~~IMPROVING THE *IN SILICO* ASSESSMENT OF PATHOGENICITY FOR COMPENSATED MUTATION~~S

Improving the *in silico* assessment of pathogenicity for compensated mutations

Luisa Azevedo[1,2,3*], Matthew Mort[4], Antonio C Costa[5], Raquel M. Silva[6], Dulce Quelhas[7,8], Antonio Amorim[1,2,3], David N. Cooper[4]

[1]Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Population Genetics and Evolution Group, Rua Alfredo Allen 208, 4200-135 Porto, Portugal
[2]IPATIMUP-Institute of Molecular Pathology and Immunology, University of Porto, Rua Júlio Amaral de Carvalho 45, 4200-135 Porto, Portugal
[3]Department of Biology, Faculty of Sciences, University of Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal
[4]Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, UK
[5]Instituto Superior de Engenharia do Porto, GECAD-ISEP Research Group, Rua Dr. Bernardino de Almeida 431, 4249-015 Porto, Portugal
[6]Department of Medical Sciences, iBiMED & IEETA, University of Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal
[7]Biochemical Genetics Unit, Centro de Genética Médica Jacinto de Magalhães, Centro Hospitalar do Porto, Porto, Portugal
[8]Unit for Multidisciplinary Research in Biomedicine (UMIB), Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto (ICBAS/UP), Porto, Portugal

[LUISA: some of the Portuguese addresses are in Portuguese, some in English. Shouldn't we adopt one or the other style?]

Running title: Pathogenicity prediction of compensated mutations

Conflict of interest: The authors are unaware of any conflicts of interests.

* To whom all correspondence should be addressed:
Dr Luisa Azevedo

Formatted:

Formatted:

Formatted:

I3S-Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Population Genetics and Evolution Group,
Rua Alfredo Allen 208, 4200-135 Porto, Portugal

ABSTRACT

Understanding the functional *sequelae* of ~~missense~~ amino acid replacements is of fundamental importance in medical genetics. Perhaps the most intuitive way to assess the potential pathogenicity of a given human missense mutation is by measuring the degree of evolutionary conservation of the substituted amino acid residue, a feature which ~~should in theory be~~generally serves as a good proxy metric for the functional/structural importance of that residue. However, the presence of ~~putative~~ compensated mutations as the wild-type alleles in orthologous proteins of other mammalian species not only challenges this classical view of amino acid essentiality but also precludes the accurate evaluation of the functional impact of this ~~ese~~ type of missense mutation~~s~~ using ~~existing~~ currently available bioinformatic prediction tools. Compensated mutations constitute at least 4% of all known missense mutations causing human inherited disease and hence represent an important potential source of error in that they may often be misclassified as benign variants. The consequent under-reporting of compensated mutations is exacerbated in the context

of next generation sequencing where their inappropriate exclusion constitutes an unfortunate natural consequence of the filtering and prioritization of the very large numbers of variants generated. Here we demonstrate the reduced performance of currently available pathogenicity prediction tools when applied to compensated mutations and ~~then~~ propose an alternative machine learning approach to assess likely pathogenicity for this particular type of mutation.

KEY-WORDS: Compensated pathogenic mutations, CPDs, human inherited disease, prediction tools, protein evolutionary conservation.

INTRODUCTION

It is almost a truism that the degree of evolutionary conservation of an amino acid residue reflects the structural and/or functional importance of that residue. Thus, when a missense mutation occurs in an amino acid residue that is highly conserved evolutionarily, purifying selection tends to act so as to prevent the deleterious allele from attaining a high frequency in the population. It is not unreasonable to expect that mutations which are disease-causing in human would also be deleterious in evolutionarily closely related

species. However, this expectation has been challenged by the realization that there are numerous examples where human mutant alleles correspond to the wild-type alleles in other mammalian species [1-7]. Such mutations have become known as *compensated pathogenic deviation*s (CPDs) following their original designation [5], because it is assumed that the apparently benign nature of these missense mutations in non-human species is due to the co-existence of other amino acid substitutions that compensates for their otherwise pathogenic ~~outcome~~consequences. Among those human mutant residues corresponding to the wild-type residue in mouse [5] are an Ala53Thr substitution at the α-synuclein (*SNCA*) locus reported to be associated with familial Parkinson disease [8]; the BRCA1-Leu892Ser and BRCA2-Val211Ala mutations associated with breast cancer; ADA-Arg142Gln causing severe combined immunodeficiency, and two mutations in the cystic fibrosis transmembrane conductance regulator (CFTR-Phe87Leu and CFTR-Val754Met) gene underlying cystic fibrosis. Another interesting CPD is EIF2B5-Arg113His, the most common lesion associated with leukoencephalopathy with vanishing white matter [9] which corresponds to the wild-type allele in the genomes of both rat and mouse [4].

When the genome sequence of the rhesus macaque was released [3], further examples of CPDs were identified. Among them were

Formatted:
Field Code
Formatted:
Formatted:
Formatted:

Tyr356His and Ile164Thr at the phenylalanine hydroxylase (PAH) responsible for the most common human inborn error of metabolism (phenylketonuria) and Arg40His at the X-linked ornithine transcarbamylase (*OTC*) locus. Although the OTC-Arg40His replacement leads to the cytosolic degradation of the human enzyme precursor [10], abnormal levels of ammonia were not evident in simian plasma. Moreover, no abnormal levels of phenylalanine were detected in macaque [3] reinforcing the notion that these mutations are only deleterious on the human genetic background. Intriguingly, a different OTC mutation (Thr125Met) associated with fatal hyperammonemia [11] was found to correspond to the wild-type allele in chimpanzees [1,12,13]. Among the CPD mutations identified through a comparison with the recently reported mountain gorilla genome [6] were NPC1-Asn961Ser that leads to Niemann-Pick disease C and GPAM-Thr447Met associated with Complex I deficiency.

Finally, two mutations associated with ciliopathies, the BBS4-Asn165His and ~~the~~ RPGRIP1L-Arg937Leu substitutions associated with Bardet-Biedl and Meckel-Gruber syndromes, respectively, constitute the wild-type alleles in the genomes of several vertebrates [7]. The same study reported a *de novo* mutation (BTG2-Val141Met) in which the disease-associated variant

(Met141) corresponded to the wild-type allele in more than 50 vertebrate species.

At this stage, it is important to mention that these are examples of a more general phenomenon that ~~affects~~ involves ~~about~~ at least 4% of all known missense mutations causing human inherited disease (see Materials and Methods section). In practice, this proportion almost certainly represents a conservative estimate because there is an intrinsic ascertainment bias against the recognition and reporting of CPDs because such mutations, by their very nature, are often predicted to be benign. This bias is exacerbated in the context of next generation sequencing (NGS) where the inappropriate exclusion of CPDs from further consideration constitutes an unfortunate natural consequence of the filtering and prioritization of the very large numbers of variants generated. It follows that we need to take urgent steps both to assess the scale of this problem and to take appropriate remedial action.

~~Therefore, i~~It is ~~important~~ obviously vital that bioinformatic predictionve tools ~~can~~ make accurate predictions when trying to assess the pathogenic impact of a putative compensated mutation. ~~Because s~~Some of the existing tools have already ~~proven to~~ been shown to ~~fail~~ make incorrect~~the~~ predictions in the case of ~~few~~ experimentally validated pathogenic mutations [7]~~.~~. Here, we demonstrate, in a larger sample of CPDs, the~~ir~~ reduced

performance of existing tools to predict the deleterious impact of CPDs when found as disease-associated variants in humans. Further, we present the prototype of a CPD-specific predictor that successfully outperformed currently available tools in terms of the prediction of the deleterious impact of these mutations in humans.

MATERIALS AND METHODS

Detection of CPDs in mammalian species

To identify the amino acid positions where a deleterious human mutation constitutes the wild-type residue in a non-human mammalian species, herein referred to as a CPD [5], we employed mutation data from the Human Gene Mutation Database (HGMD; http://www.hgmd.org; 52,765 disease-causing missense mutations, annotated as DMs; July 2013) to screen 1-to-1 orthologous mammalian protein sequences annotated at the Ensembl Genome Browser (http://www.ensembl.org; release 73) [14]. Data from a total of 39 mammalian species (35 placental, 3 marsupials and 1 monotreme) were available and sequences were automatically retrieved and submitted through a series of stringent filters before being considered for subsequent analysis. Whenever more than one sequence matched the gene symbol used in the search (due to redundancy caused by for example, alternatively spliced isoforms),

a pairwise alignment with the human reference sequence was performed and the sequence retained was that which had the highest degree of identity with its human counterpart. In order to avoid highly incomplete sequences, we calculated the pairwise identity to the human sequence and retained only those that were at least 50% identical to the human ortholog. The orthologous sequences that passed these filters were then used in the identification of CPDs by comparing the aligned sequences. For all the cases where a CPD was identified, we applied a strategy similar to that previously documented [5] viz. screening a flanking window of five amino acid residues upstream and downstream of the putative CPD site and retaining only those CPDs with no more than four differences with respect to the human sequence within the flanking region. Missense mutations at the initiator methionine residue were removed from the analysis. This strategy allowed the identification of 1,964 CPDs in a total of 684 protein-coding genes [LUISA: Are you going to make these data available online? I think we may be required to do this. If not, you should maybe say 'data available on request'], a figure which corresponds to about 3.7% of all missense mutations analysed. All the alignments were performed using ClustalO 1.2.0 (www.clustal.org/omega) [15] and other tools were developed locally using GNU/Linux-based computers with scripting tools + C language.

DM set

A set of 10,211 disease-causing missense mutations which result in an amino acid substitution (AAS) in 2,030 genes was obtained from the Human Gene Mutation Database (HGMD) [16]. This set of disease-causing mutations is representative of datasets that have typically been employed in the training and evaluation of a number of different bioinformatic prediction tools to identify disease-causing AAS. In order to allow an unbiased evaluation of bioinformatic prediction tools, an unseen test set (not used for training) should always be used, otherwise the evaluation represents in-sample error rather than out-of-sample error and hence is likely to be overly optimistic in terms of prediction performance. As some of the prediction tools evaluated here (e.g. MetaSVM) either already used, or could have used, HGMD (or other similar overlapping datasets e.g. OMIM) either as training data or in the development of the prediction method, this DM set was selected so as to contain only recently reported (2014 onwards) disease-causing missense mutations from HGMD. This dataset of ~~disease-causing~~ mutations would therefore not ~~therefore~~ have been available when the various prediction tools being tested were being developed. This set of disease-causing mutations is henceforth referred to as the DM set.

SNP set

As a negative control, –a set of 2,174 putatively 'neutral' common missense SNPs (MAF ≥ 0.4) in 1,640 genes from the *NHLBI ESP6500 Exomes (http://evs.gs.washington.edu/EVS/)* was compiled. This set is henceforth referred to as the SNP set.

Performance evaluation of bioinformatic prediction tools used to identify disease-causing mutations

Numerous different prediction methods to identify disease-causing or functional variants have been developed. The four selected here (SIFT [17], Mutation Assessor [18], PROVEAN [19] and MetaSVM [20]) were selected on the grounds that they are commonly used (e.g. SIFT) or represent a different approach to the classification problem [e.g. MetaSVM, an ensemble approach using population frequency data and the scores from 10 other prediction methods (SIFT, PolyPhen-2 HDIV, PolyPhen-2 HVAR, GERP++, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy, PhyloP)]. Prediction scores for the CPD, DM and SNP sets (where available) were then obtained from dbNSFP (version 3.2c) [20]. As a means to evaluate performance for each of the four bioinformatic prediction methods, receiver operating characteristic (ROC) curves

and the area under the ROC curve (AUC) were calculated for the CPD set versus the SNP set and the DM set versus the SNP set.

Feature subset ranking

In order to evaluate discriminative features or attributes (commonly used to identify disease-causing mutations) in the context of CPDs, an array of features relating to the AAS (e.g. solvent accessibility) were derived from SNVBox [21] (Table 1). Related features were then grouped into seven different subsets (Amino acid based, Exon based, Genomic MSA, Protein MSA, Protein structure, Regional protein composition and annotated functional sites) (Table S1). For more details, please refer to SNVBox [21]. The prediction performance of each feature subset (e.g. Eexon based features) was evaluated using cross-validation and a linear support vector machine for two different datasets (CPD&SNP, DM&SNP). This allowed us to rank how informative each feature subset was for identifying CPDs and also for identifying disease-causing mutations.

A machine learning approach to assess the pathogenicity of potential CPDs

Current methods to assess the pathogenicity of potential disease-causing mutations have not been specifically developed or evaluated in the context of CPDs. Although they do demonstrate some utility in the functional assessment of CPDs, we set out to develop a novel prototype CPD-specific predictor. Using the features employed in this study (Table S1), two different Random Forest classification models were built: the first, termed the 'CPD trained model', was trained using the CPD (positive examples) and SNP (negative examples) sets. The second model, termed the 'DM trained model', was trained using the DM set as positive examples and the SNP set as negative examples. The DM trained model also excludes any features derived from multiple sequence alignments (MSA). The area under the ROC curve (AUC) was then calculated for each model (CPD trained model and DM trained model). We also employed standard benchmarking statistics (Table S2) to evaluate performance such as the true positive rate (sensitivity), the false positive rate and the Matthew's Correlation Coefficient (MCC) [22]. The MCC was employed since it represents one of the best available measures of prediction quality. It returns a value between −1 and +1; a coefficient of −1 represents the worst possible prediction, 0 a random prediction and +1 a perfect prediction.

RESULTS AND DISCUSSION

How accurately are CPDs predicted by current bioinformatics tools?

~~Due~~ Owing to their occurrence in multiple sequence alignments (MSA), there may well be a tendency for CPDs to evade detection by the predictive tools commonly used to evaluate the functional impact of human missense variants, simply because the mutant residues in question are tolerated in other species. Most predictive methods rely to some extent on the degree of evolutionary conservation of the mutated residue but, when a mutant residue occurs as the wild-type allele in one or more orthologs, its pathogenicity in a human context may not be readily predictable. Thus, for example using three widely used predictive tools, PolyPhen [23], SIFT [17] and MutationAssessor [24], Jordan and colleagues [7]~~,~~ failed to predict the deleterious effect of three compensated mutations (BBS4-Asn165His, RPGRIP1L-Arg937Leu and BTG2-Val141Met) experimentally demonstrated to be pathogenic in human.

To assess the extent of the ability of existing predictive methods to deal with CPDs, we compared a set of more than 10,200 disease-causing mutations from the HGMD with a set of CPDs and a set of neutral missense SNPs (MAF≥0.4) (Figure 1). The results showed

that all four bionformatic prediction tools tested here exhibited reduced prediction performance (-24.2% to -8.5% AUC) in relation to CPDs as compared to disease-causing mutations. Of the four tools analysed, SIFT exhibited the largest reduction in prediction performance (-24.2%). MetaSVM, a consensus of scores from 10 different tools (SIFT, PolyPhen-2 HDIV, PolyPhen-2 HVAR, GERP++, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy, PhyloP) and the maximum frequency observed in the 1000 Genomes populations, outperformed all other tools in terms of identifying CPDs, suggesting that an ensemble approach could be useful in classifying CPDs.

Evaluation of features commonly used to identify disease-causing AAS

In order to perform an evaluation in the context of CPDs, the discriminatory power of groups of related features (e.g. structural features) used to distinguish disease-causing missense mutations from neutral polymorphic missense variants (Table S1) was computed using 10-fold cross-validation with a linear support vector machine (Figure 2). The most informative feature subsets discriminating disease-causing mutations from common putatively neutral polymorphisms (SNP set) were derived from MSA (95% AUC for genomic MSA and 81% AUC for protein MSA; Figure 2 - DM set). These MSA-derived features (genomic and protein

MSA) also demonstrated the largest reduction in performance (-20% in AUC) when used to discriminate between the CPDs and the SNP set (as compared to the DM versus the SNP set; Figure 2 - CPD set).

A machine learning approach to evaluate potential CPDs

We next developed two different machine learning approaches to assess the pathogenicity of potential CPDs using the features indicated in Table S1 (see Materials and Methods section). The CPD trained model (AUC=86.21%) outperformed the DM trained model (with no MSA features) by over 10% based on the AUC (Table S2, Figure 3). Both models shared similar false positive rates with the DM trained model exhibiting reduced sensitivity (-23.2%) as compared to the CPD trained model. In the context of identifying CPDs, the CPD model developed here outperformed all existing bioinformatic prediction methods evaluated in this study, indicating utility in developing a specific CPD predictor.

In conclusion, the scale of the compensated mutation phenomenon is such that a significant proportion of pathogenic human missense mutations (a minimum of 3.7%) are found as the wild-type allele in at least one of the other mammalian species ~~here~~ analysed here. In terms of evaluating the pathological impact of these mutations,

traditional approaches suffer from the serious drawback of relying upon the evolutionary conservation score between homologous proteins irrespective of the influence of genetic variation at other amino acid positions. Here we demonstrate the poor performance of established mutation prediction tools to assess the pathological significance of CPDs and show that the development of new tools should yield an increase in prediction accuracy by avoiding the information provided by the MSA features. The *in silico* assessment of pathogenesis for novel CPDs ~~that result from~~ identified by whole exome/genome NGS studies currently requires a different protocol from that employed for the bulk of non-compensated mutations. We propose a two-stage analysis whereby whole exome/genome data should first be screened for potential CPDs (using the strategy employed in this study and in others [5] to identify missense variants where the mutant amino acid represents the wild-type amino acid in another mammalian species). A method such as the novel CPD prediction tool developed here could then be applied to identify any high confidence CPD candidates for further analysis. Adoption of a CPD-specific protocol could help to ~~reduce~~ avoid the mis-classification of a sizable proportion of pathogenic missense mutations as benign.

ACKNOWLEDGEMENTS

IPATIMUP integrates the i3S Research Unit, which is partially supported by FCT, the Portuguese Foundation for Science and Technology. This work is funded by FEDER funds through the Operational Programme for Competitiveness Factors - COMPETE and National Funds through the FCT-Foundation for Science and Technology, under the projects "PEst-C/SAU/LA0003/2013". DNC and MM gratefully acknowledges financial support from Qiagen Inc through a License Agreement with Cardiff University.

REFERENCES

1. Azevedo L, Carneiro J, van Asch B, Moleirinho A, Pereira F, Amorim A. Epistatic interactions modulate the evolution of mammalian mitochondrial respiratory complex components. *BMC Genomics* 2009; **10:** 266.

2. Gao L, Zhang J. Why are some human disease-associated mutations fixed in mice? *Trends Genet* 2003; **19:** 678-681.

3. Gibbs RA, Rogers J, Katze MG *et al.* Evolutionary and biomedical insights from the rhesus macaque genome. *Science* 2007; **316:** 222-234.

4. Huang H, Winter EE, Wang H *et al.* Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes. *Genome Biol* 2004; **5:** R47.

5. Kondrashov AS, Sunyaev S, Kondrashov FA. Dobzhansky-Muller incompatibilities in protein evolution. *Proc Natl Acad Sci U S A* 2002; **99:** 14878-14883.

6. Xue Y, Prado-Martinez J, Sudmant PH *et al.* Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science* 2015; **348:** 242-245.

7. Jordan DM, Frangakis SG, Golzio C *et al.* Identification of *cis*-suppression of human disease mutations by comparative genomics. *Nature* 2015; **524:** 225-229.

8. Polymeropoulos MH, Lavedan C, Leroy E *et al.* Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. *Science* 1997; **276:** 2045-2047.

9. Gascon-Bayarri J, Campdelacreu J, Sanchez-Castaneda C *et al.* Leukoencephalopathy with vanishing white matter presenting with presenile dementia. *J Neurol Neurosurg Psychiatry* 2009; **80:** 810-811.

10. Mavinakere M, Morizono H, Shi D, Allewell NM, Tuchman M. The clinically variable R40H mutant ornithine carbamoyltransferase shows cytosolic degradation of the precursor protein in CHO cells. *J Inherit Metab Dis* 2001; **24:** 614-622.

11. Gilbert-Dussardier B, Segues B, Rozet JM *et al.* Partial duplication [dup. TCAC (178)] and novel point mutations (T125M, G188R, A209V, and H302L) of the ornithine transcarbamylase gene in congenital hyperammonemia. *Hum Mutat* 1996; **8:** 74-76.

Formatted:

12. Azevedo L, Suriano G, van Asch B, Harding RM, Amorim A. Epistatic interactions: how strong in disease and evolution? *Trends Genet* 2006; **22:** 581-585.

13. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 2005; **437:** 69-87.

14. Flicek P, Ahmed I, Amode MR *et al.* Ensembl 2013. *Nucleic Acids Res~~earch~~* 2012; **41**(Database issue): D48-55.

15. Sievers F, Wilm A, Dineen D *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol~~ecular~~ S~~s~~yst~~ems~~ B~~b~~iol~~ogy~~* 2011; **7:** 539-539.

16. Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet* 2014; **133:** 1-9.

17. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* 2003; **31:** 3812~~--~~3814.

18. Reva B, Antipin Y, Sander C. Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol* 2007; **8:** R232.

19. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* 2015; **31:** 2745-2747.

Formatted:

Formatted:

20. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: A oOne-sStop dDatabase of fFunctional pPredictions and aAnnotations for hHuman nNonsynonymous and sSplice-sSite SNVs. *Hum Mutat* 2016; **37:** 235-241.

21. Wong WC, Kim D, Carter H, Diekhans M, Ryan MC, Karchin R. CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. *Bioinformatics* 2011; **27:** 2147-2148.

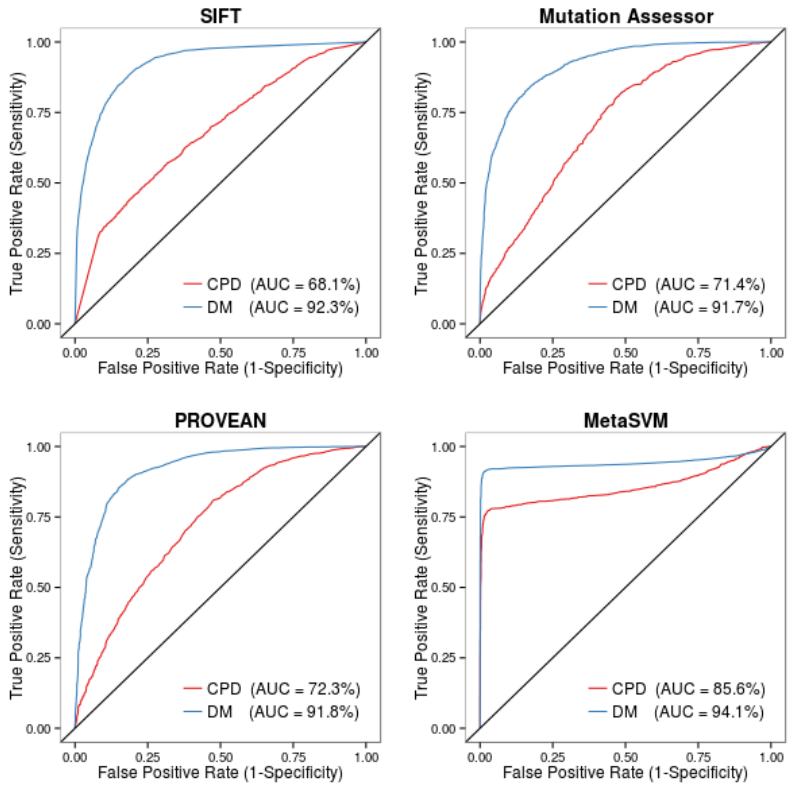
22. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim~~ica et Bb~~iophys~~ica~~ A~~a~~cta* 1975; **405:** 442-451.

23. Adzhubei IA, Schmidt S, Peshkin L *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* 2010; **7:** 248-249.

24. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* 2011; **39:** e118.

Figures

Figure 1. Performance evaluation of four different bioinformatic tools (SIFT, Mutation Assessor, PROVEAN and MetaSVM) used to identify disease-causing mutations. Each bioinformatic tool was evaluated on compensated pathogenic deviations (CPDs) and disease-causing mutations (DMs). Receiver operating characteristic (ROC) curves and the area under the ROC curve (AUC) were calculated for each mutation dataset (CPD & DM) with a different prediction tool. An AUC of 100% would represent a perfect predictor, whereas an AUC of 50% would represent a prediction tool making only random predictions (denoted by the black diagonal line). Note: MetaSVM is an ensemble prediction method based on 10 other prediction tools.



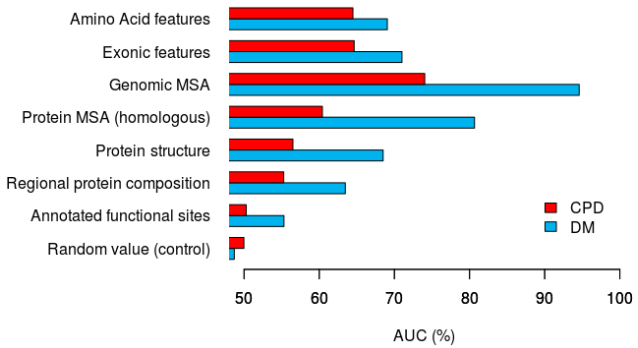


Figure 2. Evaluation of different groups of commonly used features (feature subset ranking) used to identify disease-causing amino acid substitutions (AAS). Each feature subset is then evaluated (10-fold cross-validation using a linear support vector machine) in the context of (i) discriminating between CPDs and common polymorphisms (CPD versus SNP set), (ii) discriminating between disease-causing mutations (DMs) and common polymorphisms (DM versus SNP set). The AUC of the ROC curve was then calculated for each feature subset; features employed in each subset are shown in Table 1. As a control, a random feature was computed for each training example and the AUC of the ROC calculated.

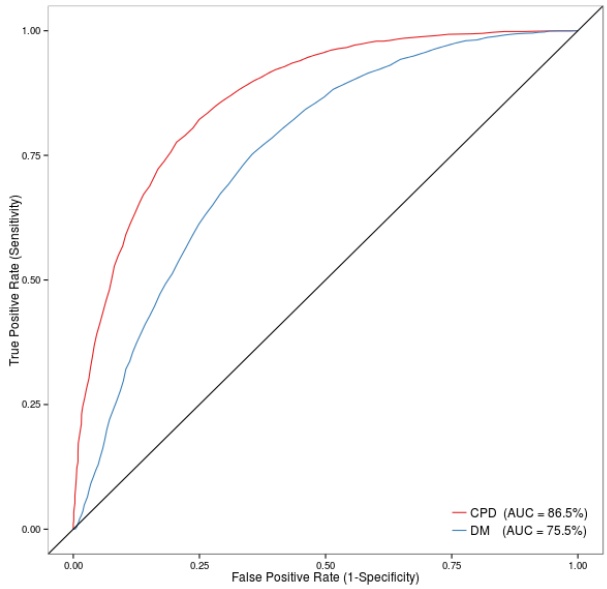


Figure 3. Performance evaluation for the identification of CPDs by means of a ROC curve using 10-fold cross-~~-~~validation of the two Random Forest prediction models developed in this study. The first model, termed the 'CPD trained model', was trained using the CPD and SNPs sets employed in this study (ROC shown in red). The second model, the 'DM trained model', was trained using disease-causing mutations (DM and SNP sets) but excludes any features derived from multiple sequence alignments (MSA). ROC shown in blue. The area under the ROC curve (AUC) was then calculated for each model (CPD trained model and DM trained model). An AUC of 100% would represent a perfect predictor, ~~whilst~~ whereas an AUC of 50% ~~represents~~ would correspond to a prediction tool making random predictions (represented by the black diagonal line).

Supplementary files:

Table S1. Summary of features commonly used used in bioinformatic tools to identify deleterious amino acid substitutions (AAS). Features were obtained using the SNVBox software for all datasets (CPD, DM and SNP sets). The descriptions of features were taken from the SNVBox User Manual (http://karchinlab.org/apps/snvbox/userdoc.pdf); for more details please refer to SNVBox [6]. [LUISA: I think the reference you mean is probably 21]

Feature name	Feature subset	Description
AA Hydrophobicity	Amino acid features	Change in hydrophobicity as a result of the AAS
AA Charge	Amino acid features	Change in formal charge as a result of the AAS
AA Volume	Amino acid features	Change in residue volume as a result of the AAS (cubic Angstroms)
AA Polarity	Amino acid features	Polarity change as a result of the AAS
AA Matrix	Amino acid features	Amino acid substitution scores from BLOSUM 62, PAM250, EX, Venkatarajan and Braun matrix & Miyazawa-Jernigan contact energy matrix

AA Transition	Amino acid features	Frequency of transition between two neighbouring amino acids based on all human proteins in SwissProt
AA Grantham score	Amino acid features	The Grantham distance from reference to mutation amino acid residue
AA Frequencies	Amino acid features	Frequency of AAS type (e.g. alanine to glycine) in HGMD (2003), HapMap (dbSNP build 129) and COSMIC (release 38)
Exon Conservation	Exonic features	Entire exon conservation computed from a 46-way genomic vertebrate alignment
Exon SNP Density	Exonic features	Number of HapMap verified SNPs in the exon where the mutation is located divided by the length of the exon
Genomic multiple sequence alignments (MSA)	Genomic MSA	Features calculated from 46-way genomic vertebrate alignments, which includes Shannon entropy and the Kullback-Leibler divergence
Protein multiple sequence alignments (MSA)	Protein MSA	Features calculated from multiple sequence alignment of diverse homologous proteins. Features computed include the Shannon entropy and Kullback-Leibler divergence

Solvent accessibility	Protein structure	Prediction that wild-type residue is buried, partially buried or exposed in terms of solvent accessibility
Secondary structure	Protein structure	Prediction that wild-type residue is helix, loop or strand
Protein stability	Protein structure	Prediction of the degree to which the wild-type residue contributes to protein stability e.g. highly stabilizing
Backbone flexibility	Protein structure	Prediction of the flexibility of the backbone of the wild-type residue
Protein composition	Regional protein composition	Features based on regional amino acid composition in a 15-amino-acid-residue window centred on the AAS
UniProt annotations of human proteins	Annotated functional sites	Includes functional sites annotated by UniProt, including binding sites (e.g. DNA, RNA, lipid, metal, carbohydrate, calcium), catalytic sites, sites of post-translational modifications, localization signals, disulphide bonds, protein-protein interaction sites

Table S2. Performance benchmarks for the identification of CPDs based on the two machine learning models developed in this study. The first model, termed the 'CPD trained model', was trained using the sets of CPDs and common SNPs employed in this study (CPD and SNP sets). The second model, the 'DM trained model' was trained using disease-causing mutations and common SNPs (DM and SNP sets) but excludes any features derived from multiple sequence alignments (MSA). The Random Forest machine-learning algorithm was employed, and evaluation was performed using a variation of 10-fold cross-validation whereby the positive evaluation set in each fold comprised unseen examples from the CPD set for both models (DM trained model and CPD trained model). TPR = true positive rate (sensitivity). FPR = false positive rate. MCC = Matthew's Correlation Coefficient; an MCC of −1 represents the worst possible prediction, 0 a random prediction and +1 a perfect prediction.

Dataset	TPR	FPR	MCC	AUC of ROC
CPD trained model	78.05	21.75	0.56	86.21
DM trained model (no MSA features)	54.84	21.98	0.34	75.10