Sequence-based Predictor Control Experiments

A set of control experiments was run using only features derived from sequence data. Amino acids were encoded using either 20 parameters (scores from the appropriate row of the BLOSUM80 matrix) or 6 parameters (3 numeric values: number of sidechain atoms; consensus hydrophobicity; charge; 3 Boolean values: ring – i.e. phe, his, try, tyr; sulphur containing – i.e. cys or met; conformational – i.e. gly or pro). As well as the native and mutant amino acid, the conservation score was considered (also used in the full predictor), as was the 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). In addition, when using the 6-parameter encoding for amino acids, in some experiments, the context of the mutation was considered (i.e. one, three or five amino acids either side of the mutated residue were also included). This was not used with the 20-parameter encoding of amino acids because of the worse ratio of training data to the number of paraneters.

In total 10 feature sets were considered and for each, four experiments were performed using different machine learning approaches in Weka: an artificial neural network using default parameters and three random forests with different parameters (see below).

For each of the 40 experiments, 10 models were trained, in each case using all 21 DCM mutations and a random selection of 21 HCM mutations. For each model, 10-fold cross-validation was performed in Weka and the MCC was calculated. The MCC values for the 10 models were averaged to give a performance metric for each experiment.

Results for these control experiments are shown in the table below. Most experiments showed random or worse-than-random performance. Only five of the 40 experiments showed an MCC > 0.1, the highest value being 0.167

Feature					**MCC (mean / S.D.)**			
Set	**Encoding**	**Pos**	**Cons**	**Context**	**ANN**	**RF20/5†**	**RF100/10**	**RF1000/20**
1	20	-	-	-	-0.129/0.117	-0.063/0.156	-0.053/0.155	-0.101/0.154
2	20	Yes	-	-	-0.099/0.181	-0.054/0.136	0.017/0.212	-0.019/0.220
3	20	-	Yes	-	-0.0147/0.106	0.005/0.212	-0.125/0.116	-0.079/0.193
4	20	Yes	Yes	-	-0.034/0.181	-0.072/0.170	0.015/0.216	0.020/0.131
5	6	-	-	-	0.062/0.247	-0.038/0.164	-0.014/0.216	0.081/0.135
6	6	Yes	Yes	-	0.024/0.173	**0.134/0.192**	-0.034/0.182	0.025/0.124
7	6	-	-	1	**0.148/0.184**	-0.005/0.206	0.058/0.241	0.072/0.212
8	6	Yes	Yes	1	0.053/0.150	0.048/0.134	**0.101/0.171**	0.042/0.221
9	6	Yes	Yes	3	**0.167/0.171**	0.018/0.107	0.004/0.125	-0.038/0.148
10	6	Yes	Yes	5	0.077/0.149	**0.139/0.158**	0.025/0.128	0.033/0.238

†RF20/5 represents T=20 (number of trees) m_{try}=5 (the number of randomly chosen variables considered at each split in the trees).