapid finite difference algorithm, utilizing successive over-relaxation to solve the Poisson-Boltzmann equation. *J Comput Chem* 12: 435–445.
  - Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, et al. (1984) A new force field for molecular mechanical simulations of nucleic acids and proteins. *J Am Chem Soc* 106: 765–784.
  - Word JM, Lovell SC, Richardson JS, Richardson DC (1999) Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation. *J Mol Biol* 285: 1735–1747.
  - Nicholls A, Sharp KA, Honig B (1991) Protein folding and association: Insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* 11: 281–296.
  - Phelan CM, Dapic V, Tice B, Favis R, Kwan E, et al. (2005) Classification of BRCA1 missense variants of unknown clinical significance. *J Med Genet* 42: 138–146.
  - Chapman MS, Verma IM (1996) Transcriptional activation by BRCA1 [letter; comment]. *Nature* 382: 678–679.
  - Horwitz AA, Sankaran S, Parvin JD (2006) Direct stimulation of transcription initiation by BRCA1 requires both its amino and carboxyl termini. *J Biol Chem* 281: 8317–8320.
  - Anderson SF, Schlegel BP, Nakajima T, Wolpin ES, Parvin JD (1998) BRCA1 protein is linked to the RNA polymerase II holoenzyme complex via RNA helicase A. *Nat Genet* 19: 254–256.
  - Chen GC, Guan LS, Yu JH, Li GC, Choi Kim HR, et al. (2001) Rb-associated protein 46 (RbAp46) inhibits transcriptional transactivation mediated by BRCA1. *Biochem Biophys Res Commun* 284: 507–514.
  - Sunyaev S, Ramensky V, Koch I, Lathe W III, Kondrashov AS, et al. (2001) Prediction of deleterious human alleles. *Hum Mol Genet* 10: 591–597.
  - Chasman D, Adams RM (2001) Predicting the functional consequences of

- non-synonymous single nucleotide polymorphisms: Structure-based assessment of amino acid variation. *J Mol Biol* 307: 683–706.
51. Mathe E, Olivier M, Kato S, Ishioka C, Vaisman I, et al. (2006) Predicting the transactivation activity of p53 missense mutants using a four-body potential score derived from Delaunay tessellations. *Hum Mutat* 27: 163–172.
  52. 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.
  53. Saunders C, Baker D (2002) Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J Mol Biol* 322: 891.
  54. Yue P, Moulton J (2006) Identification and analysis of deleterious human SNPs. *J Mol Biol* 356: 1263–1274.
  55. Bao L, Cui Y (2005) Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. *Bioinformatics* 21: 2185–2190.
  56. Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, et al. (2005) LS-SNP: Large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics* 21: 2814–2820.
  57. Ng PC, Henikoff S (2002) Accounting for human polymorphisms predicted to affect protein function. *Genome Res* 12: 436–446.
  58. Davis IW, Murray LW, Richardson JS, Richardson DC (2004) MOLPROBITY: Structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic Acids Res* 32: W615–W619.
  59. Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: A program to check the stereochemical quality of protein structures. *J Appl Cryst* 26: 283–291.
  60. Magnard C, Bachelier R, Vincent A, Jaquinod M, Kieffer S, et al. (2002) BRCA1 interacts with acetyl-CoA carboxylase through its tandem of BRCT domains. *Oncogene* 21: 6729–6739.
  61. Cantor SB, Bell DW, Ganesan S, Kass EM, Drapkin R, et al. (2001) BACH1, a novel helicase-like protein, interacts directly with BRCA1 and contributes to its DNA repair function. *Cell* 105: 149–160.
  62. Ye Q, Hu YF, Zhong H, Nye AC, Belmont AS, et al. (2001) BRCA1-induced large-scale chromatin unfolding and allele-specific effects of cancer-predisposing mutations. *J Cell Biol* 155: 911–921.
  63. Anderson SF, Schlegel BP, Nakajima T, Wolpin ES, Parvin JD (1998) BRCA1 protein is linked to the RNA polymerase II holoenzyme complex via RNA helicase A. *Nat Genet* 19: 254–256.
  64. Williams RS, Lee MS, Hau DD, Glover JN (2004) Structural basis of phosphopeptide recognition by the BRCT domain of BRCA1. *Nat Struct Mol Biol* 11: 519–525.
  65. Varma AK, Brown RS, Birrane G, Ladias JA (2005) Structural basis for cell cycle checkpoint control by the BRCA1-CtIP complex. *Biochemistry* 44: 10941–10946.
  66. Di Lullo GA, Sweeney SM, Korkko J, Ala-Kokko L, San Antonio JD (2002) Mapping the ligand-binding sites and disease-associated mutations on the most abundant protein in the human, type I collagen. *J Biol Chem* 277: 4223–4231.
  67. Becker KD, Gottshall KR, Hickey R, Perriard JC, Chien KR (1997) Point mutations in human beta cardiac myosin heavy chain have differential effects on sarcomeric structure and assembly: An ATP binding site change disrupts both thick and thin filaments, whereas hypertrophic cardiomyopathy mutations display normal assembly. *J Cell Biol* 137: 131–140.
  68. Werner M, Hermann-Le Denmat S, Treich I, Sentenac A, Thuriaux P (1992) Effect of mutations in a zinc-binding domain of yeast RNA polymerase C (III) on enzyme function and subunit association. *Mol Cell Biol* 12: 1087–1095.
  69. Baker RT, Smith SA, Marano R, McKee J, Board PG (1994) Protein expression using cotranslational fusion and cleavage of ubiquitin. Mutagenesis of the glutathione-binding site of human Pi class glutathione S-transferase. *J Biol Chem* 269: 25381–25386.
  70. Leebeek FW, Kariya K, Schwabe M, Fowlkes DM (1992) Identification of a receptor binding site in the carboxyl terminus of human interleukin-6. *J Biol Chem* 267: 14832–14838.
  71. Selent U, Ruter T, Kohler E, Liedtke M, Thielking V, et al. (1992) A site-directed mutagenesis study to identify amino acid residues involved in the catalytic function of the restriction endonuclease EcoRV. *Biochemistry* 31: 4808–4815.
  72. Raschle M, Dufner P, Marra G, Jiricny J (2002) Mutations within the hMLH1 and hPMS2 subunits of the human MutLalpha mismatch repair factor affect its ATPase activity, but not its ability to interact with hMutSalphalpha. *J Biol Chem* 277: 21810–21820.
  73. Knight JC, Udalova I, Hill AV, Greenwood BM, Peshu N, et al. (1999) A polymorphism that affects OCT-1 binding to the TNF promoter region is associated with severe malaria. *Nat Genet* 22: 145–150.
  74. Middleton SA, Johnson DL, Jin R, McMahon FJ, Collins A, et al. (1996) Identification of a critical ligand binding determinant of the human erythropoietin receptor. Evidence for common ligand binding motifs in the cytokine receptor family. *J Biol Chem* 271: 14045–14054.
  75. Carvalho MA, Couch FJ, Monteiro AN (2006) Functional assays for BRCA1 and BRCA2. *Int J Biochem Cell Biol* 39: 298–310.
  76. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. (2001) dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res* 29: 308–311.
  77. Xie L, Bourne PE (2005) Functional coverage of the human genome by existing structures, structural genomics targets, and homology models. *PLoS Comput Biol* 1: e31.
  78. Sanchez R, Sali A (1998) Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc Natl Acad Sci U S A* 95: 13597–13602.
  79. Yue P, Li Z, Moulton J (2005) Loss of protein structure stability as a major causative factor in monogenic disease. *J Mol Biol* 353: 459–473.
  80. Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH (1985) Hydrophobicity of amino acid residues in globular proteins. *Science* 229: 834–838.
  81. Otzen DE, Rheinacker M, Fersht AR (1995) Structural factors contributing to the hydrophobic effect: The partly exposed hydrophobic minicore in chymotrypsin inhibitor 2. *Biochemistry* 34: 13051–13058.
  82. Zamyatin AA (1972) Protein volume in solution. *Prog Biophys Mol Biol* 24: 107–123.
  83. Engelman DM, Steitz TA, Goldman A (1986) Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu Rev Biophys Chem* 15: 321–353.
  84. Kullback S (1959) Information theory and statistics. New York: Wiley. 395 p.
  85. Lennard-Jones JE (1931) Cohesion. *Proc Phys Soc* 43: 461–482.
  86. Sing T, Sander O, Beerwinkler N, Lengauer T (2004) ROCr: An R package for visualizing the performance of scoring classifiers.