SNPdryad: predicting deleterious non-synonymous human SNPs using only orthologous protein sequences

Ka-Chun Wong\(^1,2\) and Zhaolei Zhang\(^1,2,3,4,\)*

\(^1\)Department of Computer Science, University of Toronto, Toronto, Ontario, Canada M5S 3G4 \(^2\)The Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, Canada M5S 3E1, \(^3\)Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario, Canada M5S 3E1 and \(^4\)Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada M5S 1A8

Associate Editor: Alfonso Valencia

**ABSTRACT**

**Motivation:** The recent advances in genome sequencing have revealed an abundance of non-synonymous polymorphisms among human individuals; subsequently, it is of immense interest and importance to predict whether such substitutions are functional neutral or have deleterious effects. The accuracy of such prediction algorithms depends on the quality of the multiple-sequence alignment, which is used to infer how an amino acid substitution is tolerated at a given position. Because of the scarcity of orthologous protein sequences in the past, the existing prediction algorithms all include sequences of protein paralogs in the alignment, which can dilute the conservation signal and affect prediction accuracy. However, we believe that, with the sequencing of a large number of mammalian genomes, it is now feasible to include only protein orthologs in the alignment and improve the prediction performance.

**Results:** We have developed a novel prediction algorithm, named SNPdryad, which only includes protein orthologs in building a multiple sequence alignment. Among many other innovations, SNPdryad uses different conservation scoring schemes and uses Random Forest as a classifier. We have tested SNPdryad on several datasets. We found that SNPdryad consistently outperformed other methods in several performance metrics, which is attributed to the exclusion of paralogous sequence. We have run SNPdryad on the complete human proteome, generating prediction scores for all the possible amino acid substitutions.

**Availability and implementation:** The algorithm and the prediction results can be accessed from the Web site: http://snps.ccb.r.utoronto.ca:8080/SNPdryad/.

**Contact:** Zhaolei.Zhang@utoronto.ca

**Supplementary information:** Supplementary data are available at Bioinformatics online.

Received on August 6, 2013; revised on November 7, 2013; accepted on December 13, 2013

**1 INTRODUCTION**

Single-nucleotide polymorphisms (SNPs) are single nucleotide variations between different individuals of the same species. They account for the majority of the genetic variations among human individuals, as it is estimated that a single SNP is present in every 2000 nt between any two individuals (Altshuler et al., 2010; Stranger et al., 2007). Depending on how the amino acid is affected by the polymorphism, the SNPs in the protein coding regions can be classified into synonymous SNP (those that do not change amino acid), non-synonymous SNP (nsSNP) (those change the amino acid) and nonsense mutations (where a SNP results in a stop codon). Based on the data from the 1000 Genomes Project, it is estimated that on an average a human individual possesses 10,000–11,000 non-synonymous substitutions compared with the reference human genome sequence (Abecasis et al., 2010). A number of databases had been developed to curate and store these human SNP data, which include dbSNP, OMIM, SNPdb and dbNSFP (Amberger et al., 2009; Liu et al., 2013; Schaefer et al., 2012; Sherry et al., 2001).

In addition to SNPs found in the protein coding regions, recent genome-wide association and expression Quantitative Trait Loci (eQTL) studies and the large-scale ENCODE project also revealed many SNPs located outside of the protein coding regions (called regulatory SNPs or rSNPs) that are also implicated in human diseases. However, these rSNPs are not discussed here, as we only focus on the effect of SNPs found in the protein coding regions. Synonymous SNPs are usually assumed to be functional neutral because they do not change the protein sequence, but in rare cases they can affect protein folding, disrupt RNA secondary structure or disrupt miRNA binding sites (Johnson et al., 2011; Kimchi-Sarfaty et al., 2007; Lin et al., 2011; Shabalina et al., 2013). Although many nsSNPs are probably selectively neutral, i.e. having little functional effects, a substantial fraction of these nsSNPs are indeed predicted to be deleterious because they can potentially disrupt functional sites on a protein or affect their correct folding (Lohmueller et al., 2008). Many nsSNPs are also linked to human disorders; many of these disease associated SNPs are documented in databases, such as OMIM, pharmGKB and HGMD (Adzhubei et al., 2010; Amberger et al., 2009; Sunyaev et al., 2001). Because of the potential functional consequences of nsSNPs, several computational methods had been developed to in silico predict whether an nsSNP is deleterious. Some of these methods include SIFT, PolyPhen, PolyPhen2, SNPs3D, SNAP and MutationTaster (Adzhubei et al., 2010; Bromberg et al., 2008; Ng and Henikoff, 2003; Schwarz et al., 2010; Sunyaev et al., 2001; Yue et al., 2006). These methods usually work by estimating the likelihood that a mutation (nsSNP) is tolerated based on whether the
Amino acid residue is observed in other evolutionarily related orthologous or paralogous protein sequences or sequence fragments, and whether the mutation is tolerated based on protein structure and the physicochemical properties of the amino acids. The major differences among these methods are how these evolutionary and structural features are extracted, and what algorithm (classifier) is used in combining these features to make a decision. There also exist ensemble methods, such as Condel and Logit, which use Naive Bayes and Logistic Regression, respectively, to combine individual prediction methods (Gonzalez-Perez and Lopez-Bigas, 2011; Li et al., 2013).

As described above, an accurate and unbiased multiple sequence alignment of a protein of interest with its orthologous and paralogous sequences is essential to derive a conservation profile of the protein, which can be used to estimate how a mutation is tolerated (Hicks et al., 2011). Ideally, only orthologous sequences should be used in this step because these orthologs are expected to perform similar function in related organisms, and the corresponding amino acid position is expected to have the same evolutionary, biophysical and structural constraint. However, to the best of our knowledge, all the current nsSNP analysis algorithms include paralogous sequences in the multiple sequence alignment (MSA) step, which perhaps was a necessity a few years ago because of the scarcity of the fully sequenced proteome sequences. However, inclusion of paralogous sequences can potentially introduce noises in generating protein sequence conservation profiles because when compared with orthologs, protein paralogs are more likely to diverge in sequence and in cellular functions. On an average, the amino acid sequence identity between paralogous protein pairs is only 30% (Axelsen et al., 2007).

The recent breakthroughs in sequencing technology have generated fully sequenced genomes and proteomes for a large number of vertebrates, which potentially can eliminate the need for including paralogous protein sequences in building multiple sequence alignment. In this article, we show that, by including only orthologous protein sequences, we achieved better performance in predicting deleterious nsSNPs. Other innovations in our method include the choice of using Random Forest in classification, which had been previously shown to be effective in high-dimension data classification (Caruana et al., 2008). We benchmarked our prediction method, termed SNPdryad, against other methods and showed that SNPdryad consistently achieved better sensitivity and specificity on the datasets tested.

2 METHODS

2.1 Overview of the SNPdryad algorithm

Figure 1 summarizes the overall design of SNPdryad. The input of SNPdryad is a non-synonymous human SNP and the sequence of the human protein that the SNP is on; the output is the predicted deleterious score for the input nsSNP. The higher the score, the more deleterious the input nsSNP is predicted to be.

2.2 Collecting orthologous protein sequences

We ran the Inparanoid program to obtain orthologous protein sequences from other organisms (Ostlund et al., 2010); the human proteins were extracted from Ensembl (GRCh37). To ensure the quality of the final multiple sequence alignment, we limited our ortholog search to only mammalian species [Pongo pygmaeus abelli (PPYG2), Mus musculus (NCBI37), Macaca mulatta (MMUL_1), Canis familiaris (BROAD22), Equus caballus (EquCab2), Rattus norvegicus (RGSC3.4), Cavia porcellus (cavPor3), Bos taurus (UMD3.1) and Monodelphis domestica (BROADO5)]. Such an approach has been proven to be successful in detecting conserved regulatory elements (Xie et al., 2005).

2.3 Generating sequence alignment profile

We used the MUSCLE software for multiple-sequence alignment, using the default parameter setting (Edgar, 2004). Other alignment tools such as MAFFT and Clustal were also tested; MUSCLE was picked because of its alignment accuracy and speed (Edgar, 2004). Given a multiple sequence alignment, the next step is to derive a positional-specific conservation profile and use it to estimate the tolerance toward mutations at each position. Two important calibrations are needed at this step; how to weight each input protein sequence, and how to score the amino acid substitutions. Kumar and colleagues previously showed that the choice of sequence-weighting scheme and substitution-scoring scheme are crucial at this step in distinguishing the deleterious nsSNPs from the rest (Kumar et al., 2009); therefore, we explored four different sequence-weighting schemes and two different substitution-scoring schemes in our work (a total of eight distinct ways of measuring conservation scores). All of the eight conservation scores were input into SNPdryad as features and were automatically weighted and selected by Random Forest classifier to achieve the best performance.

2.4 Constructing feature vectors

2.4.1 Sequence-weighting schemes

(i) In the first sequence-weighting scheme, each individual input protein sequence is treated equally and given the same weight. (ii) In the second sequence-weighting scheme, a weight is calculated for each input protein sequence based on its overall sequence similarity to the human protein sequence. In particular, the sum of pair-wise BLOSUM62 alignment score is computed as the weight to quantify the pair-wise sequence similarity. The higher the sequence similarity (sum of BLOSUM62 scores), the higher weight is given to the input sequence. Such a weighting scheme is different from the ones used in PolyPhen2, SIFT and other prediction methods, which in contrast give low weight to the sequence that is highly similar to the human sequence. The rationale in their approaches is that such a low weight can eliminate redundant or highly similar sequences when searching for homologous sequences in a large sequence database. In contrast, SNPdryad does not have such drawbacks because only orthologous protein sequences are allowed in the alignment. SNPdryad can confidently assign a high
weight to a similar orthologous sequence that is believed to be function-
ally consistent with the human sequence. (iii) In the third weighting
scheme, each input sequence is weighted according to its evolutionary
distance from human. We first ran the PhyML program on the aligned
sequences to build a phylogenetic tree (Guindon et al., 2010); each se-
quence is then assigned the weight of 1 minus the additive branch length
to human sequence. (iv) In the fourth approach, we adopted the same
scheme as in the weighting method used in the ‘evolutionary trace’
(Mihalek et al., 2004). This method takes into account the sequence con-
version at multiple levels of the phylogenetic tree constructed and
assigns weights globally.

2.4.2 Substitution-scoring schemes We tested two substitution-
scoring schemes in estimating the conservation level at each amino acid
position: information entropy (Shannon Entropy) and simple
BLOSUM62 scores. The former is an unbiased measure of amino acid
conservation, whereas the latter can take into account the biophysical
properties of each amino acid. In the end, by combining four different
sequence-weighting schemes and two different substitution-scoring
schemes, we calculated eight different conservation scores for each
amino acid at each position that has an nsSNP in the human protein
sequence. Next, we calculated the differences in conservation score be-
tween the reference allele and the variant allele and incorporated it as a
feature into the classification model.

Overall Conservation Statistics: Besides the conservation score calcu-
lated at the nsSNP-containing column in the multiple-sequence alignment,
the mean and standard deviation of the entropy of the alignment is also
calculated. Z-score can thus be calculated for the nsSNP-containing
column.

2.4.3 Physiochemical properties of amino acids In addition to
sequence conservation profile, we also included the physiochemical prop-
erties of the amino acids as a feature in our classification scheme. The
following properties are included: hydropathy index (Kyte and Doolittle,
1982), polarity (Cooper and Hausman, 2007), mass (Reichert and Suhnel,
2002), volume (Zamyatnin, 1972), surface area (Chothia, 1976), residue
non-polar surface area (Karplus, 1997), estimated hydrophobic effects for
residue burial and side chain (Karplus, 1997), population percentage of
being exposed in solvent, being buried in solvent and being neither
exposed nor buried in solvent (Bordo and Argos, 1991). We note that
these features are not independent from each other.

2.4.4 Other features To alleviate the noise and fluctuation in the
data, the number of protein sequences included in the multiple-sequence
alignment and the number of distinct amino acid residues in the nsSNP-
containing column are also used as features in the classifier. The rati-
nale for such treatment is that the conservation scores derived from alignment
of higher number of sequences are deemed more reliable than scores derived
from fewer sequences. In addition, functional annotation of the
region on the protein where the SNP is present is also important. Based
on such a rationale, we have included the presence of the annotations
from PFAM, SUPERFAMILY and PROSITE as features for predic-
tions (Hulo et al., 2006; Punta et al., 2012; Wilson et al., 2009). The complete list of features used is listed in Supplementary Table S1.

2.5 Classification methods
We evaluated 10 leading classification methods, applied them onto the
aforementioned features and benchmarked their performance following a
standard 10-fold cross-validation procedure. These methods include Random Forest (Breiman, 2001), Naïve Bayes (John and Langley,
1995), Bayes Network (Cooper and Herskovits, 1992), Multilayer
Perceptron (Bishop, 1995), AdaBoost (Freund and Schapire, 1996),
Support Vector Machine using Polynomial Kernel (Burges, 1998),
Support Vector Machine using Radial Basis Kernel (Burges, 1998) and
three k-Nearest Neighbor Classifiers (Cover and Hart, 1967). These
methods are implemented in software WEKA (Hall et al., 2009); their
parameter settings are well-tuned and described in the Supplementary
Text.

2.6 Datasets
To ensure a fair comparison, we have downloaded the datasets from the
PolyPhen-2 Web site, namely HumDiv and HumVar (version 2.1.0)
(Adzhubei et al., 2010). Both datasets were compiled from UniProtKB
(Magrane and Consortium, 2011). Specifically, HumDiv was compiled
using the annotation keywords, which imply causal mutation–phenotype
relationships, whereas HumVar was compiled from all the human
disease-causing mutations annotated. HumDiv has 7070 neutral
nsSNPs and 5322 deleterious nsSNPs, whereas HumVar has 21142 neu-
tral nsSNPs and 20989 deleterious nsSNPs. In particular, HumDiv is
considered higher in quality than HumVar because the SNPs in
HumDiv were selected using a controlled set of keywords.

3 RESULTS
3.1 Classification model selections in SNPdryad
We trained and tested the aforementioned 10 classification
models on HumDiv and HumVar, following the standard
10-fold cross-validation. The results are shown in Table 1 for
HumDiv and HumVar. It is clear that, for each classification
method and for both HumDiv and HumVar, using only ortho-
logous proteins had a better performance than inclusion of par-
alogous proteins. Not only did the orthologs-only approach
achieve better prediction accuracy and Area Under Receiver
Operating Characteristics (ROC) Curve (AUC), but it also had
lower level of error. In addition, the Random Forest method had
the best performance; this is also consistent with a previous com-
parison study showing that Random Forest is the best in high-
dimensions empirically (Caruana et al., 2008). The predictive
power of the Random Forest classifier lies in its ensemble
nature with bagging and random subspace techniques. Instead
of relying on a single-decision tree, Random Forest takes into
account the votes of multiple decision trees to make the final
prediction decision.

3.2 Comparisons on the HumDiv and HumVar datasets
Next, we compared the prediction performance of SNPdryad
(using only orthologs and Random Forest classifier) with other
commonly used nsSNP analysis methods, MutationTaster,
PolyPhen2 and SIFT (Adzhubei et al., 2010; Ng and Henikoff,
2003; Schwarz et al., 2010). Some other prediction methods such
as SNPs3D were not included in the comparison because they are
no longer actively maintained (Yue et al., 2006). Ensemble meth-
ods such as Condel, which reanalyze results from other methods,
are not included in the comparison either. We used the HumDiv
and HumVar datasets for the basis of the comparison. PolyPhen2
and other methods also used the same datasets in evaluations. To
ensure an objective and unbiased comparison, we downloaded the
prediction results from the web servers of PolyPhen2, SIFT and
MutationTaster, respectively, in January 2012. Figure 2 compares
the ROC curves and the Precision-Recall curves for these four
methods. It is clear from Figure 2 that SNPdryad outperforms
all three other methods on both HumDiv and HumVar datasets,
whereas PolyPhen2 has the second best performance. We like to
Using the contrast to a previous study (Li et al., 2012), we investigated how frequently the methods agreed with each other on whether an nsSNP is considered deleterious. Because these methods use different statistical schemes to denote prediction confidence or degree of harmful effect, it is impossible to select a single statistical threshold (e.g. a single P-value) to apply to all methods and compare the predictions that are above this threshold. To overcome this problem, for each method, we first selected all the nsSNP that are predicted to be deleterious and then ranked these nsSNPs according to this specific method’s own scoring scheme, from the most deleterious to the least deleterious. We then compared the overlap among the top ranked nsSNPs, the resultant Venn diagrams are shown in Figure 3. Three observations can be made from these Venn diagrams. (i) At a lower cutoff, the methods have a greater level of overlap in their predictions; however, among the top predicted nsSNPs, the overlapping fraction becomes much smaller. (ii) The nsSNPs that are predicted to be deleterious by more than one method are always more accurate than those predicted by a single method. Notably, such observation is in contrast to a previous study (Li et al., 2012). A possible explanation is that SNPdryad and PolyPhen2 are trained on the same datasets, boosting their ensemble performance. (iii) SNPdryad predicts more unique deleterious nsSNP that are missed by the others at a higher accuracy. For example, among the top 10% predicted nsSNPs, SNPdryad predicted 260 unique deleterious nsSNPs alone (256 are true positives, accuracy = 98%).

We next ranked the nsSNPs from HumDiv by the predicted scores of each of the three methods, and calculated the pair-wise Spearman’s rank correlation coefficients. As a comparison, we also calculated the correlations with the annotated deleterious effect provided in the HumDiv dataset (0 denotes neutral and 1 denotes deleterious). Table 2 shows that among the four methods, SNPdryad has the highest correlation with Annotation (0.83) than the others, suggesting that SNPdryad has higher accuracy than other methods.

### 3.3 Exclusion of paralogs contributes to better performance in SNPdryad

After concluding that SNPdryad is accurate in predicting deleterious nsSNPs, at least on the HumDiv and HumVar datasets, we next investigated whether such a superior performance was the result of either a better classifier (Random Forest) or the fact

<table>
<thead>
<tr>
<th>HumDiv dataset</th>
<th>RF</th>
<th>NB</th>
<th>BNet</th>
<th>MLP</th>
<th>AB</th>
<th>POLY</th>
<th>RBF</th>
<th>1NN</th>
<th>3NN</th>
<th>5NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Using the orthologous sequences from Inparanoid</td>
<td>Accuracy</td>
<td>0.93</td>
<td>0.88</td>
<td>0.90</td>
<td>0.91</td>
<td>0.90</td>
<td>0.91</td>
<td>0.91</td>
<td>0.87</td>
<td>0.87</td>
</tr>
<tr>
<td>Kappa statistics</td>
<td>0.85</td>
<td>0.76</td>
<td>0.79</td>
<td>0.81</td>
<td>0.80</td>
<td>0.83</td>
<td>0.82</td>
<td>0.73</td>
<td>0.74</td>
<td>0.74</td>
</tr>
<tr>
<td>Root mean absolute error</td>
<td>0.24</td>
<td>0.34</td>
<td>0.31</td>
<td>0.29</td>
<td>0.27</td>
<td>0.29</td>
<td>0.30</td>
<td>0.36</td>
<td>0.32</td>
<td>0.31</td>
</tr>
<tr>
<td>Root relative squared error</td>
<td>0.48</td>
<td>0.68</td>
<td>0.64</td>
<td>0.58</td>
<td>0.54</td>
<td>0.59</td>
<td>0.60</td>
<td>0.73</td>
<td>0.64</td>
<td>0.63</td>
</tr>
<tr>
<td>AUC</td>
<td>0.98</td>
<td>0.94</td>
<td>0.96</td>
<td>0.96</td>
<td>0.92</td>
<td>0.91</td>
<td>0.91</td>
<td>0.87</td>
<td>0.92</td>
<td>0.93</td>
</tr>
<tr>
<td>Using the homologous sequences from UniRef100</td>
<td>Accuracy</td>
<td>0.89</td>
<td>0.78</td>
<td>0.86</td>
<td>0.87</td>
<td>0.85</td>
<td>0.88</td>
<td>0.87</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td>Kappa statistics</td>
<td>0.77</td>
<td>0.56</td>
<td>0.71</td>
<td>0.73</td>
<td>0.68</td>
<td>0.75</td>
<td>0.72</td>
<td>0.67</td>
<td>0.67</td>
<td>0.68</td>
</tr>
<tr>
<td>Root mean absolute error</td>
<td>0.28</td>
<td>0.46</td>
<td>0.37</td>
<td>0.34</td>
<td>0.33</td>
<td>0.35</td>
<td>0.37</td>
<td>0.40</td>
<td>0.35</td>
<td>0.34</td>
</tr>
<tr>
<td>Root relative squared error</td>
<td>0.57</td>
<td>0.93</td>
<td>0.75</td>
<td>0.69</td>
<td>0.67</td>
<td>0.71</td>
<td>0.74</td>
<td>0.81</td>
<td>0.71</td>
<td>0.69</td>
</tr>
<tr>
<td>AUC</td>
<td>0.96</td>
<td>0.90</td>
<td>0.93</td>
<td>0.94</td>
<td>0.92</td>
<td>0.87</td>
<td>0.86</td>
<td>0.84</td>
<td>0.90</td>
<td>0.91</td>
</tr>
<tr>
<td>HumVar dataset</td>
<td>RF</td>
<td>NB</td>
<td>BNet</td>
<td>MLP</td>
<td>AB</td>
<td>POLY</td>
<td>RBF</td>
<td>1NN</td>
<td>3NN</td>
<td>5NN</td>
</tr>
<tr>
<td>----------------</td>
<td>----</td>
<td>----</td>
<td>------</td>
<td>-----</td>
<td>----</td>
<td>------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Using the orthologous sequences from Inparanoid</td>
<td>Accuracy</td>
<td>0.83</td>
<td>0.76</td>
<td>0.80</td>
<td>0.81</td>
<td>0.79</td>
<td>0.82</td>
<td>0.82</td>
<td>0.77</td>
<td>0.79</td>
</tr>
<tr>
<td>Kappa statistics</td>
<td>0.66</td>
<td>0.53</td>
<td>0.60</td>
<td>0.62</td>
<td>0.58</td>
<td>0.64</td>
<td>0.64</td>
<td>0.54</td>
<td>0.58</td>
<td>0.60</td>
</tr>
<tr>
<td>Root mean absolute error</td>
<td>0.35</td>
<td>0.48</td>
<td>0.44</td>
<td>0.38</td>
<td>0.39</td>
<td>0.42</td>
<td>0.43</td>
<td>0.48</td>
<td>0.40</td>
<td>0.39</td>
</tr>
<tr>
<td>Root relative squared error</td>
<td>0.70</td>
<td>0.96</td>
<td>0.88</td>
<td>0.75</td>
<td>0.77</td>
<td>0.85</td>
<td>0.85</td>
<td>0.96</td>
<td>0.81</td>
<td>0.77</td>
</tr>
<tr>
<td>AUC</td>
<td>0.91</td>
<td>0.87</td>
<td>0.88</td>
<td>0.89</td>
<td>0.86</td>
<td>0.82</td>
<td>0.82</td>
<td>0.78</td>
<td>0.85</td>
<td>0.87</td>
</tr>
<tr>
<td>Using the homologous sequences from UniRef100</td>
<td>Accuracy</td>
<td>0.81</td>
<td>0.73</td>
<td>0.78</td>
<td>0.80</td>
<td>0.79</td>
<td>0.80</td>
<td>0.80</td>
<td>0.76</td>
<td>0.78</td>
</tr>
<tr>
<td>Kappa statistics</td>
<td>0.63</td>
<td>0.46</td>
<td>0.55</td>
<td>0.59</td>
<td>0.58</td>
<td>0.61</td>
<td>0.61</td>
<td>0.53</td>
<td>0.55</td>
<td>0.56</td>
</tr>
<tr>
<td>Root mean absolute error</td>
<td>0.36</td>
<td>0.52</td>
<td>0.46</td>
<td>0.38</td>
<td>0.38</td>
<td>0.44</td>
<td>0.44</td>
<td>0.49</td>
<td>0.41</td>
<td>0.40</td>
</tr>
<tr>
<td>Root relative squared error</td>
<td>0.72</td>
<td>1.04</td>
<td>0.92</td>
<td>0.76</td>
<td>0.76</td>
<td>0.89</td>
<td>0.89</td>
<td>0.97</td>
<td>0.83</td>
<td>0.80</td>
</tr>
<tr>
<td>AUC</td>
<td>0.90</td>
<td>0.83</td>
<td>0.86</td>
<td>0.88</td>
<td>0.87</td>
<td>0.80</td>
<td>0.80</td>
<td>0.78</td>
<td>0.84</td>
<td>0.85</td>
</tr>
</tbody>
</table>

Note: RF, Random Forest; NB, Naive Bayes; BNet, Bayes Network; MLP, MultiLayer Perceptron; AB, AdaBoost; POLY, Support Vector Machine using Polynomial Kernel; RBF, Support Vector Machine using Radial Basis Kernel; kNN, k Nearest Neighbor. The parameter settings can be referred to supplementary data. In each row the classification model that has the best performance is highlighted in bold.

Table 1. Results of the 10 Classification Models trained and tested on the HumDiv and HumVar dataset using the standard 10-fold cross-validation.
that we only included orthologous protein sequences. We compared the AUC values of PolyPhen2 (Figure 2A and 2B, 0.95 for HumDiv and 0.89 for HumVar) to the AUC values of different classifiers that we tested in SNPdryad (Table 1). It can be seen from Table 1 that if homologous sequences are used, PolyPhen2 has higher AUC than all the classification models except for Random Forest (0.96). However, the top half of the Table 1 shows that if only orthologous sequences are used, several other classification methods such as BNet (Bayesian Net), MLP (MultiLayer Perceptron) and AB (AdaBoost) actually have better AUC scores (0.96) than PolyPhen2 (0.95). Such comparisons showed that the decision of including only orthologous protein sequences is the primary reason for the good prediction power of SNPdryad.

Table 2. Pair-wise Spearman rank correlation coefficients between the prediction scores of SNPdryad, PolyPhen2, SIFT, MutationTaster and the annotations on the HumDiv dataset. The highest is highlighted in bold

<table>
<thead>
<tr>
<th></th>
<th>SNPdryad</th>
<th>PolyPhen2</th>
<th>SIFT</th>
<th>MutationTaster</th>
<th>Annotations</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNPdryad</td>
<td>1</td>
<td>0.8145</td>
<td>0.7084</td>
<td>0.6983</td>
<td>0.8271</td>
</tr>
<tr>
<td>PolyPhen2</td>
<td>1</td>
<td>0.8145</td>
<td>0.7084</td>
<td>0.6983</td>
<td>0.8271</td>
</tr>
<tr>
<td>SIFT</td>
<td>1</td>
<td>0.7764</td>
<td>0.7634</td>
<td>0.8044</td>
<td>0.8044</td>
</tr>
<tr>
<td>MutationTaster</td>
<td>1</td>
<td>0.7167</td>
<td>0.7167</td>
<td>0.7157</td>
<td>0.7167</td>
</tr>
<tr>
<td>Annotations</td>
<td>1</td>
<td>1</td>
<td>0.7157</td>
<td>0.7167</td>
<td>1</td>
</tr>
</tbody>
</table>
To further illustrate such an idea, two examples of deleterious nsSNPs are shown in the Supplementary Figures S1–S6, which are correctly predicted by SNPdryad but by neither PolyPhen2 nor SIFT. These cases demonstrate that the quality of the MSA is crucial in accurately predicting deleterious nsSNPs. The first case is a SNP located at the position 1161 of the Complement C3 protein (P01024), which changes a Glutamine residue (Q) to a Lysine residue (K), causing increased susceptibility to hemolytic uremic syndrome atypical type 5 (AHUS5) (UniProt variation ID: VAR_063219). Supplementary Figures S1–S3 compare the alignments generated by SNPdryad, PolyPhen2 and SIFT, respectively. The substitutions to Lysine (K) or Arginine (R) are found in the alignments generated by both PolyPhen2 (Supplementary Figure S3) and SIFT (Supplementary Figure S2), misleading these algorithms to predict this particular SNP as a neutral substitution in human. This is likely caused by the inclusion of protein sequences paralogous to the input human protein. In contrast, neither Lysine (K) nor Arginine (R) is found in the orthologous sequence alignment computed by SNPdryad, which accurately predicted this nsSNP as deleterious (Supplementary Figure S1). Supplementary Figures S4–S6 show another example of a mutation at position 104 on human Transthyretin protein (P02766), which changes an Isoleucine (I) to a Serine residue (S) (UniProt ID: VAR_007584). This deleterious mutation has been shown to contribute to transthyretin-related amyloidosis (AMYL-TTR). The sequence alignments generated by both SIFT (Supplementary Figure S5) and PolyPhen2 (Supplementary Figure S6) include residue Serine at position 104, which likely caused both programs to predict this substitution as a neutral one. SNPdryad accurately predicted this mutation as a deleterious mutation (Supplementary Figure S4).

3.4 Prediction on nsSNPs annotated in SNPdbe

Next, as an independent benchmark, we used SNPdryad to make predictions on the nsSNPs curated in the SNPdbe database, and compared the results of SNPdryad, PolyPhen2, SIFT and MutationTaster. SNPdbe is a comprehensive database that collects and annotates SNP information from multiple sources, including dbSNP, SwissProt, OMIM and 1000 Genomes.

3.5 Prediction on nsSNPs annotated in ExoVar

As another independent benchmark dataset, we have selected the latest ExoVar dataset (Li et al., 2013). ExoVar consists of not only the UniProt annotations but also the recent rare nsSNPs in the 1000 Genomes Project. To be consistent with the past study (Li et al., 2013), we have removed the variants that are not derived alleles. As shown in Figure 5, the performance of SNPdryad is similar to that of Logit and better than all the other methods, including PolyPhen2 and MutationTaster. It is worth noting that both Logit and Condel are ensemble methods, which take input predictions from individual methods and output a weighted average. SNPdryad has the best performance among all the independent methods.

3.6 Predictions on human proteome

In this section, we first trained SNPdryad on the HumDiv dataset then ran it on the entire human proteome (Ensembl version GRCh37.64) for all the possible substitutions at all the amino acid positions. Note, in this sense, we are testing the functional effect of amino acid substitutions instead of on nsSNPs. Prediction Score Distribution: Supplementary Figure S7 depicts the average prediction scores for all the possible pair-wise substitutions between 20 amino acid residues, averaged over all the possible amino acid positions in the human proteome. The complete predictions can be accessed at our Website: http://snps.ccbr.utoronto.ca:8080/SNPdryad/. In general, not surprisingly, we observe that the predicted deleterious effect of replacing one amino acid by another is consistent with their similarity in physicochemical properties. In particular, most of the non-synonymous substitutions to Tryptophan (W) are likely to be deleterious and most of the non-synonymous substitutions to Alanine (A) are likely functional neutral. Furthermore, Serine

![Fig. 4. Performance comparison of SNPdryad, PolyPhen2, SIFT and MutationTaster on the labeled SNPdbe dataset. (a) ROC curves (b) PRC curves](https://academic.oup.com/bioinformatics/article-abstract/30/8/1112/257784)
Fig. 5. Performance comparison of SNPdryad with other methods on the ExoVar dataset. (a) ROC curves (b) PRC curves. Note that the ensemble and individual methods are denoted in dotted lines and solid lines, respectively. The performance curves of the other methods are adopted from the past literature (Li et al., 2013)

(S) and Threonine (T) are mostly interchangeable to each other based on the prediction scores.

In total, we scanned 92,012 human proteins (including protein isoforms) and 36,935,804 amino acid positions; a total of 10,120,155 substitutions (~1.4%) were predicted to be fully deleterious (with the SNPdryad prediction score of 1).

Next, we investigated whether such fully deleterious SNPs are enriched or depleted among common human variants. To achieve this, we filtered the nsSNPs annotated in the SNPdbs database, and retained only those common nsSNPs that have major allele frequency >0.05, as estimated by the 1000 Genomes Project. This resulted in 14,733 nsSNPs on 6,645 proteins (designated as Query set). Among them, 627 nsSNPs were predicted by SNPdryad to be fully deleterious. As a Background control dataset, we also took the same 6,645 human proteins, and simulated all the possible non-synonymous substitutions at all positions under the constraint that their amino acid type changes [e.g. Arginine (R) to Glutamine (Q)] do exist in the aforementioned 14,733 common variants. The resultant Background had 32,246,488 nsSNPs, among which 2,529,197 nsSNPs were predicted to be fully deleterious by SNPdryad. By computing the hypergeometric cumulative distribution function between the Query set and Background set, we observed significant depletion of fully deleterious SNPs among the common nsSNPs (Query Set) with \( P < 1.0674 \times 10^{-69} \). Such depletion of highly deleterious SNPs, while fully expected, further demonstrated the value and effectiveness of prediction algorithms such as SNPdryad.

PFAM Domain Statistics: We are interested to know where these fully deleterious substitutions are located, especially which PFAM protein domains are enriched of such extremely deleterious substitutions. Supplementary Table S2 lists the top PFAM domains that contain the highest number of harmful (i.e. fully deleterious) nsSNPs. It shows that the G protein coupled receptor (GPCR) domain (PF00001) contains the highest number of harmful nsSNPs as predicted by SNPdryad. GPCR domains are important domains in cell-surface receptors, which sense molecules outside of cells in many disease-causing signaling pathways. They are also important therapeutic targets, as ~40% of all modern drugs target GPCR proteins (Filmore, 2004). We also used odds-ratio to ascertain the statistical significance of these enrichments (last column). (Assuming every sequence position in every human protein are equally likely to be hit by a harmful nsSNP, we define the sum of the sequence lengths of the domain hits of a Pfam domain \( D \) as \( l_D \) in a human proteome. We also define the sum of the sequence lengths of all the proteins in the human proteome as \( l_{Total} \). Given a harmful nsSNP, the probability that it is located at the domain \( D \) is calculated as \( \frac{l_D}{l_{Total}} \). If we have \( N \) harmful nsSNPs, the expected number of their hits located at the domain \( D \) can be calculated as \( N \times \frac{l_D}{l_{Total}} \). Based on such an idea, the expected nsSNP hits are calculated for each domain. The odds-ratio of a domain is then calculated as the number of observed deleterious nsSNPs residing in a domain, divided by the expected number of deleterious nsSNPs in the domain.)

4 DISCUSSION

In this article, we described SNPdryad, a novel computational method that can predict deleterious effect of amino acid substitutions occurred in human proteins. As elaborated in this article, SNPdryad outperforms other leading algorithms in accurately predicting deleterious nsSNPs. We demonstrated that this is primarily because SNPdryad only includes orthologous sequences in building the multiple-sequence alignment, as opposed to other contemporary methods, which include paralogous sequences as well. Such an innovation allows construction of a more accurate protein sequence conservation profile, allowing a more precise estimate on whether a substitution is tolerated at a specific position. This would not have been possible until now, when a large number of mammalian or vertebrate genome sequences have been completely sequenced, thanks to the drastically decreasing cost of genomic sequencing. The next-generation sequencing technology has generated a deluge of genomic sequences and subsequently, a wealth of genetic variation data such as non-synonymous polymorphisms. We envision that an intelligent algorithm such as SNPdryad can take advantage of this large amount of data and further improve the accuracy of predicting deleterious nsSNPs.
SNPdryad: Predicting deleterious amino acid substitutions

ACKNOWLEDGEMENT

The authors would like to thank the five anonymous reviewers for their constructive comments. This article is dedicated to the memory of S.Z.

Funding: Discovery Grant from Natural Sciences and Engineering Research Council, Canada (NSERC), grant number [327612-2009 RGPIN to Z.Z.]; Acres Inc. - Joseph Yonan Memorial Fellowship, Kwok Sau Po Scholarship, and International Research and Teaching Assistantship from University of Toronto (to K.W.).

Conflict of Interest: none declared.

REFERENCES


Condel.


