From HP Lattice Models to Real Proteins: Coordination Number Prediction Using Learning Classifier Systems

EvoBio 2006 Paper

Michael Stout, Jaume Bacardit, Jonathan Hirst, Jacek Blacewicz, Natalio Krasnogor*

* Contact author.

Outline

- Introduction
  - Proteins
  - Problem Definition
  - Technical Approach
- Experiments
- Results
- Discussion
- Related Work
- Conclusions
- Future Work

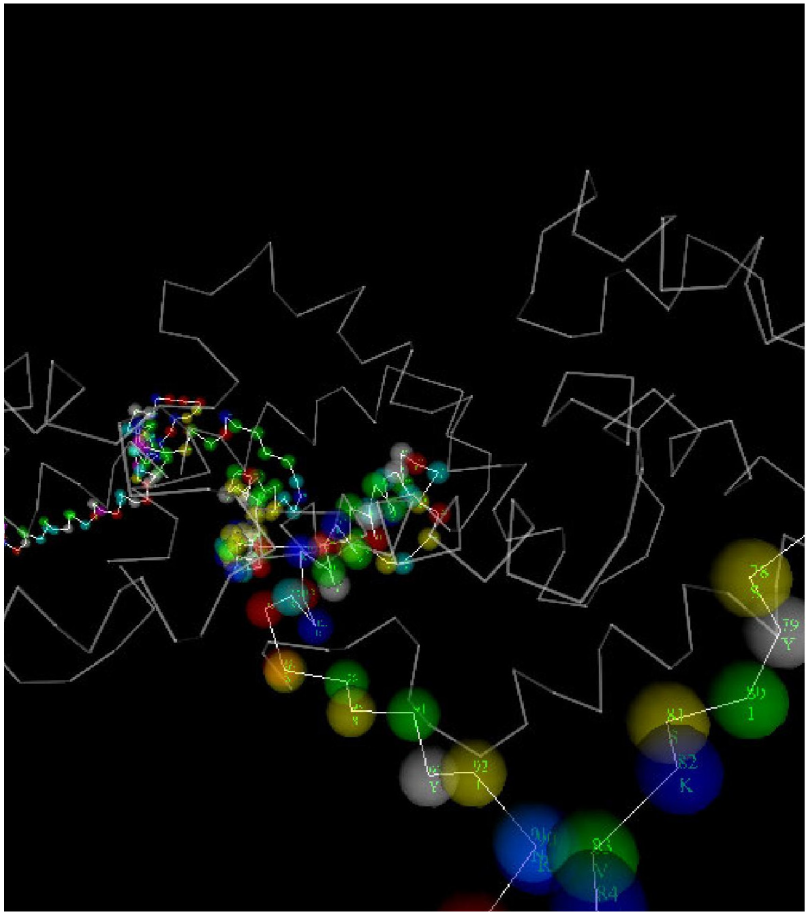
Objective

- Investigate protein Contact Number prediction
- Compare a range of
  - Representations from abstract to intermediate to real proteins
  - Machine Learning algorithms
  - Experimental Parameters

Protein Structure Prediction

- Prediction of protein 3D structures
  - Fundamental
  - Difficult
  - Unsolved

- Popular approaches
  - 1) Predict specific attributes
  - 2) Simplify representations
  - 3) Combine these to make overall predictions



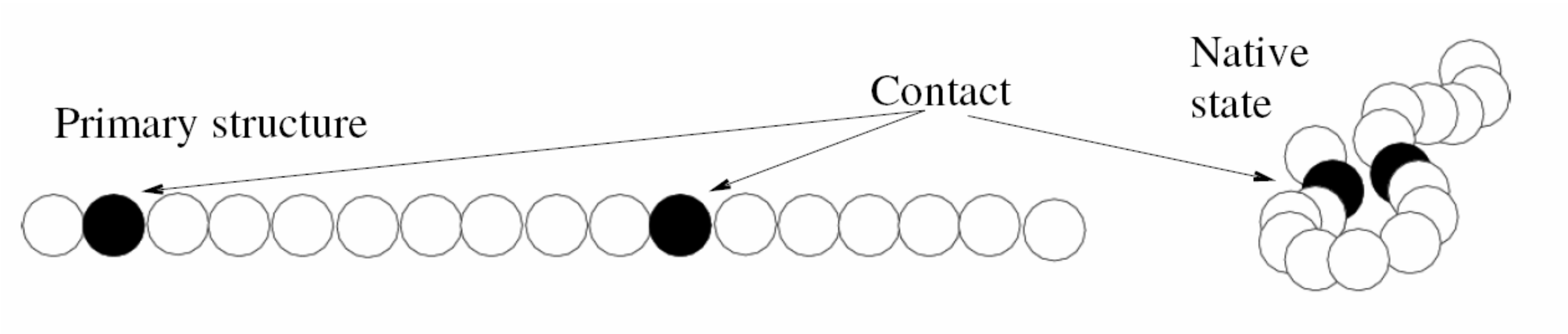
Staphylococcus aureus
virulence regulatory protein

1) Specific Attributes

- Secondary structure
- Solvent accessibility
- Disulfide (cysteine) bridges

- Coordination number (CN)
  - Functional sites in proteins are pockets of residues
  - Active sites contain buried residues ➔ high CN
  - CN studies relevant to understanding protein function

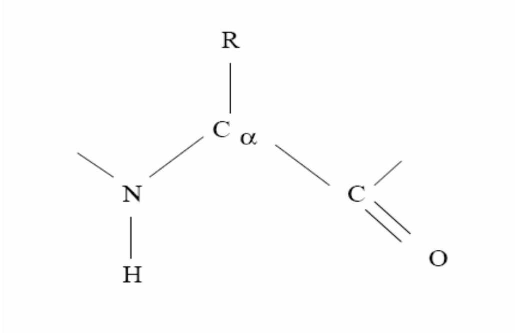
Residues Contacts



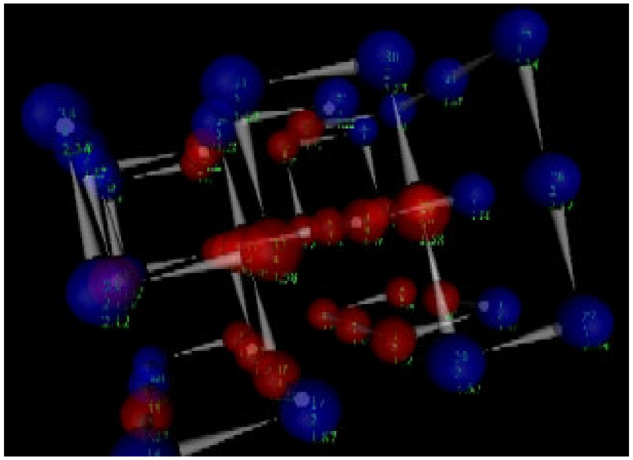
- For residue *r*, CN is the number of residues in contact with it
- Threshold distance
- Related to contact map (CM) prediction

2) Simplified Models

- Simplifications
  - Only Residues (C_α or C_β atoms) cf. all atoms



  - Fewer residue types
    - Focus on physical/chemical properties
      - hydrophobic-polar (HP) models



- Reduce spatial degrees of freedom
  - Restrict locations to lattice

    2D triangular, square etc

    3D diamond, face cantered cubic etc

Our Approach

- Use a Real Valued CN definition
- Frame prediction as a Classification Problem
- Compare several ML tool
  - Learning Classifier Systems (LCS)
  - Decision Trees
  - Naïve Bayes

- Investigate 3 levels of “simplification“

  1. Model proteins 2 letter HP alphabet on 3D cubic lattice
  2. Real proteins 2 letter HP alphabet
  3. Real proteins 20 letter AA type alphabet

  - Explore effects of experimental parameters
    - Window size
    - Number of classes

Real Valued CN Definition

- C_β atoms, distance cut off d_c10Ǻ
- Smooth boundary using sigmoid function

CN of residue *ith* protein chain *p* is:

$$O_i^p = \sum_{j:|j-i|>2} \frac{1}{1 + exp(w(r_{ij} - d_c))}$$

- where r_{ij} is distance between C_β atoms of i^{th} and j^{th} residues
- *w* determines sharpness of boundary of sphere (we use w=3)

- Minimum chain separation of 2 residues
- Kinjo *et al.* 2005

Real-Valued CN ➔ Class

- Frame problem as a classification problem

- Real-valued CN ➔ Discrete Classes (similar to “bining”)

  - Group instances with similar CN
  - Choose class boundaries ➔ uniform number of instances
  - Defining these globally for all 20 residue types

Learning Classifier Systems (LCS)

- Rule-based ML systems
- Use EC as search mechanism

- GAssist (Bacardit, 2004)
  - Pittsburgh Genetic Based Machine Learning system
  - Descendant of GABIL
  - Generates accurate, compact, highly interpretable solutions
- Applies near-standard GA
- Evolves individuals representing complete problem solutions
- Individuals are ordered, variable-length rule sets

GAssist LCS

- Use special fitness function
  - Minimum Description Length (MDL)

    Balance complexity and accuracy of rule set

- Uses windowing scheme
  - Incremental Learning with Alternating Strata (ILAS)

    Reduces run-time, especially with very large dataset

- Attribute representations
  - Nominal: GABIL rule-based knowledge representation
  - Real: Adaptive discretization intervals (ADI)

Learning Classifier Systems

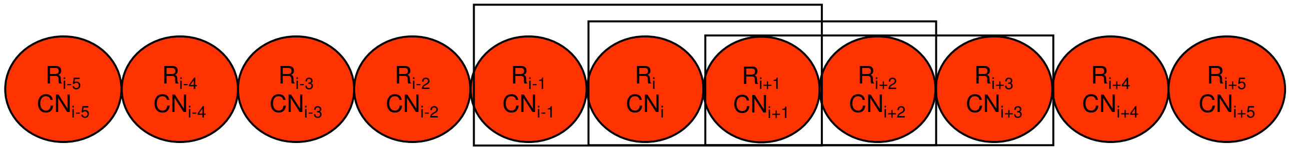
- Match process
  - Individuals are interpreted as a decision list [Rivest, 87]: an ordered rule set
  - At the end of the rule set there is an static and explicit default rule
  - The class of the default rule will not be used by the other classes, reducing the search space

1	2	3	4	5	6	7	8

Instance 1 matches rules 2, 3 and 7 → Rule 2 will be used
Instance 2 matches rules 1 and 8 → Rule 1 will be used
Instance 3 matches rule 8 → Rule 8 will be used
Instance 4 matches no rules → Instance 4 will be classified by the default rule

Experimentation design

- We have to transform the data into a regular structure so that it can be processed by standard machine learning techniques
- Each residue is characterized by several features. We use one (i.e., the AA type) or more of them as input information and one of them as target (CN)



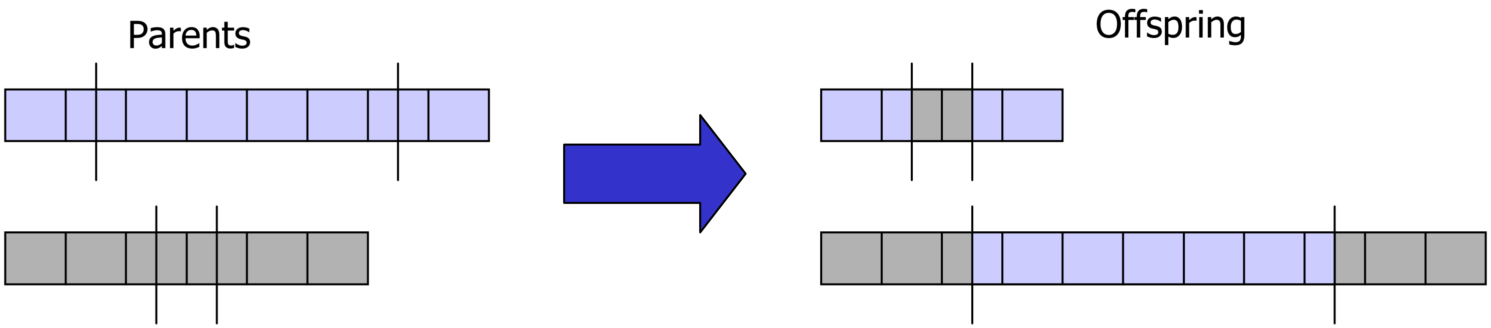
$R_{i-1},R_i,R_{i+1} \rightarrow CN_i$

$R_i,R_{i+1},R_{i+2} \rightarrow CN_{i+1}$

$R_{i+1},R_{i+2},R_{i+3} \rightarrow CN_{i+2}$

Learning Classifier Systems

- Recombination operators
  - Crossover operator

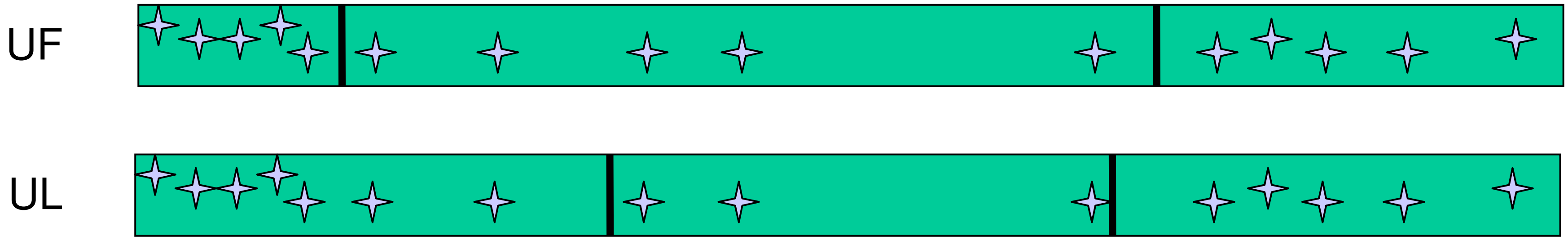


  - Mutation operator: classic GA mutation of bit inversion

Classification approach

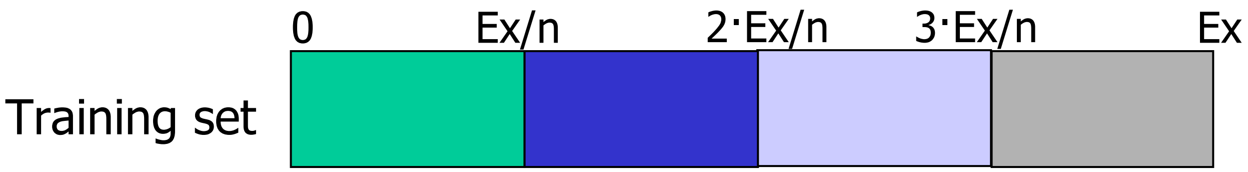
Unsupervised discretization methods:

- Uniform Frequency (UF)
- Uniform Length (UL)

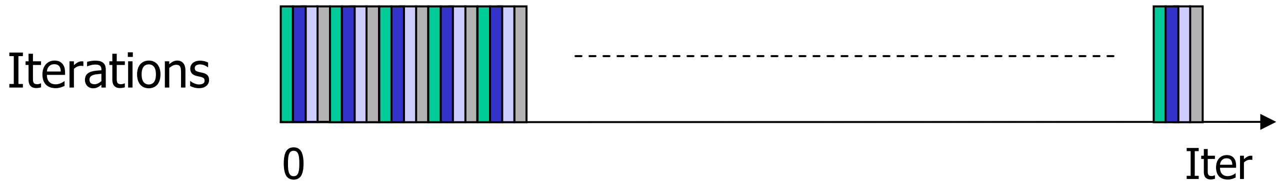


Learning Classifier Systems

- Costly evaluation process if dataset is big
- Computational cost is alleviated by using a windowing mechanism called ILAS



- This mechanism also introduces some generalization pressure



Comparison of ML Algorithms

- Compare 3 ML Algorithms:
  - GAssist: LCS
  - C4.5: rule induction system
  - Naive Bayes: Bayesian learning algorithm

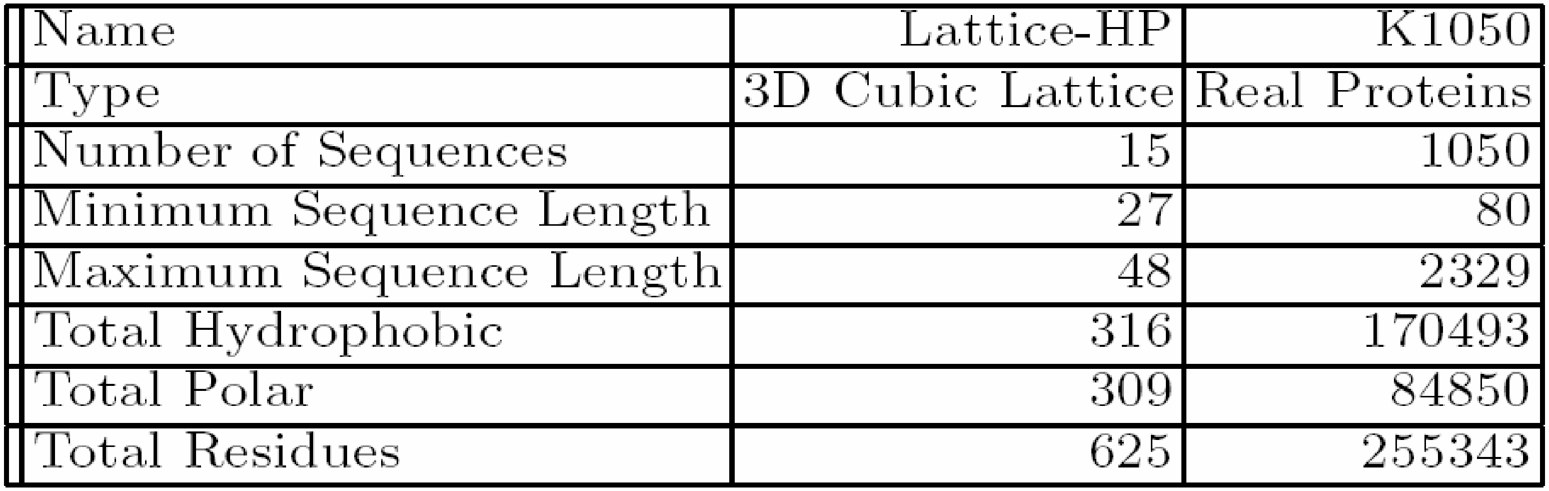
- Performance Evaluation
  - Student t-tests of mean prediction accuracies
  - Confidence interval 95%

Datasets

- Lattice-HP
  - Bill Hart's *Tortilla* Benchmark Collection
  - 15 structures on simple cubic lattice (CN=6)
- Real Proteins
  - Selected from PDB
  - Same dataset and training/test partitions as Kinjo et al 2005
  - Total of 1050 protein chains

Experimental Framework

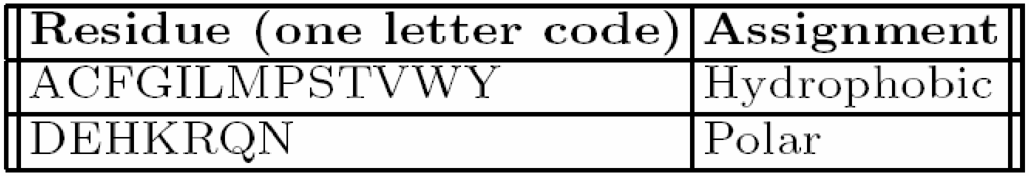
- Two datasets in this study
  - 3D HP lattice model dataset
  - Data set of real proteins



Name	Lattice-HP	K1050
Type	3D Cubic Lattice	Real Proteins
Number of Sequences	15	1050
Minimum Sequence Length	27	80
Maximum Sequence Length	48	2329
Total Hydrophobic	316	170493
Total Polar	309	84850
Total Residues	625	255343

HP Abstraction of Real Proteins Residues

- Assigning each real residue and H/P value
- Used assignments of Broome and Hecht (2000)

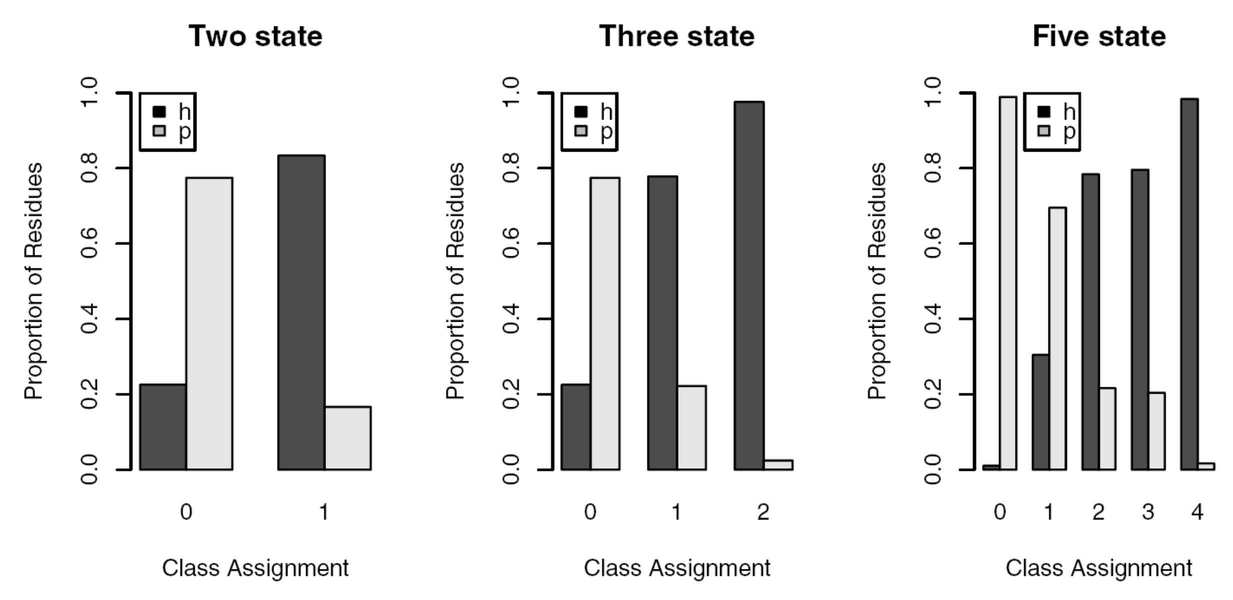


Residue (one letter code)	Assignment
ACFGILMPSTVWY	Hydrophobic
DEHKRQN	Polar

  - "Octanol : Water Partitioning" & "Binary Genetic Code" agreement

- Residue distributions ➔ baseline for prediction algorithms

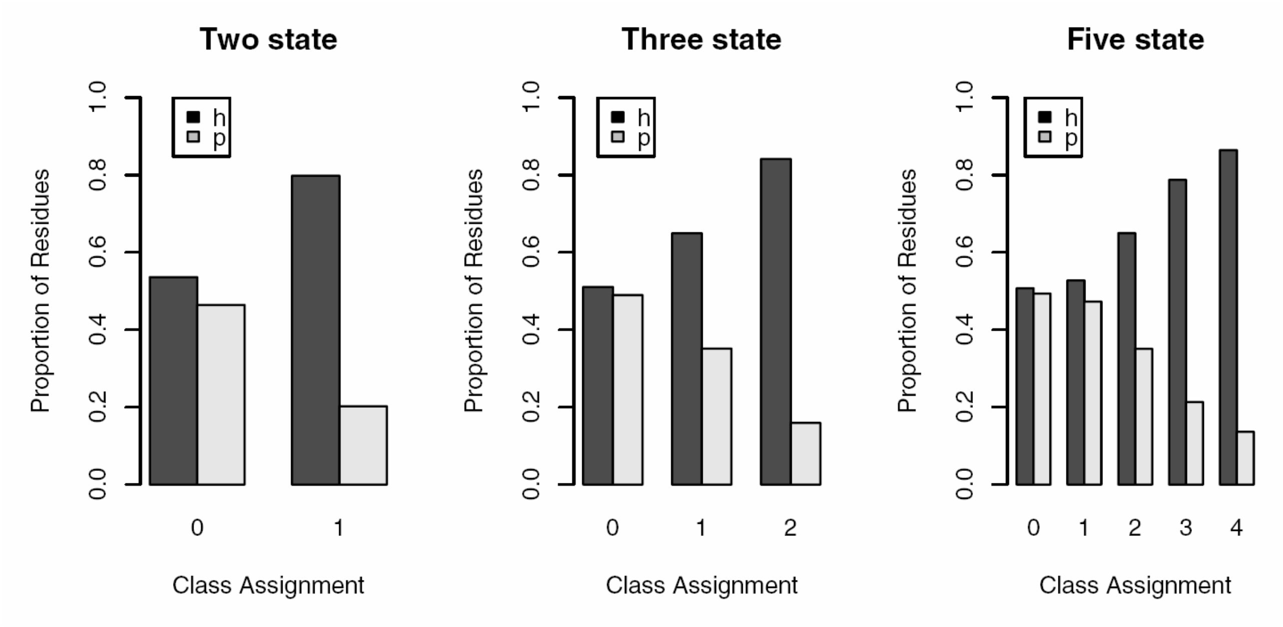
Residue Distributions: Lattice HP



- Lattice-HP
  - High CN ➔ more H residues: core of buried hydrophobic residues
  - Low CN ➔ more P residues

HP models optimized on basis of hydrophobicity …

Residue Distributions: Real-HP



- Real-HP
  - High CN ➔ more H: buried hydrophobic core
  - Low CN ➔ ~Equal distribution of H and P in (exposed) classes

  2H:1P ratio in HP assignments (above)

Creating Instances

- Window sizes
  - 1,2 and 3 residues each side of central residue (3 - 7 residue fragment)
- CN of central residue

  ➔ Class of instance
  - Lattice Models:

    Non-consecutive residues on lattice
  - Real Proteins

    Distance cut off 10Ǻ
- Instance Set divided randomly

  ➔ 10 pairs of training and test sets
  - Training == 950 proteins
  - Testing == 100
  - similar to ten-fold cross-validation

XXXRTDC

XXRTDCY

XRTDCYG

RTDCYGN

TDCYGNV

DCYGNVN

CYGNVNR

YGNVNRI

GNVNRID

Estimation of Information Loss (1/2)

- Two measures:

$$\text{redundancy} = 1 - \frac{\#\text{unique instances}}{\#\text{total instances}}$$

$$\text{inconsistency} = \frac{\left(\frac{\#\text{unique instances}}{\#\text{unique antecedents}}\right) - 1}{\#states - 1}$$

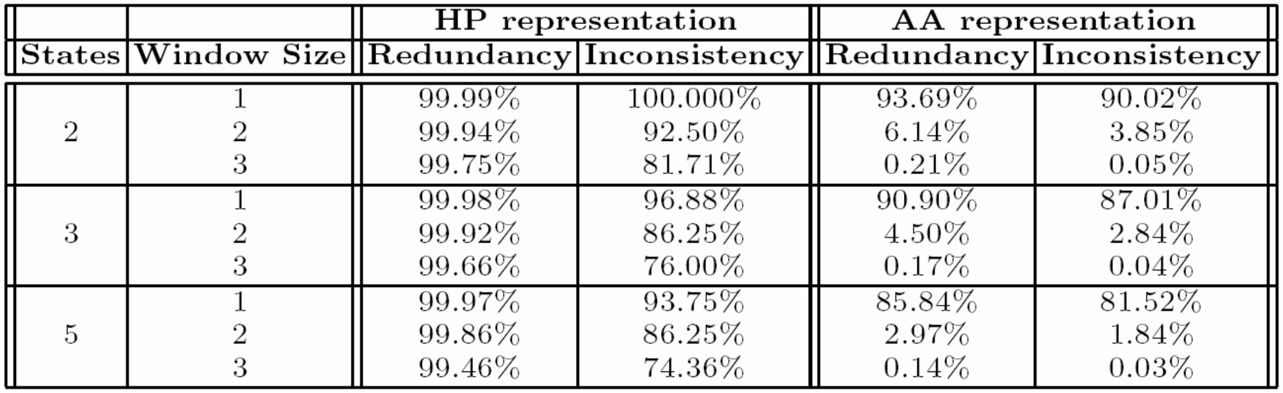
- Reducing alphabet and window size

  ==> many copies of same instances

  ==> inconsistent instances

  (Instances with equal input attributes (antecedent) but different class)

Estimation of Information Loss (2/2)



		HP representation		AA representation	
States	Window Size	Redundancy	Inconsistency	Redundancy	Inconsistency
2	1	99.99%	100.000%	93.69%	90.02%
	2	99.94%	92.50%	6.14%	3.85%
	3	99.75%	81.71%	0.21%	0.05%
3	1	99.98%	96.88%	90.90%	87.01%
	2	99.92%	86.25%	4.50%	2.84%
	3	99.66%	76.00%	0.17%	0.04%
5	1	99.97%	93.75%	85.84%	81.52%
	2	99.86%	86.25%	2.97%	1.84%
	3	99.46%	74.36%	0.14%	0.03%

(Normalized for different number of target states)

- Extreme case: s2, w1, Real-HP:
  - Any possible antecedent appears associated to both classes
  - Proportions of two classes for each antecedent are different
  - System can still learn
- Real-HP dataset is highly redundant
- w2/3 presents low redundancy and inconsistency rate ??????

Results Overview: Lattice-HP

- For all algorithms
  - Increased number of states ➔ decreased accuracy
    s2: ~80% ➔ s5: ~51%

- For each state
  - Increased window size ➔ increased accuracy (~0.1%-~0.2%)

- Best predictions:

– s2: C4.5	w1 ➔ 80%	+/- 4.9
– s3: GAssist	w2 ➔ 67%	+/- 4.1
– s5: GAssist	w3 ➔ 52.7%	+/-5.3

Results Overview: Real-HP

- For all algorithms:

  Increase in number of states ➔ decrease in accuracy

  s2: ~63% - ~64% ➔ s5: ~29% - ~30%

- For each state:

  Increased window size ➔ increased accuracy (~1%)

- Best predictions:

– s2: GAssist & C4.5	w3 ➔ 64.4%	+/-0.5
– s3: C4.5	w2 ➔ 45%	+/-0.4
– s5: C4.5	w3 ➔ 30.4%	+/-0.5

Results Overview: Real-AA

- For all algorithms:

  Increase in number of states → decrease in accuracy

  s2: ~68% → s5: ~34%

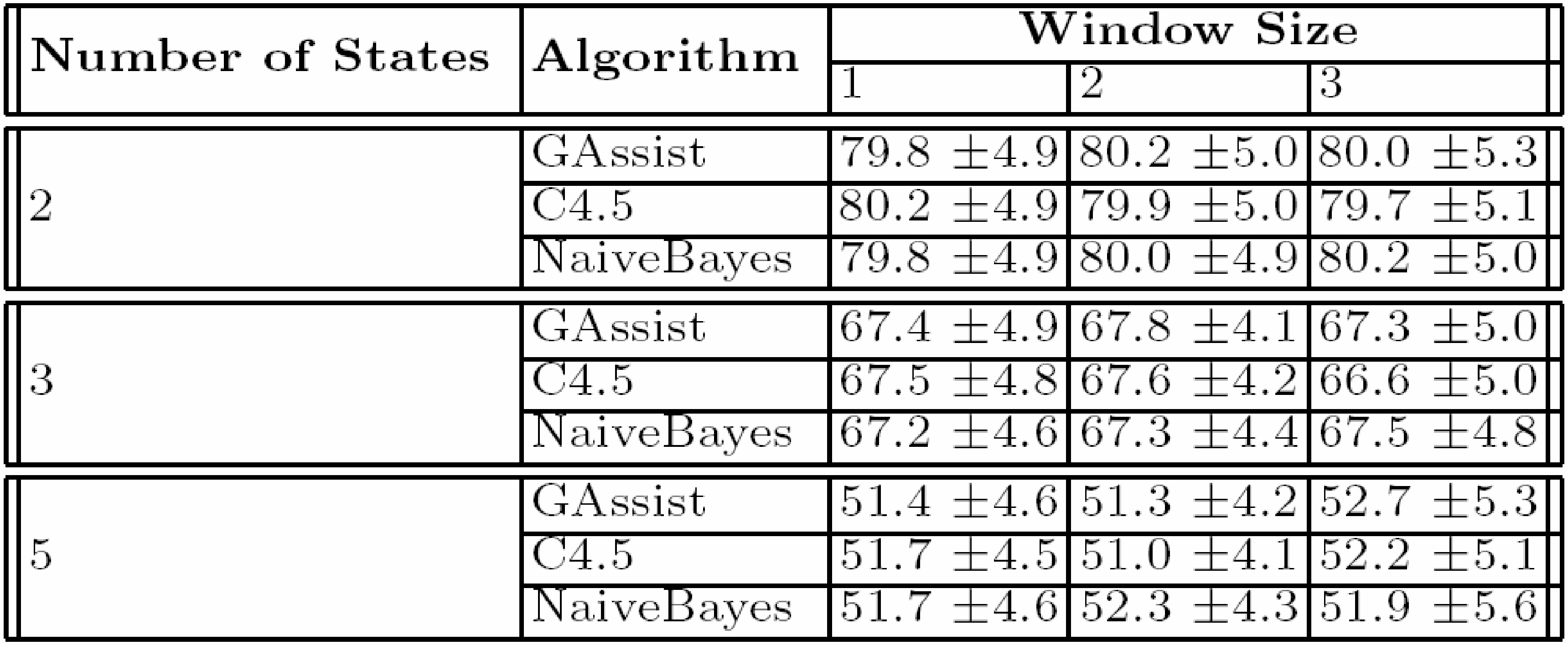
- For each state:

  Increased window size → increased accuracy (~0.5%)

- Best predictions:

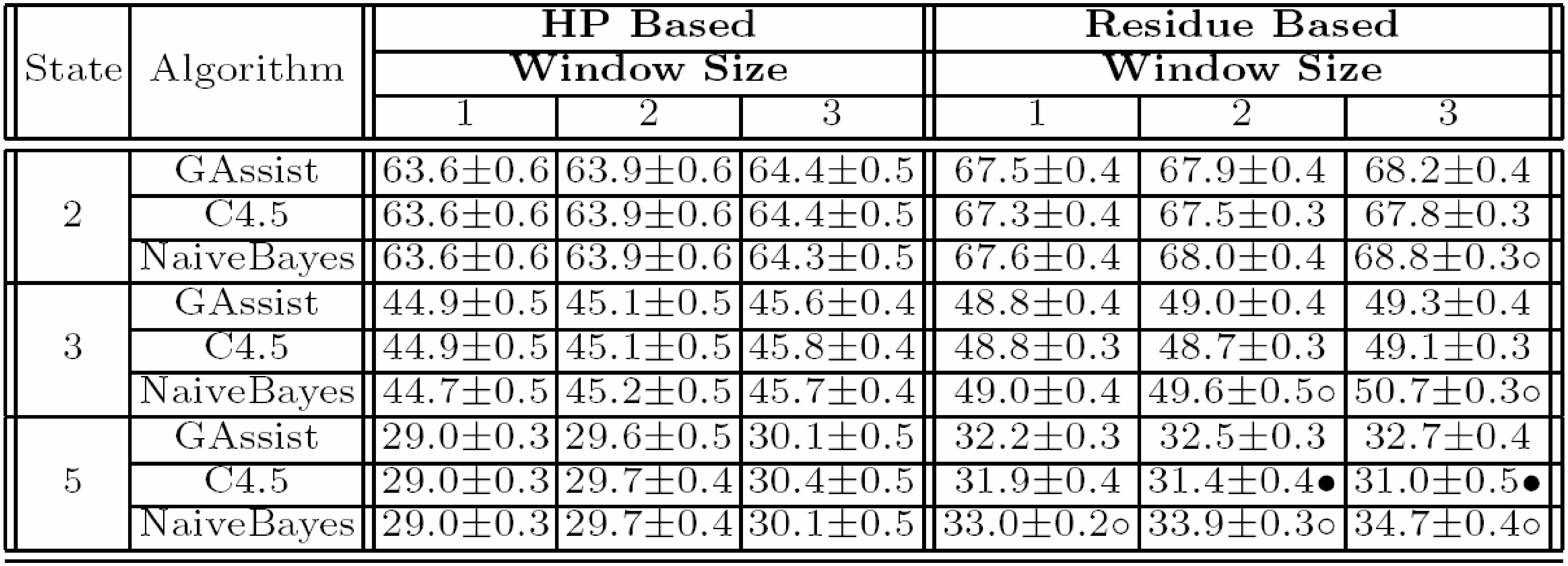
– s2: Naive Bayes	w3 → 68.8%	+-0.3
– s3: Naive Bayes	w3 → 50.7%	+-0.3
– s5: Naive Bayes	w3 → 34.7%	+-0.4

Results: Lattice-HP



Number of States	Algorithm	Window Size		
		1	2	3
2	GAssist	79.8 ±4.9	80.2 ±5.0	80.0 ±5.3
	C4.5	80.2 ±4.9	79.9 ±5.0	79.7 ±5.1
	NaiveBayes	79.8 ±4.9	80.0 ±4.9	80.2 ±5.0
3	GAssist	67.4 ±4.9	67.8 ±4.1	67.3 ±5.0
	C4.5	67.5 ±4.8	67.6 ±4.2	66.6 ±5.0
	NaiveBayes	67.2 ±4.6	67.3 ±4.4	67.5 ±4.8
5	GAssist	51.4 ±4.6	51.3 ±4.2	52.7 ±5.3
	C4.5	51.7 ±4.5	51.0 ±4.1	52.2 ±5.1
	NaiveBayes	51.7 ±4.6	52.3 ±4.3	51.9 ±5.6

Results: Real Proteins



State	Algorithm	HP Based			Residue Based		
		Window Size			Window Size		
		1	2	3	1	2	3
2	GAssist	63.6±0.6	63.9±0.6	64.4±0.5	67.5±0.4	67.9±0.4	68.2±0.4
	C4.5	63.6±0.6	63.9±0.6	64.4±0.5	67.3±0.4	67.5±0.3	67.8±0.3
	NaiveBayes	63.6±0.6	63.9±0.6	64.3±0.5	67.6±0.4	68.0±0.4	68.8±0.3○
3	GAssist	44.9±0.5	45.1±0.5	45.6±0.4	48.8±0.4	49.0±0.4	49.3±0.4
	C4.5	44.9±0.5	45.1±0.5	45.8±0.4	48.8±0.3	48.7±0.3	49.1±0.3
	NaiveBayes	44.7±0.5	45.2±0.5	45.7±0.4	49.0±0.4	49.6±0.5○	50.7±0.3○
5	GAssist	29.0±0.3	29.6±0.5	30.1±0.5	32.2±0.3	32.5±0.3	32.7±0.4
	C4.5	29.0±0.3	29.7±0.4	30.4±0.5	31.9±0.4	31.4±0.4●	31.0±0.5●
	NaiveBayes	29.0±0.3	29.7±0.4	30.1±0.5	33.0±0.2○	33.9±0.3○	34.7±0.4○

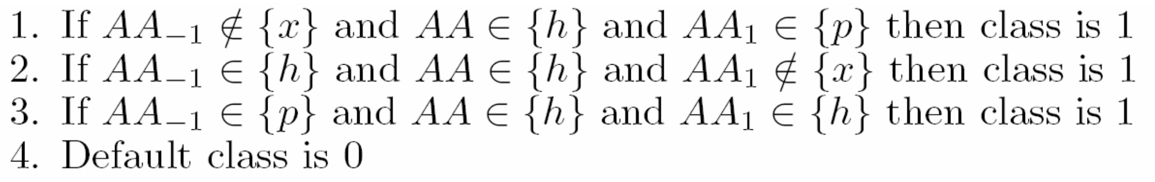
Discussion (1/2)

- All algorithms performed at similar levels
- No statistically significant differences
- Increasing number of classes (states) ➔ reduced accuracy
  - Can be offset using larger window size
- Reduced spatial degrees of freedom (lattice)

  ➔ improvement ~20%, s5
- Moving from 2 to 20 letter representation ➔ 3-5% improvement
- **Indicates hydrophobicity information is key determinant of CN**
  - Consistent with literature
- Shows studies of HP models are relevant in PSP
- LCS evolved rules from the HP representation are simpler

Discussion (2/2)

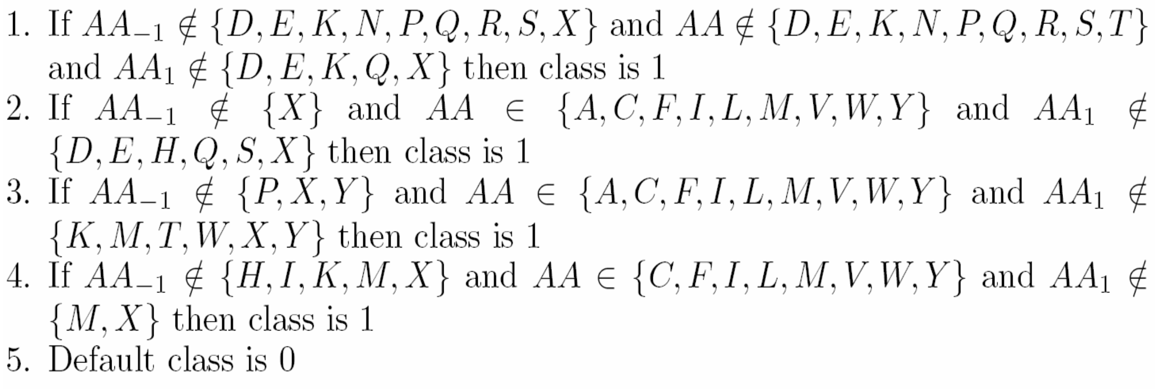
- HP-alphabet (2 letters) rules: simpler & easier to understand – e.g.. rule set with 62.9% accuracy:



1. If $AA_{-1} \notin \{x\}$ and $AA \in \{h\}$ and $AA_1 \in \{p\}$ then class is 1
2. If $AA_{-1} \in \{h\}$ and $AA \in \{h\}$ and $AA_1 \notin \{x\}$ then class is 1
3. If $AA_{-1} \in \{p\}$ and $AA \in \{h\}$ and $AA_1 \in \{h\}$ then class is 1
4. Default class is 0

  - X represents positions at end of chains
  - Class assignment: 1=high, 0=low

- AA-alphabet (20 letters) rules: rule set with 67.7% accuracy:



1. If $AA_{-1} \notin \{D, E, K, N, P, Q, R, S, X\}$ and $AA \notin \{D, E, K, N, P, Q, R, S, T\}$ and $AA_1 \notin \{D, E, K, Q, X\}$ then class is 1
2. If $AA_{-1} \notin \{X\}$ and $AA \in \{A, C, F, I, L, M, V, W, Y\}$ and $AA_1 \notin \{D, E, H, Q, S, X\}$ then class is 1
3. If $AA_{-1} \notin \{P, X, Y\}$ and $AA \in \{A, C, F, I, L, M, V, W, Y\}$ and $AA_1 \notin \{K, M, T, W, X, Y\}$ then class is 1
4. If $AA_{-1} \notin \{H, I, K, M, X\}$ and $AA \in \{C, F, I, L, M, V, W, Y\}$ and $AA_1 \notin \{M, X\}$ then class is 1
5. Default class is 0

Related Work

- Kinjo et al 2005 s2,3,10 CN prediction
  - Obtained higher accuracies
  - Used non-standard accuracy measure & more input information

  Our aim was compare performance using simpler representations

  Not trying for best accuracy

- Real Protein CN prediction by LCS compared with Kinjo Group predictions (papers accepted)

- Detailed studies of HP proteins CN and Residue Exposure prediction (paper accepted)

Conclusions (1/2)

- It is possible to predict CN (5 state, window size 3) using
  - Lattice-HP model proteins ~52%
  - Real-HP representations ~30%
  - Real-AA representation ~32%

  Reasonable since HP representation discards information

- Accuracy using 2 letter representation is close to 20 letter representation
  - 64% vs 68% (s2)
  - 45% vs 50% (s3)
  - 30% vs 33% (s5)

Conclusions (2/2)

- Indicates most information is contained in HP representation
- Hydrophobicity is a key determinant of CN
  - Consistent with earlier studies
- Information inconsistency ratio
  - “Ambiguous antecedents” : “Consequent assignments”
  - 2 letter representation has considerable inconsistency even for s=5 and larger windows
  - Algorithms may learn from distributions inconsistencies

Future Work

- Li et al 2005
  - Is there minimal residue alphabet for prediction?
  - 10 letters may be sufficient
- Investigate other reduced letter alphabets
- Quantify information loss in each

- Extend studies to prediction of other structural attributes
  - Secondary structure, relative solvent accessibility

- Ultimately, determine utility of CN for designing prediction heuristics for Real proteins

Acknowledgments

- Support:
  - UK Engineering and Physical Sciences Research Council (EPSRC) under grant GR/T07534/01
  - Biotechnology and Biological Sciences Research Council (BBSRC) under grant BB/C511764/1

Reference

- *From HP Lattice Models to Real Proteins: coordination number prediction using Learning Classifier Systems*, Stout, M., Bacardit, J., Hirst, J.D., Krasnogor, N. and Blazewicz, J., (2006) LNCS **3907** pp. 208 - 220 (forthcoming)

Questions???

HP Models

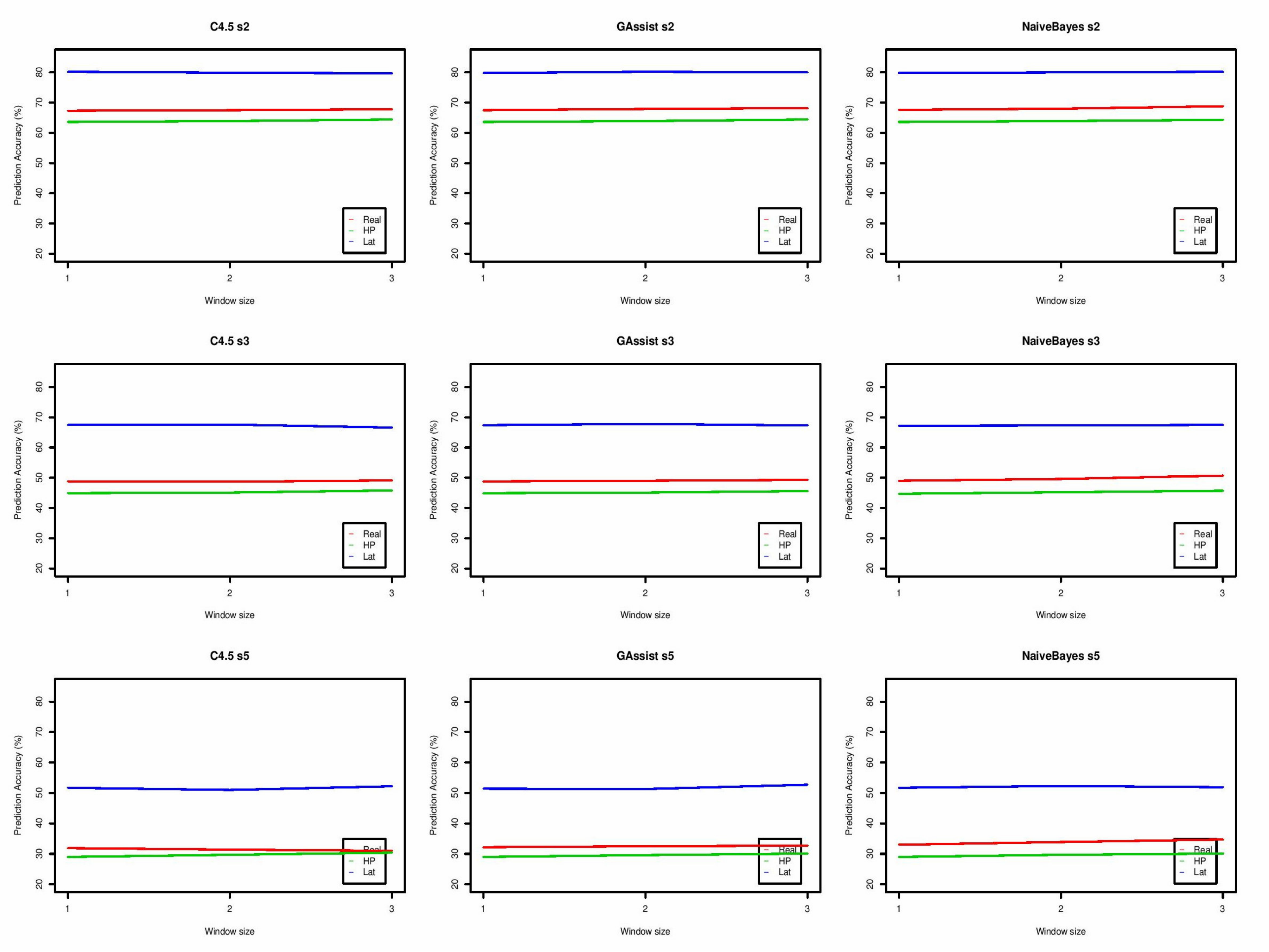
- 20 residue types reduced to 2
  - Non-polar or hydrophobic (H)
  - Polar (P) or hydrophilic
- *n* residue protein represented by sequence *s*
- Sequence is mapped to a lattice
- Each residue in *s* occupies different lattice cell
- Mapping is required to be self-avoiding

$$E(s) = \sum_{i<j\ ;\ 1\leq i,j\leq n} (\Delta_{i,j}\epsilon_{i,j})$$

- Energy potential reflects propensity of H residues to form compact core

$$\Delta_{i,j} = \begin{cases} 1 \text{ if } i, j \text{ are in contact and } |i-j| > 1 \\ 0 \text{ otherwise} \end{cases}$$

- Standard HP model
  - HP and PP assigned energy 0
  - HH contact assigned energy -1
- Optimal structures minimize energy potential


C4.5 s2
GAssist s2
NaiveBayes s2
C4.5 s3
GAssist s3
NaiveBayes s3
C4.5 s5
GAssist s5
NaiveBayes s5
Prediction Accuracy (%)
20
30
40
50
60
70
80
Window size
1
2
3
Real
HP
Lat