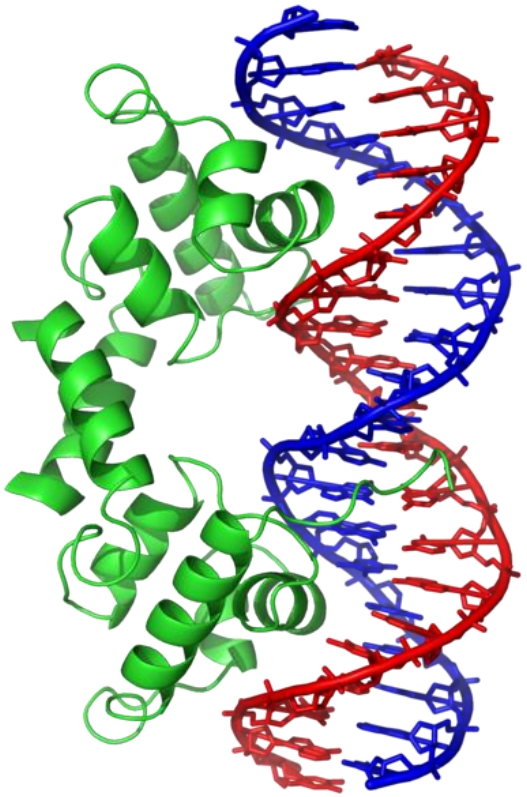
Prediction of DNA-binding Proteins from Structural Features

Andrea Szabóová
Ondřej Kuželka
Filip Železný
Jakub Tolar

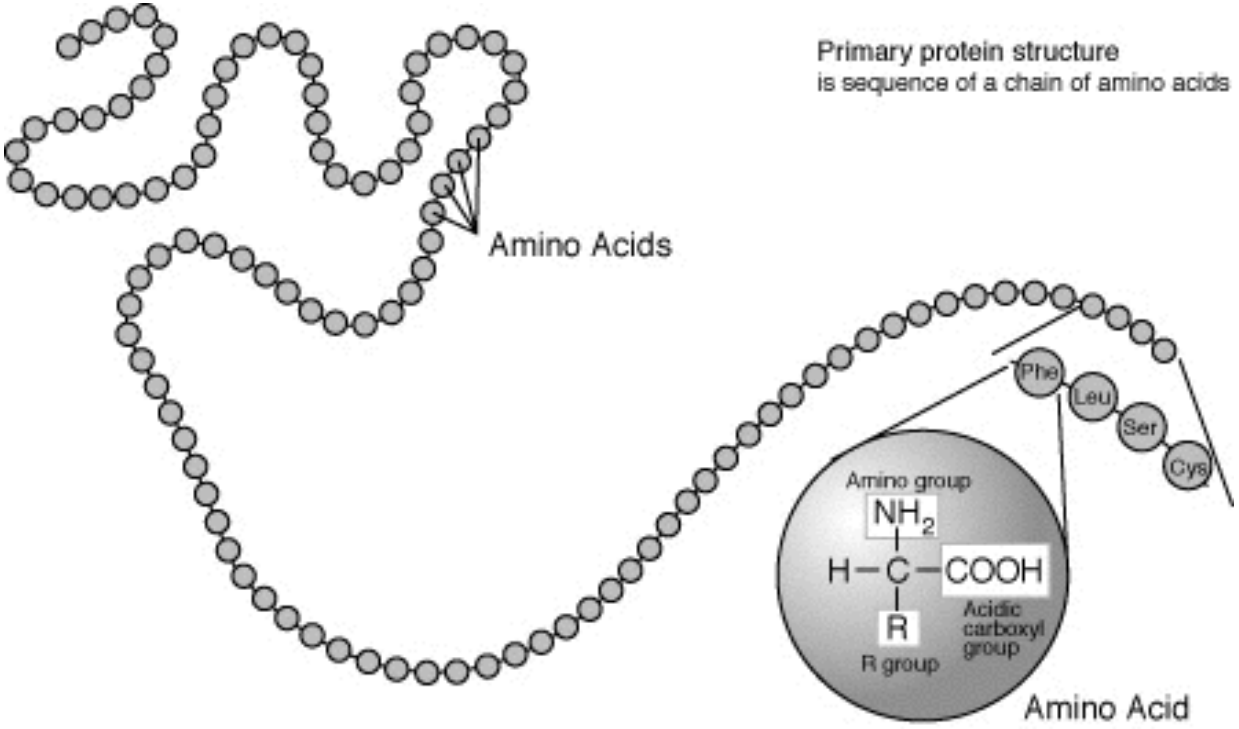
Introduction
DNA-Binding Proteins



- Proteins containing DNA-binding Domains
- DNA-binding Domain is an independently folded protein domain containing at least one motif that recognizes DNA

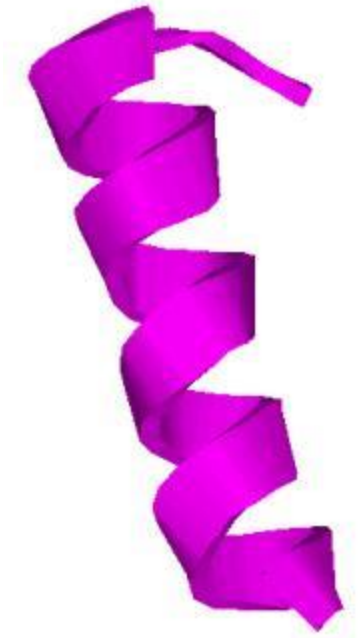
Introduction
Protein – Primary Structure

Amino Acid Sequence



Introduction
Protein – Secondary Structure

Alpha-helix



Beta-sheet

The Problem That We Try To Solve

Problem: Given a **spatial structure** of a protein, construct a classifier for predicting whether the protein binds to DNA or not.

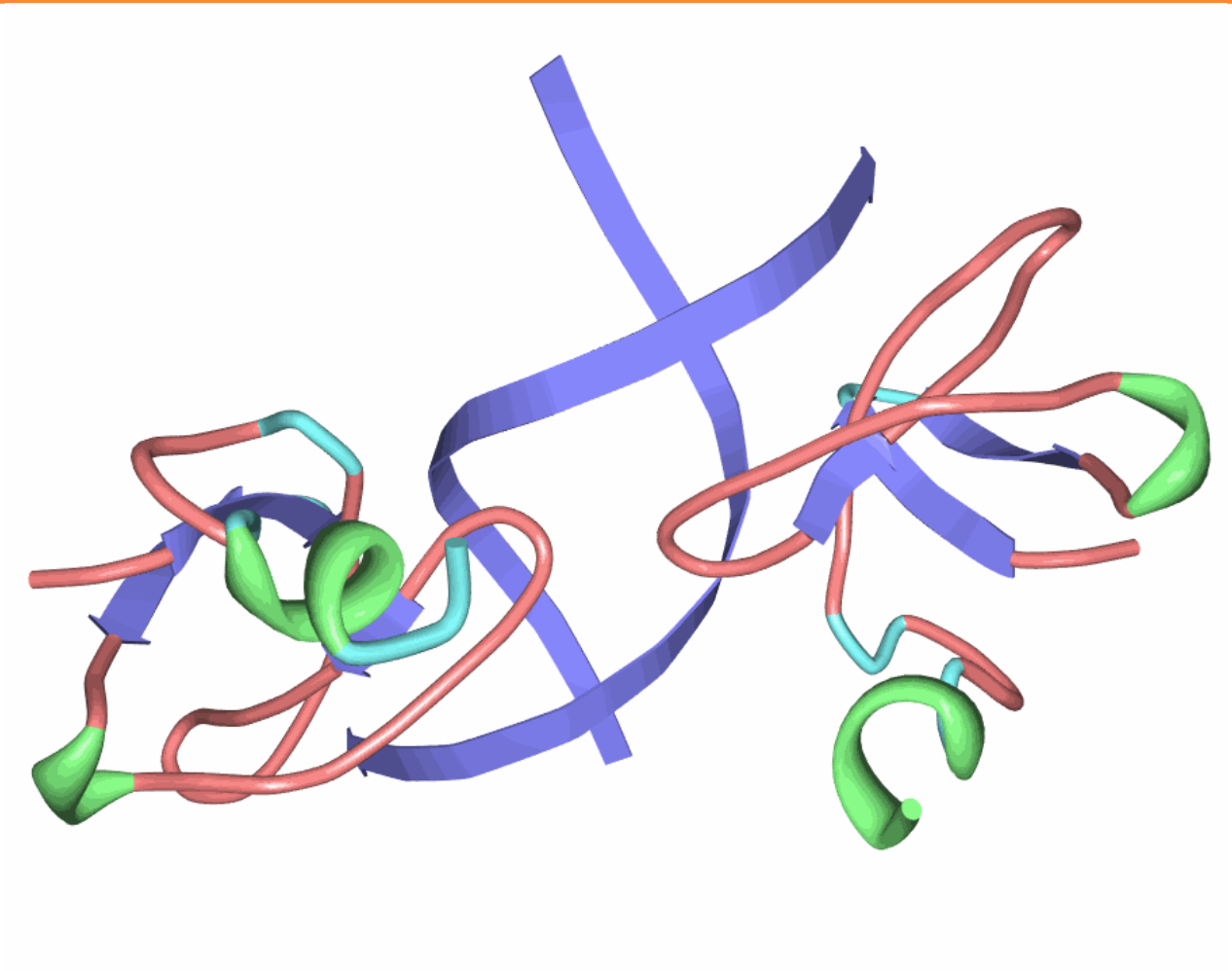
- Many approaches have been developed in literature to tackle this problem

Additional Requirements: A learning algorithm for this task should be able to bring us **additional insights** into the protein-DNA interaction mechanism.

- This has not received much attention in the literature. We will show that our approach could bring such kind of novel information.

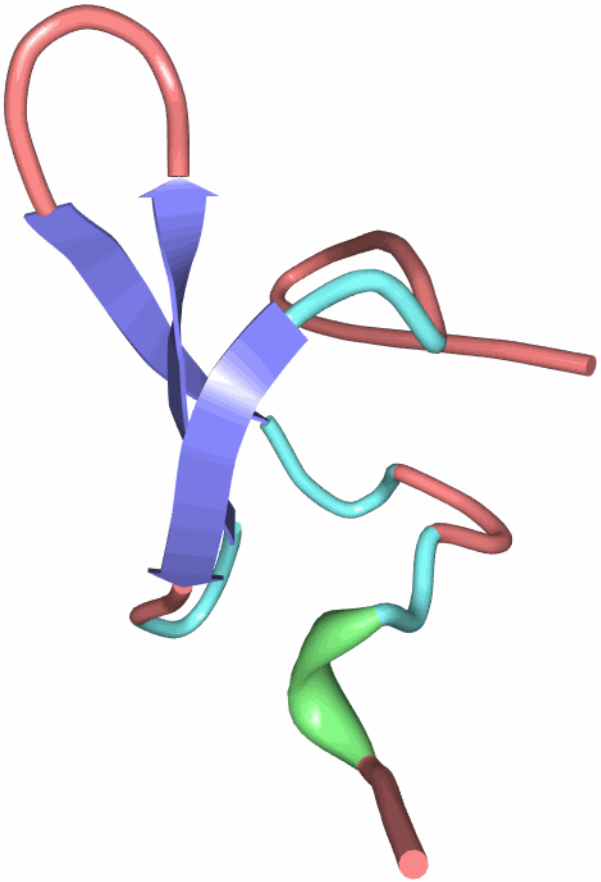
Available Training Data (Bounded Conformations)

DNA-binding Protein - Bounded conformation



Available Training Data (Unbounded Conformations)

DNA-binding Protein - Unbounded conformation



Previous Work

Approaches using Evolutionary Information:

- Some algorithms use information about evolutionary conserved parts of proteins (so-called patches).
- Rationale: If some part of a protein remains unchanged during evolution then it is likely important for the protein's function.
- e.g. Nimrod et al., Identification of DNA-binding Proteins Using Structural, Electrostatic and Evolutionary Features, *J. Mol. Biol.* (2009)
- **A problem is that these approaches would not work well with engineered proteins (what is one of the main aims of our running project)**

Previous Work

Approaches NOT using Evolutionary Information:

- These approaches use mainly structural information or sometimes only the sequence information to predict whether a protein is DNA-binding or not
- In principle, they should be more applicable to engineered proteins
- One of state-of-the-art approaches: Szilagyi A, Skolnick J, Efficient prediction of nucleic acid binding function from low-resolution protein structures. *J Mol Biol, 2006*

The Work by Szilagyi et al.

Szilagyi A, Skolnick J, Efficient prediction of nucleic acid binding function from low-resolution protein structures. *J Mol Biol, 2006*)

- A logistic regression classifier based on the amino acid composition, the asymmetry of the spatial distribution of specific residues and the dipole moment of the protein.
- The features have been hand-crafted by biologists!
- **We will show that ILP can construct better features outperforming Szilagyi's features**

Our Approach

Protein Data Bank

PDB files

```
SEQRES   1 A   71  MET SER VAL ALA CYS LEU SER CYS ARG LYS ARG HIS ILE
SEQRES   2 A   71  LYS CYS PRO GLY GLY ASN PRO CYS GLN LYS CYS VAL THR
SEQRES   3 A   71  SER ASN ALA ILE CYS GLU TYR LEU GLU PRO SER LYS LYS
SEQRES   4 A   71  ILE VAL VAL SER THR LYS TYR LEU GLN GLN LEU GLN LYS
SEQRES   5 A   71  ASP LEU ASN ASP LYS THR GLU GLU ASN ASN ARG LEU LYS
SEQRES   6 A   71  ALA LEU LEU LEU GLU ARG
SEQRES   1 B   71  MET SER VAL ALA CYS LEU SER CYS ARG LYS ARG HIS ILE
SEQRES   2 B   71  LYS CYS PRO GLY GLY ASN PRO CYS GLN LYS CYS VAL THR
 ...
HELIX    1   1 LEU A   35  LYS A   39  1                                   5
HELIX    2   2 LYS A   52  THR A   55  1                                   4
HELIX    3   3 THR A   73  LEU A   97  1                                  25
HELIX    4   4 LEU B   35  ARG B   40  1                                   6
HELIX    5   5 LYS B   52  THR B   55  1                                   4
HELIX    6   6 THR B   73  LEU B   97  1                                  25
SHEET    1   A 2 ILE A  69  THR A  73  0
SHEET    2   A 2 ILE B  69  THR B  73 -1
 ...
ATOM      1  N   MET A  30      16.046 -16.401 -31.079  1.00  0.00           N
ATOM      2  CA  MET A  30      14.761 -15.656 -30.943  1.00  0.00           C
ATOM      3  C   MET A  30      14.036 -15.604 -32.291  1.00  0.00           C
ATOM      4  O   MET A  30      14.582 -15.965 -33.315  1.00  0.00           O
ATOM      5  CB  MET A  30      15.163 -14.250 -30.497  1.00  0.00           C
ATOM      6  CG  MET A  30      15.106 -14.163 -28.971  1.00  0.00           C
ATOM      7  SD  MET A  30      14.926 -12.433 -28.472  1.00  0.00           S
ATOM      8  CE  MET A  30      15.473 -12.633 -26.758  1.00  0.00           C
ATOM      9  H1  MET A  30      15.869 -17.321 -31.528  1.00  0.00           H
ATOM     10  H2  MET A  30      16.706 -15.849 -31.666  1.00  0.00           H
ATOM     11  H3  MET A  30      16.461 -16.551 -30.138  1.00  0.00           H
ATOM     12  HA  MET A  30      14.135 -16.115 -30.196  1.00  0.00           H
ATOM     13  HB2 MET A  30      16.169 -14.040 -30.833  1.00  0.00           H
ATOM     14  HB3 MET A  30      14.482 -13.528 -30.922  1.00  0.00           H
```

Data

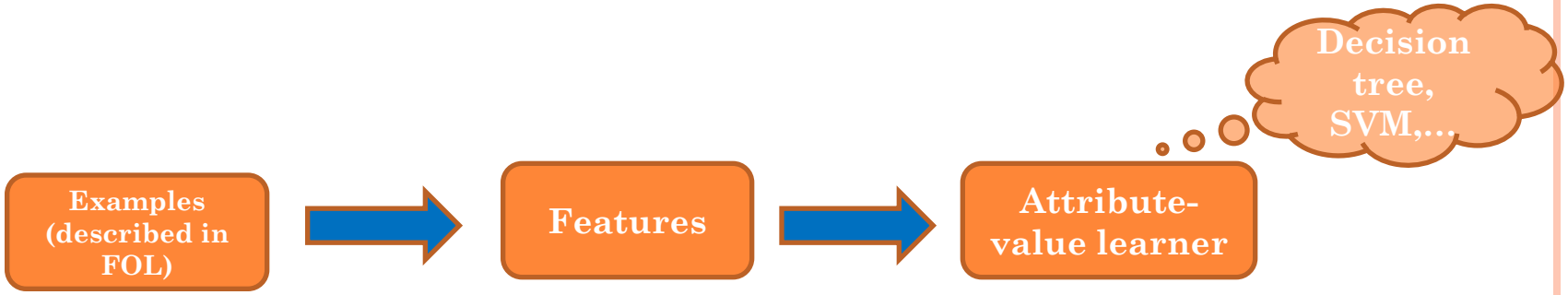
- Positive Data Set
  - 54 DNA-binding proteins in unbounded conformation
- Negative Data Set
  - 110 non-DNA-binding proteins
- Extracted Information
  - List of residues (10's of residue entries per protein)
  - List of pair wise spatial distances among all residues (100's and 1000's entries per protein)
  - Proportion of residues: ARG, LYS, ASP, ALA and GLY
  - Spatial asymmetry of residues: ARG, GLY, ASN and SER
  - Dipole moment

FOL Representation

- Proteins described by formal-logic assertions, e.g.:
  - res('1AYJ', r1, 'CYS')
  - dist(r1, r2, 10)
- Complete Description of a Protein – logical conjunction of formal-logic assertions for all residues and their all pair wise spatial distances
- Real protein – conjunction of 10 000's of literals
- Features
  - Feature F – conjunction of first order literals
  - *F = res(P,R1,'CYS'),res(P,R2,'HIS'),dist(R1,R2,8)*

Propositionalisation

- Propositionalisation – process of transforming a multi-relational dataset into a propositional dataset with derived attribute-value features



- The employed feature construction algorithm RelF* constructs a set of features which are not redundant and have a frequency higher than a given threshold

*Kuzelka O., Zelezny F.: Block-Wise Construction of Tree-like Relational Features with Monotone Reducibility and Redundancy. *Machine Learning*, 2010

Counting Features vs. Existential Features

- Two options how to create the attribute value table:
  - 1. Existential* – a feature acquires value *T (true)* for a protein if it is present in the protein at least once
  - 2. Counting – a feature acquires value *n (integer)* for a protein if it is present in the protein exactly *n*-times

Existential Features

	F_1	…	F_N
Protein 1	T	…	F
…	…	…	…
Protein M	T	…	T

Counting Features

	F_1	…	F_N
Protein 1	5	…	0
…	…	…	…
Protein M	3	…	1

*Houssam N. et al., An Inductive Logic Programming Approach to Validate Hexose Binding Biochemical Knowledge. *ILP* 2009

Classification

- Once we have a sufficiently rich set of features, we may feed the features into any attribute-value learning algorithm
- 7 state-of-the-art attribute-value learning algorithms:
  - Linear SVM
  - SVM with radial basis kernel
  - Simple logistic regression
  - L_2-regularized logistic regression
  - Ada-boost (with decision stamps)
  - Random forest
  - J48 decision tree

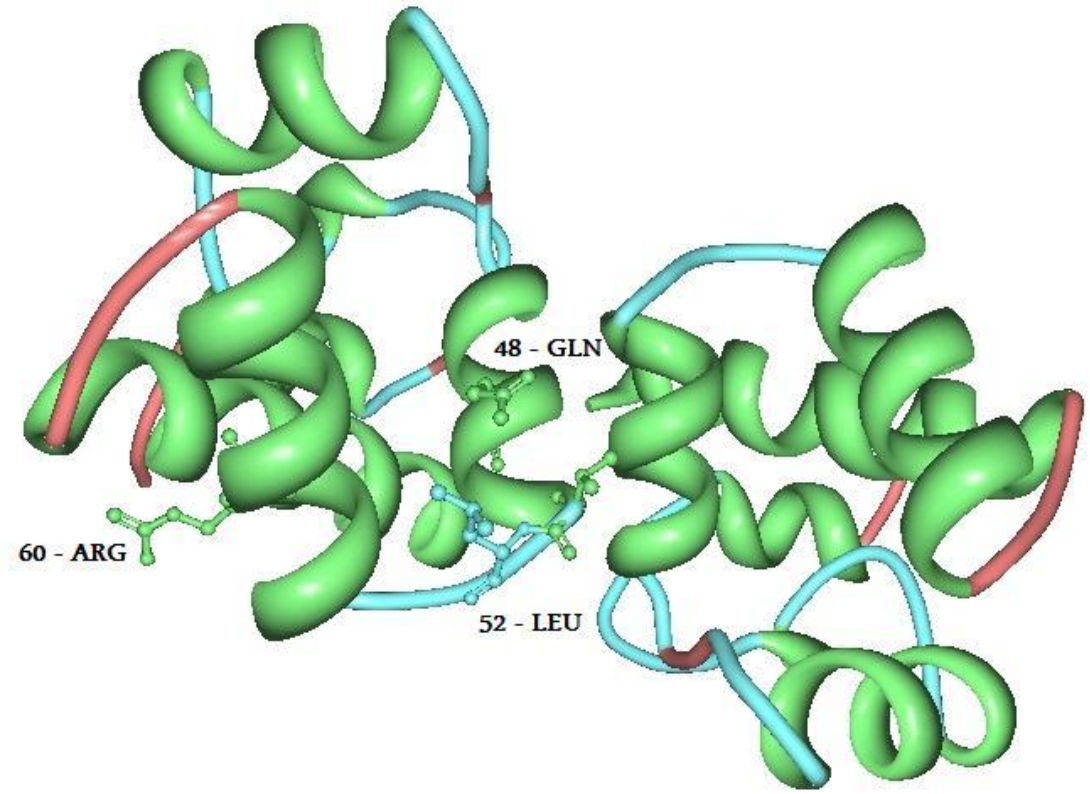
Results

- We obtained about 1500 structural patterns for 54 unbounded DNA-binding proteins
- Accuracies obtained by stratified 10-fold crossvalidation using coarse-grained features (F1), our structural pattern features (F2) and combination of both of them (F1+2)

Classifier	F1	F2(NC)	F1+2(NC)	F2(C)	F1+2(C)
Linear SVM	84.0 (2)	77.5 (5)	78.1 (4)	83.0 (3)	**84.2** (1)
SVM with radial basis kernel	81.6 (3)	67.1 (4-5)	67.1 (4-5)	83.0 (2)	**85.4** (1)
Simple logistic regression	81.6 (3)	73.9 (5)	78.8 (4)	**87.6** (1)	82.3 (2)
L_2-regularized logistic regression	84.0 (2)	78.7 (5)	80.5 (4)	82.4 (3)	**84.2** (1)
Ada-boost (with decision stamps)	77.4 (4)	73.2 (5)	83.0 (2)	79.3 (3)	**84.7** (1)
Random forest	78.6 (4)	76.8 (5)	**83.6** (1)	80.5 (2)	79.9 (3)
J48 decision tree	75.0 (3)	70.7 (4)	75.6 (2)	68.1 (5)	**76.2** (1)
Average ranking:	3	4.79	3.07	2.71	**1.43**

Features Give Us Insight!

Example of A Discovered Structural Feature



- F = res(P,R1,'ARG'), res(P,R2,'GLN') , res(P,R3,'LEU'), dist(R1,R2,10.0), dist(R1,R3,10.0)

Conclusions and Future Work

- It turns out that an important factor contributing to the high predictive accuracies is that the features are not Boolean but rather are assigned values counting the occurrences of the corresponding spatial pattern in the protein.

- We are currently trying to further improve the predictions by incorporating further background knowledge describing the protein domains.

Thank You For Your Attention!