Bioinformatics, 2016, 1–9
doi: 10.1093/bioinformatics/btw362
Advance Access Publication Date: 17 June 2016
Original Paper

OXFORD

Structural bioinformatics

The structural effects of mutations can aid in differential phenotype prediction of beta-myosin heavy chain (Myosin-7) missense variants

Nouf S. Al-Numair[1], Luis Lopes[2], Petros Syrris[2], Lorenzo Monserrat[3], Perry Elliott[2] and Andrew C. R. Martin[1,*]

[1]Institute of Structural and Molecular Biology, Division of Biosciences, University College London, London WC1E 6BT, UK, [2]Institute of Cardiovascular Science, UCL, London, UK and [3]Complejo Hospitalario Universitario de A Coruña, Insituto de Investigación Biomédica, Coruña, Spain

*To whom correspondence should be addressed.

Associate Editor: Anna Tramontano

Received on January 7, 2016; revised on May 3, 2016; accepted on June 6, 2016

Abstract

Motivation: High-throughput sequencing platforms are increasingly used to screen patients with genetic disease for pathogenic mutations, but prediction of the effects of mutations remains challenging. Previously we developed SAAPdap (Single Amino Acid Polymorphism Data Analysis Pipeline) and SAAPpred (Single Amino Acid Polymorphism Predictor) that use a combination of rule-based structural measures to predict whether a missense genetic variant is pathogenic. Here we investigate whether the same methodology can be used to develop a differential phenotype predictor, which, once a mutation has been predicted as pathogenic, is able to distinguish between phenotypes—in this case the two major clinical phenotypes (hypertrophic cardiomyopathy, HCM and dilated cardiomyopathy, DCM) associated with mutations in the beta-myosin heavy chain (*MYH7*) gene product (Myosin-7).

Results: A random forest predictor trained on rule-based structural analyses together with structural clustering data gave a Matthews' correlation coefficient (MCC) of 0.53 (accuracy, 75%). A *post hoc* removal of machine learning models that performed particularly badly, increased the performance (MCC = 0.61, Acc = 79%). This proof of concept suggests that methods used for pathogenicity prediction can be extended for use in differential phenotype prediction.

Availability and Implementation: Analyses were implemented in Perl and C and used the Java-based Weka machine learning environment. Please contact the authors for availability.

Contacts: andrew@bioinf.org.uk or andrew.martin@ucl.ac.uk

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

1 Introduction

Mutations in proteins generally result in loss of function, but in some cases can lead to a gain of function. Generally this is not gain of a *novel* function, but an increased activity, often through loss of some type of control mechanism. In general, predictors of pathogenicity do not try to distinguish between loss-of-function and gain-of-function mutations, but simply predict whether or not there will be *some* effect on function leading to a pathogenic state.

In some cases however, the situation is more complex, with mutations in a single protein leading to a number of distinct phenotypes. For example, inherited heart muscle diseases, or

cardiomyopathies, which are a major cause of sudden cardiac death in the young and an important cause of heart failure at all ages (Hughes and McKenna, 2005) are, as a group, very heterogeneous in genotype and phenotype. Radically different phenotypes can result from mutations in the same gene (Arad *et al.*, 2002).

The widespread application of Single Nucleotide Polymorphism (SNP) chips and high-throughput sequencing has generated an urgent need for informatics tools that can help predict the effects of the many sequence variants that these platforms identify. More than 20 groups have devised methods to predict whether a given mutation will have a deleterious effect (Al-Numair and Martin, 2013; Bao *et al.*, 2005; Bromberg and Rost, 2007; Bromberg *et al.*, 2008; Calabrese *et al.*, 2009; Dantzer *et al.*, 2005; González-Pérez and López-Bigas, 2011; Karchin *et al.*, 2005; Kircher *et al.*, 2014; Li *et al.*, 2009; Reumers *et al.*, 2005; Reva *et al.*, 2011; Schwarz *et al.*, 2010; Shihab *et al.*, 2013; Stitziel *et al.*, 2004; Uzun *et al.*, 2007; Worth *et al.*, 2011; Yates *et al.*, 2014; Yip *et al.*, 2004; Yue *et al.*, 2006), the best known methods being SIFT (Ng and Henikoff, 2003), an evolutionary method which calculates a sophisticated residue conservation score from multiple alignment, and PolyPhen-2 (Adzhubei *et al.*, 2010, 2013), which uses machine learning on a set of eight sequence- and three structure-based features. A more complete list of methods is provided on our web site at http://www.bioinf.org.uk/saap/methods/. However, these tools are generally not validated for individual diseases where most available datasets are too small to train machine-learning methods and tend to be heavily unbalanced. An additional problem is that it is often very difficult to obtain reliable validated data on neutral mutations. One of the few cases where a predictor has been produced for an individual class of proteins is the work on voltage-gated potassium channels by Stead *et al.* (2011).

Attempting to distinguish between mutations in a single protein that result in different pathogenic phenotypes is a difficult problem that, unlike pathogenicity prediction, has not been widely addressed. There have been a small number of attempts to distinguish loss-of-function and gain-of-function mutations at a molecular level, but (as stated above) typically gain-of-function mutations result from loss of regulation making the protein constitutively active. For example, mutations that cause the VAB-1 tyrosine kinase to become constitutively active cause severe axon defects (Mohamed and Chin-Sang, 2006). Some of the challenges in the 'Comparative Assessment of Genome Interpretation' (CAGI) experiment have required the prediction of the level of enzyme activity (e.g. genomeinterpretation.org/content/4-NAGLU) and some have been related to familial combined hyperlipidemia or channelopathies (genomeinterpretation.org/content/FCH, genomeinterpretation.org/content/scn5a), but, to our knowledge, there have been no clear cases where predictions have focused on mutations in the same protein resulting in different phenotypes other than through loss of function *versus* loss of regulation.

Initially our own focus was on trying to understand the effects that mutations have on protein structure and then to use this information to compare the effects of non-pathogenic mutations and pathogenic deviations (Hurst *et al.*, 2009). Our approach has been to map mutations onto protein structure and to perform a rule-based analysis of the likely structural effects of these mutations in order to 'explain' the known functional effect (if any) of the mutation. Since we map mutations to structure, we only consider mutations in proteins for which a structure has been solved. With the recent growth in the amount of mutation data, we have moved from updating a database of analysis of mutations, to providing a server (SAAPdap—Single Amino Acid Polymorphism Data Analysis Pipeline) for analysis of the effects of mutations (http://www.bioinf.org.uk/saap/dap/) (Al-Numair and Martin, 2013). The approach has been used to study structural differences between disease-causing mutations and neutral polymorphisms (Al-Numair and Martin, 2013; Hurst *et al.*, 2009), and to analyze mutations in glucose-6-phosphate dehydrogenase (Kwok *et al.*, 2002) and in the tumour suppressor P53 (Martin *et al.*, 2002).

While SAAPdap uses a combination of rule-based structural measures to assess whether a mutation is likely to alter the local structural environment, we have also developed SAAPpred (Single Amino Acid Polymorphism Predictor) which exploits the results of the structural analysis and uses a Random Forest machine-learning method to predict whether mutations are pathogenic (Al-Numair and Martin, 2013). SAAPpred is restricted to analyzing mutations in proteins for which a native structure is available, but appears to outperform methods such as SIFT (Ng and Henikoff, 2003), PolyPhen-2 (Adzhubei *et al.*, 2010, 2013) and FATHMM (Shihab *et al.*, 2013).

In this paper, we investigate whether having predicted a mutation as being pathogenic, the approach that we developed for SAAPdap and SAAPpred can be used for differential phenotype prediction, specifically for mutations in the beta-myosin heavy chain (Myosin-7, UniProtKB/SwissProt accession P12883, http://www.uniprot.org/uniprot/P12883), encoded by the *MYH7* gene (OMIM *160760). Mutations in *MYH7* lead to a number of phenotypes, the most common being hypertrophic cardiomyopathy (HCM, OMIM #192600) and dilated cardiomyopathy (DCM, OMIM #613426). The numbers of mutations available for other phenotypes are very small and consequently, for this proof-of-concept paper, we have attempted to distinguish just between HCM and DCM.

Myosin-7 is part of the force-generating molecular motor of the sarcomere and parts of the structure have been solved. It is divided into three main domains as shown in Supplementary Figure S1: a globular 'head', which includes the ATP-binding site and the actin-binding site; the 'neck' which is composed of an α-helical domain to which the myosin light chains bind and which is further subdivided into a converter region and a lever arm involved in the amplification of mechanical energy; and the 'tail' or 'rod' region. Together with *MYBPC3* (the gene encoding myosin binding protein C), mutations in *MYH7* are the major cause of HCM as well as being a cause of DCM and left ventricular non-compaction (LVNC) (Haas *et al.*, 2014). In contrast to *MYBPC3*, where most pathogenic variants cause mRNA and protein truncation, the large majority of *MYH7* variants are missense (Carrier *et al.*, 1997; Richard *et al.*, 2003) which often makes prediction of pathogenicity problematic (Kumar *et al.*, 2013; Walsh *et al.*, 2010).

2 Materials and methods

2.1 Dataset of variants

A dataset of *MYH7* variants was built from (i) disease-causing or likely-pathogenic variants for which phenotypic data are available in the Human Genome Mutation Database (HGMD) (Stenson *et al.*, 2002); (ii) variants found in a curated dataset extracted from the literature and used for commercial gene testing reports (*Health in Code SL*); and (iii) variants detected in a cohort of consecutively evaluated unrelated HCM/DCM patients at the UCLH Heart Hospital. Genetic analysis was approved by the UCLH review board (IRB) and informed written consent was obtained from all subjects (Lopes *et al.*, 2013). Although there are no co-segregation data or functional studies that can 'prove' the causality of mutations, selected variants from all three datasets were rare as defined by a minor allele frequency (MAF) $< 0.5\%$ in the ESP6500 NIH Heart, Lung and Blood Institute (NHLBI) exome sequencing project database (Andreasen *et al.*, 2013; Pan *et al.*, 2012). Consequently as it is not possible to know whether variants are truly

pathogenic, we treat mutations *associated* with an HCM or DCM cardiomyopathy phenotype in the above-mentioned databases, or in the literature, as actual positives. This dataset is larger and more comprehensive than the data available from other sources and contains approximately twice the number of Myosin-7 mutations available in Swissvar/Humsavar. The complete dataset has been provided as `SupplementaryFile1.xls`. Proprietary data from HGMD, where the mutations are not available in other datasets, have been indicated solely by their HGMD accession code. The numbers of mutations for each phenotype are summarized in Supplementary Table S2.

A total of 395 unique mutations (i.e. distinct mutations, different from one another at the protein level) were identified in the *MYH7* gene. More than two-thirds of them have previously been published in the literature as being associated with disease and the others are novel variants. Since we map mutations to protein structure and therefore require a structure to be solved of the protein of interest, we are not able to analyze all mutations. Of the 395 mutations, 157 (39.7%) did not map to structure and therefore could not be analyzed (see Supplementary Table S2). This situation should improve as further structures become available. 382 of the 395 unique mutations had a recorded phenotype and of these 228 mapped to at least one Protein DataBank (PDB) chain. Supplementary Table S3 lists the PDB structures that were identified for human Myosin-7. When preparing the dataset in 2014, five structures were available and three (PDB IDs: 2fxm, 2fxo and 4db1) were used in this work. The two other structures (IDs: 1ik2 and 3dtp) were eliminated since 1ik2 is a theoretical model and 3dtp is a human-chicken fusion protein. Preliminary experiments that included this fusion protein, which covers the same region as 2fxm and 2fxo, degraded the results. Since this dataset was built, a number of other structures have become available in the PDB—some very recently—all but one of which map to the myosin tail (see Supplementary Figure S1), but all are also chimeric fusion proteins (see Supplementary Table S3). Consequently none of these structures has been included at this stage.

Most mutations were associated with HCM ($n = 290$), whereas all other phenotypes were associated with fewer than 50 mutations each, including DCM with the next highest number of mutations ($n = 46$). Of the unique HCM and DCM mutations, 190 and 21 respectively mapped to structure (see Supplementary Table S2). Since mutations related to these phenotypes were the most abundant, for this proof of concept, further analyses were restricted to HCM and DCM, grouping the remaining phenotypes as 'other'.

2.2 SAAPdap structural analysis and SAAPpred

Our previous software, SAAPdap (Al-Numair and Martin, 2013) performs a set of 14 structural analyses (using software written in Perl and C), plus the calculation of solvent accessibility (Lee and Richards, 1971). SAAPdap provides cutoffs for each of the analyses to suggest whether these are likely to be damaging (Al-Numair and Martin, 2013; Hurst *et al.*, 2009). To predict pathogenicity, a total of 47 features are derived from these analyses (Supplementary Table S1) and are used as input to SAAPpred, a machine learning method that uses Random Forests to predict whether a mutation is pathogenic (Al-Numair and Martin, 2013). In this paper, the same methodology is used but, rather than using a dataset of pathogenic and phenotypically silent mutations, a dataset of HCM and DCM mutations in Myosin-7 is used.

2.3 A machine-learning approach for *MYH7* differential phenotype prediction

As described above, for machine learning, all mutations associated with multiple phenotypes, or causing phenotypes other than HCM or DCM, were discarded leaving 190 unique HCM and 21 unique DCM mutations which map to structure (Supplementary Table S2).

Using the results of the SAAPdap structural analysis described above, of the 47 features used to describe the mutations, 14 were found to be redundant (i.e. they had the same value for all examples in the dataset: the 13 UniProtKB/SwissProt features and the disulphide (SSGeom) analysis), thus reducing the number of informative features to 33.

Since multiple structures have been solved for *MYH7*, for a given mutation, the numeric values of the features derived for each version of the structure can be slightly different. Although a single structure was used with SAAPpred, because of the limited size of the available dataset for differential phenotype prediction, it was desirable to exploit the variability in multiple structures to expand and enrich the dataset. PDB files 4db1 and 2fxm contain two copies of the protein while 2fxo contains four copies. For each mutation, the feature vectors, defined from analysis of the structure, were described using the Weka Attribute-Relation File Format (ARFF). These data were then used to train Random Forest predictors implemented in WEKA version 3.6.7 (Witten *et al.*, 2011).

The Weka Random Forest gives a classification based on a jury vote from the trees and produces a confidence score which is the fraction of trees that gave that prediction. Thus the confidence score is always between 0.5 and 1.0. In the prediction phase, the scores for the two classes are averaged separately and the higher average score is selected as the prediction. The confidence scores were then rescaled to run from –1.0 (DCM) to + 1.0 (HCM).

The parameter space described by the number of features used in each tree decision point (m_{try}) and the number of trees (T) was explored to find the best parameters for machine learning.

2.4 Cross-validation

A given mutation has multiple feature vectors describing the mutation in different copies of the structure. From the machine-learning perspective, each is a separate data point. Thus the use of multiple structures for each mutation meant that cross-validation could not be performed within WEKA since it is possible that WEKA could select the same mutation (in a different structure) to be in both training and testing sets.

To address the cross-validation problem and to deal with the severe imbalance of the dataset (there being many more HCM mutations than DCM), Perl code was written to limit the size of each class by selecting examples at random and to divide the 190 HCM and 21 DCM unique mutations with available PDB structures into sets of approximately the same size. For example, if the data were split into 21 sets, each of these 21 sets in turn (each containing one DCM mutation) was chosen as a test set and the remaining 20 sets (together containing the remaining 20 DCMs) were used for training. In each case, the datasets were enlarged with all the available PDB chain structures and balanced training datasets were generated by retaining all the DCM mutations and randomly drawing the same number of mutations from the HCM dataset. The random draws from the HCM dataset were taken 10 times over to provide a representative sample of the HCM class and the results from the trained predictors were averaged.

2.5 Structural clustering of mutations

Anecdotal evidence suggested that HCM- and DCM-associated mutations tend to be distributed differently across the Myosin-7 structure. This observation was exploited to provide additional features for the machine learning.

PDB files 2fxm and 2fxo, which represent the C-terminal region, contain only two DCM mutations compared with 35 HCM, indicating that DCM mutations are very rare in this domain. For the N-terminal domain (PDB file 4db1), the Cα positions of the mutated residues were clustered using single linkage hierarchical clustering. For each of 2…10 clusters, a χ^2 test was performed to see how well the clustering separated HCM from DCM mutations. As described in the Results, three clusters have the best significance for the clustering.

To use this information in machine learning, the centroid of each cluster was calculated and the feature vector for each mutation was expanded by the addition of the three distances from the C-alpha of the mutated residue to each of the three centroids. Mutations that were in the C-terminal domain (and mapped to PDB files 2fxm and 2fxo rather than 4db1) were given distances of 100.0, 100.0, 100.0 Å from the three clusters. To use this information in the Random Forests, these three additional features were added to the feature vectors in the ARFF files.

2.6 Optimizing the machine learning: feature selection

As well as using the full set of 33 non-redundant features from SAAPdap (the 'All' set) with or without the three clustering features, five reduced feature sets were explored. The first two of these were chosen to change the way that voids are treated, while the remaining sets were generated using feature selection to identify the most informative features.

- **All** is the full set of 33 informative features (47 from SAAPdap, but with the 14 redundant features, which were identical for all mutations, removed): BuriedCharge, Binding, CorePhilic, CisPro, Clash, Glycine, HBond, ImPACT, Interface, MutantLargestVoid1…MutantLargestVoid10, NativeLargestVoid1…NativeLargestVoid10, Proline, RelAccess, SurfacePhobic, Void. (See Supplementary Table S1 for explanation of the feature names.)
- **Top 5 voids** uses the top five largest voids (before and after mutation) instead of the standard top 10: BuriedCharge, Binding, CorePhilic, CisPro, Clash, Glycine, HBond, ImPACT, Interface, MutantLargestVoid1…NativeLargestVoid10 NativeLargestVoid1…NativeLargestVoid10 Proline, RelAccess, SurfacePhobic, Void.
- **Delta Voids** uses the differences in the sizes of the top 10 voids in native and mutant structures rather than absolute values: BuriedCharge, Binding, CorePhilic, CisPro, Clash, Glycine, HBond, ImPACT, Interface, DeltaLargestVoid11… NativeLargestVoid1 Proline, RelAccess, SurfacePhobic, Void.
- **Set1** uses the three features from the 'All' set found, individually, to be most discriminatory together with the relative solvent accessibility. A χ^2 test was performed applying the default 'damaging' threshold (Al-Numair and Martin, 2013) to each feature to determine how well the feature could separate mutations associated with HCM and DCM. See Supplementary Table S4. The three most informative features were found to be residue conservation, mutations affecting glycine residues and those affecting residues involved in specific binding interactions. Accessibility was also included since our observations of the clustering of HCM and DCM residues showed clear differences in accessibility within the clusters. Thus the feature set used was: Binding, RelAccess, ImPACT and Glycine.
- **Set2** was generated using the 'BestFirst' feature selection method within Weka. The 'BestFirst' algorithm searches the space of attribute subsets by greedy hillclimbing augmented with backtracking. The feature set based on feature selecting from the 'hemc set was: Binding, RelAccess, SurfacePhobic, CorePhilic, TotalVoidVolume, MutantLargestVoid, NativeLargestVoid, Clash, Proline, CisPro.
- **Set3** was also generated using the 'BestFirst' feature selection but on the 'Delta Voids' set. Selected features were: Binding, Interface, RelAccess, ImPACT, Hbond, BuriedCharge, DeltaVoidTotal, DeltaVoidLargest1,…NativeLargestVoid Clash, Glycine.

Initially, the number of machine-learning models was tested using the full feature set ('All'), plus those feature sets that reduced the amount of void data ('Top 5 voids' and 'Delta voids'), with and without the clustering features. Having established that 11 models was the most effective, the reduced feature sets were explored using a smaller value of m_{try} owing to the much reduced number of features.

2.7 Control experiments

Three control experiments were performed to ensure that the use of structural information in addition to sequence information or clustering was worthwhile.

First, to demonstrate improved prediction over a simple sequence-based predictor, a set of control experiments was run using only features derived from sequence data. These experiments are described in detail in `SupplementaryFile3.pdf`. Briefly, two different amino acid encodings were used, with and without conservation score, residue number (since position in the sequence can be regarded as a proxy for domain information, given that it is known that some phenotypes correlate with certain domains) and contextual information (one, three or five amino acids either side of the mutated residue). In total, 10 feature sets were considered and for each, four experiments were performed using different machine learning approaches.

Second, to ensure that the performance of the predictor does not come only from the structural clustering, we also tested the performance using the structural clusters alone. Using the 2–10 structural clusters described above, each cluster was assigned as a DCM or HCM cluster based on that phenotype having a higher observed/expected ratio in that cluster. An additional cluster was created to represent the mutations that map to the C-terminal domain (PDB code 2fxm or 2fxo) which has a very small number of DCM mutations. Each mutation was then predicted as DCM or HCM based on its cluster membership. For a real prediction problem, cluster membership would need to be assigned based on the distance to the closest cluster centre (average linkage) or closest cluster member (single linkage). Performance was then calculated for each level of clustering.

Finally, as a control on the overall prediction, the testing was repeated using two of the test sets, but the labels were randomly shuffled five times over.

3 Results

3.1 *MYH7* mutation data analysis and prediction of pathogenicity

The distribution of the variants amongst the structural and functionally-annotated domains of the beta-myosin heavy chain protein was analyzed. Supplementary Figure S1 shows the regions for which structures are known and the distribution of observed mutations together with the domains of the Myosin-7 sequence as annotated by UniProtKB/SwissProt (UniProt Consortium, 2014)

(http://www.uniprot.org/uniprot/P12883#section_features), Pfam (Finn *et al.*, 2014) (http://pfam.xfam.org/protein/P12883), SMART (Letunic *et al.*, 2012) (http://smart.embl.de/smart/show_motifs.pl?ID=P12883) and InterPro (Hunter *et al.*, 2012) (http://www.ebi.ac.uk/interpro/protein/P12883). All of the 238 unique variants that mapped to structure were located in the myosin globular 'head' domain or the 'neck' region with no mutations seen in the 'tail' or 'IQ motif' regions. 99.1% of mutations were in annotated domains or regions, while just two mutations (0.9%, at positions 82 and 838) were in un-annotated parts of the sequence. The numbers of HCM and DCM mutations seen in each of the annotated domains are shown in Supplementary Table S5.

The individual structural effects for the 228 unique mutations which mapped to structure and for which a phenotype was also recorded (see Supplementary Table S2) were analyzed using SAAPdap. 175 variants (76.8%) had one or more individual structural effects classified as likely to be damaging by the individual SAAPdap analyses while for 55 variants, no significant individual structural effect was detected (see Table 1). The features affected most frequently were: mutation of a highly conserved residue (ImPACT) occurring in 138 variants; mutation of an interface amino acid occurring in 48 variants; and disruption of hydrogen-bonds occurring in 42 variants. Other significant mutation effects occurred less frequently, with no observed mutations causing voids.

Before attempting to perform differential phenotype prediction, it would be necessary to predict that a mutation is pathogenic. The output from SAAPdap for the 228 unique mutations that mapped to structure was fed into SAAPpred (Al-Numair and Martin, 2013) and 93.0% of mutations were predicted as pathogenic (i.e. Sn = 0.930). This compares with 69.51% predicted to be pathogenic using SIFT and 90% predicted to be pathogenic using PolyPhen-2. Other metrics such as specificity (Sp), accuracy (Acc), the F1-score and the Matthews' Correlation Coefficient (MCC) could not be calculated since no set of validated non-pathogenic single amino acid mutations is available—even in the ESP 5K and 1000 Genomes data there are very few missense variants in MYH7 with a frequency >5% that could comfortably be classified as benign.

Table 1. SAAPdap Structural Analysis for the 228 unique Myosin-7 mutations with a recorded phenotype which mapped to structure (see Supplementary Table S2)

SAAPdap structural analysis	Number of mutations
Individual significant structural effect	55
At least one significant structural effect	175
• HBond	42
• BuriedCharge	31
• SProtFT	2
• Interface	48
• Clash	14
• Proline	2
• ImPACT	138
• Binding	20
• Void	0
• SurfacePhobic	15
• Glycine	8
• CisPro	1
• CorePhilic	26
• SSGeom	0

3.2 Initial machine learning results for differential phenotype prediction

As described in the Materials and Methods, machine learning was performed using random forests implemented in Weka with the 33 non-redundant features from SAAPdap structural analysis (Supplementary Table S1). Since each mutation mapped to multiple structures, cross-validation was performed outside Weka to ensure that the same mutation was not included in the training and testing sets (but mapped to different structures). The parameter space described by the number of features used in each tree decision point (m_{try}) and the number of trees (T) was explored and, as shown in Table 2, the best results were obtained using 1000 trees with 20 features (accuracy of 70% and MCC = 0.41).

3.3 Structural clustering of mutations

As described in the Materials and Methods, the Cα positions of the mutated residues were clustered using single linkage hierarchical clustering and a χ^2 test was performed for each of 2...10 clusters, to see how well the clustering separated HCM from DCM mutations. Results are shown in Table 3. Apart from two clusters, these are all clearly significant at the $P < 0.05$ level. However, as the number of clusters gets larger, one needs to take care with the significance levels, because no more than 20% of expected values should be <5 and none <1 (significance will be over-estimated if either of these is true). For ≥ 4 clusters, the first of these fails and for ≥ 6 clusters the second also fails. However, between three and six clusters the significance is so good, that (while it will be over-estimated for 4–6 clusters) it is clearly still better than $P < 0.05$ with 3 clusters passing both of the validity criteria and giving a highly significant result even if a Bonferroni correction is made for multiple testing. Consequently we clearly have clusters of residues in the N-terminal region that are over/under populated with DCM and HCM mutations compared with what is expected.

Figure 1 illustrates the three clusters in the N-terminal domain contained in PDB file 4db1. Cluster members are listed in `SupplementaryFile2.txt` and shown on the sequence in Supplementary Figure S2. In particular, DCM is highly over-represented in the third (blue/cyan) cluster. DCM mutations in clusters 1 and 2 (orange and yellow) are hardly visible and therefore mostly buried. On the other hand, the DCM mutations in cluster 3 (cyan) are largely on the surface.

Table 2. Exploring the number of features and number of trees in HCM *versus* DCM prediction

Number of folds/models	T	m_{try}	Acc	MCC
10	1000	10	0.6229	0.2463
10	1000	15	0.6750	0.3590
10	**1000**	**20**	**0.7000**	**0.4103**
10	1000	25	0.6916	0.3851
10	50	20	0.6833	0.3681
10	100	20	0.6916	0.3872
10	500	20	0.6937	0.4023
10	**1000**	**20**	**0.7000**	**0.4103**
10	2000	20	0.6812	0.3686
10	5000	20	0.7000	0.4005

T is the number of trees; m_{try} is the number of randomly chosen attributes in every split. Initially m_{try} was explored using $T = 1000$ and an optimum value of 20 was identified (shown in bold). T was then explored retaining the optimum value of 1000. Performance measures: accuracy (Acc) and Matthews' correlation coefficient (MCC). All scores are averaged over 10-folds of 'manual' (non-WEKA) cross-validation.

Table 3. Significance calculated from χ^2 tests on the ability of 3D clustering to separate HCM from DCM mutations

Number of clusters	Significance	Percentage of Expecteds < 5
2	$p < 0.4384$	0
3	**$p < 0.0003755$**	**16.7%**
4	$p < 0.001256$	37.5%
5	$p < 0.002577$	50%
6	$p < 0.005057$	50%
7	$p < 0.01013$	50%
8	$p < 0.01778$	56.25%
9	$p < 0.03044$	55.56%
10	$p < 0.03116$	60%

The highest significance result is shown in bold. For the *P*-value to be reliable, there must be no more than 20% of expected counts less than five. Consequently the *P*-values for ≥ 4 clusters will be over-estimated.

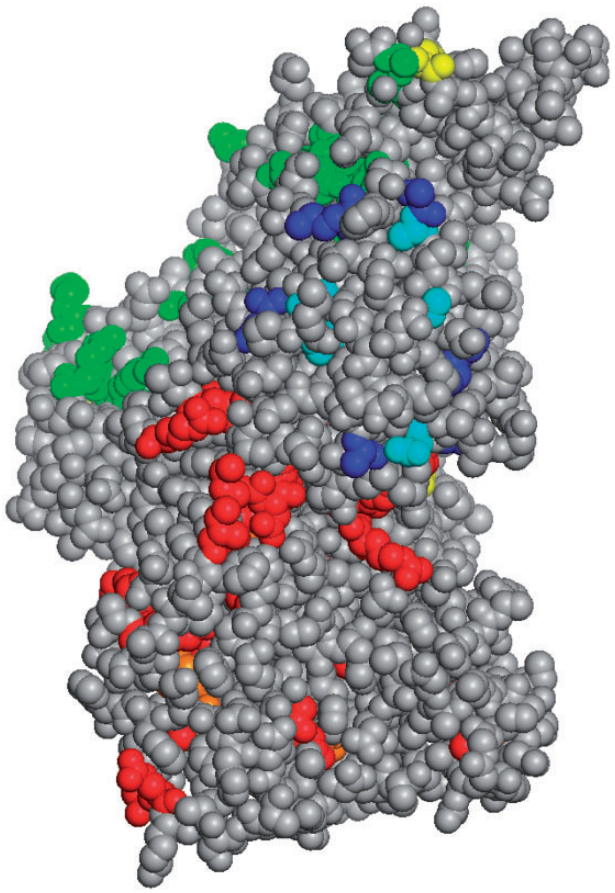

Fig. 1. Clustering Myosin-7 mutations in the N-terminal region using PDB file 4db1. For the three clusters, HCM mutations are shown in 1: red, 2: green and 3: blue, while DCM mutations are shown in 1: orange, 2: yellow and 3: cyan. DCM mutations are over-represented in cluster 3 (cyan); when they appear in clusters 1 and 2, (orange and yellow) they are mostly buried

As a control, to ensure that the significance of the clustering was not a random effect, we also permuted the labels randomly for the three clusters 1000 times over and calculated the average random *P*-value ($P = 0.5133$, $\sigma_{n-1} = 0.2859$) from a χ^2 test. This is clearly not significant and compares with the true labels which gave a $P < 0.0003755$. This p-value is 1.794 standard deviations away from the mean on the distribution of random *P*-values which is significant at the $P < 0.05$ level.

3.4 Optimizing the machine learning

Initial training to explore the number of trees and features considered per decision point was performed as described in the materials and methods using 10 machine-learning models (equivalent to cross-validation folds, each with a random selection of the HCM data) with the prediction results averaged across the 10. After determining the optimum number of features considered per descision point and number of trees, the different feature subsets were explored together with different numbers of machine-learning models (5, 11 and 21 models). Addition of the 'clustering' feature described above was also explored.

As shown in Table 4, the best performance was obtained using 11 machine-learning models with 'Set2' plus the clustering features. Cross-validation with 11 models used 19 of the 21 DCMs in each training set with 2 held back for testing. This gave an accuracy of 75% and MCC = 0.531. By removing two machine-learning models that performed particularly badly and did not predict any DCM mutations (whether correct or incorrect), this increased to an accuracy of 79% and MCC = 0.61. It appears that these particularly bad machine-learning models have failed to learn the characteristics of DCM mutations. To apply the method to novel mutations, we would remove these two bad machine-learning models and use the remaining nine to make predictions.

3.5 Control experiments

First, a set of 40 control experiments were performed to demonstrate improved prediction over a simple sequence-based predictor. Only five of the 40 experiments showed a mean MCC >0.1 with the best performance being a mean MCC = 0.167 which is clearly considerably worse than our full predictor (MCC = 0.53, or MCC = 0.61 with the worst machine-learning models removed)—See `SupplementaryFile3.pdf`.

Second, a control experiment was performed to ensure that the performance of the predictor does not come only from the structural clustering, by assigning a prediction of DCM or HCM based purely on cluster membership. The best performance was achieved with three clusters (plus the C-terminal domain cluster): MCC = 0.33, ACC = 0.89, $Sn_{HCM} = 0.95$, $Sn_{DCM} = 0.33$. Clearly this performance is considerably worse than our full predictor as judged by MCC (full predictor MCC = 0.53, or MCC = 0.61 with the worst machine-learning models removed).

This is also a good example to illustrate the well-known problem in machine learning that accuracy is a poor indicator of performance with unbalanced datasets (the cluster-only prediction gives ACC = 0.89 while the full predictor gives ACC = 0.75, or ACC = 0.79 with the worst models machine-learning removed). However, simply predicting everything as HCM would give ACC = 0.90 and, by definition, $Sn_{HCM} = 1.00$ and $Sn_{DCM} = 0.00$, while the MCC would be a much better indicator of overall performance giving a value of MCC = 0.12 (adding 1 to TP,FP,TN,FN since TN = FN = 0 results in a divide-by-zero error and treating HCM as positive and DCM as negative.).

Finally, the overall prediction control experiment, shuffling the labels on two of the test sets randomly, as expected, gave essentially random prediction performance with an MCC = –0.123 for the first test set and MCC = –0.115 for the second test set.

4 Discussion

It is logical to assume that the functional consequences of mutations in the same gene depend on the specific domain or region where the

Table 4. Summary results of machine learning performance using different features of HCM/DCM dataset and using different numbers of folds of cross-validation

Number of folds /models	Features used	T	m_{try}	Sn_{HCM}	Sn_{DCM}	F1	Acc	MCC
5	All	1000	20	0.572	0.611	0.576	0.576	0.152
5	All + Clustering	1000	20	0.755	0.481	0.679	0.648	0.311
5	Top 5 voids + Clustering	1000	20	0.735	0.611	0.688	0.681	0.368
5	10 delta void + Clustering	1000	20	0.785	0.407	0.676	0.608	0.205
11	All	1000	20	0.705	0.648	0.673	0.682	0.429
11	All + Clustering	1000	20	0.739	0.463	0.662	0.608	0.220
11	Top 5 voids + Clustering	1000	20	0.830	0.481	0.741	0.699	0.427
11	10 delta voids + Clustering	1000	20	0.830	0.519	0.730	0.676	0.521
21	All	1000	20	0.619	0.648	0.585	0.631	0.357
21	All + Clustering	1000	20	0.746	0.463	0.684	0.623	0.293
21	Top 5 voids + Clustering	1000	20	0.690	0.463	0.610	0.627	0.374
21	10 delta voids + Clustering	1000	20	0.619	0.426	0.584	0.560	0.133
11	Set1 + Clustering	1000	5	0.659	0.593	0.603	0.625	0.314
11	**Set2 + Clustering**	**1000**	**5**	**0.795**	**0.574**	**0.737**	**0.750**	**0.531**
11	Set3 + Clustering	1000	5	0.852	0.519	0.746	0.699	0.520

The best performing predictor is shown in bold. (T: the number of trees; m_{try}: the number of randomly chosen attributes in every split; Sn_{HCM}: Sensitivity for HCM mutations; Sn_{DCM}: Sensitivity for DCM mutations; F1: The F1-score; Acc: Accuracy; MCC: Matthews' Correlation Coefficient).

variant is localized (Woo *et al.*, 2003), but the hypothesis that the structural impact of a missense variant influences differential pathogenic phenotype or outcome has not previously been tested.

In practice, a novel mutation would be tested for predicted pathogenicity before an HCM/DCM prediction was performed. We confirmed that the SAAPpred approach performs well in identifying pathogenic mutations in *MYH7* and went on to test a machine-learning method that discriminated between pathogenic variants associated with an HCM or DCM phenotype (accuracy of 75% and MCC = 0.531). This was achieved by averaging 11 machine-learning models using feature Set2 (Binding, RelAccess, SurfacePhobic, CorePhilic, Voids, MutantLargestVoid1, NativeLargestVoid1, Clash, Proline, CisPro and Clustering) and using 1000 trees with 5 features. These differential phenotype prediction results are surprisingly good considering the limited size of the dataset used in training. Indeed the results are as good as the overall performance of some methods used for general pathogenicity prediction—for example, our assessment (Al-Numair and Martin, 2013) of MutationAssessor showed an overall accuracy of 69.8% and MCC = 0.453, while SIFT showed an overall accuracy of 76.3% and MCC = 0.528. Clearly these results are comparable with what we are able to achieve for HCM/DCM differential phenotype prediction which is a more difficult problem owing to the small unbalanced dataset. By removing two machine-learning models that performed particularly badly, the performance was increased to an accuracy of 79% and MCC = 0.61.

Because the SAAPdap structural analysis relies on having a crystal structure of the protein in question, our predictions are limited to mutations in regions of the protein for which a structure has been solved. Consequently, we are only able to look at 190 of 290 unique mutations leading to HCM and 21 of 46 mutations leading to DCM. As structures become available for more of the protein, then this situation will improve and some new structures have become available since our dataset was built. However, for mutations that are present in disordered regions of structure, different methods of prediction will be required. It is also possible that the performance of the method may be further improved by taking into account missing parts of the structure. However, since all the structural parameters included in the prediction are the results of local interactions, this is unlikely to have a significant effect.

Our analysis of the structural distribution of HCM- and DCM-associated mutations showed that there was a highly statistically significant difference in the locations of these mutations. Referring to Figure 1 and Supplementary Figure S2, DCM is highly over-represented in the blue/cyan cluster and largely on the surface, while DCM mutations present in the remaining clusters are mostly buried. The functional consequences of this distribution warrant further *in vitro* studies.

4.1 Conclusions and future directions

Missense single nucleotide variants in *MYH7* lead to a dominant negative effect in which the mutated protein is not degraded but rather integrates into the sarcomere, leading to the disease phenotype. The various effects of individual variants on fibre contractile velocity, force and calcium sensitivity have been proposed as an explanation for the existence of dramatically different phenotypes arising from genetic variation in the same molecule. A paradigm has been proposed whereby mutations that increase motor activity and power output lead to HCM, while those that diminish motor function and decrease power output lead to DCM (Spudich, 2014).

Our SAAPpred predictor currently relies on having a structure available for the protein in question, but planned enhancements include the use of modelled structures where no experimental structure is available and exploitation of structural information from homologues. In the same way, we plan to expand the data points for our differential phenotype predictor by including information from homologous proteins. In future work, we will also explore the newly available chimeric structures to see if they can be used for prediction of additional mutations.

As more mutation data become available, we also intend to integrate a validated pathogenicity predictor with a three-class differential phenotype predictor (HCM *versus* DCM *versus* other) although there is no *a priori* reason to believe that all 'other' mutations will have shared properties, or indeed that they will have properties that are very different from HCM or DCM. As a preliminary experiment,

we selected 10 'other' mutations (some collected after the main dataset was built) at random and found that nine of them were predicted as pathogenic using SAAPpred. If the properties of these mutations are significantly different from HCM and DCM we might expect the confidence scores provided by the differential phenotype predictor to be very low. We analyzed all 10 mutations and found that eight were predicted as HCM and two as DCM. For two of the HCM predictions (including the one predicted as SNP by SAAPpred), the confidence score was indeed very low (<0.05), but for the others, the confidence was >0.3, typical of other predictions (see Supplementary Table S6). Nonetheless, we intend to explore this further.

This work confirms the hypothesis that structural data can be used with machine learning to create a differential phenotype predictor, in this case able to distinguish between HCM and DCM mutations in *MYH7*. The performance exceeds that of the well-known SIFT program in the problem of predicting pathogenic *versus* neutral mutations. Differential phenotype prediction has all the challenges of pathogenicity prediction with the added complications of having a small unbalanced dataset. This work provides the basis for differential phenotype prediction and with further work could be used to guide clinical genetic testing strategies and further clinical investigations.

Funding

NSAN was funded by the Saudi Arabian Ministry of Higher Education (MOHE) and the King Faisal Specialist Hospital & Research Centre (KFSH&RC). LRL was supported by a grant from the Gulbenkian Doctoral Programme for Advanced Medical Education, sponsored by Fundação Calouste Gulbenkian, Fundação Champalimaud, Ministério da Saúde and Fundação para a Ciência e Tecnologia, Portugal. This work was undertaken at UCLH/UCL who received a proportion of funding from the UK Department of Health's National Institute for Health Research Biomedical Research Centres funding scheme. LM received funding from grant FIS 2011: PI11/02604, Instituto de Salud Carlos III, Madrid, Spain.

Conflict of Interest: LM is a shareholder in *Health in Code SL*. The remaining authors have no interests/relationships to declare.

References

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

Adzhubei,I.A. *et al.* (2013) Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.*, **76**, 7.20.

Al-Numair,N.S. and Martin,A.C.R. (2013) The SAAP pipeline and database: tools to analyze the impact and predict the pathogenicity of mutations. *BMC Genomics*, **14**, 1–11.

Andreasen,C. *et al.* (2013) New population-based exome data are questioning the pathogenicity of previously cardiomyopathy-associated genetic variants. *Eur. J. Hum. Genet.*, **21**, 918–928.

Arad,M. *et al.* (2002) Phenotypic diversity in hypertrophic cardiomyopathy. *Hum. Mol. Genet.*, **11**, 2499–2506.

Bao,L. *et al.* (2005) nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucleic Acids Res.*, **33**, W480–W482.

Bromberg,Y. and Rost,B. (2007) SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.*, **35**, 3823–3835.

Bromberg,Y. *et al.* (2008) SNAP predicts effect of mutations on protein function. *Bioinformatics*, **24**, 2397–2398.

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

Carrier,L. *et al.* (1997) Organization and sequence of human cardiac myosin binding protein C gene (MYBPC3) and identification of mutations predicted to produce truncated proteins in familial hypertrophic cardiomyopathy. *Circulation Res.*, **80**, 427–434.

Dantzer,J. *et al.* (2005) MutDB services: interactive structural analysis of mutation data. *Nucleic Acids Res.*, **33**, W311–W314.

Finn,R.D. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.

González-Pérez,A. and López-Bigas,N. (2011) Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am. J. Hum. Genet.*, **88**, 440–449.

Haas,J. *et al.* (2014) Atlas of the clinical genetics of human dilated cardiomyopathy. *Eur. Heart J.*, **36**, 1123–1135.

Hughes,S.E. and McKenna,W.J. (2005) New insights into the pathology of inherited cardiomyopathy. *Heart*, **91**, 257–264.

Hunter,S. *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**, D306–D312.

Hurst,J.M. *et al.* (2009) The SAAPdb web resource: a large-scale structural analysis of mutant proteins. *Hum. Mutat.*, **30**, 616–624.

Karchin,R. *et al.* (2005) LS-SNP: large-scale annotation of coding nonsynonymous SNPs based on multiple information sources. *Bioinformatics*, **21**, 2814–2820.

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

Kumar,A. *et al.* (2013) Roadmap to determine the point mutations involved in cardiomyopathy disorder: a Bayesian approach. *Gene*, **519**, 34–40.

Kwok,C.J. *et al.* (2002) G6PDdb, an integrated database of glucose-6-phosphate dehydrogenase (G6PD) mutations. *Hum. Mutat.*, **19**, 217–224.

Lee,B.K. and Richards,F.M. (1971) The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.*, **55**, 379–400.

Letunic,I. *et al.* (2012) SMART 7: Recent updates to the protein domain annotation resource. *Nucleic Acids Res.*, **40**, D302–D305.

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

Lopes,L.R. *et al.* (2013) Genetic complexity in hypertrophic cardiomyopathy revealed by high-throughput sequencing. *J. Med. Genet.*, **50**, 228–239.

Martin,A.C.R. *et al.* (2002) Integrating mutation data and structural analysis of the TP53 tumor-suppressor protein. *Hum. Mutat.*, **19**, 149–164.

Mohamed,A.M. and Chin-Sang,I.D. (2006) Characterization of loss-of-function and gain-of-function Eph receptor tyrosine kinase signaling in *C. elegans* axon targeting and cell migration. *Dev. Biol.*, **290**, 164–176.

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

Pan,S. *et al.* (2012) Cardiac structural and sarcomere genes associated with cardiomyopathy exhibit marked intolerance of genetic variation. *Circ. Cardiovasc. Genet.*, **5**, 602–610.

Reumers,J. *et al.* (2005) SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs. *Nucleic Acids Res.*, **33**, D527–D532.

Reva,B. *et al.* (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.*, **39**, e118–e118.

Richard,P. *et al.* for the EUROGENE Heart Failure Project. (2003) Hypertrophic cardiomyopathy: Distribution of disease genes, spectrum of mutations, and implications for a molecular diagnosis strategy. *Circulation*, **107**, 2227–2232.

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

Shihab,H.A. *et al.* (2013) Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.*, **34**, 57–65.

Spudich,J.A. (2014) Hypertrophic and dilated cardiomyopathy: four decades of basic research on muscle lead to potential therapeutic approaches to these devastating genetic diseases. *Biophys. J.*, **106**, 1236–1249.

Stead,L.F. *et al.* (2011) Kvsnp: accurately predicting the effect of genetic variants in voltage-gated potassium channels. *Bioinformatics*, **27**, 2181–2186.

Stenson,P.D. *et al.* (2002) *The Human Gene Mutation Database (HGMD) and Its Exploitation in the Fields of Personalized Genomics and Molecular Evolution*, Current Protocols in Bioinformatics, 39, 1.13:1.13.11.13.20.

Stitziel,N.O. *et al*. (2004) topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. *Nucleic Acids Res*., **32**, D520–D522.

UniProt Consortium (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res*., **42**, 7486–7486.

Uzun,A. *et al*. (2007) Structure SNP (StSNP): a web server for mapping and modeling nsSNPs on protein structures with linkage to metabolic pathways. *Nucleic Acids Res*., **35**, W384–W392.

Walsh,R. *et al*. (2010) Cardiomyopathy: a systematic review of disease-causing mutations in myosin heavy chain 7 and their phenotypic manifestations. *Cardiology*, **115**, 49–60.

Witten,I.H. *et al*. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. Morgan Kaufmann, Burlington, MA.

Woo,A. *et al*. (2003) Mutations of the beta myosin heavy chain gene in hypertrophic cardiomyopathy: critical functional sites determine prognosis. *Heart*, **89**, 1179–1185.

Worth,C. *et al*. (2011) SDM — a server for predicting effects of mutations on protein stability and malfunction. *Nucleic Acids Res*., **39**, W215–W222.

Yates,C.M. *et al*. (2014) SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. *J. Mol. Biol*., **426**, 2692–2701.

Yip,Y.L. *et al*. (2004) The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants. *Hum. Mutat*., **23**, 464–470.

Yue,P. *et al*. (2006) SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*, **7**, 166166.