Topic

Prediction of the burial status of transmembrane residues of helical membrane proteins
(with support vectors)

by Thorsten Will

Overview

Introduction

- helical membrane proteins
- needed definitions: burial status ,(r)SASA
- a two step architecture: the TMX method

Statistical methods used

- prediction:
    - basic principles of a SVM
    - Support Vector Regression in detail

- assessment:
    - measuring regression performance
    - cross-validation

The practical part

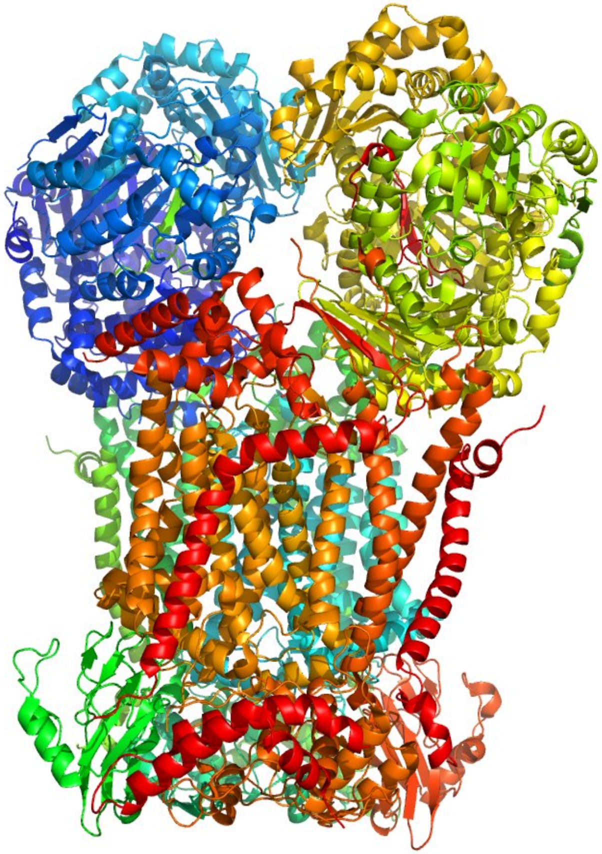
- the dataset and the problem
- workflow
- results and discussion

Introduction
Why helical membrane proteins are important

Facts:

- crucial role in fundamental cellular processes

- account for 20-30% of the ORFs of sequenced genomes



Cytochrome bc1 complex
(respiratory chain) (from PDB: 1PP9)

Introduction
Why helical membrane proteins are interesting

Facts:

- crucial role in fundamental cellular processes → structure determination desirable

so far very difficult with current experimental techniques

- account for 20-30% of the ORFs of sequenced genomes

Less than 1% of the proteins with known structure are HMPs !

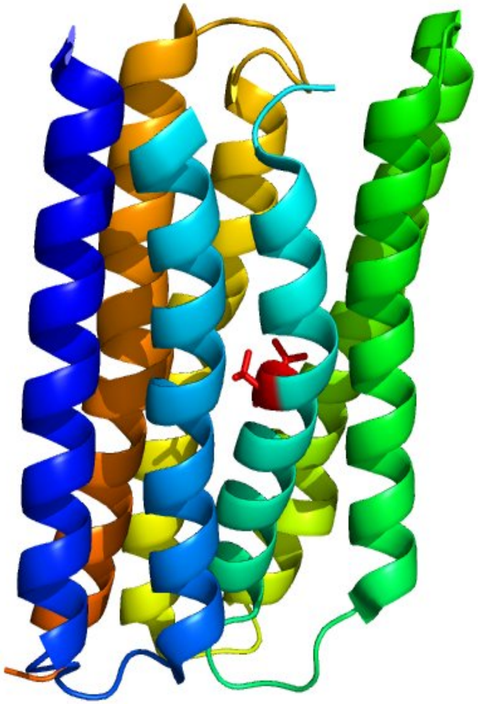
sequence-based predictors invaluable!

Introduction

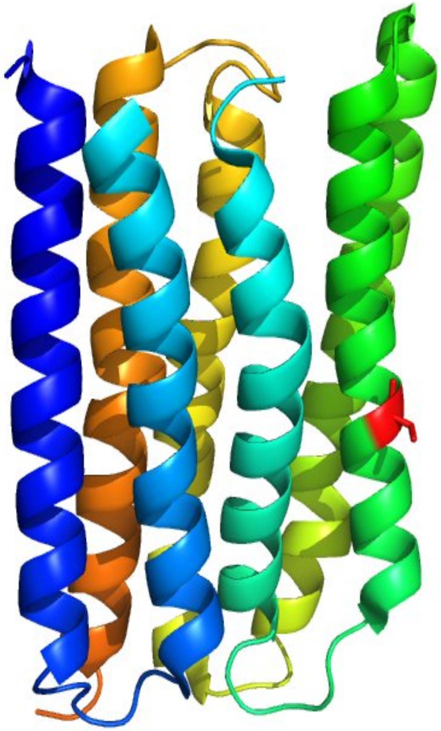
What is the burial status and why is it useful

In the case of membrane proteins:

Buried in the protein core vs **exposed** to the membrane



two burried threonine



an exposed serine

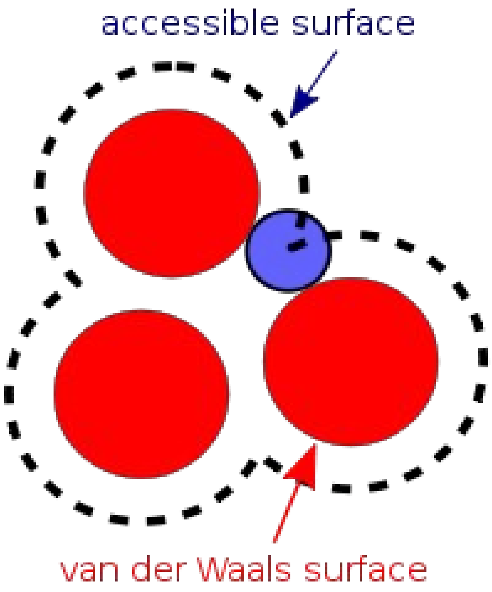
Example:

a bacterial rhodopsin from the dataset (from PDB: 1xio)

Introduction
Definition of the rSASA

(S)ASA: (Solvent) Accessible Surface Area

- the accessible area of the surface for a solvent of a specific size



from Wikipedia:
http://en.wikipedia.org/wiki/Accessible_surface_area

rSASA: relative SASA

- normalized measurement for the SASA:
  → dividation by reference values

- rSASA = 0.00 → residue defined as burried

Introduction
The TMX method

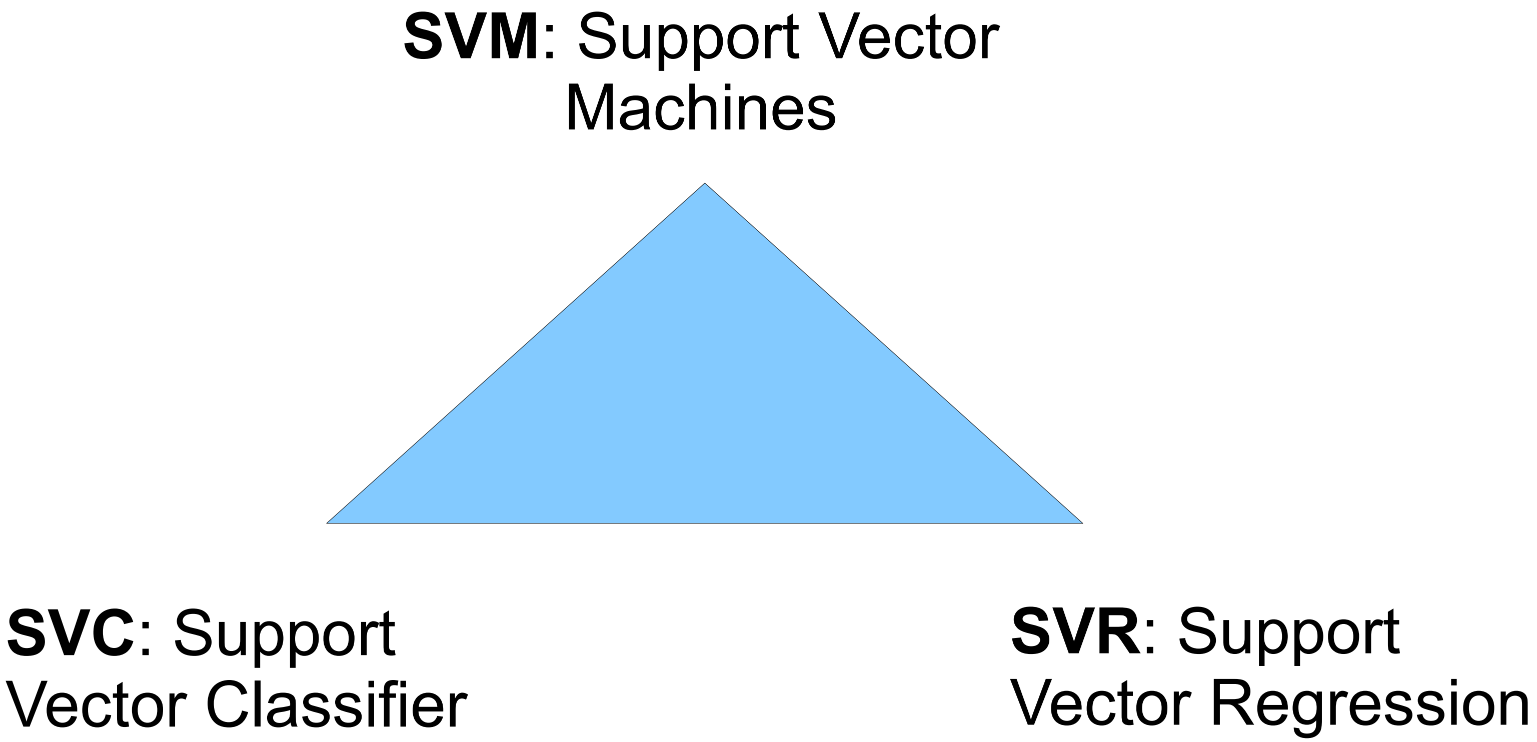
few words about...

TMX: TransMembrane eXposure

A two step approach

1. positional score
2. classification

Introduction
SVM in general



SVM: Support Vector Machines

SVC: Support Vector Classifier

SVR: Support Vector Regression

Statistical methods used

SVM basics: the formal problem simplified

SVC: classification

SVR: regression

$$\{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathcal{X} \times \{-1, +1\}$$

Input / training data

$$\{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathcal{X} \times \mathbb{R}$$

Statistical methods used

SVM basics: the formal problem simplified

SVC: classification

$\{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathcal{X} \times \{-1, +1\}$

Input / training data

SVR: regression

$\{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathcal{X} \times \mathbb{R}$

Search for a function $f(x_i) \approx y_i$ for “many“ i. → construct specific **hyperplane H**

H does best possible class separation

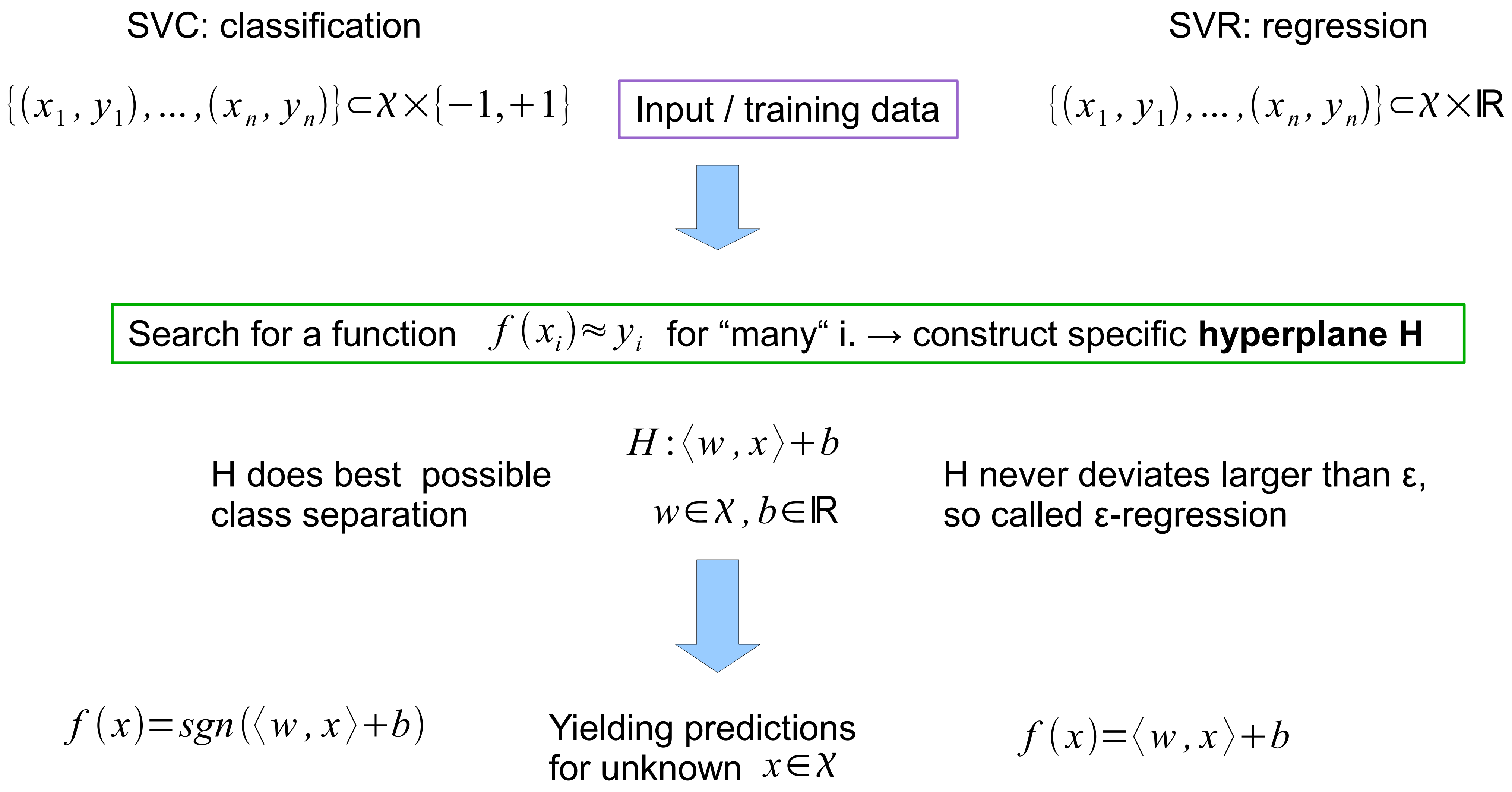
$$H : \langle w, x \rangle + b$$

$$w \in \mathcal{X}, b \in \mathbb{R}$$

H never deviates larger than ε, so called ε-regression

Statistical methods used

SVM basics: the formal problem simplified



SVC: classification

SVR: regression

$\{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathcal{X} \times \{-1, +1\}$

Input / training data

$\{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathcal{X} \times \mathbb{R}$

Search for a function $f(x_i) \approx y_i$ for “many“ i. → construct specific **hyperplane H**

H does best possible class separation

$H : \langle w, x \rangle + b$

$w \in \mathcal{X}, b \in \mathbb{R}$

H never deviates larger than ε, so called ε-regression

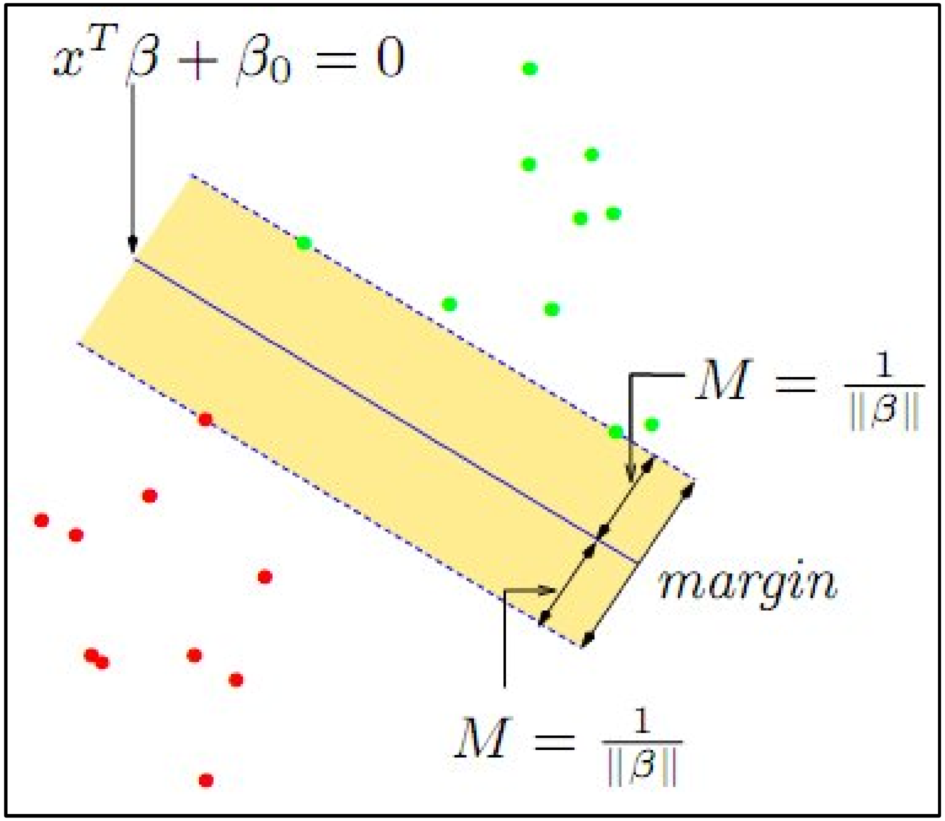
$f(x) = sgn(\langle w, x \rangle + b)$

Yielding predictions for unknown $x \in \mathcal{X}$

$f(x) = \langle w, x \rangle + b$

Statistical methods used
SVM basics: optimal hyperplanes

How to choose the **best** hyperplane



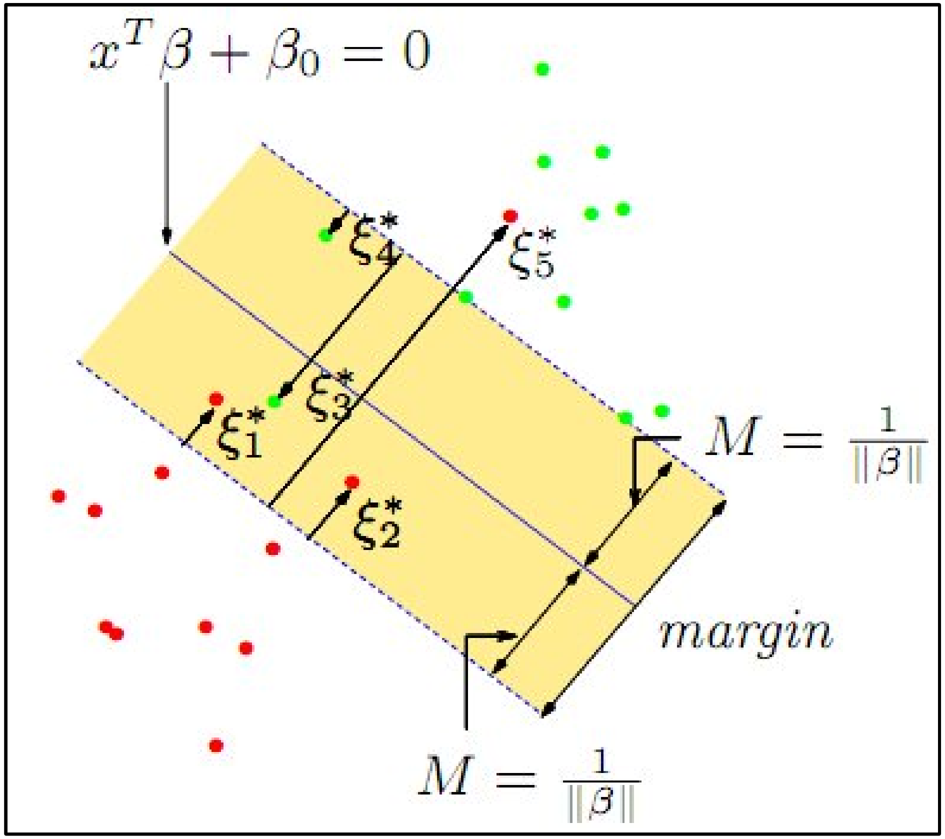
from “The elements of statistical learning“

Choose the hyperplane that **maximizes the margin**

→ minimize ||β||

Statistical methods used
SVM basics: introducing slack variables

Getting around with **non-feasible** problems



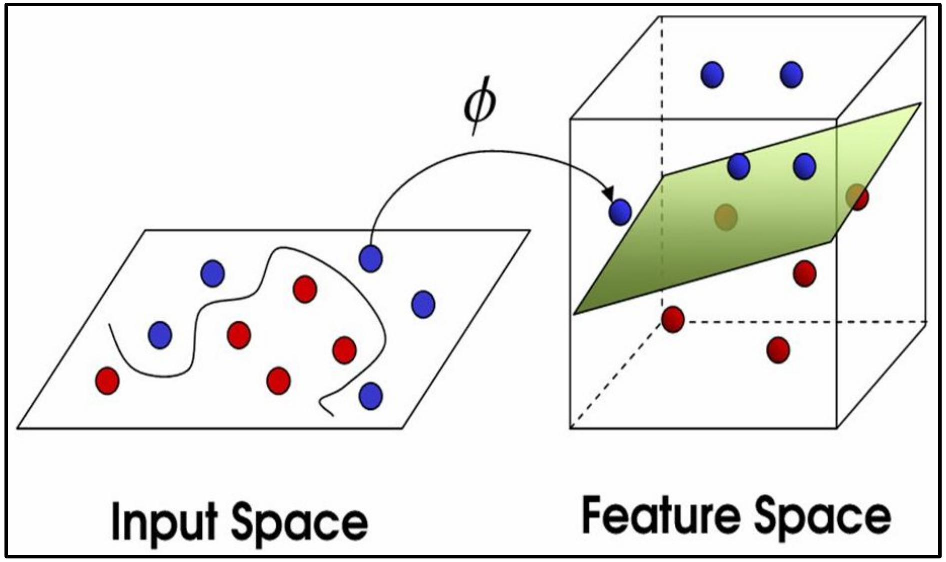
from “The elements of statistical learning“

Allow violations of the margin constraint but minimize the extent of the violations

Statistical methods used
SVM basics: the kernel-trick

Getting **non-linear**

Every dot product is replaced by a **non-linear kernel function**
→ input space transported into a high-dimensional feature space

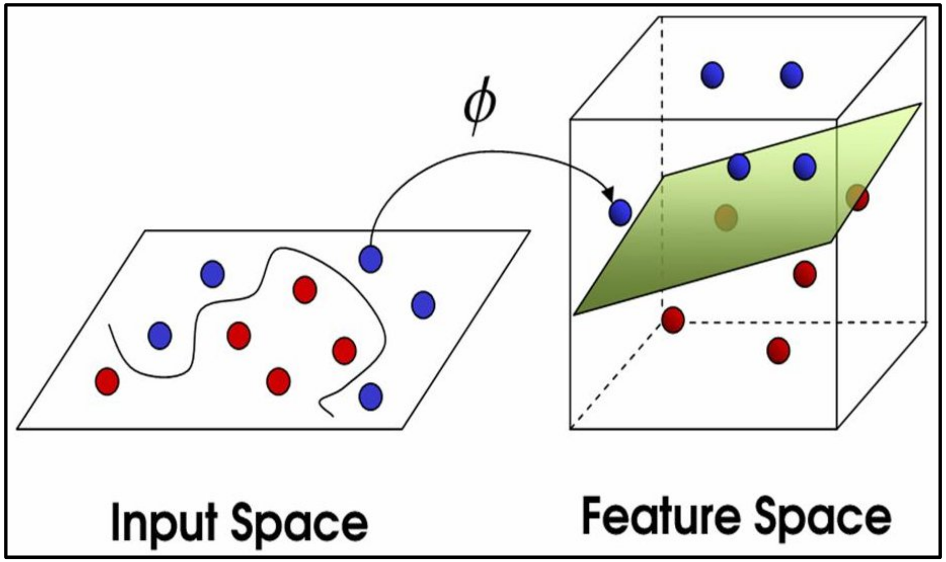


from www.imtech.res.in: rice blast prediction
http://www.imtech.res.in/raghava/rbpred/svm.jpg

Statistical methods used
SVM basics: the kernel-trick

Getting **non-linear**

Every dot product is replaced by a **non-linear kernel function**
→ input space transported into a high-dimensional feature space



from www.imtech.res.in: rice blast prediction
http://www.imtech.res.in/raghava/rbpred/svm.jpg

Some kernel functions:

polynomial (homogenous):

$$k(x,x')=\langle x,x'\rangle^d$$

polynomial (inhomogeneous):

$$k(x,x')=(\langle x,x'\rangle+1)^d$$

radial basis function:

$$k(x,x')=\mathrm{e}^{-\gamma*\langle x-x',x-x'\rangle}\ for\ \gamma>0$$

sigmoid:

$$k(x,x')=\tanh(\kappa\langle x,x'\rangle+c)$$
$$for\ some\ \kappa>0\wedge c<0$$

Statistical methods used
SVR in particular: from abstract to application

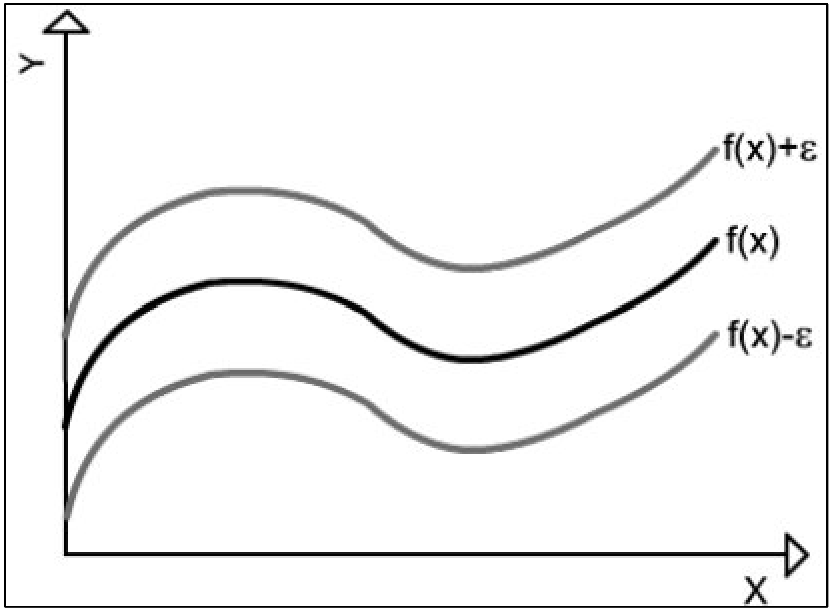
simplified:

$$\{(x_1, y_1), \dots, (x_n, y_n)\} \subset \mathcal{X} \times \mathbb{R} \xrightarrow{f(x_i) \approx y_i} f(x) = \langle w, x \rangle + b$$

Seeking for best w,b :

$$\text{minimize } \frac{1}{2} \|w\|^2$$

$$\text{subject to } \begin{cases} y_i - \langle w, x_i \rangle - b \leq \epsilon \\ \langle w, x_i \rangle + b - y_i \leq \epsilon \end{cases}$$



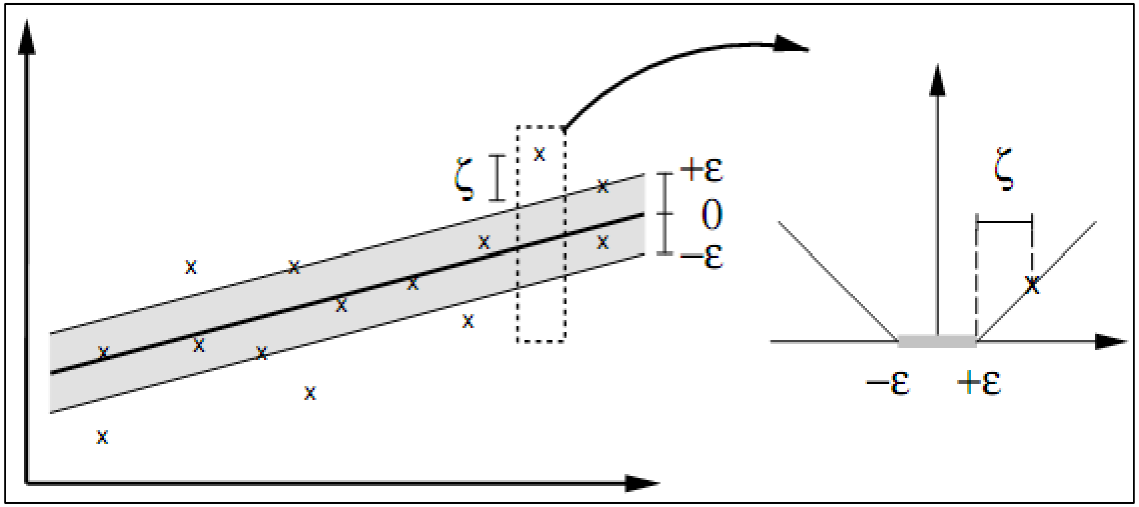
from “Online SVR“, Parrella
http://onlinesvr.altervista.org/Theory/Images/03-17.png

Statistical methods used
SVR in particular: from abstract to application

Adding the „soft-margin“ loss function to obtain the ε-regression:

$$\text{minimize } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

$$\text{subject to } \begin{cases} y_i - \langle w, x_i\rangle - b \le \epsilon + \xi_i \\ \langle w, x_i\rangle + b - y_i \le \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases}$$



from “A Tutorial on SVR“, Smola / Schölkopf

Quadratic Programming Problem

Behaviour: ε-insensitive loss-function:

$$|\xi|_\epsilon := \begin{cases} 0 & \text{for } |\xi| \le \epsilon \\ |\xi| - \epsilon & \text{otherwise} \end{cases}$$

Statistical methods used

SVR in particular: building the Lagrangian

$$L := \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

objective / primal function

$$-\sum_{i=1}^{n}(\eta_i \xi_i + \eta_i^* \xi_i^*)$$

$$-\sum_{i=1}^{n}\alpha_i(\epsilon + \xi_i - y_i + \langle w, x_i\rangle + b)$$

$$-\sum_{i=1}^{n}\alpha_i^*(\epsilon + \xi_i^* + y_i - \langle w, x_i\rangle - b)$$

$\alpha_i, \alpha_i^*, \eta_i, \eta_i^* \geq 0$ are Langrangian multipliers

w, b, ξ_i, ξ_i^* are the primal variables

⇓

$$\partial_b L = \sum_{i=1}^{n}(\alpha_i^* - \alpha_i) = 0$$

$$\partial_w L = w - \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)x_i = 0$$

$$\partial_{\xi_i^{(*)}} L = C - \alpha_i^{(*)} - \eta_i^{(*)} = 0$$

Statistical methods used

SVR in particular: gaining insight

$$\partial_w L = w - \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) x_i = 0$$

can be rewritten as $w = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) x_i$

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \langle x_i , x \rangle + b$$

w can be **completely** described by a linear combination of the training patterns!

Statistical methods used

SVR in particular: optimization problem

Substitution yields the dual optimization problem:

$$\text{maximize} \begin{cases} -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)\langle x_i, x_j\rangle \\ -\epsilon\sum_{i=1}^{n}(\alpha_i+\alpha_i^*)+\sum_{i=1}^{n}y_i(\alpha_i-\alpha_i^*) \end{cases}$$

$$\text{subject to } \sum_{i=1}^{n}(\alpha_i-\alpha_i^*)=0 \wedge \alpha_i, \alpha_i^* \in [0, C]$$

easier Quadratic Programming Problem
(solvable by several optimization algorithms)

Statistical methods used
Model assessment

A performance measure for regression:

Pearson's Correlation Coefficient

$$Corr(X,Y)=\frac{Cov(X,Y)}{\sqrt{Var(X)}\cdot\sqrt{Var(Y)}}=\frac{Cov(X,Y)}{\sigma_X\cdot\sigma_Y}\in[-1,1]$$

Statistical methods used
Model assessment

A performance measure for regression:

Mean Squared Error

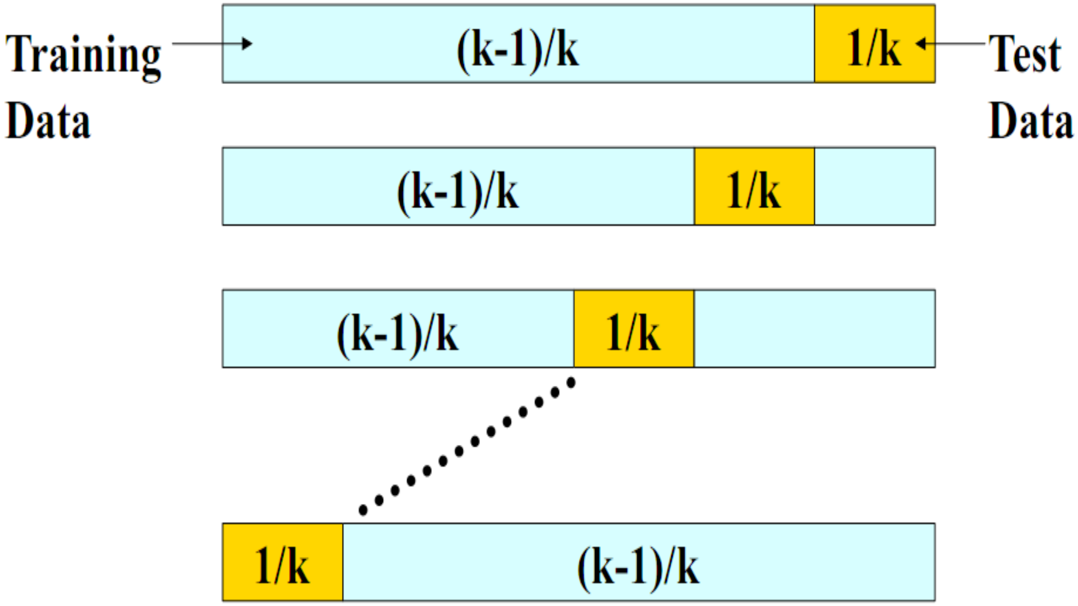
$$MSE(f(X)) = E[(X - f(X))^2] = Var(f(X)) + Bias^2(f(X), X)$$

$$f(X): predictor\ \ for\ X$$

Statistical methods used
Model assessment

Estimating the prediction quality with limited data:

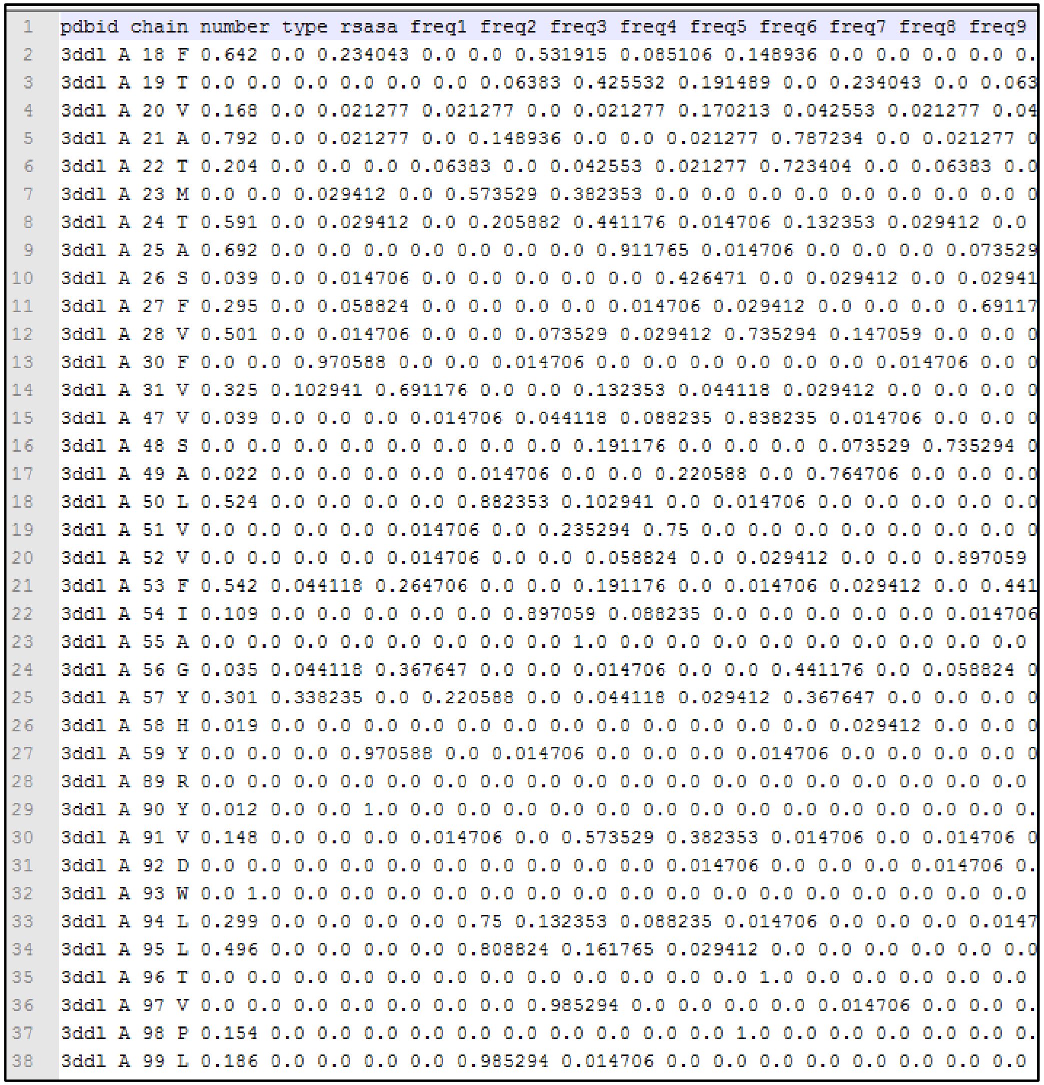
k-fold cross-validation



From Lecture Bioinformatik I by H.-P. Lenhof

The practical part

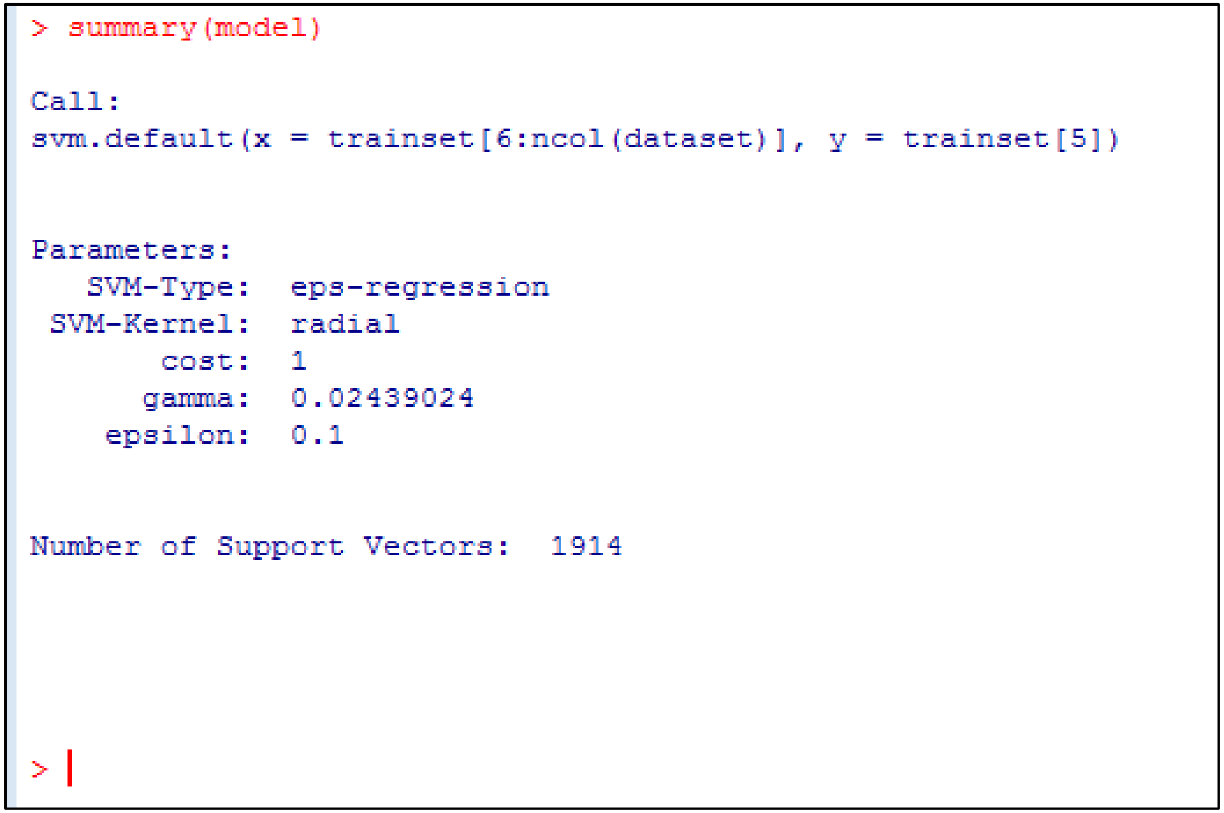
The dataset and the problem



```
pdbid chain number type rsasa freq1 freq2 freq3 freq4 freq5 freq6 freq7 freq8 freq9
3ddl A 18 F 0.642 0.0 0.234043 0.0 0.0 0.531915 0.085106 0.148936 0.0 0.0 0.0 0.0 0.
3ddl A 19 T 0.0 0.0 0.0 0.0 0.0 0.0 0.06383 0.425532 0.191489 0.0 0.234043 0.0 0.063
3ddl A 20 V 0.168 0.0 0.021277 0.021277 0.0 0.021277 0.170213 0.042553 0.021277 0.04
3ddl A 21 A 0.792 0.0 0.021277 0.0 0.148936 0.0 0.0 0.021277 0.787234 0.0 0.021277 0
3ddl A 22 T 0.204 0.0 0.0 0.0 0.06383 0.0 0.042553 0.021277 0.723404 0.0 0.06383 0.0
3ddl A 23 M 0.0 0.0 0.029412 0.0 0.573529 0.382353 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0
3ddl A 24 T 0.591 0.0 0.029412 0.0 0.205882 0.441176 0.014706 0.132353 0.029412 0.0
3ddl A 25 A 0.692 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.911765 0.014706 0.0 0.0 0.0 0.073529
3ddl A 26 S 0.039 0.0 0.014706 0.0 0.0 0.0 0.0 0.0 0.426471 0.0 0.029412 0.0 0.02941
3ddl A 27 F 0.295 0.0 0.058824 0.0 0.0 0.0 0.0 0.014706 0.029412 0.0 0.0 0.0 0.69117
3ddl A 28 V 0.501 0.0 0.014706 0.0 0.0 0.073529 0.029412 0.735294 0.147059 0.0 0.0 0
3ddl A 30 F 0.0 0.0 0.970588 0.0 0.0 0.014706 0.0 0.0 0.0 0.0 0.0 0.0 0.014706 0.0 0
3ddl A 31 V 0.325 0.102941 0.691176 0.0 0.0 0.132353 0.044118 0.029412 0.0 0.0 0.0 0
3ddl A 47 V 0.039 0.0 0.0 0.0 0.014706 0.044118 0.088235 0.838235 0.014706 0.0 0.0 0
3ddl A 48 S 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.191176 0.0 0.0 0.0 0.073529 0.735294 0
3ddl A 49 A 0.022 0.0 0.0 0.0 0.0 0.014706 0.0 0.0 0.220588 0.0 0.764706 0.0 0.0 0.0
3ddl A 50 L 0.524 0.0 0.0 0.0 0.0 0.882353 0.102941 0.0 0.014706 0.0 0.0 0.0 0.0 0.0
3ddl A 51 V 0.0 0.0 0.0 0.0 0.014706 0.0 0.235294 0.75 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0
3ddl A 52 V 0.0 0.0 0.0 0.0 0.014706 0.0 0.0 0.058824 0.0 0.029412 0.0 0.0 0.897059
3ddl A 53 F 0.542 0.044118 0.264706 0.0 0.0 0.191176 0.0 0.014706 0.029412 0.0 0.441
3ddl A 54 I 0.109 0.0 0.0 0.0 0.0 0.0 0.897059 0.088235 0.0 0.0 0.0 0.0 0.0 0.014706
3ddl A 55 A 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
3ddl A 56 G 0.035 0.044118 0.367647 0.0 0.0 0.014706 0.0 0.0 0.441176 0.0 0.058824 0
3ddl A 57 Y 0.301 0.338235 0.0 0.220588 0.0 0.044118 0.029412 0.367647 0.0 0.0 0.0 0
3ddl A 58 H 0.019 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.029412 0.0 0.0 0
3ddl A 59 Y 0.0 0.0 0.0 0.970588 0.0 0.014706 0.0 0.0 0.0 0.014706 0.0 0.0 0.0 0.0 0
3ddl A 89 R 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
3ddl A 90 Y 0.012 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.
3ddl A 91 V 0.148 0.0 0.0 0.0 0.014706 0.0 0.573529 0.382353 0.014706 0.0 0.014706 0
3ddl A 92 D 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.014706 0.0 0.0 0.0 0.014706 0.
3ddl A 93 W 0.0 1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
3ddl A 94 L 0.299 0.0 0.0 0.0 0.0 0.75 0.132353 0.088235 0.014706 0.0 0.0 0.0 0.0147
3ddl A 95 L 0.496 0.0 0.0 0.0 0.0 0.808824 0.161765 0.029412 0.0 0.0 0.0 0.0 0.0 0.0
3ddl A 96 T 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0
3ddl A 97 V 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.985294 0.0 0.0 0.0 0.0 0.014706 0.0 0.0 0.
3ddl A 98 P 0.154 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0 0.
3ddl A 99 L 0.186 0.0 0.0 0.0 0.0 0.985294 0.014706 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
```

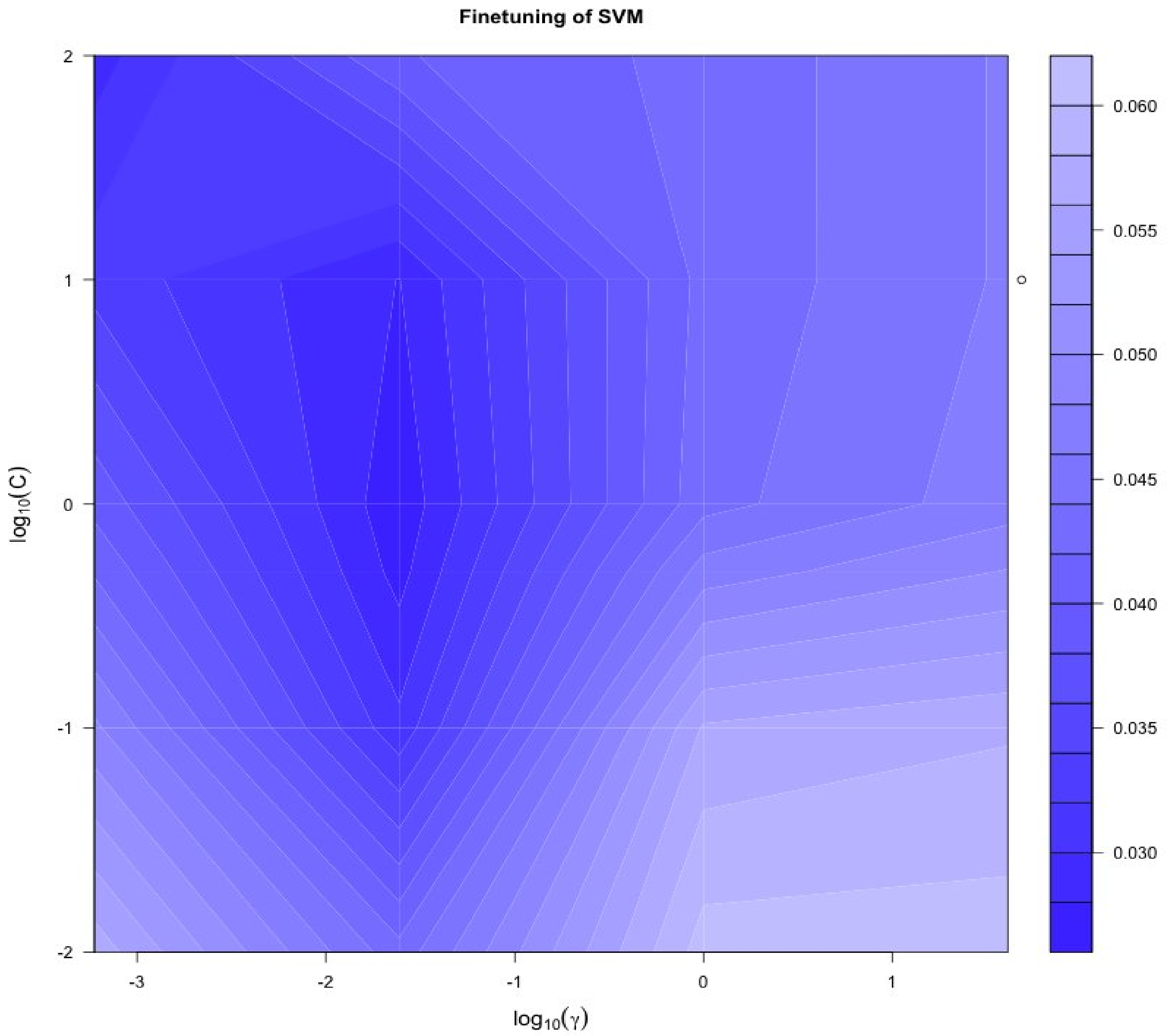
- 2595 residues with computed rSASA
  - of the transmembrane regions of 28 different proteins
- 41 features for each residue
  - frequencies per aa
  - PSIBLAST-score per aa
  - conservation-score

The practical part
Workflow



```
> summary(model)

Call:
svm.default(x = trainset[6:ncol(dataset)], y = trainset[5])


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.02439024
    epsilon:  0.1


Number of Support Vectors:  1914




> 
```

The practical part
SVR Setup



The practical part
Results

after 2595-fold cross-validation on training data:

Total Mean Squared Error: 0.02627267
Squared Correlation Coefficient: 0.4464628

Correlation of **0.668**

The practical part

Accuracy per amino acid

Amino acid	Mean squared error $[10^{-2}]$
R	0.055
H	0.058
Q	0.100
D	0.243
N	0.530
E	0.585
K	0.879
S	1.035
T	1.307
G	1.789
Y	1.902
M	2.160
P	2.571
W	2.603
C	2.872
A	3.003
V	3.306
I	3.327
L	3.328
F	3.704

The practical part

Accuracy per amino acid

Amino acid	Mean squared error [10^{-2}]	Burried [%]	
R	0.055	100.0	hydrophilic
H	0.058	90.9	
Q	0.100	100.0	
V	3.306	39.9	hydrophobic
I	3.327	30.0	
L	3.328	27.7	
F	3.704	28.4	

The practical part

Accuracy per amino acid

Amino acid	Mean squared error $[10^{-2}]$	Burried [%]	Mean conservation score	
R	0.055	100.0	1.673	
H	0.058	90.9	1.413	hydrophilic
Q	0.100	100.0	1.040	
V	3.306	39.9	-0.230	
I	3.327	30.0	-0.255	hydrophobic
L	3.328	27.7	-0.204	
F	3.704	28.4	0.008	

The practical part
Contribution of individual features

feature(s)	Correlation with rSASA
aa-freq.	0.0806
PSSM-score	0.1481
conservation	0.4393

Other probably interesting features:

hydrophobicity	0.2862
vdW - volumes	0.2044

The practical part

Adding new features to the model

Features used for prediction	Correlation of prediction and rSASA
standard	0.668
- conservation	0.644
+ hydrophobicity	0.671
+ hydrophobicity + VdW - volumes	0.669

Summary

Membrane proteins are interesting for bioinformatics

Support Vector Machines are **useful tools** in machine learning

Features that are **most likely useful** in sequence based methods:

→ neighborhood (windowing)
→ profiles / PSSM
→ conservation

But, like always:

it is **not** about the more the better!

Used literature

Hastie, Tibshirani, Friedman:
The Elements of Statistical Learning (sec. edition, 2008)

Park, Hayat, Helms:
Prediction of the burial status of transmembrane residues of helical membrane proteins (paper, BMC Bioinformatics, 2007)

Smola, Schölkopf:
A Tutorial on Support Vector Regression (paper, 2003)

Thank you for your attention!