Published in final edited form as:

Hum Mutat. 2012 August ; 33(8): 1267–1274. doi:10.1002/humu.22110.

Phenotype-Optimized Sequence Ensembles Substantially Improve Prediction of Disease-Causing Mutation in Cystic Fibrosis

David L. Masica¹, Patrick R. Sosnay^{2,3}, Garry R. Cutting², and Rachel Karchin^{1,*}

¹Department of Biomedical Engineering and Institute for Computational Medicine, Johns Hopkins University, Baltimore, Maryland ²McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland ³Perdana University Graduate School of Medicine, Serdang, Malaysia

Abstract

Cystic fibrosis transmembrane conductance regulator (CFTR) mutation is associated with a phenotypic spectrum that includes cystic fibrosis (CF). The disease liability of some common CFTR mutations is known, but rare mutations are seen in too few patients to categorize unequivocally, making genetic diagnosis difficult. Computational methods can predict the impact of mutation, but prediction specificity is often below that required for clinical utility. Here, we present a novel supervised learning approach for predicting CF from CFTR missense mutation. The algorithm begins by constructing custom multiple sequence alignments called phenotype-optimized sequence ensembles (POSEs). POSEs are constructed iteratively, by selecting sequences that optimize predictive performance on a training set of CFTR mutations of known clinical significance. Next, we predict CF disease liability from a different set of CFTR mutations (test-set mutations). This approach achieves improved prediction performance relative to popular methods recently assessed using the same test-set mutations. Of clinical significance, our method achieves 94% prediction specificity. Because databases such as HGMD and locus-specific mutation databases are growing rapidly, methods that automatically tailor their predictions for a specific phenotype may be of immediate utility. If the performance achieved here generalizes to other systems, the approach could be an excellent tool to help establish genetic diagnoses.

© 2012 Wiley Periodicals, Inc.

*Correspondence to: Rachel Karchin, Johns Hopkins University, Biomedical Engineering, 217A Hackerman Hall. CSEB, 3400 N. Charles St., Baltimore, Maryland. karchin@jhu.edu.

Communicated by Marc S. Greenblatt

Additional Supporting Information may be found in the online version of this article.

Supplemental Data: A data supplement accompanies this manuscript, and includes the following: a schematic illustrating our new score function (Supp. Fig. S1); the 10 most enriched sequences for domain-specific CFTR POSEs across three independent trials (Supp. Table S1); domain-specific performance achieved by our method in three independent trials (Supp. Table S2); a list of all amino-acid attributes used during scoring (Supp. Table S3); a list of all CF-causing and CF-neutral missense mutations used for algorithmic training and testing (Supp. Table S4)

Keywords

phenotype-optimized sequence ensembles; POSE; CFTR; cystic fibrosis; prediction

Introduction

Enhanced technology and a shifting focus to personalized medicine have led to greater investigation of the genetic contributions to disease [Chan and Ginsburg, 2011; Hamburg and Collins, 2010]. In the case of Mendelian traits influenced by high-effect alleles that contribution is generally easy to recognize. Increasingly however, genetic variants are identified that have uncertain consequence on disease trait [Borg et al., 2010; Chao et al., 2008; Karchin et al., 2008; Sosnay et al., 2011]. Individuals identified with these variants face uncertain prognosis, and/or have to make reproductive decisions with incomplete information. Characterizing variant disease liability is typically informed by carrier clinical data, but many variants occur in too few patients to make statistical comparisons from clinical features. Furthermore, there are too many variants to make functional studies practical. Computational methods can address the clinical need to interpret these variants and examine the effect of genetic diversity on a much wider scale than traditional Mendelian genetics.

Missense mutations present an interpretive challenge because the possible impact of amino acid substitution on protein activity, cellular behavior, and disease phenotype can range from benign to damaging. Computational methods exist for predicting the impact of amino acid substitution, and typically rely on sequence, structure, and/or annotation [Cline and Karchin, 2011; Jordan et al., 2010; Ng and Henikoff, 2006]. Sequence-based prediction utilizes patterns derived from a multiple sequence alignment (MSA) of the gene of interest and a collection of homologous sequences that are putatively descended from a common ancestral sequence [Karchin et al., 2005; Mi et al., 2007; Ng and Henikoff, 2003; Ramensky et al., 2002]. There are many approaches for selecting the sequences to include in the MSAs [Hicks et al., 2011; Karchin, 2009]. Structure-based methods assess the impact of mutation on the estimated stability of protein native structure [Potapov et al., 2009], but in the absence of a high-resolution experimental structure these methods are severely limited. Annotation assigns known biological function to specific protein regions [Karchin et al., 2005; Ramensky et al., 2002]. Annotation can enhance prediction because mutation in structurally and functionally important locations can be deleterious (e.g., catalytic residue, ligand-binding site, or protein-protein interface); unfortunately, this information is typically sparse.

The ability of computational approaches to predict the impact of missense mutation can be assessed by comparing predictions with a gold standard. Locus-specific databases (LSDBs) that inventory the consequence of genetic mutation may be used as such a standard [Cotton et al., 1996]. For the disease cystic fibrosis (CF; MIM# 219700) [Collins, 1992; Rosenstein and Cutting, 1998], caused by mutation in the cystic fibrosis transmembrane conductance regulator (CFTR; MIM# 602421) [Riordan, 2008; Sosnay et al., 2011], the Clinical and Functional TRanslation of CFTR (CFTR2) database (www.CFTR2.org) contains CFTR

mutation and clinical data for 39,545 CF patients, including 392 distinct missense mutations. To confirm the disease liability of 56 of the most commonly seen missense mutations, clinical information was compared with cell-based functional assays to ensure that mutations met independent clinical and functional criteria to be identified as disease causing. Although the CFTR protein lacks a high-resolution structural model, this database and the well-described protein domains can both be utilized to inform a computational predictor.

Several groups have predicted CF from CFTR missense mutation using computational approaches [Dorfman et al., 2010; Gaucher et al., 2006; George Priya Doss et al., 2008]. Priya-Doss et al. (2008) examined publically available computational tools. Their analysis combined sequence-based approaches, including SIFT and PolyPhen, and structure-based methods, including NOMAD, SRide, and FoldX. The predictions made in this study were not correlated against a definitive LSDB, and contained several incorrect calls, highlighting the difficulty of prediction and the limitation of existing methods. A potential complicating factor, with respect to structure-based prediction, is the absence of a high-resolution experimental structure for full-length CFTR.

Gaucher et al. (2006) developed DETECTER, a method based on phylogenetic analysis of fast- and slow-evolving codons, and tested performance using the CF Mutation Database (CFMD), a precursor to the CFTR2 database. To enhance their prediction, they examined patterns of polar and nonpolar regions of the membrane-spanning domains of CFTR. This approach resulted in slightly improved prediction performance relative to SIFT. However, the CFMD is a catalog of genetic variance, rather than a true LSDB, and the disease liability of many mutations is unknown. Therefore, it is unclear to what extent, if any, DETECTER outperformed SIFT.

In 2010, Dorfman et al. (2010) compared the performance of several popular sequenced-based approaches, namely, SIFT, PolyPhen, and PANTHER, using a carefully hand-curated set of known CF-causing and CF-neutral missense mutations. Owing to poor prediction specificity, the authors argue against using the tested algorithms to confirm or refute CF diagnosis. Unfortunately, classifying missense mutation with computational approaches often results in poor prediction specificity, limiting clinical utility. Of the methods tested, PANTHER had the best specificity. Notably, PANTHER uses a score derived from an alignment of a small number of orthologs, whereas SIFT and PolyPhen use larger, more diverse alignments that include both orthologs and paralogs. Hicks et al. (2011) recently showed that specificity, and prediction performance in general, can vary drastically as a function of sequence alignment employed. That study used four popular methods to predict impact of missense mutation on four cancer-related genes.

Here, we develop a new sequence-based computational method to predict disease from missense mutation, and apply it to predict CF from CFTR missense mutation. This method builds on our previous work using whole-genome MULTIZ alignments [Blanchette et al., 2004] to estimate conservation at a codon [Carter et al., 2009]. The MULTIZ alignments provide a pool of CFTR orthologs and paralogs. From this pool, an optimized sequence ensemble is then iteratively reestimated by optimizing performance of a scoring function on a training set of mutations of well-known disease liability. In contrast to standard sequence

selection algorithms, the sequence ensemble can contain multiple copies of the same sequence. The score function integrates physiochemical properties of amino acid residues with evolutionary relationships between the target disease protein and its homologs, including additional weights generated by sequence duplicates in the ensemble. This method aims to enhance specificity that has been lacking in computational predictors.

Materials and Methods

We developed a novel supervised learning approach to predict mutation disease liability in high-penetrance genes, based on protein sequence evolution and amino acid residue physiochemistry. The algorithm isolates collections of sequences from a precomputed MSA, generated from a comparative genomics dataset of orthologs and paralogs. These collections are optimized to enable our score function to discriminate disease-causing mutations; we refer to such a collection of sequences as a *phenotype-optimized sequence ensemble* (POSE). POSEs can be used for blind prediction on mutations of known or unknown clinical significance.

Algorithm Development and Implementation

The algorithm (Fig. 1) begins by selecting a random ensemble of 150 or fewer sequences from the initial MSA. Next, the algorithm scores (see section Score Function) a training set of known disease causing and other mutations; a score greater than or equal to 0.0 is considered disease causing. Because the true class of each mutation in the training set is known, we can evaluate the classification performance (defined as the sum of sensitivity and specificity) of the scoring function, given the selected sequences (see section Performance Metrics). This process of randomly selecting a sequence ensemble, using that ensemble to score mutations, and evaluating the scores is repeated many times (iterations). Every 100 iterations, the top-performing 1% of random sequence ensembles, are appended to the MSA, resulting in multiple copies of top-performing sequences. A randomly chosen MSA will perform well if it accurately captures the selection inherent in the disease-positive and disease-negative mutations being scored. Therefore, as the algorithm progresses through many hundreds of iterations, the MSA becomes enriched for sequences that best discriminate between disease-causing and other mutations. In contrast to classic MSAs used in missense mutation function prediction, a POSE may contain as many as 100+ copies of a single ortholog or paralog sequence (see Supp. Table S1); that is, the influence of individual sequences on mutation discrimination is differentially weighted.

For this study, one complete execution of the above-described algorithm consisted of 25,000 iterations; from each execution, we selected the 10 top-performing POSEs for subsequent analysis and blind prediction. Because CFTR is composed of evolutionarily distinct domains, we identified POSEs for each of five CFTR domains independently. And, because optimization of this type is not exhaustive and can be prone to sampling only local minima, we performed 100 independent algorithmic executions; this number of independent executions resulted in excellent performance (Supp. Table S2) and sequence (Supp. Table S1) convergence. Therefore, for this study, we sampled 2.5×10 sequence ensembles for each of five domains, resulting in a total of 1.25×10^7 random sequence ensembles sampled.

During blind prediction, we averaged predictions from the 10 top-performing POSEs, from each of the 100 independent executions, for final prediction. Finally, we tested algorithmic stability (obtaining the same prediction performance, for example) by repeating this entire process in triplicate.

The program was written in the python programming language. The optimization process (i.e., constructing CFTR POSEs) was executed on an IBM iDataPlex cluster with 500 Quad Core Intel Xeon E5472 processors (2,000 processor cores), at the Institute for Computational Medicine, Johns Hopkins University.

Score Function

To predict the impact of mutation, we developed a score function that considers both biophysical molecular properties and evolutionary distance of related proteins to a target disease protein (Supp. Fig. S1). The score (S) for a specific amino acid residue at a specific position in the alignment is calculated as the sum of an amino acid conservation score (S_{aa}) and a biophysical properties conservation score (S_{prop}), minus a molecular weight conservation score (S_{mw} ; Equation 1). A score can be calculated for any amino acid residue at any alignment column, regardless of whether the amino acid appears in that column.

$$S = S_{aa} + S_{prop} - S_{mw} \quad (1)$$

We weight the contribution of each sequence in the alignment, by an estimate of its evolutionary distance from the target sequence. Each sequence contributes an amount proportional to its sequence identity, relative to the target. Thus, to calculate the amino acid conservation score (S_{aa}) for a specific amino acid, the algorithm sums the sequence identities (Ω) of all sequences in the alignment that harbor that specific amino acid (n_{aa}), at the position being scored (Equation 2):

$$S_{aa} = \frac{1}{L} \sum_{i=1}^{n_{aa}} \Omega_i, \quad (2)$$

$$L = \sum_{i=1}^N \Omega_i. \quad (3)$$

In Equation 2, L is a normalizing constant, which is the sum of sequence identities for all N sequences in the alignment (Equation 3). Similarly, to calculate the biophysical properties conservation score for a specific amino acid, the algorithm sums the sequence identities of all sequences in the alignment that harbor that specific property, at the position being scored (see Equation 4; Supp. Table S3); the biophysical properties conservation score is calculated for each property that defines the amino acid being scored (m_{prop}). This score is additionally normalized by the total number of properties defining the amino acid being scored (n_{prop}). The additional normalization constant assures that amino acids with a greater number of properties are not artificially favored relative to amino acids with fewer properties. The molecular weight conservation score (S_{mw}) is the difference between the molecular weight

of the amino acid being scored (x) and the mean for that column in the alignment (μ), divided by the standard deviation of molecular weight (σ) for that column in the alignment (i.e., a Z-score; Equation 5). In Equation 5, the molecular weight coefficient (C_{mw}) is 0.1. This coefficient was chosen because we observed higher values of C_{mw} resulting in a score function dominated by S_{mw} . The score for an amino acid substitution (S_{sub} ; Equation 6) is this difference between the wild type (S_{WT}) and mutant amino acids scores (S_{Mut}), where S_{WT} and S_{Mut} are each calculated from Equation 1. Supp. Figure S1 is a schematic illustration of the score function developed here:

$$S_{prop} = \frac{1}{n_{prop}} \sum_{i=1}^{n_{prop}} \frac{1}{L} \sum_{j=1}^{m_{prop}} \Omega_{ij}, \quad (4)$$

$$S_{mw} = C_{mw} + Z_{mw}, \text{ where } Z_{mw} = \frac{x - \mu}{\sigma}, \quad (5)$$

$$S_{sub} = S_{WT} - S_{Mut}. \quad (6)$$

Supp. Table S3 shows the amino acid properties and molecular weights used by the score function. The properties include standard biophysical properties associated with the side chains of each of the 20 naturally occurring amino acids (hydrogen bond donating or accepting, negatively, or positively charged, etc.). Histidine is unique because the side chain titrates at physiological pH, resulting in varied charge and hydrogen bond capacity; therefore, we define histidine uniquely (see Supp. Table S3). Similarly, cysteine is unique in its ability to form inter-side-chain covalent bonds and glycine is unique because it lacks a side chain; therefore, we define cysteine and glycine uniquely. Proline mutation has a unique capacity to perturb well-defined protein secondary structure because it lacks a backbone amide proton and has a restricted backbone Φ torsion angle; therefore, mutating to a proline and mutating from a proline are handled distinctly by our score function (see Supp. Table S3). Molecular weights are in units of g/mol.

Polar amino acid substitutions in membrane-spanning regions of proteins convey a higher-than-average propensity for disease [Partridge et al., 2002, 2004]. Several studies suggest this relationship also applies to CFTR [Choi et al., 2004; Partridge et al., 2004]. Aligning CFTR membrane-spanning regions to Schiffer–Edmunson helical wheels reveals many highly polar residues spaced three to four sites from each other, suggesting these residues face one another and interact electrostatically [Gaucher et al., 2006]. To capture the importance of charge in the membrane-spanning regions, we added a penalty of +5.0 for any amino acid substitution resulting in a change of charge (e.g., positive to neutral, negative to positive) in the membrane spanning regions. Boundaries defining the segments are the same as in Mornon et al. (2008).

Performance Metrics

Sequence selection during POSE construction relied on the specificity and sensitivity performance metrics. Sensitivity is defined as number of true positives (TP) divided by the

sum of TP and false negatives (FN). Specificity is defined as the number of true negatives (TN) divided by the sum of TN and false positives (FP). During blind prediction, we additionally considered the positive predictive value (PPV) and negative predictive value (NPV), where the PPV is defined as $TP/(TP + FP)$ and NPV as $TN/(TN + FN)$.

CFTR Orthologs and Paralogs

CFTR is member 7 of subfamily C of the large family of evolutionarily related ABC transporters. For this study, we constructed the MSA's from all CFTR (ABCC7) orthologs, and CFTR paralogs (ABCC1-6 and 8–12) from the UCSC 46-vertebrate MULTIZ genome alignments [Blanchette et al., 2004]; this resulted in a total of 547 CFTR homologs.

CFTR Mutations

During algorithmic training, we built CFTR POSEs using 65 CF-causing CFTR mutations and 75 CF-neutral CFTR mutations. We constructed the training set of CF-causing mutations from the CFTR2 database (<http://www.cftr2.org>). Among the many clinical manifestations of CF curated with CFTR genotype in the CFTR2 database is patient sweat chloride. A high sweat chloride level is synonymous with the CF phenotype [Rosenstein and Cutting, 1998; Wilschanski et al., 1995], with a value of 60 mM/L or greater being diagnostic of disease and values between 40 and 60 defined as borderline. Inclusion in our CF-causing training set required a mutation be observed in at least three patients, the difference of the mean and standard deviation sweat chloride level was greater than or equal to 50 mM/L, and the mutation could not be present in the test set (see below). These requirements helped assure training-set CF-causing CFTR missense mutations were consistently CF causing.

To construct a high-quality set of 75 CF-neutral mutations, we curated variants that are seen in the general population with frequency greater than 0.1%, but not reported in the CFTR2 database (drawn from CF patients) [Castellani et al., 2008; Durbin et al., 2010; Sherry et al., 2001]. We also added mutations that had been extensively studied in vitro and found to have no trafficking or chloride conductance defects, and are not seen in the CFTR2 patient database [Baldursson et al., 2001; Park et al., 2002; Vankeerberghen et al., 1998; Wilkinson et al., 1997; Zegarra-Moran et al., 2007].

To benchmark algorithmic performance and compare with existing methods, we tested our method using the same CFTR mutations as in Dorfman et al. (2010). In that study, Dorfman and colleagues tested the ability of SIFT [Ng and Henikoff, 2003], PolyPhen [Adzhubei et al., 2010], and PANTHER [Thomas and Kejariwal, 2004] to predict CF from CFTR missense mutation. This dataset includes 43 CF-causing and 60 CF-neutral CFTR missense mutations; the 60 CF-neutral mutations comprise 54 mutations associated with CFTR-related diseases and six mutations that present no known clinical manifestations. There was no overlap between the training set and this test set. Since the study by Dorfman et al. (2010) was completed, updated versions of SIFT and PolyPhen have been released (SIFT v. 1.03; PolyPhen v. 2.2.2); we used these updated versions for comparison in this study. Also, because SIFT and PolyPhen-2 have been reported to perform better when utilizing custom MSAs [Hicks et al., 2011], rather than the default MSA provided by these algorithms, we

also compare SIFT and PolyPhen-2 results when a POSE sequence alignment or the hand-curated CFTR alignment of Gaucher et al. (2006) is provided.

Results

We constructed phenotype-optimized sequence ensembles (POSEs) to predict CF from CFTR missense mutation, using a novel score function (see section Materials and Methods). First, we constructed POSEs using a set of known CF-causing and CF-neutral mutations (*Training Set*, Supp. Table S4). Next, we used those POSEs to make blind predictions on a different set of known CF-causing and CF-neutral mutations (*Test Set*, Supp. Table S4).

The first row in Table 1 shows four performance metrics (see section Materials and Methods) achieved by our method, during blind prediction. Using our score function with POSEs, prediction sensitivity was 74%, which is the fraction of CF-causing mutations correctly identified. The fraction of CF-neutral mutations correctly classified is the test specificity, which was 94%. For all mutations predicted to cause disease, the PPV is the fraction that actually cause disease; our method achieved 89% PPV during blind prediction. Conversely, for all mutations predicted CF-neutral, the NPV is the fraction that is truly CF-neutral; our method achieved an NPV of 85% during blind prediction.

Figure 2A shows the receiver-operating characteristic (ROC) curve resulting from blind prediction, using our score function with POSEs obtained during training. The area under the ROC curve for these predictions was 84%, and is the probability that our classifier would score a randomly chosen CF-causing mutation higher than a randomly chosen CF-neutral mutation. Figure 2B is a strip chart for results obtained during blind prediction, and is useful for showing how well the score function discriminates CF-causing and CF-neutral mutations. Scores for test-set mutations range from -2.75 (most benign) to 11.37 (greatest disease liability), and had a mean value of 2.58 with a standard deviation of 2.86. Of the 103 test-set mutations, two CF-neutral mutations had a score higher than one standard deviation above the mean (G550R and R933G), and two CF-causing mutations had scores lower than one standard deviation below the mean (G85V and T338I); these four mutations were consistently the most difficult to accurately rank with our method.

SIFT, PolyPhen-2, and PANTHER are three of the most widely used computational methods for predicting the impact of missense mutation. Table 1 shows the prediction performance of these three methods, using the same test-set mutations (*Test Set*, Supp. Table S4). All three methods had greater than 80% sensitivity, which was greater than the 74% sensitivity achieved by our approach. PANTHER achieved reasonable specificity at 74%, but SIFT and PolyPhen-2 had low specificity (35% and 15%, respectively); our approach shows a drastic increase in specificity at 94%. Comparing the mean of specificity and sensitivity, our method outperforms the other three at 84%, with PANTHER, SIFT, and PolyPhen-2 at 78%, 66.5%, 57.5%, respectively. SIFT, PolyPhen-2, and PANTHER all had poor PPVs, with PANTHER doing the best at 63%; our method achieved a significantly increased PPV of 89%. All three popular methods had reasonable NPV. Our method had the third-ranked NPV at 85% and PolyPhen-2 and SIFT had an NPV of 100% and 95%, respectively. Comparing

the mean of PPV and NPV, our method outperforms the other three at 87%, with PANTHER, SIFT, and PolyPhen-2 at 73%, 73.5%, and 73%, respectively.

To see if the increased performance achieved by our new algorithm was the result of using POSEs, or resulted solely from the novel score function, we scored the same test-set mutations using our score function and several MSAs of CFTR homologs not optimized to predict CF from CFTR missense mutation. First, we scored test-set mutations using our score function and an MSA comprising only the 46 CFTR (ABCC7) orthologs from the UCSC 46-vertebrate alignment (row 5, Table 1). While this approach achieved reasonable sensitivity and NPV, the specificity and PPV were low. In general, use of an ortholog-only MSA with our score function resulted in performance slightly greater than PolyPhen-2, slightly less than SIFT or PANTHER, and significantly reduced performance relative to our score function utilizing POSEs. Next, we scored test-set mutations using our score function and each of 11 CFTR paralogous groups (ABCC1-6 and 8–12) independently, using sequences from the UCSC 46-vertebrate alignment. Of the 11 paralogous MSAs, use of ABCC9 resulted in the best overall performance (row 6, Table 1). Using our score function with an ABCC9-only MSA resulted in better performance than using an ABCC7-only MSA, but reduced performance relative to our score function with POSEs. Finally, we scored the test-set mutations using our score function and all 547 ABCC sequences (ABCC1-12) from the UCSC 46-vertebrate alignment. This approach drastically increased specificity and PPV relative to SIFT, PolyPhen-2, and PANTHER; however, sensitivity using this approach was low (65%). And, by all performance metrics, our score function achieves increased performance when using POSEs, compared with an MSA comprising all ABCC-member sequences. This analysis shows that the increased performance of our method, relative to previous methods, results from both the novel score function and use of POSEs.

Providing SIFT or PolyPhen-2 with custom alignments can result in increased performance relative to the default alignments employed by the algorithms [Hicks et al., 2011]. Therefore, we predicted the impact of the same test-set mutations using SIFT and PolyPhen-2, substituting the default alignments with the hand-curated CFTR alignment of Gaucher et al. (2006). Table 1 (rows 8 and 10) show prediction performance for SIFT and PolyPhen-2, respectively, using the hand-curated alignment. Whereas both algorithms make many false-positive calls utilizing their default alignments (i.e., high sensitivity and low specificity), the hand-curated alignments facilitate greater balance in predictive performance. While the hand-curated alignment improves true negative prediction and maintains reasonable true positive prediction for both SIFT and PolyPhen-2, their sum of sensitivity and specificity is still significantly lower than PANTHER's, and in particular, that achieved by our new method.

To investigate the impact of using a POSE in conjunction with an existing missense mutation score functions, we scored the test-set mutations using the SIFT and PolyPhen-2, but substituted their default MSAs for human CFTR with POSEs from this study (Table 1, rows 9 and 11, respectively). Because our POSEs contained more sequences than SIFT allows, MSAs submitted to the SIFT server contained the 10 most enriched POSE sequences, for each of five CFTR domains (see Supp. Table S1, *Trial 1*). For both SIFT and PolyPhen-2, substituting the default sequence alignments with POSE sequences resulted in

increased sensitivity, but decreased specificity. For instance, use of POSE sequences resulted in a 35% decrease in sensitivity, but a 47% increase in specificity. PolyPhen-2 achieved a 32% increase in specificity, at the cost of a 14% decrease in sensitivity. Similarly, SIFT had a 19% increase in PPV using POSE sequences, rather than SIFT's default CFTR MSA, but suffered a 20% decrease in NPV. And, PolyPhen-2 had an 8% increase in PPV, but a decrease in NPV of 18%. This result suggests that POSEs may be useful to improve the specificity of other missense mutation scoring functions, although at the cost of reducing sensitivity.

The exhaustive exploration of all sequence combinations during POSE construction is computationally intractable. Also, POSE construction is stochastic, rather than deterministic. Taken together, these limitations could result in POSEs of inconsistent sequence composition and variable performance, which would reduce generalizability. To address these potential limitations we ran the entire prediction process in triplicate. Aside from a slight deviation in the second transmembrane domain (TMD2), performance metrics were identical (within 0.1%) for all three trials, for each CFTR domain (Supp. Table S2). For TMD2, trial 2 had increased sensitivity but decreased specificity, relative to trials 1 and 3, which were identical (Supp. Table S2). Comparing prediction performance among the CFTR domains, TMD2 was generally the lowest. Prediction performance was high (100% accuracy) for the second nucleotide-binding domain (NBD2) for all metrics, across all three trials.

Consistency in POSE sequence composition was also remarkably stable across the three independent trials (Supp. Table S1). For instance, horse and guinea pig ABCC11, tarsier ABCC4, and treeshrew ABCC9 were among the 10 most enriched sequences for TMD1 (the first transmembrane domain) POSEs in all three trials. For the first nucleotide-binding domain (NBD1), POSEs from all three trials were composed exclusively of ABCC3 and ABCC6 members, and a single ABCC1 member. The regulatory domain (R domain) POSEs contained only ABCC10 and ABCC11 members, except for trial 3, which additionally contained a single ABCC2 sequence. While the 10 most enriched sequences from TMD2 POSEs, for each trial, contained chicken ABCC2 and cat ABCC8, these POSEs had the most variability across trials; this variability might be related to the observed performance variability for TMD2 predictions (see the preceding paragraph). NBD2 POSEs were dominated by two sequences—hedgehog and kangaroo rat ABCC7—for each independent trial; this unique solution to constructing NBD2 POSEs might be related to the high-prediction performance, discussed in the previous paragraph.

Discussion

With the rapid increase in genetic screening of both healthy and disease populations, biomedical researchers are compiling comprehensive genetic catalogues of human disease [Cotton et al., 1996; Sherry et al., 2001; Stenson et al., 2003]. Direct experimental and clinical validation of disease genotypes will remain the hallmark of definitive classification in the foreseeable future, but the large number of coding variants exceeds the time and resources available to perform such validation. Therefore, bioinformatics analysis needs to be evaluated in genes where it is possible to assess the accuracy of predictions. These well-

characterized genes will provide a benchmark to determine how well bioinformatics methods can address the diversity of genetic variance as it relates to disease.

The utility of disease-genotype databases is dependent on our ability to assess disease liability from genetic mutation. While predictive capability is increasing, computational characterization of genetic mutation is not yet reliable enough to confirm or refute molecular diagnosis [Dorfman et al., 2010; Jordan et al., 2010; Ng and Henikoff, 2006]. In this study, we developed a novel sequence-based approach for predicting disease from missense mutation, and applied it to predicting CF from CFTR missense mutations. We hypothesized that there exists one or more MSAs, whose composition of orthologs and paralogs optimally capture the selective pressures that define biologically important amino acid residues in a disease gene (such as CFTR). These are the residues whose perturbation yields disease (such as CF). Indeed, MSA composition can drastically affect the performance of sequence-based methods [Hicks et al., 2011]. Our method determines an optimal MSA (POSE) in an automated fashion, which resulted in improved prediction performance in this study.

MSAs comprising only orthologous sequences can have high sensitivity because high sequence conservation predicts low-tolerance for substitution, but this comes at the expense of high false-positive rates. Unfortunately, these high false-positive rates translate to poor specificity. The converse of this problem can occur when large super-family alignments are used. Because this type of alignment contains diverse sequences, the probability of observing the mutation of interest in the alignment increases, and so does prediction false-negative rate (i.e., reduced sensitivity). These limitations, in part, prevent sequence-based computational approaches from having real clinical utility. Our method of selecting sequences that maximize the sum of specificity and sensitivity appears to balance sequence conservation and diversity enough to partially address these shortcomings.

In clinical scenarios, both high-sensitivity and high-specificity predictors are needed for the molecular diagnosis of disease [Kraft et al., 2009; Pepe et al., 2004]. Screening of asymptomatic individuals (for example carrier screening or newborn screening) in particular necessitates that the post-test risk be well understood [Watson et al., 2006]. The clinical utility of this or any diagnostic tool is best interpreted by experienced clinicians and genetic counselors. Because the observed manifestations of disease vary considerably and are subject to genetic and environmental modifiers, the classification of mutation is not expected to be 100% accurate. Therefore, any computational methods trained to recover clinical and experimental findings will likely have accuracy limits. In this study, we constructed POSEs that maximized the sum of specificity and sensitivity. This choice was made, in part, to facilitate direct comparison with previous studies that used the same performance benchmark. However, an important flexibility of our method is the ability to select POSEs that optimize any desired performance metric. Another approach might be to combine multiple methods, accounting for individual strengths and weaknesses. In either case (i.e., using multiple methods or a single method with multiple POSE selection criteria), one predictor could be used for sensitivity and the other for specificity, and each predictor could be favored in the appropriate clinical setting.

An additional flexibility inherent in the design of our algorithm is the ability to use different classification schemes to construct POSEs. In this study, we constructed POSEs that ultimately optimized the binary classification of CF causing or CF neutral. An alternative implementation of the algorithm could be to classify a continuous phenotype. For instance, sweat chloride concentration is a commonly used clinical diagnostic that reflects the degree of CFTR dysfunction [Rosenstein and Cutting, 1998; Wilschanski et al., 1995]. Our algorithm could be modified to predict this spectrum of disease phenotypes, rather than a binary classification of CF causing versus CF neutral. This could be achieved, for example, by fitting predictions (scores) and sweat chloride concentrations to a linear regression and selecting sequences that maximize the correlation coefficient. This type of prediction would be useful because sweat chloride concentration can vary significantly among CF patients, and is proportional to disease severity.

A potential improvement to our algorithm might be the inclusion of structural data. In this study, structural information was only used to identify charged residues that locate in the cell membrane. There is no high-resolution X-ray crystal structure for full-length CFTR, and we considered the possibility that existing low-resolution structural data could confound prediction. However, several groups have acquired high-resolution structural data for CFTR. At least three groups developed high-resolution homology models for human CFTR [Kalid et al., 2010; Mornon et al., 2009; Mornon et al., 2008; Serohijos et al., 2008]; these models are based on X-ray crystal structures of bacterial ABC transporters homologous to CFTR. Considering orientation and dynamics information from CFTR NMR experiments might improve prediction [Baker et al., 2007; Kanelis et al., 2010]. This type of data has the potential to afford annotation of CFTR regions important for dynamic processes such as regulation and allostery. Also, hydrogen-exchange mass spectrometry can highlight protein residues as solvent exposed or buried, and can also report solvent accessibility changes based on external cues such as ligand binding. At least one group acquired this type of data for CFTR [Lewis et al., 2010], which could be used to predict the impact of mutation on structure, and improve our method.

A remaining question is the generalizability of our new method. In this study, we predicted CF from CFTR missense mutation, and the POSE method should currently be considered a contribution to CF diagnostics, and not yet a general predictor of genetic disease. Because the resulting prediction performance was high, our approach might be useful for considering other genes and associated phenotypes. Importantly, phenotypes could be diverse, such as clinical manifestations, catalytic rates, protein stability, and so on. Furthermore, constructing POSEs for a single gene, but multiple phenotypes, might result in POSEs of distinct sequence composition. For instance, sequences that optimize the prediction of disease may differ from those that optimize prediction of protein stability changes. One limitation to algorithmic generalizability is the requirement for training-set mutations, and algorithmic performance is expected to have a dependence on size and quality of the training set employed. The CF-CFTR system provides an excellent model, because the CFTR2 database contains many carefully curated CFTR mutations with corresponding clinical information. Furthermore, CFTR has an extensive and well-characterized evolutionary history, accounting for many orthologous and paralogous sequences that can be used by sequence-based classifiers. The growth of high quality, curated mutation databases such as HGMD

[Stenson et al., 2003] and LSDBs [Cotton et al., 1996] could facilitate algorithmic generalizability to other genes and phenotypes, but rigorous testing on other systems is required confirm this possibility.

Another remaining question regards evolutionary inference from POSE composition. POSEs for several CFTR domains had conspicuous sequence composition (Supp. Table S1). For instance, NBD2 POSEs were dominated by two orthologous sequences (hedgehog and kangaroo rat ABCC7) in each of three independent trials. Similarly, NBD1 POSEs were always composed exclusively of ABCC3 and ABCC6 paralogs, and a single ABCC1 paralog. It is not clear if the observed domain-specific sequence selection has any evolutionary origin. While phylogenetic analysis is beyond the scope of this study, POSEs might provide insight into what stages of evolution were under similar constraints pursuant to a specific phenotype. However, over interpretation is cautioned because POSE sequence selection can arise from nonbiological artifacts. In the case of the CFTR R domain, the algorithm exclusively selected ABCC10 and ABCC11 paralogs. This result is counterintuitive, because CFTR (ABCC7) is the only ABC transporter that has an R domain [Riordan, 2008]. But, combining ABCC10 and ABCC11 sequences results in enough overlap that the resulting alignment spans all positions being predicted in this study. Of alignments sampled, this configuration results in optimized prediction performance for R-domain mutations. This seemingly spurious result may arise from a relatively small training-set for the R domain, which included only a single CF-causing mutation (Supp. Table S4).

The method developed for this study should be an important contribution to the molecular diagnosis of CF. The score function provides a new approach to integrating physiochemical impact of amino acid substitution and the evolutionary distance between a human disease gene and homologous sequences. The POSE construction algorithm is the first supervised learning approach, to our knowledge, to select the most evolutionarily relevant sequences to a disease gene. The method should be further developed and tested using different performance metrics and classification schemes for constructing POSEs. Additionally, an expanded mutation training set could increase performance, and an expanded mutation test set would provide a more conclusive assessment of performance. Demonstrating algorithmic generalizability to other systems is also an important next step. If the performance obtained in this study generalized to other systems, this method would be an excellent contribution to the genetics community at large.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Contract grant sponsors: National Science Foundation (DBI 0845275 to RK); US CF Foundation (CUTTING 09A, SOSNAY10Q0, and KARCH12I0).

References

- Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010; 7:248–249. [PubMed: 20354512]
- Baker JMR, Hudson RP, Kanelis V, Choy WY, Thibodeau PH, Thomas PJ, Forman-Kay JD. CFTR regulatory region interacts with NBD1 predominantly via multiple transient helices. *Nat Struct Mol Biol*. 2007; 14:738–745. [PubMed: 17660831]
- Baldursson O, Ostedgaard LS, Rokhlina T, Cotten JF, Welsh MJ. Cystic fibrosis transmembrane conductance regulator cl. Æ channels with R domain deletions and translocations show phosphorylation-dependent and -independent activity. *J Biol Chem*. 2001; 276:1904–1910. [PubMed: 11038358]
- Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*. 2004; 14:708–715. [PubMed: 15060014]
- Borg Å, Haile RW, Malone KE, Capanu M, Diep A, Törngren T, Teraoka S, Begg CB, Thomas DC, Concannon P, et al. Characterization of BRCA1 and BRCA2 deleterious mutations and variants of unknown clinical significance in unilateral and bilateral breast cancer: the WECARE study. *Hum Mutat*. 2010; 31:E1200–E1240. [PubMed: 20104584]
- Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, Vogelstein B, Karchin R. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res*. 2009; 69:6660–6667. [PubMed: 19654296]
- Castellani C, Cuppens H, Macek M Jr, Cassiman J, Kerem E, Durie P, Tullis E, Assael B, Bombieri C, Brown A. Consensus on the use and interpretation of cystic fibrosis mutation analysis in clinical practice. *J Cystic Fibrosis*. 2008; 7:179–196.
- Chan IS, Ginsburg GS. Personalized Medicine: Progress and Promise. *Annu Rev Genomics Hum Genet*. 2011; 12:217–244. [PubMed: 21721939]
- Chao EC, Velasquez JL, Witherspoon MSL, Rozek LS, Peel D, Ng P, Gruber SB, Watson P, Rennett G, Anton-Culver H, et al. Accurate classification of MLH1/MSH2 missense variants with multivariate analysis of protein polymorphisms–mismatch repair (MAPP-MMR). *Hum Mutat*. 2008; 29:852–860. [PubMed: 18383312]
- Choi MY, Cardarelli L, Therien AG, Deber CM. Non-native interhelical hydrogen bonds in the cystic fibrosis transmembrane conductance regulator domain modulated by polar mutations. Å†. *Biochemistry*. 2004; 43:8077–8083. [PubMed: 15209503]
- Cline MS, Karchin R. Using bioinformatics to predict the functional impact of SNVs. *Bioinformatics*. 2011; 27:441–448. [PubMed: 21159622]
- Collins F. Cystic fibrosis: molecular biology and therapeutic implications. *Science*. 1992; 256:774–779. [PubMed: 1375392]
- Cotton R, Scriver C, McKusick V. Locus-specific mutation databases: A resource. *Genome Digest*. 1996 Jan.6
- Dorfman R, Nalpathamkalam T, Taylor C, Gonska T, Keenan K, Yuan XW, Corey M, Tsui LC, Zielenski J, Durie P. Do common in silico tools predict the clinical consequences of amino-acid substitutions in the CFTR gene. *Clin Genet*. 2010b; 77:464–473. [PubMed: 20059485]
- Durbin RM, Altshuler DL, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Collins FS, Francisco M, Donnelly P, Egholm M. A map of human genome variation from population-scale sequencing. *Nature*. 2010; 467:1061–1073. [PubMed: 20981092]
- Gaucher E, De Kee D, Benner S. Application of DETECTER an evolutionary genomic tool to analyze genetic variation to the cystic fibrosis gene family. *BMC Genomics*. 2006; 7:44. [PubMed: 16522197]
- George Priya Doss C, Rajasekaran R, Sudandiradoss C, Ramanathan K, Purohit R, Sethumadhavan R. A novel computational and structural analysis of nsSNPs in <i>CFTR</i> gene. *Genomic Med*. 2008; 2:23–32. [PubMed: 18716917]
- Hamburg MA, Collins FS. The path to personalized medicine. *N Engl J Med*. 2010; 363:301–304. [PubMed: 20551152]

- Hicks S, Wheeler DA, Plon SE, Kimmel M. Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. *Hum Mutat.* 2011; 32:661–668. [PubMed: 21480434]
- Jordan DM, Ramensky VE, Sunyaev SR. Human allelic variation: perspective from protein function structure, and evolution. *Curr Opin Struct Biol.* 2010; 20:342–350. [PubMed: 20399638]
- Kalid O, Mense M, Fischman S, Shitrit A, Bihler H, Ben-Zeev E, Schutz N, Pedemonte N, Thomas P, Bridges R, et al. Small molecule correctors of F508del-CFTR discovered by structure-based virtual screening. *J Comput-Aided Mol Des.* 2010; 24:971–991. [PubMed: 20976528]
- Kanelis V, Hudson RP, Thibodeau PH, Thomas PJ, Forman-Kay JD. NMR evidence for differential phosphorylation-dependent interactions in WT and [Delta]F508 CFTR. *EMBO J.* 2010; 29:263–277. [PubMed: 19927121]
- Karchin R. Next generation tools for the annotation of human SNPs. *Briefings Bioinformatics.* 2009; 10:35–52.
- Karchin R, Agarwal M, Sali A, Couch F, Beattie MS. Classifying variants of undetermined significance in BRCA2 with protein likelihood ratios. *Cancer Informatics.* 2008; 6:203. [PubMed: 19043619]
- Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, Eswar N, Haussler D, Sali A. LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics.* 2005; 21:2814–2820. [PubMed: 15827081]
- Kraft P, Wacholder S, Cornelis MC, Hu FB, Hayes RB, Thomas G, Hoover R, Hunter DJ, Chanock S. Beyond odds ratios [mdash] communicating disease risk based on genetic profiles. *Nat Rev Genet.* 2009; 10:264–269. [PubMed: 19238176]
- Lewis HA, Wang C, Zhao X, Hamuro Y, Connors K, Kearins MC, Lu F, Sauder JM, Molnar KS, Coales SJ, et al. Structure and dynamics of NBD1 from CFTR characterized using crystallography and hydrogen/deuterium exchange mass spectrometry. *J Mol Biol.* 2010; 396:406–430. [PubMed: 19944699]
- Mi H, Guo N, Kejariwal A, Thomas PD. PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. *Nucleic Acids Res.* 2007; 35:D247–D252. [PubMed: 17130144]
- Mornon JP, Lehn P, Callebaut I. Molecular models of the open and closed states of the whole human CFTR protein. *Cell Mol Life Sci.* 2009; 66:3469–3486. [PubMed: 19707853]
- Mornon JP, Lehn P, Callebaut I. Atomic model of human cystic fibrosis transmembrane conductance regulator: membrane-spanning domains and coupling interfaces. *Cell Mol Life Sci.* 2008; 65:2594–2612. [PubMed: 18597042]
- Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 2003; 31:3812–3814. [PubMed: 12824425]
- Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet.* 2006; 7:61–80. [PubMed: 16824020]
- Park M, Ko SBH, Choi JY, Muallem G, Thomas PJ, Pushkin A, Lee MS, Kim JY, Lee MG, Muallem S, et al. The cystic fibrosis transmembrane conductance regulator interacts with and regulates the activity of the HCO salvage transporter human Na⁺-HCO cotransport isoform 3. *J Biol Chem.* 2002; 277:50503–50509. [PubMed: 12403779]
- Partridge AW, Therien AG, Deber CM. Polar mutations in membrane proteins as a biophysical basis for disease. *Pept Sci.* 2002; 66:350–358.
- Partridge AW, Therien AG, Deber CM. Missense mutations in transmembrane domains of proteins: phenotypic propensity of polar residues for human disease. *Proteins.* 2004; 54:648–656. [PubMed: 14997561]
- Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol.* 2004; 159:882–890. [PubMed: 15105181]
- Potapov V, Cohen M, Schreiber G. Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Protein Eng Des Sel.* 2009; 22:553–560. [PubMed: 19561092]

- Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* 2002; 30:3894–3900. [PubMed: 12202775]
- Riordan JR. CFTR function and prospects for therapy. *Annu RevBiochem.* 2008; 77:701–726.
- Rosenstein BJ, Cutting GR. The diagnosis of cystic fibrosis: a consensus statement. *J Pediatrics.* 1998; 132:589–595.
- Serohijos AWR, Hegedüs Ts, Aleksandrov AA, He L, Cui L, Dokholyan NV, Riordan JR. Phenylalanine-508 mediates a cytoplasmic. Åmembrane domain contact in the CFTR 3D structure crucial to assembly and channel function. *Proc Natl Acad Sci.* 2008; 105:3256–3261. [PubMed: 18305154]
- Sherry S, Ward MH, Kholodov M, Baker J, Phan L, Smigielski E, Sirotkin K. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001; 29:308–311. [PubMed: 11125122]
- Sosnay PR, Castellani C, Corey M, Dorfman R, Zielenski J, Karchin R, Penland CM, Cutting GR. Evaluation of the disease liability of CFTR Variants. *Methods Mol Biol (Clifton, NJ).* 2011; 742:355.
- Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NST, Abeyasinghe S, Krawczak M, Cooper DN. Human Gene Mutation Database (HGMD®): 2003 update. *Hum Mutat.* 2003; 21:577–581. [PubMed: 12754702]
- Thomas PD, Kejariwal A. Coding single-nucleotide polymorphisms associated with complex versus Mendelian disease: evolutionary evidence for differences in molecular effects. *Proc Natl Acad Sci USA U S A.* 2004; 101:15398–15403.
- Vankeerberghen A, Wei L, Jaspers M, Cassiman JJ, Nilius B, Cuppens H. Characterization of 19 disease-associated missense mutations in the regulatory domain of the cystic fibrosis transmembrane conductance regulator. *Hum Mol Genet.* 1998; 7:1761. [PubMed: 9736778]
- Watson MS, Mann MY, Lloyd-Puryear MA, Rinaldo P, Howell RR, Group ACoMGNSE. Newborn Screening: Toward a Uniform Screening Panel and System. ÅExecutive Summary. *Pediatrics.* 2006; 117:S296–S307. [PubMed: 16735256]
- Wilkinson DJ, Strong TV, Mansoura MK, Wood DL, Smith SS, Collins FS, Dawson DC. CFTR activation: additive effects of stimulatory and inhibitory phosphorylation sites in the R domain. *Am J Physiol - Lung Cell Mol Physiol.* 1997; 273:L127–L133.
- Wilschanski M, Zielenski J, Markiewicz D, Tsui LC, Corey M, Levison H, Durie PR. Correlation of sweat chloride concentration with classes of the cystic fibrosis transmembrane conductance regulator gene mutations. *J Pediatr.* 1995; 127:705–710. [PubMed: 7472820]
- Zegarar-Moran O, Monteverde M, Galletta LJ, Moran O. Functional analysis of mutations in the putative binding site for cystic fibrosis transmembrane conductance regulator potentiators. *J Biol Chem.* 2007; 282:9098–9104. [PubMed: 17244607]

Figure 1.

Algorithmic flow chart for one complete execution of the method developed here. Given an input MSA (A) the algorithm selects a random ensemble of up to 150 sequences (B). Next, the algorithm uses the selected random ensemble to score a set of known disease-negative and disease-positive mutations (B). The score function's ability to discriminate disease-positive and disease-negative mutations, using a particular random sequence ensemble, is the prediction performance, and is calculated as the sum of sensitivity (SENS) and specificity (SPEC). The entire loop defined in B is repeated 25,000 times (iterations); every 100 iterations (C), sequences that resulted in the best prediction performance are appended to the input MSA (D). After 25,000 iterations, the 10 sequence ensembles that resulted in the best prediction performance are saved for subsequent blind prediction and analysis, and the corresponding performance metrics and plots are calculated (E). See section Materials and Methods for complete explanation.

Figure 2. Performance and predictions obtained using POSEs. Receiver-operating characteristic analysis (A) and prediction scores for CF-neutral (*CF*⁻) and CF-causing (*CF*⁺) mutations (B). In (A), plot color indicates the performance at different score cutoffs and is defined in the color bar (right-hand side of plot). The area under the ROC curve (AUC) is 0.84.

Table 1
Sensitivity, Specificity, PPV, and NPV Calculated for Different Score Function Sequence Alignment Combinations

Method	Sensitivity	Specificity	PPV	NPV
POSE [†]	74%	94%	89%	85%
SIFT	98%	35%	52%	95%
PolyPhen2	100%	15%	46%	100%
PANTHER [‡]	82%	74%	63%	83%
Orthologs only [†]	84%	48%	54%	81%
Best paralog (ABCC9) [†]	84%	67%	64%	85%
All homologs [†]	65%	92%	85%	79%
SIFT+Hand-curated align.	67%	78%	69%	77%
SIFT+POSE	63%	82%	71%	75%
PolyPhen2 + Hand-curated align.	95%	22%	47%	87%
PolyPhen2 + POSE	86%	47%	54%	82%

[†] Different alignments used with our new score function (see Materials and Methods).

[‡] Values taken from Dorfman et al. (2010), using the same CFTR missense mutations as in this study. Performance for +POSE calculations are for the respective algorithms utilizing a POSE sequence alignment. Performance for +Hand-curated align. are for the respective algorithms utilizing a the hand-curated alignment of Gaucher et al. (2006).