Pathogenicity prediction of non-synonymous single nucleotide variants in dilated cardiomyopathy

Sabine C. Mueller, Christina Backes, Jan Haas, the INHERITANCE Study Group, Hugo A. Katus, Benjamin Meder, Eckart Meese and Andreas Keller

Corresponding author. Sabine C. Mueller, Chair for Clinical Bioinformatics, Saarland University, University Hospital, 66123 Saarbrücken, Germany, Tel.: +49 681 302 68610, Fax: +49 6841 16 26185. E-mail: sabine.mueller@uniklinikum-saarland.de

Abstract

Non-synonymous single nucleotide variants (nsSNVs) in coding DNA regions can result in phenotypic differences between individuals; however, only some nsSNVs are causative for a certain disease. As just a fraction of respective nsSNVs is annotated in databases, computational biology tools are applied to predict the pathogenicity in silico. In addition to applications on oncology, novel molecular diagnostic tests have been developed for cardiovascular disorders as a leading cause of morbidity and mortality in industrialized nations. We explored the concordance and performance of 13 nsSNV pathogenicity prediction tools on panel sequencing results of dilated cardiomyopathy. The analyzed data set from the INHERITANCE study contained 842 nsSNVs discovered in 639 patients, screened for the full sequence of 76 genes related to cardiomyopathies. The single tools prediction revealed a surprisingly high heterogeneity and discordance based on the implemented prediction method. Known disease associations were not reported by the tools, limiting usability in clinics. Because different tools have different advantages, we combined their results. By clustering of correlated methods using similar prediction strategies and calculating a majority vote-based consensus, we found that the prediction accuracy and sensitivity can be further improved. Although challenges remain, different in silico tools bear the potential to predict the malignancy of nsSNVs, especially if different algorithms are combined. Most tools rely mainly on sequence features; beyond these, structural information is important to analyze the relationship of nsSNVs with disease phenotypes. Likewise, current tools consider single nsSNVs, which may, however, show a cumulative effect and turn neutral mutations in an ensemble into pathogenic variants.

Key words: nsSNVs; DCM; pathogenicity prediction; concordance; performance quality
Introduction

Advances in high-throughput DNA sequencing techniques and computational methods have enabled the reliable detection of individual sequence variants in the human genome [1]. Next-generation sequencing (NGS) allows discovery, sequencing and genotyping of hundreds to thousands of genetic variants in different species. The most common type of human genetic variation refers to a single-nucleotide polymorphism (SNP), a position where two alternative bases occur with >1% in the human population [2]. In the context of this study, we refer to single nucleotide variants (SNVs), as the analyzed mutations do not necessarily fit the 1% criterion. Synonymous SNVs that do not change the amino acid sequence because of the degeneracy of the genetic code constitute the majority of genetic variation. From a medical point of view, non-synonymous single nucleotide variants (nsSNVs) in coding DNA can be neutral, associated with a disease by exerting a small effect on a specific trait, or they can be the sole cause for a distinct disease [3]. Single amino acid substitutions can influence the stability of the native protein structure or change catalytic residues and binding properties, probably altering protein function [4]. At present, ~82,176 disease-associated nsSNVs are published [3]. Because of the advent of high-throughput variant detection, however, the amount of identified nsSNVs is growing rapidly. Experimentally gained knowledge is deposited and curated in different databases, e.g. the protein knowledgebase SwissProt [5] and the Single Nucleotide Polymorphism Database (dbSNP) [6], currently the largest database concerning SNV annotation with 29,901,117 deposited SNVs (dbSNP build 138). The most popular database for known pathogenic mutations is the Human Gene Mutation Database (HGMD), a comprehensive repository of mutations associated with human inherited disease [7].

To gain knowledge concerning the pathogenicity of nsSNVs via experimental analysis such as complex association studies is laborious and time-consuming, and often even not possible. This intricate problem requires novel approaches, such as the prediction of the biological impact on a protein’s function in silico [8]. In consequence, many computational methods have been developed over the past decade using different algorithms and features to predict whether an nsSNV is associated with a specific disease. Moreover, genome-wide association studies (GWAS) aim to identify associations of SNVs with phenotypic traits. The published studies and the identified SNVs are collected in the GWAS catalog [9]. Although various tools for predicting the functional significance of nsSNVs are available, evaluating the reliability of prediction results is difficult. On the one hand, some approaches are closely correlated, producing highly similar predictions. On the other hand, the obtained results can vary critically when using different prediction strategies with different advantages. Because of the limitations of available data sets, the majority of methods used similar training data based on SwissProt or neutral pseudo mutations, complicating the validation of prediction results. For validation, however, only nsSNVs already identified and annotated in one of the established databases can be considered, restricting validation strategies.

In consequence, evaluation studies have been performed to compare the available pathogenicity prediction methods and to identify the best method to prioritize nsSNVs as disease cause. In 2010, Thusberg et al. [10] analyzed the performance of nine prediction tools on neutral (from dbSNP) and disease-associated (from PhenCode database and IDbases [11]) variant data sets [12]. Castellana and Mazza further studied the uniformity of the predictions of six methods for whole-exome sequencing data [8]. Moreover, Frousiotis et al. [1] evaluated the prediction performance of nine methods on data from the HGMD and the 1000 Genomes Project Pilot project, and developed a consensus tool integrating four available prediction methods [13].

Our study aims at both, comprehensive evaluation of prediction concordance and prediction quality of existing tools on a high-quality clinical data set. Therefore, we extended the number of tested tools and applied these tools to data with a coverage rate exceeding classical exome capture studies by several orders of magnitude. The underlying approach is sketched in Figure 1. We first report the congruency of common state-of-the-art prediction tools on published NGS data from 639 dilated cardiomyopathy (DCM) patients who have been genetically characterized using NGS [14]. Specifically, we analyzed 842 (339 annotated) nsSNVs in 76 genes relevant for DCM to evaluate the performance and concordance of single nsSNV prediction approaches.

Materials and methods

Data sets

We analyzed a data set containing 842 nsSNVs in 76 genes that are clinically relevant for DCM (known causes and likely candidate genes for DCM) found by studying the genetics in 639 patients with sporadic (51%) or familial (49%) DCM [14]. The sequencing was performed on IlluminaHiSeq instruments. Per patient, roughly 2 billion bases have been sequenced. To ensure diagnostic quality for clinical application, ~99.1% of the targeted genomic region is covered at least 50-fold. In average, each patient carried ~32 nsSNVs in the investigated target region.

In a first step, we collected available information concerning the data set nsSNVs deposited in the databases SwissProt [5], dbSNP [6] and the HGMD [7]. SwissProt provides a collection of human polymorphisms and disease mutations (HUMSAVAR) assigned according to literature reports on probable disease association [15]. Tightly coupled with dbSNP, ClinVar accessions report human variations and interpretations of the relationship of these variations to human health [16]. Entries are labeled according to clinical significance. Moreover, the HGMD collates known (published) gene lesions responsible for human inherited disease. In a second step, we built a test set from the DCM data including only nsSNVs with at least one annotation in dbSNP, SwissProt and/or the HGMD as well as benign or disease-linked information. When information from several sources was available (only 5% have information in all three), we built a majority vote-based consensus. While ~60% are deposited in dbSNP with an rs ID, only 45% have pathogenicity information available. Supplementary File 1 represents the detected pathogenicity annotations in ClinVar, HUMSAVAR and the HGMD for the DCM data set. The neutral labeled set and the disease-associated set comprise 192 and 147 nsSNVs, respectively. This annotated cardio test set, however, refers to only ~45% of our whole data. A total of 55% of the nsSNVs in the DCM data set have no available clinical significance information, and ~40% have neither an rs ID nor other known identifiers and annotations.

Prediction methods

Several methods to assess the effects of amino acid mutations on proteins and their function have been developed in recent years. These approaches rely on physicochemical properties of amino acids and the character of their side chains. Referring to the assumption of survived natural selection, evolutionarily
conserved amino acid positions across protein families are, however, likely to be functionally important. Conservation information usually is obtained from alignments of homologous or somehow related sequences, including position-specific profiles. Some prediction methods also incorporate available annotations, e.g. Gene Ontology (GO) or Pfam [17]. Because protein structure encodes protein function, information concerning the three-dimensional structural environment, such as solvent accessibility, electrostatics and hydrophobicity, is also a crucial criterion to assume a variant's functional impact. Finally, the in silico-derived information about protein structure and function, including essential properties of both the original and substituted residues, is combined into features. These features are used to train a classifier distinguishing between disease-linked and presumably neutral variants. Classification methods usually are machine learning techniques, such as support vector machines, neural networks or random forests, or refer to Bayesian methods, empirically derived rules or mathematical operations. Most techniques were trained on data deposited in databases such as dbSNP or SwissProt, and with artificially constructed test sets. Table 1 summarizes the basic information for each of these tools.

Here, we tested the state-of-the-art tools on the complete data set of 842 nsSNVs to analyze prediction congruency and our generated test set, including 339 nsSNVs to assess prediction quality. For all calculations, the default parameters proposed and preset by each tool were used (see Supplementary Table S1). In addition, we tested SNEffect [18], nsSNP Analyzer [19] and LS-SNP [20], but these prediction methods are not suitable for large-scale studies or require prerequisites that are not generally available, such as a homologous 3D structure or a dbSNP ID for prediction. Interestingly, some prediction methods, such as MutPred [21] and SNPs&GO [22], even incorporate information from other prediction tools within their own pipeline, increasing the already existing correlation between all the available approaches.

### Statistical analysis

The statistical analysis comprised two components, the prediction congruency analysis and the prediction quality analysis of the 13 state-of-the-art pathogenicity prediction tools. In a first step, we calculated for all 842 nsSNVs in our data set the pathogenicity prediction of the 13 tools. There are a dozen computational tools aimed at functional prediction of nsSNVs; thus, there are also approaches trying to build a unified consensus classification score from them [23, 24]. To avoid adding to the complexity, we evaluated the straightforward majority vote to build a consensus: for each nsSNV, we determined the most frequent prediction result among the single predicted ones. To evaluate the concordance of prediction, we computed for each single nsSNV a consensus prediction out of all 13 prediction results. For each pair of prediction methods, we calculated the similarity score as follows:

\[
\text{Similarity} = \frac{\sum_{n: \text{comparison}(n)} 1}{\text{number of comparisons}}
\]

where the comparison of an nsSNV is 1 if the compared prediction results are equal and −1 if the results contradict each other. Via Cytoscape [25], we built a network with prediction methods referring to nodes and their pair-wise similarity scores defining the edges. For clarity, only edges marking at least 70% similarity...
<table>
<thead>
<tr>
<th>Prediction method</th>
<th>Input</th>
<th>Classifier</th>
<th>Evolutionary analysis</th>
<th>Structural attributes</th>
<th>Annotations</th>
<th>Running modality</th>
</tr>
</thead>
<tbody>
<tr>
<td>MutPred</td>
<td>Fasta sequence</td>
<td>Random forest</td>
<td>SIFT, Pfam, PSI-BLAST</td>
<td>Predictions of: secondary structure, solvent accessibility, transmembrane helices, stability, etc.</td>
<td>/</td>
<td>Web server</td>
</tr>
<tr>
<td>PMut</td>
<td>Fasta sequence, UniProt ID</td>
<td>Neural network</td>
<td>PSI-BLAST, multiple sequence alignments (MSA), Pfam</td>
<td>Homolog mapping/predictions</td>
<td>/</td>
<td>Web server</td>
</tr>
<tr>
<td>PROVEAN</td>
<td>Ensembl ID, NCBI RefSeq ID, UniProt ID</td>
<td>Alignment scores</td>
<td>BLAST</td>
<td>/</td>
<td>/</td>
<td>Web server, stand-alone</td>
</tr>
<tr>
<td>SNPs&amp;GO</td>
<td>UniProt ID</td>
<td>Support vector machines</td>
<td>Sequence environment, sequence profiles, PANTHER</td>
<td>/</td>
<td>GO</td>
<td>Web server</td>
</tr>
<tr>
<td>SNAP</td>
<td>Fasta sequence</td>
<td>Neural network</td>
<td>PSI-BLAST, position-specific independent counts (PSIC) profiles, Pfam</td>
<td>Predictions: secondary structure, solvent accessibility, chain flexibility</td>
<td>SwissProt</td>
<td>Web server</td>
</tr>
<tr>
<td>SIFT</td>
<td>Comma separated: chromosome, coordinate, orientation, alleles</td>
<td>Alignment scores</td>
<td>MSA</td>
<td>/</td>
<td>/</td>
<td>Web server, stand-alone</td>
</tr>
<tr>
<td>PANTHER</td>
<td>Fasta sequence</td>
<td>Alignment scores</td>
<td>PANTHER library, hidden Markov Models</td>
<td>/</td>
<td>GO</td>
<td>Web server, stand-alone</td>
</tr>
<tr>
<td>PhD-SNP</td>
<td>Fasta sequence</td>
<td>Support vector machines</td>
<td>Sequence environment, sequence profiles, MSA</td>
<td>/</td>
<td>/</td>
<td>Web server, stand-alone</td>
</tr>
<tr>
<td>SNPs3D</td>
<td>SNP ID, sequence ID</td>
<td>Support vector machines</td>
<td>PSI-BLAST, position-specific scoring matrix, MSA</td>
<td>Structure stability model (solvent accessibility, electrostatics, hydrophobicity,…)</td>
<td></td>
<td>Web server</td>
</tr>
<tr>
<td>PolyPhen2 (PPh2)</td>
<td>Fasta sequence, SNP ID, UniProt ID</td>
<td>Bayesian classification</td>
<td>PSIC profiles</td>
<td>Homolog mapping/predictions</td>
<td>Pfam</td>
<td>Web server, stand-alone</td>
</tr>
<tr>
<td>MutationAssessor</td>
<td>UniProt ID</td>
<td>Alignment scores</td>
<td>MSA</td>
<td>Predictions: secondary structure, solvent accessibility, chain flexibility (SNAP, nsSNPAnalyzer), homolog mapping/predictions (PolyPhen2)</td>
<td>/</td>
<td>Web server</td>
</tr>
<tr>
<td>PredictSNP</td>
<td>Fasta sequence</td>
<td>Confidence-based random forest consensus</td>
<td>MAPP, nsSNPAnalyzer, PANTHER, PhD-SNP, PolyPhen, PolyPhen2, SIFT, SNAP</td>
<td>/</td>
<td>SwissProt (SNAP), GO (PANTHER), Pfam (PolyPhen2)</td>
<td>Web server, batch script available with many dependencies</td>
</tr>
<tr>
<td>Condel</td>
<td>Uniprot ID, chromosome + start + end + mutant</td>
<td>Weighted average score</td>
<td>SIFT, PolyPhen2, MutationAssessor</td>
<td>Homolog mapping/predictions (PolyPhen2)</td>
<td>Pfam (PolyPhen2)</td>
<td>Web server</td>
</tr>
</tbody>
</table>
Table 2. AUC values and confidence intervals

<table>
<thead>
<tr>
<th>Prediction tool</th>
<th>AUC (%)</th>
<th>Confidence interval (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMut</td>
<td>54.9</td>
<td>48.3–61.6</td>
</tr>
<tr>
<td>MutPred</td>
<td>79</td>
<td>73.6–84.3</td>
</tr>
<tr>
<td>SNAP</td>
<td>51.2</td>
<td>44.6–57.8</td>
</tr>
<tr>
<td>SIFT</td>
<td>58.9</td>
<td>52.3–65.4</td>
</tr>
<tr>
<td>SNPs&amp;GO</td>
<td>63.1</td>
<td>56.8–69.4</td>
</tr>
<tr>
<td>PhD-SNP</td>
<td>54.9</td>
<td>48.3–61.5</td>
</tr>
<tr>
<td>PROVEAN</td>
<td>64.8</td>
<td>58.5–71.2</td>
</tr>
</tbody>
</table>

AUC values and confidence intervals were computed only for tools with numerical output and <50 NA values. Confidence intervals were calculated using \( n = 1000 \) bootstrap samples. For better comparison, only those 292 nsSNVs, where all seven tools returned an output, were considered.

To make the ROC analysis better comparable between the different tools, we used just those 292 nsSNVs from the test set, where all seven tools were able to return an output. Thereby, we minimize the risk of potential bias toward variants, which were hard to predict.

Results
Concordance of prediction methods

To evaluate the concordance of the 13 prediction methods, we determined the distribution of obtained prediction results on the complete data set of 842 nsSNVs (Figure 2). Because no pathogenicity information for the complete data set is available, we first focused on the analysis of congruency and association of the state-of-the-art methods. We studied both, the overall concordance of one prediction tool compared with the consensus of all prediction methods and the mutual agreement among all methods. Although SIFT and MutationAssessor [31] predicted ~50% of the 842 nsSNVs to be disease-associated, the other methods proposed the majority of the nsSNPs to be neutral. SNPs3D [32] (90%), PANTHER [33] (35%), MutationAssessor (29%) and PolyPhen2 [34] (23%) failed to predict all of the 842 nsSNVs. In contrast, MutPred, SNPs&GO, PhD-SNP, SNAP, PMut, PredictSNP [24] and PROVEAN had a prediction failure rate of <4%. Figure 3 illustrates the comparison of each prediction tool to the consensus prediction result built using all methods. PredictSNP, PROVEAN, PhD-SNP, SNPs&GO and Condel achieved the best conformity to the overall consensus.

To evaluate the mutual agreement among the tested methods, we created a network based on pair-wise comparison of similarity scores of the prediction results (Figure 4). SNPs3D, PANTHER, MutationAssessor and PolyPhen2 showed the worst conformity with all other tested tools. In contrast, PredictSNP, PROVEAN, MutPred, SNAP, PMut, SNPs&GO and PhD-SNP had the best concordance values with >80%. Except PROVEAN, these tools use machine learning classifiers for pathogenicity prediction, and at least four include structural annotations. Condel and PredictSNP build consensus predictions based on the single predictions of other methods. Interestingly, machine learning-based applications cluster well, indicating the chosen classifier method to be essential for the prediction outcome (see Figure 5).

The underlying classification method even reveals greater influence on the overall concordance than tools integrating further prediction methods. Moreover, tools including others do not necessarily show equal performance. These findings agree with previous studies [12].

Performance of prediction methods

Based on the generated test set of 339 nsSNVs with available pathogenicity annotations, we calculated the statistical measures, accuracy, specificity, sensitivity, balanced accuracy and MCC to evaluate the prediction performance of the 13 published tools included in our study and computed a majority vote-based consensus prediction (Figure 6, detailed values in Supplementary Table S2). The best performance concerning balanced accuracy and sensitivity in combination with prediction ability (NA ratio) was reached by MutPred with 66% accuracy, 96% specificity, 28% sensitivity, 62% balanced accuracy and 0.32 MCC. Despite promising balanced accuracy values, SNPs3D could classify only ~5% of our generated test set. The remaining tools frequently showed a low hit ratio. In addition to SNPs3D and MutationAssessor, most methods revealed much higher

Calculating receiver operator characteristics curves and area under curve values

For 10 tools, we obtained a numerical output that generally allows for calculating receiver operator characteristics (ROC) curves beyond the classification accuracy, specificity, sensitivity, balanced accuracy and MCC. We excluded those tools that showed substantially many NA values (only tools that showed lesser than 50 NA values were kept). For the following seven tools, area under curve (AUC) values were calculated along with confidence intervals: PMut [26], SNAP [27], MutPred, PROVEAN [28], SIFT [29], PhD-SNP [30] and SNPs&GO (Table 2). ROC curves, AUC values and confidence intervals have been determined by functions relying on the pROC package. For the confidence intervals, \( n = 1000 \) bootstrap samples have been calculated.

or connecting methods that incorporate other tools were recognized.

To measure the quality of prediction, we tested the 13 prediction tools on our generated test set of 339 nsSNVs with available pathogenicity annotations. We calculated the confusion matrices \([\text{true positives (TP)}, \text{true negatives (TN)}, \text{false positives (FP)}, \text{false negatives (FN)}]\) and consequently accuracy, specificity and sensitivity for the results of each single prediction tool. We also calculated the balanced accuracy and Matthews correlation coefficient (MCC) because the distribution of available annotations concerning neutral and disease-associated nsSNVs is imbalanced.

\[
\begin{align*}
\text{accuracy} & = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \\
\text{specificity} & = \frac{\text{TN}}{\text{FP} + \text{TN}} \\
\text{sensitivity} & = \frac{\text{TP}}{\text{TP} + \text{FN}} \\
\text{balanced accuracy} & = \frac{0.5 \times \text{TP} + 0.5 \times \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \\
\text{MCC} & = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FN})(\text{TP} + \text{FP})(\text{TN} + \text{FN})(\text{TN} + \text{FP})}}
\end{align*}
\]

Furthermore, we determined the performance quality of the consensus prediction result, computed for each nsSNV during congruency analysis. Cases with a balanced 50% vote were excluded within calculation of prediction measures. Referring to the methods’ close correlations, we also clustered related methods and calculated a kind of nested majority vote. We first determined the majority vote for related methods and then determined the majority vote over all these sub-majority votes.
Figure 2. Distribution of obtained prediction results for the complete data set.

Figure 3. Comparison of single prediction with consensus prediction results.

Figure 4. Network of prediction concordance. Nodes represent prediction tools whereas edges mark the pair-wise prediction similarity for connected nodes. The dashed edges refer to a similarity of ~70–80% and the bold line edges to a similarity of >80%. For clarity, edges with less similarity values are neglected, except if one method includes another (pointed edges). Pink arrow edges point to included tools. The color of the nodes codes for concordance: purple = high, turquoise = moderate and blue = low. A colour version of this figure is available at BIB online: http://bib.oxfordjournals.org.
specificity values compared with sensitivity. The mean values for accuracy (60%), specificity (69%), sensitivity (49%), balanced accuracy (59%) and MCC (0.20) show the current limitations of nsSNV pathogenicity prediction.

In addition, we calculated ROC curves and AUC values including confidence intervals for the seven tools with numerical output and <50 NA values (Supplementary File 2). The used data set comprises only the 292 nsSNVs from the test set, where all seven tools were able to return an output. The tools and the corresponding values are listed in Table 2. Besides (balanced) accuracy, MutPred also received the best AUC value of 79% (Figure 7). Moreover, SIFT (59%), SNPs&GO (63%) and PROVEAN (65%) obtain significantly better AUC values compared with balanced accuracy. Interestingly, MutPred uses SIFT for defining evolutionary attributes.

Because close correlations between some tools exist, we clustered related methods to build a consensus prediction of non-overlapping tools, as well as comparing structure-based and sequence-based methods to try to improve the prediction results. We discriminated between the structure-based group (MutPred, PMut, SNAP, SNPs3D and PolyPhen2) and the sequence-based group (PROVEAN, SNPs&GO, SIFT, PANTHER, PhD-SNP and MutationAssessor). We also clustered related methods, namely, methods using the same classification method or prediction features: machine learning-based group (MutPred, SNPs&GO, SNPs3D, PhD-SNP, SNAP and PMut) and the non-machine learning-based group (PANTHER, PROVEAN, Condel, PolyPhen2, SIFT and MutationAssessor). As PredictSNP builds a consensus of sequence- and structure-based methods as well as incorporates machine learning and non-machine
learning-based tools, we excluded this method from our consensus calculation. Table 3 depicts the resulting quality measures. In Figure 8, all prediction results including single tool predictions as well as consensus results are gathered. The consensus predictions of clustered and structure information including methods reveal slightly improved accuracy values. We also built a consensus of the results obtained by structure-based and sequence-based methods, which yielded the best consensus prediction result with 65% accuracy (balanced accuracy 63%) and ~63% sensitivity. In addition, the sequence structure consensus and the clustered consensus were able to return a result for each sample in the data set.

**Discussion**

In this study, we analyzed the performance and concordance of state-of-the-art nsSNV prediction tools on a real-life high-quality data set. The vast amount of genetic variant data requires development of automatic procedures to predict the functional and phenotypic effects of nsSNVs. In recent years, a number of approaches dealing with the functional impact of genetic variants on protein function have been developed. However, one realizes that many existing prediction methods are not well suited for large-scale studies with real-life data. Often only server-based application is possible, and/or only input of single sequences is allowed. For example, MutPred, which marginally outperforms all other approaches concerning balanced accuracy (62%), AUC (79%) and prediction ability (only in ~3% of the cases no prediction was possible), allows only mutation submission of one single sequence in a server query. Some methods even require rs IDs (SNPs3D) or similar annotations, which are in most cases not available for a huge part of real-life data. Even when rs IDs were available, SNPs3D, MutationAssessor and PolyPhen2 failed to predict in 20–95% of all cases with available annotations. Our data set of 842 nsSNVs in 76 DCM genes contained 45% pathogenically annotated nsSNVs with rs IDs, and thus, methods such as SNPs3D requiring an rs ID are not suitable for the prediction of the remaining 55%.

Moreover, the prediction result seems to critically depend on the chosen method. Approaches based on machine learning techniques clustered together very well, independent of whether a method included structure information. The underlying classification technique even reveals greater influence on the overall prediction concordance than tools incorporating further prediction methods. Different prediction tools yielded different, even conflicting, results. SNPs3D, PANTHER, MutationAssessor and PolyPhen2 showed worse concordance with the remaining state-of-the-art methods. A combination of different prediction tools, however, is laborious because of incompatible input formats; furthermore, the obtained results cannot be parsed automatically because the results are in most cases not textual. In addition, methods such as MutationAssessor, Condel, SNPs3D, SIFT and
SNPs&GO are restricted to ambiguous input formats such as the Uniprot ID. They do not allow a differentiation of different transcripts. In consequence, a compatible, generalized format for nsSNVs simplifying the combination and automation of several prediction methods, as well as using an unambiguous input such as fasta sequences, is required.

We further tested a straightforward majority vote based on clustering related methods to analyze whether a combination of different prediction tools can improve performance quality. The most promising results of 64.43% balanced accuracy were obtained with a consensus of structure-based majority and sequence-based majority results. In general, clustering correlated prediction methods improved accuracy and sensitivity values compared with single prediction results and returned a prediction for every sample in the data set. The mean accuracy of all tested tools was ~60% (balanced accuracy 59%) and the mean sensitivity ~50%, whereas the clustering of related methods resulted in 65% accuracy (64% balanced accuracy) and 63% sensitivity. Most methods, except SNPs3D and MutationAssessor, yielded much higher specificity values compared with sensitivity. Depending on the clinical application scenario, higher sensitivity values can be more important than higher specificity, because classification of a neutral nsSNV as disease-associated may be less problematic from a clinical point of view than labeling a disease nsSNV as neutral.

Table 3. Prediction results of clustered prediction methods

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Specificity</th>
<th>Sensitivity</th>
<th>Balanced</th>
<th>MCC</th>
<th>NA accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence consens</td>
<td>57.7</td>
<td>63.44</td>
<td>50.34</td>
<td>56.89</td>
<td>0.14</td>
<td>2.4</td>
</tr>
<tr>
<td>Structure consens</td>
<td>62.73</td>
<td>76.24</td>
<td>45.39</td>
<td>60.82</td>
<td>0.23</td>
<td>5.0</td>
</tr>
<tr>
<td>Sequence structure consens</td>
<td>65.0</td>
<td>66.0</td>
<td>62.86</td>
<td>64.43</td>
<td>0.27</td>
<td>0</td>
</tr>
<tr>
<td>Clustered consens</td>
<td>63.74</td>
<td>64.43</td>
<td>61.76</td>
<td>63.1</td>
<td>0.24</td>
<td>0</td>
</tr>
</tbody>
</table>

Sequence consens: consensus prediction of sequence-based group (PROVEAN, SNPs&GO, SIFT, PANTHER, PhD-SNP and MutationAssessor); structure consens: consensus prediction of structure-based group (MutPred, PMut, SNAP, SNPs3D and PolyPhen2); sequence structure consens: consensus from sequence and structure consens; clustered consens: consensus prediction of related methods into machine learning group (MutPred, PMut, SNPs&GO, SNPs3D, SNAP and PhD-SNP); and non-machine learning group (PROVEAN, PANTHER, PolyPhen2, SIFT, Condel and MutationAssessor).

Incorporating the numerical output, where available, in future enhanced statistical analysis such as different classification and subset selection techniques may however further improve the performance quality measures—in particular the sensitivity—of the prediction tools. The calculated AUC values represent the potential of a more elaborate statistical analysis. The straightforward consensus calculation, which is based on an intuitive majority vote, is applied just to provide evidence that consensus approaches can potentially further add to a diagnosis and especially to the sensitivity of the prediction. Hence, respective tools can improve the clinical value. Optimization and validation of the consensus calculation in future studies may also result in further improvement of pathogenicity prediction. However, in such cases, more complex models have to be trained and tested on different data sets to minimize the common risk of overtraining of the machine learning approaches.

In addition, a major problem concerning all computational methods and databases is the maintenance of the developed software. Rare updates lead to obsolete annotation linkages and can even negatively influence classification results. In fact, some of the available supposedly neutral nsSNV data sets used in former studies contain disease-associated mutations according to actual database entries. We identified, for example, some variants in the neutral VariBench data set of Thusberg et al. [12] as disease-associated, with entries in the HGMD. The major problem in general refers to the limited availability of suitable data and, particularly, high-quality data. Often data sets are constructed from information contained in one particular database. Frousios et al. [11] constructed their data sets based on information of the 1000 genomes pilot project and the HGMD without cross-checking these information in additional databases such as e.g. dbSNP or SwissProt. However, some information missing in one database might be available in another. Sometimes, even annotations from different data sources disagree. In particular, the construction of a neutral labeled data set is highly challenging. Hence, we constructed our positive and negative data sets on a consensus out of the three most popular databases for nsSNVs. Based on our findings, we recommend using a consensus of different available data sources to avoid biased data sets in future studies.

Besides pathogenicity annotation, the quality of the used data in evaluation studies is often highly varying. Frousios et al.

Figure 8. Prediction results of the 13 state-of-the-art tools and the applied consensus calculations.
The study of nsSNVs as genetic factors in human diseases, their pathological effect on protein function and consequently the resulting phenotypes is essential in human health care. In our study, we provide an overview of the performance and concordance of state-of-the-art pathogenicity prediction tools on a real data set of DCM patients. Different prediction methods, especially if combined, reveal the potential to predict the malignancy of nsSNVs. Capturing diagnostic requirements and improving clinical treatment, however, requires further optimization. Although several mutations in one patient can influence pathogenicity and thus the diagnosis, the tested approaches predict single nsSNVs without consideration of neighboring nsSNVs and mutual effects. Rule-based statistical or structure-based methods to identify and analyze dependencies of nsSNVs to improve computational diagnostics may further increase the accuracy of diagnostic and prognostic approaches to human diseases.

Conclusion

The study of nsSNVs as genetic factors in human diseases, their pathological effect on protein function and consequently the resulting phenotypes is essential in human health care. In our study, we provide an overview of the performance and concordance of state-of-the-art pathogenicity prediction tools on a real data set of DCM patients. Different prediction methods, especially if combined, reveal the potential to predict the malignancy of nsSNVs. Capturing diagnostic requirements and improving clinical treatment, however, requires further optimization. Although several mutations in one patient can influence pathogenicity and thus the diagnosis, the tested approaches predict single nsSNVs without consideration of neighboring nsSNVs and mutual effects. Rule-based statistical or structure-based methods to identify and analyze dependencies of nsSNVs to improve computational diagnostics may further increase the accuracy of diagnostic and prognostic approaches to human diseases.

Acknowledgments

This work was supported by the Best Ageing grant 306031 from the European Union. The manuscript was revised by the professional language polishing service San Francisco Edit.

Funding

This work was funded by the Best Ageing grant 306031 from the European Union.

References


