PHACT: Phylogeny-Aware Computing of Tolerance for Missense Mutations

Nurdan Kuru, Onur Dereli, Emrah Akkoyun, Aylin Bircan, Oznur Tastan, and Ogun Adebali *

[1]Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul 34956, Turkey

***Corresponding author:** E-mail: oadebali@sabanciuniv.edu.

Associate Editor: Michael Rosenberg

Abstract

Evolutionary conservation is a fundamental resource for predicting the substitutability of amino acids and the loss of function in proteins. The use of multiple sequence alignment alone—without considering the evolutionary relationships among sequences—results in the redundant counting of evolutionarily related alteration events, as if they were independent. Here, we propose a new method, PHACT, that predicts the pathogenicity of missense mutations directly from the phylogenetic tree of proteins. PHACT travels through the nodes of the phylogenetic tree and evaluates the deleteriousness of a substitution based on the probability differences of ancestral amino acids between neighboring nodes in the tree. Moreover, PHACT assigns weights to each node in the tree based on their distance to the query organism. For each potential amino acid substitution, the algorithm generates a score that is used to calculate the effect of substitution on protein function. To analyze the predictive performance of PHACT, we performed various experiments over the subsets of two datasets that include 3,023 proteins and 61,662 variants in total. The experiments demonstrated that our method outperformed the widely used pathogenicity prediction tools (i.e., SIFT and PolyPhen-2) and achieved a better predictive performance than other conventional statistical approaches presented in dbNSFP. The PHACT source code is available at https://github.com/CompGenomeLab/PHACT.

Key words: **phylogenetics, Mendelian diseases, pathogenicity scoring, amino acid substitution.**

Introduction

Advancements in the characterization of single-nucleotide polymorphisms (SNPs) have significantly facilitated our understanding of the genomic differences between individuals (Kwok and Gu 1999). In various hereditary diseases, SNPs determine the differential susceptibility to the condition. Single-nucleotide variations in the coding regions might cause a single amino acid change in the encoded protein (i.e., missense mutations). Although some missense mutations are tolerable, in some cases, these amino acid substitutions may disrupt protein function and lead to diseases (Castellana and Mazza 2013).

Understanding the deleterious effect of a missense mutation facilitates the diagnosis of Mendelian diseases. Although the reduction in the cost of genome sequencing has enabled a massive sequence data generation in clinical settings, the assessment of the functional impact of variants, experimentally, in a high-throughput fashion remains challenging. Therefore, many computational techniques have been developed to predict the effects of missense mutations (Garber et al. 2009; McVicker et al. 2009; Adzhubei et al. 2010; Davydov et al. 2010; Pollard et al. 2010; Choi et al. 2012; Sim et al. 2012; Carter et al. 2013; Schwarz et al. 2014; Dong et al. 2015; Gulko et al. 2015; Lu et al. 2015; Ioannidis et al. 2016; Ionita-Laza et al. 2016; Vaser et al. 2016; Feng 2017; Raimondi et al. 2017; Alirezaie et al. 2018; Rogers et al. 2018; Rentzsch et al. 2019; Malhis et al. 2020). Although these methods have not attained the desired level of accuracy and are not recommended for use in clinical studies, clinicians tend to use them to prioritize and reduce the number of variants to be analyzed (Eilbeck et al. 2017). Therefore, it remains critical to advance the methods aimed at pathogenicity prediction.

The pathogenicity prediction methods for missense mutations can be grouped into two main categories: (1) conventional statistical methods (Siepel et al. 2005; Garber et al. 2009; McVicker et al. 2009; Davydov et al. 2010; Pollard et al. 2010; Choi et al. 2012; Sim et al. 2012; Gulko et al. 2015; Lu et al. 2015; Vaser et al. 2016; Malhis et al. 2020) and (2) machine learning-based methods (Adzhubei et al. 2010; Carter et al. 2013; Schwarz et al. 2014; Dong et al. 2015; Ioannidis et al. 2016; Ionita-Laza et al. 2016; Feng 2017; Raimondi et al. 2017; Alirezaie et al. 2018; Rogers et al. 2018; Rentzsch et al. 2019; Jiang et al. 2021; Qi et al. 2021). The predictions of most of the conventional statistical methods rely on the conservation level of the protein position obtained from multiple sequence alignment (MSA). It is naturally expected that a position that has been conserved for millions of years through evolution is unlikely to tolerate a substitution. Thus, a substitution that disrupts a conserved position increases the risk of pathogenicity (Sunyaev et al. 2000). Nevertheless, substituting amino acids in homologous sequences is unlikely to

Article

reduce the evolutionary fitness and to be deleterious for the query species (Jordan et al. 2010). Therefore, the conservation level is crucial to understand the tolerance of a specific position to amino acid substitutions.

Machine learning-based variant-effect prediction algorithms improve the predictive ability of pathogenicity prediction methods through the inclusion of structural, functional, and physicochemical features, in addition to the sequence-conservation-related features (Adzhubei et al. 2010; Carter et al. 2013; Schwarz et al. 2014; Dong et al. 2015; Ioannidis et al. 2016; Ionita-Laza et al. 2016; Feng 2017; Raimondi et al. 2017; Alirezaie et al. 2018; Rogers et al. 2018; Rentzsch et al. 2019; Jiang et al. 2021; Qi et al. 2021). Because many of these algorithms use conservation-related scores as input features, the use of more accurate conservation-based pathogenicity scoring methods would improve their predictive performances, thus highlighting the benefits of developing such methods.

Most evolution-based tools (Siepel et al. 2005; Garber et al. 2009; McVicker et al. 2009; Davydov et al. 2010; Pollard et al. 2010; Choi et al. 2012; Sim et al. 2012; Lu et al. 2015; Vaser et al. 2016; Malhis et al. 2020) rely on MSA to measure the conservation of a position and predict the substitutability of an amino acid. This approach has two main drawbacks: the MSA (1) cannot distinguish between the independent and dependent amino acid substitutions at a single position and (2) ignores the evolutionary distance between genes (see supplementary fig. S1, Supplementary Material online for sample cases). To circumvent these problems, we introduce PHACT, which uses both MSA and the corresponding phylogenetic tree. PHACT employs the reconstructed amino acid probabilities of each tree node to predict amino acid substitution tolerance. Experiments on two datasets demonstrated that PHACT accurately predicts the deleterious effect of missense mutations.

New Approaches

We introduce a novel phylogeny-dependent probabilistic approach to predict the functional effects of missense mutations. Our approach exploits the phylogenetic tree to measure the deleteriousness of a given variant. PHACT considers various factors in the establishment of the decision of whether a substitution is tolerated at the corresponding position based on the following considerations:

- Conservation scores based solely on MSAs disregard the evolutionary dependencies among alteration events that have taken place during evolution. Multiple amino acid alterations observed might stem from a single evolutionary event in a common ancestral state, which might affect multiple descendants' leaves (sequences). The evaluation of these variations as independent alteration events causes redundancies and the overcounting of these changes. Similarly, when using MSA alone, independent evolutionary events do not receive sufficient scoring, which increases the probability of substitution tolerance. PHACT traverses the phylogenetic tree to compute the probability differences for all connected nodes and ignores the dependent substitutions.
- Evolutionary information gathered from species located on the different nodes of a phylogenetic tree should not be treated equally in substitutability scoring. Amino acids observed in evolutionarily close species are better indicators of neutrality for the query species. We assign a weight for each node proportionally to their phylogenetic distance to the query sequence.
- The physicochemical properties of amino acids are also instrumental in deciding on the neutrality or pathogenicity of a substitution. Depending on the property constraints on a position, amino acids with similar physicochemical properties can replace each other without affecting the function of the protein (Kumar et al. 2009). PHACT incorporates the physicochemical properties of amino acids into the ancestral reconstruction step. In turn, ancestral reconstruction reports the resulting probability distribution of amino acids per node by considering the observed amino acids at MSA, as well as their physicochemical properties, with the help of amino acid replacement matrices. These matrices, such as LG, are designed to capture the biological and physicochemical properties of amino acids. We used LG4X, a more advanced model consisting of four LG-based matrices defined by considering different substitution rates and site heterogeneity (Le et al. 2012). With the help of LG matrices, rather than observations alone, the expected amino acid substitutions are also included in probabilities, and thus, in our score computation.

PHACT derives independent evolutionary events and phylogenetic relationships among species from gene-based phylogenetic trees. The probability distribution at each internal node of the phylogenetic tree is obtained using ancestral reconstruction. We summarize the workflow of PHACT in figure 1 and present the pseudocode used for scoring in Algorithm 1. PHACT takes the MSA of the gene, the phylogenetic tree, and the probability distribution of amino acids at each ancestral node as input (fig. 1A). Starting from the query species, which was *Homo sapiens* in this work, we traversed the tree and recorded the probability of change for each amino acid, to determine where the substitutions occurred. We then used these probabilities to predict the effect of the amino acid change on the query species according to the phylogenetic location of the alteration. The arrows on the phylogenetic tree in figure 1*B* represent the direction of the probability subtraction process. Here, the rationale for using probability differences was to identify the point at which the probability of amino acid substitution increases; that is, we determined the phylogenetic nodes at which missense mutations have emerged using the positive probability differences. Although a positive change in

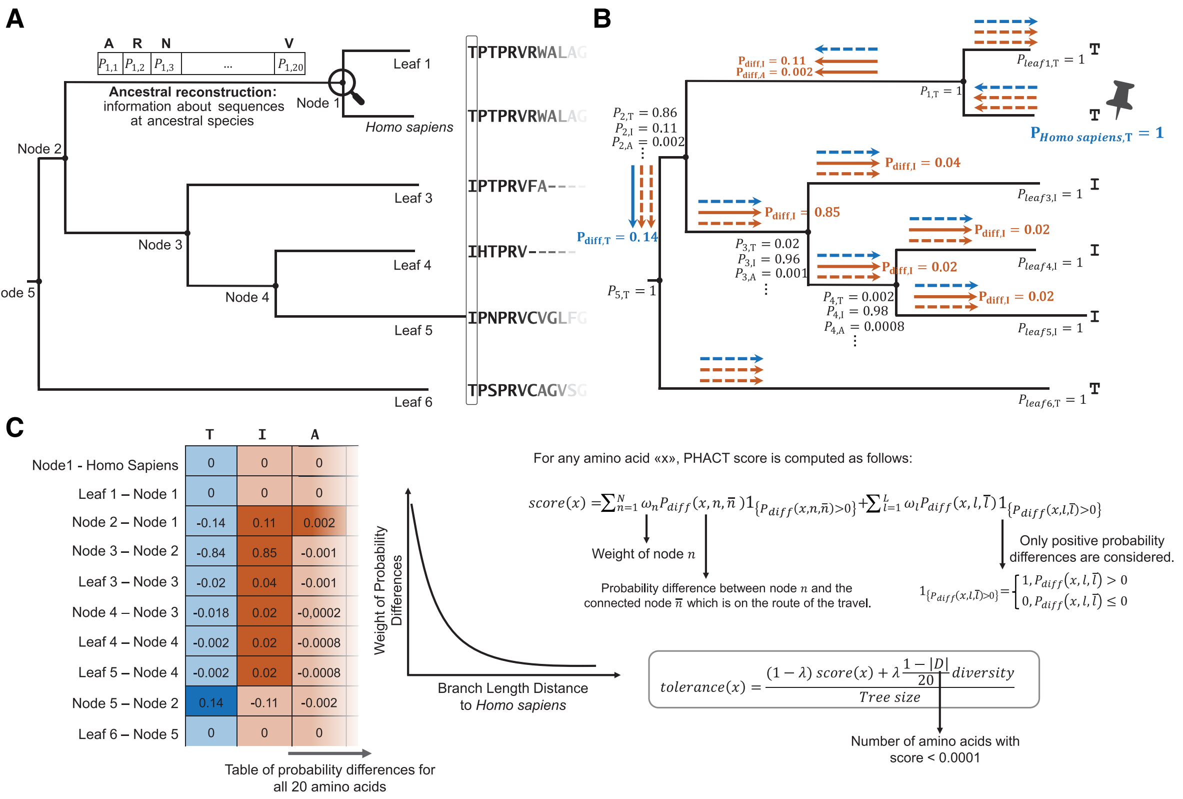
A
A R N V
$P_{1,1}$ $P_{1,2}$ $P_{1,3}$ ... $P_{1,20}$
Ancestral reconstruction: information about sequences at ancestral species
Node 1
Node 2
Node 3
Node 4
Node 5
Leaf 1
Homo sapiens
Leaf 3
Leaf 4
Leaf 5
Leaf 6
TPTPRVRWALAG
TPTPRVRWALAG
IPTPRVFA----
IHTPRV-----
IPNPRVCVGLFG
TPSPRVCAGVSG
B
$P_{diff,I} = 0.11$
$P_{diff,A} = 0.002$
$P_{1,T} = 1$
$P_{leaf1,T} = 1$ T
T
$P_{Homo\ sapiens,T} = 1$
$P_{2,T} = 0.86$
$P_{2,I} = 0.11$
$P_{2,A} = 0.002$
$P_{diff,T} = 0.14$
$P_{diff,I} = 0.85$
$P_{diff,I} = 0.04$
$P_{leaf3,I} = 1$ I
$P_{3,T} = 0.02$
$P_{3,I} = 0.96$
$P_{3,A} = 0.001$
$P_{diff,I} = 0.02$
$P_{diff,I} = 0.02$
$P_{leaf4,I} = 1$ I
$P_{4,T} = 0.002$
$P_{4,I} = 0.98$
$P_{4,A} = 0.0008$
$P_{diff,I} = 0.02$
$P_{leaf5,I} = 1$ I
$P_{5,T} = 1$
$P_{leaf6,T} = 1$ T
C
T I A
Node1 - Homo Sapiens 0 0 0
Leaf 1 – Node 1 0 0 0
Node 2 – Node 1 -0.14 0.11 0.002
Node 3 – Node 2 -0.84 0.85 -0.001
Leaf 3 – Node 3 -0.02 0.04 -0.001
Node 4 – Node 3 -0.018 0.02 -0,0002
Leaf 4 – Node 4 -0.002 0.02 -0.0008
Leaf 5 – Node 4 -0.002 0.02 -0.0008
Node 5 – Node 2 0.14 -0.11 -0.002
Leaf 6 – Node 5 0 0 0
Table of probability differences for all 20 amino acids
Weight of Probability Differences
Branch Length Distance to Homo sapiens
For any amino acid «x», PHACT score is computed as follows:
$score(x) = \sum_{n=1}^{N} \omega_n P_{diff}(x,n,\bar{n}) 1_{\{P_{diff}(x,n,\bar{n})>0\}} + \sum_{l=1}^{L} \omega_l P_{diff}(x,l,\bar{l}) 1_{\{P_{diff}(x,l,\bar{l})>0\}}$
Weight of node n
Probability difference between node n and the connected node $\bar{n}$ which is on the route of the travel.
Only positive probability differences are considered.
$1_{\{P_{diff}(x,l,\bar{l})>0\}} = \begin{cases} 1, P_{diff}(x,l,\bar{l}) > 0 \\ 0, P_{diff}(x,l,\bar{l}) \leq 0 \end{cases}$
$tolerance(x) = \frac{(1-\lambda)\, score(x) + \lambda \frac{1-|D|}{20} diversity}{Tree\ size}$
Number of amino acids with score < 0.0001

FIG. 1. Workflow of PHACT. (*A*) Input of the algorithm includes MSA, phylogenetic tree, and probability distribution of amino acids at each ancestral node. (*B*) Calculation of all probability differences between consecutive nodes and leaves, starting from query species. The blue and orange arrows correspond to the reference and alternating amino acids, respectively. Positive probability differences are represented as solid lines, whereas dashed arrows indicate that the probability difference for the corresponding amino acid is negative or zero. The values of positive probability differences are also indicated next to the corresponding arrows. (*C*) The weighted summation of positive probability differences yields an individual score per amino acid. The final score is obtained by eliminating the effect of the tree size (total number of nodes) and including the diversity of the position, which is obtained by summing the scores for all amino acids, with the exception of the score with the maximum value. The maximum weight of diversity, λ, is set as 0.1 in the final formula, which gives a tolerance score per amino acid. The weight of the diversity term increases proportionally with the number of amino acids that contributes to the summation.

probability indicates an alteration, negative probability changes are observed because of a substitution belonging to the previously visited part of the tree. The decrease in probability for a specific amino acid means that the expectation of encountering this amino acid decreases in the ensuing travel steps. The negative probability changes are ignored in score computation, to prevent repetitive counting of the dependent substitutions (fig. 1C).

It has been previously hypothesized that a variant in a human gene is more likely to be benign when it is present in closely related species. In contrast, it is more likely to be deleterious when it only exists in distant ones (Ionita-Laza et al. 2016; Malhis et al. 2020). In PHACT, during the tree traversal process, all positive changes in amino acid probabilities are added via a weighted summation, in which, the weights are considered to be inversely proportional to the distance between the corresponding nodes of change and the query sequence. Although we investigated several weighting approaches to determine the contribution of each node to the substitution score, the best performing weighting scheme was obtained by assigning a weight of 0.5 to the species with the closest evolutionary distance to the query sequence. The formal definition of weight at any node n for a phylogenetic tree with L leaves is as follows:

$$w_n = \frac{1}{1 + [d_n/\min(d_l)]} \tag{1}$$

where d_n is the distance between the query sequence and node n and d_l is the $(L-1)$-dimensional vector of the distance between the query sequence (human) and leaves, with the exception of the leaf of the query species. We employed various weighting approaches, including the Gaussian function; however, as explained in detail in the PHACT—Results section, the approach outlined in equation (1) performed better than the other functions. The details of the different weighting approaches can be found in the Supplementary Material online.

After completing a traversal on the tree, we obtained a weighted summation of probability differences for each of the 20 amino acids at the corresponding position. In addition to the individual scores of amino acids, the variability of a position is also an important factor in terms of the effects of an alteration. To include the variability of a position in the score, we rescaled the PHACT score based on the diversity of the position. The diversity information per position is obtained by summing all amino acid scores that are obtained in the traveling process through the internal nodes of the phylogenetic tree, with the exception of the one with the highest score. The amino acid with the highest score most often corresponds to the reference amino acid, and we eliminate this value in diversity computation because it generally affords an incorrect signal related to the variability of the position. The second misleading source of information about position diversity stems from the leaves of the phylogenetic tree. As the amino acid probabilities at leaves are set as 1 or 0, depending on the observed amino acid, they can dominate the score and shadow the substitution information obtained from the internal nodes of the tree. To prevent this, PHACT uses internal nodes and the leaf of query species exclusively in diversity computation by ignoring the remaining leaves of the tree. The weight of the diversity term is computed over the number of amino acids that contributes to the total score. We count the number of amino acids that have a higher score than a predefined threshold (set at 0.0001) and increase the weight of the diversity term by considering this count. The weight of the diversity term, $w_{\text{diversity}}$, is defined as follows:

$$w_{\text{diversity}} = \lambda\left(1 - \frac{\text{number of amino acids with score} < 0.0001}{20}\right) \tag{2}$$

where λ is the maximum possible weight (set at 0.1 in the design of PHACT). The increase in the number of amino acids with an individual score >0.0001 leads to the increase in the variability of the position and to the assignment of a higher weight to the diversity, up to 0.1. The final PHACT score, which includes both the individual scorings of each amino acid and the diversity of the position, is shown in Lines 12–15 of Algorithm 1. The score is classified according to the tree size, which corresponds to the total number of nodes, to eliminate the bias associated with obtaining larger tolerance values for larger trees. Because we classified the score according to the tree size at the end of the procedure, the final score became a small number that is difficult to interpret. To overcome this scaling problem, we shifted the alternating amino acid scores to a [0, 1] interval with the help of the formula given in Line 16 of Algorithm 1 and obtained the "tolerance score," which is mentioned as the "PHACT score" throughout the manuscript. The final formula of tolerance for an amino acid x with respect to PHACT is computed as in equation (3):

$$\text{tolerance}(x) = 1 - \frac{\log\left(\dfrac{(1-\lambda)\text{score}(x) \ + \lambda\left(1 - \dfrac{|score_{wol} < 0.0001|}{20}\right)\text{diversity}}{\text{Tree size (number of nodes)}} + \varepsilon\right)}{\log(\varepsilon)} \tag{3}$$

where

$$\text{diversity} = \text{sum}(score_{wol}) - \ \max(score_{wol})$$

and λ is 0.1, ε is the small number used for scaling (set at 10^{-15}), $score(x)$ is obtained at the end of tree traversal for amino acid x, and $score_{wol}$ corresponds to the 20-dimensional vector representing the summation of weighted probability differences, with the exception of the leaves of the phylogenetic tree. $|score_{wol} < 0.0001|$ corresponds to the number of components of $score_{wol}$ having a smaller value than 0.0001.

The tolerance/PHACT score measures the possible deleterious or neutral effect of a missense mutation. Substitutions of reference amino acids with amino acids having a high tolerance score (close to 1) tended to yield

Algorithm 1: PHACT—Phylogeny-Aware Computing of Tolerance

```
Input: MSA M, Phylogenetic tree T(N;L) with N nodes and L leaves, Probability distribution matrix A (N + L by 20), Leaf of human $l_{human}$, The amino acid at the leaf of human $aa_{human}$, Individual score per amino acid score, Score over internal nodes of the tree per amino acid $score_{wol}$
1  Compute the node weights $\omega_n$, $\forall n$ in $\mathcal{N} = (1, \ldots, N, \ N+1, \ldots, N+L)$
2  $score = score_{wol} = 0$
3  $score(aa_{human}) = score_{wol}(aa_{human}) = 1$
4  for aa in 1:20
5      for n in $\mathcal{N} - \{l_{human}\}$
6          Find the connected node $\bar{n}$ on the direction of travel
7          $P_{diff}(aa, n, \bar{n}) = A(n, aa) - A(\bar{n}, aa)$
8          $score(aa) = score(aa) + \omega_n P_{diff}(aa, n, \bar{n}).1_{\{P_{diff}(aa,n,\bar{n})>0\}}$
9          $score_{wol(aa)} = score_{wol}(aa) + \omega_n P_{diff}(aa, n, \bar{n}).1_{\{P_{diff}(aa,n,\bar{n})>0 \ \& \ n \leq N\}}$
10     end
11 end
12 $diversity = sum(score_{wol}) - \max(score_{wol})$
13 Determine $D = \{score_{wol}(aa) | \ aa$ in 1:20 and $score_{wol}(aa) < 0.0001\}$
14 The weight of diversity term $w_{diversity} = 0.1*(1 - \frac{|D|}{20})$
15 $score_{upd}(aa) = \frac{(0.9score(aa) + w_{diversity}diversity)}{Tree\ size\ (number\ of\ nodes)}$
16 $tolerance(aa) = 1 - \frac{\log(score_{upd}(aa)+10^{-15})}{\log(10^{-15})}$
```

a neutral effect. A lower tolerance score corresponded to a deleterious effect on the protein function. Although the query sequence was human in our experiments, the algorithm can easily be used for other species.

Results

To assess the predictive performance of our algorithm, we performed computational experiments on two sets of data: (1) HCG, which stands for Humsavar, **C**linVar, **g**nomAD and includes the missense variants in Humsavar (The UniProt Consortium 2021), ClinVar (Landrum et al. 2016), and gnomAD (Karczewski et al. 2020); and (2) Grimm datasets (see the Materials and Methods section for details), which were constructed by us based on databases that are frequently used in the literature (Sasidharan Nair and Vihinen 2013; Grimm et al. 2015; Landrum et al. 2016; Karczewski et al. 2020; Liu et al. 2020; The UniProt Consortium 2021). The resulting vector of 20 components obtained at the end of PHACT computation (Algorithm 1) represents a composite of the individual substitution score for each amino acid. This vector is used to predict the possible effect of an alteration from the reference amino acid to the changed amino acid in question. The summed score correlates with the overall variability of the corresponding position regarding the query organism. This score is different from the conventional conservation scores, in/for which, the query species does not affect this general diversity measurement. In the PHACT—Results section, we present the performance measures for PHACT on the HCG dataset by explaining the contribution of different PHACT components. The second subsection is dedicated to a benchmark comparison with SIFT (Sim et al. 2012), PolyPhen-2 (Adzhubei et al. 2010), LIST-S2 (Malhis et al. 2020), and other statistical tools included in the dbNSFP database (Liu et al. 2020). In our experiments, we relied on the area under the receiver operating characteristic (ROC) curve (AUC), the area under the precision-recall (PR) curve (AUPR), F1 score, balanced accuracy, and Matthews correlation coefficient (MCC). In these analyses, positive and negative labels are used to define measures such that true positives and true negatives correspond to pathogenic and neutral variants, respectively.

PHACT—Results

We present the resulting ROC and PR curve figures and the AUC and AUPR values for the 2,836 proteins and 13,420 neutral and 15,728 pathogenic variants obtained from the HCG dataset. The details of the HCG dataset are provided in the Materials and Methods section. The ROC and PR curves are graphical plots that are used to understand the performance of a binary classifier that assigns the elements of a set into two groups. The *x*-axis and *y*-axis of the ROC curve indicate the true-positive rate (TPR, the ratio of the variants that are correctly classified as positive among all positives) and the false-positive rate (FPR, the ratio of variants that are mislabeled as positive among all negatives), respectively; moreover, under various thresholds, the ROC curve illustrates how the variations of TPR and FPR are related. Similarly, the PR curve is created by plotting precision (ratio of true positives to all variants that are labeled as positive) against recall (the same as TPR) under different threshold values. In our analyses, we used the areas under the ROC and PR curves to compare the performances of various missense mutation classifiers, as reported in the literature (Carter et al. 2013; Malhis et al. 2020; Jiang et al. 2021; Qi et al. 2021). Overall, larger AUC and AUPR values are associated with better performance.

Distance-Based Node-Weighting Approaches

Figure 2 presents the performance of PHACT for various weighting approaches. For the best performing approach, that is, Max05, we assigned a weight of 0.5 to the closest leaf to the query sequence, which in our experiments was *H. sapiens*. The details and explicit mathematical functions of weights are given in the Supplementary Material online. As shown in figure 2, Max05 achieved AUC and AUPR values of 0.949 and 0.949, respectively. An alternative version of Max05 that assigns the 0.5 weight to the closest node instead of the closest leaf (Max05 [Node]) yielded high AUC (0.935) and AUPR (0.934) values.

As an alternative weighting scheme, we experimented with the Gaussian functions using various bandwidth parameters, such as the mean and median of the distances, for assigning the importance of each node in the prediction of the functional consequence of a missense mutation. However, we observed that the Gaussian function did not perform well in the detection of the possible effect of an alteration on our query species. The bell-curved shape of the function resulted in the assignment of higher weights to the remote homologs, unless there was a drastic decrease in evolutionary distance. For example, in the tree of P10826, 328 out of 999 nodes were within a one-unit distance (namely, less than or equal to one substitution per site) to the query sequence (human). The branch length distances and corresponding weights for this sample protein are shown in supplementary figure S2, Supplementary Material online.

Conversely, we observed that the closer homologs to the query were more important than the distant ones for neutrality prediction. Because the Gaussian function does not entirely cover this importance, we mainly utilized a linearly decreasing approach depending on the distance to the query. We also checked whether the combination of Max05 and the Gaussian function with the mean as the bandwidth parameter affected the accuracy, and combined them using their geometric mean. This approach was labeled "Mix (Gaussian and Max05)," and the resulting performance can be found in figure 2*A* and *B*. Although a 3.1% increase in AUC and a 2.4% increase in AUPR versus the Gaussian function (mean) were achieved, this approach did not perform well as in Max05. We naturally

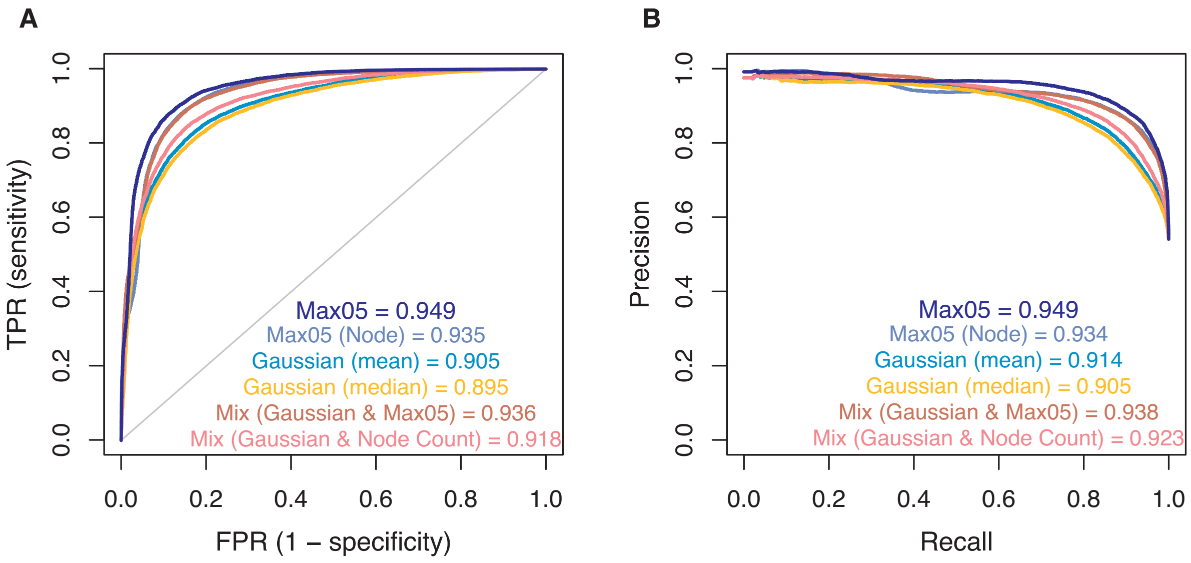

FIG. 2. Comparison of PHACT calculations using various weight functions. (*A*) ROC curve. AUC values are shown. Sensitivity (specificity) on the axis refers to the rate of positive (negative) predictions that are truly positive (negative). (*B*) PR curve. AUPR values are shown. Precision on the *y*-axis of the plot refers to the ratio of positive predictions among all predictions that were labeled as positive. Recall is the same as sensitivity. The best performing weight is Max05, with 0.949 AUC and 0.949 AUPR levels.

expected that the Gaussian function would detect the neutral variants better than Max05, as it attributes a higher score to the amino acids observed in distant species. However, both the Gaussian function and the new weights obtained via a combination of the two weights (Mix [Gaussian and Max05]) failed to outperform Max05, which demonstrates that Max05 not only is a good predictor of pathogenic variants, but also performs better than the Gaussian function in the prediction of neutral variants.

We also constructed a new weight, that is, Mix (Gaussian and Node Count), based on the Gaussian function, with the mean being used as the bandwidth parameter. However, because the Gaussian function did not lead to good predictions, we used the average of the Gaussian function and the number of nodes between query species and each internal node, to determine the weight of the corresponding node. This combined weight performed better than the Gaussian function. Nevertheless, Max05 outperformed this combined weight as well, with a 5.7% total difference in the AUC and AUPR levels (the performance can be found in fig. 2).

Baseline Comparisons

Our approach was based on an understanding that using evolutionary information is essential for explaining the functional consequences of amino acid alterations. We compared PHACT with some simple approaches to better grasp the contributions from different information sources. The set of alternative approaches included the ones that solely employ either MSA or the physicochemical properties of amino acids, to understand the effects of different information sources. Formal definitions of these simple approaches are given in the Supplementary Material online. The details of the resulting performances are illustrated in figure 3.

Blosum62 Score

The first baseline comparison relied on the substitution matrices. We computed a score based solely on the Blosum62 matrix (the details are given in the Supplementary Material online). The resulting AUC and AUPR levels were the lowest among all the baseline comparisons, at 0.648 and 0.657, respectively.

MSA-Based Diversity

Position diversity provides an insight into the prediction of pathogenicity of the unobserved amino acid substitutions (Sim et al. 2012; Malhis et al. 2020). In PHACT, we used diversity to adjust the individual amino acid scores by including position dynamics. However, we hypothesized that the exclusive use of MSA is inappropriate to correctly identify position diversity. Here, the MSA-based diversity measure relied on the number of different amino acids observed at the position in question. If the position was diverse in terms of the observed amino acids in the MSA, we assigned a higher score to substitutions in that position, which made them close to neutral. As shown in figure 3, by yielding 0.779 AUC and 0.790 AUPR levels, this approach resulted in one of the lowest scores in baseline comparisons.

Amino Acid Frequency

To establish a baseline, we compared our results with the most straightforward possible approach, which is

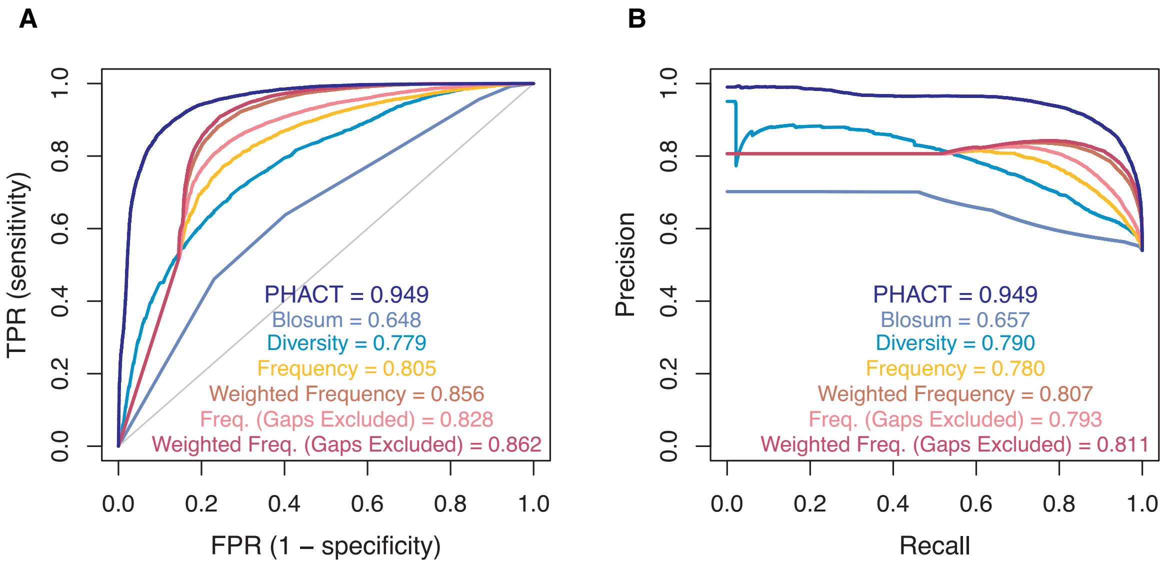

Fig. 3. Baseline comparisons. The performance differences between PHACT (Max05) and simple approaches are presented. (*A*) ROC curve. AUC values are shown. (*B*) PR curve. AUPR values are shown.

determination of the conservation level based on the frequency of change, that is, the ratio of the total number of alternative amino acids observed in the MSA position to the total number of sequences. Figure 3 shows that the exclusive use of this frequency information resulted in one of the lowest performances among the six approaches, with 0.805 AUC and 0.780 AUPR levels.

Weighted Amino Acid Frequency

In this approach, the distance between the query species and the remaining species in the alignment is used to compute a frequency score. Here, by eliminating the traveling through the tree process, we aimed to highlight its contribution by comparing the resulting performance with the original result of PHACT. Although this approach performed better than the classical frequency approach, it underperformed compared with PHACT, with a 9.3% decrease in AUC scores and a 14.2% decrease in AUPR scores.

Frequency and Weighted Frequency by Excluding Gaps

Here, in addition to employing frequency and weighted frequency approaches, we excluded the number of gaps in frequency computation; that is, the frequency and weighted frequency were computed over the total number of sequences after the sequences with a gap at the corresponding position were eliminated. The average ratio of gaps was 0.377, with a standard deviation of 0.306 for all proteins on DS1. Despite an increase in the AUC and AUPR values with respect to the versions that included gaps, an important performance difference remained relative to PHACT. As shown in figure 3, when the gaps were counted, the frequency and weighted approaches resulted in AUC values of 0.828 and 0.862 and AUPR values of 0.793 and 0.811, respectively.

Contribution of the Algorithm Steps

PHACT calculation relies on various components, such as evolutionary relationships between species, ancestral reconstruction, and traversal of the phylogenetic tree, which consider the probability differences and the weighting approach described above. To delineate the contribution of each of these components to the resulting performance, we employed several alternative versions of score computation by including one or more steps of PHACT, as presented in figure 4. We divided these alternative scoring strategies into two main groups: the intermediate steps based on the tree traversal process, and the final step based on the perturbation of the individual score per amino acid with respect to the diversity of the position. The details of these approaches are given in the Supplementary Material online. Figure 4*A* and *B* shows the ROC and PR curves, respectively, for the algorithms considering the intermediate steps, and figure 4*C* and *D* aims to explain the contribution of the final step to the predictions.

PHACT After Exclusion of the Leaves

As explained in the New Approaches section, PHACT starts from a query species and travels through the ancestral nodes and leaves of the phylogenetic tree to detect the location of substitutions and predict their effect on the query. An alternative approach is to exclude all the leaves from score computation and utilize only the internal nodes. We did not pursue this approach in the finalized version of PHACT for two reasons. First, although the ancestral reconstruction tools, such as RaxML-NG (Kozlov et al. 2019), compute the probability of each amino acid at ancestral nodes using the observed amino acids at the leaves (Yang 2006), ignoring the probability difference at the leaves of the phylogenetic tree can result in missing

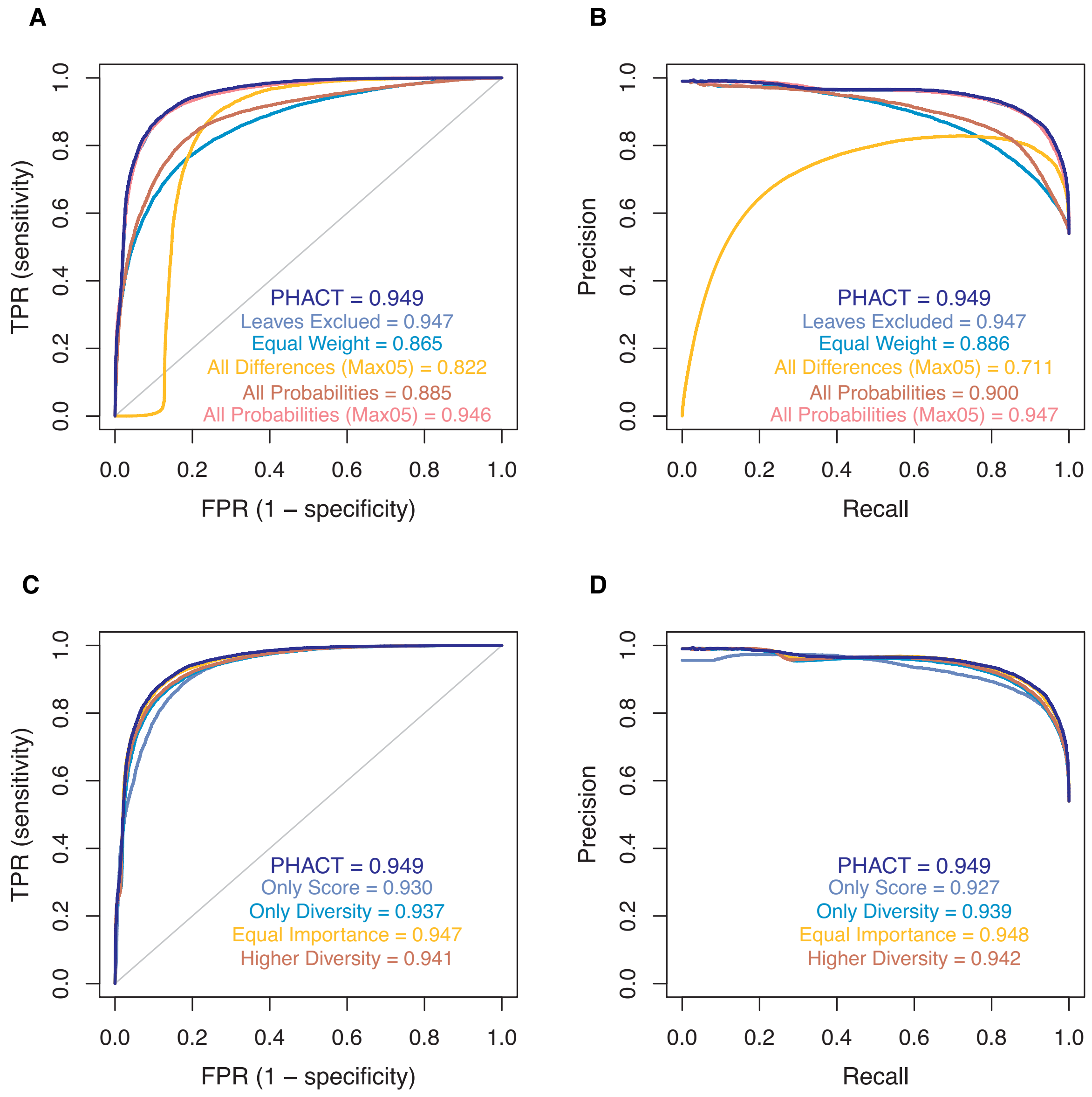

Fig. 4. Effects of the different components of PHACT on the performance of the algorithm. (*A*) and (*B*): ROC and PR curves for the methods related to the intermediate steps of PHACT. (*C*) and (*D*): ROC and PR curves for the final steps of PHACT.

some substitutions that occurred in the species that exist today. Second, we observed that the inclusion of the leaves yielded better performance in predicting neutral and pathogenic mutations over different datasets and proteins. The same trend is also observed in figure 4*A* and *B*. PHACT performed 0.2% better at both the AUC and AUPR levels than the approach in which the leaves were excluded.

PHACT Without Weighting

In this approach, we removed the weighting step of PHACT and assumed that substitutions are equally important regardless of the distance to the query species. Eliminating the weighting approach yielded an 8.4% lower AUC level and a 6.3% lower AUPR level compared with PHACT ("Equal-Weight" in fig. 4*A* and *B*). This result indicates that tree traversal is an essential component of this process.

Inclusion of Negative Probability Differences

PHACT uses the positive probability differences and disregards negative ones to factor in the substitution dependence, as mentioned in the New Approaches section. Instead, here we probed the performance change after incorporating all of the differences into PHACT. We observed a substantial decrease in the AUC and AUPR levels when both positive and negative differences were considered ("All Differences [Max05]" in fig. 4*A* and *B*), which highlighted the benefit of the inclusion of positive differences exclusively.

Inclusion of All Ancestral Reconstruction Probabilities

In this approach, we employed all of the full probabilities obtained from ancestral reconstruction, rather than the probability differences between nodes, to calculate the tolerance score for each amino acid. The major conceptual problem of this approach is the overcounting of dependent substitutions. We presented the performances achieved by using all of the probabilities with equal weights and the Max05 weight as two versions ("All Probabilities" and "All Probabilities [Max05]" in fig. 4*A* and *B*). We did not consider this version further because both approaches yielded a performance lower than that of PHACT. Moreover, the scores were highly correlated with the observed frequencies of amino acids in the MSA; thus, this was incompatible with our aim of accounting for the dependence and independence of substitutions.

PHACT Score Without the Diversity Term

As discussed in the New Approaches section, the PHACT score consists of two parts: individual amino acid scores and diversity of the position in MSA. Although individual scores represent the acceptability of each amino acid, the conservation level and variability of the corresponding position are also essential for predicting the effect of amino acid alterations. To analyze the effect of the diversity term further, we investigated the performance of individual scores exclusively, without considering the variability of the position. The score alone resulted in a 1.9% lower AUC level and a 2.2% lower AUPR level compared with PHACT ("Only Score" in fig. 4*C* and *D*).

Only PHACT Diversity Score

Similar to the PHACT score without diversity, the diversity information without considering the individual score of amino acids was also deficient for variant-effect prediction. To demonstrate this further, we presented the resulting performance of the algorithm that relied exclusively on the diversity of the position. The AUC and AUPR levels decreased by 1.2% and 1%, respectively ("Only Diversity" in fig. 4*C* and *D*). We noted that this diversity score was also computed using the tree traveling process and weighting approach of PHACT. The MSA-based diversity score was not as successful as the PHACT diversity term, with a 15.8% lower AUC and a 14.9% lower AUPR level, respectively ("Diversity" in fig. 3 vs. "Only Diversity" in fig. 4). Moreover, SIFT, which is another popular tool that employs position variability by considering the number of observed amino acids in the MSA, could not outperform PHACT in the ROC and PR curve comparisons, and other important metrics, such as MCC, F1 score, and balanced accuracy (the details can be found in the Comparisons with Benchmark Tools section). These results demonstrated the success of the diversity term, which is obtained from the individual PHACT scores of amino acids.

Equal Importance of Score and Diversity

This approach was based on the assignment of equal weight to the individual score per amino acid and the variability of the position. Although attributing equal importance to the score and diversity did not outperform PHACT, it yielded a similar performance ("Equal Importance" in fig. 4*C* and *D*). We also observed a 1% higher AUC and a 0.9% higher AUPR compared with the diversity score of PHACT ("Only Diversity" in fig. 4), thus illustrating the importance of the individual score of amino acids in variant-effect prediction.

Higher Importance of Diversity

In the final alternative approach, we examined the variations in the AUC and AUPR levels when a higher weight was assigned to diversity rather than to individual amino acid scores. The parameter λ in equation (3) was set as 0.9. Decreasing the importance of the individual amino acids scores resulted in a lower performance, with 0.941 AUC and 0.942 AUPR levels, compared with PHACT ("Higher Diversity" in fig. 4*C* and *D*).

Comparisons with Benchmark Tools

We compared PHACT with SIFT (Sim et al. 2012), PolyPhen-2 (Adzhubei et al. 2010), and several other statistical methods included in dbNSFP. SIFT utilizes MSA and defines the probability of an amino acid substitution based on the position diversity, in addition to the observed amino acids at the position in question. The scoring scheme of SIFT relies on the determination of the acceptable substitutions with the help of the physicochemical properties of the observed amino acids at a position of interest. The second benchmark tool, PolyPhen-2 (Adzhubei et al. 2010), computes a naive Bayes posterior probability for variants using various sequences and structural features. We obtained the precomputed scores for the benchmark algorithms from dbNSFP v4.1 (Liu et al. 2020). To avoid circularity and training biases, which result in overly optimistic predictive performances (Grimm et al. 2015), we eliminated the proteins that were used for training or parameter optimization of the predictive models from the test datasets. The quality and depth of MSAs highly affected the biological conclusions inferred from the conservation-based methods. Therefore, we also presented the SIFT scores computed using PHACT alignment. Although we constructed a bias-free dataset for PolyPhen-2 by eliminating its training set from our dataset, it is not feasible to build a bias-free dataset for all algorithms presented in dbNSFP. Most of the machine learning algorithms in dbNSFP are ensemble methods that use either the conservation-based features or the prediction scores obtained by other machine learning algorithms (Adzhubei et al. 2010; Carter et al. 2013; Schwarz et al. 2014; Dong et al. 2015; Ioannidis et al. 2016; Ionita-Laza et al. 2016; Feng 2017; Raimondi et al. 2017; Alirezaie et al. 2018; Rogers et al. 2018; Rentzsch et al. 2019). The training sets of these machine learning algorithms generally include variants annotated in UniProt (The UniProt Consortium 2021), ClinVar (Landrum et al. 2016), and VariBench (Sasidharan Nair and Vihinen 2013), which mainly overlap with our datasets. In addition, some of these algorithms utilize the allele frequency information obtained from various databases, such as those of gnomAD (Karczewski et al. 2020), ExAC (Lek et al. 2016),

Table 1. Summary of the Subdatasets Constructed Here From the HCG and Grimm Datasets to Evaluate the Discriminative Ability of PHACT and Benchmark Algorithms.

Subdataset	Main Dataset	Subset of	Eliminated Proteins	Number of Proteins	Number of Neutral Variants	Number of Pathogenic Variants
DS1	HCG	—	—	2,836	13,420	15,728
DS2	Grimm	—	—	2,325	13,401	30,412
DS3	HCG	DS1	PolyPhen-2 training set	645	2,234	1,401
DS4	Grimm	DS2	PolyPhen-2 training set	450	1,414	557
DS5	HCG	DS1	LIST-S2 optimization set	1,155	5,014	7,165
DS6	Grimm	DS2	LIST-S2 optimization set	1,023	5,859	11,871

and the 1000 Genomes Project (The 1000 Genomes Project Consortium 2015). The neutral variants included in our datasets were labeled based on the allele frequency information; thus, the training sets of these machine learning algorithms overlapped with our datasets to a large extent. Therefore, it was not possible to obtain a fair comparison with the machine learning algorithms in dbNSFP using the proteins for which we had already built the phylogenetic trees. As a result, we did not include these algorithms, with the exception of PolyPhen-2, in our benchmark comparisons. PolyPhen-2 is one of the most frequently used tools; therefore, comparing PHACT with it is of interest. Among the conventional statistical methods presented in dbNSFP, in addition to SIFT, LIST-S2 (Malhis et al. 2020) is also a notable pathogenicity scoring tool, particularly because it uses taxonomy distances for estimating the pathogenicity of missense mutations. To eliminate the problem of obtaining overly optimistic results for benchmarking as much as possible, we constructed subdatasets that did not include the proteins in the training set of PolyPhen-2 (Adzhubei et al. 2010) and the optimization set of LIST-S2. Table 1 summarizes the datasets used here.

PHACT Discriminates Pathogenic and Neutral Missense Variants

We built phylogenetic trees for 3,380 proteins and constructed our datasets using these proteins. Dataset 1 (DS1) and Dataset 2 (DS2) contained variants for proteins that could be mapped to the HCG and Grimm datasets, respectively. Each dataset was also mapped to dbNSFP, to obtain the scores of benchmark algorithms. We obtained the DS3 and DS4 datasets by discarding the proteins in the training set of PolyPhen-2 from DS1 and DS2, respectively. Similarly, DS5 and DS6 were constructed by eliminating the optimization set of LIST-S2 from DS1 and DS2, respectively. The datasets used in our study are available in the Supplementary Material online.

Figure 5 shows the AUC and AUPR values obtained by PHACT, SIFT, and SIFT using PHACT alignment, and PolyPhen-2 algorithms on the DS3 and DS4 datasets. We observed that PHACT achieved higher AUC and AUPR values on both datasets than the SIFT and PolyPhen-2 algorithms. Compared with the other algorithms, PHACT yielded higher TPRs for any FPR on both datasets (fig. 5A and C); moreover, it attained a higher precision for all or almost all TPRs on these datasets, respectively (fig. 5B and D). These results indicate that PHACT comfortably outperformed SIFT and PolyPhen-2 in predicting neutral and pathogenic variants. The comparison of PHACT with PolyPhen-2 and other statistical pathogenicity prediction tools included in dbNSFP on the DS4 dataset is provided in supplementary figure S3, Supplementary Material online. The success of PHACT can be explained by the utilization of a phylogenetic tree for scoring. The integration of evolutionary relatedness of the sequences enables PHACT to attribute less importance to the variations observed in distant species. Furthermore, the inclusion of two additional factors that are ignored in SIFT and PolyPhen-2 allowed PHACT to distinguish the neutral and pathogenic variants more accurately; these factors are the effect of (1) whether a substitution is observed at different time points independently and (2) whether a substitution at one ancestral species is the cause of multiple alterations.

Figure 6 shows the AUC and AUPR values of the PHACT and LIST-S2 algorithms on the DS5 and DS6 datasets. On the DS5 dataset, PHACT and LIST-S2 exhibited a similar performance regarding AUC and AUPR values. We observed a 0.2% improved performance with PHACT in ROC and a 0.2% improved performance with LIST-S2 in the PR curve comparison. Conversely, on the DS6 dataset, PHACT outperformed LIST-S2 by 1.4% at the AUC level and by 1% at the AUPR level. These results indicate that PHACT affords a slightly better or comparable performance against LIST-S2 for distinguishing the neutral variants from the pathogenic variants.

In table 2, we compared the performance of PHACT, SIFT, PolyPhen-2, and LIST-S2 with other known metrics, such as the F1 score, balanced accuracy, and MCC. The F1 score and balanced accuracy are well-known assessment metrics for binary classification problems. We also report MCC, which is a more reliable performance measure than the F1 score and accuracy for unbalanced datasets that considers true positives, false negatives, true negatives, and false positives proportionally to the class sizes (Chicco and Jurman 2020). The computation of all of these performance metrics requires the definition of a cutoff value that discriminates neutral from pathogenic variants. The cutoff value for PHACT scores was determined over DS1 because it includes more proteins and variants by maximizing the geometric mean of sensitivity and specificity. The substitutions with scores below and above this threshold (0.679) were assumed to be

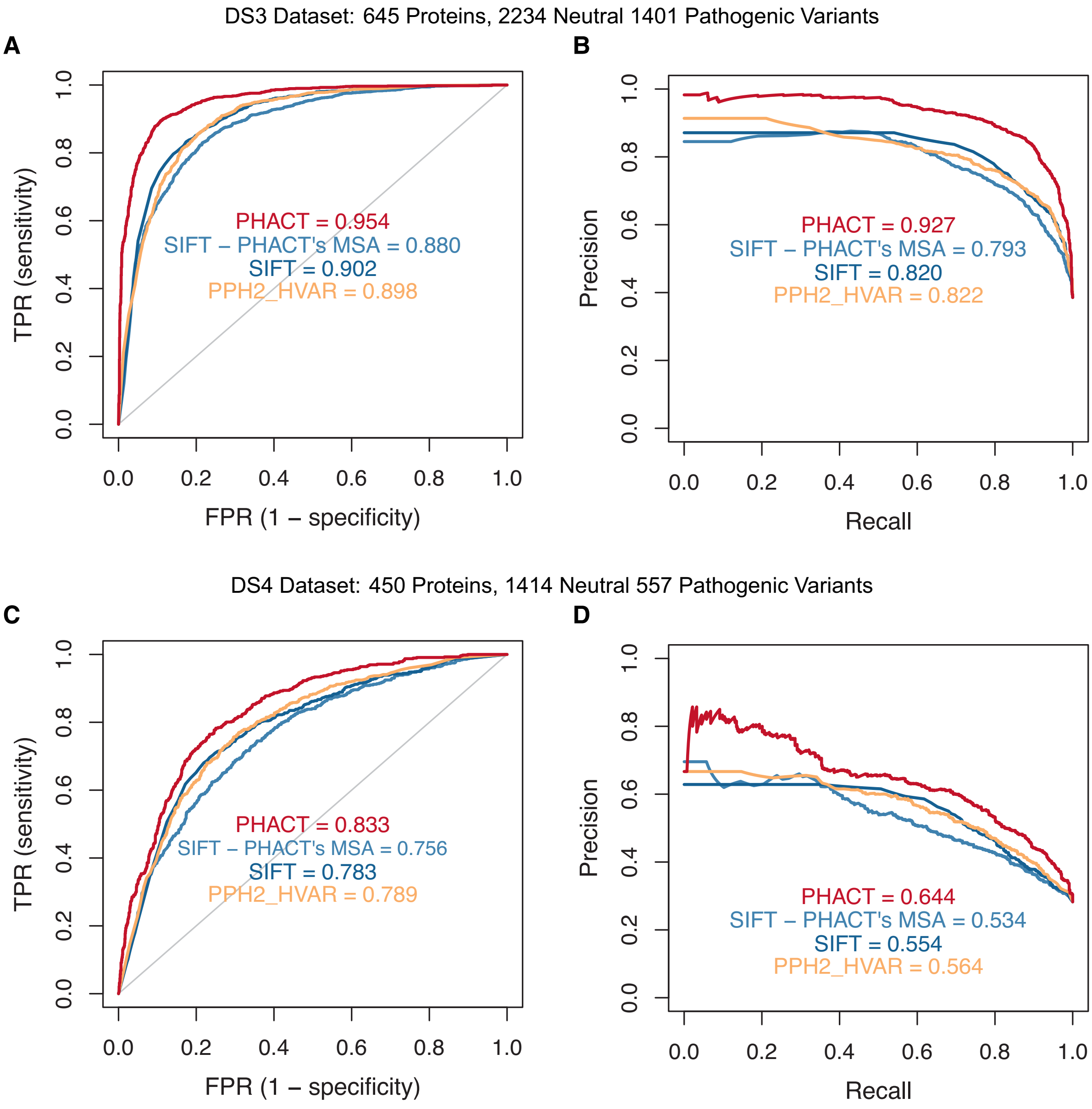

FIG. 5. Discriminative performance comparisons of PHACT, SIFT, and SIFT with PHACT's MSA and PolyPhen-2 (PPH2_HVAR) algorithms. (*A*) and (*B*): ROC and PR curves for the DS3 dataset (HCG dataset excluding the PolyPhen-2 training set). (*C*) and (*D*): ROC and PR curves for the DS4 dataset (Grimm dataset excluding the PolyPhen-2 training set).

pathogenic and neutral, respectively. The cutoff value is set at 0.05 for SIFT, 0.493 for PolyPhen-2, and 0.85 for LIST-S2, as reported in related studies (Adzhubei et al. 2010; Sim et al. 2012; Malhis et al. 2020). The resulting performance metrics are given in table 2. These results illustrate the fact that PHACT outperformed the tools in comparison with significant differences over almost all metrics including MCC, F1 score, and balanced accuracy.

The cutoff value for PHACT was determined over the largest dataset, DS1, which also includes DS3 and DS5. To avoid a potential bias of the cutoff value of PHACT, which was derived from the dataset used to assess comparative performances, we present the results based on the best performing cutoff values derived from the dataset in the comparison of all algorithms. The resulting performance measures are shown in table 3. PHACT achieved the best performance regarding MCC, F1 score, and balanced accuracy. Moreover, PHACT performed better than LIST-S2 regarding MCC and balanced accuracy, and exhibited a similar performance for the F1 score. These results underscore the advantage of using phylogenetic information. Although we computed a cutoff value for all tools by maximizing the geometric mean of sensitivity and specificity for table 3, a known alternative approach is to pick the threshold by maximizing the F1 score, the harmonic mean of sensitivity and specificity (Lipton et al. 2014). Therefore, we also calculated the performance metrics when the cutoff value that maximized the F1 score was

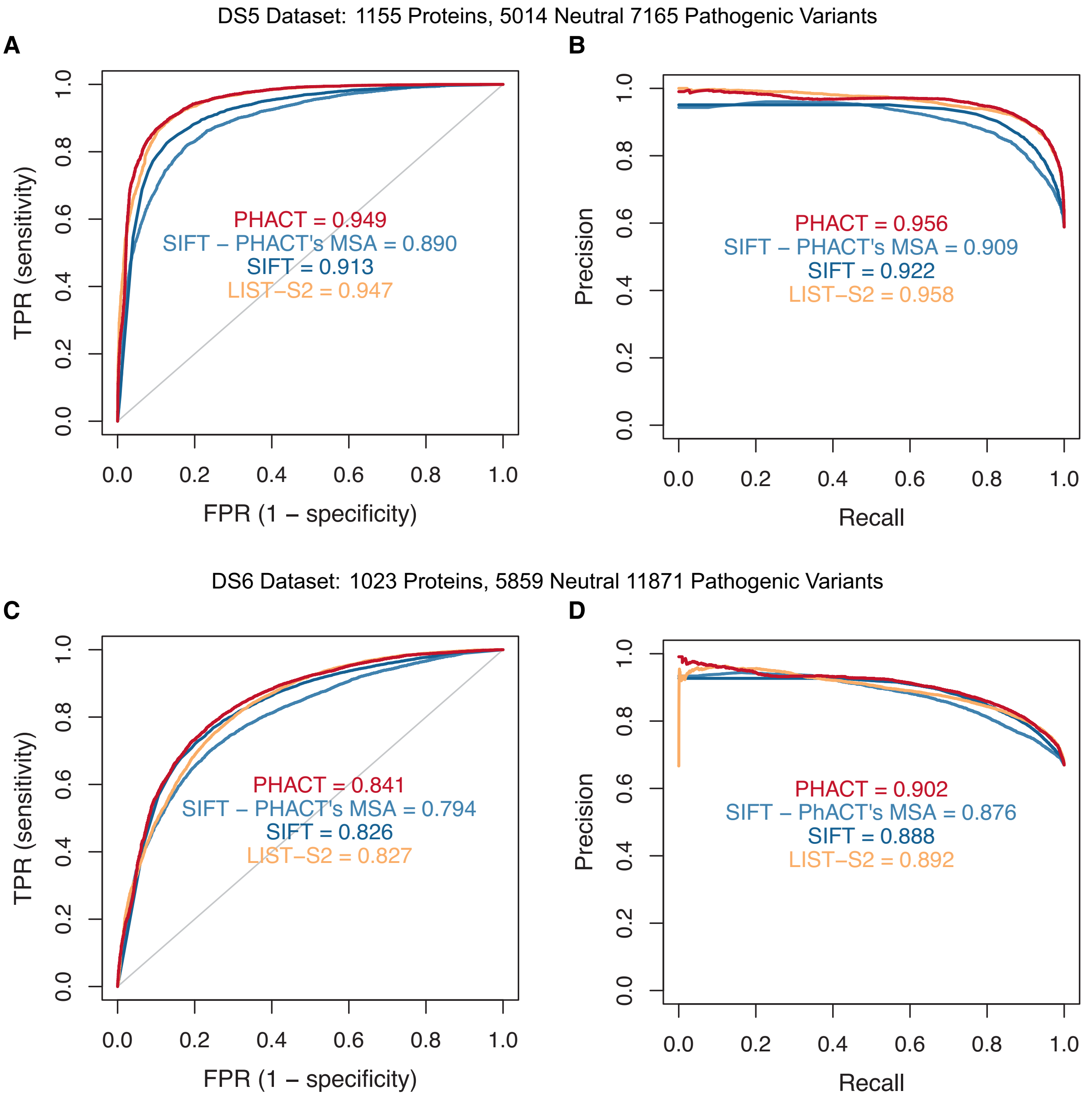

Fig. 6. Discriminative performance comparisons of PHACT, SIFT, and SIFT with PHACT's MSA and LIST-S2 algorithms. (*A*) and (*B*): ROC and PR curves for the DS5 dataset (HCG dataset excluding the LIST-S2 optimization set). (*C*) and (*D*): ROC and PR curves for the DS6 dataset (Grimm dataset excluding the LIST-S2 optimization set).

selected (supplementary table 1, Supplementary Material online). Additionally, true positive, false negative, true negative, and false positive values used for calculating the performance metrics in tables 2 and 3 are reported in supplementary table 2, Supplementary Material online.

Figure 7 and supplementary figures S3 and S4, Supplementary Material online compare the discriminative performance of PHACT with those of the other statistical pathogenicity prediction tools included in dbNSFP over the DS5, DS4, and DS6 datasets, respectively. On DS5, PHACT outperformed 18 conservation-based pathogenicity scoring methods by attaining higher AUC and AUPR values and exhibited an equal performance with LIST-S2 (fig. 7). On DS4 and DS6, PHACT performed better than all the conservation-based pathogenicity scoring methods, with higher AUC and AUPR levels (supplementary figs. S3 and S4, Supplementary Material online). LIST-S2 yielded comparable results against PHACT on DS5 in terms of the F1 score. However, PHACT outperformed LIST-S2 on DS6 in terms of the F1 score, MCC, and balanced accuracy, and showed a slightly better performance in terms of MCC and balanced accuracy on the DS5 dataset (tables 2 and 3). We also compared the ROC and PR curve performances of the two approaches when LIST-S2 scores were computed using PHACT MSA on both DS1 and a subset of DS1, when the variants in the optimization set of LIST-S2 were eliminated (supplementary fig. S5, Supplementary Material online). The results indicated that PHACT outperformed LIST-S2 with 1.4% higher AUC and AUPR levels on DS1.

It should be noted that LIST-S2 also utilizes a taxonomy tree as an evolutionary relationship of species, which likely accounts for the high performance of LIST-S2 (see more details of LIST-S2 in Discussion). These results emphasize the importance of the incorporation of evolutionary relationships between species and the effect of different mutational patterns into the final model. We significantly improved the AUC and AUPR values by traversing the phylogenetic tree and weighting approaches over frequency (fig. 3). The observation of a better or comparable performance against the known conservation-based tools supports our idea of expanding the analysis, rather than making predictions based on MSA alone.

In the final analyses, we compared the performance of tools in terms of the execution time over 50 proteins (see supplementary table 3, Supplementary Material online for details). We measured the time for the score computation step of the tools alone; the time required for obtaining MSA, phylogenetic tree, and the training process of PolyPhen-2 were ignored because they are all one-time processes. For a fair comparison, we computed PHACT, SIFT, PolyPhen-2, and LIST-S2 scores using the same MSAs used by PHACT and ran all these algorithms on the same computing system. The average execution times for PHACT, SIFT, PolyPhen-2, and LIST-S2 were 0.006, 0.0002, 0.166, and 0.0004 sec. per position, respectively, with a standard deviation of 0.005, 0.00006, 0.142, and 0.0003, respectively. We noted that the alignment format of LIST-S2 includes extra information, such as the number of common and different amino acids between proteins and the query sequence; because we provided the alignment file, these computations were not included in the reported time for LIST-S2. These results indicate that all of these tools can be run within seconds. Although the PHACT algorithm involves traveling through the nodes of the given phylogenetic tree, the computation time of the tools was within a comparable time scale.

Table 2. Comparison of Various Metrics Against SIFT, PolyPhen-2, and LIST-S2.

Subdataset	Algorithm	MCC	F1 Score	Balanced Accuracy
DS3	PHACT	**0.760**	**0.856**	**0.887**
	SIFT (PHACT's MSA)	0.596	0.761	0.805
	SIFT	0.596	0.761	0.804
	PPH2	0.630	0.782	0.823
DS4	PHACT	**0.459**	**0.628**	**0.754**
	SIFT (PHACT's MSA)	0.350	0.556	0.689
	SIFT	0.365	0.570	0.702
	PPH2	0.415	0.601	0.728
DS5	PHACT	**0.761**	**0.899**	**0.883**
	SIFT (PHACT's MSA)	0.628	0.843	0.816
	SIFT	0.664	0.870	0.819
	LIST-S2	0.754	0.896	0.879
DS6	PHACT	**0.515**	0.821	**0.768**
	SIFT (PHACT MSA)	0.435	0.772	0.730
	SIFT	0.485	**0.835**	0.739
	LIST-S2	0.481	0.809	0.750

Note.—The cutoff values for SIFT, PolyPhen-2, and LIST-S2 were set as 0.05, 0.493, and 0.85, respectively. The cutoff for PHACT was computed over DS1 as 0.679. The highest score is indicated in bold font in the table.

Table 3. Comparison of Various Metrics Against SIFT, PolyPhen-2, and LIST-S2.

Subdataset	Algorithm	MCC	F1 Score	Balanced Accuracy
DS3	PHACT	**0.778**	**0.866**	**0.893**
	SIFT (PHACT's MSA)	0.600	0.763	0.807
	SIFT	0.648	0.789	0.829
	PPH2	0.636	0.783	0.824
DS4	PHACT	**0.489**	**0.647**	**0.766**
	SIFT (PHACT's MSA)	0.355	0.562	0.694
	SIFT	0.428	0.606	0.730
	PPH2	0.419	0.603	0.730
DS5	PHACT	**0.760**	**0.898**	**0.883**
	SIFT (PHACT's MSA)	0.633	0.842	0.820
	SIFT	0.690	0.865	0.849
	LIST-S2	0.757	**0.898**	0.880
DS6	PHACT	**0.515**	**0.817**	**0.769**
	SIFT (PHACT MSA)	0.435	0.778	0.729
	SIFT	0.499	0.813	0.760
	LIST-S2	0.479	0.803	0.751

Note.—The cutoff values for all algorithms were determined over the corresponding dataset. The highest score is indicated in bold font in the table.

Discussion

In this study, we propose a novel phylogenetic tree-based missense mutation scoring approach, PHACT, to discriminate the pathogenic and neutral missense variants. We tested PHACT on two datasets composed of Humsavar (The UniProt Consortium 2021), ClinVar (Landrum et al. 2016), gnomAD (Karczewski et al. 2020), and Grimm circularity datasets (Grimm et al. 2015). Our comparative analyses showed that PHACT afforded a better predictive performance compared with SIFT (Sim et al. 2012) and PolyPhen-2 (Adzhubei et al. 2010), which are tools that are widely adopted to prioritize variants in clinical studies. This improvement is noteworthy because PHACT, which is a phylogenetic tree-based evolutionary conservation scoring method, outperformed a conservation-based method that was developed using evolutionary relatedness calculated directly from MSA (i.e., SIFT; Sim et al. 2012), as well as a machine learning algorithm that uses structural and physicochemical features in addition to the conservational measures (i.e., PolyPhen-2; Adzhubei et al. 2010).

Additional comparisons with SIFT (Sim et al. 2012), PolyPhen-2 (Adzhubei et al. 2010), and LIST-S2 (Malhis et al. 2020) based on the F1 score, MCC, and balanced accuracy showed that PHACT can achieve better predictive performances than the benchmark algorithms over different datasets (tables 2 and 3, supplementary table 1 Supplementary Material online). PHACT outperformed SIFT and PolyPhen-2 because those algorithms utilize MSA and cannot fully consider the evolutionary relationship between species. As depicted in figure 3, the use of MSA alone is not a suitable measure for pathogenicity prediction, even when the distance between species is considered ("Weighted Frequency" and "Weighted Freq. [Gaps

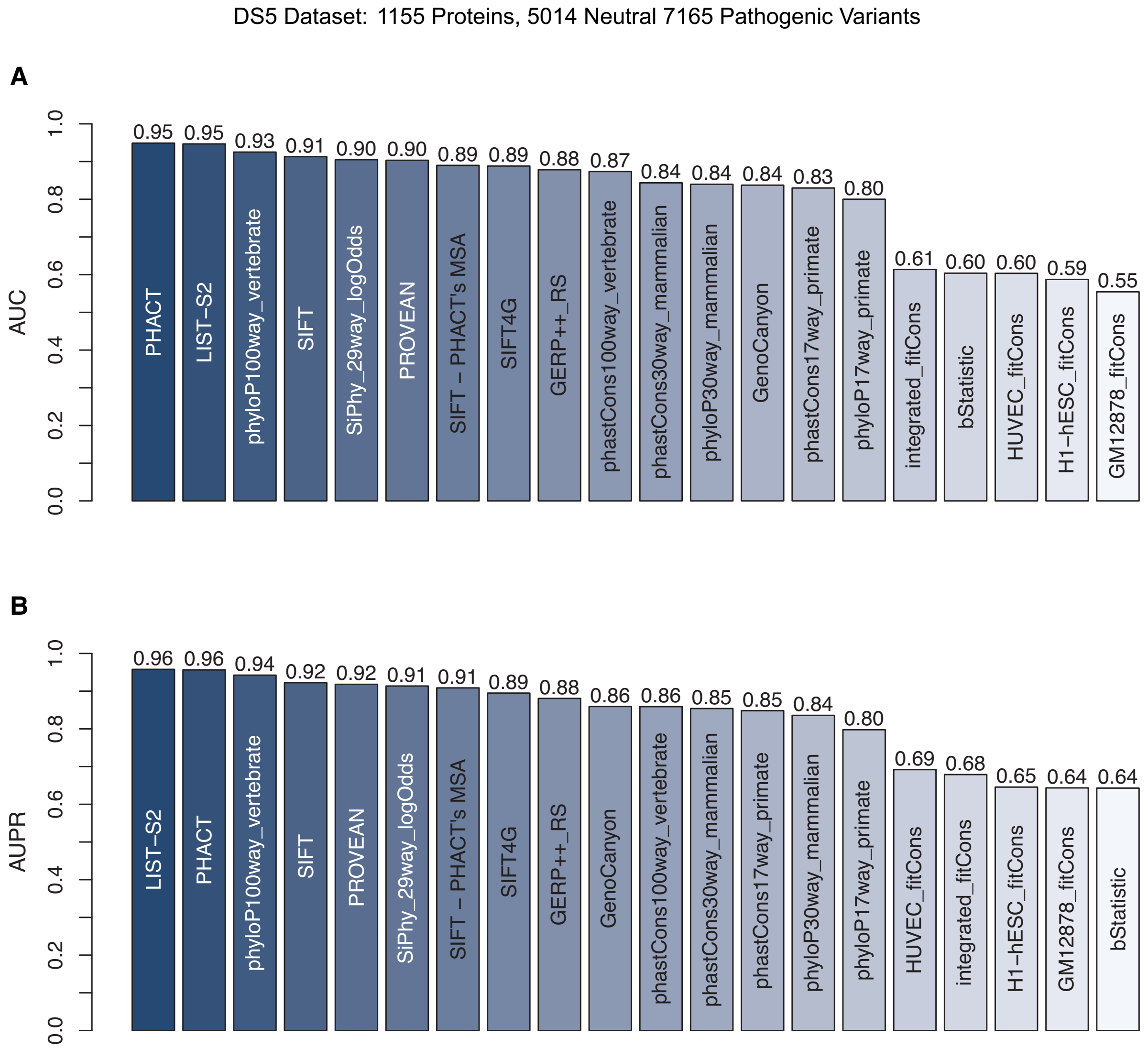

Fig. 7. AUC and AUPR comparisons of PHACT against the statistical pathogenicity prediction algorithms presented in dbNSFP (LIST-S2, Malhis et al. 2020; phyloP, Pollard et al. 2010; SIFT, Sim et al. 2012; PROVEAN, Choi et al. 2012; SiPhy, Garber et al. 2009; SIFT4G, Vaser et al. 2016; GERP, Davydov et al. 2010; GenoCanyon, Lu et al. 2015; phastCons, Siepel et al. 2005; fitCons, Gulko et al. 2015; and bStatistic, McVicker et al. 2009 on the DS5 [HCG] dataset).

Excluded]" in fig. 3). We observed that both SIFT and PolyPhen-2 tended to mislabel (1) the neutral variants that are observed at a species close to the query sequence but are rare in terms of frequency and (2) the pathogenic variants that are detected at multiple species because of a single mutation in their ancestral species. The sample positions that were mislabeled by SIFT and PolyPhen-2 are given in supplementary figure S1, Supplementary Material online. Conversely, we observed a similar performance for LIST-S2 and PHACT over various subsets. Similar to PHACT, LIST-S2 also considers the distance between species and the vulnerability of the position to the substitutions. We believe that a similar performance observed between these two approaches resulted from these similarities, as the distance between species and the diversity of the position are two important factors that are predictive of the consequences of substitutions. However, LIST-S2 relies on BLASTP (Camacho et al. 2009) pairwise alignments, which mainly align the sequences with respect to the query sequence and are not suitable for constructing phylogenetic trees; in turn, LIST-S2 also utilizes a taxonomy tree to compute the number of edges between species. Although taxonomy trees represent the classification of species, they explain only the general groups that are obtained using the similarities of species. Based on the data presented in this study, we suggest that phylogenetic trees are a more reliable input for pathogenicity prediction tools because they are constructed by considering the unique evolutionary history of the genes in question.

The success of PHACT against the existing statistical pathogenicity prediction methods is also presented in figure 7 and supplementary figures S3 and S4, Supplementary Material online. Unlike PHACT, these methods compute the evolutionary conservation directly from MSA. Here, we illustrated the success of our approach, which exploits the information stemming from the phylogenetic tree through the higher predictive performances attained in analyses.

Although PHACT yielded a higher accuracy than the existing conservation-based methods, it can be further improved by considering other evolutionary events, such as epistasis and coevolution. Here, PHACT computed a score per substitution assuming the independence of the sequence positions. However, it is known that the coevolving positions in a clade can be tolerated when they change together. This type of relationship between positions could fundamentally affect the resulting prediction. Although a few tools (EVmutation, Hopf et al. 2017; GEMME, Laine et al. 2019; and DeepSequence, Riesselman et al. 2018) exist that incorporate the coevolution between residues, whereas predicting the functional consequences of missense mutations, they utilize the frequency and conservation level of amino acids to determine the coevolving positions by ignoring the evolutionary history. As a future direction, we aim to incorporate the coevolution of positions into PHACT. The second future research avenue pertains to the consideration of paralog information to determine the weights of ancestral nodes. Gene duplication is one of the major mechanisms in the implementation of new functions (Long et al. 2003). After gene duplication, one of the two paralogs might accumulate more mutations and diverge from the other one (Ohno 1970). Adding the functionally diverged sequences to the MSA and the tree would render amino acid substitutions that are intolerable in the functionally diverged lineages tolerable. In the current version of our algorithm, we do not consider this discrimination, but cover the divergence of one copy by assigning a lower weight to that node with the help of our weight function. In future studies, we aim to include the duplication process in PHACT and decrease the importance of the second copy by detecting the one that conserves the ancestral function.

Based on the evidence that the inclusion of conservation-based features in machine learning algorithms for pathogenicity prediction improves their predictive performance (Adzhubei et al. 2010; Carter et al. 2013; Schwarz et al. 2014; Dong et al. 2015; Ioannidis et al. 2016; Ionita-Laza et al. 2016; Feng 2017; Raimondi et al. 2017; Alirezaie et al. 2018; Rogers et al. 2018; Rentzsch et al. 2019), we plan to develop a supervised machine learning algorithm based on both PHACT scores and features derived from the phylogenetic tree in the future. By doing so, we aim to increase the discriminative ability of our approach for the pathogenic and neutral missense variants. Finally, we plan to develop a web server on which users can access the prediction scores and phylogenetic tree information for a given protein.

Materials and Methods

Datasets

The reference human protein sequences used in the present study were obtained from the UniProtKB/Swiss-Prot Knowledgebase database (The UniProt Consortium 2021) which was released in April 2019 (Release 2019_04). Two different datasets, including missense variants, were constructed to assess the performance of our algorithm.

Disease-related variants of the first dataset were obtained from the UniProtKB/Swiss-Prot Knowledgebase (The UniProt Consortium 2021) and ClinVar database (Landrum et al. 2016). The former is a high-quality, manually annotated and reviewed protein sequence database. All missense variants annotated in the human UniProtKB/Swiss-Prot entries are listed in the Humsavar dataset (https://www.uniprot.org/downloads). We used the Humsavar dataset released in February 2022 to extract the variants reportedly associated with diseases (i.e., pathogenic variants). ClinVar is a freely available archive of interpretations of the clinical significance of variants in reported conditions (https://www.ncbi.nlm.nih.gov/clinvar/). We downloaded the ClinVar database released in February 2022 to obtain the disease-associated variants. In the present study, we selected ClinVar's germline missense variants with clinical significance as pathogenic or likely to be pathogenic. ClinVar provides a review status for each entry indicating the trustworthiness of assertions. The review status values range from 0 to 4 (i.e., 0: lowest level; 4: highest level). We excluded the entries scoring 0 to obtain a more reliable dataset. Due to the complex underlying mechanisms of cancer-related diseases, in the present study, we did not consider the pathogenic variants associated with the cancer types listed on the National Cancer Institute's website (https://www.cancer.gov/types). We generated the neutral variants of the first dataset by extracting the missense mutations that had alternate allele frequencies above 0.01 from the Genome Aggregation Database (gnomAD v3.1) (Karczewski et al. 2020). We merged the pathogenic variants selected from the Humsavar and ClinVar datasets and extracted the neutral variants from the gnomAD database to form our first dataset (referred to as the HCG dataset). We excluded the variants that had conflicting clinical significance (i.e., the ones labeled as pathogenic in one dataset, but labeled as neutral in another dataset) from the HCG dataset. Before the elimination of variants, we matched variants from each dataset with their corresponding chromosome coordinates to map each of them to our reference protein sequence database. To match the variants, we benefited from the homo_sapiens_variation.txt.gz file from the UniProtKB/Swiss-Prot Knowledgebase database, which lists variants identified on protein isoforms. Through this approach, we could include the variants identified on protein isoforms in the ClinVar and gnomAD datasets into our variant dataset, using the corresponding amino acid positions of these variants in our reference amino acid sequences.

To further test our algorithm, we aggregated five publicly available datasets presented in the Grimm circularity dataset (Grimm et al. 2015) that we obtained from VariBench (Sasidharan Nair and Vihinen 2013). We obtained all the corresponding variants for the given chromosome coordinates and amino acid substitutions in the Grimm dataset. We included the ones that matched the reference protein sequence database of our study and eliminated the variants with conflicting class labels from the combined dataset (named as the Grimm dataset).

In this study, we used dbNSFP v4.1 (Liu et al. 2020) to obtain the precomputed ranked scores of benchmark algorithms. Building a phylogenetic tree is a time-consuming process. To include as many proteins as possible in our research, we built phylogenetic trees starting with proteins containing the highest number of variants in the HCG dataset. We performed our analyses on 3,023 proteins included in the HCG and Grimm datasets (see Supplementary Material online for the list of proteins used in this study).

BLAST and MSA

The homologs of each query sequence were searched through the PSI–BLAST (Altschul et al. 1997) against a nonredundant database of 14,010,480 proteins obtained from the reference proteomes in the UniProtKB/Swiss-Prot Knowledgebase (The UniProt Consortium 2021). We performed two PSI–BLAST iterations with 5,000 maximum target sequences. Due to computational limitations of building the phylogenetic trees, we limited the hits to 1,000 sequences with a minimum identity of 30% and *E*-value of 0.00001. The sequences were aligned using MAFFT FFTNS (Katoh and Standley 2013), and the MSA was trimmed with the trimAl tool gappyout method (Capella-Gutierrez et al. 2009).

Maximum-Likelihood Phylogenetic Tree

The resulting MSA was used to generate a maximum-likelihood phylogenetic tree through the RaxML-NG tool (Kozlov et al. 2019) via the LG4X model, leaving the remaining parameters at default settings.

Ancestral Reconstruction

Positions with a "gap" character in the query sequence were removed from the original MSA (without trimming). The resulting MSA was used to perform ancestral sequence reconstructions using the RaxML-NG tool (Kozlov et al. 2019) via the LG4X model, maintaining the remaining parameters at default settings.

Workflow Engine and High-Performance Computing

Large-scale data analysis involving the chained execution of many command-line applications requires a workflow engine that helps to automate human-readable pipeline runs and ensure reproducibility. This study effectively used a workflow management system tool, referred to as Snakemake Field (Koster and Rahmann 2012), which offers high-performance computing cluster-level scalability, to perform reproducible and scalable data analysis. In total, 330K CPU hours were consumed for 3,380 proteins, and almost 2 TB of data were created during the long-term analyses.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

This work was supported by the Health Institutes of Turkey (TUSEB) (Project no: 4587 to O.A.) and EMBO Installation Grant (Project no: 4163 to O.A.) funded by the Scientific and Technological Research Council of Turkey (TÜBİTAK). Most of the numerical calculations reported in this paper were performed at the High Performance and Grid Computing Center (TRUBA resources) of TÜBITAK. The TOSUN cluster at Sabanci University was also used for computational analyses. We also want to thank Nehircan Özdemir for his art illustration of the PHACT approach.

Data Availability

The PHACT source code and the entire data generated in this study are available at https://github.com/CompGenomeLab/PHACT.

References

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526**:68–74.

The UniProt Consortium. 2021. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*. **49**:D480–D489.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Methods*. **7**: 248–249.

Alirezaie N, Kernohan KD, Hartley T, Majewski J, Hocking TD. 2018. ClinPred: prediction tool to identify disease-relevant nonsynonymous single-nucleotide variants. *Am J Hum Genet*. **103**:474–483.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. **25**:3389–3402.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinform*. **10**:421.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**:1972–1973.

Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. 2013. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* **14**:1–16.

Castellana S, Mazza T. 2013. Congruency in the prediction of pathogenic missense mutations: state-of-the-art web-based tools. *Brief Bioinform*. **14**:448–459.

Chicco D, Jurman G. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**:6.

Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. 2012. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* **7**:e46688.

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol*. **6**: e1001025.

Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, Liu X. 2015. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet*. **24**:2125–2137.

Eilbeck K, Quinlan A, Yandell M. 2017. Settling the score: variant prioritization and Mendelian disease. *Nat Rev Genet*. **18**:599–612.

Feng BJ. 2017. PERCH: a unified framework for disease gene prioritization. *Hum Mutat*. **38**:243–251.

Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. 2009. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**:i54–i62.

Grimm DG, Azencott CA, Aicheler F, Gieraths U, MacArthur DG, Samocha KE, Cooper DN, Stenson PD, Daly MJ, Smoller JW, *et al.* 2015. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum Mutat*. **36**:513–523.

Gulko B, Hubisz MJ, Gronau I, Siepel A. 2015. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat Genet*. **47**:276–283.

Hopf TA, Ingraham JB, Poelwijk FJ, Scharfe CP, Springer M, Sander C, Marks DS. 2017. Mutation effects predicted from sequence co-variation. *Nat Biotechnol*. **35**:128–135.

Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D, *et al.* 2016. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am J Hum Genet*. **99**:877–885.

Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. 2016. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet*. **48**:214–220.

Jiang T, Fang L, Wang K. 2021. MutFormer: a context-dependent transformer-based model to predict pathogenic missense mutations. arXiv preprint arXiv:2110.14746.

Jordan DM, Ramensky VE, Sunyaev SR. 2010. Human allelic variation: perspective from protein function, structure, and evolution. *Curr Opin Struct Biol*. **20**:342–350.

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, *et al.* 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**:434–443.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. **30**:772–780.

Koster J, Rahmann S. 2012. Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics* **28**:2520–2522.

Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**: 4453–4455.

Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc*. **4**:1073–1081.

Kwok P-Y, Gu Z. 1999. Single nucleotide polymorphism libraries: why and how are we building them? *Mol Med Today*. **5**:538–543.

Laine E, Karami Y, Carbone A. 2019. GEMME: a simple and fast global epistatic model predicting mutational effects. *Mol Biol Evol*. **36**: 2604–2619.

Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, *et al.* 2016. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res*. **44**:D862–D868.

Le SQ, Dang CC, Gascuel O. 2012. Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Mol Biol Evol*. **29**:2921–2936.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, *et al.* 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**:285–291.

Lipton ZC, Elkan C, Naryanaswamy B. 2014. Optimal thresholding of classifiers to maximize F1 measure. *Mach Learn Knowl Discov Databases*. **8725**:225–239.

Liu X, Li C, Mou C, Dong Y, Tu Y. 2020. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med*. **12**:103.

Long M, Betrán E, Thornton K, Wang W. 2003. The origin of new genes: glimpses from the young and old. *Nat Rev Genet*. **4**:865–875.

Lu Q, Hu Y, Sun J, Cheng Y, Cheung K-H, Zhao H. 2015. A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Sci Rep*. **5**:1–13.

Malhis N, Jacobson M, Jones SJ, Gsponer J. 2020. LIST-S2: taxonomy based sorting of deleterious missense mutations across species. *Nucleic Acids Res*. **48**:W154–W161.

McVicker G, Gordon D, Davis C, Green P. 2009. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet*. **5**:e1000471.

Ohno S. 1970. *Evolution by gene duplication*. Heidelberg (Berlin): Springer-Verlag.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*. **20**:110–121.

Qi H, Zhang H, Zhao Y, Chen C, Long JJ, Chung WK, Guan Y, Shen Y. 2021. MVP predicts the pathogenicity of missense variants by deep learning. *Nat Commun*. **12**:1–9.

Raimondi D, Tanyalcin I, Ferte J, Gazzo A, Orlando G, Lenaerts T, Rooman M, Vranken W. 2017. DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. *Nucleic Acids Res*. **45**:W201–W206.

Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. 2019. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res*. **47**:D886–D894.

Riesselman AJ, Ingraham JB, Marks DS. 2018. Deep generative models of genetic variation capture the effects of mutations. *Nat Methods*. **15**:816–822.

Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR, Campbell C. 2018. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics* **34**:511–513.

Sasidharan Nair P, Vihinen M. 2013. VariBench: a benchmark database for variations. *Hum Mutat*. **34**:42–49.

Schwarz JM, Cooper DN, Schuelke M, Seelow D. 2014. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods*. **11**:361–362.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*. **15**:1034–1050.

Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. 2012. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res*. **40**:W452–W457.

Sunyaev S, Ramensky V, Bork P. 2000. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet*. **16**:198–200.

Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. 2016. SIFT missense predictions for genomes. *Nat Protoc*. **11**:1–9.

Yang Z. 2006. *Computational molecular evolution*. Oxford: Oxford University Press.