Outline

- CSPs
- Protein Structure Prediction Problem
- Algorithm Selection Problem
- Machine learning && Features
- Experimental results
- Conclusions and future work
A Constraint Satisfaction Problem (CSP) is a triple \((X, D, C)\):

- **Variables**: \(X_1, X_2, \ldots, X_n\)
- **Domains**: \(D_1, D_2, \ldots, D_n\)
- **Constraints**:
  
  \[
  \begin{align*}
  X_1 + 5 & \leq X_2 \\
  X_6 + X_2 & = X_1 \times X_9 \\
  \vdots & \\
  X_7 - X_3 & \geq 10
  \end{align*}
  \]
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

Minimize the energy function

protein ID=1ZDDP
Protein Structure Prediction Problem

Sequence of amino-acids

$s_1, s_2, \ldots, s_n$

20 amino-acids

3D conformation

Minimize the energy function

Energy contribution

protein ID=1ZDDP
Protein Structure Prediction Problem

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

\[ E(w) = \sum_{1 \leq i < n} \sum_{i+2 \leq j \leq n} \text{contact}(w(i), w(j)) \times \text{Pot}(s_i, s_j) \]
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
- ...

Wednesday, July 21, 2010
Algorithm Selection

For each training instance:

- Compute the best strategy based on algorithm’s cost solution.
- Build the classifier on the training set.

Features

Select the best algorithm

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

Off-line

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

Online

1. Target problem Distribution
2. Compute training information
3. Features Pre-processing
4. Learn a classification model

- Compute features
- Feature Normalization
- Best Heuristic

Wednesday, July 21, 2010
Decision Trees

A well-known learning algorithm for classification

- **Training set:**
  \[ \text{Inst} = \{x_1, x_2, \ldots, x_n\} \rightarrow \{h_1, h_2, \ldots, h_m\} \]

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

features
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

The image shows a set of features represented as $S_1, S_2, S_3, \ldots, S_n$ with corresponding indices $\{1, 2, 3\}$.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
</tr>
</thead>
</table>

- **Hydrophobicity** includes R, K, E, D, Q, N.
- **Volume** includes G, A, S, C, T, P, D.
- **Polarity** includes L, I, F, W, C, M, V, Y.
- **Polarizability** includes G, A, S, D, T.

These attributes are categorized into different groups for further analysis.
# Features

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polarizability</td>
<td><strong>G,A,S,D,T</strong></td>
<td><strong>C,P,N,V,E,Q,I,L</strong></td>
<td><strong>K,M,H,F,R,Y,W</strong></td>
</tr>
</tbody>
</table>
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:
  - $<\text{lexico}, \text{min-val}>, <\text{domFD}+, \text{med-val}>, <\text{wdeg}, \text{med-val}>, <\text{wdeg}+, \text{med-val}>$

Best heuristics
Experiments

- We perform 10-fold cross validation

<table>
<thead>
<tr>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
<th>P7</th>
<th>P8</th>
<th>P9</th>
<th>P10</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test</strong></td>
<td>Train</td>
<td>Test</td>
<td>Train</td>
<td>Test</td>
<td>Train</td>
<td>Test</td>
<td>Train</td>
<td>Test</td>
<td>Train</td>
</tr>
</tbody>
</table>
Experiments

Black points
automatic alg. selection

Red points
best single heuristic

Better in 110 instances
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?