Disease Risk of Missense Mutations Using Structural Inference from Predicted Function

Jeremy A. Horst[1,2,3], Kai Wang[4], Orapin V. Horst[1,5,6], Michael L. Cunningham[1,7,8] and Ram Samudrala[1,3,*]

[1]Department of Oral Biology, [2]Department of Oral Medicine, School of Dentistry, University of Washington, USA; [3]Department of Microbiology School of Medicine, University of Washington, USA; [4]Center for Applied Genomics, Children's Hospital of Philadelphia; [5]Department of Dental Public Health Sciences, [6]Department of Endodontics, School of Dentistry, University of Washington, USA; [7]Department of Pediatrics, School of Medicine, University of Washington, USA; [8]Craniofacial Clinic, Seattle Children's Hospital 1959 NE Pacific St #357132, Seattle, WA 98195, USA

Abstract: Advancements in sequencing techniques place personalized genomic medicine upon the horizon, bringing along the responsibility of clinicians to understand the likelihood for a mutation to cause disease, and of scientists to separate etiology from nonpathologic variability. Pathogenicity is discernable from patterns of interactions between a missense mutation, the surrounding protein structure, and intermolecular interactions. Physicochemical stability calculations are not accessible without structures, as is the case for the vast majority of human proteins, so diagnostic accuracy remains in infancy. To model the effects of missense mutations on functional stability without structure, we combine novel protein sequence analysis algorithms to discern spatial distributions of sequence, evolutionary, and physicochemical conservation, through a new approach to optimize component selection. Novel components include a combinatory substitution matrix and two heuristic algorithms that detect positions which confer structural support to interaction interfaces. The method reaches 0.91 AUC in ten-fold cross-validation to predict alteration of function for 6,392 *in vitro* mutations. For clinical utility we trained the method on 7,022 disease associated missense mutations within the Online Mendelian inheritance in man amongst a larger randomized set. In a blinded prospective test to delineate mutations unique to 186 patients with craniosynostosis from those in the 95 highly variant Coriell controls and 1000 age matched controls, we achieved roughly 1/3 sensitivity and perfect specificity. The component algorithms retained during machine learning constitute novel protein sequence analysis techniques to describe environments supporting neutrality or pathology of mutations. This approach to pathogenetics enables new insight into the mechanistic relationship of missense mutations to disease phenotypes in our patients.

Keywords: Computational biology, protein stability, machine learning, missense mutation, nonsynonymous SNP, sequence analysis.

INTRODUCTION

The majority of single amino acid changes that increase risk of clinical manifestations (e.g. developmental malformations, neoplasms, or infections) are caused by alteration of structural stability rather than the otherwise attractive notion that disease causing missense mutations directly disrupt ligand interactions sites (e.g. metabolite, protein, or autogenous substrate) [1-3]. Wang and Moult proposed a taxonomy for disease causing missense mutations and deleterious effects on protein function generally as: directly effecting protein stability, ligand binding, catalysis, allosteric regulation, or post-translational modification [1]. Mechanisms of disrupted stability are observed as the most abundant phenotypic missense mutations (~80%) because more residues contribute to structure than specific interaction interfaces [1,4,5].

*Address correspondence to this author at the Department of Microbiology School of Medicine, University of Washington, USA,
Tel: ??????????????????????; Fax: ????????????????????????;
E-mail: ram@compbio.washington.edu

Differences in stability are further categorized into effects on hydrogen bonds, hydrophobic exclusion, salt bridges, burial of charged residues, overpacking, induction of internal cavities, electrostatic repulsion, burial of polar residues, metal ion binding, disulfide bonding, backbone strain, and the effects of all these on multimer stability [1]. Many of these mutations disrupt the stability of a protein folding transition state. They induce a thermodynamic shift to decrease the proportion of properly folded gene products, which describes a fundamental constraint of protein evolution [6]. However, a spectrum of effects can also occur for a single mutation, from maintaining function at decreased rates, to making no difference (neutral), to gain of functions such as enhancing catalysis or signal propagation [7]. For example the functional conformation sometimes still exists in a smaller proportion of gene products and the physiologic actions of the protein remains possible.

Alternatively an effect on stability can maintain proper fold topology but produce a slightly different interface, affecting interaction specificity and sensitivity as seen in pathogen drug resistance [8]. This effect is estimated to comprise 5% of disease causal mutations in humans [1]. The

1389-2037/10 $55.00+.00

balance of flexibility and stability across interaction sites controls progression of reactions by both kinetic and thermodynamic mechanisms. The support residues in spatial shells surrounding the active site are far more abundant than the interface residues themselves (estimated as $n^3/2$), and therefore are more likely to become mutated.

The variance in amino acid types within the positions that support interface structures may enable interaction specificity for a given protein family. The interface positions which directly contribute electrostatic interactions (electron sharing) commonly maintain interaction with a family of ligands or facilitate a type of chemical reaction. They are highly conserved and therefore finding them within a protein sequence appears tractable [9-10]. Like hydrophobic ligand binding residues (guiding residues), the structural support positions differ from the electron sharing positions in that they are difficult to detect by automated residue conservation algorithms. Therefore automated methods to identify phenotypic mutations from the larger set of nondisruptive missense mutations should directly consider the conservation of spatial neighbors rather than just the position itself.

Parameters of Protein Structure

The absence of tertiary and quaternary structure data for the vast majority of human proteins limits the ability to assess the environment of each residue. Sequences are available for the most relevant proteins, and so protein structure prediction offers the possibility of ascertaining these structural parameters.

Templates available for comparative knowledge based modeling (template-based modeling) are already nearly sufficient to answer this problem for proteins like those which have been crystallized [11-12]. Yet the available templates required for consistently accurate modeling [13], might only present detectable similarity to 40% of the human or terrestrial proteome [14-15]. So it is plausible that only 40% of the proteome can be modeled based on structures determined with existing structure determination methods. Many proteins seem to not be accessible to NMR or crystallography. For example a 4 in 5 failure rate for target proteins by the protein structure initiatives [16], indicates contemporary experimental methods for assessing protein structure may not be able to interrogate the remaining ~60% of protein families [14]. In other words, the subset of proteins in a genome for which we can accurately predict protein structure may be equivalent to the group of proteins that can be characterized by contemporary experimental methods. So, while structure prediction does offer long term utility to many other problems [17-18], the relevance of template-based protein structure prediction to the problem of phenotypic missense mutations may be limited to the time until the assessable human proteins are experimentally characterized. Currently only ~40% of human mutations can currently be mapped to corresponding or modeled structures [19-20].

Therefore we posit parameters of protein structure predictable from sequence as a substitute for 3D structure (tertiary and quaternary, experimental and predicted) in the investigation of phenoytypic missense mutations generally and interaction interface support residues specifically. We use existing sequence analytic knowledge based algorithms to predict secondary structure, solvent exposure, burial, disorder, domain restraints, and nonlocal contact prediction at multiple shell radii to substitute tertiary structure information. All sequence analytic methods applied here are implemented on the results from a single default PSI-BLAST run [21]. The structure features are predicted using the suite of software kindly provided to the community by Jianlin Cheng, selecting *ab initio* methods where available [22-27]. These methods performed as the best or near best in each related category of the 8th Community wide experiment on the critical assessment of methods for protein structure prediction (CASP8) [28]. In this work we demonstrate how these predicted structural parameters can derive functional importance, thereby finessing dependence on high quality structural data for the problem of separating insignificant missense mutations from disease risk inducing mutations.

Relation to other Methods for Predicting Phenotypic Missense Mutations

Amino Acid Substitution Matrices

It is unclear what data set first led to the observation of a differentiable profile of amino acid types in disruptive missense mutations, but the work relating the genetic code to amino acid replacement in missense suppression seems to have been the groundwork [29-30]. The probabilities of disruption for mutation of each amino acid type is now discernable from large datasets. For example the distribution of disruptive and silent mutations in ASEdb describes an order for the likelihood of disruption for mutating each amino acid: WYRIDNPKHQEFVMSTLC (single letter amino acid code), for which the first three residues stand out with respect to the others [31]. Observations of trends for certain wild to mutant amino acid type pairs to be disruptive or permitted led to substitution matrices specifically trained for effects on functional stability [32-33]. However, amino acid substitution matrices have always been designed to estimate the significance of different amino acid types at the same position [30,34]. So all matrices can be relevant to this problem.

The significantly distinct substitution matrices have been conveniently summarized in the AAindex [35]. Additionally, substitution matrices created within PSI-BLAST iterations hold unique information as they are customized to the query protein [19,36]. A substitution matrix specific to types of predicted structure was first applied to this problem in SNAP, for predicted transmembrane domains [36]. Here we elaborate on the concept of exploiting separable substitution patterns by allowing inclusion of multiple matrices specific to structural contexts, and posit a way to achieve balance between minimizing overtraining and maximizing power by combining multiple noncontextual matrices.

Structural Analysis

A thorough discussion of the events leading to our understanding of destabilizing mutations is far beyond the scope of this paper, but brief summary informs a framework to understand the logic built into the heuristic algorithms designed to address this problem. Analysis began with modeling the free energy change by adjusting side chain rotamers in known X-ray crystal diffraction structures [37-38]. Estimations of free

energy change calculated through knowledge based functions does reach clinically relevant accuracy for the case of assessing physiologic ligand interactions such as drug resistance, when enough data is available to specifically model the particular system [8].

Structural analysis delineated the importance of the hydrophobic effect to this problem [39], the corollary trends for specific amino acid types [40], the predicted degree of solvation, and types of nonlocal contacts [41]. Structural analysis of disruptive mutations highlighted measurable patterns, including distance from the active site and changes in torsion angles, hydrogen bonding, solvent exposed hydrophobicity, and stability at progressive stages of minimization simulations [42]. Simply considering the quantity of each amino acid type within a structural shell becomes useful with machine learning [43,44], which we abstract here to the sequence inferred structural environment.

Descriptors of the structural environment can be more useful than direct measurements of stability, particularly in the case of completely modeled protein structures. In one algorithmic combination we applied to the CASP8 function prediction experiment, we employed the *in silico* mutation analysis part of our meta-functional signature protocol (MFS) [10]. This analysis was designed from the observation that unbound interaction interface residues are more unstable than other conserved residues, which tend to be structurally important [45]. Instability facilitates the thermodynamics of binding by compensating entropy loss with enthalpy [46]. However, the minimization steps used to improve the models remove the native instability of the interface residues! As a result, the MFS sequence limited analysis was more accurate than adding the destabilizing filter. Instead using a simple spatial cluster heuristic that sought other high MFS score residues in the vicinity enriched predictions by 24%. As a result our group performed as the second and third best for metal and ligand binding respectively [47]. Our observations that heuristics can outperform structural measurements and the successful designs of MUpro [44] and SNAP [36] suggest that many of the structure analytic techniques may be extended through sequence analysis, allaying the need to build full structural models.

Sequence Analysis

As mentioned above, conservation detects residues that facilitate structural integrity or direct interactions. The earliest sequence based public tools attempted to jointly model structural context and conservation. Conservation was combined with position-specific scoring matrices (with Dirichlet priors) by SIFT [48]. Conservation was also combined with measurements of the structural environment within completely modeled structures by SNPs3D [1].

Expanding the philosophical basis of analysis beyond conservation within the contemporary snapshot of evolution improves accuracy. Modeling physicochemical conservation by deviation of six physical parameters from ortholog alignment position alone is complimentary to standard sequence analysis [49]. Modeling positive selection through DNA sequence [50] or phylogenetic branch deviation [51] are two examples of more directly evolutionary approaches. We offer a third here, similar to branch deviation but instead analyzing only the proportion of state changes to branch points, or steps (SSR) [10].

Annotation mapping [36,44] and text mining [52] are nontrivial tasks which directly make use of the exponentially growing scientific information base. These data mining tools have the capacity to add subtly relevant evidence, not accessible through traditional protein sequence informatics. The difficulty of accurate automated mapping of mutation sites to protein structures is familiar to anyone who has read a protein structure file, but seems to have been solved by two groups [19-20].

Subdivision of proteins in superfamily or functional type was found to be useful, perhaps by improving the context of algorithmic learning [53]. Portability of methods learning from bench experimental data was shown to be enhanced greatly by including the experimental temperature and pH [44], which could be fruitful for application of these methods to specific *in vivo* tissues and organisms.

The profile of amino acids contained within a sequential residue window (sliding window) becomes highly useful with machine learning [38]. This amino acid context was found to maintain significance considering only three positions before and after the residue [44]. Fragment analysis was further extended by modeling transitional frequencies within matching three residue fragments from a sequence database [36].

Application of machine learning techniques approximately followed the order in which they were developed. Simple decision trees combined structure and sequence measurements [54]. Neural networks and support vector machines facilitated combination of many parameters [55-56]. Hidden Markov models enabled modeling of complex chain relationships, creating perhaps the best approach to model gain of function mutations [57].

Sequence Inferred Structural Analysis

All other improvements of sequences analysis have come from inference of structure. Early systematic analyses for both structure and sequence features revealed the utility of the following predictable features: solvent accessibility and disorder (B-factor) [58]; and specific involvement hydrogen bonds, salt bridges, metal ion binding, disulfide bonding, and multimer interface [1]. Considering the predicted solvation in concert with other component algorithms specifically was found to improve predictions [36]. Secondary structure [55], changes in predicted secondary structure upon mutation, and transmembrane locality have also been found to effect patterns in fragility [36].

Nonsystematic approaches to template based modeling were useful as soon as protein structure prediction became accurate [59]. The quality of the predicted model obviously effects the ability to draw meaningful inference. Sequence similarity between query protein and template structure helps to estimate applicability to this problem, estimated as between 40% and 60% [2-3]. Model quality assessment now far surpasses the accuracy of sequence similarity [60], so these measurements need to be redone, albeit with consideration that much can be derived from low resolution models with accurate topology.

Here we continue the exploration of sequence inferred structural inferences of potentially destabilizing mutations.

METHODS

For a given protein sequence, the residues and their degree of functional importance can be thought of as a signature representing the function of the protein. We previously developed a combination of knowledge- and biophysics-based function prediction techniques to elucidate the relationships between the structural and functional roles of individual protein residues [10]. Such a meta-functional signature (MFS) may be used to study proteins of known function in greater detail and to aid experimental characterization of proteins of unknown function [10].

Here we extend the MFS philosophy to evaluate the contribution of residues to functional, structural, or interactive stability by amino acid type, amino acid substitution, functional importance scores based on multiple sequence alignments, structural features, and the scores of residues predicted to be nearby in 3D space (nonlocal contacts). We use backwards stepwise multiple regression to remove score types that do not add weight to the predictions with statistical significance, i.e. include all scores then remove one at a time with cycles of training by logistic regression until all add significant improvement on the training set. We employ supervised learning only by forcing the maintenance of all amino acid types, as a base from which to improve. We then train a support vector machine (SVM) on the resultant set of score types, as this training approach creates hyperplanes between combinations of score types to refine accuracy, instead of the single best fit polynomial line of logistic regression. In training, an SVM will attempt to draw curved lines between scores from different algorithms and derivable features thereof, and even enclosing circles for clusters of datapoints. The resulting model yields a continuous spectrum of thresholds with corresponding specificity and sensitivity which the user can balance as appropriate to the particular application.

Note: we use the marker [Novel] to denote new software or algorithmic changes presented here.

Protein Sequence Analysis Using Multiple Sequence Alignments

We use the position specific iterative basic local alignment search tool (PSI-BLAST) [21] to find similar protein sequences from the non-redundant database [61]. More sensitive and specific methods have emerged, such as the context sensitive iterative BLAST (CSI-BLAST) [62], HMM-HMM predictive comparison method (HHpred) [63], and PSI-BLAST intermediate sequence search (PSI-BLAST-ISS) [64], which are reviewed by us previously [65]. While PSI-BLAST results have inherent limitations of sensitivity compared to these newer tools, we do overcome the specificity problem in part by applying the multiple sequence comparison by log-expectation algorithm (MUSCLE) [66] to the PSI-BLAST output, and keep the top 250 nearest neighbors in the resulting multiple sequence alignment (MSA).

For each protein we use a single pass of PSI-BLAST and MUSCLE calculations (each with multiple internal iterations: 3 for PSI-BLAST, and the default selection of MUSCLE) to drive the entire prediction pipeline. Each of the following algorithms calculates functional importance given this single MSA.

HMMRE

We train a hidden Markov model (HMM) from the MSA using the Hmmer package [67], then compare emission frequency estimates from the model with the amino acid background frequency in nature, given by karlin.c of the BLAST program package [21], to produce the HMM relative entropy score for each amino acid position [10,68]. [Novel] Here we make a significant change by constraining the Markov chain architecture to the form of protein sequence, rather than using the chain apparent from conservation measured across the entire MSA.

SSR

We model the evolutionary context of each position by creating a maximum parsimony phylogenetic tree for the surrounding sequence of each position using the PHYLIP platform [69]. Each protein in the MSA is considered as a leaf in the tree, and the root represents the theoretical ancestral sequence. We quantify the evolutionary divergence of the position by taking the ratio of different amino acid states appearing at the particular position, to the total number of step changes in the modeled evolution between the input and ancestral protein within the phylogenetic tree, termed the state to step ratio (SSR) [10].

MAPP

The multivariate analysis of protein polymorphisms algorithm (MAPP) uses an MSA of protein sequence orthologs (the matching protein in another species) to estimate a mean for each of six physicochemical values for each position (MSA column) [49]. For each physicochemical value, deviation from the mean is calculated for all twenty amino acids, and a single composite value is generated by a center of mass calculation on a principal component transformation, wherein each physicochemical property is taken as a coordinate axis. Then the Euclidean distance of each amino acid from this center of mass composite value is taken to estimate the effect of a mutation at that position [49].

Str from Seq

We employ sequence based predictions of structural features including secondary structure, level of solvent exposure, disorder, disulfide bonds, domain breaks, and nonlocal contacts. All of these structural features are predicted using the suite of software kindly provided to the community by the Jianlin Cheng group [22-27].

CloseSS - [Novel]

Protein residues come together in 3D space to form functional sites. We have created a method to consider the probability of concordant function for a residue one through five positions away, related to the secondary structure predicted for the evaluated position. For example, side chains in the n+2 position of an extended beta strand will tend to be nearby the considered position (n), as will the side chains of n+3 and n+4 for an alpha helix.

Shells - [Novel]

We hypothesize that many modelable features of the structural environment affect the stability of a position. We use nonlocal sequence contact prediction to select residues more than 5 positions away, as those that can contribute to the 3D environment of each position. Algorithms from the Jianlin Cheng group are separately trained for nonlocal contact prediction at distances of 0-8Å or 0-12Å. We use these methods to resolve virtual concentric contact shells at 0-8Å, 0-12Å, and 8-12Å. Within each shell we measure the count of amino acid types, the mean and distribution of HMMRE conservation scores, the probability of the nonlocal contact prediction, and the simple number of contacts. We thereby dissect contact shells by progressively indirect effects on stability. The philosophy underlying this method arises from detailed analysis on the patterns of interresidue distances on stability, wherein patterns of <8Å contacts vary considerably from those further out [70].

Fxn from Str

This term refers to the combination of the novel methods CloseSS and Shells with predictions of structural features (Str from Seq).

Sequence Independent Algorithms

Matrices - [Novel]

We train a simple look up table by considering the 94 matrices curated into the AAindex database [35] in our regression protocol. All matrices kept after the reverse stepwise logistic regression steps are combined into a single matrix using the regression coefficients for weights. This approach decreases overtraining common to matrices derived directly from the training set data. Predictions for mutation types disproportionately abundant or absent in the training set are rectified by the analyses and data sets used to build into each component matrix. We also separately consider the similarity matrix produced within the last PSI-BLAST iteration as described in SNPs3D [19], and the position independent matrix as described in SNAP [36].

Machine Learning Techniques

Logistic Regression

The simplest approach to combining float point predictions is regression. Linear or logistic regression often do not display significant differences in performance in protein informatics, as for this problem (data not shown).

Reverse Stepwise Logistic Regression

To improve the signal to noise ratio we evoke the reverse stepwise approach to removing component algorithm predictions. Forward regression adds one component at a time until the significance of an added component is lost. When considering many components the search is either over inclusive or non-exhaustive, depending upon the significance threshold (many components combinations are never considered). By reversal of the search direction, overly similar predictions are prioritized for removal. Stepwise single component removal allows the most significant of the similar algorithms to be retained. Any inclusive combination of algorithms can be considered, while maintaining highly stringent significance requirements. The component set selected by the forward approach varies depending upon the order of inclusion. The tractable search of the reverse direction is deterministic, which is valuable for testing hypotheses in informatics experiments. To avoid overtraining we do not consider reentry. We use the removal statistic (P-value > 0.001), which describes the probability of observing the component data when unrelated to the neutral or deleterious condition.

Support Vector Machine

The two most popular machine learning techniques in protein informatics are neural networks and support vector machines (SVM) [71-72]. Neural networks create exhaustive regressions between all possible combinations of the component set, where the number of components to be combined is defined by the operator determined number of hidden layer nodes. SVM vectors are not limited to the same amount of components. SVM solves maximal separation between combinations of components through multiple hyperplane vectors. The hyperplanes can be solved using Lagrange multipliers, thus enabling complex inferred relationships, for example separating case from control with circles instead of lines. We selected the radial basis kernel function (RBF), because it can match the behavior of linear or sigmoid kernel functions with parameter training [73-74], and the polynomial kernel function has more hyperplane parameters and thus is more likely to be over-trained. We train the RBF parameters by internal ten fold cross validation grid searches (possible values for cost 1-10 and gamma 0-5), within each of the ten training samples (cross within cross). In other words the cost and gamma values were trained for each cross validation set, and the training involved subdivision of the subset. For the final score we use the decision probability estimate [75]. Finally, we checked for specific effects of the software by using multiple SVM packages: libsvm [76], svmvia [77], and svmlight [78].

Experiment Data Sets

In Vitro Set

We use the *in vitro* deleterious point mutation standard benchmarking set assembled by the creators of SIFT [48]. Although the assay values from the original papers were binned into four categories, we use a binary approach as a more rigorous test of whether our algorithms can discern any measurable effect on protein function. The set is comprised by the ability for 336 point mutations in HIV-1 protease to maintain Gag-Pol precursor cleavage [79], 2,015 mutations in Bacteriophage T4 lysozyme to maintain plaque formation when exposed to *Salmonella typhimurium* [80], and 4,044 in *Escherichia coli* LacI repressor to maintain repression of the Lac operon in the absence of allolactose (or IPTG) and release in its presence [81]. Other high quality *in vitro* data sets not used here include Protherm [82] and ASEdb [83].

Clinical Set

The online Mendelian inheritance in man of the National Library of Medicine at the National Institutes of Health (OMIM) [84] registers 7,022 missense mutations as etiologic or a contributory risk factor to disease, while 35,434 more

human nsSNPs have been described in the literature and curated by the Protein mutation database [85] but are not clearly delineated for effect in patients. As the clinicians and scientists who generated the instances within the PMD mutation set focus on identifying difference rather than similarity, we assume the latter list to be skewed significantly towards disease causing mutations relative to a hypothetical set wherein clinical and bench data would be collected for a random distribution of proteins and patients. With limited sequence characterization resources, scientists interrogate genes likely to cause problems in their patients, cells, or proteins. Meanwhile, those mutations for which the data do not demonstrate linkage, could disrupt function in ways not measured by the particular experiment. So reliable true positives exist, but true negatives do not. [Novel] Thus we created a negative set necessary to train our knowledge based function, reproducing the distribution that would be produced by random single SNPs at random positions, matching the abundance of nsSNPs for each protein in the PMD (e.g. if 5 are demonstrated in protein X, we create 5 in X).

Craniosynostosis Data Sets

The only true test of a method is prospective verification. Our collaborators sequenced 27 genes in a set of 186 patients with single suture craniosynostosis, and the 95 highly variant Coriell controls. Seventy eight missense mutations were found, of which 49 were novel and unique to the patients. They genotyped these novel 49 missense mutations in 1000 age and geographic matched controls, resulting in 15 unique novel mutations which are taken here as potentially causal mutations [86]. Patients were collected based on primary diagnosis, through a network of four craniofacial clinics in Children's Hospitals within the United States. The genes were chosen based on previously demonstrated causal mutations for craniosynostosis or syndromes including this phenotype, and functionally related genes [86]. For example similar transcription factors, or if a mutation in a protein ligand was known to cause the phenotype, they added the receptor for the ligand, other proteins in the family of the ligand, and inhibitors of the ligand. Although further work remains to delineate relation and causality, the fifteen cases are taken as causal for our purposes here.

Prospective Clinical Test

We prospectively tested a group of algorithms on the set of 78 missense mutations observed in patients. We were blinded to assignment of the mutations to case or control, including whether a mutation was seen in multiple patients, or whether multiple mutations were seen in an individual patient. No attempts were made by us to check for presence of the mutations in public databases. We predicted disease causality as at least three scores above predetermined thresholds for any of MFS, HMMRE, SSR, SIFT, PMUT, MAPP, or PAM250. Thresholds were taken from the related publications for SIFT, PMUT, and MAPP. Those for MFS, HMMRE, and SSR were taken from previous observations of functional importance in other proteins. Positive scores were used for PAM250.

Retrospective Clinical Test

The prospective test design skews for noncausal mutations, as more mutations known to cause craniosynostosis or syndromes with craniosynostosis were present in the set than other previously observed mutations. As well, presentation of the set predated completion of the HUSCY methods presented here. Thus we considered a second set of all 105 mutations found in patients with craniosynostosis or age matched controls seen in the Craniofacial clinic at Seattle Children's over roughly fifteen years [86]. The thirty cases were all characterized elsewhere to be causative mutations. We removed the mutations present in the OMIM database to create a blinded validation scenario in which these mutations are hypothetically not known, and retrained the HUSCY algorithm for this retrospective test.

Evaluation Methods

Ten Fold Cross Validation

A knowledge based (informatic) algorithm assessment protocol wherein the training set is used for testing. The training set is divided randomly into ten equivalent subsets, each of ten versions of the algorithm is trained on the remaining 90% subset and assessed for accuracy based on predictions for the 10% subset.

ROC

The receiver operator characteristic (ROC) displays the balance of specificity and sensitivity across the range of possible score thresholds for deciding what predictions are positive versus negative [87]. This plot is valuable in enabling critical assessment of weaknesses of a method, demonstrating where method accuracies are separable, and in informing selection of a cutoff threshold appropriate to the particular application.

AUC

Accuracy across the range of score thresholds is summarized by measuring the area under the ROC curve (AUC), for which 50% is equivalent to random expectation and 100% is perfect prediction. To estimate AUC from the data, we employ the rectangular method (draw nonoverlapping rectangles down from the ROC data points), which tends to underestimate; we prefer this over the trapezoidal approach which can overestimate in some instances.

Two State Accuracy

A single value that describes total accurate predictions is useful to describe total predictive ability for a particular threshold or machine learning output. Two state accuracy is calculated as the fraction of correct predictions (positive or negative cases).

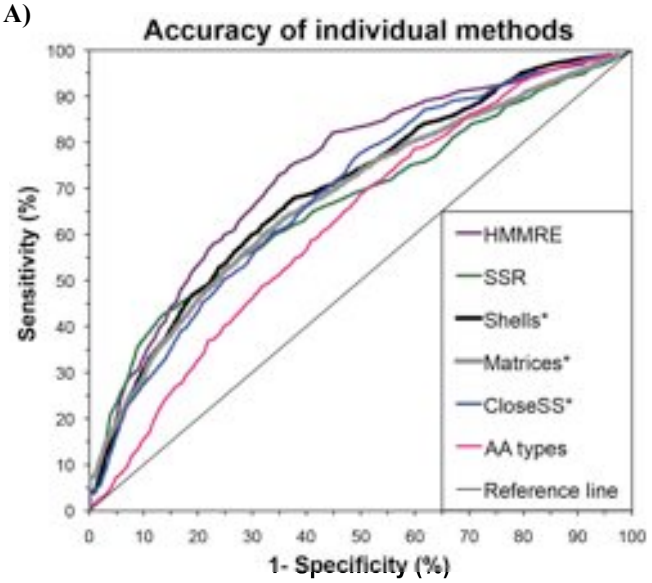
RESULTS AND DISCUSSION

Performance of Novel Algorithms to Predict Functional Disruption by Artificial Missense Mutations in the Standard *In Vitro* Mutation Test Set

The receiver operator characteristic in Fig. (**1**) demonstrates performance of the five algorithms separately. Each algorithm performs 1.4 - 2 times better than only considering amino acid type (AUC=61.9). HMMRE is most accurate (AUC=73.2) except in high specificity cases, for which SSR displays higher sensitivity. Correlations between the predic-

tions for each pair of these methods are sufficiently low to motivate combination see Fig. (**1b**). Thus we have collected and created a novel set of tools useful for protein sequence analysis.

A)



B)

Inter method correlations

	Shells	Matrices	CloseSS	HMMRE	SSR	AA type
Shells	---	0.19	0.34	0.34	0.24	0.24
Matrices	0.19	---	0.21	0.26	0.10	0.31
CloseSS	0.34	0.21	---	0.45	0.24	0.18
HMMRE	0.34	0.26	0.45	---	0.50	0.20
SSR	0.24	0.10	0.24	0.50	---	0.12
AA type	0.24	0.31	0.18	0.20	0.12	---

Fig. (1). New methods for prediction of mutational disruption. Methods assessing conservation (HMMRE, SSR), sequence derived structural patterns (Shells, CloseSS), and a combinatory amino acid substitution matrix (Matrices) are novely applied to the problem of predicting functional disruption by artificial missense mutations in the standard *in vitro* mutation test set assembled by the creators of SIFT [48]. This set is comprised by *in vitro* assay results for 336 mutations in HIV protease [79], 2015 in Bacteriophage T4 lysozyme [80], and 4044 in the *E coli* Lac repressor [81]. Parameters are trained for Shells (AUC=69.1), Matrices (AUC=67.4), CloseSS (AUC=68.3), and AA type (AUC=61.9). **(A)** The receiver operator characteristic of the five algorithms in ten fold cross validation (* indicates novel algorithms). Each algorithm performs better than random (Reference line) in all cases, each between 1.4-2 times more accurate than only considering amino acid type. HMMRE is most accurate (AUC=73.2) except in high specificity cases, for which SSR (AUC=66.5) performs better. **(B)** Low correlation between predictions of the different algorithms indicates additive predictive ability can be achieved by combination see Fig. (**2**).

Additive Prediction Value from Combining Novel Algorithms

Sequential regression combination demonstrates separable improvements for sequence derived parameters of structure and function Fig. (**2**).

Addition of structure parameter predictions to MFS (AUC=76.9) adds some predictive ability (AUC=78.2). Including the heuristic algorithms derived solely from these structural features (Shells and CloseSS) does add 5-10% sensitivity below 85% specificity (AUC=80.7, see Fig. **2**). Combination of predicted structural features with predicted function from predicted structural features, reaches accuracy comparable to MFS trained on this set, yet the structure from function algorithms do not directly include the conservation of the residue nor the amino acid type (Fxn from Str, AUC=76.3). The MFS StrFxn combination represents the total structure and functional importance of the residue in maintaining protein function (AUC=80.7).

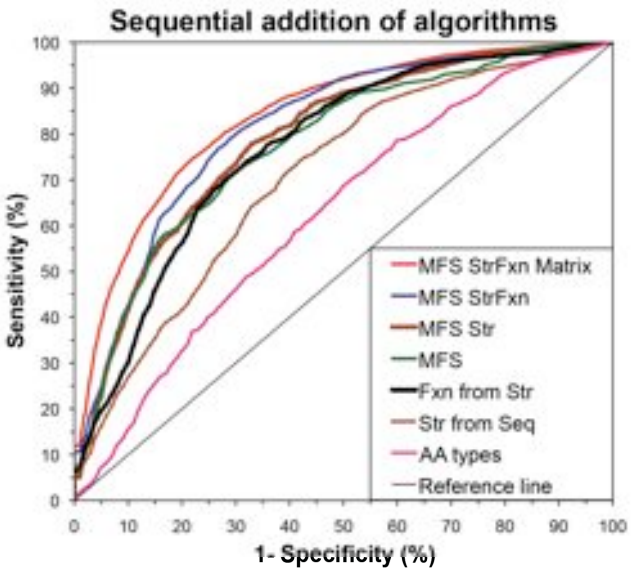


Fig. (2). Additive prediction value of combining novel algorithms. The philosophical derivation of predictive algorithms demonstrates separable improvements for combining sequence derived parameters of structure and function. Predicted structural features (Str from Seq, AUC=70.1) include disorder, secondary structure, solvation, contribution to disulfide bonds, and domain break points. Adding in predicted function from predicted structure (Fxn from Str, AUC=76.3) includes the Shells and CloseSS methods. Regression combination of HMMRE and SSR conservation methods and amino acid type is synoymous to our approach to predict residues with direct functional contribution measured by contacts with any interacting molecule [10], but here instead we consider functional contribution as positions for which mutations will disrupt protein function (MFS, AUC=76.9). Combining MFS with Str from Seq adds improvement (MFS Str, AUC=78.2). More sensitivity is added when adding the Fxn from Str algorithms (MFS StrFxn, AUC=80.7) which use only data already present in MFS Str, suggesting that the model of the structural environment by this sequence based algorithm is significant. Finally including the substitution matrices into the regression increases predictive ability (MFS StrFxn Matrix, AUC=83.2).

The Matrices combination models the importance of the particular wild type to mutation amino acid type. Adding this feature to MFS StrFxn increases sensitivity dramatically for specificities above 85% (AUC=83.2).

Amino Acid Substitution Scoring Matrix

It is easy to train a substitution matrix to a particular data set, but such a matrix tends to not be robust to situations outside of the data set. By combining the matrices of the AAin-

dex database [35] which do not necessitate extraneous structural features (e.g. solvation and secondary structure) we build upon the diversity of the set and the power of the analyses used to build each of them see Fig. (**1**) Supplemental Fig. (**1**).

The mutations that result in aspartic acid or tryptophan uniformly disrupt function, yet there are only a few of each in the standard *in vitro* set see Supplemental Fig. (**2**). As such these mutations do not hold the strongest scores in our matrix. None of the mutation types with uniform disruptive or neutral effects result in the strongest or weakest mutation scores, which is suggestive of not overtraining. The approach implicitly favors those mutation types with favorable power calculations. Thus Matrices derives scores for mutations seldom observed in the training set from the analyses used to build the other matrices, using weights trained on abundant mutation types.

Evidence of robustness is found in superior performance of application to the clinical data set as compared to other contemporarily popular and historically important matrices see Supplemental Fig. (**3**).

Improved Prediction Using Less Information During Machine Learning

Employing reverse stepwise logistic regression as a sample preparation technique to decrease information greatly improves support vector machine training see Fig. (**3**). Reverse stepwise logistic regression removed noisy data types but did not significantly alter prediction outcomes (AUC=83.25), as compared to simple logistic regression (AUC=83.22). Application of the exact same support vector machine training protocol see Fig. (**3**) and even the same gamma and cost function values (data not shown), results in substantially improved predictive ability by support vector machine predictions (AUC=90.6 versus 83.8). While obviously indicating that noise gets in the way of signal, the difference also suggests valuable algorithmic combinations not intentionally designed by us. For example the SVM models connect the predicted degree of solvation and nonlocal algorithmic components [41], and proline as the mutant amino acid to alpha helix predicted for wild type [36]. Neural networks will be employed in future work to further disentangle these algorithmic improvements.

Rendering the support vector machine regularization path shows gradual slopes for changes in the kernel parameters cost (penalty) and gamma (exponent coefficient) to produce stable accuracy in the regions from which the values were derived see Supplemental Fig. (**4**). The ten fold cross validation sets resolved to cost values 4±1 and gamma 1.5±0.5. As expected, steep slopes are found for cost below 1.0. No significant differences were found from training with different support vector machine programs (SVMlight, libSVM, SVMvia; data not shown).

The novel learning approach applied to the clinical data s*et al*so reaches sensitivity to specificity combinations not attained by logistic regression, which appropriately favor the 4.5x abundant neutral instances see Fig. (**6a**). The reverse stepwise logistic regression into support vector machine approach is generally novel to bioinformatics, creating perhaps the first example of improved performance of internal cross validation by avoiding overtraining.

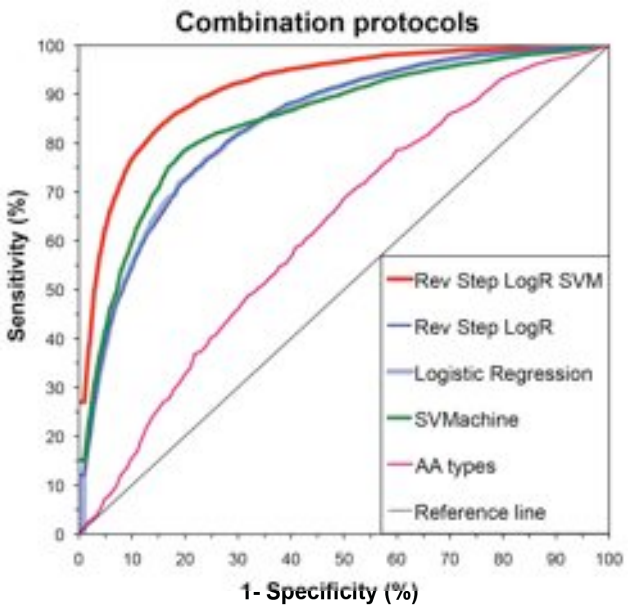
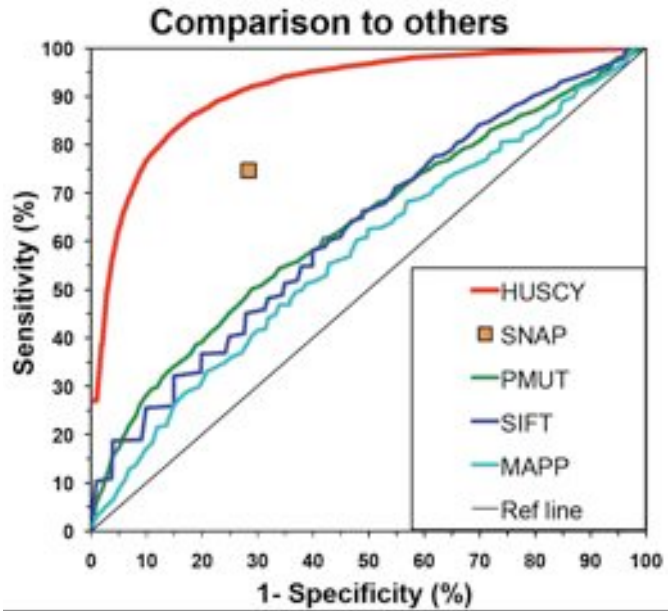


Fig. (3). Improved prediction using less information; a novel approach to machine learning. Machine learning techniques address the challenge of combining predictive scores from unique individual algorithms into a unified prediction. Previous approaches to data type selection in protein informatics assume to combine all available data or follow an expert's intuition. We demonstrate that employing a sample preparation technique to decrease information greatly improves the predictions of a more complex machine learning method. For the preparation technique we employ reverse stepwise logistic regression (Rev Step LogR), which removes data types but does not significantly alter prediction outcomes. The blue markers demonstrate extremely similar accuracy profiles of logistic regression, before (AUC=83.22) and after (AUC=83.25) filtering insignificantly contributing information types with Rev Step LogR. The green line depicts accuracy of support vector machine (SVM) training without Rev Step LogR filtration steps (AUC=83.8). The red line shows the exact same SVM method applied after filtration (AUC=90.6), demonstrating far better specificity and sensitivity than reached when including all data types. The Rev Step LogR SVM (referred to as HUSCY in later figures) depiction highlights that the approach is generally novel to bioinformatics, creating perhaps the first example of improved performance of internal cross validation by avoiding overtraining.

Novel Predictors of Structural and Functional Importance

Improvements in protein sequence analysis can be informed by the set of algorithms and parameters thereof which were consistently retained during reverse stepwise logistic regression training across ten fold cross validation in both the standard *in vitro* mutation set and the clinical mutation set (see Table **1**). The conservation methods include HMMRE, SSR, and MAPP, which suggests meaningful differences between the respective philosophical bases of contemporary sequence, evolutionary, and physicochemical conservation. The profile of specific amino acid types predicted to occur within nonlocal contact shells describe neutral and

A)



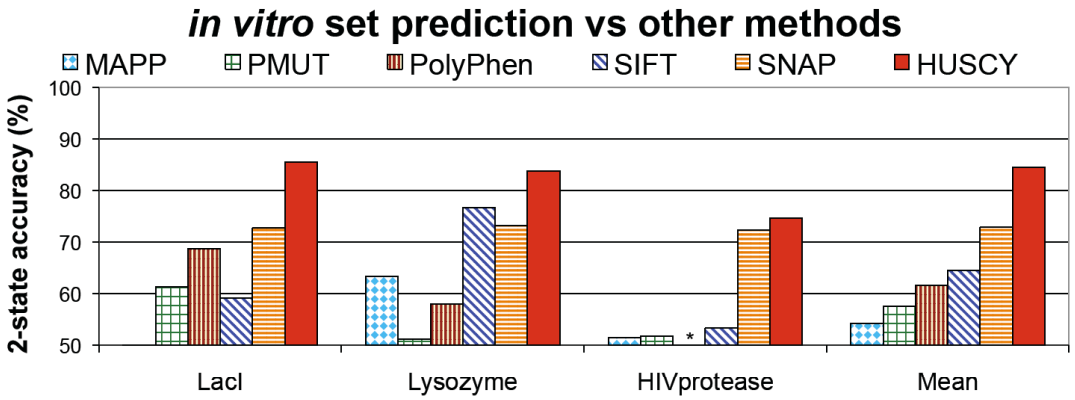


Fig. (4). Comparison to other methods for missense mutation phenotype prediction**.** Comparison of performance on the standard *in vitro* dataset for HUSCY (Rev Step LogR SVM in Fig. **3**) to approaches previously published in the field: SNAP [36], SIFT [48], PolyPhen, [88], PMUT [55], and MAPP [49]. PMUT was designed to predict human mutations, not the microbial systems assessed here. Other methods including SIFT, SNAP, PMUT, and PolyPhen were not trained on this specific data set, and thus would not be anticipated to perform as well. Nonetheless the set is used to standardize comparison for the methods. Performance for PolyPhen taken from the SNAP paper. **(A)** ROC accuracy profiles. **(B)** Two state accuracy separated by the three protein reporter systems comprising the standard set. HUSCY and SNAP methods perform stably across the three proteins. These data demonstrate a contribution to the field of characterizing mechanisms of protein function on the stringent test of picking out single mutations that produce any experimentally measurable change in the assayed functions.

deleterious environments. For example prediction of nonlocal prolines, alanines, or serines to be within 8Å of a mutation increases the probability of altering function. Meanwhile the presence of histidines within any measured contact shell is protective for maintenance of protein function (see Table **1**). The hypothesized patterns derived from patterns of side chain three dimensional proximity appeared exactly in the retained positions: two and four away within beta strands, three and four away in alpha helices, and three away in random coil.

The variance (standard deviation) of HMMRE conservation across residues predicted to be within the 8Å nonlocal contact shell was a positive predictor, as was the probability of those residues being within the nonlocal contact shell, and solvation or the total number of residues predicted to be in the shell. Meanwhile higher averages of HMMRE conservation for these residues were predictive of neutral mutation effects (see Table **1**). These observations together imply that for positions not picked up by the substitution matrix or the conservation scores of the residue itself, those mutations directly within the flexible functional site tend to be tolerated, while it is more generally deleterious to mutate residues occurring between functional sites and the rest of the protein. Thus our model appears to have caught the subtle effects of interface support residues.

Comparison to other Methods for Missense Mutation Phenotype Prediction

From here forth we will use the term HUSCY to refer to the model resulting from the reverse stepwise logistic regres-

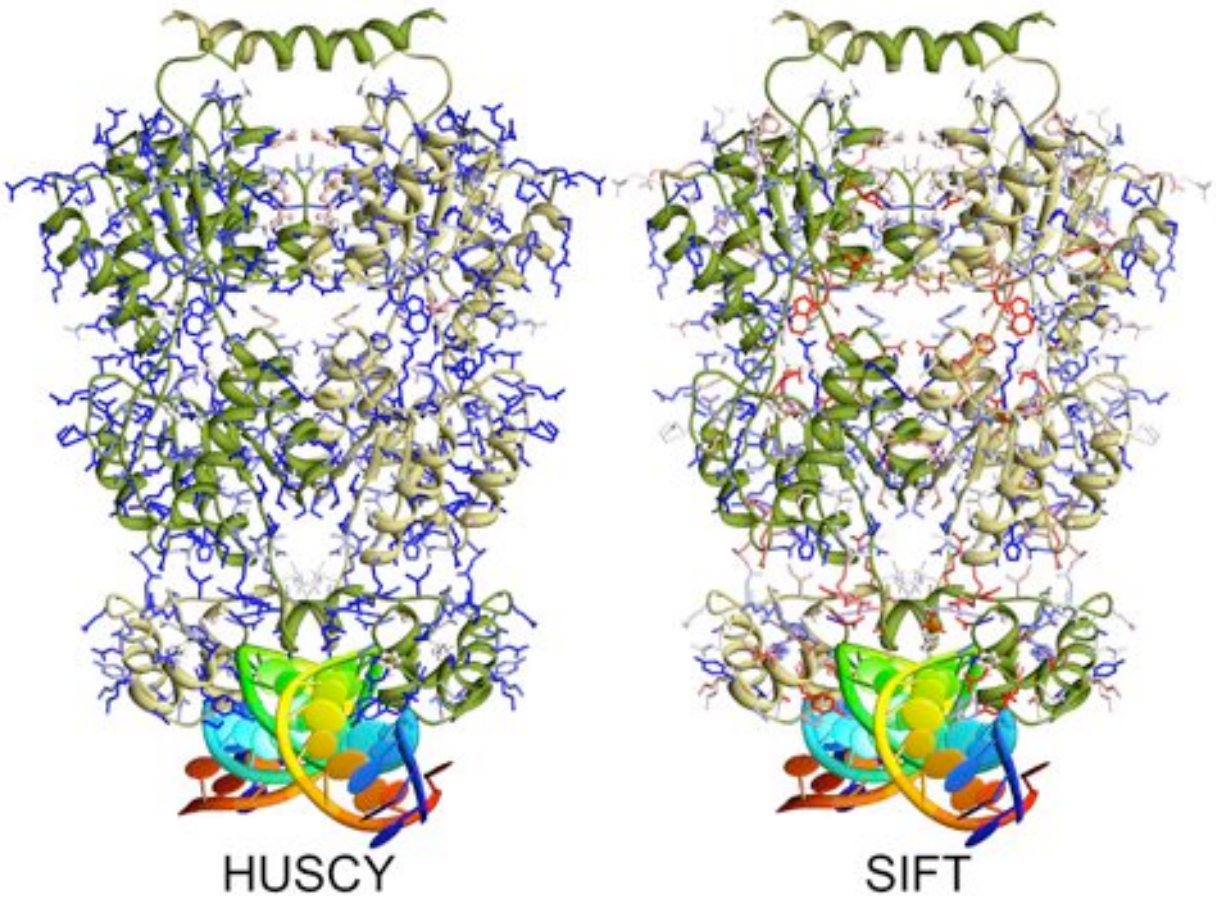


Fig. (5). Accuracy profile for prediction of deleterious effects by mutations in Lac Repressor**.** The two state prediction accuracy of HUSCY (left), and SIFT (right) [48] for all 12 or 13 mutations at each position (of 4044 in *E coli* Lac repressor) [81], mapped onto the homodimer structure of LacRepressor bound to the operator DNA (PDBid 1lbg). Side chains built by SCWRL4 [96] are shown for all 328 residues for which mutations were made, colored as heat map from blue for perfect selection to red for no correct selections. Main chains colored to differentiate homodimer chains. DNA shown as simplified ellipsoids in 5'-3' rainbow map. Two state accuracy includes correct prediction of either deleterious effects or no effects. Residues with <50% accuracy by HUSCY are shown as ball and stick in both renderings. The residues for which HUSCY displays poor performance are clustered at the protein homodimer interface and the allolactose binding site. Future improvements are directed by this analysis to include terms for interaction interface prediction. Clearly we already achieve our goal of accurate prediction for the interface support residues, bringing forward the field of sequence based prediction of destabilizing mutations.

sion into the support vector machine. The HUSCY method achieves far greater accuracy (AUC=90.6) than previous efforts to train on the *in vitro* dataset, e.g. MAPP (AUC=57.2; see Fig. **4**) [49].

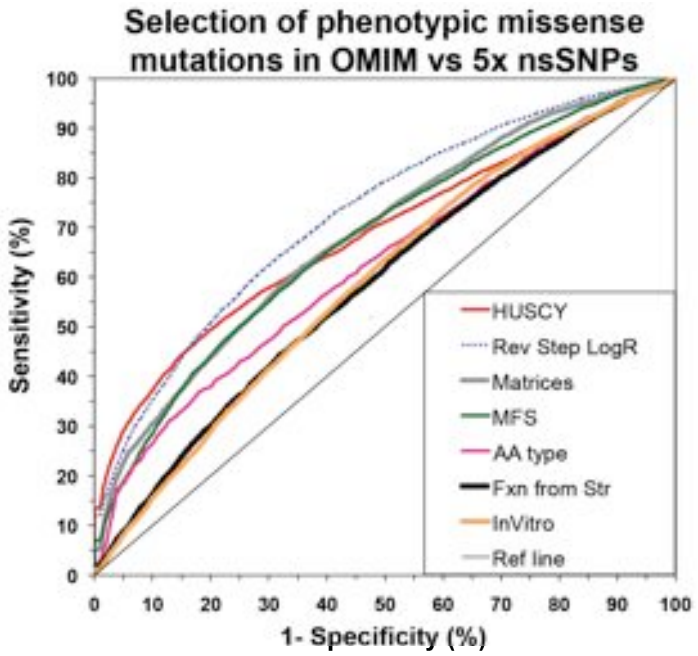
This set has been used in previous work multiple times to standardize comparison. Thus we also compare performance to other methods designed to predict disruption of missense mutations: SIFT [48], SNAP [36], PolyPhen [88], and PMUT (designed for human not *in vitro* systems; AUC= 62.7; see Fig. **4**) [55]. Conclusions based on comparison to this latter set should be drawn cautiously. The HUSCY method holds as the best performing upon dissection by the three proteins which comprise the standard set see Fig. (**4b**). SNAP and the HUSCY methods maintain significant accuracy across each of the three proteins.

These data demonstrate a contribution to the field of characterizing mechanisms of protein function by reproducing bench point mutations with high accuracy, on the stringent test of picking out single mutations that produce any experimentally measurable change in the assayed functions.

Three Dimensional Profile of Prediction Accuracy for Mutations in LacRepressor

We conceived a novel graphic design to display trends in performance across a protein to inform future improvement and comparison of the prediction algorithms see Fig. (**5**). Patterns in performance are communicated by the distribution of amino acid side chains colored according to the two state prediction accuracy. The small multiple (parallel figure construction) draws attention to the difference in two state accuracy between HUSCY and SIFT for intermolecular interaction residues and the support residues thereof see Fig. (**5**) albeit taking attention away from the excellent performance for peripheral mutations by both methods. Comparison to SNAP draws attention to the buried residues and interaction support residues which are more accurate for HUSCY see Supplemental Fig. (**5**) while both accurately predict the DNA interface. Similar side chain depiction for the 19 resi-

A)



B)

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	---		366	251		19							231			106	143	159		
C		---			218	397									755	295			444	535
D	362		---	84		392	420					130						225		213
E	231		78	---		340			380					89				178		
F		218			---			65		68						356		90		154
G	23	395	388	340		---									395	268		180	445	
H			419				---			261		126	360	206	522					402
I					65			---	369	1	95	380			276	381	351	110		
K				372				367	---		402	254		269	239		168			
L					70		262	1		---	7		585	287	398	344		69	108	
M								98	405	8	---				421		347	233		
N			133				127	381	263			---				119	185			156
P	223						361			588			---	401	361	369	340			
Q				92			205		277	292			403	---	768					
R		754				392	520	275	227	395	421		359	756	---	195	248		998	
S	104	295			358	269		382		344		115	376		202	---	24		249	245
T	146							355	172		347	180	342		252	20	---			
V	149		225	177	91	181		104		70	229							---		
W		444				444				108					999	249			---	
Y		535	213		155		401					158				244				---

Fig. (6). Prediction of disease related nonsynonymous SNPs in OMIM. **(A)** Receiver operating characteristic for selection of the 7,022 nonsynonymous SNPs recognized by the Online Mendelian inheritance in man (OMIM) [84] as contributory to human disease versus a negative control set of 31,698 randomly generated nonsynonymous SNPs we created to match the distribution of occurrence to all those observed in patients (PMD human *in vivo* subset) [85]. Predictive ability is gained from training the combination on this data set in ten fold cross validation. It might be surprising from the figure that HUSCY reaches an two state accuracy of 85% (98.5% specificity, 17.5% sensitivity; AUC=67.7), but there are 4.5 times more neutral instances than deleterious cases. This prediction value of 70% above random for clinical data has not been achieved previously. Rev Step LogR results in consistent selection of parameters across the ten derivations, which suggests stability of the algorithm (AUC=71.9). The HUSCY method trained on the standard *in vitro* set (InVitro, AUC=58.7) does not perform as well as simply considering the amino acid type (AUC=61.9), which highlights the difference in these problems (discussed in Results). Further disclosures for this data set include: passive nonsynonymous SNPs in humans are not yet known and so are modeled here, therefore many of the instances taken as negative would actually effect function as positives; many of the positive instances have not been thoroughly evaluated, e.g. in multiple prospectively studied populations. **(B)** We trained a specific amino acid substitution scoring matrix to select disease related nonsynonymous SNPs (gray circles in (A), AUC=67.8) as a combination of those in the AAindex database [35] which do not require other features such as secondary structure or solvation. The matrix demonstrates marginally higher accuracy than a sophisticated conservation measure trained for this purpose (MFS; green line in (A), AUC=67.0). Higher values are predictive of disease relation. Coloring is presented as a heat map, with red representing stronger predictions of disease and green representing minimal chance of causing disease. Only the mutations possible from a single nucleotide change are shown (i.e. nonsynonymous SNPs). The matrix values converge to two significant figures across the ten cross validation training sets. This matrix can be applied instantaneously as a simple look up table for clinicians not familiar with protein informatics.

Table 1. Novel Predictors Consistently Retained in Reverse Stepwise Logistic Regression

Positive Predictors	
1. Conservation methods	HMMRE, SSR, MAPP
2. CloseSS conservation, α-helix position	3, 4
3. CloseSS conservation, β-extended position	2, 4
4. CloseSS conservation, random coil position	3
5. Quantity of residues within the 0-8Å shell	ALA, PRO, SER
6.Quantity of residues within the 0-12Å shell	CYS, GLY, TRP
7. Quantity of residues within the 8-12Å shell	CYS, GLU, GLY
8. Quantity of neighbors within 0-12Å shell	
9. Standard deviation of the HMMRE conservation within 0-12Å shell	
10. Standard dev of product of HMMRE conservation and probability of being in 0-8Å shell	
11. Degree of solvation	
Negative Predictors	
1. Quantity of residues within the 0-8Å shell	CYS, GLU, HIS, ASN
2. Quantity of residues within the 0-12Å shell	HIS, ARG
3. Quantity of residues within the 8-12Å shell	HIS, MET
4. Mean HMMRE conservation within 0-8Å shell	
5. Quantity of disulfide bonds within 0-8Å shell	

dues on which HUSCY performs worse than random (<50%) shows better performance for a few by each of the other two methods, suggesting that this information can be captured.

The HUSCY depiction viewed alone highlights the cluster of inaccurate predictions at the distal homodimer interface and the allolactose binding site, and a minimum per residue accuracy of 25% (n=1; see Fig. **5**). The graphic design clearly demonstrates that we reached our goal to predict mutation effects for interface support residues. Simultaneously the clustering of non blue residues informs future work to interface residues perhaps by incorporating interface prediction as a specific pretrained parameter.

Prediction of Disease Related Missense Mutations in OMIM.

We retrained the entire HUSCY approach on the problem of selecting clinically phenotypic missense mutations. Although this clinical data set is far from ideal as described in Methods, our algorithms and machine learning approach demonstrate sequentially additive progress in prediction accuracy similar to that for the standard *in vitro* set see Fig. (**6**). It should be noted that the ROC plot skews the data by representing neutral and deleterious instances as equivalent in amount; rather the former are 4.5 times more abundant. Thus it might be surprising from Fig. (**6**) that the HUSCY method reaches 85% two state accuracy on the clinical set (98.5% specificity, 17.5% sensitivity), as for the *in vitro* set (87.3% specificity, 80.2% sensitivity). This prediction value of 70% above random for clinical data has not been achieved previously. Meanwhile the complete profile of predictions is worse with SVM training (AUC=67.7) than without (AUC=71.9). The amino acid substitution matrix trained from these data may be useful for clinicians see Supplemental Fig. (**6**).

The low sensitivity rate for disease prediction limits clinical relevance for this tool. As well, the HUSCY method trained on the standard *in vitro* set does not perform well on the clinical data set see Fig. (**6**, InVitro line; AUC=58.7). Nonetheless, there is value to the HUSCY predictions trained on the clinical set. The clinician and scientist can titrate the cutoff for each application see Supplemental Fig. (**7**). For example considering mutations with scores above 0.82 result in predominantly disease instances, a useful subset for designing expensive bench experiments (cutoff used for craniosynostosis cases below). Meanwhile the larger set of mutations with scores above 0.33 removes many instances vastly dominated by neutral mutations, which would robustly inform selection of SNPs for follow up screening.

Predicted Function from Predicted Structure is Limited for the Clinical Set

The majority of the OMIM and PMD proteins do not have available structures or templates with which to model them. Meanwhile most of the mutations that are in proteins for which structures or templates are available, similarly occur in structurally undefined regions of the protein [52].

Minimal significant predictions were made for core structural features (nonlocal contacts, secondary structure, and solvation) for many of the same proteins and positions. Thus many of the clinical set mutations may be located in proteins or regions that simply do not fold into globular domains in contemporary structure determination conditions. This pattern may be concordant with and descriptive of mutations that allow life, i.e. mutations in canonical globular domains very often result in severe abrogation of function, failure to thrive, and therefore are not observed clinically. This effect would make the modeled neutral cases carry more phenotypic instances than expected at random from well studied *in vitro* systems. A structure from sequence approach may not bring these algorithms to clinically relevant sensitivity. Such sensitivity may await a data set with more tersely resolved phenotypic cases, more abundant neutral instances, and advancements in the understanding of the unknown human protein structures, which may be as many as ~40% [15].

Nonsynonymous SNP Amino Acid Substitution Scoring Matrix

We trained a specific amino acid substitution scoring matrix to select disease related nonsynonymous SNPs as a combination of those in the AAindex database [35] which do not require input other features such as secondary structure or solvation. The matrix (AUC=67.8) demonstrates marginally higher accuracy than a sophisticated conservation measure trained for this purpose (MFS, AUC=67.0, see Fig. (**6b**). The matrix values converge to two significant figures for most mutation types in each of ten cross validation training sets. The matrix may be useful to guide a general understanding of the prevalence and therefore probability of each mutation causing phenotypic changes in patients. For example a clinician can apply these scores instantaneously as a simple look up table for mutation observed in patients.

Perfect Specificity but Weak Sensitivity for Clinical Cases of Craniosynostosis

We tested the approach before completion in a prospective test of 78 novel missense mutations found in 28 genes of our patients with single suture craniosynostosis. We predicted disease causality for all mutations with at least three scores above predetermined thresholds for any of MFS, HMMRE, SSR, SIFT, PMUT, MAPP, or PAM250. All four mutations matching these criteria described novel cases of craniosynostosis.

We also compiled a larger set of 105 missense mutations seen in the same control patients and patients with craniosynostosis, adding the previously known, multi-suture, and syndromic mutations. We applied the HUSCY method trained on the clinical data set, with the 0.82 score threshold described above. All test mutations found in the training set were removed prior to retraining. The 75 missense mutations observed in multiple nonaffected control patients were assumed to not be related to disease, while the 30 other mutations are accepted as causal in the literature. Our predictions select ten mutations of which all cause craniosynostosis. No neutral mutations were falsely predicted to be deleterious, yet twenty disease causal mutations remained undetected. The accuracy here is motivating, but again our results on clinical data favor specificity rather than sensitivity - it would be more attractive for clinical purposes to detect rather than reject.

CONCLUSIONS

We postulated that the context of the structural environment for a missense mutation could be inferred from sequence analysis without building structure models, challenging ourselves to use as much information as possible before progressing to direct structural analysis. Our objective to target effects on interface support residues by modeling nonlocal structural context appears successful in directing the automated learning of protein systems.

Novel missense mutations are going to keep coming. Roughly 50,000 nsSNPs are publicly reported in humans. An estimated 40,000 to 200,000 more are anticipated to already exist in populations under study en masse [89], through projects including the HapMap [90]. External populations such as Indians appear to be rich in genetic variation not yet considered in most projections of human variation [91]. Meanwhile, novel mutations will keep springing up throughout the course of our species, as they are under positive evolutionary selection to do so [92-95].

Our approach uses two machine learning systems to tersely combine sequence, evolutionary, and physicochemical approaches to measure conservation; protein specific, nonprotein specific, structural context specific, and non structural context specific substitution matrices; predictable aspects of protein structure; and patterns of conservation and amino acid type within separable parts of the structural environment inferred from the structural parameters. In effect we build on the vast work of others to model importance to interaction, structure, and thereby stability, using only primary sequence. This work offers novel tools for predicting disease risk from missense mutations and for interpreting the mechanistic basis of disease.

ACKNOWLEDGEMENTS

We would like to thank Yana Bromberg for the friendly help in running the *in vitro* data set through the SNAP method. We thank those who made their software and data sets publicly available for the advancement of science, particularly Jianlin Chang. We also thank the many members of the Samudrala computational biology group who gave advice on the science and data presentation, particularly Brady Bernard, Michal Guerquin, Kajohnkiart Janebodin, Aaron Goldman, Adrian Laurenzi, and Michael Zhao. This work was supported by

NIH NIDCR F30 DE 017522 to JAH ("Individual predoctoral dentist scientist fellowship"), an NSF CAREER award to RS, and NIH NIDCR R01 DE 018227 to MLC ("Single suture craniosynostosis gene expression and discovery").

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publishers Web site along with the published article.

REFERENCES

[1] Wang, Z.; Moult, J. SNPs, Protein Structure, and Disease. *Hum. Mutat.*, **2001**, *17(4)*, 263-270.

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

[3] Yue, P.; Li, Z.; Moult, J. Loss of protein structure stability as a major causative factor in monogenic disease. *J. Mol. Biol.*, **2005**, *353(2)*, 459-473.

[4] Pakula, A.A.; Sauer, R.T. Genetic analysis of protein stability and function. *Annu. Rev. Genet.*, **1989**, *23*, 289-310.

[5] Allali-Hassani, A.; Wasney, G.A.; Chau, I.; Hong, B.S.; Senisterra, G.; Loppnau, P.; Shi, Z.; Moult, J.; Edwards, A.M.; Arrowsmith, C.H.; Park, H.W.; Schapira, M.; Vedadi, M. A survey of proteins encoded by non-synonymous single nucleotide polymorphisms reveals a significant fraction with altered stability and activity. *Biochem. J.*, **2009**, *424(1)*, 15-26.

[6] Tokuriki, N.; Tawfik, D.S. Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol.*, **2009**, *19(5)*, 596-604.

[7] Bloom, J.D.; Arnold, F.H. In the light of directed evolution: pathways of adaptive protein evolution. *Proc. Natl. Acad. Sci. USA*, **2009**, *1*, 9995-10000.

[8] Jenwitheesuk, E.; Samudrala, R. Prediction of HIV-1 protease inhibitor resistance using a protein-inhibitor flexible docking approach. *Antiviral Therapy*, **2005**, *10*, 157-166.

[9] Gutteridge, A.; Thornton, J.M. Understanding nature's catalytic toolkit. *Trends Biochem. Sci.,* **2005**, *30*, 622-629.

[10] Wang, K.; Horst, J.A.; Cheng, G.; Nickle, D.; Samudrala, R. Protein meta-functional signatures from combining sequence, structure, evolution and amino acid property information. *PLoS Comp. Bio.*, **2008**, *4*, e1000181.

[11] Zhang, Y.; Skolnick, J. The protein structure prediction problem could be solved using the current PDB library. *Proc. Natl. Acad. Sci. USA*, **2005**, *102(4)*, 1029-34.

[12] Zhang, Y. Progress and challenges in protein structure prediction. *Curr. Opin. Str. Biol.*, **2008**, *18*, 342-348.

[13] Moult, J.; Fidelis, K.; Kryshtafovych, A.; Rost, B.; Tramontano, A. Critical assessment of methods of protein structure prediction - Round VIII. *Proteins*, **2009**, *9*, 1-4.

[14] Chandonia, J.M.; Brenner, S.E. The impact of structural genomics: expectations and outcomes. *Science*, **2006**, *311*, 347-351.

[15] Dessailly, B.H.; Nair, R.; Jaroszewski, L.; Fajardo, J.E.; Kouranov, A.; Lee, D.; Fiser, A.; Godzik, A.; Rost, B.; Orengo, C. PSI-2: structural genomics to cover protein domain family space. *Structure*, **2009**, *17(6)*, 869-881.

[16] Protein Structure Inititiative. Structural Genomics Knowledgebase: TargetDB Statistics Summary Report. http://targetdb.pdb.org/ statistics/TargetStatistics.html (accessed November 11, 2009). Chen, L.; Oughtred, R.; Berman, H.B.; Westbrook, J. TargetDB: a target registration database for structural genomics projects. *Bioinformatics*, **2004**, *20*, 2860-2862.

[17] Schwede, T.; Sali, A.; Honig, B.; Levitt, M.; Berman, H.M.; Jones, D.; Brenner, S.E.; Burley, S.K.; Das, R.; Dokholyan, N.V.; Dunbrack, R.L.; Fidelis, K.; Fiser, A.; Godzik, A.; Huang, Y.J.; Humblet, C.; Jacobson, M.P.; Joachimiak, A.; Krystek, S.R.; Kortemme, T.; Kryshtafovych, A.; Montelione, G.T.; Moult, J.; Murray, D.; Sanchez, R.; Sosnick, T.R.; Standley, D.M.; Stouch, T.; Vajda, S.; Vasquez, M.; Westbrook, J.D.; Wilson, I.A. Outcome of a workshop on applications of protein models in biomedical research. *Structure*, **2009**, *17(2)*, 151-159.

[18] Michino, M.; Abola, E.; GPCR Dock Participants; Brooks, C.L.; Dixon, J.S.; Moult, J.; Stevens, R.C. Community-wide assessment of GPCR structure modelling and ligand docking: GPCR Dock 2008. *Nat. Rev. Drug. Discov.*, **2009**, *8(6)*, 455-463.

[19] Yue, P.; Melamud, E.; Moult, J. SNPs3D: candidate gene and SNP selection for association studies. BMC *Bioinformatics*, **2006**, *7*, 166.

[20] Burke, D.F.; Worth, C.L.; Priego, E.M.; Cheng, T.; Smink, L.J.; Todd, J.A.; Blundell, T.L. Genome bioinformatic analysis of nonsynonymous SNPs. *BMC Bioinformatics*, **2007**, *8*, 301.

[21] Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **1997**, *25*, 3389-3402.

[22] Cheng, J.; Randall, A.Z.; Sweredoski, M.J.; Baldi, P. SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res.*, **2005**, *33*, W72-W76.

[23] Cheng, J.; Saigo, H.; Baldi, P. Large-scale prediction of disulphide bridges using kernel methods, two-dimensional recursive neural networks, and weighted graph matching. *Proteins*, **2006**, *62(3)*, 617-629.

[24] Cheng, J.; Sweredoski, M.; Baldi, P. DOMpro: Protein Domain Prediction Using Profiles, Secondary Structure, Relative Solvent Accessibility, and Recursive Neural Networks. *Data. Min. Knowl. Discov.*, **2006**, *13(1)*, 1-10,.

[25] Cheng, J.; Baldi, P. Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics*, **2007**, *8*, 113.

[26] Tegge, A.N.; Wang, Z.; Eickholt, J.; Cheng, J. NNcon: improved protein contact map prediction using 2D-recursive neural networks. *Nucleic Acids Res.*, **2009**, *37*, W515-W518.

[27] Deng, X.; Eickholt, J.; Cheng, J. PreDisorder: ab initio sequence-based prediction of protein disordered regions. *BMC Bioinformatics*, **2009**, *10*, 436.

[28] Cheng, J.; Wang, Z.; Tegge, A.N.; Eickholt, J. Prediction of global and local quality of CASP8 models by MULTICOM series. *Proteins*, **2009**, *9*, 181-184.

[29] Berger, H.; Yanofsky, C. Suppressor selectrion for amino acid replacements expected on the basis of the genetic code. *Science*, **1967**, *156(773)*, 394-397.

[30] Dayhoff, M.O. Computer analysis of protein sequences. *Fed. Proc.*, **1974**, *33(12)*, 2314-2316.

[31] Thorn, K.S.; Bogan, A.A. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*, **2001**, *17(3)*, 284-285.

[32] Miyazawa, S.; Jernigan, R.L. Protein stability for single substitution mutants and the extent of local compactness in the denatured state. *Protein Eng.*, **1994**, *7*, 1209-1220.

[33] Topham, C.M.; Srinivasan, N.; Blundell, T.L. Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. *Protein Eng.*, **1997**, *101*, 46-50.

[34] Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, **1992**, *89*, 10915-10919.

[35] Kawashima, S.; Pokarowski, P.; Pokarowska, M.; Kolinski, A.; Katayama, T.; Kanehisa, M. AAindex: amino acid index database, progress report. *Nucleic Acids Res.*, **2008**, *36*, D202-D205.

[36] Bromberg, Y.; Rost, B. Correlating protein function and stability through the analysis of single amino acid substitutions. *BMC Bioinformatics*, **2009**, *8*, S8.

[37] Dang, L.X.; Merz, K.M.; Kollman, P.A. Free-energy calculatons on protein stability: Thr-157 val-157 mutation of t4 lysozyme. *J. Am. Chem. Soc.*, **1989**, *111*, 8505-8508.

[38] Capriotti, E.; Fariselli, P.; Casadio, R. A neural network-based method for predicting protein stability changes upon single point mutations. In: *Proceedings of the 2004 conference on intelligent systems for molecular biology* (ISMB04), Bioinformatics. New York: Oxford University,Press, *S1*, 190-201.

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

[40] Villegas, V.; Viguera, A.R.; Aviles, F.X.; Serrano, L. Stabilization of proteins by rational design of alpha-helix stability using helix/coil transition theory. *Fold Des.*, **1996**, *1*, 29-34.

[41] Gillis, D.; Rooman, M. Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence. *J. Mol. Biol.*, **1997**, *272*, 276-290.

[42] Terp, B.N.; Cooper, D.N.; Christensen, I.T.; Jørgensen, F.S.; Bross, P.; Gregersen, N.; Krawczak, M. Assessing the relative importance of the biophysical properties of amino acid substitutions associated with human genetic disease. *Hum. Mutat.*, **2002**, *20(2)*, 98-109.

[43] Stitziel, N.O.; Tseng, Y.Y.; Pervouchine, D.; Goddeau, D.; Kasif, S.; Liang, J. Structural location of disease-associated single-nucleotide polymorphisms. *J. Mol. Biol.*, **2003**, *327*, 1021-1030.

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

[45] Cheng, G.; Qian, B.; Samudrala, R.; Baker, D. Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. *Nucleic Acids Res.*, **2005**, *33*, 5861-5867.

[46] Shoemaker, B.A.; Portman, J.J.; Wolynes, P.G. Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. *Proc. Natl. Acad. Sci. USA*, **2000**, *97*, 8868-8873.

[47] López, G.; Ezkurdia, I.; Tress, M.L. Assessment of ligand binding residue predictions in CASP8. *Proteins*, **2009**, *9*, 138-146.

[48] Ng, P.C.; Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.*, **2001**, *11*, 863-874.

[49] Stone, E.A.; Sidow, A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res.*, **2005**, *15*, 978-986.

[50] Fleming, M.A.; Potter, J.D.; Ramirez, C.J.; Ostrander, G.K.; Ostrander, E.A. Understanding missense mutations in the BRCA1 gene: an evolutionary approach. *Proc. Natl. Acad. Sci. USA*, **2003**, *100*, 1151-1156.

[51] Santibanez, K.o.r.e.f.; Gangeswaran, R.; Santibanez, K.o.r.e.f.; Shanahan, N.; Hancock, J.M. Aphylogenetic approach to assessing the significance of missense mutations in disease genes. *Hum. Mutat.*, **2003**, *22*, 51-58

[52] Yue, P.; Moult, J. Identification and analysis of deleterious human SNPs. *J. Mol. Biol.*, **2006**, *356(5)*, 1263-1274.

[53] del, S.o.l.; Pazos, F.; Valencia, A. Automatic methods for predicting functionally important residues. *J. Mol. Biol.*, **2003**, *326*, 1289-1302.

[54] Saunders, C.T.; Baker, D. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J. Mol. Biol.*, **2002**, *322*, 891-901.

[55] Ferrer-Costa, C.; Orozco, M.; de, l.a. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *J. Mol. Biol.*, **2002**, *315*, 771-786.

[56] Krishnan, V.G.; Westhead, D.R. A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. *Bioinformatics*, **2003**, *19*, 2199-2209.

[57] Thomas, P.D.; Kejariwal, A. Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. *Proc. Natl. Acad. Sci. USA*, **2004**, *101*, 15398-15403.

[58] Chasman, D.; Adams, R.M. Predicting the functional consequences of nonsynonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol.*, **2001**, *307*, 683-706.

[59] Lee, C. Testing homology modeling on mutant proteins: predicting structural and thermodynamic effects in the ala98-val mutants of t4 lysozyme. *Fold Des.*, **1995**, *1*, 1-12.

[60] Cozzetto, D.; Kryshtafovych, A.; Tramontano, A. Evaluation of CASP8 model quality predictions. *Proteins*, **2009**, *9*, 157-166.

[61] Pruitt, K.D.; Tatusova, T.; Maglott, D.R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucl. Acids. Res.*, **2005**, *33*, D501-D504.

[62] Biegert, A.; Söding, J. Sequence context-specific profiles for homology searching. *Proc. Natl. Acad. Sci. USA*, **2009**, *106*, 3770-3775.

[63] Söding, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **2005**, *21(7)*, 951-960.

[64] Margelevicius, M.; Venclovas, C. PSI-BLAST-ISS: an intermediate sequence search tool for estimation of the position-specific alignment reliability. *BMC Bioinformatics*, **2005**, *6*, 185.

[65] Horst, J.A.; Samudrala, R. Diversity of protein structures and difficulties in fold recognition: the curious case of protein G. *F1000 Biology Reports*, **2009**, *1*, 69.

[66] Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **2004**, *32*, 1792-1797.

[67] Eddy, S.R. Profile hidden Markov models. *Bioinformatics*, **1998**, *14*, 755-763.

[68] Wang, K.; Samudrala, R. Incorporating background frequency improves entropy-based residue conservation measures. *BMC Bioinformatics*, **2006**, *7*, 385.

[69] Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **1981**, *17*, 368-376.

[70] Gromiha, M.M.; Selvaraj, S. Inter-residue interactions in protein folding and stability. Prog Biophys Mol Biol, **2004**, 86, 235-277.

[71] Boser, B.E.; Guyon, I.; Vapnik, V. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. ACM Press, **1992**, 144-152.

[72] Cortes, C.; Vapnik, V. Support-vector network. *Machine Learning*, **1995**, *20*, 273-297.

[73] Keerthi, S.S.; Lin, C.-J. Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Comput.*, **2003**, *15(7)*, 1667-1689.

[74] Lin, H.-T.; Lin, C.-J. A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods. *Technical report*, Department of Computer Science, National Taiwan University. **2003**.

[75] Schölkopf, B.; Platt, J.; Shawe-Taylor, J.; Smola, A.J.; Williamson, R.C. Estimating the support of a high-dimensional distribution. *Neural. Comput.*, **2001**, *13*, 1443-1471.

[76] Chang, C.C.; Lin, C.J. LIBSVM: a library for support vector machines, **2001**. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[77] Serang, O.; Noble, W.S. A fast regularization pathway for the support vector machine. **2009**. Software available at http://noble.gs.washington.edu/proj/svmvia.

[78] Joachims, T. In: *Making large-Scale SVM Learning Practical.* Schölkopf, B.; Burges, C.; Smola, A. Eds.; MIT-Press, Cambridge, MA. 1999.

[79] Loeb, D.D.; Swanstrom, R.; Everitt, L.; Manchester, M.; Stamper, S.E.; Hutchison, C.A. Complete mutagenesis of the HIV-1 protease. *Nature*, **1989**, *340(6232)*, 397-400.

[80] Rennell, D.; Bouvier, S.E.; Hardy, L.W.; Poteete, A.R. Systematic mutation of bacteriophage T4 lysozyme. *J. Mol. Biol.*, **1991**, *222(1)*, 67-88.

[81] Markiewicz, P.; Kleina, L.G.; Cruz, C.; Ehret, S.; Miller, J.H. Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence. *J. Mol. Biol.*, **1994**, *240(5)*, 421-433.

[82] Gromiha, M.; An, J.; Kono, H.; Oobatake, M.; Uedaira, H.; Prabakaran, P.; Sarai, A. Protherm, version 2.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res.*, **2000**, *28*, 283-285.

[83] Thorn, K.S.; Bogan, A.A. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*, **2001**, *17(3)*, 284-285.

[84] Online Mendelian Inheritance in Man, OMIM (TM). McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), {date of download}. World Wide Web URL: http://www.ncbi.nlm.nih.gov/omim/

[85] Kawabata, T.; Ota, M.; Nishikawa, K. The protein mutant database. *Nucleic Acids Res.*, **1999**, *27*, 355-357.

[86] Cunningham, M.L.; Horst, J.A.; Rieder, M.; Hing, A.; Park, S.; Speltz, M. IGF1R variants associated with isolated single suture craniosynostosis. *Am. J. Hum. Genet.* (*in review*).

[87] Zweig, M.H.; Campbell, G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin. Chem.*, **1993**, *39(4)*, 561-577.

[88] Ramensky, V.; Bork, P.; Sunyaev, S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.*, **2002**, *30(17)*, 3894-900.

[89] Ng, P.C.; Henikoff, S. Predicting the Effects of Amino Acid Substitutions on Protein Function. *Annu. Rev. Genomics Hum. Genet.*, **2006**, *7*, 61-80

[90] International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **2007**, *449(7164)*, 851-861.

[91] Reich, D.; Thangaraj, K.; Patterson, N.; Price, A.L.; Singh, L. Reconstructing Indian population history. *Nature*, **2009**, *461(7263)*, 489-494.

[92] Blundell, T.L.; Wood, S.P. Is the evolution of insulin Darwinian or due to selectively neutral mutation? *Nature,* **1975**, *257*, 197-203.

[93] Sawyer, S.A.; Kulathinal, R.J.; Bustamante, C.D.; Hartl, D.L. Bayesian analysis suggests that most amino acid replacements in Drosophila are driven by positive selection. *J. Mol. Evol.*, **2003**, *1*, 154-164.

[94] Bazykin, G.A.; Kondrashov, F.A.; Ogurtsov, A.Y.; Sunyaev, S.; Kondrashov, A.S. Positive selection at sites of multiple amino acid replacements since rat-mouse divergence. *Nature*, **2004**, *429*, 558-562.

[95] DePristo, M.; Weinreich, D.; Hartl, D. Missense meanderings in sequence space: A biophysical view of protein evolution. *Nat. Rev. Genet.*, **2005**, *6(9)*, 678-687.

[96] Krivov, G.G.; Shapovalov, M.V.; Dunbrack, R.L. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*, **2009**, *77(4)*, 778-795.

Received: ???????????????? **Revised:** ???????????????? **Accepted:** ????????????????