The University of Nottingham

asap automated scheduling optimisation & planning research

The Following Talks are Based On:

- **Bacardit, J., Stout, M., Hirst, J.D., and Krasnogor, N***, *Coordination number prediction using Learning Classifier Systems: Performance and interpretability*, accepted for GECCO 2006 Seatle, USA, 8th-12th July 2006.

- **Stout, M., Bacardit, J., Hirst, J.D., Blazewicz, J, and Krasnogor, N***, *Prediction of Residue Exposure and Contact Number for Simplified HP Lattice Model Proteins Using Learning Classifier Systems*, accepted for the 7th International FLINS Conference on Applied Artificial Intelligence, Genova, Italy, 29-31 August 2006.

- **Stout, M., Bacardit, J., Hirst, J.D., Blazewicz, J. and Krasnogor, N.*** *From HP Lattice Models to Real Proteins: coordination number prediction using Learning Classifier Systems*, accepted for the 4th European Workshop on Evolutionary Computation and Machine Learning in Bioinformatics, Budapest, Hungary, 10-12 April 2006. (download)

- Corresponding author: Natalio.Krasnogor@Nottingham.ac.uk
- www.cs.nott.ac.uk/~nxk

The University of Nottingham

asap automated scheduling optimisation & planning research

Prediction of Residue Exposure and Contact Number for Simplified HP Lattice Model Proteins Using Learning Classifier Systems

M Stout, J. Bacardit, J.D. Hirst, and N. Krasnogor*

7th International FLINS Conference
on Applied Artificial Intelligence
Genova, Italy

* Corresponding author

CIBB 29th August 2006

Summary

- Simplified hydrophobic/polar (HP) lattice model proteins
- Compare machine learning (ML) algorithms
- CN and RE prediction

- Explore Learning Classifier Systems (LCS)
- LCS apply Evolutionary Computation to Machine Learning problems

Introduction

- Predicting structural properties of proteins from sequence
  - fundamental
  - important
- Rather than predicting the complete tertiary (native) structure we divide-and-conquer and try to predict some attributes/properties of the native state
  - residue exposure (RE)
  - coordination number (CN)
- Simplified protein models (e.g. HP models)
- Represent sequence using two residue types
  - hydrophobic and polar
- Restrict the residue locations to those of a lattice

HP Models

- 20 amino acids reduced to two classes
  - Non-polar (H) -- hydrophobic
  - Polar (P) -- hydrophilic

- n amino acid protein represented by sequences
- Sequence s is mapped to lattice
- Each residue in s occupies a different lattice cell

- Mapping is self-avoiding

Lattices Geometries

- 2D: triangular, square
- 3D: diamond, face centered cubic

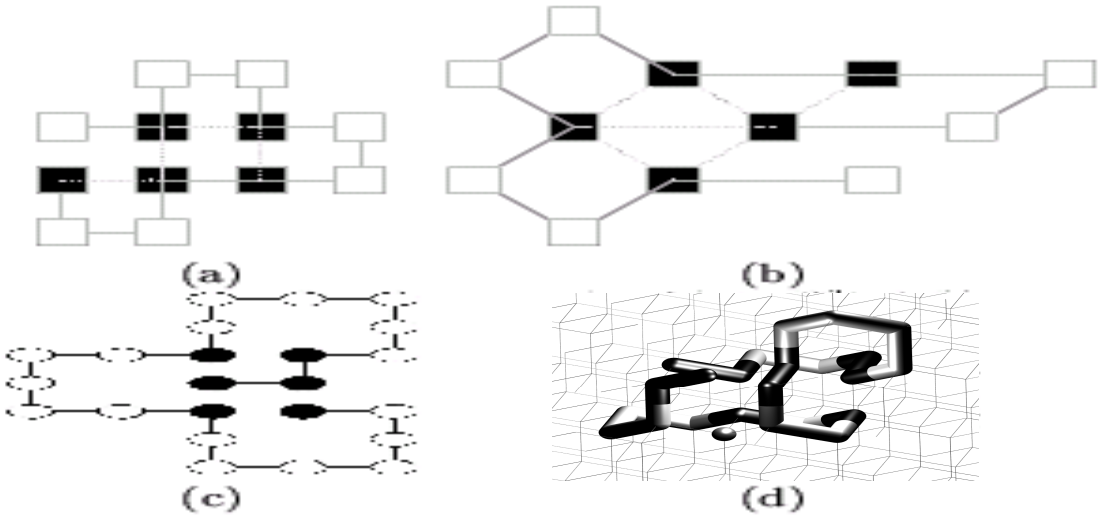


FIG. 1: *HP protein embedded in the square lattice (a) and triangular lattice(b). Functional Model protein embedded in the square lattice(c) and diamond (3D) lattice(d). In (c) and (d) native structures are not maximally compact as they must have a "binding pocket".*

Energy Potentials

- HP model ==> hydrophobic amino acids have propensity to form hydrophobic core
- Standard HP model
  - HP and PP assigned 0
  - HH contacts assigned -1
- Functional model protein (FMP)
  - HP and PP receive a value of 1
  - HH value of -1
- FMP must fold into unique native state
- Dill's HP model sequences have variety of minimum energy states
- Native structure is required to have a binding pocket
  - At least one hole in conformation
- Energy gap between minimum energy conformation and next excited state

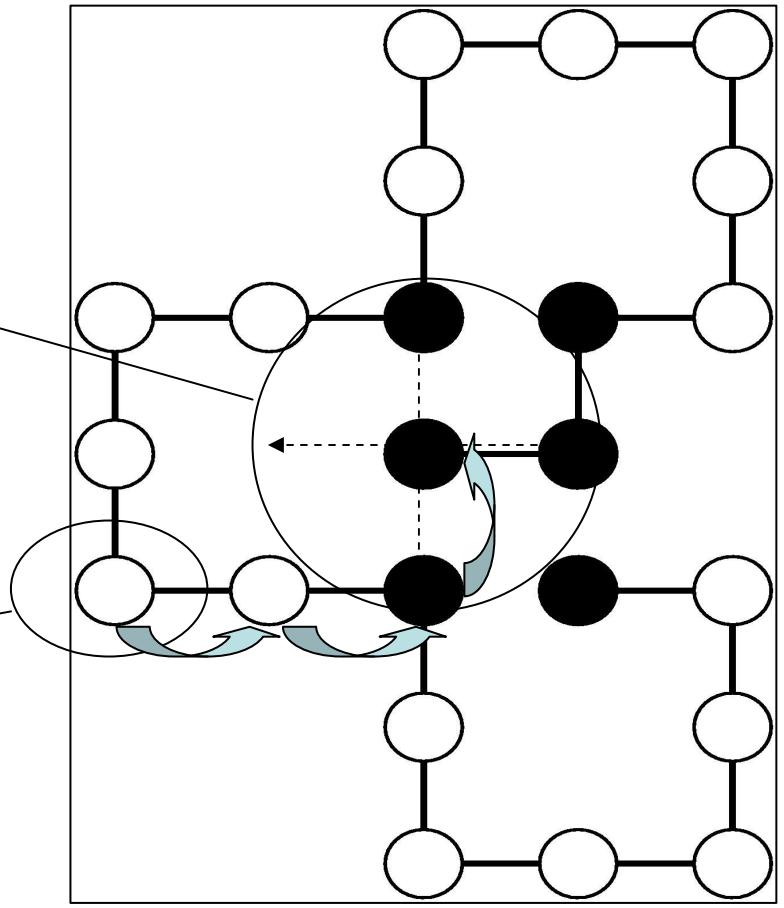
Definitions

- CN: number of non-contiguous residues within radius (r=1.0 lattice unit)

CN=2

- RE: distance of residue from center of mass of protein

RE=3



Questions

1) Is it possible to predict, from sequence alone, which proteins will and will not fold? (remember discussion on degenerate sequences). This is the Fold/Non Fold prediction problem.

2) Is it possible to predict, from sequence alone, which residues have above or below average CN and RE? This is the *relative* CN/RE prediction problem.

3) Is it possible to predict, from sequence alone, the precise CN/RE states? This is the CN/RE prediction problem.

4) Are LCS a suitable tool for these tasks?

Methodology

- Three datasets were employed

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[8]	Taken from [9]	Taken from[10]

Experimental Design (1/2)

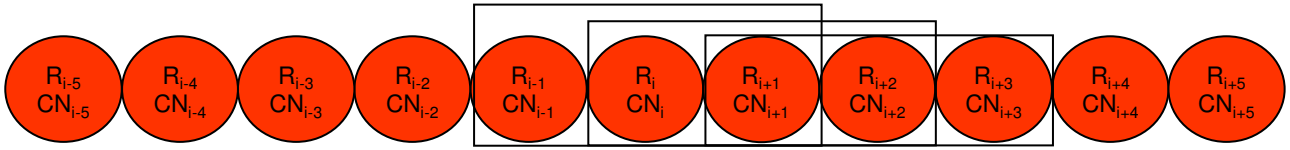
- 1) Calculate CN and RE from the existin data sets
- 2) Move fixed length window over sequence → attribute vectors
  - Assigning class to each instance:
    - CN/RE value for central residue of window
- 3) Split instance sets into Training and Test sets
- 4) Apply machine learning tools to learn to predict classes in training set
- 5) Apply learning knowledge in Test Sets
- 6) Extract classification accuracies for each algorithm

Experimental Design (2/2)

- 6) For non-deterministic algorithms (e.g. GAssist) iterate 10 times with different random number seeds
- 7) Calculate mean prediction accuracy
- 8) Perform student t-tests on mean prediction accuracies
  - determine which algorithms significantly outperformed the others (using a confidence interval of 95 and Bonferroni correction for multiple pair-wise comparisons)

Instance Generation

- Windows generated 1,2,3 residues each side of central residue



$R_{i-1},R_i,R_{i+1} \rightarrow RE_i/CN_i$
$R_i,R_{i+1},R_{i+2} \rightarrow RE_i/CN_{i+1}$
$R_{i+1},R_{i+2},R_{i+3} \rightarrow RE_i/CN_{i+2}$

- 3 class assignment schemes (Q2)
  - Two State: Class 1 (high) or 2 (low): below or above average for that protein
- Three State (Q3)
  - 1 (low), 2 (intermediate) or 3 (high) for the lower, middle or upper third of the range respectively
- Five State (Q5)
  - 1, 2, 3, 4 or 5 for the first, second through fifth portion of the range respectively

Learning Classifier Systems

- Composed of rule learning algorithm
- Rule inference engine
- Balance multiple, potentially conflicting, constraints
  - e.g. formation of local structures vs global structures
- Produce high quality predictions
- Produce human understandable explanations of rules evolved

GAssist LCS

- Pittsburgh learning classifier system
- Standard Genetic Algorithm (GA)
- Individuals represent complete problem solutions
- Individual consists of variable length rule set
- Rule-based knowledge representation of GABIL
- Default parameters except for larger datasets (2DCNRE), 25 strata

ML Algorithms

- GAssist LCS
- WEKA implementations
  - Naive Bayes
  - C4.5
  - IBk (k=3)
  - JRip

GAssist Results Overview

- Folding or non-folding prediction:
  - Functional HP model proteins
  - 3D diamond lattice
  - ~88.3% accuracy
  - Outperformed significantly three out of four other methods

- Contact Number/Residue Exposure prediction:
  - HP model protein instances
  - 2D square and 3D cubic lattices
  - ~27.8% - ~80.9%
  - Level comparable to other ML technologies
  - Outperforming significantly them in 24 out of 180 cases
  - Outperformed just six times

Detailed Results
Fold Non-Fold Prediction

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

● Significantly outperformed

- Overall average and deviation of test accuracy
- GAssist best method on this dataset
- Outperformed significantly three of four other tested methods

Detailed Results: CN and RE 3DHPCNRE, 2DHPCNRE

- GAssist performed at similar or better level than others ML Algorithms
- Significantly outperformed other methods 24 times
- Outperformed in six tests

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○

Discussion

- GAssist equal or better than other methods
  - Especially on fold/non fold dataset
- Out peformed significantly very few times
- CN easier to classify than RE
- 2D lattice data more difficult than 3D

The University of Nottingham

asap automated scheduling optimisation & planning research

In Particular

- 2D lattice
  - CN ~65%, 71% and 59% for two, three and five states
  - RE ~62%, 47% and 33% for two, three and five states

- 3D lattice
  - CN ~80%, 67% and 52% for two, three and five states
  - RE ~78%, 62% and 38%

Rule Sets Details

- GAssist rule sets consist ~52.8, 9.6 and 3.5 rules: 3DFNF, 2DCNRE and 3DCNRE
- A classifier with 3 rules for CN in 3D lattice (W=7):
  - Remember that "X" = positions at the end of the chains, "H" = high CN, "L" = low CN

(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

Achieves 87.3% accuracy for two state prediction Rule set only had three rules

- At most three of seven input attributes were expressed, that is, it ignores useless features → concise theory

Rule Sets Details

- Rules are interpreted in order
- All examples not matched by first or second rules are assigned default class
- Moving from highly abstract (s2) to more informative predictions (s5) more input data (larger windows) required to facilitate learning

- 3D structures on cubic lattice have less than 50 residues

  ==> training data has high proportion of exposed/low-CN residues

  including hydrophobic residues -- more usually found buried

- Distribution of residues by class showed 2D square lattice structures bias in input data distributions is less pronounced

Conclusions (1/2)

- a) ~80% accuracy for Fold/Non-fold
- b) Can predict residues having above or below average CN and RE
-
- c) Can predicted detailed CN and RE
- d) GAssist LCS performs at level comparable to other ML algorithms
- Algorithms based on orthogonal representations perform slightly better than those which are not
- Lattice structure models focus on essential details

Conclusions (2/2)

- When moving from highly abstract predictions, e.g. above/below mean for a given attribute towards more detailed structural predictions, e.g. five state CN or RE → Accuracy is increased by increased window size

- This is related to the fact that:
  - In real proteins only some contacts arise from local residue sequence patterns (secondary structure contacts)
  - Others arise from long-range global features of proteins
  - This may not be evident in short local sequence patterns

Next

- We extend these studies from HP model proteins, to HP representations of real proteins, to real proteins