Protein Fold Recognition with Combined SVM-RDA Classifier

Wiesław Chmielnicki[1]
Katarzyna Stąpor[2]

[1]Jagiellonian University, Kraków, Poland
[2]Silesian University of Technology, Gliwice, Poland

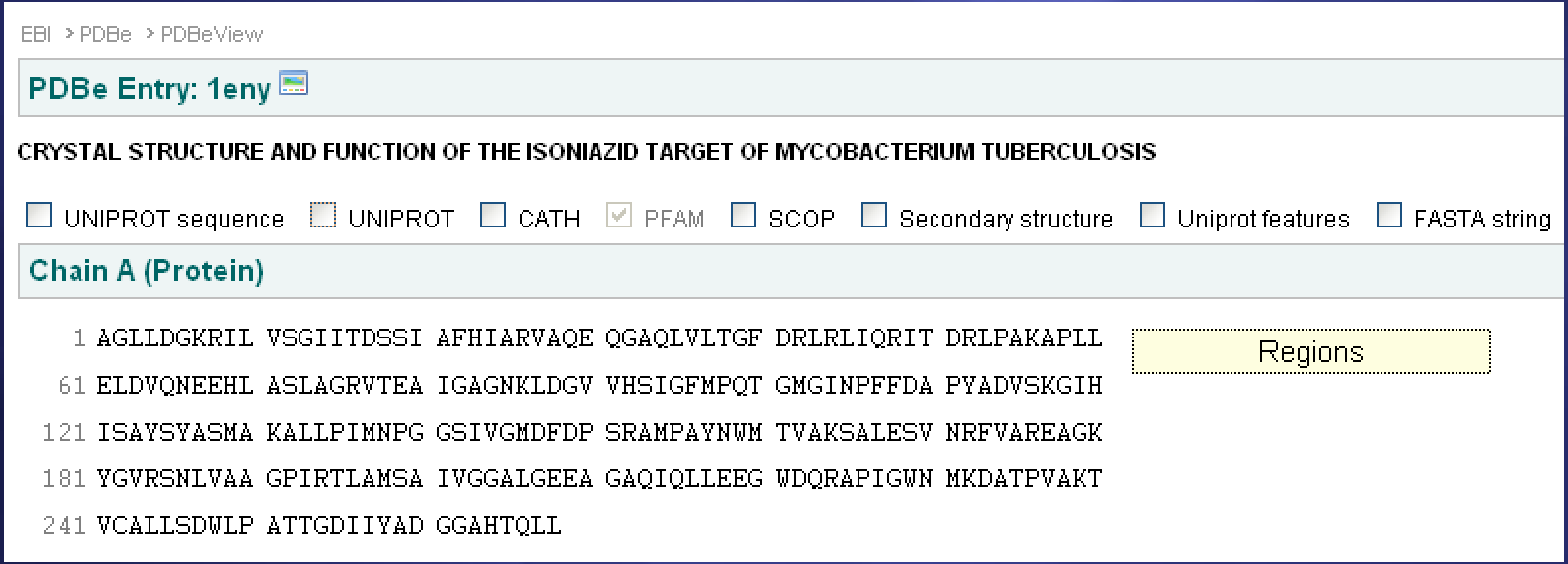
HAIS 2010, San Sebastian, 23rd - 25th June 2010

Outline

- Protein structure
- Methods of protein structure predition
- The database and the feature vectors
- First approach: an RDA classifier
- Second approach: an SVM classifier
- A binary and a multi-class problems
- The proposed hybrid SVM-RDA classifier
- Results and conclusions

Protein structure

- Primary protein structure – the sequence of amino acid residues



EBI > PDBe > PDBeView

PDBe Entry: 1eny

CRYSTAL STRUCTURE AND FUNCTION OF THE ISONIAZID TARGET OF MYCOBACTERIUM TUBERCULOSIS

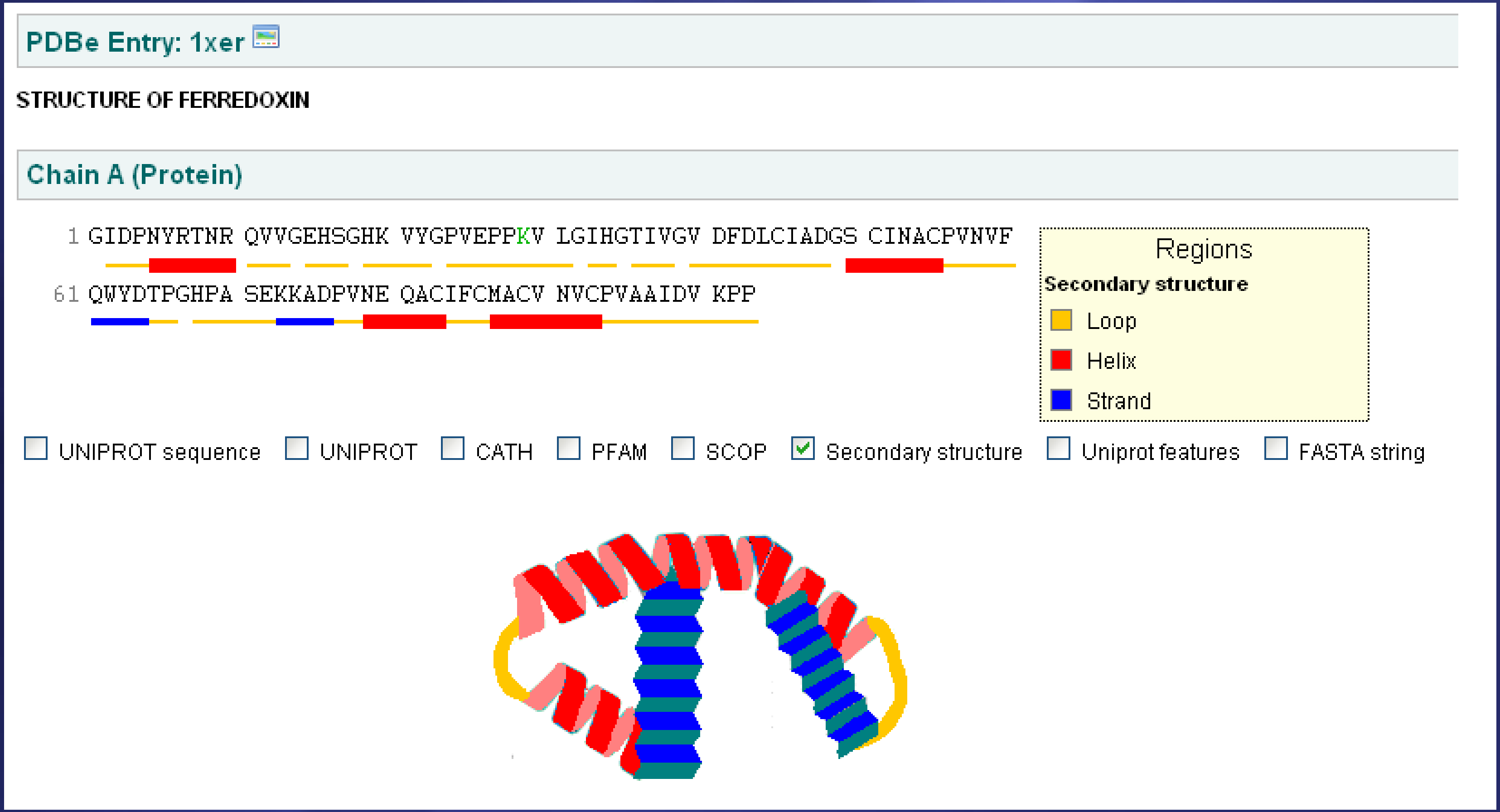
☐ UNIPROT sequence ☐ UNIPROT ☐ CATH ☑ PFAM ☐ SCOP ☐ Secondary structure ☐ Uniprot features ☐ FASTA string

Chain A (Protein)

```
  1 AGLLDGKRIL VSGIITDSSI AFHIARVAQE QGAQLVLTGF DRLRLIQRIT DRLPAKAPLL
 61 ELDVQNEEHL ASLAGRVTEA IGAGNKLDGV VHSIGFMPQT GMGINPFFDA PYADVSKGIH
121 ISAYSYASMA KALLPIMNPG GSIVGMDFDP SRAMPAYNWM TVAKSALESV NRFVAREAGK
181 YGVRSNLVAA GPIRTLAMSA IVGGALGEEA GAQIQLLEEG WDQRAPIGWN MKDATPVAKT
241 VCALLSDWLP ATTGDIIYAD GGAHTQLL
```

Protein structure

- Secondary protein structure



PDBe Entry: 1xer

STRUCTURE OF FERREDOXIN

Chain A (Protein)

```
 1 GIDPNYRTNR QVVGEHSGHK VYGPVEPPKV LGIHGTIVGV DFDLCIADGS CINACPVNVF
61 QWYDTPGHPA SEKKADPVNE QACIFCMACV NVCPVAAIDV KPP
```

Regions

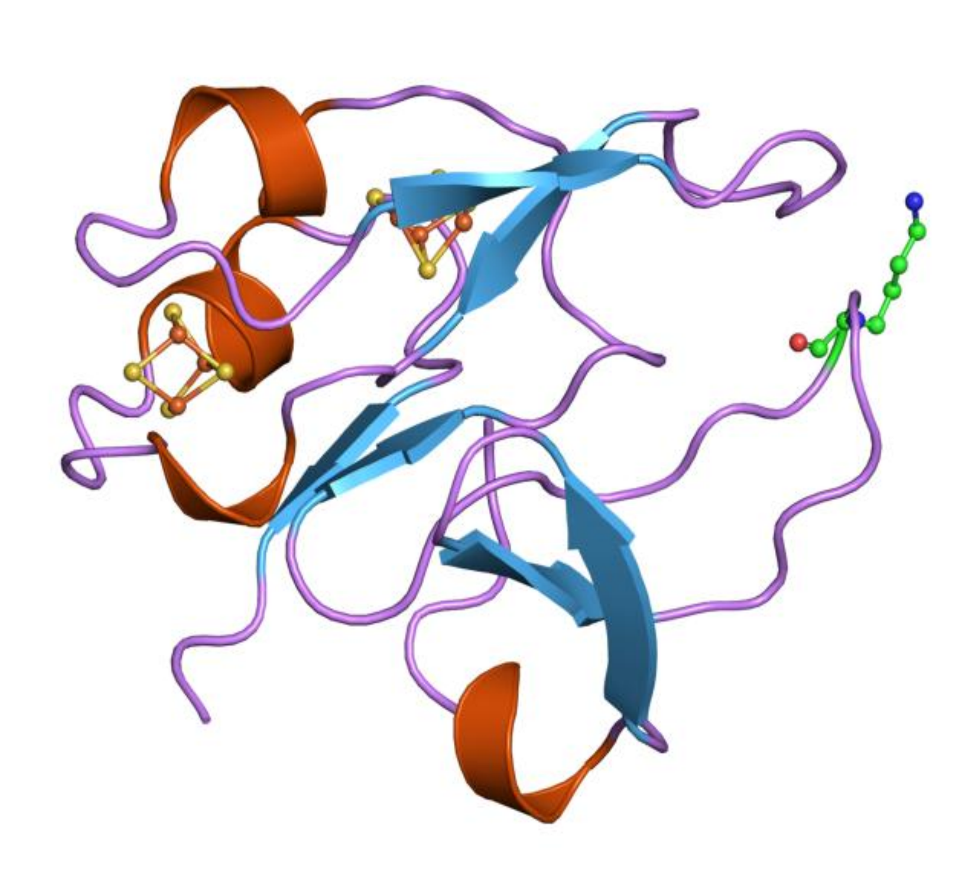
Secondary structure

- 🟨 Loop
- 🟥 Helix
- 🟦 Strand

□ UNIPROT sequence □ UNIPROT □ CATH □ PFAM □ SCOP ☑ Secondary structure □ Uniprot features □ FASTA string

Protein structure

- Tertiary (3D) protein structure - protein fold

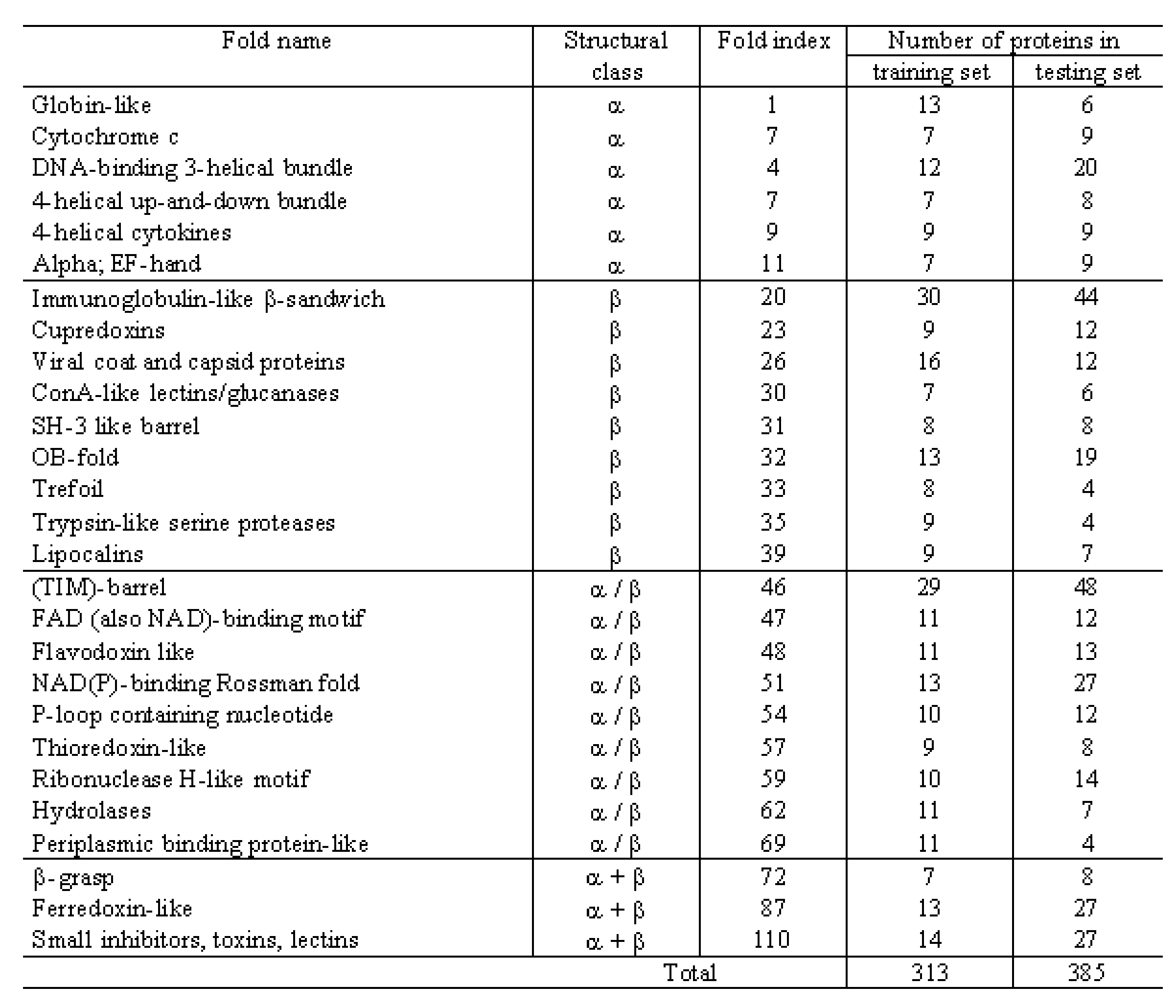


Methods of protein fold prediction

- Ab initio protein modelling
  - Based on physical principles
- Comparative protein modelling
- Side chain geometry prediction
- Statistical methods
  - Based on amino acid composition
  - And other protein parameters
- The recognition ratios varied from 50 to 60 percent

The database

- Training set and testing set



Fold name	Structural class	Fold index	Number of proteins in	
			training set	testing set
Globin-like	α	1	13	6
Cytochrome c	α	7	7	9
DNA-binding 3-helical bundle	α	4	12	20
4-helical up-and-down bundle	α	7	7	8
4-helical cytokines	α	9	9	9
Alpha; EF-hand	α	11	7	9
Immunoglobulin-like β-sandwich	β	20	30	44
Cupredoxins	β	23	9	12
Viral coat and capsid proteins	β	26	16	12
ConA-like lectins/glucanases	β	30	7	6
SH-3 like barrel	β	31	8	8
OB-fold	β	32	13	19
Trefoil	β	33	8	4
Trypsin-like serine proteases	β	35	9	4
Lipocalins	β	39	9	7
(TIM)-barrel	α / β	46	29	48
FAD (also NAD)-binding motif	α / β	47	11	12
Flavodoxin like	α / β	48	11	13
NAD(P)-binding Rossman fold	α / β	51	13	27
P-loop containing nucleotide	α / β	54	10	12
Thioredoxin-like	α / β	57	9	8
Ribonuclease H-like motif	α / β	59	10	14
Hydrolases	α / β	62	11	7
Periplasmic binding protein-like	α / β	69	11	4
β-grasp	α + β	72	7	8
Ferredoxin-like	α + β	87	13	27
Small inhibitors, toxins, lectins	α + β	110	14	27
	Total		313	385

The feature vectors

- The feature vectors are based on six parameters
  - Amino acids composition
  - Predicted secondary structure
  - Hydrophobity
  - Normalized Van der Walls volume
  - Polarity
  - Polarizability
- The detailed description can be found in Ding and Dubchak papers

An RDA classifier

- Quadratic Discriminant Analysis
  - Discriminant function

$$d_k(\mathbf{X}) = (\mathbf{X} - \mu_k)^T \Sigma_k^{-1} (\mathbf{X} - \mu_k) + \log|\Sigma_k| - 2\log\pi(k)$$

  - Estimates

$$\hat{\mu}_k = \overline{X}_k = \frac{1}{N_k}\begin{bmatrix} \sum_{i=1}^{N} X_{n1} \\ \vdots \\ \sum_{i=1}^{N} X_{np} \end{bmatrix} = \begin{bmatrix} \overline{x}_1 \\ \vdots \\ \overline{x}_p \end{bmatrix}$$

$$\hat{\Sigma}_k = \frac{S_k}{N_k} = \frac{1}{N_k}\sum_{c(v)=k} (X - \overline{X}_k)(X - \overline{X}_k)^T$$

An RDA classifier

- Covariance matrix regularization
  - Let's replace the individual class covariance matrices by their average

$$\hat{\Sigma} = \frac{\sum_{k=1}^{K} S_k}{\sum_{k=1}^{K} N_k}$$

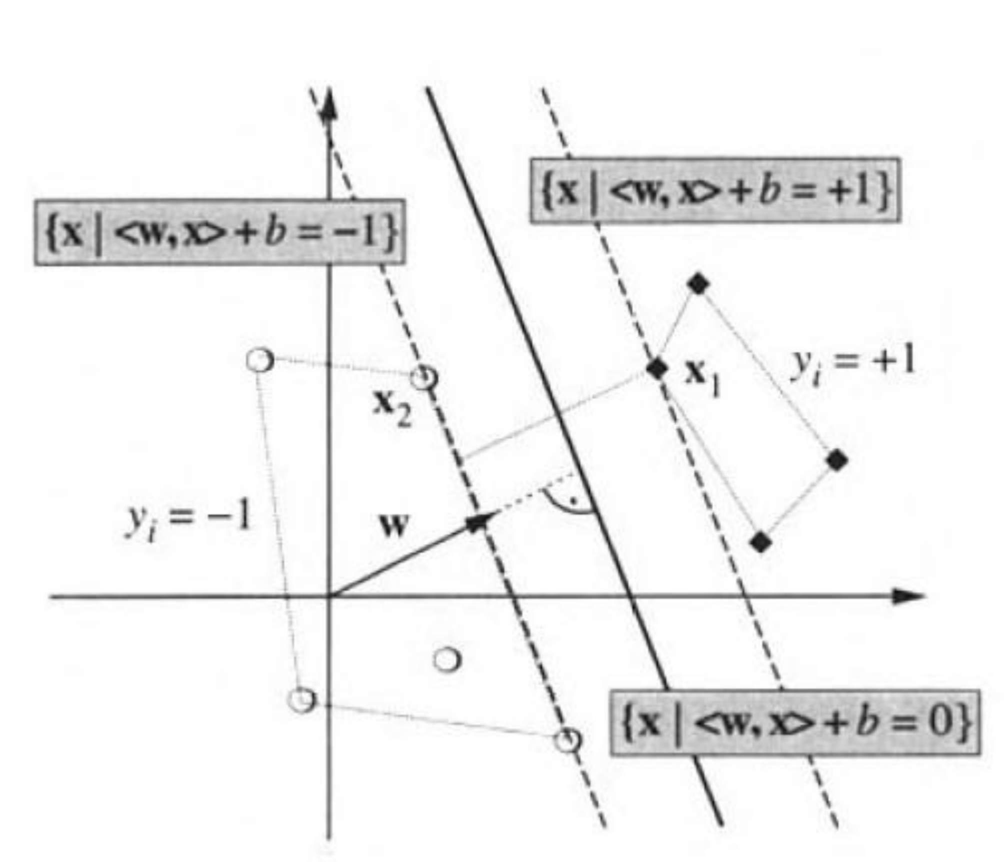
  - A less limited approach

$$\hat{\Sigma}_k(\lambda) = (1-\lambda)\hat{\Sigma}_k + \lambda\hat{\Sigma}$$

- The recognition ratio is 55.6%

An SVM classifier

- Maximun-margin hyperplane



An SVM classifier

- Discriminant function

$$f(x) = sign\Big(\sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b\Big) \; ,$$

- where $0 \leq \alpha_i \leq C, i = 1, 2, \ldots, N$

- The RBF kernel

$$K(x_i, x) = -\gamma ||x - x_i||^2, \gamma > 0$$

An SVM classifier

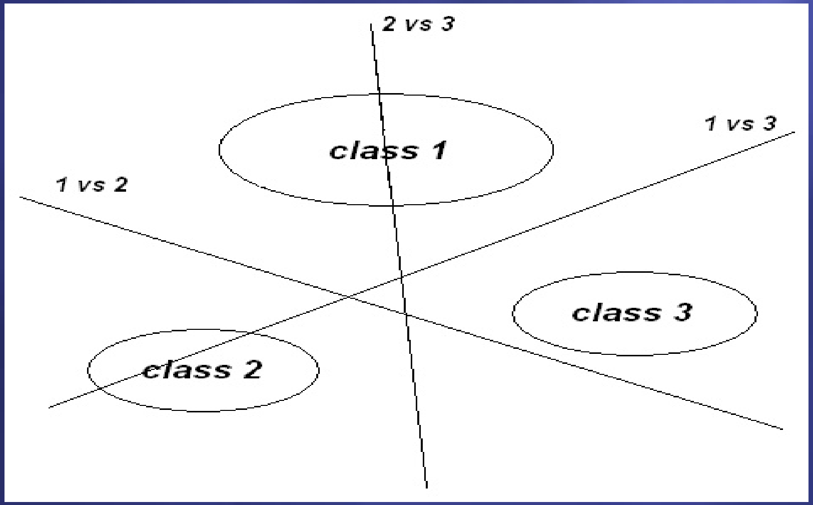
- Advantages of an SVM
  - Maximization of generalization ability
  - No local minima
  - Robustness to outliers
- Disadvantages of an SVM
  - Long training time
  - The selection of a kernel parameters
  - It is a binary classifier

An SVM classifier

- Extension to the a multiclass problem
  - We can consider all classes in one optimization
  - Or cover one n-class problem with several binary problems
- The approach with binary problems
  - One-versus-others strategy
  - One-versus-one strategy
  - Others: DAG, ADAG, BDT, DB2, pairwise coupling
- The recognition ratio is 58.7%

Combined SVM-RDA classifier

- The reliability of the binary classifiers



Combined SVM-RDA classifier

- Discriminant function of an RDA classifier

$$d_k(\mathbf{X}) = (\mathbf{X} - \mu_k)^T \Sigma_k^{-1} (\mathbf{X} - \mu_k) + \log|\Sigma_k| - 2\log\pi(k)$$

- Let's define

$$d_{min}(x) = \min\{d_k(x)\},\ k = 1, 2, \ldots, n$$

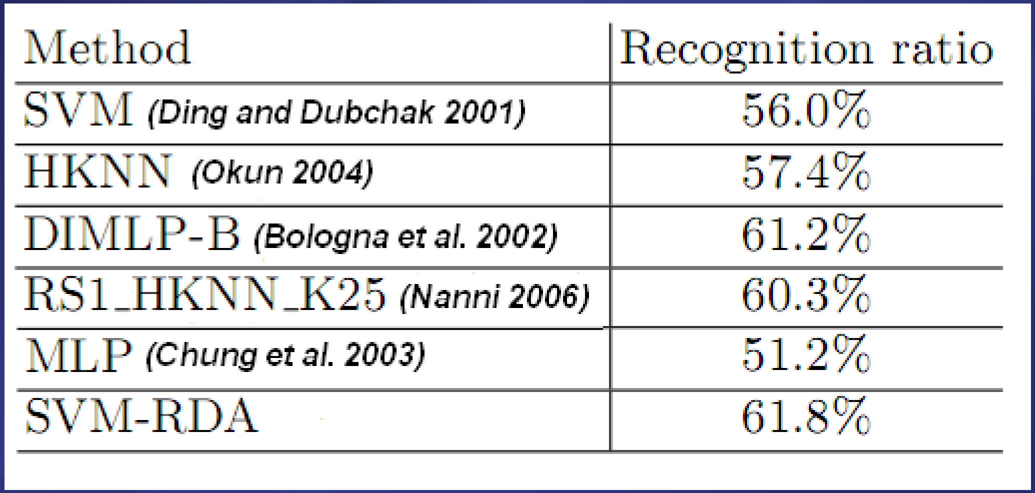
- Then, for every binary classifier

$$1 - \frac{d_i(x) - d_{min}(x)}{d_{min}(x)}$$

- Now, the value defined above will be a weight of the vote of the binary classifier

Combined SVM-RDA classifier

- Results
  - RDA classifier – 55,6%
  - SVM classifier– 58,7%
  - Combined SVM-RDA classifier – 61,8%
- Comparison with other methods



Method	Recognition ratio
SVM *(Ding and Dubchak 2001)*	56.0%
HKNN *(Okun 2004)*	57.4%
DIMLP-B *(Bologna et al. 2002)*	61.2%
RS1_HKNN_K25 *(Nanni 2006)*	60.3%
MLP *(Chung et al. 2003)*	51.2%
SVM-RDA	61.8%

Thank you