

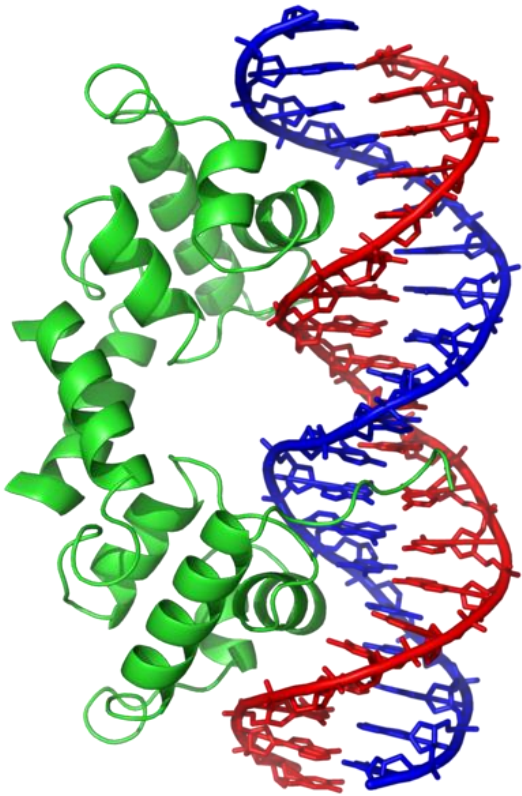


# **PREDICTION OF DNA-BINDING PROPENSITY OF PROTEINS**

**Andrea Szabóová  
Ondřej Kuželka  
Sergio Morales E.  
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




# 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

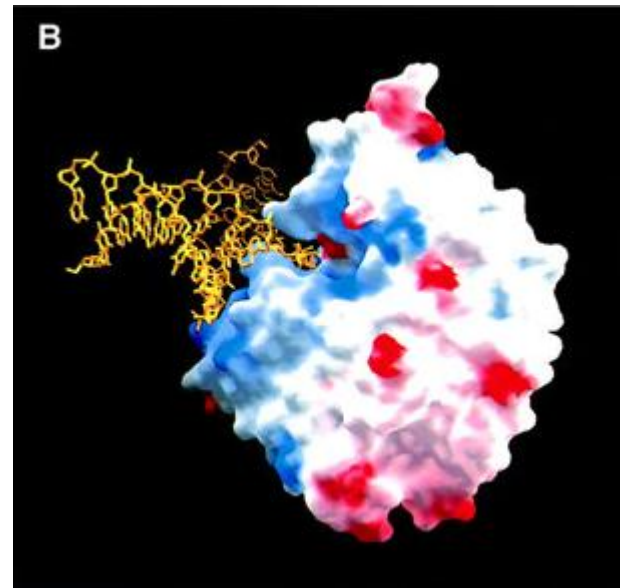
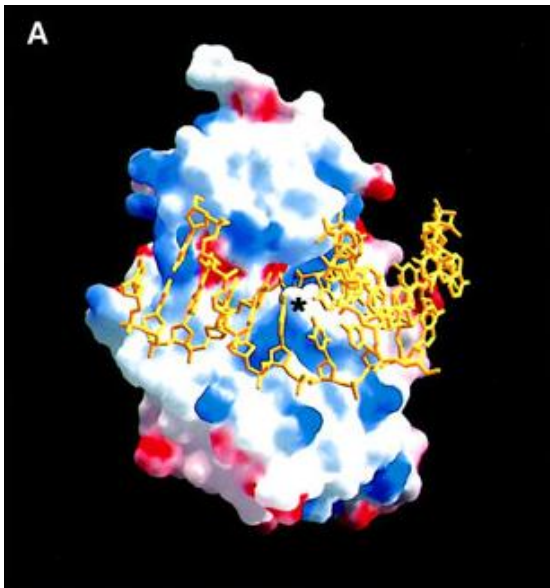
**Additional Requirements:** A learning algorithm for this task should be able to bring us **additional insights** into the protein-DNA interaction mechanism.

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

# PREVIOUS WORK

## Approaches using Electrostatic Information:

- Ohlendorf and Matthews (1985): the formation of protein-DNA complexes driven by the electrostatic interaction of asymmetrically distributed charges on the surface of the proteins complement the charges on DNA



# THE WORK BY SZILAGYI ET AL.

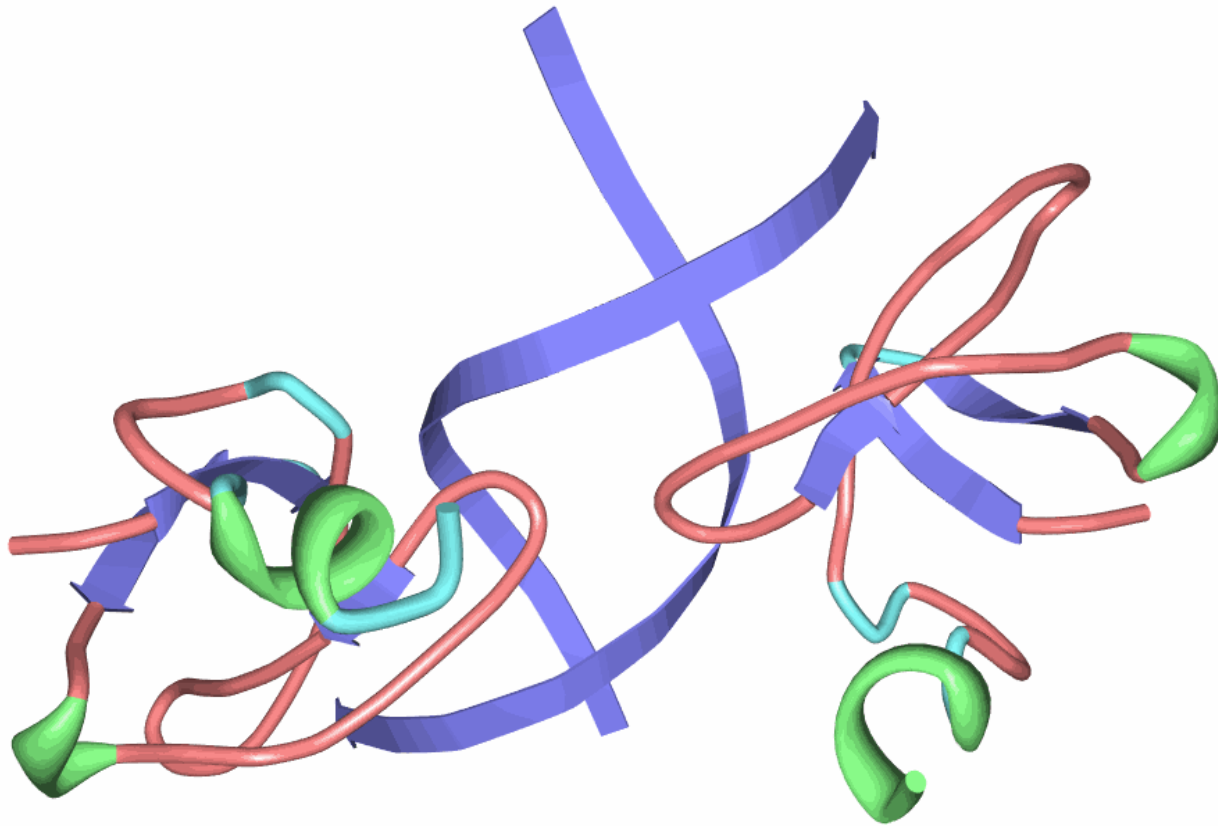
**Szilagyi A, Skolnick J, Efficient prediction of nucleic acid binding function from low-resolution protein structures. *J Mol Biol*, 2006)**

- A logistic regression classifier based on the amino acid composition, the asymmetry of the spatial distribution of specific amino acids and the dipole moment of the protein.
- Most important variables:
  - Arginine content (Positively charged amino acid)
  - Glycine content
  - Lysine content (Positively charged amino acid)
  - Dipole moment



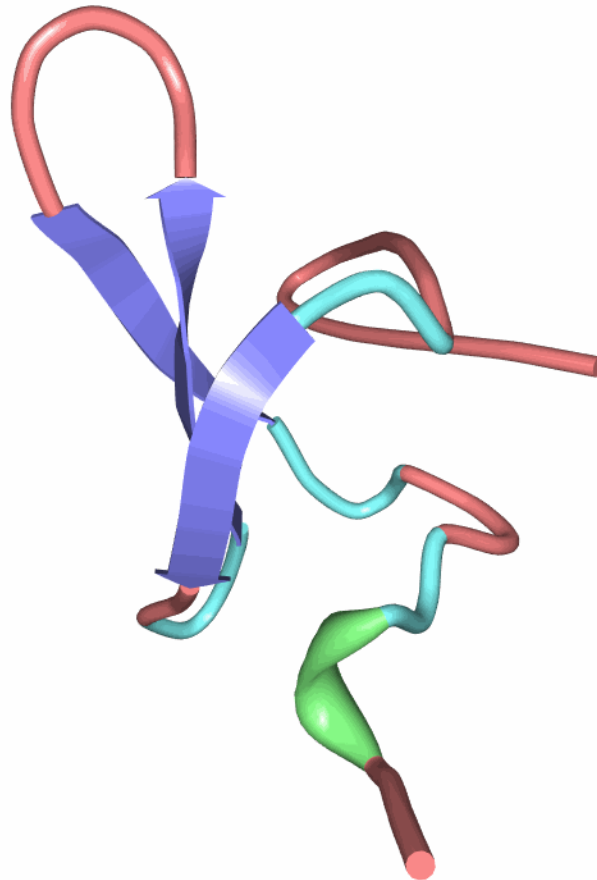
# AVAILABLE TRAINING DATA

DNA-binding Protein - Bounded conformation



# AVAILABLE TRAINING DATA

DNA-binding Protein - Unbounded conformation



# 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
```





# AN APPROACH BASED ON LOGICAL FEATURES



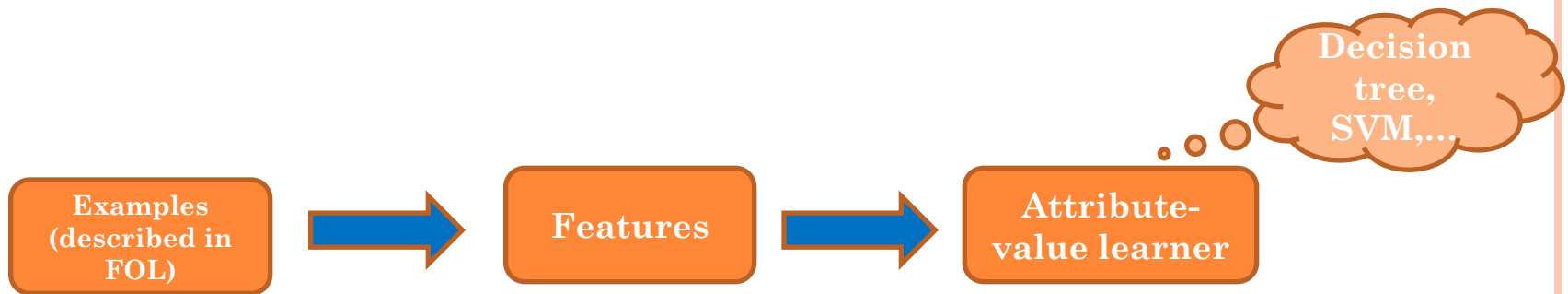
# FOL REPRESENTATION

- Proteins described by formal-logic assertions, e.g.:
  - $\text{res}('1\text{AYJ}', r1, 'CYS')$
  - $\text{dist}(r1, r2, 10)$
- 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
  - $F = \text{res}(P, R1, 'CYS'), \text{res}(P, R2, 'HIS'), \text{dist}(R1, R2, 8)$



# 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

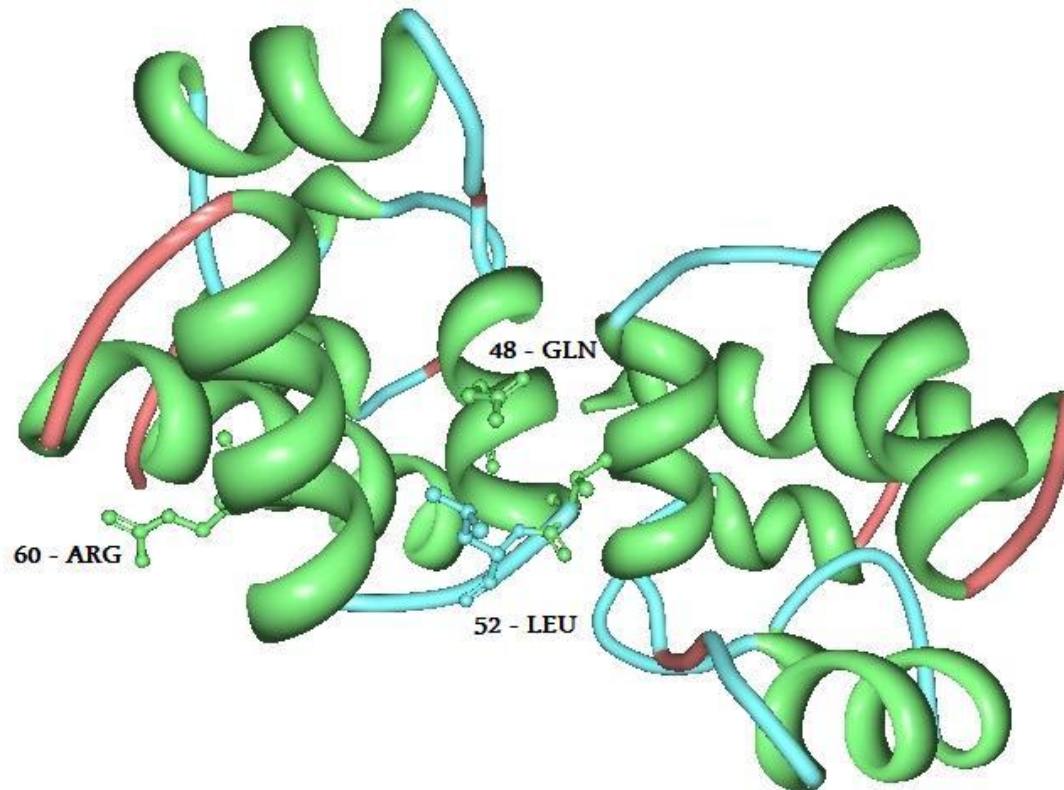
- We obtained about 1500 structural patterns for 54 unbounded DNA-binding proteins
- Accuracies obtained by stratified 10-fold crossvalidation using coarse-grained features (F1), our structural pattern features (F2) and combination of both of them (F1+2)

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)	<b>84.2 (1)</b>
SVM with radial basis kernel	81.6 (3)	67.1 (4-5)	67.1 (4-5)	83.0 (2)	<b>85.4 (1)</b>
Simple logistic regression	81.6 (3)	73.9 (5)	78.8 (4)	<b>87.6 (1)</b>	82.3 (2)
L <sub>2</sub> -regularized logistic regression	84.0 (2)	78.7 (5)	80.5 (4)	82.4 (3)	<b>84.2 (1)</b>
Ada-boost (with decision stamps)	77.4 (4)	73.2 (5)	83.0 (2)	79.3 (3)	<b>84.7 (1)</b>
Random forest	78.6 (4)	76.8 (5)	<b>83.6 (1)</b>	80.5 (2)	79.9 (3)
J48 decision tree	75.0 (3)	70.7 (4)	75.6 (2)	68.1 (5)	<b>76.2 (1)</b>
<b>Average ranking:</b>	<b>3</b>	<b>4.79</b>	<b>3.07</b>	<b>2.71</b>	<b>1.43</b>



# FEATURES GIVE US INSIGHT!

EXAMPLE OF A DISCOVERED STRUCTURAL FEATURE



- $F = \text{res}(P, R1, 'ARG'), \text{res}(P, R2, 'GLN'),$   
 $\text{res}(P, R3, 'LEU'), \text{dist}(R1, R2, 10.0), \text{dist}(R1, R3, 10.0)$



# AN APPROACH BASED ON BALL HISTOGRAMS





# THE BALL-HISTOGRAM METHOD

- The classification method consists of three main steps:



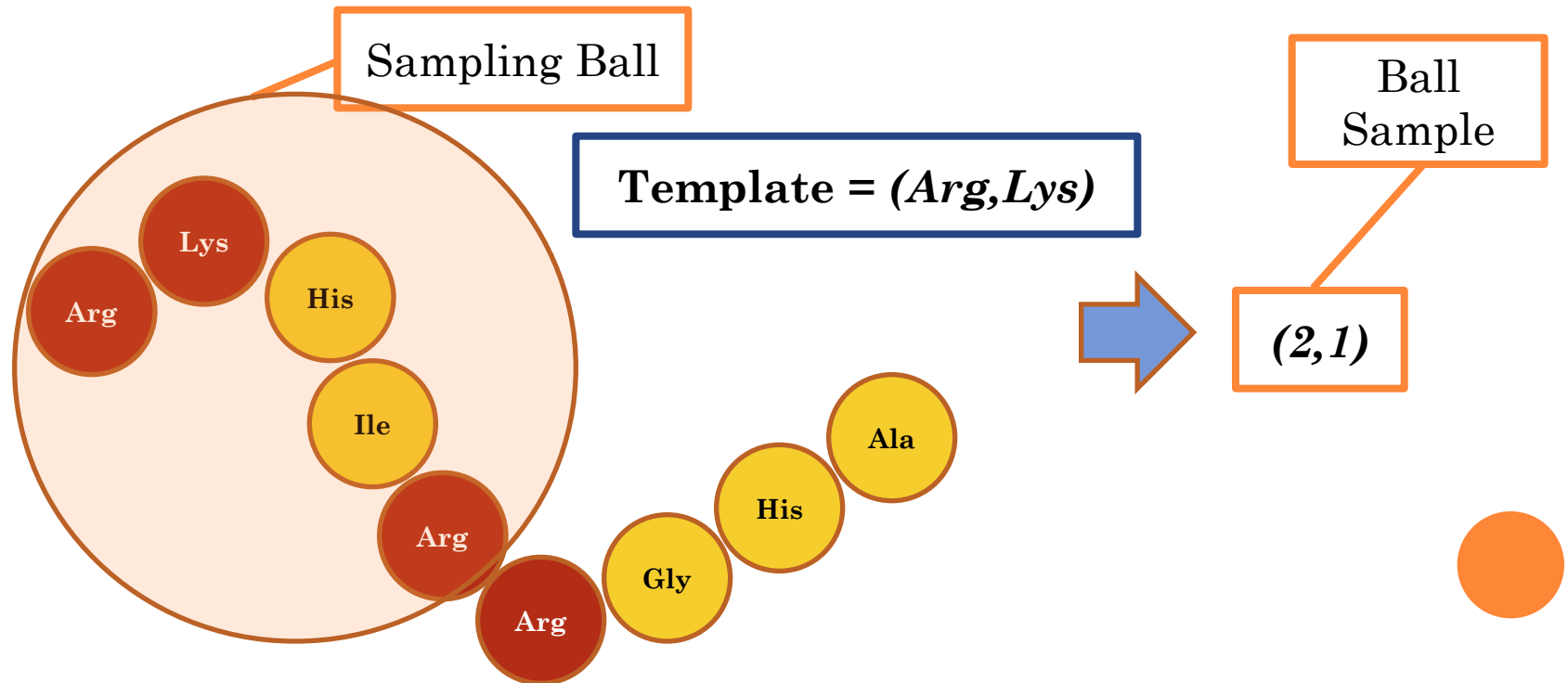
- Random forest classifier is used for the classification



# TEMPLATE AND BALL SAMPLE

**Template:** A list of amino-acid types or properties, for example (Arg, Lys)

**Ball Sample:** Place a ball (*sampling ball*) randomly on the protein and count the amino-acids in it which are listed in a given template – the resulting vector is a *ball sample*.

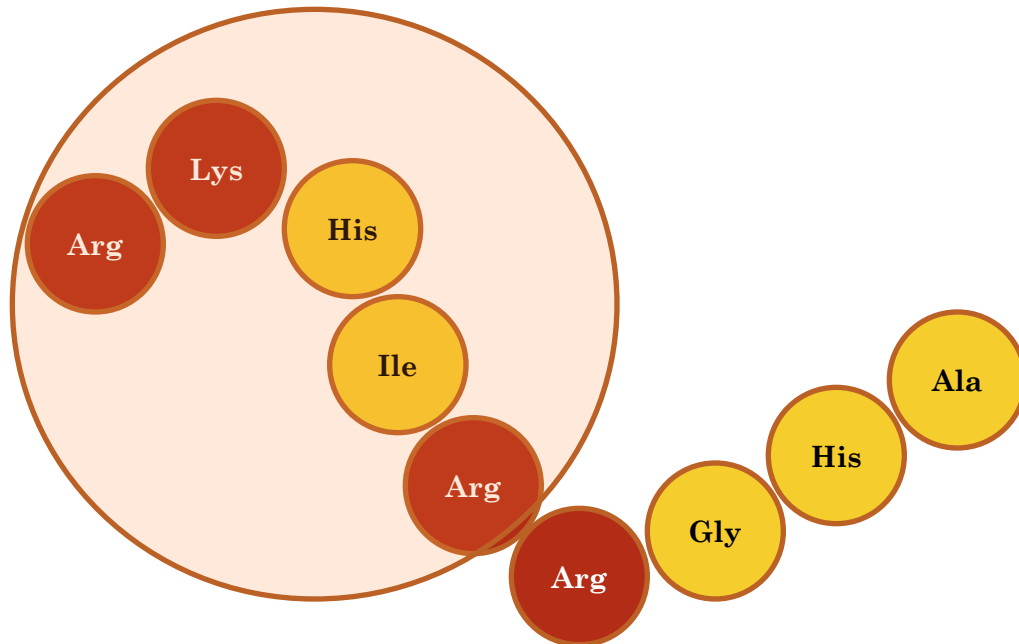


# BALL HISTOGRAM

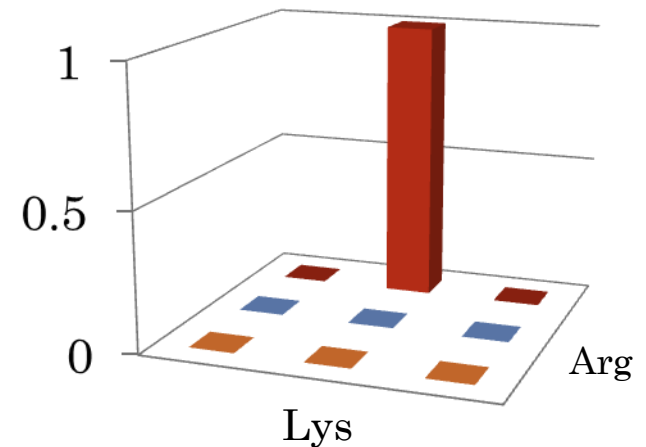
**Ball Histogram** is a multi-dimensional histogram constructed from ball samples randomly placed on a protein.

Template = (*Arg*, *Lys*)

Ball Sample = (2, 1)



Histogram:

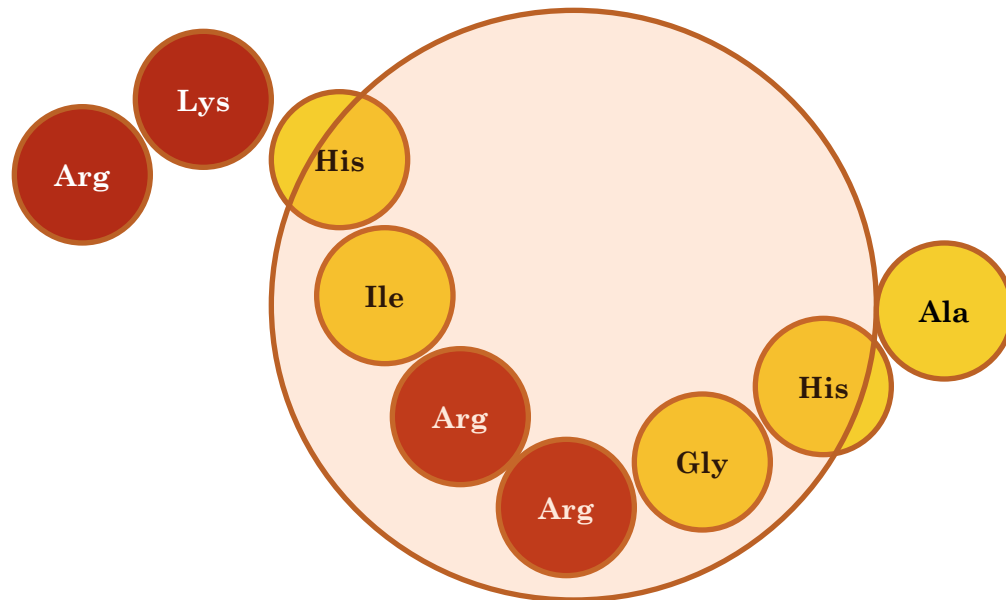


# BALL HISTOGRAM

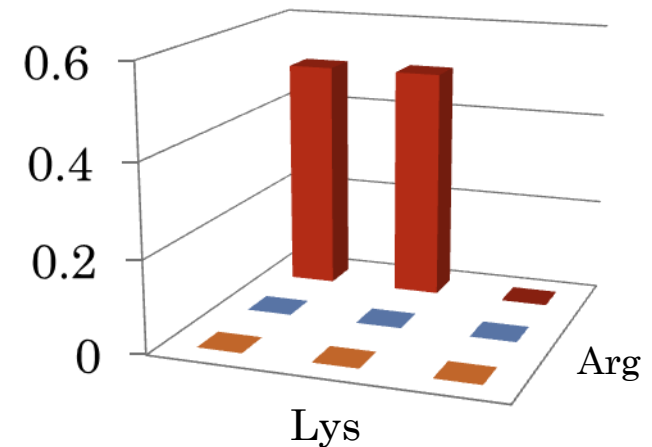
**Ball Histogram** is a multi-dimensional histogram constructed from ball samples randomly placed on a protein.

Template = (*Arg*, *Lys*)

Ball Sample = (2, 0)



Histogram:



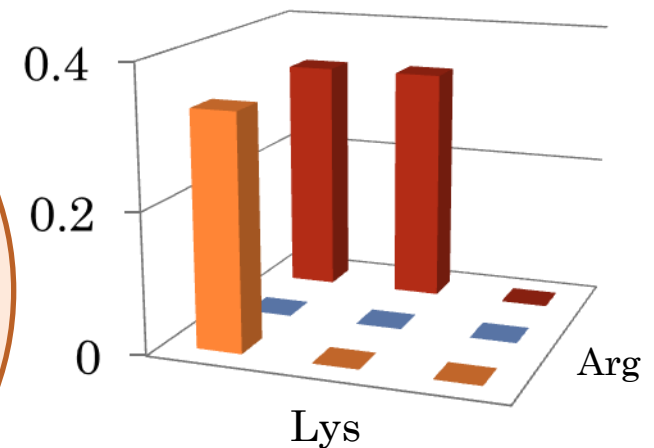
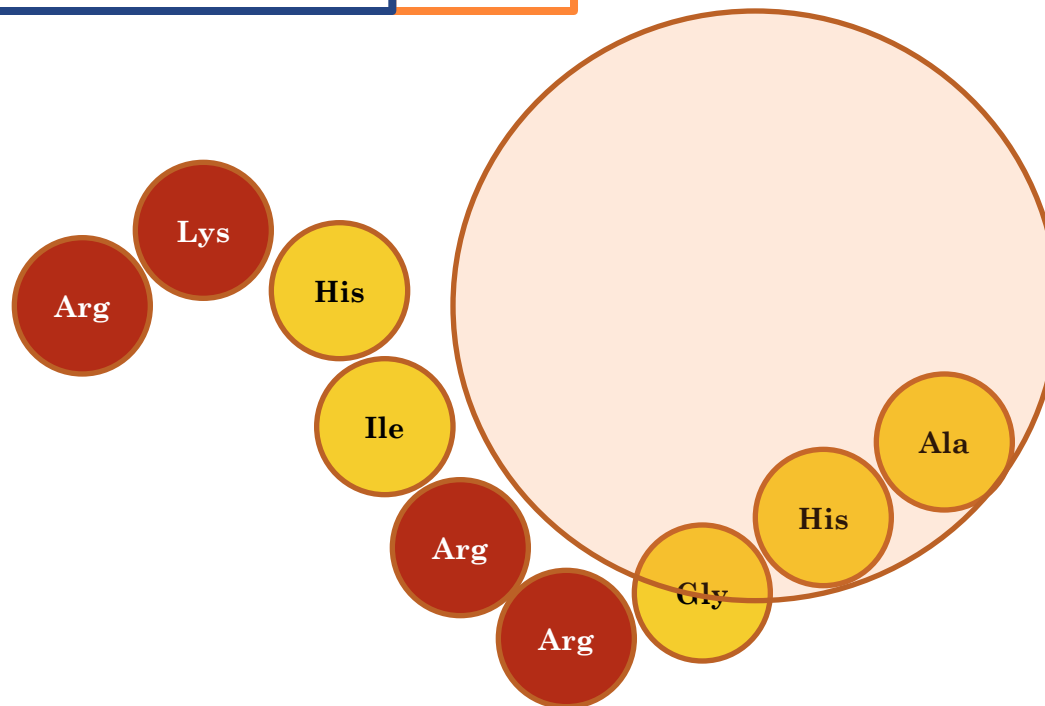
# BALL HISTOGRAM

**Ball Histogram** is a multi-dimensional histogram constructed from ball samples randomly placed on a protein.

Template = (*Arg*, *Lys*)

Ball Sample = (0, 0)

Histogram:

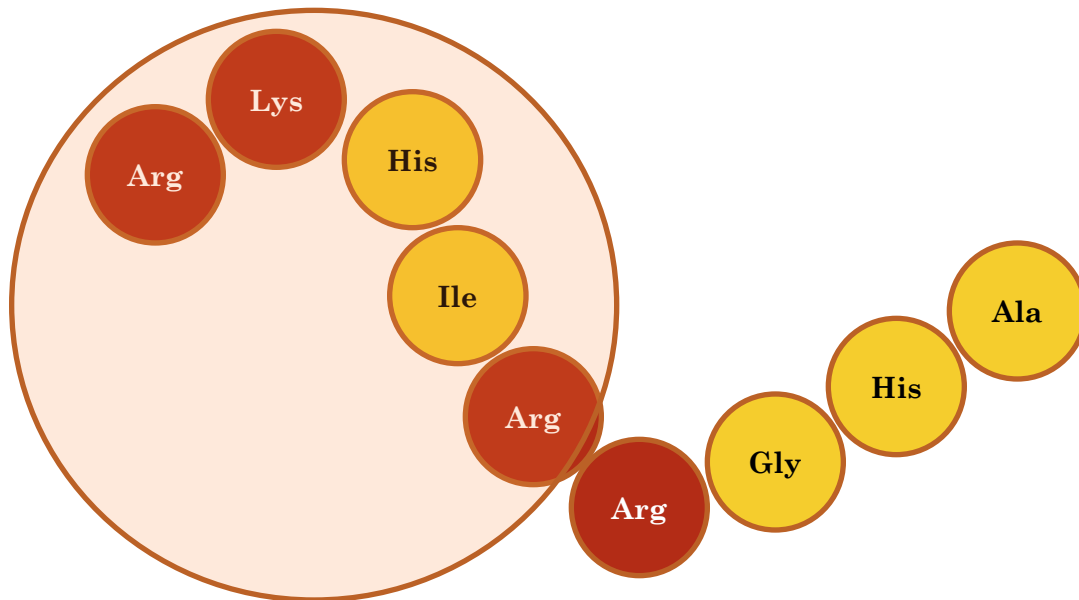


# BALL HISTOGRAM

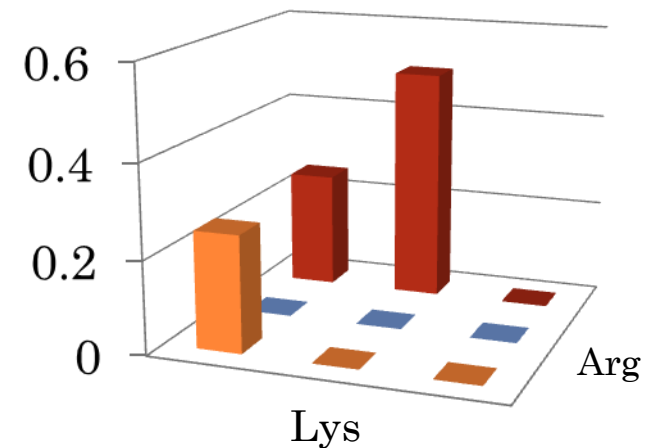
**Ball Histogram** is a multi-dimensional histogram constructed from ball samples randomly placed on a protein.

Template = (*Arg*, *Lys*)

Ball Sample = (2, 1)



Histogram:



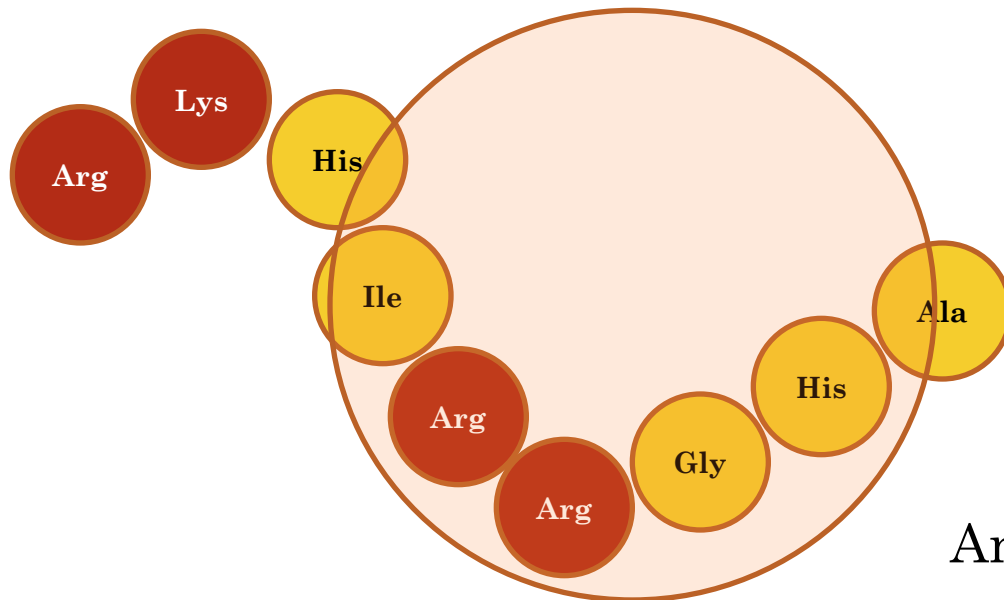
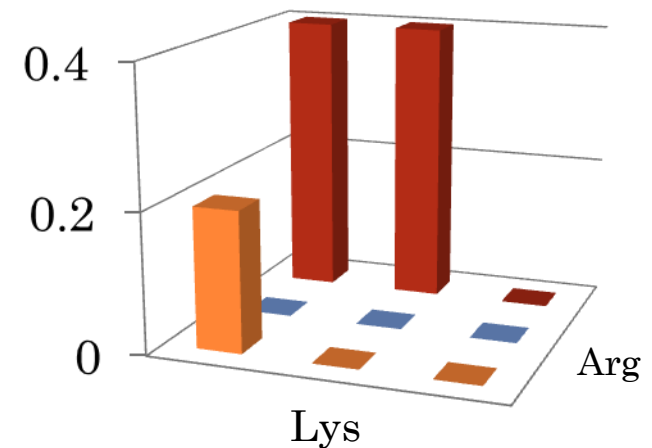
# BALL HISTOGRAM

**Ball Histogram** is a multi-dimensional histogram constructed from ball samples randomly placed on a protein.

Template = (*Arg*, *Lys*)

Ball Sample = (2, 0)

Histogram:



And so on until convergence...

# BALL HISTOGRAM PROPERTIES

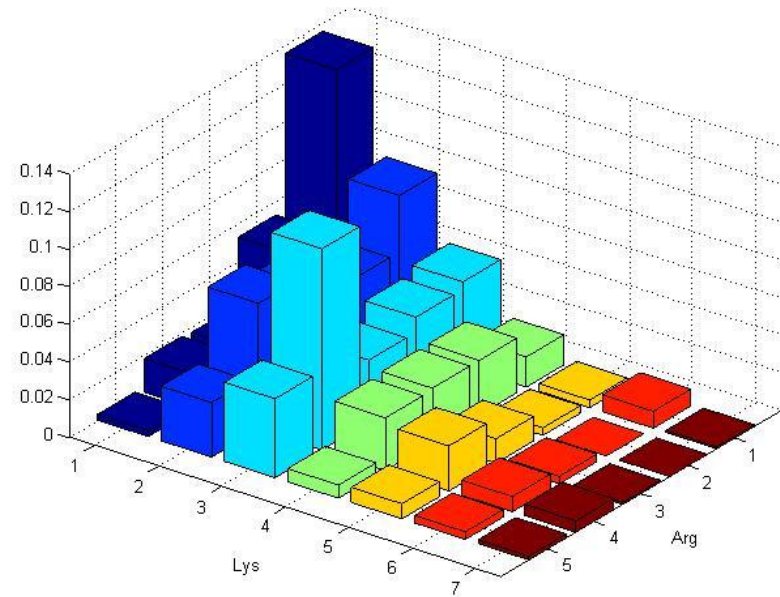
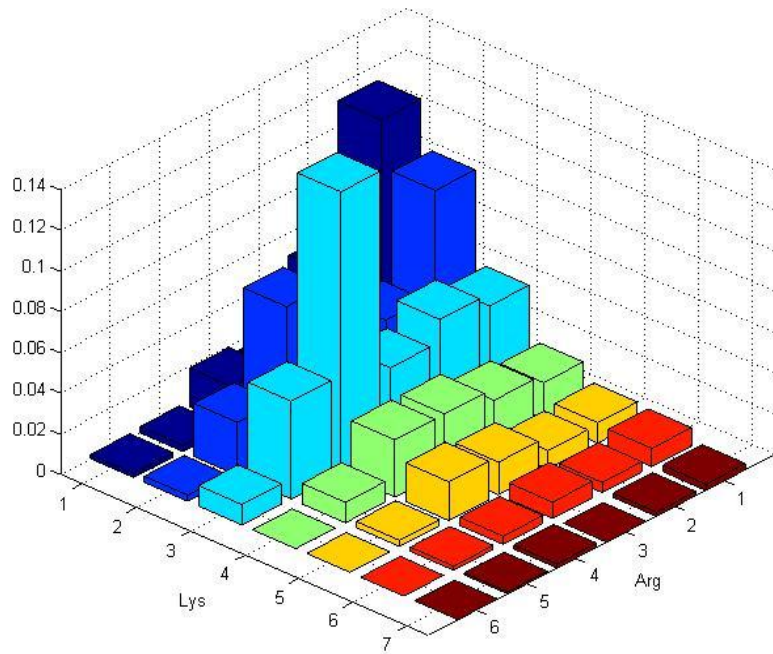
Intuitively, a ball histogram captures the joint probability that certain numbers of amino acids listed in a template will appear in a randomly selected ball region

1. Ball histogram is invariant with respect to rotation and translation of the protein  
(this is achieved by appropriately choosing the space from which the samples are drawn – so *called sampling sphere*)
2. Sampling balls which contain no amino acids are ignored





# ACTUAL BALL HISTOGRAM FOR PROTEINS 1A31 & 1A3Q



# PREDICTIVE CLASSIFICATION USING BALL HISTOGRAMS

1. For each protein in the dataset we can construct a ball histogram
2. We convert the histograms to attribute-value form readable by Weka – each bin in a histogram becomes an attribute
3. We apply the standard machine learning *random forest* algorithm

(Number of trees and the optimal sampling-ball radius is selected using internal cross-validation)



# RESULTS

- Two types of experiments:
  - Ball Histograms using Charged Amino Acids  
(*Arg, Lys, Glu, Asp*)
  - Ball Histograms using the Second Set of Amino Acids  
(*Arg, Lys, Gly, Asp, Asn, Ser, Ala*)
- Accuracies estimated by 10-fold cross-validation on PD138-NB110:

Method	Accuracy [%]
Szilágyi et al.	81,4
Ball Histogram using Charged Amino Acids	80,2
Ball Histogram using the Second Set of Amino Acids	84,7



# RESULTS

- The three most informative *features* according to the  $\chi^2$  criterion - charged amino acids:

	Arg	Lys	Glu	Asp
1 <sup>st</sup> feature	1	1	0	0
2 <sup>nd</sup> feature	2	0	0	0
3 <sup>rd</sup> feature	1	0	0	0

- The three most informative *features* according to the  $\chi^2$  criterion – second set of amino acids:

	Arg	Lys	Gly	Asp	Asn	Ser	Ala
1 <sup>st</sup> feature	1	1	0	0	0	0	0
2 <sup>nd</sup> feature	1	0	0	0	0	0	0
3 <sup>rd</sup> feature	2	0	0	0	0	0	0



# CONCLUSIONS

- We contribute a novel histogram-based method to predict the DNA-binding propensity of the proteins
- We base our approach on a systematic exploration of spatial distribution of selected amino acids
- Our *Ball Histogram* method improves on the state-of-the-art approaches



# FEATURES GIVE US INSIGHT!

## EXAMPLE OF A DISCOVERED STRUCTURAL FEATURE

- The most informative feature according to the  $\chi^2$  criterion:  
(Arg = 1, Lys = 1, Glu = 0, Asp = 0)

