Building Portfolios for the Protein Structure Prediction Problem

Alejandro Arbelaez Youssef Hamadi Michele Sebag

Workshop on Constraint Based Methods for Bioinformatics (WCB) - July 2010

Outline

- CSPs
- Protein Structure Prediction Problem
- Algorithm Selection Problem
- Machine learning && Features
- Experimental results
- Conclusions and future work

CSP

- A Constraint Satisfaction Problem (**CSP**) is a triple (X, D, C):

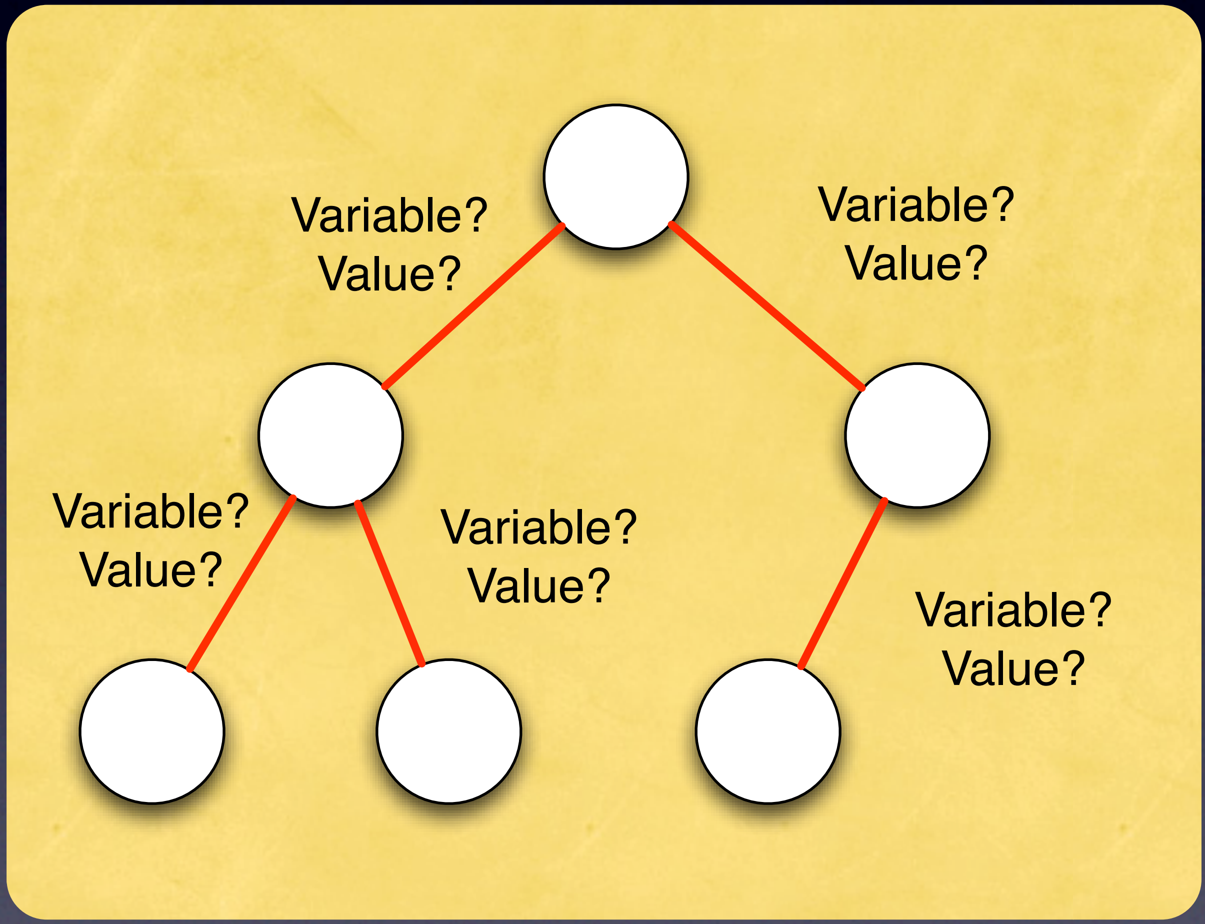
$$
\begin{aligned}
&Variables: \\
&\qquad X_1, X_2, \ldots, X_n \\
&Domains: \\
&\qquad D_1, D_2, \ldots, D_n
\end{aligned}
$$

Constraints:

$$
\begin{array}{lcl}
X_1 + 5 & \leq & X_2 \\
X_6 + X_2 & = & X_1 * X_9 \\
 & \vdots & \\
X_7 - X_3 & \geq & 10
\end{array}
$$

Solution

- Backtracking algorithm:
  - Which Value?
  - Which Variable?



- Depends on the (family of) problems.
- Conditions the effectiveness of the algorithm

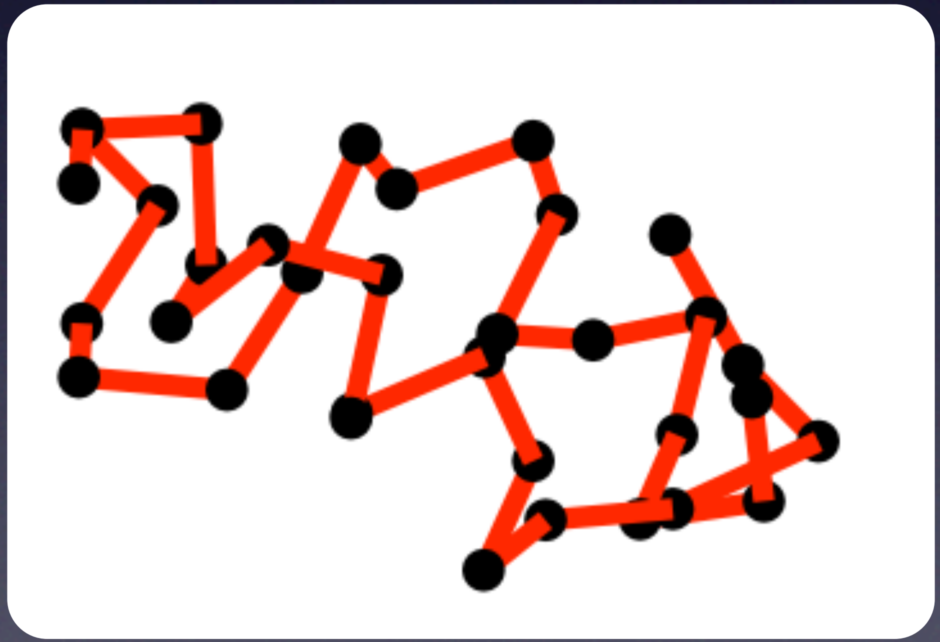
→ In this paper

Protein Structure Prediction Problem

Sequence of amino-acids → 3D conformation

$s_1, s_2, \ldots, s_n$ →

20 amino-acids



protein ID=1ZDDP

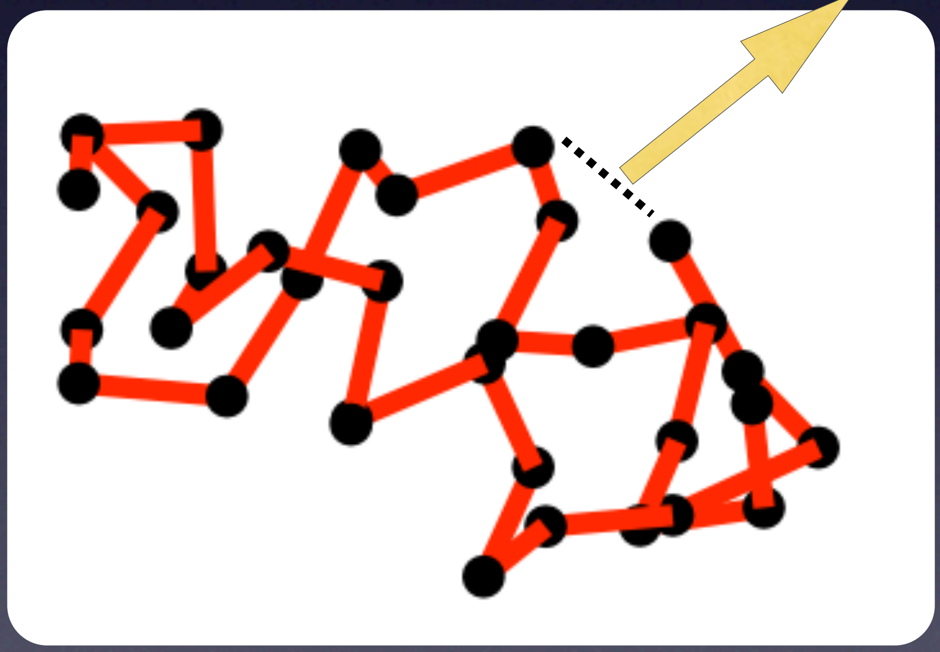
Minimize the energy function

Protein Structure Prediction Problem

Sequence of amino-acids

3D conformation

Energy contribution



protein ID=1ZDDP

$s_1, s_2, \ldots, s_n$

20 amino-acids

Minimize the energy function

Protein Structure Prediction Problem

- HP Models
  - 20 symbols alphabet => 2 symbols alphabet
- Lattice Models (a FCC lattice)

$$E(w) = \sum_{1 \leq i < n} \sum_{i+2 \leq j \leq n} contact(w(i), w(j)) \times Pot(s_i, s_j)$$

Protein Structure Prediction Problem

- Which heuristic to use?

- dom/wdeg
- wdeg
- domFD
- min-dom
- ...

Well known CSP heuristics

Protein Structure Prediction Problem

- Which heuristic to use?

We can use Paul the octopus to predict the best heuristic

Protein Structure Prediction Problem

- Which heuristic to use?

We can use Paul the octopus to predict the best heuristic

but ... Paul is now retired!

What about using machine learning to select the most appropriate heuristic?

Protein Structure Prediction Problem

- Of course, one could also use a problem domain heuristic, but....

Algorithm Selection

Classification
problem

Algorithm Selection

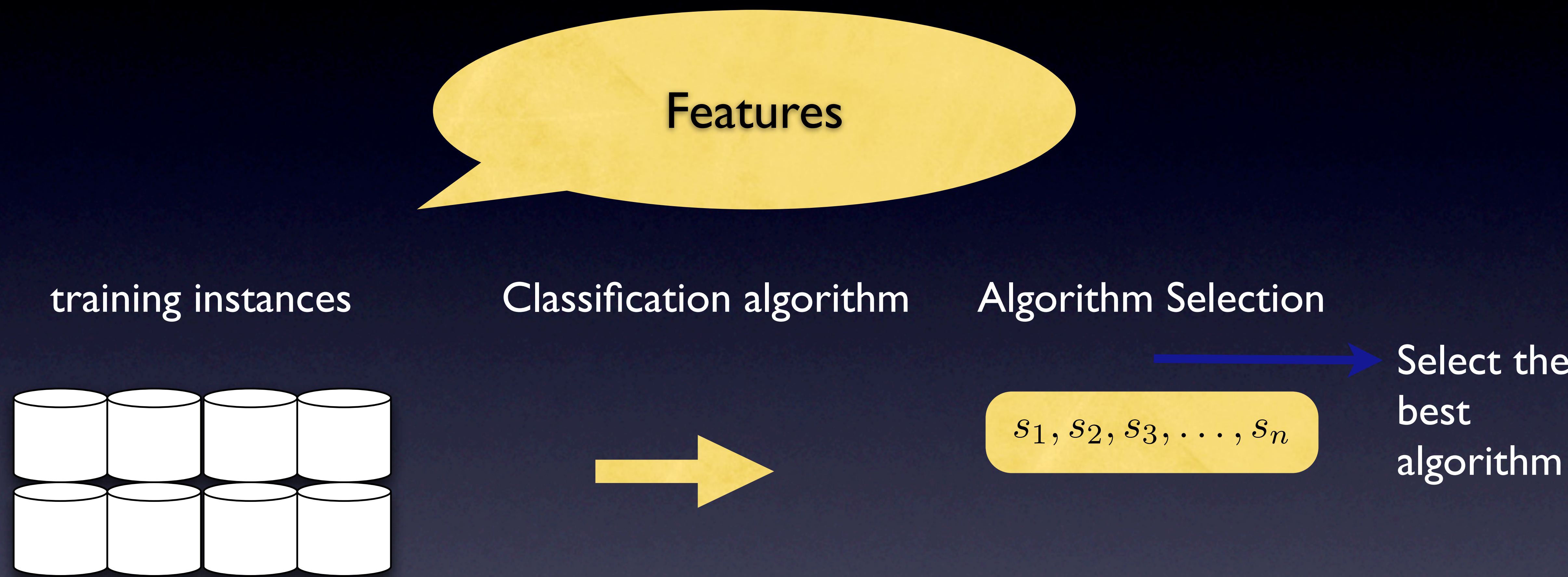
Classification
problem

Features $\rightarrow \mathbb{R}^d$

$s_1, s_2, s_3, \ldots, s_n$

- dom/wdeg
- wdeg
- domFD
- min-dom
- ...

Algorithm Selection



- For each training instance:
  - Compute the best strategy based on algorithm's cost solution.
  - Build the classifier on the training set

General Methodology

Off-line

- Try various heuristics
- Record the corresponding solutions
- Compute features

Target problem Distribution → Compute training information → Features Pre-processing → Learn a classification model

Online

Compute features → Feature Normalization → Best Heuristic

Decision Trees

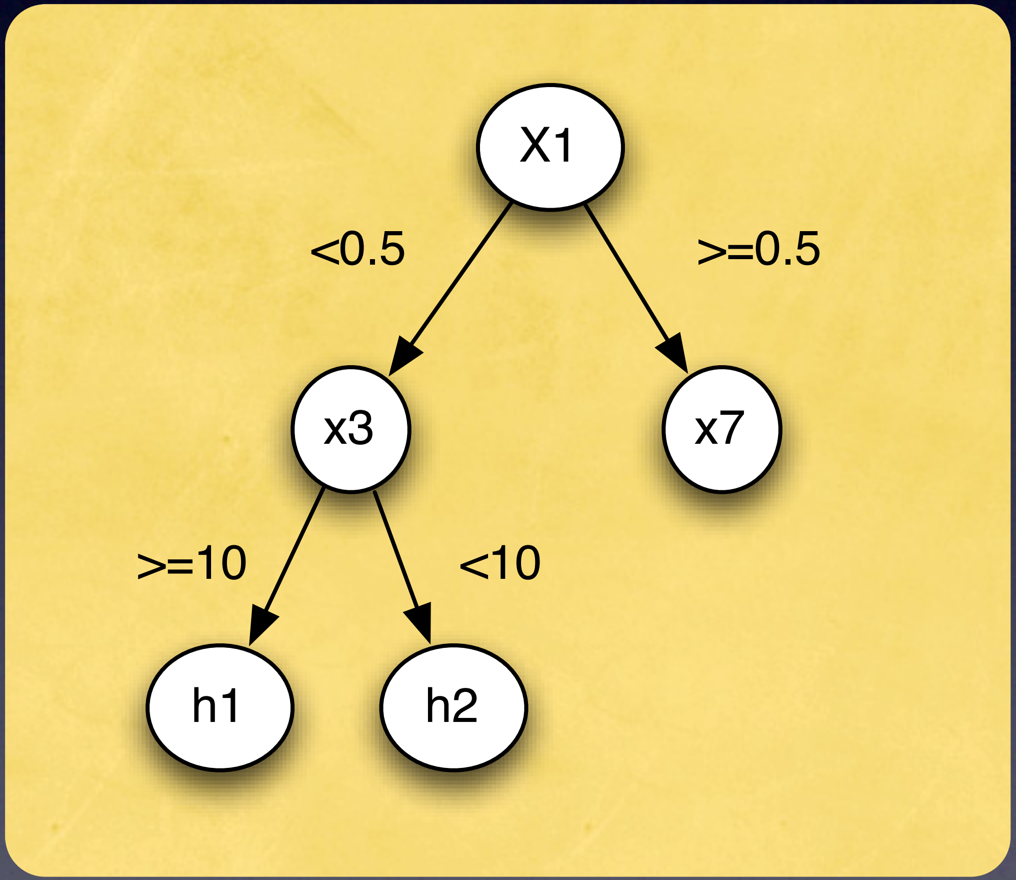
A well-known learning algorithm for classification

- Training set:

features

$$Inst = \{x_1, x_2, \ldots, x_n\} \rightarrow \{\ h_1, h_2, \ldots, h_m\}$$

- Tree structure:
  - Node => Feature
  - Branching => Decision
  - leaf node => Label



Algorithm Selection

- dom/wdeg
- wdeg
- domFD
- min-dom
- ...

→

- Algorithm with best solution cost is labeled as winner during the training phase

Features

- Machine Learning && Protein Classification
  - Highly studied problem in Computational biology

Let's use classical descriptors to build a portfolio algorithm

Features

$$\begin{tikzcd} s_1, \arrow[d] & s_2, \arrow[d] & s_3, \arrow[d] & \ldots, & s_n \arrow[d] \\ \{1,2,3\} & \{1,2,3\} & \{1,2,3\} & & \{1,2,3\} \end{tikzcd}$$

Attribute	Group 1	Group 2	Group 3
Hydrophobicity	R,K,E,D,Q,N	G,A,S,T,P,H,Y	C,V,L,I,M,F,W
Volume	G,A,S,C,T,P,D	N,V,E,Q,I,L	M,H,K,F,R,Y,W
Polarity	L,I,F,W,C,M,V,Y	P,A,T,G,S	H,Q,R,K,N,E,D
Polarizability	G,A,S,D,T	C,P,N,V,E,Q,I,L	K,M,H,F,R,Y,W

Features

$$
\begin{array}{cccccc}
R, & S, & T, & V, & V, & H \\
\downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
1 & 2 & 2 & 3 & 3 & 2
\end{array}
$$

Attribute	Group 1	Group 2	Group 3
Hydrophobicity	**R,K,E,D,Q,N**	**G,A,S,T,P,H,Y**	**C,V,L,I,M,F,W**
Volume	G,A,S,C,T,P,D	N,V,E,Q,I,L	M,H,K,F,R,Y,W
Polarity	L,I,F,W,C,M,V,Y	P,A,T,G,S	H,Q,R,K,N,E,D
Polarizability	G,A,S,D,T	C,P,N,V,E,Q,I,L	K,M,H,F,R,Y,W

Features

- Composition: 3 descriptors representing the percentage of each group in the sequence
- Transition: 3 descriptors representing the frequency with which a residue from group(i) is followed by a residue from group(i+1), or vise-versa
- Distribution: 15 Descriptors representing the fraction in the sequence where the first residue, 25%, 50%, 75% and 100% of the residues are contained.

105 Descriptors: 84 ((15+3+3)*4))
20 (amino-acids)
1 (size)

Experiments

- 400 Random sequences
- 10 fold-cross validation
- Machine Learning Algorithm => C4.5
- We have used the Gecode model proposed in Cipriano, Dal Palu, Dovier. WCB'08

Experiments

- Experimented with 18 heuristics candidates to build the portfolio.
- Manual selection of Heuristics candidate:
  - <lexico,min-val>, <domFD+, med-val>, <wdeg, med-val>, <wdeg+, med-val>

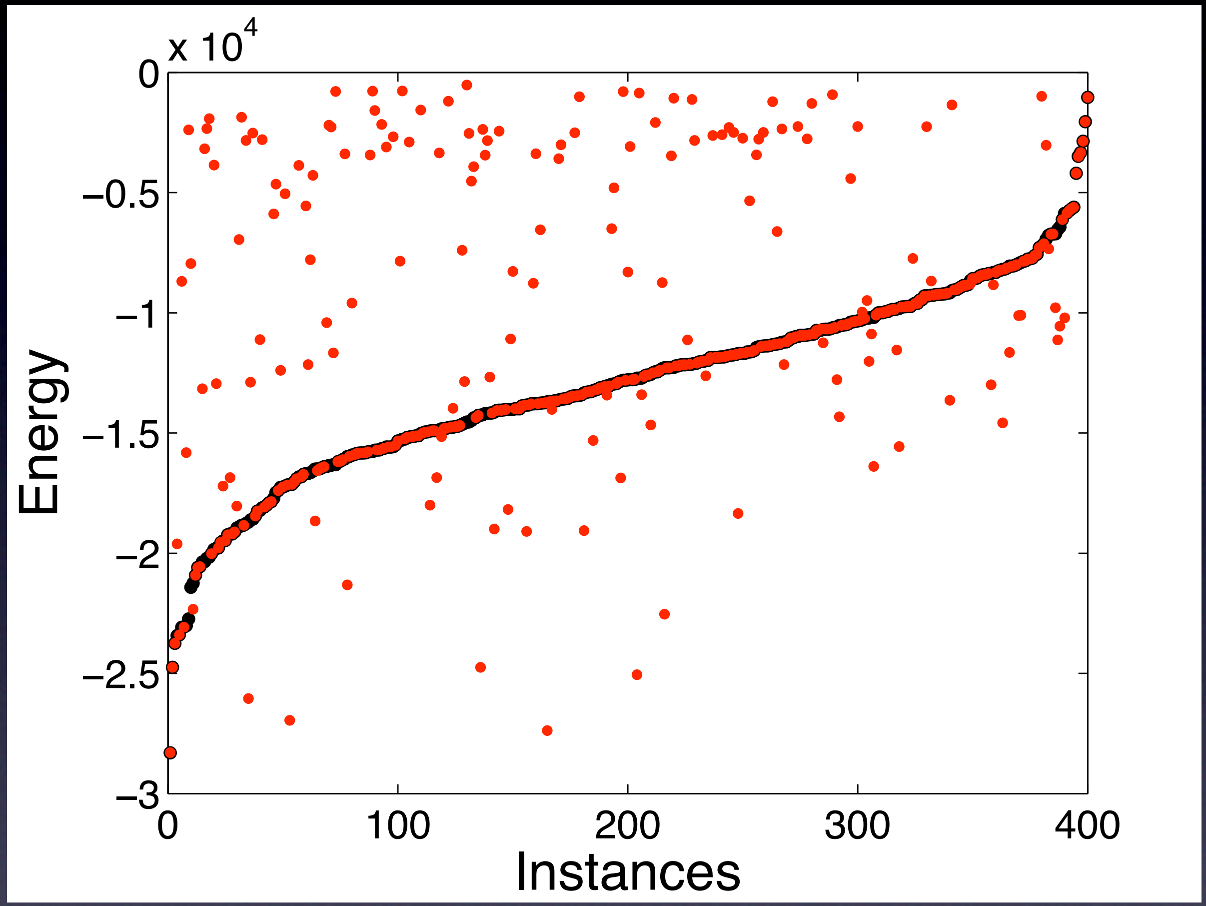
Best heuristics

Experiments

- We perform 10-fold cross validation

P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Test	Train								
Train	Test	Train							
Train		Test	Train						
Train			Test	Train					
Train				Test	Train				
Train					Test	Train			
Train						Test	Train		
Train							Test	Train	
Train								Test	Train
Train									Test

Experiments



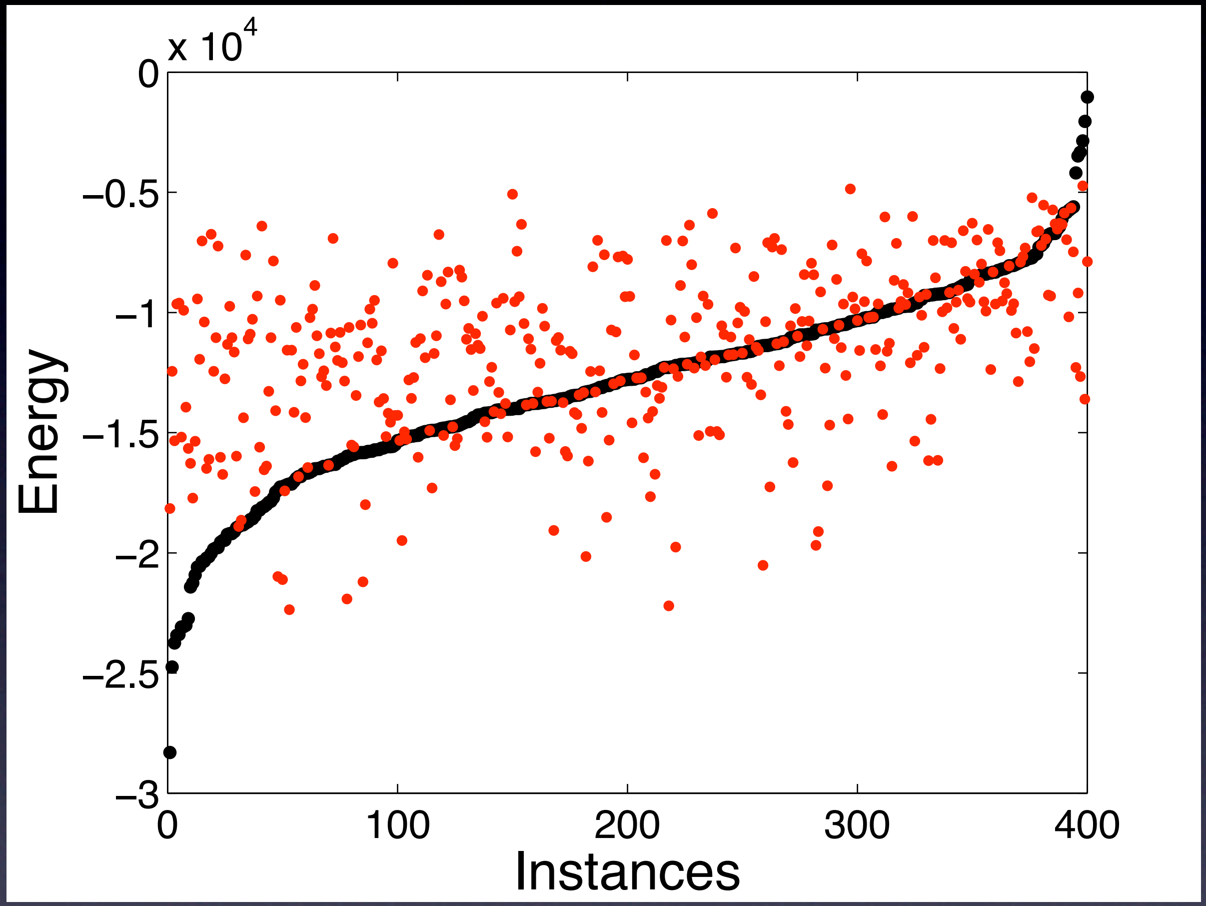
Black points
automatic alg. selection

Better in 110 instances

Red points
best single heuristic

Better in 43 instances

Experiments



Black points
automatic alg. selection

Better in 213 instances

Red points
2nd best single heuristic

Better in 127 instances

Conclusions

- A CP Solver can automatically choose feasible heuristics considering features of the original problem
- We need to select good heuristics for building the portfolio

Future work

- Future work => Ongoing work
- Experimenting with real sequences
- Automatic selection of the algorithms candidates
- Using features based on the CSP codification of the problem

Thanks for your attention

Questions and Comments?