Bioinformatics, 32(14), 2016, 2230–2232
doi: 10.1093/bioinformatics/btw222
Advance Access Publication Date: 18 May 2016
Applications Note

OXFORD

Genetics and population analysis

PANTHER-PSEP: predicting disease-causing genetic variants using position-specific evolutionary preservation

Haiming Tang and Paul D. Thomas*

Division of Bioinformatics, Department of Preventive Medicine, University of Southern California, Los Angeles, CA 90033, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on October 7, 2015; revised on March 3, 2016; accepted on April 18, 2016

Abstract

Summary: PANTHER-PSEP is a new software tool for predicting non-synonymous genetic variants that may play a causal role in human disease. Several previous variant pathogenicity prediction methods have been proposed that quantify evolutionary conservation among homologous proteins from different organisms. PANTHER-PSEP employs a related but distinct metric based on 'evolutionary preservation': homologous proteins are used to reconstruct the likely sequences of ancestral proteins at nodes in a phylogenetic tree, and the history of each amino acid can be traced back in time from its current state to estimate how long that state has been preserved in its ancestors. Here, we describe the PSEP tool, and assess its performance on standard benchmarks for distinguishing disease-associated from neutral variation in humans. On these benchmarks, PSEP outperforms not only previous tools that utilize evolutionary conservation, but also several highly used tools that include multiple other sources of information as well. For predicting pathogenic human variants, the trace back of course starts with a human 'reference' protein sequence, but the PSEP tool can also be applied to predicting deleterious or pathogenic variants in reference proteins from any of the ∼100 other species in the PANTHER database.

Availability and implementation: PANTHER-PSEP is freely available on the web at http://pantherdb.org/tools/csnpScoreForm.jsp. Users can also download the command-line based tool at ftp://ftp.pantherdb.org/cSNP_analysis/PSEP/.

Contact: pdthomas@usc.edu

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

1 Introduction

Recent advances in sequencing technology have increased the rate of genetic variation discoveries, thus exacerbating the problem of interpreting these genetic variations. Here, we focus on non-synonymous single nucleotide variants (NSVs), a type of genetic variation that results in a single amino acid substitution in an encoded protein. Most computational tools for predicting the functional consequences of NSVs, including SIFT (Ng and Henikoff, 2003), PolyPhen (Adzhubei *et al.*, 2010; Ramensky *et al.*, 2002) and PANTHER-PSEC (Thomas *et al.*, 2003), rely at least in part on the assumption that crucial amino acids for protein families are conserved through evolution due to negative selection (i.e. amino acid changes at these sites were *deleterious* in the past), and that mutations at these sites have an increased likelihood of being *pathogenic* (causing disease) in humans. In general, homologous sequences of a target protein are collected and aligned, and a metric of conservation is computed based on the weighted frequencies of different amino acids observed in the target position in the alignment.

Here, we present a new tool, PANTHER-PSEP (position-specific evolutionary preservation) for predicting potentially pathogenic or deleterious NSVs. PANTHER-PSEP is different from other variant prediction tools, in that PSEP uses a metric based on evolutionary *preservation* (Marini *et al.*, 2010). Preservation is related to conservation, as it also reflects the effect of negative selection that has acted to prevent evolutionary change at a given site in a protein. It is also related to earlier methods that utilize a phylogenetic tree in addition to a multiple sequence alignment, pioneered by the Evolutionary Trace method (Lichtarge *et al.*, 1996) and later applied to variant pathogenicity prediction in the PANTHER-PSECv2 (Thomas and Kejariwal, 2004) and Evolutionary Action (Katsonis and Lichtarge, 2014) methods. But unlike these other approaches, preservation uses *ancestral sequence reconstruction* to assess evolutionary change specifically in the *lineage* leading to the sequence of interest. We have previously shown evolutionary preservation to be superior to conservation for identifying deleterious mutations (in a growth-based assay) in the human Mthfr protein, as well as the *E.coli* lacI protein (Marini *et al.*, 2010). Here, we describe a software tool based on the same concept of ancestral preservation, but using a simple, standardized metric for quantifying preservation. In PANTHER-PSEP, preservation is measured as the approximate length of *time* a site has been preserved, tracing back the lineage from the relevant human protein, or potentially any other protein in the PANTHER library (Mi *et al.*, 2013). Because PANTHER trees are reconciled with the known species tree, we can use best estimates of speciation times (Hedges *et al.*, 2006) to date the ancestral preservation of each site.

2 Methods

2.1 Trees, alignments and ancestral sequence reconstruction

Sequence alignments and phylogenetic trees were obtained from the PANTHER database (Mi *et al.*, 2013), version 9.0. For each family in PANTHER, we reconstructed the ancestral sequences for all common ancestors (internal nodes) in the tree using PAML (Yang, 1997), with the PANTHER tree and alignment as input. The reconstruction was performed at the amino acid level, using default parameters and the WAG substitution model (Whelan and Goldman, 2001).

2.2 Input and output

Input is a reference protein sequence and one or more single-site substitutions in that sequence. The input protein sequence is searched against the PANTHER sequences using BLASTP (Camacho *et al.*, 2009) to find the best-matching PANTHER sequence (tree leaf node). The target amino acid is traced from the leaf back through increasingly older ancestral proteins in the tree. This trace stops if the target amino acid is different from the corresponding amino acid in the ancestral sequence, or if the amino acid reconstruction probability is less than a predefined threshold, and the age (in millions of years) of the last preserved ancestor is reported. The reconstruction probability threshold is the only free parameter in the PSEP method. We determined the optimal threshold on the HumVar dataset (Capriotti *et al.*, 2006), a commonly used dataset for pathogenic human variant prediction methods, but we find that the choice of threshold has very little effect on the performance of PSEP (Supplemental Fig. S1).

2.3 Testing and comparing to other methods

We validated PANTHER-PSEP on the SwissVarSelected dataset (Grimm *et al.*, 2015; Mottaz *et al.*, 2010), which was recently shown to be the most appropriate set for this purpose as it avoids major issues that can confound comparisons between different methods. This dataset is independent of the HumVar dataset (as well as other datasets that are commonly used for training and parameterization of NSV prediction methods), and is well balanced in representing pathogenic and benign variants across a variety of proteins (Grimm *et al.*, 2015). Using SwissVarSelected, we can compare PANTHER-PSEP with similar, sequence alignment-based methods as well as with 'combined' methods that include multiple sources of information in addition to sequence alignments (Fig. 1 and Supplemental Fig. S2). Finally, in order to facilitate comparisons with previously published results, we also assessed PANTHER-PSEP on the HumVar, SwissVar (Capriotti *et al.*, 2013; Mottaz *et al.*, 2010) and VariBench (Schwarz *et al.*, 2010) datasets (Supplemental Figs S3–S5; Supplemental Table S1).

3 Implementation

PANTHER-PSEP is implemented both as a standalone tool and a web server. The software tool is written in Perl. The web server (available at http://pantherdb.org/tools/csnpScoreForm.jsp) uses the same code base as the standalone tool, with an interface written in Java and HTML.

4 Results and conclusions

Figure 1 shows that on the SwissVarSelected dataset, PANTHER-PSEP performs better than previously published sequence alignment-based methods. This performance is somewhat surprising, especially as PSEP does not use any prior information commonly used in evolutionary conservation methods, such as amino acid chemical similarity. The superior performance of PSEP suggests that preservation may be a more informative way to use homologous sequence data than conservation. Perhaps even more surprisingly, PSEP outperforms many trained, multi-feature prediction

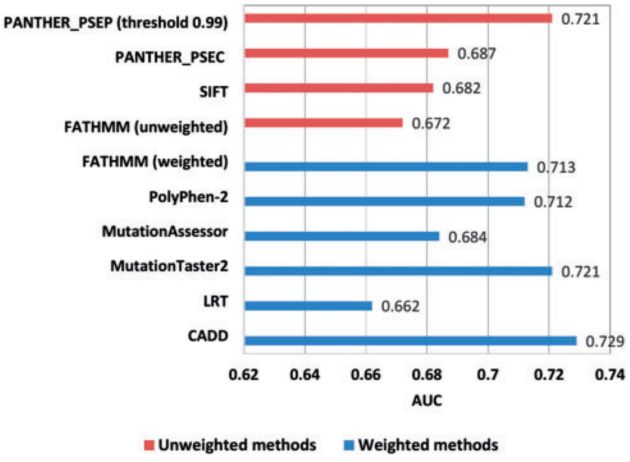

Fig.1. Comparison of the performance of PANTHER-PSEP with other tools for predicting pathogenic variants, on the SwissVarSelected dataset. The most closely comparable tools are shown in red; these are all 'unweighted' tools that, like PSEP, use only multiple sequence alignment information (PANTHER-PSEC, SIFT and FATHMM). Other tools are shown in blue; these are weighted methods that use multiple additional features (FATHMM, PolyPhen-2, MutationAssessor, MutationTaster2, LRT and CADD). Data are from (Grimm *et al.*, 2015), except for PSEP and PSEC. Bar lengths show the area under the ROC curve for each method. Full ROC curves for all tools are plotted in Supplemental Figure S2

methods. The high performance is also reflected in comparisons on the HumVar and SwissVar datasets; performance on VariBench is similar to current sequence-based methods, though the significance of this is unclear due to biases recently uncovered in this benchmark (Grimm *et al.*, 2015). PSEP scores are correlated with, but in many cases are complementary to, scores from CADD (Kircher *et al.*, 2014) (Supplemental Fig. S6). Taken together, these results suggest that PANTHER-PSEP could be a useful addition to multi-feature prediction methods such as CADD, PolyPhen-2 (Adzhubei *et al.*, 2010), MutPred (Li *et al.*, 2009) and SNPS&GO (Calabrese *et al.*, 2009). Finally, we emphasize that the metric for preservation employed here is an extremely simple one. It considers only preservation of the exact amino acid, and therefore chemically conservative substitutions are treated the same as radical changes. It also reports a preservation time, rather than considering how this metric might scale for lineages other than human, that have different generation times and population histories. Future developments could consider more sophisticated metrics based on the same evolutionary model.

Acknowledgements

We thank Dr. Huaiyu Mi for help with the PANTHER data, Perl programming and providing scripts to access the USC high-performance computing center.

Conflict of Interest: none declared.

References

Adzhubei,I.A. *et al.* (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.

Calabrese,R. *et al.* (2009) Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum. Mutat.*, **30**, 1237–1244.

Camacho,C. *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Capriotti,E. *et al.* (2006) Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics (Oxford, England)*, **22**, 2729–2734.

Capriotti,E. *et al.* (2013) Collective judgment predicts disease-associated single nucleotide variants. *BMC Genomics*, **14**, S2.

Grimm,D.G. *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.

Hedges,S.B. *et al.* (2006) TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics (Oxford, England)*, **22**, 2971–2972.

Katsonis,P. and Lichtarge,O. (2014) A formal perturbation equation between genotype and phenotype determines the Evolutionary Action of protein-coding variations on fitness. *Genome Res.*, **24**, 2050–2058.

Kircher,M. *et al.* (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.

Li,B. *et al.* (2009) Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics (Oxford, England)*, **25**, 2744–2750.

Lichtarge,O. *et al.* (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.

Marini,N.J. *et al.* (2010) The use of orthologous sequences to predict the impact of amino acid substitutions on protein function. *PLoS Genet.*, **6**, e1000968.

Mi,H. *et al.* (2013) PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.*, **41**, D377–D386.

Mottaz,A. *et al.* (2010) Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. *Bioinformatics (Oxford, England)*, **26**, 851–852.

Ng,P.C. and Henikoff,S. (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.

Ramensky,V. *et al.* (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.*, **30**, 3894–3900.

Schwarz,J. *et al.* (2010) MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods*, **7**, 575–576.

Thomas,P.D. *et al.* (2003) PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.*, **13**, 2129–2141.

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

Whelan,S. and Goldman,N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.*, **18**, 691–699.

Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.: CABIOS*, **13**, 555–556.