PREDICTION OF RESIDUE EXPOSURE AND CONTACT NUMBER FOR SIMPLIFIED HP LATTICE MODEL PROTEINS USING LEARNING CLASSIFIER SYSTEMS

MICHAEL STOUT, JAUME BACARDIT, JONATHAN D. HIRST, JACEK BLAZEWICZ AND NATALIO KRASNOGOR*

Automated Scheduling, Optimisation and Planning Research Group, School of Computer Science and IT, University of Nottingham, Jubilee Campus, Wollaton Road, Nottingham, NG8 1BB, UK
Email: {jqb,mqs,nxk}@cs.nott.ac.uk
School of Chemistry, University of Nottingham, University Park, Nottingham NG7 2RD, UK
Email: jonathan.hirst@nottingham.ac.uk
Poznan University of Technology, Institute of Computing Science, ul. Piotrowo 3a, 60-965 Poznan, Poland
Email: jblazewicz@cs.put.poznan.pl

The performance of a Learning Classifier System (LCS) applied to the classification of simplified hydrophobic/polar (HP) lattice model proteins was compared to other machine learning (ML) algorithms. The GAssist LCS classified functional HP model proteins on the 3D diamond lattice as folding or non-folding at 88.3% accuracy, outperforming significantly three out of the four other methods. GAssist correctly classified HP model protein instances on the basis of Contact Number (CN) and Residue Exposure (RE) on both 2D square and 3D cubic lattices at a level of between 27.8% and 80.9%. Again, the LCS performed at a level comparable to the other ML technologies in this task outperforming significantly them in 24 out of 180 cases, and being outperformed just six times. The benefits of using LCS for this problem domain are discussed and examples of the LCS generated rules are described.

1. Introduction

Prediction of structural properties of proteins such as residue exposure (RE) and coordination number (CN) based solely on protein sequence has recently received renewed attention. In other studies, simplified protein

*corresponding author

models such as the HP model have been used to understand protein folding and protein structure prediction. These models represent the sequence of a protein using two residue types: hydrophobic and polar restricting the residue locations to those of a lattice. This paper compares CN and RE prediction for simplified HP model proteins using machine learning technologies, in particular Learning Classifier Systems (LCS). LCS apply Evolutionary Computation to Machine Learning problems. Four questions were examined: 1) Is it possible to predict, from sequence alone, which proteins will and will not fold? 2) Is it possible to predict which residues have above or below average CN and RE? 3) Is it possible to predict the detailed CN and RE states? and 4) Are LCS suitable tools for these tasks?

2. Background

2.1. *Protein Structure Prediction*

The prediction of the 3D structures of proteins is a fundamental and difficult problem in computational biology. Popular approachs include predicting specific attributes of proteins, such as secondary structure, solvent accessibility or coordination number. The contact/coordination number (CN) problem is defined as the prediction, for a given residue, of the number of residues from the same protein that are in contact with it. Two residues are said to be in contact when the distance between the two is below a certain threshold. This problem is closely related to contact map (CM) prediction.

While protein structure prediction remains unsolved, researchers have resorted to simplified protein models to try to gain understanding of both the process of folding and the algorithms needed to predict it [1]. Approaches have included fuzzy sets, cellular automata, L-systems and memetic algorithms (for references see [2]). One common simplification is to focus only on the residues (C-alpha or C-beta atoms) rather than all the atoms in the protein. A further simplification is to reduce the number of residue types to less than twenty by using residue sequence representations based, for instance, on physical properties such as hydrophobicity, as in the so called hydrophobic/polar (HP) models. Another simplification is to reduce the number of spatial degrees of freedom by restricting the atom or residue locations to those of a lattice [1,3,4]. Lattices of various geometries have been explored, e.g., two-dimensional triangular and square geometries or three-dimensional diamond and face centered cubic. Idealized models have been used, among other things, to study the nature of the energy landscape, the uniqueness of the native state or associated degenerate sequences, the origin of the two-state thermodynamic behavior of globular proteins (i.e.

first folding into secondary structures and later into a three dimensional shape), the existence of cooperative folding (i.e. an energy gap between the native conformation and the closest non-native one) and structure-function relations (for further references see [5], [2])

2.2. *HP Models*

In the HP model (and its variants) the 20 amino acids are reduced to two classes: non-polar or hydrophobic (H) and polar (P) or hydrophilic amino acids. An n amino acid protein is represented by a sequence $s \in \{H, P\}^+$ with $|s| = n$. The sequence s is to be mapped to a lattice, where each residue in s occupies a different lattice cell and the mapping is required to be self-avoiding. The energy potential in the HP model reflects the fact that hydrophobic amino acids have a propensity to form a hydrophobic core.

In the standard HP model, contacts that are HP and PP are assigned an energy of 0 and an HH contact is assigned an energy of -1. Whilst in the functional model protein (FMP), HP and PP receive a value of 1 and HH a value of -1. For an FMP sequence to be viable it must fold into a unique native state (unlike Dill's model [6] where the same sequence could have a variety of minimum energy states), the native structure is required to have a binding pocket, i.e. at least one hole in the conformation [5]. Moreover, there must exist an energy gap between the minimum energy conformation and the next excited state.

In this paper, rather than applying optimisation methods [7,8] to minimise the energy of the structures we concentrate on classification of models. We employ a class of machine learning techniques called Learning Classifier Systems, and in particular we use the GAssist system [9] which is based on a binary representations of rules (see section 3 for more details). [10]

3. Methodology

Three datasets were employed (Table 1). A 3D HP diamond lattice data set used for the Fold/Non-fold experiments (3DFNF), a 3D HP cubic lattice dataset used for the CN and RE experiments (3DCNRE) and a 2D square lattice dataset used for the CN and RE experiments (2DCNRE). Datasets are available on-line at `http://www.cs.nott.ac.uk/~nxk/hppdb.html`.

The experimental design was as follows: 1) For all residues, calculate CN and RE. CN is typically defined as the number of non-contiguous residues within a given radius (r=1.0 lattice unit) of each residue. RE was defined as the distance of each residue from the center of mass of the protein. 2) Create instance sets by moving a window of fixed length over the sequence-attribute

Table 1. Details of the data sets used in these experiments.

Dataset Identifier	3DFNF	3DCNRE	2DCNRE
Lattice Dimensions	3D	3D	2D
Lattice Type	Diamond	Cubic	Square
Coordination Number	4	6	4
Model Type	FMP	HP	FMP
Number of Sequences	4196352	15	4428
Number of Structures	893	15	4428
Maximum Sequence Length	23	48	20
Minimum Sequence Length	23	27	20
Total Residues	96516096	640	92988
Total Hydrophobic	48258049	316	42638
Total Polar	48258047	309	45922
Source	Taken from[11]	Taken from [12]	Taken from[13]

vectors, assigning a class to each instance: the value of that attribute for central residue in the window. 3) Split the instance sets into Training and Test sets. 4) Apply machine learning tools to predict the classes in Test Sets. 5) Extract classification accuracies for each algorithm. 6) For the non-deterministic algorithms (GAssist) iterate 10 times with different random number seeds. 7) Calculate the mean prediction accuracy. 8) Perform student t-tests on the mean prediction accuracies to determine which algorithms significantly outperformed the others (using a confidence interval of 95 and Bonferroni correction [14] for multiple pair-wise comparisons).

Windows were generated for one, two and three residues at each side of a central residue. For each attribute and for each window size, three class assignment levels (Two State, Three State and Five State) were explored. For two state assignment residues were assigned the class 1 (high) or 2 (low) according to whether their attribute value was below or above the average for that attribute value in that particular the protein. For three states the class assignments were 1 (low), 2 (intermediate) or 3 (high) for the lower, middle or upper third of the range respectively. In five state assignments the classes were 1, 2, 3, 4 or 5 for the first, second through fifth portion of the range respectively.

Composed of a rule learning algorithm and a rule inference engine, LCSs have the ability to balance multiple, potentially conflicting, constraints (e.g. formation of local structures vs global structures) and can produce high quality predictions. Moreover, LCS can produce human understandable explanations of the rules they have used to make their classifications, unlike, for example, neural network based systems. GAssist [9] is a Pittsburgh learning classifier system descended from GABIL [10]. The system applies a near-standard Genetic Algorithm (GA) that evolves individuals that represent complete problem solutions. Each individual consists of a variable length rule set. We used the rule-based knowledge representation of the

GABIL [10] system (see section 5 for an example of a generated rule set). The experimental parameters used for the GAssist experiments were the default values [9] except that for the larger datasets (2DCNRE), where 25 strata were used rather than the two strata used by default. One thousand iterations of the LCS were used. GAssist was compared against Naive Bayes, C4.5, IBk (k=3) and JRip, all of them taken from the WEKA machine learning package.

4. Experimental Results

4.1. *Results of Fold Non-Fold Classification Experiments*

Table 2 summarises the results of the Fold/Non-fold classification experiments on the 3D Diamond Lattice Structure dataset. For each algorithm the overall average and deviation of test accuracy is shown. GAssist was the best method on this dataset, outperforming significantly three of the four other tested methods.

Table 2. Averaged Classification Accuracies (%) for 3D HP Fold/Non-Fold Experiments. A • means that GAssist significantly outperformed the Algorithm to the left

Algorithm	**Total**
Naive Bayes	74.8±3.1 •
GAssist	88.3±1.7
IBk	81.8±2.7 •
JRip	86.9±3.1 •
C4.5	87.9±2.5

4.2. *Results of CN and RE Classification Experiments*

Table 3 summarises the results of the classification experiments for CN and RE for the 3DHPCNRE the 2DHPCNRE datasets. For each algorithm the overall average and deviation of test accuracy is shown. GAssist performed at a similar or better level than the other tested machine learning methods. It significantly outperformed other methods 24 times and it was outperformed in just six of the tested datasets.

5. Discussion

The performance of the GAssist LCS was equal or better than the other tested methods, especially on the fold/non fold dataset. It was outperformed significantly very few times. From a general point of view we can say that CN is easier to classify than RE, and that the 2D lattice data are also more difficult to classify than the 3D data. On the 3D lattice, CN can be classified around 80%, 67% and 52% for two, three and five states, and

Table 3. Averaged Classification Accuracies (%) for 2D and 3D HP CN and RE Experiments. A • means that GAssist significantly outperformed the Algorithm to the left, a ○ means that the Algorithm on the left outperformed GAssist

Exper.	States	Alg.\Win. Size	3D Data			2D Data		
			3	5	7	3	5	7
CN	2	Naive Bayes	79.7±5.8	79.9±5.2	80.2±4.5	61.2±0.3	63.9±0.4	62.6±0.4•
		GAssist	79.9±6.0	80.2±5.4	79.6±4.7	61.2±0.3	64.1±0.4	64.9±0.3
		IBk	80.1±6.0	79.0±5.4	78.0±5.1	61.2±0.3	64.1±0.4	65.1±0.4
		JRip	80.1±6.0	80.1±5.8	79.9±5.0	61.2±0.3	63.8±0.4	64.7±0.4
		C4.5	80.2±6.0	79.9±5.7	79.8±4.6	61.2±0.3	64.0±0.4	65.1±0.4
	3	Naive Bayes	67.1±5.6	67.2±4.6	67.3±4.9	70.9±0.2	70.9±0.2	68.5±0.2•
		GAssist	67.1±6.0•	67.7±4.6	67.3±5.0	70.8±0.4	71.0±0.4	71.0±0.4
		IBk	66.1±6.3	66.7±5.3	64.9±5.7	70.9±0.2	71.1±0.3	71.0±0.2
		JRip	60.7±5.2	64.8±5.2	64.5±4.9	70.9±0.2	70.5±0.3•	70.5±0.3•
		C4.5	67.5±5.6	67.7±4.7	65.8±5.1	70.9±0.2	71.1±0.3	71.0±0.2
	5	Naive Bayes	51.6±4.4	52.2±4.4	51.8±5.8	58.1±0.2	56.8±0.2•	56.4±0.3•
		GAssist	51.4±4.5	51.3±4.4	52.9±5.3	58.1±0.2	58.7±0.3	58.8±0.3
		IBk	51.3±4.6	49.6±4.6	48.8±5.8	58.1±0.2	58.7±0.3	58.9±0.3
		JRip	45.5±3.7•	46.9±4.3•	49.0±6.0	58.1±0.2	57.6±0.3•	57.6±0.3•
		C4.5	51.7±4.5	50.7±4.2	52.3±5.1	58.1±0.2	58.6±0.3	58.8±0.2
RE	2	Naive Bayes	77.8±5.5	78.6±4.4	79.7±4.4	56.9±0.5	60.0±0.4•	58.7±0.5•
		GAssist	77.9±5.5	78.1±4.8	78.2±4.2	56.9±0.4	60.4±0.5	61.4±0.5
		IBk	78.2±5.3	76.7±5.1	76.2±4.3	56.9±0.4	60.5±0.5	61.9±0.6○
		JRip	78.1±5.3	77.8±4.8	78.3±4.6	56.9±0.4	60.2±0.5	61.1±0.5
		C4.5	77.8±5.4	77.6±4.2	77.9±4.1	56.9±0.4	60.5±0.4	61.7±0.6
	3	Naive Bayes	63.0±5.7	63.3±5.2	62.5±5.5	43.3±0.3	45.4±0.3•	44.2±0.3•
		GAssist	62.0±5.5	61.7±5.5	62.1±4.7	43.3±0.3	46.5±0.3	47.2±0.6
		IBk	61.1±4.9	61.0±5.0	61.8±5.2	43.3±0.3	46.5±0.3	47.8±0.5○
		JRip	59.7±3.0	59.0±3.3•	61.4±3.9	43.3±0.3	45.6±0.3•	46.5±0.4•
		C4.5	61.6±5.2	61.7±5.3	64.1±4.1	43.3±0.3	46.5±0.3	47.8±0.4○
	5	Naive Bayes	37.3±6.6	38.6±6.1	37.6±6.1	27.8±0.2	27.8±0.3•	28.1±0.4•
		GAssist	37.6±5.9	36.2±5.9	39.2±5.3	27.8±0.3	30.8±0.5	32.0±0.6
		IBk	37.0±5.7	36.7±5.9	38.5±6.1	27.8±0.3	31.1±0.5	33.1±0.4○
		JRip	34.5±2.9	33.6±3.8•	36.2±5.6	25.3±0.0•	28.4±0.3•	28.0±0.3•
		C4.5	38.2±6.8	36.8±6.3	38.9±4.9	27.8±0.3	31.2±0.5○	33.0±0.4○

RE can be classified around 78%, 62% and 38%. For the 2D lattice data, CN can be classified around 65%, 71% and 59%, and RE can be classified around 62%, 47% and 33% for two, three and five states. The fold/non fold domain can be classified with an 88% accuracy.

Beside its performance, GAssist has another advantage, which is the generation of compact and interpretable solutions. GAssist generated on average rule sets consisting of 52.8, 9.6 and 3.5 rules for the 3DFNF, 2DCNRE and 3DCNRE datasets, respectively. As an example, we show a rule set from an individual generating 87.3% accuracy for two state prediction with a window size of seven (three residues either side of the residue being predicted) for the CN domain using 3D lattice. An X symbol is used to represent positions at the end of the chains, that is beyond the central residue being studied, H means high CN, L means low CN. The rule set only had three rules, and at most three of the seven input attributes were expressed. The rules are interpreted in order, therefore all examples not matched by the first or second rules are assigned class L.

(1) If $Position_{i-1} \notin \{p\}$, $Position_i \in \{h\}$, $Position_{i+1} \notin \{h\}$ then class is H
(2) If $Position_{i-2} \notin \{X\}$, $Position_i \in \{h\}$ then class is H
(3) Default class is L

Moving from highly abstract (2 class) to more informative predictions (5 class) more input data (larger windows) are required in order to facilitate learning. The 3D structures on the cubic lattice have less than 50 residues, as a result the training data has an unnaturally high proportion of exposed/low-CN residues (including hydrophobic residues which are more usually found buried). Analysis (not shown) of the distribution of residues by class showed that for the 2D square lattice structures this bias in the input data distributions is less pronounced. We have extended these studies to real proteins (papers submitted) and HP representations of real proteins [2]. In the future we will investigate computation and prediction of other structural properties such secondary structures and disulfide bridges.

6. Conclusions

These studies have shown that: a) it was possible to discriminate at around 80% accuracy, from sequence alone, which proteins will and will not fold b) It was also possible to predict which residues have above or below average CN and RE c) it is possible to predict the detailed CN and RE states of residues and d) The GAssist LCS performs at a level comparable to other ML algorithms on these problems. Of the WEKA algorithms studied, those based on orthogonal representations perform slightly better than those which are not. Minimalist lattice structure models focus on the essential details of protein structure prediction. Moving from highly abstract predictions (above/below mean for a given attribute) to more detailed structural predictions (eg. five state CN), accuracy can be increased by incorporating more local residue pattern information in the inputs (increased window size). However, in real proteins, only some contacts (secondary structure contacts) arise from local residue sequence patterns that may be recognizable in short fragments/windows. Other contacts arise from long-range global features of proteins and these may not be evident in short local sequence patterns. Future studies will extend these investigations with classifications based on other structural attributes and studies of real protein datasets.

7. Acknowledgments

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

GR/62052/01 and the Biotechnology and Biological Sciences Research Council (BBSRC) under grant BB/C511764/1.

References

1. Yue, K., Fiebig, K.M., Thomas, P.D., Sun, C.H., Shakhnovich, E.I., Dill, K.A.: A test of lattice protein folding algorithms. Proc. Natl. Acad. Sci. USA **92** (1995) 325–329
2. Stout, M., Bacardit, J., Hirst, J.D., Krasnogor, N., Blazewicz, J.: From hp lattice models to real proteins: coordination number prediction using learning classifier systems. In: 4th European Workshop on Evolutionary Computation and Machine Learning in Bioinformatics 2006 (to appear). (2006)
3. Blazewicz, J., Dill, K., Lukasiak, P., Milostan, M.: A tabu search strategy for finding low energy structures of proteins in hp-model. (Computational Methods in Science and Technology)
4. Blazewicz, J., Lukasiak, P., Milostan, M.: Application of tabu search strategy for finding low energy structure of protein. Artificial Intelligence in Medicine **35** (2005) 135–145
5. Hirst, J.D.: The evolutionary landscape of functional model proteins. Protein Engineering **12** (1999) 721–726
6. Dill, K., Bromberg, S., Yue, K., Fiebig, K., Yee, D., Thomas, P., Chan, H.: Principles of protein folding: A perspective from simple exact models. Prot. Sci. **4** (1995) 561
7. Krasnogor, N., Blackburne, B., Burke, E., Hirst, J.: Multimeme algorithms for protein structure prediction. In: Proceedings of the Parallel Problem Solving from Nature VII. Lecture Notes in Computer Science. Volume 2439. (2002) 769–778
8. Hart, W., Istrail, S.: Fast protein folding in the hydrophobic-hydrophilic model within three-eighths of optimal. Journal of Computational Biology **3** (1996) 53–96
9. Bacardit, J.: Pittsburgh Genetics-Based Machine Learning in the Data Mining era: Representations, generalization, and run-time. PhD thesis, Ramon Llull University, Barcelona, Catalonia, Spain (2004)
10. DeJong, K., Spears, W., Gordon, D.: Using genetic algorithms for concept learning. Machine Learning **13** (1993) 161–188
11. Blackburne, B.P., Hirst, J.D.: Three dimensional functional model proteins: Structure, function and evolution. Journal of Chemical Physics **119** (2003) 3453–3460
12. Hart, W.: (www.cs.sandia.gov/tech_reports/compbio/tortilla-hp-benchmarks.html) Tortilla HP Benchmarks.
13. Blackburne, B.P., Hirst, J.D.: Evolution of functional model proteins. Journal of Chemical Physics **115** (2001) 1935–1942
14. Miller, R.G.: Simultaneous Statistical Inference. Springer Verlag, New York (1981) Heidelberger, Berlin.