Bioinformatics Advance Access published February 5, 2008

Prediction of Recursive Convex Hull Class Assignments for Protein Residues

Michael Stout[1], Jaume Bacardit[1], Jonathan D. Hirst[2] and Natalio Krasnogor[1*]

[1] Automated Scheduling, Optimization and Planning research group, School of Computer Science,
University of Nottingham, Jubilee Campus, Wollaton Road, Nottingham, NG8 1BB, UK
Email: {jqb,mqs,nxk}@cs.nott.ac.uk
[2] School of Chemistry, University of Nottingham, University Park, Nottingham NG7 2RD, UK
Email: jonathan.hirst@nottingham.ac.uk

Associate Editor: Prof. Burkhard Rost

ABSTRACT

Motivation: We introduce a new method for designating the location of residues in folded protein structures based on the Recursive Convex Hull (RCH) of a point set of atomic coordinates. The RCH can be calculated with an efficient and parameterless algorithm.

Results: We show that residue RCH class contains information complementary to widely studied measures such as solvent accessibility (SA), residue depth (RD), and to the distance of residues from the centroid of the chain, the residues' exposure (Exp). RCH is more conserved for related structures across folds and correlates better with changes in thermal stability of mutants than the other measures. Further, we assess the *predictability* of these measures using three types of machine learning technique: decision trees (C4.5), Naive Bayes and Learning Classifier Systems (LCS) showing that RCH is more easily predicted than the other measures. As an exemplar application of predicted RCH class (in combination with other measures) we show that RCH is potentially helpful in improving prediction of residue contact numbers.

Contact: nxk@cs.nott.ac.uk

Supplementary Information: Datasets: www.infobiotic.net/datasets, RCH Prediction Servers: www.infobiotic.net

1 INTRODUCTION

Prediction of the three-dimensional structure of proteins from their constituent amino acid sequences continues to be one of the key goals of structural biology and a wide range of predictive strategies has been investigated. Steady improvements in predictive accuracy have resulted from decomposition of the problem into subproblems, such as prediction of secondary structural elements (approaching a theoretical prediction limit of 80% (Dor and Zhou, 2007; Wood and Hirst, 2005)), of residue coordination number (at over 80% (Bacardit *et al.*, 2006)) and of residue solvent accessibility (at over 77% using consensus predictors (Gianese and Pascarella, 2006)). Burial of hydrophobic groups within the protein core is a primary driving force for protein structure formation. Characterizations of residue accessibility to solvent are, therefore, important for protein structure prediction (PSP), potentially helping to constrain the search space to be explored using *de novo* methods (Baldi and Pollastri, 2002). Whilst classifying residue neighbourhood density as high or low will generally assign the high class to residues buried within the structure and the low class to residues exposed on the surface, residues lining cavities in the structure (that may be functionally significant (Chen *et al.*, 2007) can have a low coordination number even when located far from the surface. Incorporation of complementary residue solvent accessibility and residue depth information improves fold recognition (Liu *et al.*, 2007). A range of measures of residue location have been studied. Lee and Richards (1971) used a spherical probe method to measure the solvent accessible surface of residues and recently Kawabata and Go (2007) have used adjustable probe parameters to identify putative ligand binding pockets on protein surfaces. Solvent accessibility, however, is difficult to compute and does not distinguish between residues below the surface. Hence, atom/residue depth (RD), the distance of an atom/residue from its nearest solvent accessible neighbour, was introduced (Chakravarty

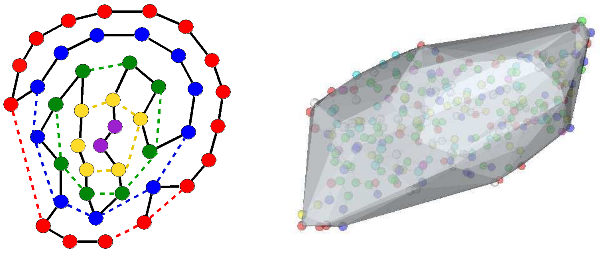

Fig. 1. Left: Recursive Convex Hulls of a 2D off-lattice protein model. The backbone is represented by coloured circles joined by solid black lines. Residues on the outermost recursive convex hull are coloured red, subsequent recursive convex hulls are coloured blue, green, and yellow, with residues on the innermost recursive convex hull coloured purple. Right: A graphical representation of the outer Recursive Convex Hull of residues in a 3D model of a natural protein chain (PDB Id. 1P4X).

[*] to whom correspondence should be addressed

and Varadarajan, 1999) and efficient algorithms are available to compute RD for a given structure (Pintar *et al.*, 2003; Vlahovicek *et al.*, 2005). Whilst SA emphasises burial, RD emphasises exposure and depends on the method used to identify surface atoms/residues. Hence, Half Sphere Exposure (HSE), has been recently proposed (Hamelryck, 2005). HSE, like CN, counts neighbouring residues but distinguishes two regions (half spheres) around each residue based on the C_α-C_β vector, i.e. a 2D measure of residue location. In addition, the distance (exposure) of residues from the chain centroid is a potentially interesting measure being related to the location of catalytic residues in enzyme structures (Ben-shimon and Eisenstein, 2005). Measures of atom/residue location typically depend on specific parameters such as probe size for SA or contact radius for CN.

In this paper, we introduce a new approach to stratifying residues in protein structures by *recursively* identifying the convex hull layer to which each residue belongs. The convex hull of a set of points is a parameterless, mathematically rigorous and unambiguous approach to identifying the points on the exterior of a point set, analogous to identifying those points that contact the enclosing surface when the point set is tightly wrapped. The convex hull is simple and efficient ($O(n * logn)$) to compute (Preparata and Hong, 1977). The recursive convex hull (RCH) of a point set is obtained by identification of the minimal point set that generates the convex hull (the vertices) and removal of these points from the point set followed by recursively applying these steps to the remaining points to identify subsequent hulls. Applied to the point set of coordinates of residues in a protein chain, a series of hulls is obtained that group residues by their distance from the convex surface of the structure. The recursive convex hulls of a 2D off-lattice protein model are shown in Figure 1 along with a representation of the outer convex hull of a 3D point set derived from the C_β atomic coordinates of residues in a real protein chain.

Convex hulls have found a wide range of applications in studies of molecular structure. Here we give a brief, by no means complete, review. Badel-chagnon and colleagues introduced a notion of the ”molecular surface convex hull” to define the depth of any molecular surface point (Badel-chagnon *et al.*, 1994) and Lin and colleagues used convex hulls to align 11 randomly generated bioactive tachykinin peptides, finding that 3D convex hulls can be used to align even these flexible structures (Lin *et al.*, 1999; Lin and Lin, 2001). Meier *et al.* proposed a convex hull based segmentation technique (that makes few assumptions about the underlying surface) to find characteristically shaped regions of molecular surfaces for prediction of possible protein docking sites (Meier *et al.*, 1995). Liang and Dill used convex hulls to define the boundaries of surface pockets and depressions in studies of packing densities in proteins (Liang and Dill, 2001). Holmes and Tsai tackled protein side-chain packing and interactions by measuring variation in convex hulls constructed around these groups (Holmes and Tsai, 2005). Coleman and Sharp introduce the notion of travel depth (the physical distance a solvent molecule would have to travel from a surface point to a suitably defined reference surface) using convex hulls of surface points (Coleman and Sharp, 2006). Recently, Lee and colleagues have employed 3D convex hulls around complementarity regions of antibodies to analyse binding sites (Lee *et al.*, 2006) and Wang *et al.* have used convex hulls of protein backbones in neural network based classification of protein structures. (Wang *et al.*, 2006). However, dissection of protein structures by *recursively* assigning convex hull numbers to residues, as we propose here, does not appear to have been previously reported.

This paper has two parts. In the first part we analyse RCH as a new computable property of proteins. We compare the information content of RCH to that of residue solvent accessibility (SA), residue depth (RD) and exposure (Exp) and show that, although not totally unrelated, these properties are indeed complementary. We show that RCH correlates better with structural conservation than the other measures of residue location and that RCH is also better correlated with changes in protein thermal stability in the presence of cavity forming mutations. We turn, in Part 2, to the question of how easy/difficult it is, in practical terms, to learn to predict these measures. The relative predictability of RCH, RD, SA and Exp using four different machine learning algorithms was assessed using six different, progressively richer, sets of input attributes at three levels of precision. The relative benefits of using these various inputs are described. C4.5 (Quinlan, 1992), Naive Bayes (John and Langley, 1995), GAssist (Bacardit, 2004) and BioHEL (Bacardit *et al.*, 2007) are the machine learning methods employed in this paper. Finally, we demonstrate the usefulness of RCH by using the predicted RCH class of residues as input for prediction of residue Coordination Number (CN) showing that, in combination with predicted residue SA and Exp class, predicted RCH information increases predictive accuracy for CN.

2 MATERIALS AND METHODS

2.1 Datasets and Features Studied

We describe next the datasets and algorithms employed to assess the novelty of RCH and its relation to previously studied measures. All of the measures studied are based on atomic coordinates. Two polypeptides that have similar structures when represented using C_α coordinates may have distinct structures when represented using C_β coordinate (Eidhammer *et al.*, 2003). Throughout this paper C_β atom coordinates are used (C_α for glycyl residues) as these are sensitive to the orientation of side chain atoms.

Protein dataset The dataset used here are those described by Bacardit et al. (Bacardit *et al.*, 2006), originally proposed by Kinjo (Kinjo *et al.*, 2005). Protein chains were selected from PDB-REPRDB (a non-redundant curated subset of the Protein Data Bank (PDB) (Noguchi *et al.*, 2001), covering the space of possible folds) using the following criteria: less than 30% sequence identity, sequence length greater than 50 residues, no membrane proteins, no non-standard residues, no chain breaks, resolution better than 2Å and a crystallographic R factor better than 20%. Chains that had no entry in the HSSP (Sander and Schneider, 1991) database were discarded. The final dataset contains 1050 protein chains (257560 residues).

Identification of Residue Recursive Convex Hulls Convex hulls were identified from the residue C_β atomic coordinates using the QHull package (Barber *et al.*, 1996). Hulls were iteratively identified, hull residues were assigned a hull number and removed from the point set. This being repeated until all residues had been assigned a hull number. The mean RCH number in this dataset was 2.6 (s.d. 2.3). Assignment of RCH numbers to the 1050 chains took 52 minutes. We term this numbering of hulls, from the outermost inward, residue RCH. An alternative numbering scheme, from innermost hull outward, termed RCHr are given in the Supplementary Material (Section 2.1). The mean RCHr number in this dataset was 5.1 (s.d. 2.7). Assigning RCHr numbers to all chains took 58 minutes.

Calculation of Residue Solvent Accessibility (SA) Solvent accessible surface values for each residue were extracted from the DSSP (Holm and Sander, 1993) file for each structure. These values were divided by the solvent accessible surface values for each amino acid as defined in Rost and

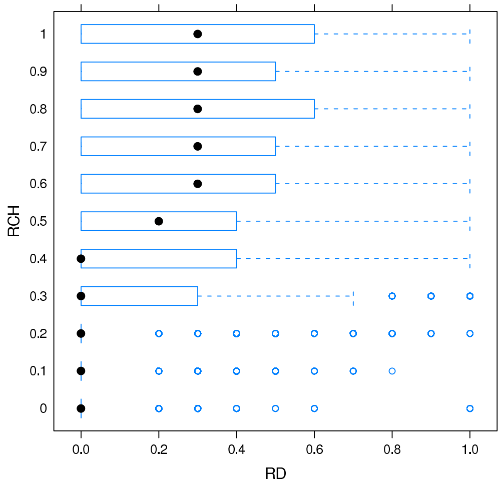

Fig. 2. Box and whisker plots of RD against RCH for 257560 residues from 1050 proteins. Black dots indicate median values. Values were normalized and rounded to one decimal place.

Sander (1994) to obtain the relative solvent accessibility of each residue. The mean SA value in this dataset was 0.27 (s.d. 0.27).

Calculation of Residue Exposure (Exp) In this study, we characterise residue exposure as the distance of residues from the centroid of each chain. (Ben-shimon and Eisenstein, 2005). The chain centroid was determined from the coordinates of the residues and the euclidean distance of each residue from this point was calculated to obtain the residues exposure value. The mean Exp value in this dataset was 19.1$\mathring{A}$ (s.d. 7.8). Determination of Exp values for the whole dataset took less than 2 minutes.

Calculation of Residue Depth (RD) Residue depth (RD) values were obtained from the DPX server (Pintar *et al.*, 2003) using default settings. RD values were positively skewed with a mean RD of 0.86 (s.d. 1.41).

Normalization In Section 2.2 both unnormalized and normalized values are reported for characterisation of the measures studied using box plots (Figure 2), correlation coefficients (Table 1), structural conservation (Table 2), thermal stability (Table 3) and mutual information between class assignments (Table 4). The value for each residue was divided by the maximum value for that measure in the corresponding chain to obtain the normalized value. Histograms of unnormalized and normalized measures are shown in the Supplementary Materials (Figures 5 and 6). After normalization RCH and RCHr are symmetric.

2.2 Comparison Between RCH and Other Measures of Residue Location

BoxPlots Figure 2 plots RD versus RCH for each residue in the dataset using the statistically robust Box and Whisker technique. Boxes cover 50% of the data points, whiskers extend to 1.5 times the interquartile range with outliers plotted as blue dots and median values indicated with black dots. Median values for RD are positively correlated with RCH yet RCH makes finer distinctions between degrees of burial and exposure. Further box plots for these measures are available in the Supplementary Materials (Figure 3).

Correlation coefficients Pairs of measures that have a low correlation coefficient are likely to be unrelated and potentially provide complementary information for PSP. Table 1 shows the Pearson correlation coefficients between the measures studied. RD has low correlation with the other measures. RCH is most highly anti-correlated with SA (-0.62) and has a higher correlation with SA and Exp than RD. RCH is not highly correlated with RD, suggesting that these are distinct characterisations of residue location. RCH appears to be the measure that correlates closely to many of the other measures. Hence, we would like to determine whether it is relatively more learnable than these other measures.

Table 1. Correlation Coefficients between Measures Studied. Norm. indicates coefficients based on normalized measures.

	SA	RD	Exp	RCH	RCHr	
SA	1.00	-0.51	0.39	-0.62	0.41	
	1.00	-0.50	0.55	-0.68	0.68	Norm.
RD		1.00	-0.26	0.43	-0.30	
		1.00	-0.34	0.48	-0.48	Norm.
Exp			1.00	-0.41	0.85	
			1.00	-0.81	0.81	Norm.
RCH				1.00	-0.42	
				1.00	-1.00	Norm.
RCHr					1.00	
					1.00	Norm.

Table 2. Conservation of Measures. Correlation of the Measures Studied between aligned residues in related structures. Norm. indicates coefficients based on normalized measures.

	RD	Exp	RCH	RCHr	SA
	0.37	0.38	0.46	0.48	0.52
Norm.	0.37	0.46	0.55	0.55	0.50

Table 3. Correlation of Structural Features with Thermal Stability. Correlation of the measures studied with changes in thermal stability of mutant proteins. Norm. indicates coefficients based on normalized measures.

	RD	Exp	RCH	RCHr	ΔASA
	-0.22	0.29	-0.38	0.29	-0.34
Norm.	-0.20	0.44	-0.35	0.35	-0.37

Conservation of RCH For related proteins, aligned residues are potentially conserved even in the absence of strong sequence homology. Measures that have relatively high correlation for aligned residue pairs potentially reflect conserved aspects of protein structure. We, therefore, assess to what degree these measures are correlated between aligned residues in pairs of superimposed structures from a range of folds. Following (Hamelryck, 2005), the conservation of RCH and the other measures was calculated for 15621 aligned residues (BLAST E-value $>=$ 1.0) in 218 pairs of structures from the SABmark version 1.63 Twilight Zone database (Van Walle *et al.*, 2005). This dataset comprises pairs of superimposed structures covering 236 folds. These pairs are structurally similar yet are without probable common evolutionary origin, effectively, a hard dataset to predict. Table 2 reports the correlation coefficients for both unnormalized and normalized measures. RCH and RCHr have higher conservation correlation coefficients than RD, Exp and SA indicating that, for such aligned residues, RCH is more highly correlated with structurally conserved locations than RD, Exp and (after normalisation) SA. As we used C_β coordinates, values for RD and SA are around 0.1 lower than those previously reported (Hamelryck, 2005).

Table 4. Pairwise Mutual Information. MI between two class (Q2) assignments for pairs of measures. Norm. indicates MI for class assignments based on normalized measures.

	SA	RD	Exp	RCHr	RCH	
SA	1.00	0.21	0.06	0.08	0.26	
	1.00	0.21	0.12	0.26	0.26	Norm.
RD		0.91	0.04	0.05	0.14	
		0.91	0.06	0.14	0.14	Norm.
Exp			1.00	0.38	0.07	
			1.00	0.29	0.29	Norm.
RCHr				0.99	0.08	
				1.00	1.00	Norm.
RCH					0.99	
					1.00	Norm.

Relationship of RCH to Changes in Thermal Stability of Mutant Proteins Changes in thermal stability of proteins after mutations of core hydrophobic residues (that potentially lead to cavity formation) has been correlated with changes in SA and residue depth (for references see Hamelryck (2005)). For such residues, measures that correlate relatively highly with changes in the proteins thermal stability reflect structurally important features. We, therefore, assess for these residues the degree to which these measures are correlated with changes in protein thermal stability. The correlation of these measures of residue location (both normalized and unnormalized) with changes in the thermal stability ($\Delta\Delta G$ in kcal/mol) of 91 Ile/Leu/Val to Ala point mutations was measured. 16 protein structures from the Protherm database (Gromiha *et al.*, 1999; Bava *et al.*, 2004; Kumar *et al.*, 2006) were employed, again following the approach of Hamelryck (2005). The correlation coefficients for RD, SA, Exp, RCH, RCHr and ΔASA (related to the change in accessible surface upon folding) are shown in table 3. RD values were similar to those previously reported. RCH is more highly correlated with changes in thermal stability upon mutation than the other measures. Exp and ΔASA showed higher correlation when the data was normalized. RD showed the lowest correlation of the measures studied. This data indicates that (unnormalized) RCH is correlated more strongly with residues in the hydrophobic core (that are related to structural stability) than are the other measures.

Mutual Information The degree to which the classes assigned to residues using these measures are mutually informative was assessed using Mutual Information (MI) (Cover and Thomas, 2006). For discrete data, MI is defined as:

$$I(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\frac{p(x,y)}{p(x)\,p(y)}, \qquad (1)$$

where $p(x)$ and $p(y)$ are the probabilities of x and y occurring in the dataset, and $p(x,y)$ is the probability of the combination of x and y occurring together in the dataset. MI is used here to measure the quantity of information that one measure (e.g. SA) tells us about another (e.g. RCH).

Table 4 shows the MI between pairs of measures for all 257560 residues studied. When the MI between the class assignments for a pair of measures is high they represent closely related problems (the MI between a measure and itself is maximal, and is 1.00 if the classes assigned to the measure are well balanced). SA shares 0,26 MI with RCH whilst Exp shares 0.38 MI with RCHr and all other pairwise MI values are less than 0.10. This indicates that the RCH class of residues provides information distinct to SA, RD and Exp class information. MI for Q3 and Q5 class assignments is given in the Supplementary Materials (Table 3) along with a detailed pairwise examination of the Q5 class assignments for SA vs. RCH, and RCHr vs. Exp, where increased levels of MI were observed (Supplementary Materials, Tables 4, 6) along with RD vs. RCH (and 5). Frequent differences in class assignments are observed for measures with greater than 0.20 MI.

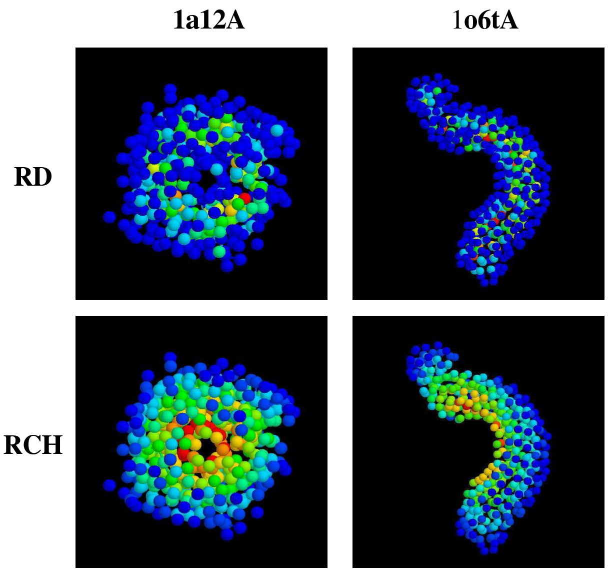

Fig. 3. Space filling C_β atom models of proteins coloured by RCH and RD. 'Core' residues are coloured red/yellow and 'surface' residues blue/green (rendered using RasMol).

To further highlight the distinction between RD and RCH, visualisations of two space filling C_β atom models of protein structures are shown in figure 3. The values for each measure were normalized and the colour assigned, in both measures, to indicate values from "exposed" (blue) to "buried" (red). These models provide visual confirmation that residue RCH assignments are distinct to those for RD. Further examples are available in the Supplementary Materials (Figures 1 and 2).

3 LEARNABILITY OF RCH AND OTHER MEASURES

Having demonstrated that residue RCH is a new and distinct characterisation of residue location, we turn to the predictability of these measures and assess, in practical terms, which of these characterisations of residue location is easier to learn. Hence, potentially more useful for PSP.

3.1 Prediction Experiments

Inputs to Predictions For each measure (RCH, RCHr, RD, SA and Exp) predictions were made using six types of input information and three levels of precision: two, three and five class partitions (Q2, Q3 and Q5). Table 5 summarises the six different types of input information used for predictions of the measures studied. Combinations of both local (neighbourhood of the target in the chain) and global (protein-wise) information were used. A window of four residues either side of the target residue has been shown to lead to high CN predictive accuracy using LCS (Bacardit *et al.*, 2006) and was used in this study also to facilitate comparison of results. For each representation (RCH, SA etc.) these inputs were labeled 1-6 in the rest of this paper, for example, RCH-3 denotes RCH predicted using input dataset 3. For each measure a total of 18 datasets was evaluated (six sets of input attributes each at three levels of class assignment). A detailed description of these inputs appears in Stout *et al.* (2007).

Table 5. Datasets. For each dataset (1-6) the input information type included in that dataset is indicated by •. The two types of local (target and its closest neighbours) and three types of global (protein-wise) input information were investigated are shown.

		Dataset					
Scope	**Input Information**	**1**	**2**	**3**	**4**	**5**	**6**
Local	AA Types in Window of target±4 residues	•	•	•	•	•	•
	Pred. Secondary Structure of target		•		•		•
Global	Chain Length			•	•		
	Residue Frequencies			•	•		
	Pred. average of target measure					•	•

Table 6. Summary of the highest predictive accuracies for each measure studied, in descending order of accuracy. Mean ± s.d. for ten fold cross validated predictions based on the input datasets that gave the best results for each measure: namely type 4 or 6 (indicated by ○).

Alg	**C4.5**	**BioHEL**	**GAssist**	**Naive Bayes**
RCHr	79.8±1.5	78.5±1.5	78.4±1.5○	77.9±1.7○
RCH	77.3±1.0	75.9±1.2○	75.7±1.1	76.1±1.1○
RD	76.0±0.4○	75.3±0.3	75.2±0.3○	75.1±0.4
Exp	73.9±1.4	72.8±1.6	72.5±1.4○	73.4±1.3○
SA	73.3±0.3	72.2±0.4○	72.2±0.4○	72.3±0.4○

In order to determine the degree to which RCH and RCHr vary in their learnability and capture properties of protein structures, in what follows we use their unnormalized versions.

Predicted secondary structure information of the target residue was obtained using the PSI-PRED predictor (Jones, 1999). This consists the secondary structure type (helix, strand or coil) and a confidence level ([0..9]) of the prediction.

For each measure, the average value of that measure was determined for each chain and ten pairs of training and test folds were prepared. For each instance, the inputs were the chain length (one integer value) and amino acid composition of each chain (20 real values) and the target class was the measured average value for the particular measure (partitioned into 10 classes using a uniform frequency cut-point strategy). Cut-points were determined separately for each training fold and used to assign classes to the values in the corresponding training and test folds. The predicted average value of the measure under consideration (termed PredAveRCH, PredAveRD etc.) was predicted using the GAssist LCS (details below) prior to preparation of the data sets for the full measure predictions. 950 instances (chains) were used for training and 100 instances for testing. 10 iterations were performed for each prediction using different random number seeds and the 10 rule sets generated were combined as an ensemble using a majority vote to predict the measure.

Class Assignments In order to predict measures using classification techniques, the calculated values for each measure were partitioned into two, three and five classes (bins) here termed Q2, Q3 and Q5 respectively. For imbalanced measures, such as SA, a class boundary that leads to more balanced classes is traditionally chosen, e.g. for SA a cut point of 25% is widely used. We apply class balancing for all measures and levels of discretization (Q2, Q3 and Q5) in this study, adopting a uniform frequency classification procedure. For our data, balanced classes for SA were obtained using, for example, a cut point of 18%. Class boundaries were determined individually for each training/test set pair using the corresponding training fold. Details of the cut points used are given in the Supplementary Materials (Table 1).

Definition of the training and tests sets Datasets were divided randomly into ten training and test set pairs (950 chains for training and 100 for testing), using bootstrap (Kohavi, 1995). We have placed a copy of the datasets at www.infobiotic.net.

Performance measures Different protein chains have different lengths and it is prediction accuracy on chains that is typically reported (Kinjo *et al.*, 2005; Jones, 1999). Prediction accuracies for each chain were, therefore, averaged to obtain the protein-wise accuracy reported here.

Machine Learning Methods We use four different machine learning (ML) methods. The first two are popular ML systems taken from the WEKA package (Witten and Frank, 2005): C4.5 (Quinlan, 1992), a decision tree rule induction system, Naive Bayes (John and Langley, 1995), a Bayesian learning algorithm. Learning systems belonging to the Learning Classifier Systems (LCS) (Holland and Reitman, 1978) class of ML techniques were also studied. These systems are rule-based machine learning systems that employ evolutionary computation (Holland, 1975) as the search mechanism. Two LCS methods have been employed: GAssist (Bacardit, 2004) and BioHEL (Bacardit *et al.*, 2007) that implement different rule induction paradigms. A detailed description of both systems is included in the Supplementary Material (Section 3.1).

Analysis of Results For each experiment, the mean prediction accuracy (as defined in section 2) over the test sets is reported. Student t-tests were applied to the ten results from each experiment to determine the best method for each dataset at a confidence level of 95%. Standard deviations and any significant differences are indicated in each table. The conservative Bonferroni correction (Miller, 1981) for multiple pair-wise comparisons was applied.

In addition, the contributions of global input information were assessed as follows: for each learning system and precision (Q2, Q3 and Q5), the maximum of (Dataset4, Dataset6)-Dataset2 was computed. As a base for the performance gap, the dataset with predSS was used, because in certain situations the Dataset1 performed poorly, distorting the comparisons. Finally, the contribution of predicted secondary structure was also assessed as follows: for each learning system and number of states the value of the maximum of (Dataset2-Dataset1, Dataset4-Dataset3, Dataset6-Dataset5) was determined.

3.2 Prediction Results

For each measure studied, Table 6 summarizes the best Q2 predictive accuracy (in descending order) for each measure (using the best possible input dataset in each case). Detailed results for the predictions are given in the Supplementary Materials (Tables 7 - 10). Predictive accuracy was higher on the two RCH based representations than on the SA, RD or Exp representation. The predictive accuracies for RCHr being statistically significantly higher than those for the other measures (p-value=0.5).

For all representations, higher predictive accuracies were seen when fewer classes were predicted (lower precision – Q2). Q5 predictive accuracy for RCH was between 30% and 40%, Q3 was approximately 20% higher, between 55% and 60% whilst for Q2 prediction accuracies exceeded 77%. The LCS's performed best on the RCHr representation when using input dataset RCHr-4. This dataset combines local information (a window of residues around the target and its predicted secondary structure) with global chain information (chain length and chain residue composition). The more compact RCHr-6 was frequently the most learnable dataset for C4.5 and Naive Bayes. This dataset comprises local information (window and predicted secondary structure) and global information (predicted average RCHr of the chain).

3.3 Predicted RCH Improves CN Prediction

Table 7. Coordination Number prediction (by BioHEL) using amino-acid sequence and various combinations of the predicted measures. ● indicates input information that leads to statistically significant increases in predictive accuracy compared to the baseline CN1 inputs. The group of best performing methods (▼) all have statistically similar performance.

Dataset	Protein-wise acc.
CN1	77.2±0.8
CN1+RD	77.4±0.8
CN1+RCHr	77.6±0.7
CN1+Exp	77.7±0.8
CN1+Exp+RCHr	77.7±0.7
CN1+RCH	78.5±0.9
CN1+RCH+RCHr	78.8±0.7●
CN1+Exp+RCH	78.9±0.8●
CN1+SA	78.9±0.8●
CN1+Exp+RCH+RCHr	78.9±0.7●
CN1+Exp+SA	79.1±0.8●▼
CN1+Exp+SA+RCHr	79.1±0.8●▼
CN1+SA+RCHr	79.1±0.8●▼
CN1+SA+RCH	79.7±0.8●▼
CN1+Exp+SA+RCH	79.8±0.8●▼
CN1+SA+RCH+RCHr	79.8±0.8●▼
CN1+Exp+SA+RCH+RCHr	79.8±0.7●▼

Finally, we assess the utility of predicted RCH as an input to prediction of other aspects of protein structure, specifically Coordination Number (CN). For each of the measures studied, the Q5 predictions (using input dataset 4) made by BioHEL (which was, in general, the best performing method) are fed back into prediction of CN (Bacardit *et al.*, 2006). The CN of a residue is a count of the number of other residues from the chain that are located within a certain threshold distance. Specifically, we have used the Kinjo *et al.* (2005) definition of CN. We predict whether the CN of a residue is above or below the midpoint of the CN domain, using as input information the AA type of a window of ±4 residues around the target (equivalent to the first set of input attributes used to predict the other features), *CN1*.

The contribution of SA, Exp, RCH and RCHr to CN prediction (individually and in combination with one another) was evaluated by extending the CN1 dataset with 16 combinations of input attributes that correspond to all combinations of these measures. Using predicted RD as input gave the lowest improvement (0.2%) over the CN1 (local window) input alone and was, therefore, not included in predictions made with combinations of inputs. Table 7 shows the results of these experiments. As a baseline, the performance of the original CN1 is included. The table has been sorted by accuracy to help identify the combinations of predicted measures that give the biggest performance boost.

The results of these experiments were analyzed using paired t-test with 95% confidence level and the Bonferroni correction. Two types of results were identified: the datasets in which BioHEL performed significantly better than when learning from CN1 (marked with a ●) and the (statistically indistinguishable) group of datasets that resulted in the highest predictive accuracies are indicated (▼).

There are two groups of measures: those that only provide a small performance boost over CN1 (RD, Exp and RCHr), and others that provide a larger boost (SA and RCH). Furthermore, combining Exp and RCHr (together and with other measures) only marginally improves the performance, indicating that these measures are mostly redundant (consistent with the observed increase in MI between these two measures, table 4). On the other hand, combining SA and RCH resulted in much better accuracy, showing that these measures complement each other. The best combinations of measures (Exp+SA+RCH, SA+RCH+RCHr and Exp+SA+RCH+RCHr) lead to a performance increase of 2.6% over the baseline CN1 dataset.

4 DISCUSSION

4.1 Prediction Results

The results show some general trends across all measures studied and all learning methods. Predictive accuracy is increased when richer input information is employed. Inclusion of local information in the form of predicted secondary structure typically leads to an increase in Q2 predictive accuracy of 2-3% on most datasets for the learning systems used, whilst using global protein information (chain length and composition) can boost Q2 predictive accuracy by more than 10% (in the case of RCH and RCHr). The type 3 and 4 datasets, containing 21 real valued global protein attributes, in particular present a considerably larger search space, and the mixture of real-valued and nominal attributes makes the learning problem more difficult than for purely nominal knowledge representations. The type 5 and 6 datasets use less attributes than the type 3 and 4 datasets enabling the LCS's to generate rule sets that are more easily interpreted. The RCHr representation was the most predictable measure this resulting from use of input dataset 4 (chain composition and length information) and the BioHEL LCS. C4.5 performed best when using datasets 4 and 6, benefiting from predicted local and global information. Naive Bayes generally performed best when using the more compact dataset 6.

There were interesting differences between these representations of residue location; RCH was easier to learn than the other representations perhaps due to this representation correctly assigning classes to residues for anisotropic (elongated) structures. For such structures, the Exp representation may assign some surface residues to the buried class and some buried residues to the exposed class. Similarly, RCH was more predictable than SA; assignment of residues deep within structures but on the surface of cavities (high solvent accessibility) to the exposed class may have lowered the predictability of the SA representation.

RD was more predictable than SA and Exp but was not as easily predicted as RCH and RCHr. As SA emphasises buried residues, so RD emphasises exposed residues. The highly imbalanced nature of the RD measure (Supplementary Materials, Figures 5 and 6) leads to imbalanced class assignments even when using a uniform frequency cut-point strategy (Supplementary Materials, Table 2). This imbalance is likely to have made this measure relatively more difficult to learn. Furthermore, when fed back as inputs for prediction of CN, RD in particular provided little additional information over the CN1 inputs resulting in only marginal improvements in CN prediction. Prediction accuracy for all measures is likely to be boosted by including additional (e.g. non-local) input information. For example, 79.3% ten-fold cross validated accuracy has been reported for two state SA prediction at a 25% cut point using an integrated system of neural networks (Dor and Zhou, 2007) with position specific scoring matrices and a range of other input data.

4.2 ”White Box” Prediction: Interpretable Analysis and Performance Considerations

Understanding the basis on which a prediction is made may be more valuable than making relatively accurate predictions in a blind manner. Decision trees can be interpreted, however, on these problems C4.5, produced pruned trees that ranged in mean size from 1068 (sd=331) to 138977 (sd=2735) nodes limiting their explanatory value. Naive Bayes, on the other hand, generates compact solutions, however, these are probabilistic models that have no immediate physico-chemical interpretation. In contrast, it is much easier to relate the compact rule sets evolved by the LCS algorithms to the underlying physical and chemical properties of proteins. The following is an example of a rule set evolved by the GAssist LCS that produced 71.2% two state predictive accuracy on the RCH-6 dataset.

1. If $PredSSConf < 7.5$ and $PredAveAtt > 8.64$ and $Res_{-4} \notin \{K, x\}$ and $Res_{-2} \notin \{x\}$ and $Res \notin \{D, E, K, N, Q\}$ and $Res_{+1} \notin \{x\}$ and $Res_{+2} \notin \{K\}$ and $Res_{+4} \notin \{x\}$ $\longrightarrow$ class is 1
2. If $PredSS \notin \{C\}$ and $PredAveAtt > 4.8$ and $Res_{-4} \notin \{E, Q\}$ and $Res_{-3} \notin \{D, T\}$ and $Res_{-2} \notin \{E\}$ and $Res_{-1} \notin \{D, P, V\}$ and $Res \notin \{D, E, H, K, N, P, Q, R, T\}$ and $Res_{+1} \notin \{x\}$ and $Res_{+2} \notin \{H\}$ and $Res_{+3} \notin \{K\}$ and $Res_{+4} \notin \{E, K, Q, R, x\}$ $\longrightarrow$ class is 1
3. If $PredSS \notin \{C\}$ and $PredAveAtt > 4.8$ and $Res_{-4} \notin \{E, R\}$ and $Res_{-3} \notin \{D, E\}$ and $Res_{-1} \notin \{P, W\}$ and $Res \notin \{A, D, E, K, N, P, Q, R\}$ and $Res_{+1} \notin \{W, x\}$ and $Res_{+2} \notin \{V, W\}$ and $Res_{+3} \notin \{K, R\}$ and $Res_{+4} \notin \{K, x\}$ $\longrightarrow$ class is 1
4. If $PredSS \in \{E\}$ and $PredAveAtt > 4.5$ and $Res_{-4} \notin \{C, K, x\}$ and $Res_{-3} \notin \{R\}$ and $Res_{-2} \notin \{K, R\}$ and $Res_{-1} \notin \{D, Q, S, x\}$ and $Res \notin \{E, F, K, M, R, T\}$ and $Res_{+1} \notin \{F, L, x\}$ and $Res_{+2} \notin \{K\}$ and $Res_{+3} \notin \{C\}$ and $Res_{+4} \notin \{x\}$ $\longrightarrow$ class is 1

   ....

   Default class is 0

The rule set structure is a decision list, rules are tested in order starting with the first and rule evaluation continues until a match is found or the default (last) rule is reached. This rule set comprised 15 rules, the first four rules are shown (the full rule set is shown in the Supplementary Materials, Section 4.3). In the first rule, the confidence level of the secondary structure prediction is tested then the predicted average value for the property under consideration (predicted average RCH > 8.64). Subsequent predicates test the amino acid types of residues surrounding the target in the sequence. $AA_{\pm}M$ means the amino acid type of the residue at position $\pm M$ in respect to the target residue. Amino acids are represented by their one letter code, plus the symbol x representing positions after the start/end of the chain, for cases where the window of neighbouring residues overlaps the beginning or the end of the chain. For positions relative to the target residue (eg. $Res_{-}2$) these predicates restrict the amino acid types for the residue at that position (membership or non-membership of a set). In subsequent rules, the predicted Secondary Structure class (PredSS) of the target residue; either Helix (H), Sheet (E) or Coil (C) is tested. In many cases, the target residue type is largely restricted to hydrophobic residues that are often found buried in the protein core and, therefore, have higher RCH numbers. The LCS has correctly identified this, predicting the above average RCH class (1) for these residues.

The accuracy, simplicity and interpretability of the rule sets generated by the LCS's must, however, be balanced against the computational expense needed to generate them. Run times for the learning phase of the BioHEL algorithm ranged from three minutes on the smaller input datasets to almost seven hours on the larger ones. In contrast, the worst case for C4.5 was 32 minutes and for Naive Bayes was less than one minute. Moreover, for each problem set, the LCS algorithms, BioHEL and GAssist, were run multiple times to produce ten classifiers for input to the ensemble procedure. Once trained, however, run times for the resulting classifiers on the entire test set was around two minutes for C4.5 and Naive Bayes but less then one minute for the LCS's, an indication, for instance, of how these LCS evolved classifiers would perform as part of a prediction web server.

5 CONCLUSIONS

In this paper, a new measure of residue location in folded protein chains, the recursive convex hull (RCH), was introduced. RCH is a parameterless, simple to compute and mathematically rigorous method that situates residues in layers within protein structures. We show that RCH is distinct to other widely studied measures of residue location and that RCH distinguishes a range of degrees of residue burial/exposure, correlates better with residue conservation and changes in protein stability under mutation than measures such as solvent accessibility, residue depth or residue distance from chain centroid. Further, we assess the predictability of these measures using three types of machine learning technique: decision trees (C4.5), Naive Bayes and Learning Classifier Systems (LCS), employing a range of predictive inputs. We show that an LCS that employs iterative rule learning, BioHEL, predicts RCH at 77.3%, 60.6% and 39.0% accuracy for Q2, Q3 and Q5, respectively. We present examples of the competent yet simple and interpretable LCS classification rules, showing how they relate to the underlying physical and chemical properties of the residues. As an exemplar application of predicted RCH class (in combination with other measures) we show that prediction of contact number can be improved by up to 2.6%.

6 ACKNOWLEDGMENTS

We acknowledge the support of the UK Engineering and Physical Sciences Research Council (EPSRC) under grants GR/ T07534 /01.

We are grateful for the use of the University of Nottingham's High Performance Computer. We are grateful to the anonymous reviewers whose insightful comments have helped to improve this manuscript. *Conflict of Interest:* none declared.

REFERENCES

Bacardit, J. (2004). *Pittsburgh Genetics-Based Machine Learning in the Data Mining era: Representations, generalization, and run-time*. Ph.D. thesis, Ramon Llull University, Barcelona, Catalonia, Spain.

Bacardit, J., Stout, M., Krasnogor, N., Hirst, J. D., and Blazewicz, J. (2006). Coordination number prediction using learning classifier systems: Performance and interpretability. In *Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation (GECCO '06)*, pages 247–254. ACM Press, New York, NY.

Bacardit, J., Stout, M., Hirst, J. D., Sastry, K., Llorà, X., and Krasnogor, N. (2007). Automated alphabet reduction method with evolutionary algorithms for protein structure prediction. In *GECCO '07: Proceedings of the 9th annual conference on Genetic and evolutionary computation*, volume 1, pages 346–353, London. ACM Press.

Badel-chagnon, A., Nessi, J., Buffat, L., and Hazout, S. (1994). "iso-depth contour map" of a molecular surface. *J Mol Graph*, **12**(3), 162–8, 193.

Baldi, P. and Pollastri, G. (2002). A machine-learning strategy for protein analysis. *IEEE Intelligent Systems*, **17**(2), 28–35.

Barber, C. B., Dobkin, D. P., and Huhdanpaa, H. (1996). The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software*, **22**(4), 469–483.

Bava, K. A., Gromiha, M. M., Uedaira, H., Kitajima, K., and Sarai, A. (2004). Protherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res*, **32**(Database issue), D120–1.

Ben-shimon, A. and Eisenstein, M. (2005). Looking at enzymes from the inside out: the proximity of catalytic residues to the molecular centroid can be used for detection of active sites and enzyme-ligand interfaces. *J Mol Biol*, **351**(2), 309–26.

Chakravarty, S. and Varadarajan, R. (1999). Residue depth: a novel parameter for the analysis of protein structure and stability. *Structure*, **7**(7), 724–32.

Chen, B. Y., Bryant, D. H., Fofanov, V. Y., Kristensen, D. M., Cruess, A. E., Kimmel, M., Lichtarge, O., and Kavraki, L. E. (2007). Cavity scaling: automated refinement of cavity-aware motifs in protein function prediction. *J Bioinform Comput Biol*, **5**(2a), 353–82.

Coleman, R. G. and Sharp, K. A. (2006). Travel depth, a new shape descriptor for macromolecules: application to ligand binding. *J Mol Biol*, **362**(3), 441–58.

Cover, T. and Thomas, J. A. (2006). *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. John Wiley & Sons Inc.

Dor, O. and Zhou, Y. (2007). Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training. *Proteins*, **66**(4), 838–45.

Eidhammer, I., Jonassen, I., and Taylor, W. R. (2003). *Protein Bioinformatics: An Algorithmic Approach to Sequence and Structure Analysis*. John Wiley and Sons Ltd.

Gianese, G. and Pascarella, S. (2006). A consensus procedure improving solvent accessibility prediction. *J Comput Chem*, **27**(5), 621–6.

Gromiha, M. M., An, J., Kono, H., Oobatake, M., Uedaira, H., and Sarai, A. (1999). Protherm: Thermodynamic database for proteins and mutants. *Nucleic Acids Res*, **27**(1), 286–8.

Hamelryck, T. (2005). An amino acid has two sides: a new 2d measure provides a different view of solvent exposure. *Proteins*, **59**(1), 38–48.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*. University of Michigan Press.

Holland, J. H. and Reitman, J. S. (1978). Cognitive systems based on adaptive algorithms. In D. A. Waterman and F. Hayes-Roth, editors, *Pattern directed inference systems*, pages 313–329. Academic Press, New York, NY.

Holm, L. and Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J Mol Biol*, **233**(1), 123–38.

Holmes, J. B. and Tsai, J. (2005). Characterizing conserved structural contacts by pair-wise relative contacts and relative packing groups. *J Mol Biol*, **354**(3), 706–21.

John, G. H. and Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345. Morgan Kaufmann Publishers, San Mateo.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*, **292**(2), 195–202.

Kawabata, T. and Go, N. (2007). Detection of pockets on protein surfaces using small and large probe spheres to find putative ligand binding sites. *Proteins*, **68**(2), 516–29.

Kinjo, A. R., Horimoto, K., and Nishikawa, K. (2005). Predicting absolute contact numbers of native protein structure from amino acid sequence. *Proteins*, **58**(1), 158–65.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In C. S. Mellish, editor, *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1137–1145, San Mateo. Morgan Kaufmann.

Kumar, M. D., Bava, K. A., Gromiha, M. M., Prabakaran, P., Kitajima, K., Uedaira, H., and Sarai, A. (2006). Protherm and pronit: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res*, **34**(Database issue), D204–6.

Lee, B. and Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J Mol Biol*, **55**(3), 379–400.

Lee, M., Lloyd, P., Zhang, X., Schallhorn, J. M., Sugimoto, K., Leach, A. G., Sapiro, G., and Houk, K. N. (2006). Shapes of antibody binding sites: qualitative and quantitative analyses based on a geomorphic classification scheme. *J Org Chem*, **71**(14), 5082–92.

Liang, J. and Dill, K. A. (2001). Are proteins well-packed? *Biophys J*, **81**(2), 751–66.

Lin, T. H. and Lin, J. J. (2001). Three-dimensional quantitative structure-activity relationship for several bioactive peptides searched by a convex hull-comparative molecular field analysis approach. *Comput Chem*, **25**(5), 489–98.

Lin, T. H., Lin, J. J., and Lu, Y. J. (1999). A comparative molecular field analysis study on several bioactive peptides using the alignment rules derived from identification of commonly exposed groups. *Biochim Biophys Acta*, **1429**(2), 476–85.

Liu, S., Zhang, C., Liang, S., and Zhou, Y. (2007). Fold recognition by concurrent use of solvent accessibility and residue depth. *Proteins*, **68**(3), 636–45.

Meier, R., Ackermann, F., Herrmann, G., Posch, S., and Sagerer, G. (1995). Segmentation of molecular surfaces based on their convex hull. In *ICIP '95: Proceedings of the 1995 International Conference on Image Processing (Vol. 3)-Volume 3*, pages 3552–. IEEE Computer Society.

Miller, R. G. (1981). *Simultaneous Statistical Inference (Springer Series in Statistics)*. Springer-Verlag New York Inc.

Noguchi, T., Matsuda, H., and Akiyama, Y. (2001). Pdb-reprdb: a database of representative protein chains from the protein data bank (pdb). *Nucleic Acids Res*, **29**(1), 219–20.

Pintar, A., Carugo, O., and Pongor, S. (2003). Dpx: for the analysis of the protein core. *Bioinformatics*, **19**(2), 313–4.

Preparata, F. P. and Hong, S. J. (1977). Convex hulls of finite sets of points in two and three dimensions. *Communications of the ACM*, **20**(2), 87–93.

Quinlan, J. R. (1992). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Rost, B. and Sander, C. (1994). Conservation and prediction of solvent accessibility in protein families. *Proteins*, **20**(3), 216–26.

Sander, C. and Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**(1), 56–68.

Stout, M., Bacardit, J., Hirst, J. D., Smith, R. E., and Krasnogor, N. (2007). Prediction of topological contacts in proteins using learning classifier systems. *Soft Computing, Special Issue on Evolutionary and Metaheuristic-based Data Mining (EMBDM), (in press)*.

Van Walle, I., Lasters, I., and Wyns, L. (2005). Sabmark–a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics*, **21**(7), 1267–8.

Vlahovicek, K., Pintar, A., Parthasarathi, L., Carugo, O., and Pongor, S. (2005). Cx, dpx and pride: Www servers for the analysis and comparison of protein 3d structures. *Nucleic Acids Res*, **33**(Web Server issue), W252–4.

Wang, Y., Wu, L.-Y., Zhang, X.-S., and Chen, L. (2006). Automatic classification of protein structures based on convex hull representation by integrated neural network. In *TAMC*, pages 505–514.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition: Practical Machine Learning Tools and Techniques (The Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann.

Wood, M. J. and Hirst, J. D. (2005). Protein secondary structure prediction with dihedral angles. *Proteins*, **59**(3), 476–81.