A novel pathogenicity and phenotype predictor for functional evaluation of beta-myosin heavy chain missense variants and distinguishing between HCM and DCM associated mutations

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, Darwin Building, Gower Street, London WC1E 6BT
[2]Institute of Cardiovascular Science, UCL, London.
[3]Complejo Hospitalario Universitario de A Coruña, Insituto de Investigación Biomédica

[*]Corresponding author

Abstract

High-throughput sequencing platforms are increasingly used to screen patients with genetic disease for pathogenic mutations, but prediction of the effects of mutations on protein structure and phenotype 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 the effect of missense generic variants on protein function. Here we determine the ability of SAAPpred to predict the pathogenicity of single missense mutations in the beta-myosin heavy chain (MYH7) gene product (Myosin-7) and extend the system to predict the major associated clinical phenotypes. Final prediction results had an accuracy of 92.7% for all MYH7 mutations considered together, 99.1% for dilated cardiomyopathy (DCM) mutations and 91.4% for hypertrophic cardiomyopathy (HCM) mutations. The novel predictor using multiple random forest models to distinguish between HCM and DCM mutations using SAAPdap analysis had a best performance accuracy of 75% and a *post hoc* removal of models that performed particularly badly, raised the accuracy of 79%. These results suggest that SAAPdap analysis improves on existing tools used to predict pathogenicity of MYH7 mutations in clinical practice. If confirmed in independent analyses, this approach to phenotype prediction has the potential significantly to improve diagnostic genetic testing.

Keywords: cardiomyopathy; MYH7; Myosin-7; genetics; machine learning; protein structure.

1 Introduction

Inherited heart muscle diseases or cardiomyopathies are a major cause of sudden cardiac death in the young and an important cause of heart failure at all ages[1]. As a group, they are very heterogeneous in genotype and phenotype and radically different phenotypes can result from mutations in the same gene[2].

The widespread application of SNP chips and high-throughput sequencing has generated an urgent need for informatics tools that can predict the effects of the many sequence variants that these platforms identify. More than a dozen groups have devised methods to predict whether a given mutation will have a deleterious effect[3–16], the best known methods being SIFT[17] (an evolutionary method which calculates a sophisticated residue conservation score from multiple alignment) and PolyPhen-2[18,19], which uses machine learning on a set of eight sequence- and three structure-based features.

However, these tools are not designed for specific diseases and frequently give conflicting results. In addition, most available datasets for individual diseases are too small to train machine-learning methods and tend to be heavily unbalanced, it being particularly difficult to obtain reliable data on neutral mutations. As described previously[20], another major limitation of most existing prediction software is that it makes limited use of structural information.

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[21]. 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/`)[20] and have developed SAAPpred (Single Amino Acid Polymorphism Predictor) which takes the results of the structural analysis and uses a random forest machine learning method to predict whether mutations are pathogenic[20]. SAAPpred is restricted to analyzing mutations in proteins for which a native structure is available, but appears to outperform methods such as SIFT[17], PolyPhen-2[18,19] and FATHMM[16].

SAAPdap and SAAPpred use a combination of rule-based structural measures to assess whether a mutation is likely to alter the local structural environment and use this information to predict whether the function of a protein will be affected and, in turn, lead to disease. The approach has been used to study structural differences between disease-causing mutations and neutral polymorphisms[20,21], and to analyse mutations in glucose-6-phosphate dehydrogenase[22] and in the tumour suppressor P53[23].

The beta-myosin heavy chain (Myosin-7, UniProtKB/SwissProt accession P12883, `http://www.uniprot.org/uniprot/P12883`), encoded by the MYH7 gene, is part of the force-generating molecular motor of the sarcomere and much of the structure has been solved. Together with MYBPC3 (the gene encoding myosin binding protein C), mutations in MYH7 are the major cause of hypertrophic cardiomyopathy (HCM) as well as a cause of dilated cardiomyopathy (DCM) and left ventricular non-compaction[24]. In contrast to MYBPC3, where most pathogenic variants cause mRNA and protein truncation, the large majority of MYH7 variants are missense[25,26] which often makes prediction of pathogenicity problematic[27,28].

Here, the performance of SAAPpred on missense mutations in MYH7, leading to changes in the Myosin-7 protein, is tested to determine their potential as a predictive tool in patients with cardiomyopathy. Further, the possibility of using the same approach (together with an additional set of features describing structural clustering) to investigate genotype/phenotype relationships at a more detailed level is investigated by attempting to distinguish mutations that cause HCM from those that cause DCM.

2 Materials and Methods

2.1 Dataset of variants

A dataset of MYH7 variants detected in a cohort of consecutively evaluated unrelated HCM patients was selected for study. All selected variants were rare as defined by a minor allele frequency (MAF) $< 0.5\%$ in the NIH Heart, Lung and Blood Institute (NHLBI) exome sequencing project database[29,30]. Genetic analysis was approved by the UCLH review board (IRB) and informed written consent was obtained from all subjects[31]. To increase the number of variants analysed, the data were enriched with other established disease-causing or likely-pathogenic variants in MYH7, for which phenotypic data are available in the Human Genome Mutation Database (HGMD)[32] or in a curated dataset of MYH7 variants extracted from the literature and used for commercial gene testing reports (*Health in Code SL*).

2.2 Prediction of *in silico* pathogenicity

Prediction of mutation pathogenicity was performed using PolyPhen-2, SIFT, and SAAPpred[20]. SAAPpred exploits SAAPdap which analyses a set of 14 structural features (**Interface:** residue is in an interface according to difference in solvent accessibility between complexed and uncomplexed forms; **Binding:** residue makes specific interactions with a different protein chain or ligand; **SProtFT:** residue is annotated as functionally relevant by UniProtKB/SwissProt; **Clash:** mutation introduces a steric clash with an existing residue; **Void:** mutation introduces a destabilizing void in the protein core; **CisPro:** mutation from cis-proline, introducing an unfavourable ω torsion angle; **Glycine:** mutation from glycine, introducing unfavourable torsion angles; **Proline:** mutation to proline, introducing unfavourable torsion angles; **HBond:** mutation disrupts a hydrogen bond; **CorePhilic:** introduction of a hydrophilic residue in the protein core; **SurfacePhobic:** introduction of a hydrophobic residue on the protein surface; **BuriedCharge:** mutation causes an unsatisfied charge in the protein core; **SSGeom:** mutation disrupts a disulphide bond; **ImPACT:** residue is significantly conserved). From these analyses, together with relative accessibility (**RelAccess**), 47 features are derived (using software written in Perl and C) and used in SAAPpred to predict pathogenicity using Random Forests implemented in WEKA[33], trained as described in Al-Numair *et al.*[20].

The SAAPdap analysis indicates local structural effects and suggests those mutations for which one or more individual analyses are likely to be damaging by themselves. The SAAPpred predictor makes use of these analyses to make a prediction of pathogenicity; thus a mutation with no analyses individually expected to be damaging can still be predicted to be pathogenic as a result of the accumulation of a number of more subtle effects.

The same approach was used in separating mutations associated with HCM and DCM. However, three additional features were used that represent distances from cluster centres identified by clustering the coordinates of HCM and DCM mutations using single linkage clustering and finding the number of clusters that gave the most significant separation of HCM and DCM mutations between the clusters (χ^2 test).

3 Results

3.1 MYH7 mutation data analysis

MYH7 mutations associated with various cardiomyopathy phenotypes are shown in Table 1. Note that it is not possible to know whether variants are truly pathogenic;

Disease (Phenotype)	Total mutations	Unique mutations	Mutations mapped to PDB
HCM	298	292	188
DCM	46	46	21
RCM	1	1	1
LVNC	17	17	1
LVNC/ASD	1	1	1
DCM/Endocardial Fibroelastosis	1	1	1
DCM/LVNC	3	3	2
HCM/LVNC	1	1	1
HCM/DCM/LVNC	2	2	2
HCM/DCM	3	3	3
HCM/RCM/DCM	2	2	2
Laing distal myopathy	4	4	2
Ebstein	5	5	1
Cardiomyopathy and distal myopathy	3	3	2
Myosin storage myopathy	3	3	1
Hyaline body myopathy	1	1	1
No recorded phenotype	11	11	5
Total	403	396	235

Table 1: Numbers of MYH7 mutations for each phenotype. Abbreviations: PDB, Protein DataBank; DCM, Dilated Cardiomyopathy; HCM, Hypertrophic Cardiomyopathy; RCM, Restrictive Cardiomyopathy; LVNC, Left Ventricular Non-compaction; ASD, Atrial Septal Defect. The mutations for which there was no recorded phenotype were excluded from structural analysis, meaning that 230 mutations which mapped to PDB structures were analysed.

PDB ID	Description	Residues
2fxm	Structure of the human beta-myosin S2 fragment	A: 838–961 B: 850–961
2fxo	Structure of the human beta-myosin S2 fragment	A: 838–963 B: 838–961 C: 838–962 D: 838–963
4db1	Cardiac human myosin S1DC, beta isoform complexed with Mn-AMPPNP	A: 2-777 B: 2-775

Table 2: PDB structures for UniProtKB/SwissProt accession code P12883. PDB files may be accessed at `http://www.pdb.org/`. Note that PDB file 2fxo contains a mutation Glu924Lys.

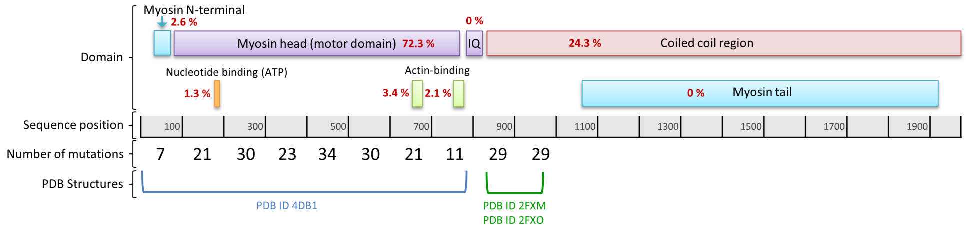

Figure 1: Annotated regions of the Myosin-7 sequence. Regions for which structures are known are indicated, together with the number of known mutations from Table 1 in each 100 amino acids of the sequence. The percentage of the total 235 mutations in each region is indicated: two of the mutations (at positions 82 and 838) do not correspond to any annotated regions. **Myosin N-terminal** Pfam annotation, residues 34–75; **Myosin head (motor domain)** Pfam and InterPro annotation, residues 85–778; **IQ motif** UniProtKB/SwissProt and InterPro annotation, residues 781–810, SMART annotation, residues 780-802; **Coiled coil region** UniProtKB/SwissProt annotation, residues 839–1935, SMART annotation, residues 841-1927; **Nucleotide binding (ATP) region** UniProtKB/SwissProt annotation, residues 178–185; **Actin-binding region** UniProtKB/SwissProt annotation, residues 655–677; **Actin-binding region** UniProtKB/SwissProt annotation, residues 757–771; **Myosin tail** Pfam and InterPro annotation, residues 1068–1926.

rather we treat mutations *associated* with an HCM or DCM cardiomyopathy phenotype in the above-mentioned databases, or in the literature, as actual positives. A total of 403 mutations 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. Of the total mutations for which a phenotype was recorded, 385 were unique and 230 mapped to at least one Protein DataBank (PDB) chain. Table 2 lists three PDB structures which were identified for human Myosin-7. Two other PDB files (IDs 1ik2 and 3dtp) were eliminated since one was a model and the other was a human-chicken fusion protein. Most mutations were associated with HCM ($n = 298$), whereas all other phenotypes were associated with fewer than 50 mutations each, including DCM with the next highest number of mutations ($n = 46$). The majority of mutations in both HCM and DCM were unique (292 and 46 respectively). Since mutations related to these phenotypes were the most abundant, further analyses were conducted, looking specifically at HCM and DCM and grouping the remaining phenotypes as 'other'.

The distribution of the variants amongst the structural and functionally-annotated domains of the beta-myosin heavy chain protein were analysed. Figure 1 shows the domains of the Myosin-7 sequence as annotated by UniProtKB/SwissProt[34] (`http://www.uniprot.org/uniprot/P12883#section_features`), Pfam[35] (`http://pfam.xfam.org/protein/P12883`), SMART[36] (`http://smart.embl.de/smart/show_motifs.pl?ID=P12883`), and InterPro[37] (`http://www.ebi.ac.uk/interpro/protein/P12883`), the regions for which structures are known and the distribution of observed mutations. All of the 235 unique variants were located in the myosin globular 'head' domain or the 'neck' region with no mutations seen in the 'Myosin tail' region or the 'IQ motif' region. 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.

SAAPdap Structural Analysis	Number of mutations
No PDB structure available	166
No 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

Table 3: SAAPdap Structural Analysis for the 230 unique Myosin-7 mutations with a recorded phenotype which mapped to structure (see Table 1).

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 analyse all mutations. Of the 396 distinct mutations in MYH7, 166 (41.9%) did not map to structure and therefore could not be analysed. This situation should improve as further crystal structures become available.

The 230 unique mutations for which a phenotype was recorded and which mapped to structure (see Table 1) were analysed using the SIFT and PolyPhen-2 prediction software. Of these, 69.51% were predicted to be pathogenic using SIFT and 90% were predicted to be pathogenic using PolyPhen-2. Since all mutations are associated with HCM and DCM, this corresponds to accuracies of 69.51% and 90% respectively.

Analysing the data with SAAPdap shows that a total of 175 variants were classified as likely to be damaging by at least one individual SAAPdap analysis. For 55 variants, no significant individual structural effect was detected by SAAPdap analysis and, as explained above, 166 could not be analysed by SAAPdap because they did not map to a PDB structure (see Table 3). The most frequent features affected were: mutation of a highly conserved residue (ImPACT) occurring in 138 variants; mutation of an interface amino acid (Interface) occurring in 48 of the variants; disruption of H-bonds occurring in 42 of the variants. Other significant mutation effects occurred less frequently, with no observed mutations causing voids or disrupting disulphide bonds.

3.2 Pathogenicity prediction

Pathogenicity prediction was performed using the SAAPpred predictor trained on HumVar as described by Al-Numair *et al.*[20]. Ten pre-built models were used and the performance results were averaged. Since all mutations in the dataset were associated with HCM and DCM, true negatives (TN) and false positives (FP) could not be calculated and consequently, the Matthews' Correlation Coefficient (MCC) could not be calculated. Table 4 shows a summary of results. Overall accuracy for

	One PDB			Multi PDB		
	Sn	$F1$	Acc	Sn	$F1$	Acc
HCM	0.914	0.955	0.914	0.795	0.883	0.795
DCM	0.991	0.995	0.991	0.789	0.878	0.789
Other	0.967	0.983	0.967	0.797	0.884	0.797
All	0.927	0.962	0.927	0.794	0.882	0.794

Table 4: Sensitivity (Sn), $F1$-measure and Accuracy (Acc) for the pathogenicity prediction for the distinct mutations that could be analysed using SAAPdap and SAAPpred. Note that there were no negative examples in the dataset and consequently, Sn and accuracy values are identical.

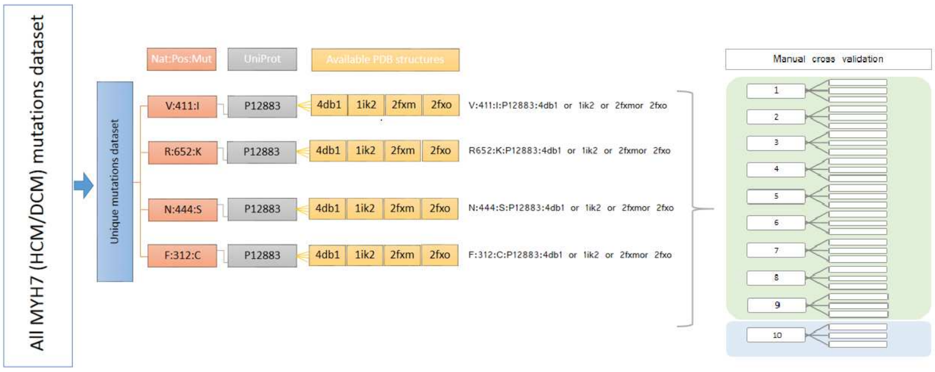
All MYH7 (HCM/DCM) mutations dataset
Unique mutations dataset
Nat:Pos:Mut
UniProt
Available PDB structures
V:411:I
R:652:K
N:444:S
F:312:C
P12883
4db1
1ik2
2fxm
2fxo
V:411:I:P12883:4db1 or 1ik2 or 2fxmor 2fxo
R652:K:P12883:4db1 or 1ik2 or 2fxmor 2fxo
N:444:S:P12883:4db1 or 1ik2 or 2fxmor 2fxo
F:312:C:P12883:4db1 or 1ik2 or 2fxmor 2fxo
Manual cross validation
1
2
3
4
5
6
7
8
9
10

Figure 2: MYH7 (HCM/DCM) dataset selection for machine learning. A unique mutation level filtering is used, where the same mutation (UniProtKB/SwissProt:Native:Number:Mutant) does not occur in training and testing sets. This was achieved using a 'manual' (non-WEKA) cross-validation that splits the dataset into N sets, each one in turn was chosen as the testing set and the remaining $N-1$ were used for training.

all phenotypes (HCM/DCM/Other) was 92.7% when using a single best-resolution PDB chain, but was reduced to 79.4% when using all PDB chains. In this instance, it is clearly important only to use the best resolution chain for each mutation rather than using data from multiple chains. Thus SAAPpred (accuracy = 92.7%) shows a clear performance improvement over the results shown earlier for SIFT (accuracy = 69.51%) and PolyPhen-2 (accuracy = 90%).

3.3 A machine learning approach to predict MYH7 phenotype

All mutations associated with multiple phenotypes, or causing phenotypes other than HCM or DCM (i.e. Restrictive Cardiomyopathy (RCM), Left Ventricular Noncompaction (LVNC), Atrial Septal Defect (ASD), Ebstein's anomaly, distal skeletal myopathies, etc.), were discarded leaving the 188 unique HCM and 21 unique DCM mutations which map to structure.

Of the 47 'features' from the SAAPdap structural analysis 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 to 33 features. Although in the pathogenicity prediction, using a single structure was more effective than using multiple structures, because of the limited size of the available dataset for pheno-

Number of folds	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

Table 5: Exploring the number of features and number of trees in HCM *vs.* DCM prediction. 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 Matthew's correlation coefficient (MCC). All scores are averaged over 10-folds of 'manual' (non-WEKA) cross-validation.

type prediction, it was desirable to exploit multiple structures to enrich the dataset. These data were then used to train Random Forest models in WEKA. 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 188 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 were used for training. In each case, the data sets were enlarged with all the available PDB chain structures and balanced datasets were generated by retaining all the DCM mutations and randomly drawing the same number of mutations from the HCM dataset (see Figure 2). 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.

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 5, the best results were obtained using 1000 trees with 20 features (accuracy of 70% and MCC=0.41).

3.4 Clustering mutations

Anecdotal evidence suggested that HCM and DCM associated mutations tend to be distributed differently across the Myosin-7 structure. This observation was exploited in an attempt to improve the results,

PDB file 2fxm, which represents the C-terminal region, contains 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

Number of clusters	Significance
2	$p < 0.4384$
3	$p <$ **0.0003755**
4	$p < 0.001256$
5	$p < 0.002577$
6	$p < 0.005057$
7	$p < 0.01013$
8	$p < 0.01778$
9	$p < 0.03044$
10	$p < 0.03116$

Table 6: Significance calculated from χ^2 tests on the ability of 3D clustering to separate HCM from DCM mutations. The highest significance result is shown in bold.

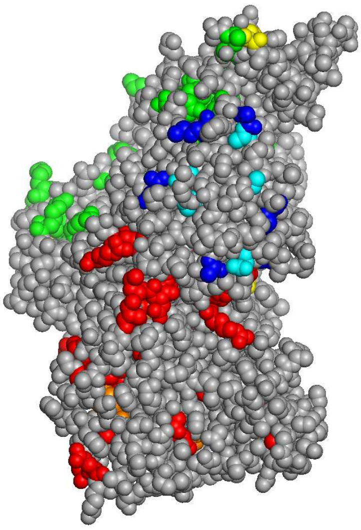

Figure 3: Clustering Myosin-7 mutations in the N-terminal region using PDB file 4db1. For the three clusters, HCM mutations are shown in 1: red, 2: green and 3: blue, while DCM mutations are shown in 1: orange, 2: yellow and 3: cyan. DCM mutations are over-represented in cluster 3 (cyan); when they appear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

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 shown in Table 6. Apart from 2 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 ≥ 3 clusters, the first of these fails and for ≥ 6 clusters the second also fails. However, between 3 and 6 clusters the significance is so good, that (while it will be over-estimated) it is probably still better than 0.05 and 3 clusters is clearly the most significant result. 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 3 illustrates the three clusters in the N-terminal domain contained in PDB file 4db1. Note that the clustering was done on one chain and the results are then shown on the two chains in the 4db1 crystal structure. 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.

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 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 100.0Å, 100.0Å, 100.0Å from the three clusters.

3.5 Optimizing the machine learning

As described above, initial training to explore the number of trees and features was performed using 10 models (each with a random selection of the HCM data) with the prediction results averaged across the 10. Using a larger number of models allows more of the HCM data to be exploited in each model while maintaining balanced datasets. Using 20 models, only one unique DCM mutation can be held back from training for test purposes. However, the number of models is not limited to 20 because it is possible to hold one DCM back and then build several models using different sets of HCMs.

After determining the optimum number of features and trees, the most informative features were explored together with different numbers of models (5, 11 and 21 models). Odd numbers were used to allow a jury vote in predictions if required. Addition of the 'clustering' feature described above was also explored. The different feature sets are described in detail together with summary results in Table 7. In brief the feature subsets were as follows: 'All' the full standard set of 33 informative features (47 from SAAPdap, but with the 14 redundant features, which were identical for all mutations, removed); 'Top 5 voids' uses only the top five largest voids (before and after mutation) instead of the standard 10; 'Delta voids' uses differences between void sizes in native and mutant structures; 'Set1' was a selection of the five features found to be most discriminatory using χ^2 tests on each of the features; 'Set2' and 'Set3' were sets of features randomly generated within WEKA, 'Set2' being based on the 'All' dataset and 'Set3' being based on the 'Delta voids' set.

Initially, the number of models was tested using the full feature set ('All'), plus those 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 more-reduced feature sets were explored using a smaller value of m_{try} owing to the much reduced number of features.

Number of folds	Features used	T	m_{try}	Acc	MCC
5	All	1000	20	0.576	0.152
5	All + Clustering	1000	20	0.648	0.311
5	Top 5 voids + Clustering	1000	20	0.681	0.368
5	10 delta void + Clustering	1000	20	0.608	0.205
11	All	1000	20	0.682	0.429
11	All + Clustering	1000	20	0.608	0.220
11	Top 5 voids + Clustering	1000	20	0.699	0.427
11	10 delta voids + Clustering	1000	20	0.676	0.521
21	All	1000	20	0.631	0.357
21	All + Clustering	1000	20	0.623	0.293
21	Top 5 voids + Clustering	1000	20	0.627	0.374
21	10 delta voids + Clustering	1000	20	0.560	0.133
11	Set1 + Clustering	1000	5	0.625	0.314
11	**Set2 + Clustering**	**1000**	**5**	**0.750**	**0.531**
11	Set3 + Clustering	1000	5	0.699	0.520

Table 7: Summary results of machine learning performance using different features of HCM/DCM dataset and using different numbers of folds of cross-validation. (Acc: Accuracy; MCC: Matthews' Correlation Coefficient; T: the number of trees; m_{try}: the number of randomly chosen attributes in every split)

• 'All': BuriedCharge, Binding, CorePhilic, CisPro, Clash, Glycine, HBond, ImPACT, Interface, MutantLargestVoid1...MutantLargestVoid10, NativeLargestVoid1...NativeLargestVoid10, Proline, RelAccess, SurfacePhobic, Void.

• 'Top 5 voids': BuriedCharge, Binding, CorePhilic, CisPro, Clash, Glycine, HBond, ImPACT, Interface, MutantLargestVoid1...MutantLargestVoid5, NativeLargestVoid1...NativeLargestVoid5, Proline, RelAccess, SurfacePhobic, Void.

• 'Delta Voids': BuriedCharge, Binding, CorePhilic, CisPro, Clash, Glycine, HBond, ImPACT, Interface, DeltaLargestVoid1...DeltaLargestVoid10, Proline, RelAccess, SurfacePhobic, Void.

• 'Set1': Uses the most informative features based on χ^2 tests: Binding, RelAccess, ImPACT and Glycine.

• 'Set2': A WEKA randomly selected dataset: Binding, RelAccess, SurfacePhobic, CorePhilic, TotalVoidVolume, MutantLargestVoid, NativeLargestVoid, Clash, Proline, CisPro.

• 'Set3': A WEKA randomly selected dataset based on the 'Delta Voids' set: Binding, Interface, RelAccess, ImPACT, Hbond, BuriedCharge, DeltaVoidTotal, DeltaVoidLargest1...DeltaVoidLargest5, Clash, Glycine.

As shown in Table 7, the best performance was obtained using 11 models with 'Set2' plus the clustering features. This gave an accuracy of 75% and MCC = 0.531. By removing models that performed particularly badly, we reached an accuracy of 79% and MCC=0.61. In these particularly bad models where mutations map to multiple PDB chains, it appears that some of the structures make the performance worse.

4 Discussion

4.1 Added value of structural data to face the challenges in predicting pathogenicity of a missense variant

Patients with heritable cardiomyopathies often carry novel and missense genetic variants that can be difficult to interpret. Ideally, novel variants should be tested using functional studies and/or co-segregation analysis within families[38], but both approaches are difficult to apply and interpret in the clinical setting, particularly when using high-throughput genetic screening strategies that identify hundreds of potentially pathogenic variants in patients as well as normal control populations.

In the case of sarcomere protein genes, previously reported and probable pathogenic variants in MYH7 and MYBPC3 have been reported at a frequency higher than expected for the prevalence of cardiomyopathy in the 1000 Genomes database[39–41], the NHLBI exome sequencing population data[29,30] and the Framingham and Jackson Heart Study cohorts[42]. Possible explanations for these findings include a combination of reduced penetrance for some of these alleles, digenic and oligogenic models of inheritance and erroneous attribution of pathogenicity to previously reported variants[43].

4.2 Proof of concept evidence for an *in silico* phenotype-prediction tool for beta-myosin heavy chain variants associated with cardiomyopathy

It is logical to assume that the functional consequences of mutations in the same gene depend on the specific domain or region where the variant is localized[44], but the hypothesis that the structural impact of a missense variant influences phenotype or outcome has not previously been tested.

In this study, we show that the SAAPpred approach was able to discriminate between pathogenic and neutral MYH7 variants with a much higher level of accuracy (92.7% for all mutations; 99.1% for DCM and 92.4% for HCM) than that of commonly used prediction models (SIFT: 69.5% and PolyPhen-2 90% for all mutations).

We were also able to develop a model that discriminated between pathological variants associated with an HCM or DCM phenotype (accuracy of 75% and MCC=0.531). This was achieved by averaging 11 models using feature Set2 (Binding, RelAccess, SurfacePhobic, CorePhilic, Voids, MutantLargestVoid1, NativeLargestVoid1, Clash, Proline, CisPro and Clustering) and using 1000 trees with 5 features. By removing models that performed particularly badly, we achieved an accuracy of 79% and MCC=0.61. While not as good as the general pathogenicity prediction, these 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 pathogenicity prediction — for example, our assessment 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[20]. Clearly these results are comparable with what we are able to achieve for HCM/DCM phenotype prediction which is a much more complex problem.

In creating this predictor, we analyzed the structural distribution of HCM- and DCM-associated mutations and found that there was a highly statistically significant difference in the locations of these mutations. Referring to Figure 3, 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.

4.3 Conclusions and future directions

The inclusion of structural data as part of an *in silico* pathogenicity prediction model increases the accuracy of pathogenicity modelling for MYH7. These preliminary and proof-of-concept data suggest that it is possible to develop an iterative gene-specific prediction tool for patients with cardiomyopathy.

5 Funding

NSAN was funded by the Saudi Arabian Ministry of Higher Education (MOHE) and the King Faisal Specialist Hospital & Research Centre (KFSH&RC) for funding. 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 finding 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.

6 Conflicts of interest

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

References

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

2. Arad, M., Seidman, J. and Seidman, C. E. (2002). Phenotypic diversity in hypertrophic cardiomyopathy. Human Molecular Genetics. *11*, 2499–2506.

3. Yue, P., Melamud, E. and Moult, J. (2006). SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinformatics. *7*, 166–166.

4. Uzun, A., Leslin, C. M., Abyzov, A. and Ilyin, V. (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.

5. Yip, Y. L., Scheib, H., Diemand, A. V., Gattiker, A., Famiglietti, L. M., Gasteiger, E. and Bairoch, A. (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.

6. Dantzer, J., Moad, C., Heiland, R. and Mooney, S. (2005). MutDB services: Interactive structural analysis of mutation data. Nucleic Acids Res. *33*, W311–W314.

7. Karchin, R., Diekhans, M., Kelly, L., Thomas, D. J., Pieper, U., Eswar, N., Haussler, D. and Sali, A. (2005). LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics. *21*, 2814–2820.

8. Stitziel, N. O., Binkowski, T. A., Tseng, Y. Y., Kasif, S. and Liang, J. (2004). topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. Nucleic Acids Res. *32*, D520–D522.

9. Reumers, J., Schymkowitz, J., Ferkinghoff-Borg, J., Stricher, F., Serrano, L. and Rousseau, F. (2005). SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs. Nucleic Acids Res. *33*, D527–D532.

10. Bao, L., Zhou, M. and Cui, Y. (2005). nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. Nucleic Acids Res. *33*, W480–W482.

11. Reva, B., Antipin, Y. and Sander, C. (2011). Predicting the functional impact of protein mutations: Application to cancer genomics. Nucleic Acids Res. *39*, e118–e118.

12. Schwarz, J. M., Rödelsperger, C., Schuelke, M. and Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. *7*, 575–576.

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

14. Bromberg, Y., Yachdav, G. and Rost, B. (2008). SNAP predicts effect of mutations on protein function. Bioinformatics. *24*, 2397–2398.

15. 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.

16. Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L. A., Edwards, K. J., Day, I. N. M. and Gaunt, T. R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat. *34*, 57–65.

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

18. Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S. and Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. Nature methods. *7*, 248–249.

19. Adzhubei, I. A., Jordan, D. M. and Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. Current Protococols in Human Genetetics. *76*, 7.20.

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.

21. Hurst, J. M., McMillan, L. E. M., Porter, C. T., Allen, J., Fakorede, A. and Martin, A. C. R. (2009). The SAAPdb web resource: a large-scale structural analysis of mutant proteins. Hum Mutat. *30*, 616–624.

22. Kwok, C. J., Martin, A. C. R., Au, S. W. N. and Lam, V. M. S. (2002). G6PDdb, an integrated database of glucose-6-phosphate dehydrogenase (G6PD) mutations. Human Mutation. *19*, 217–224.

23. Martin, A. C. R., Facchiano, A. M., Cuff, A. L., Hernandez-Boussard, T., Olivier, M., Hainaut, P. and Thornton, J. M. (2002). Integrating mutation data and structural analysis of the TP53 tumor-suppressor protein. Hum Mutat. *19*, 149–164.

24. Haas, J., Frese, K. S., Peil, B., Kloos, W., Keller, A., Nietsch, R., Feng, Z., Müller, S., Kayvanpour, E., Vogel, B., *et al.* (2014). Atlas of the clinical genetics of human dilated cardiomyopathy. Eur Heart J. *EPub ahead of print*, ehu301.

25. Carrier, L., Bonne, G., Bahrend, E., Yu, B., Richard, P., Niel, F., Hainque, B., Cruaud, C., Gary, F., Labeit, S., *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 Research. *80*, 427–434.

26. Richard, P., Charron, P., Carrier, L., Ledeuil, C., Cheav, T., Pichereau, C., Benaiche, A., Isnard, R., Dubourg, O., Burban, M., *et al.* (2003). Hypertrophic cardiomyopathy: Distribution of disease genes, spectrum of mutations, and implications for a molecular diagnosis strategy. Circulation. *107*, 2227–2232.

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

28. Kumar, A., Rajendran, V., Sethumadhavan, R. and Purohit, R. (2013). Roadmap to determine the point mutations involved in cardiomyopathy disorder: a Bayesian approach. Gene. *519*, 34–40.

29. Pan, S., Caleshu, C. A., Dunn, K. E., Foti, M. J., Moran, M. K., Soyinka, O. and Ashley, E. A. (2012). Cardiac structural and sarcomere genes associated with cardiomyopathy exhibit marked intolerance of genetic variation. Circ Cardiovasc Genet. *5*, 602–610.

30. Andreasen, C., Nielsen, J. B., Refsgaard, L., Holst, A. G., Christensen, A. H., Andreasen, L., Sajadieh, A., Haunsø, S., Svendsen, J. H. and Olesen, M. S. (2013). New population-based exome data are questioning the pathogenicity of previously cardiomyopathy-associated genetic variants. Eur J Hum Genet. *21*, 918–928.

31. Lopes, L. R., Zekavati, A., Syrris, P., Hubank, M., Giambartolomei, C., Dalageorgou, C., Jenkins, S., McKenna, W., 1000 Genomes Project Consortium, Plagnol, V., *et al.* (2013). Genetic complexity in hypertrophic cardiomyopathy revealed by high-throughput sequencing. J Med Genet. *50*, 228–239.

32. Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shaw, K. and Cooper, D. N. (2002). The Human Gene Mutation Database (HGMD) and Its Exploitation in the Fields of Personalized Genomics and Molecular Evolution: Chapter 1, Unit 1.13. John Wiley & Sons, Inc.

33. Witten, I. H., Frank, E. and Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann: Burlington, MA 01803, USA: Third edition.

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

35. Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., *et al.* (2014). Pfam: the protein families database. Nucleic Acids Res. *42*, D222–D230.

36. Letunic, I., Doerks, T. and Bork, P. (2012). SMART 7: Recent updates to the protein domain annotation resource. Nucleic Acids Res. *40*, D302–D305.

37. Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T. K., Bateman, A., Bernard, T., Binns, D., Bork, P., Burge, S., *et al.* (2012). InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res. *40*, D306–D312.

38. MacArthur, D. G., Manolio, T. A., Dimmock, D. P., Rehm, H. L., Shendure, J., Abecasis, G. R., Adams, D. R., Altman, R. B., Antonarakis, S. E., Ashley, E. A., *et al.* (2014). Guidelines for investigating causality of sequence variants in human disease. Nature. *508*, 469–476.

39. 1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E. and McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. Nature. *467*, 1061–1073.

40. 1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T. and McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature. *491*, 56–65.

41. Golbus, J. R., Puckelwartz, M. J., Fahrenbach, J. P., Dellefave-Castillo, L. M., Wolfgeher, D. and McNally, E. M. (2012). Population-based variation in cardiomyopathy genes. Circ Cardiovasc Genet. *5*, 391–399.

42. Bick, A. G., Flannick, J., Ito, K., Cheng, S., Vasan, R. S., Parfenov, M. G., Herman, D. S., DePalma, S. R., Gupta, N., Gabriel, S. B., *et al.* (2012). Burden of rare sarcomere gene variants in the Framingham and Jackson Heart Study cohorts. Am J Hum Genet. *91*, 513–519.

43. Cooper, D. N., Krawczak, M., Polychronakos, C., Tyler-Smith, C. and Kehrer-Sawatzki, H. (2013). Where genotype is not predictive of phenotype: Towards an understanding of the molecular basis of reduced penetrance in human inherited disease. Hum Genet. *132*, 1077–1130.

44. Woo, A., Rakowski, H., Liew, J. C., Zhao, M.-S., Liew, C.-C., Parker, T. G., Zeller, M., Wigle, E. D. and Sole, M. J. (2003). Mutations of the beta myosin heavy chain gene in hypertrophic cardiomyopathy: Critical functional sites determine prognosis. Heart. *89*, 1179–1185.