# iDNA-Prot|dis

Critical paper review by

Jasleen Grewal
Ph.D. Student, Steven Jones Lab
Canada's Michael Smith Genome Sciences Centre

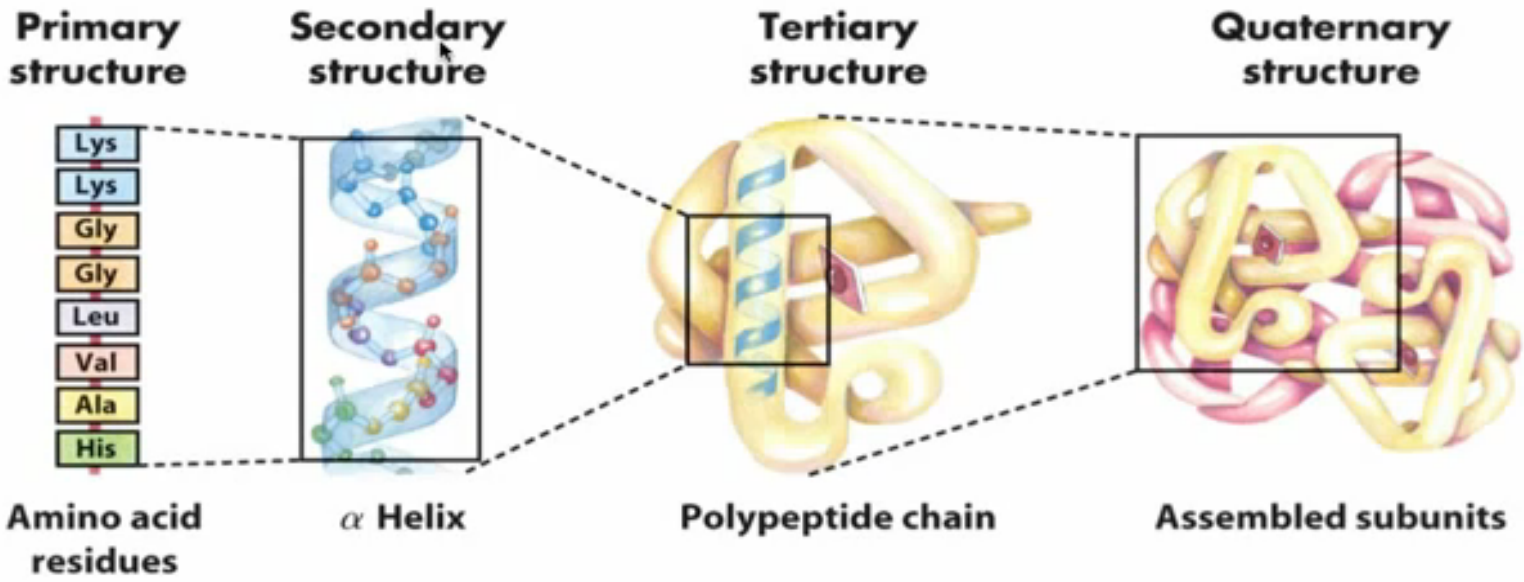
# iDNA-Prot|dis: Identifying DNA-Binding Proteins by Incorporating Amino Acid Distance-Pairs and Reduced Alphabet Profile into the General Pseudo Amino Acid Composition

**Bin Liu[1,2,3,4]*, Jinghao Xu[1], Xun Lan[5], Ruifeng Xu[1,2], Jiyun Zhou[1], Xiaolong Wang[1,2], Kuo-Chen Chou[4,6]***

**1** School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong, China, **2** Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong, China, **3** Shanghai Key Laboratory of Intelligent Information Processing, Shanghai, China, **4** Gordon Life Science Institute, Belmont, Massachusetts, United States of America, **5** Stanford University, Stanford, California, United States of America, **6** Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia

- DNA-binding proteins (vs non DNA-binding proteins)
- Modification of general pseudo amino acid composition
- Classify using Support Vector Machine (SVM)

# Protein Structure




David L. Nelson, Michael M. Cox. *Lehninger Principles of Biochemistry.* New York: W. H.Freeman and Company, 2008.

# Finding protein sequence & **structure**

- Chromatin Immuno Precipitation
- X-ray crystallography
- Mass Spectrometry (Nordhoff, 1999)
- Machine Learning
  - Random Forest (DNA-Prot)
  - Grey Models (iDNA-Prot)
  - Support Vector Machines (iDNA-Prot|dis)
  - Ensemble classifiers (nDNA-Prot)

# Data

**Source**

- Protein Data Bank (PDB). Dec 2013.
- *mmCIF keyword "(non) DNA binding protein"*

**Filters**

- Keep proteins > 50 residues long
- Remove proteins containing residue 'X' (unknown)
- Remove proteins with sequence similarity > 25% (PISCES)

**Final set**

- 525 DNA binding proteins ($S^+$)
- 550 non DNA binding proteins ($S^-$)

# Sequence vs Structure

- Protein Structure
  - Energy state models
  - Identifying functional domains
- Amino Acid Sequence
  - Sequence evolutionary profiles
  - Chou's Pseudo Amino Acid Compositions (PseAAC)
  - Combination of the two (ex. iDNA-prot)

# (Good fences make) Good neighbors

- Pseudo Amino Acid Compositions

Consider a protein chain of $L$ amino acid residues:

$$R_1R_2R_3R_4R_5R_6R_7 \cdots R_L \tag{1}$$

effect can be approximately reflected with a set of sequence order-correlated factors as defined below:

$$\begin{cases} \theta_1 = \dfrac{1}{L-1}\displaystyle\sum_{i=1}^{L-1} \Theta(R_i, R_{i+1}) \\ \theta_2 = \dfrac{1}{L-2}\displaystyle\sum_{i=1}^{L-2} \Theta(R_i, R_{i+2}) \\ \theta_3 = \dfrac{1}{L-3}\displaystyle\sum_{i=1}^{L-3} \Theta(R_i, R_{i+3}), \quad (\lambda < L) \\ \cdots \\ \theta_\lambda = \dfrac{1}{L-\lambda}\displaystyle\sum_{i=1}^{L-\lambda} \Theta(R_i, R_{i+\lambda}) \end{cases} \tag{2}$$

PROTEINS: Structure, Function, and Genetics 43:246-255 (2001)

# (Good fences make) Good neighbors

- Pseudo Amino Acid Compositions
- Pairwise Distance

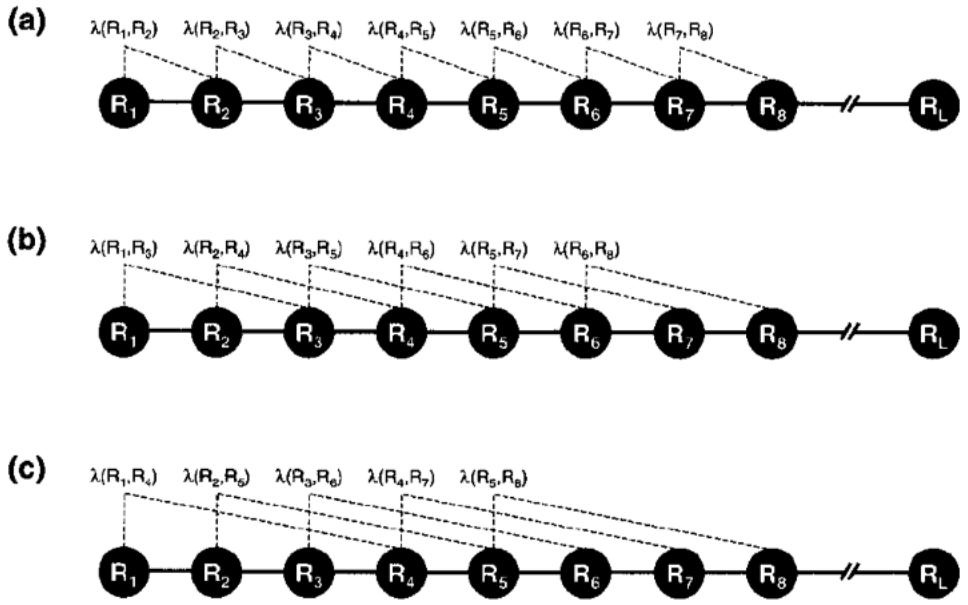



Fig. 1. A schematic drawing to show (a) the first-tier, (b) the second-tier, and (3) the third-tier sequence order correlation mode along a protein sequence. Panel (**a**) reflects the correlation mode between all the most contiguous residues, panel (**b**) that between all the secon-most contiguous residues, and panel (**c**) that between all the third-most contiguous residues.

PROTEINS: Structure, Function, and Genetics 43:246-255 (2001)

# The high dimensionality conundrum

- 20 Amino Acids
  - Studying neighbors at distance ‘1’, in a protein of length ‘L’, we get (20 x 20) x (L-1) pairwise comparisons
- Average protein length in eukaryotes?
- Feature set is massive!
- Increased computational time and resources
- Redundancy in associations -> Overfitting

# Solutions?

- Limit scope of classification?
  - Restrict to a sub-region of proteins – informative?
- Reduce the feature vector size?
  - Use amino acid cluster profiles -> Reduced alphabet

# Representative AA cluster profiles

- Different proteins have different functions
- Functions derived from single amino acids (ex. Lys, Arg)
- Functions derived from a cluster of amino acids (ex. Leucine zippers (2 x heptads))
- Analyze protein sequence in terms of these ‘function defining’ amino acids
  - Fewer ‘units’ to compare

# Summary - the 'dis' in iDNA-Prot|dis

- Reduced Amino Acid Alphabets
  - Clustering of AAs based on some measure of their relative similarity[1]
- Selected 3 different alphabet profiles for DNA binding proteins (after testing 164 reduced alphabet schemes)[1]
  - cp(13)= {MF; IL; V; A; C; WYQHP; G; T; S; N; RK; D; E}
  - cp(14)= {EIMV; L; F; WY; G; P; C; A; S; T; N; HRKQ; E; D}
  - cp(19)= {P; G; E; K; R; Q; D; S; N; T; H; C; I; V; W; YF; A; L; M}

*n(c) = x, where x defines cp(x)*

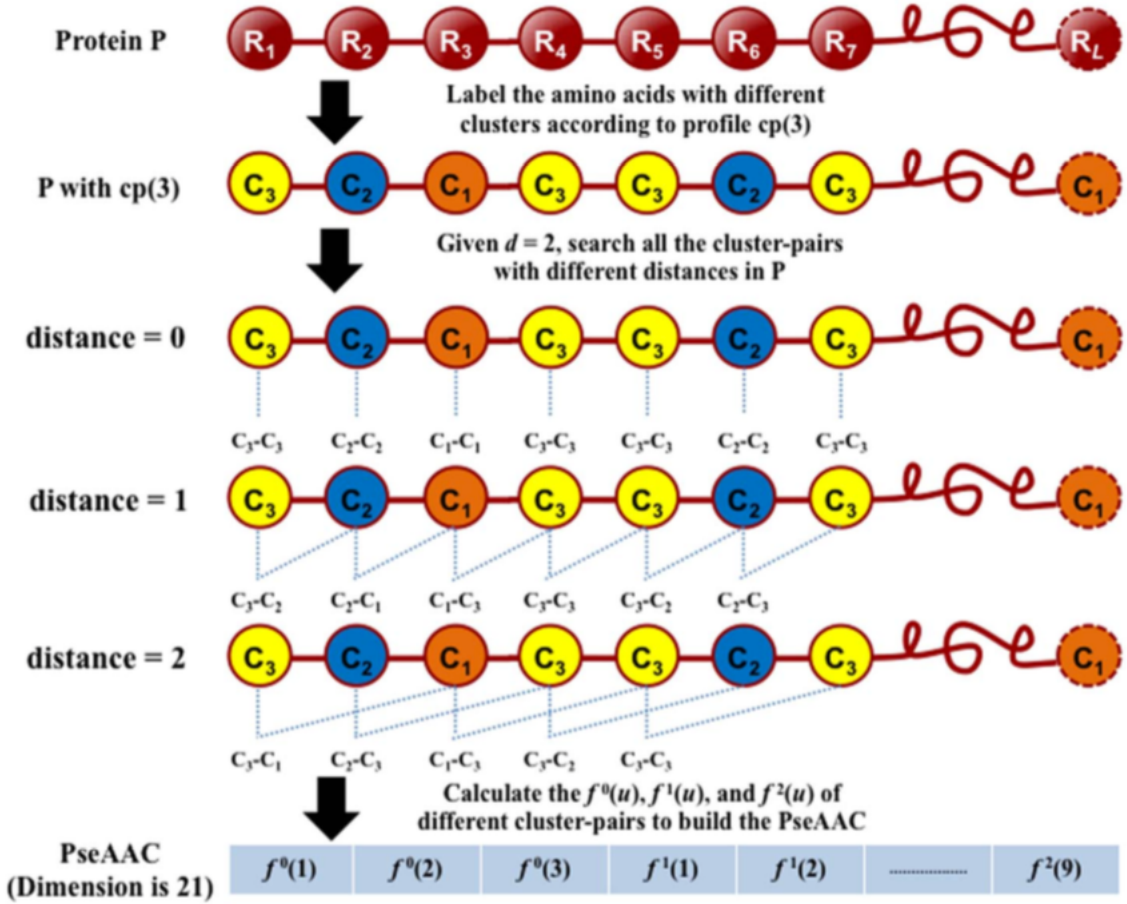
- New dimension size = $n(c) + n^2(c)d$

*when considering max pairwise distance = 'd'*

1. Peterson et al. Reduced Amino Acid Alphabets Exhibit an Improved Sensitivity and Selectivity in Fold Assignment. 2009, Bioinformatics.

# (Fewer = Good?) neighbors

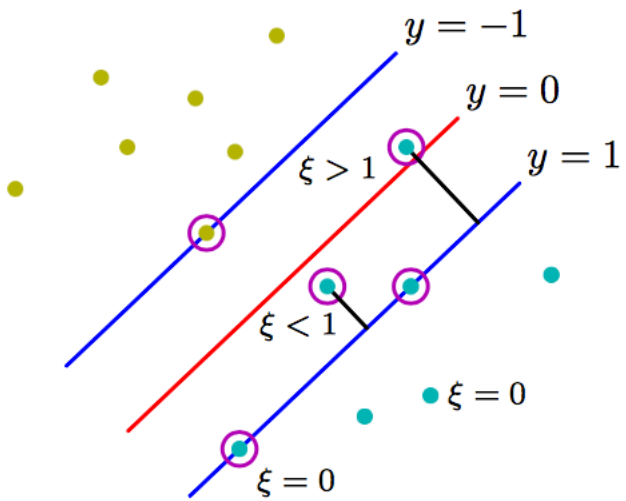
- Pseudo Amino Acid Compositions
- Pairwise Distance (ex. For n(c) = 3)

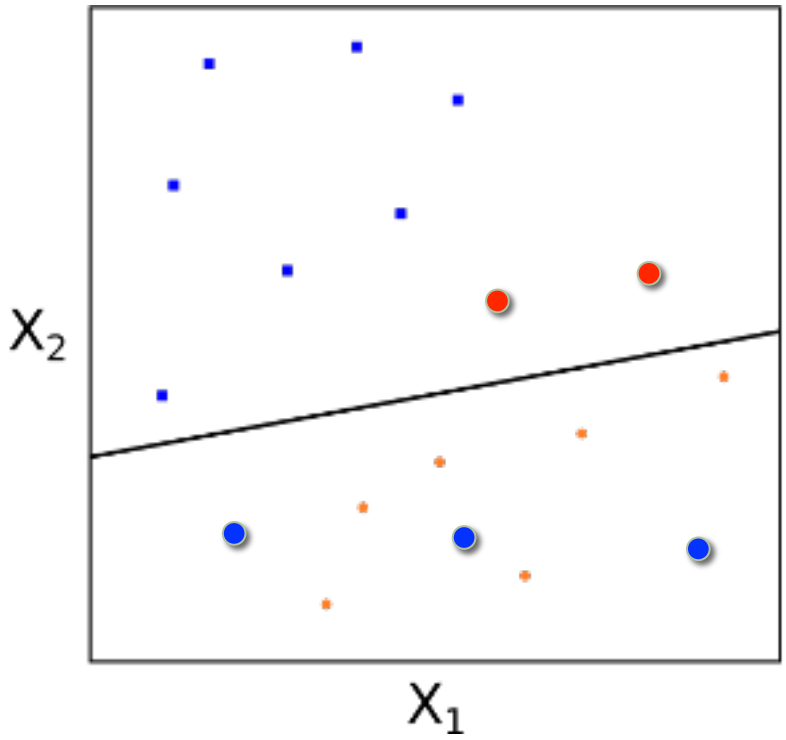



doi:10.1371/journal.pone.0106691.g001

# Classification of data - Perceptrons

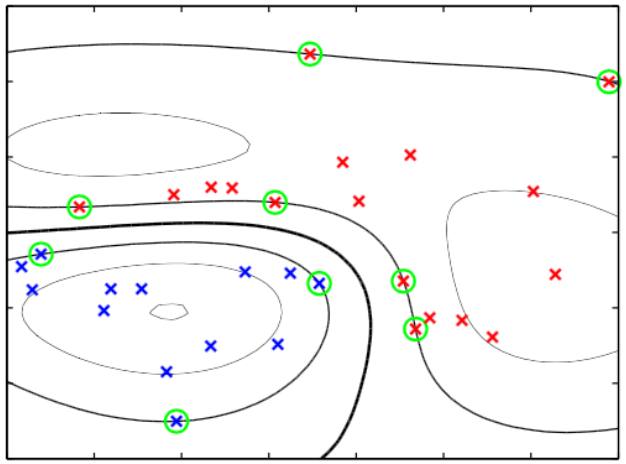
- 2 class problem
  - Minimize error over ‘mis-classified’ examples only
- Only classify linearly separable data



# Support Vector Machines

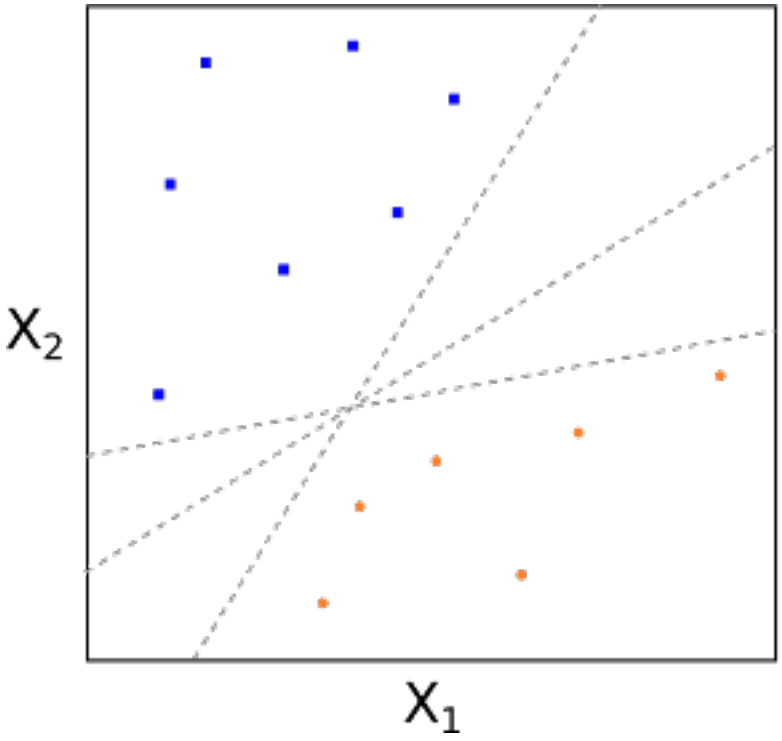
- An example of an SVM applied to a non-linear feature space



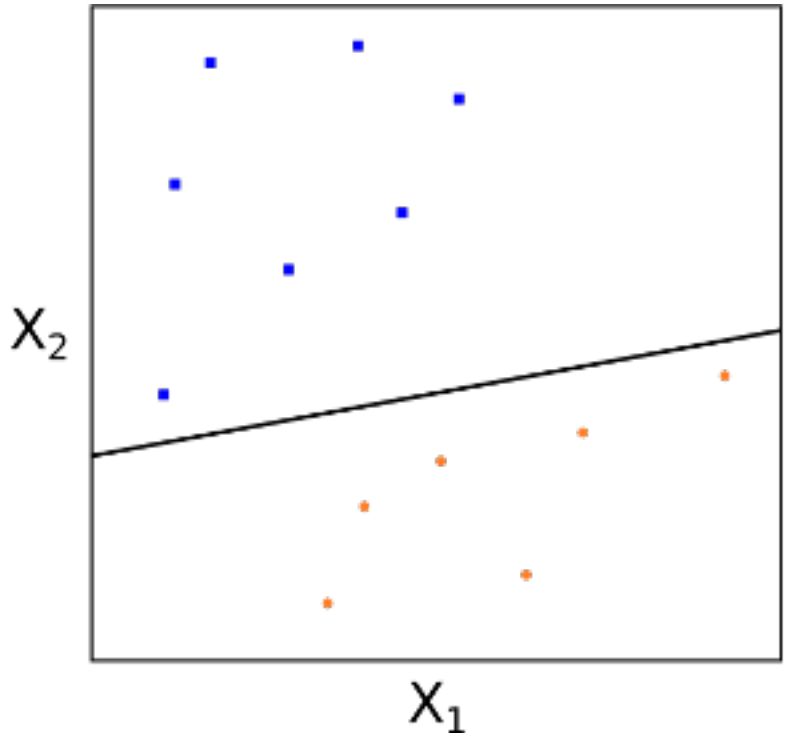
SVM trained using Gaussian kernel

Support vectors circled

Note non-linear decision boundary in $\boldsymbol{x}$ space



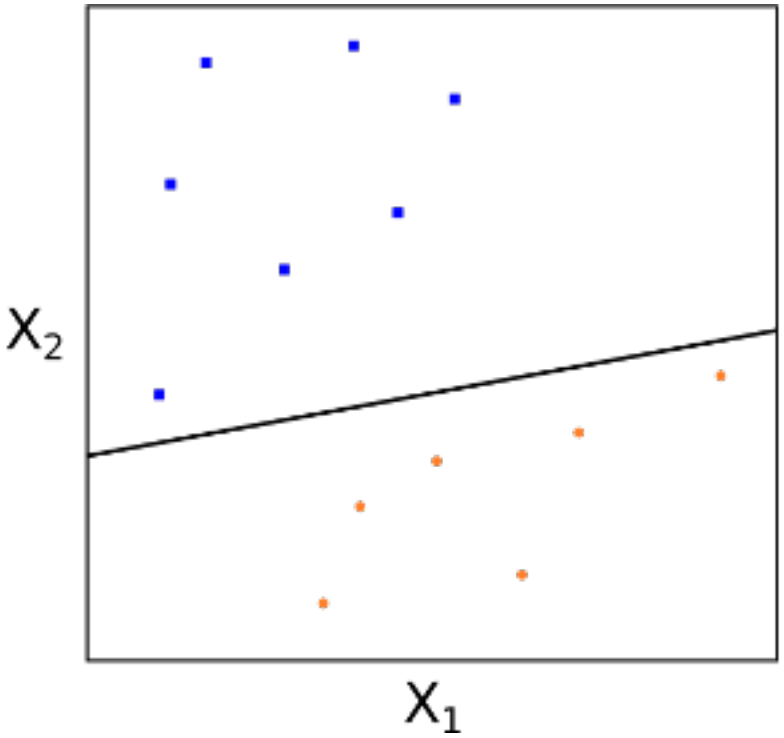

Multiple possible ‘hyperplanes’



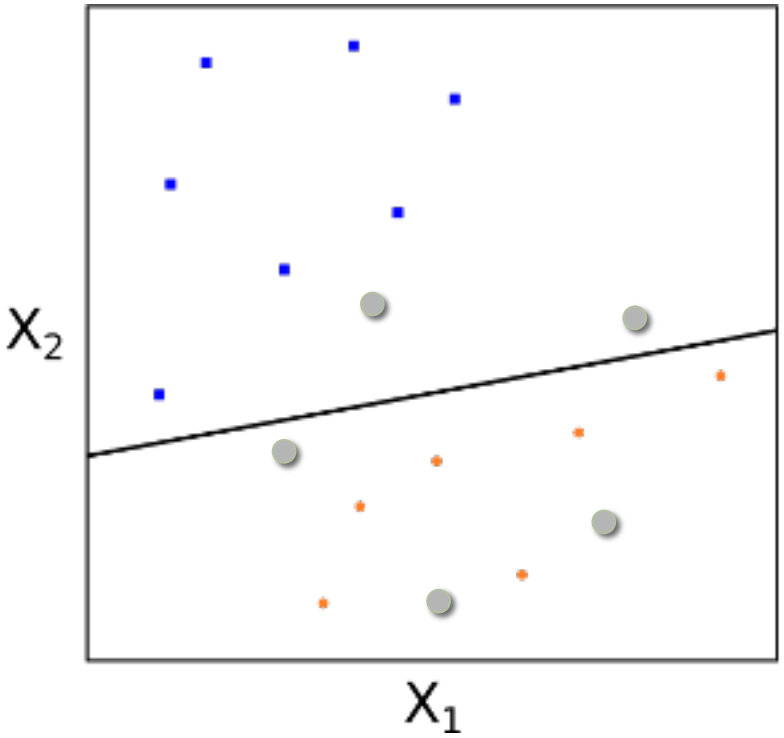

Hyperplane with perfect separation

Training dataset

# Good Classifier



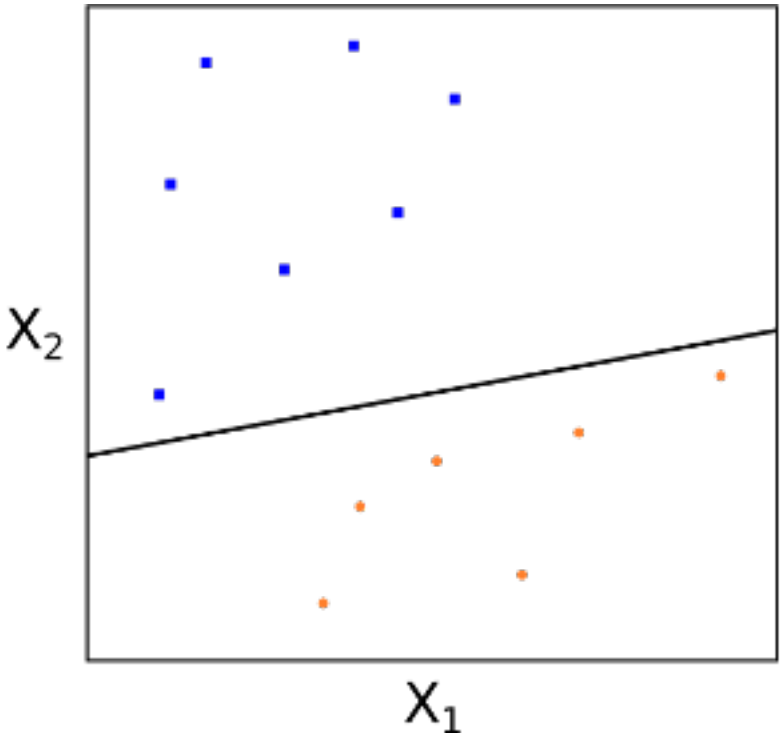

Learnt optimal hyperplane



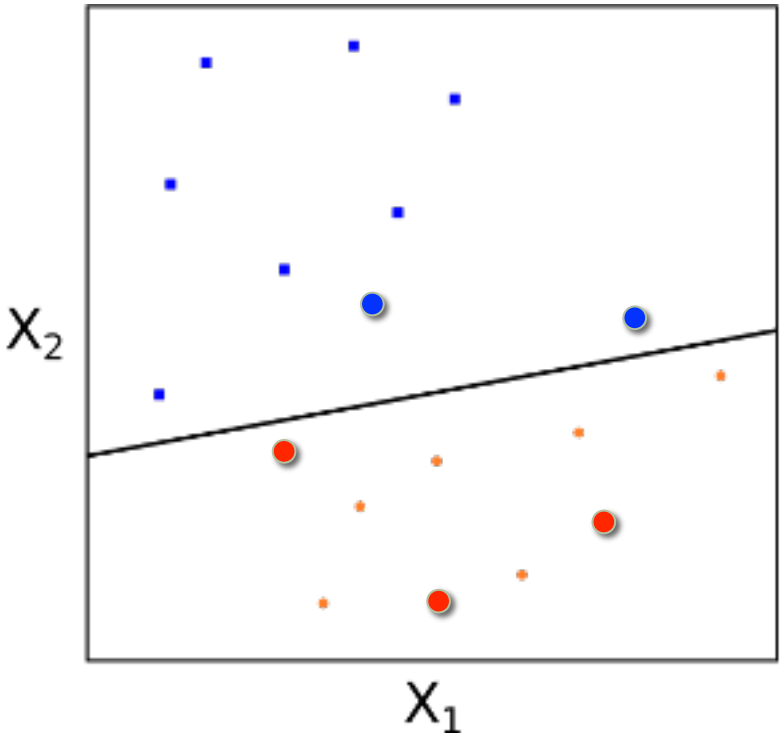

Test datapoints

Test dataset 1

# Good Classifier



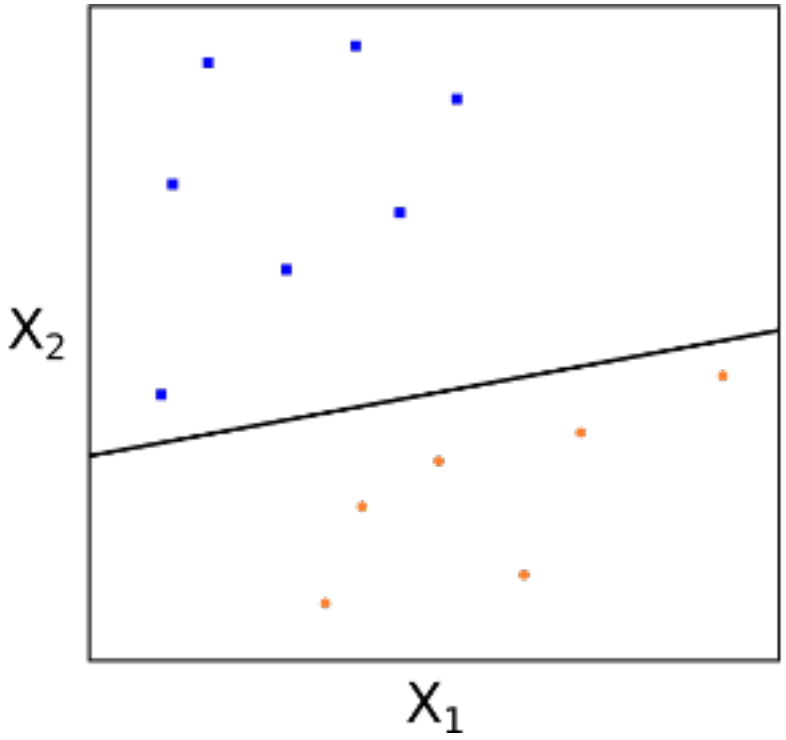

Learnt optimal hyperplane



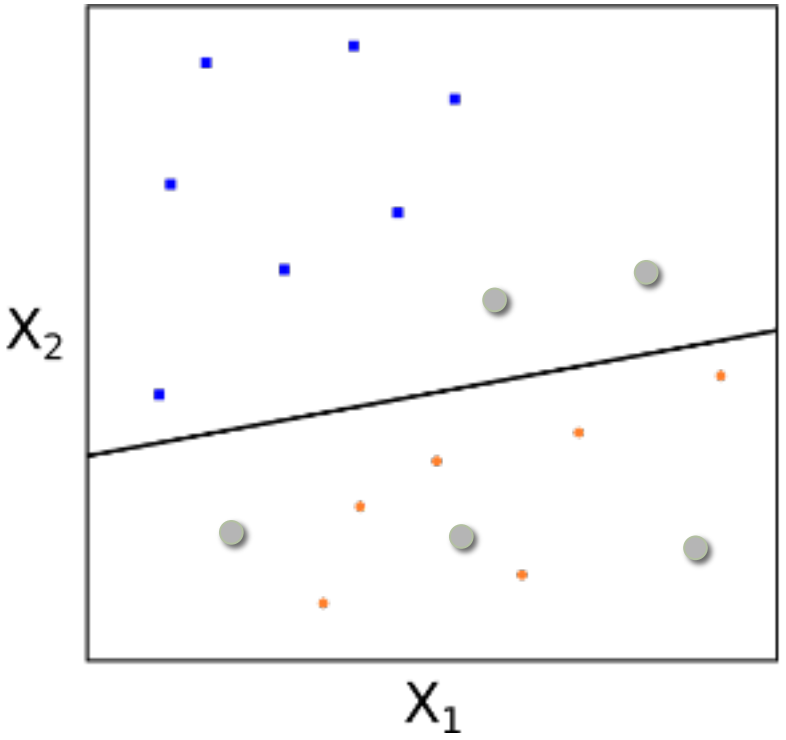

Test datapoints classified

Test dataset 1

# Bad Classifier



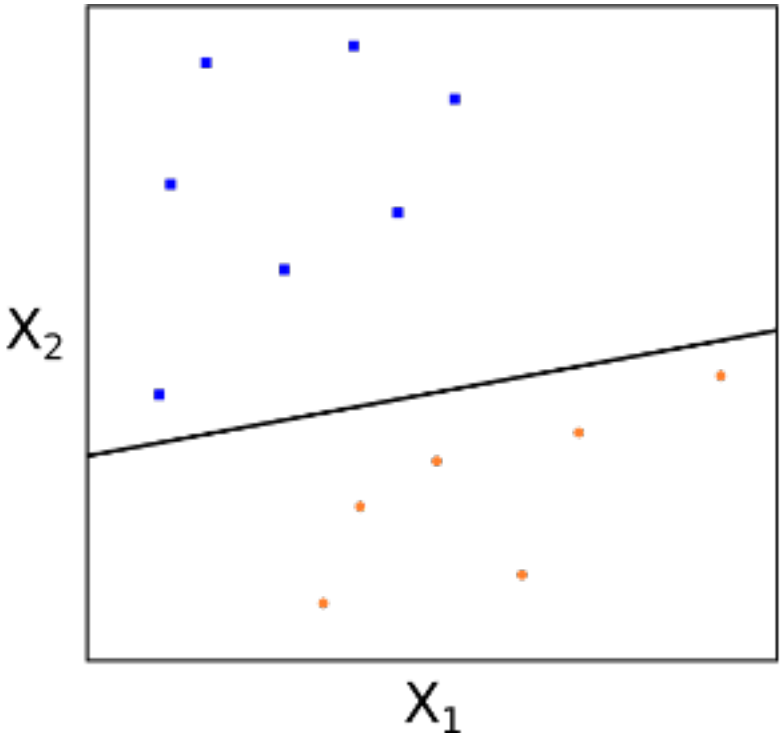

Learnt optimal hyperplane



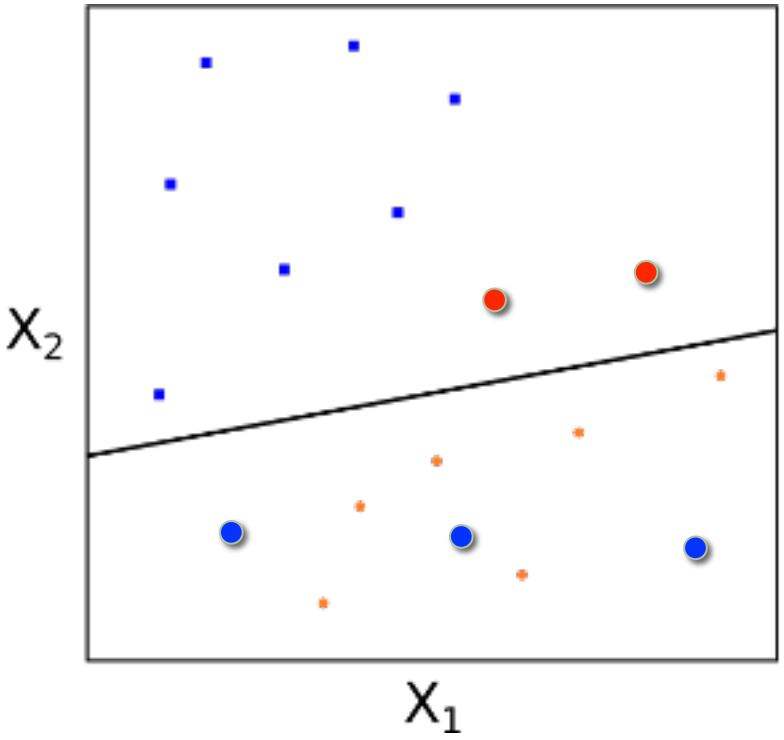

Test datapoints

Test dataset 2

# Bad Classifier




Learnt optimal hyperplane




Test datapoints classified

Test dataset 2

# Assessing classifier performance

- Method to assess predictor
  - 4 metrics (single label systems – right (1), wrong(0))
- Sensitivity – Call 1, 1
- Specificity – Call 0, 0
- Accuracy – Call 1,1; Call 0,0
- Mathew's correlation coefficient
  - 1 : High sensitivity, and High specificity
  - 0 : No different from random prediction
  - -1 : Total disagreement between prediction and observation

# Optimal Maximum Distance ‘d’

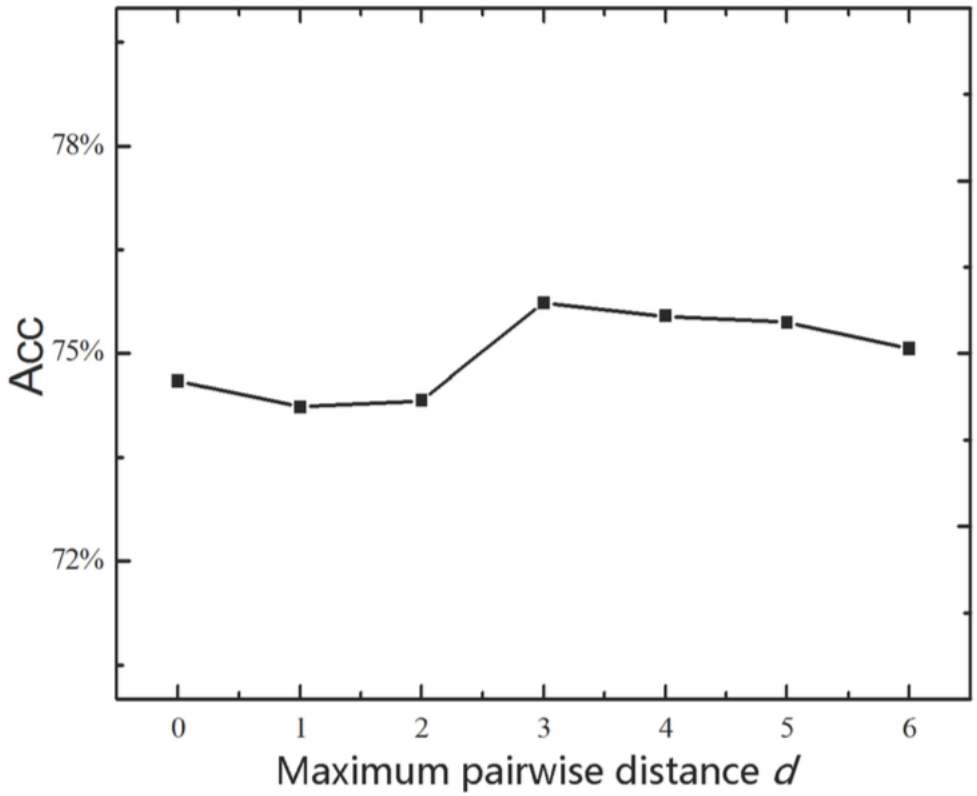
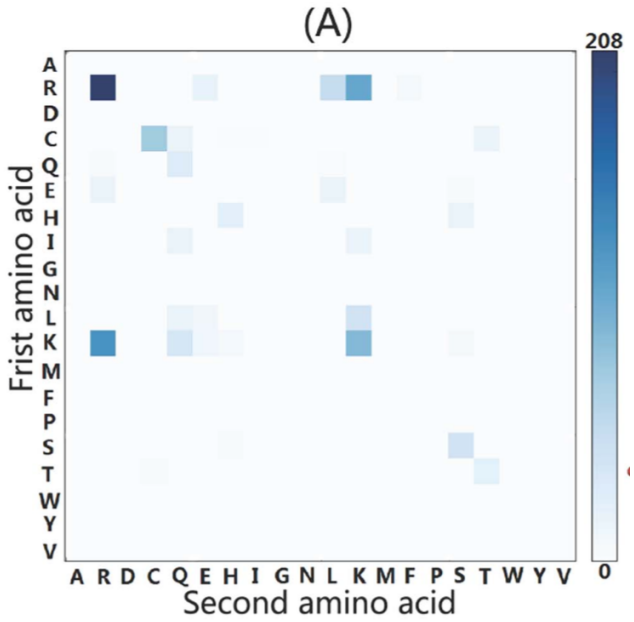



**Figure 2. The overall Acc values achieved by iDNA-Prot|dis for cp(20) with different *d* values based on the benchmark dataset through five-cross validation.**
doi:10.1371/journal.pone.0106691.g002

$$\mathbf{W} = \mathbf{A} \cdot \mathbf{M} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix}^{\mathbf{T}} \begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1j} \\ m_{21} & m_{22} & \cdots & m_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ m_{N1} & m_{N2} & \cdots & m_{Nj} \end{bmatrix}$$




cp(20) with max distance = 3 (400 AA pairs)

N = 1075, j = 1220

# Results – reduced alphabet space

**Table 1.** The jackknife test results by iDNA-Prot|dis with different amino acid alphabet profiles (cf. Eqs. 9–13) on the benchmark dataset of Eq. 1 (cf. Supporting Information S1).

| Cluster profile | Acc (%) | MCC | Sn(%) | Sp(%) | AUC(%) |
|---|---|---|---|---|---|
| cp(20)[a] | 75.81 | 0.52 | 81.14 | 70.72 | 83.40 |
| cp(19)[b] | 76.46 | 0.53 | 82.28 | 70.90 | 83.30 |
| cp(14)[c] | **77.30** | **0.54** | 79.40 | 75.27 | 82.60 |
| cp(13)[d] | 77.20 | 0.54 | 80.76 | 73.81 | 83.10 |

[a]The parameters used: $d=3$, $C=4$, $\gamma=2^{-13}$.
[b]The parameters used: $d=3$, $C=4$, $\gamma=2^{-13}$.
[c]The parameters used: $d=3$, $C=2$, $\gamma=2^{-12}$.
[d]The parameters used: $d=3$, $C=64$, $\gamma=2^{-17}$.
doi:10.1371/journal.pone.0106691.t001

- Feature space reduction
  - cp(20) = 1220 features
  - **cp(14) = 602 features (**$14+(14^2)(3)$)

# Mirror mirror on the wall…

- Comparison with existing methods
  - Dataset = Benchmark dataset (525 DNABP, 550 non-DNABP)

**Table 2.** A comparison of the jackknife test results by iDNA-Prot|dis with the other methods on the benchmark dataset of Eq. 1.

| Method | Acc(%) | MCC | Sn(%) | Sp(%) | AUC(%) |
|---|---|---|---|---|---|
| iDNA-Prot\|dis (cp(14))[a] | 77.30 | 0.54 | 79.40 | 75.27 | 82.60 |
| DNAbinder (dimension 21)[b] | 73.95 | 0.48 | 68.57 | 79.09 | 81.40 |
| DNAbinder (dimension 400)[c] | 73.58 | 0.47 | 66.47 | 80.36 | 81.50 |
| DNA-Prot[d] | 72.55 | 0.44 | 82.67 | 59.76 | 78.90 |
| iDNA-Prot[e] | 75.40 | 0.50 | 83.81 | 64.73 | 76.10 |

# Mirror mirror on the wall…

- Comparison with existing methods
  - Dataset = PDB186 (93 DNABP, 93 non-DNABP)

**Table 3.** A comparison of the results[a] obtained by iDNA-Prot|dis and the other methods on the independent dataset PDB186.

| Methods | Acc(%) | MCC | Sn(%) | Sp(%) | AUC(%) |
|---|---|---|---|---|---|
| iDNA-Prot\|dis | **72.00** | 0.445 | 79.50 | 64.50 | 78.60 |
| iDNA-Prot | 67.20 | 0.344 | 67.70 | 66.70 | N/A |
| DNA-Prot | 61.80 | 0.240 | 69.90 | 53.80 | N/A |
| DNAbinder | 60.80 | 0.216 | 57.00 | 64.50 | 60.70 |
| DNABIND | 67.70 | 0.355 | 66.70 | 68.80 | 69.40 |
| DNA-Threader | 59.70 | 0.279 | 23.70 | 95.70 | N/A |
| DBPPred | 76.90 | 0.538 | 79.60 | 74.20 | 79.10 |

[a]The results of iDNA-Prot [15], DNA-Prot [14], DNAbinder [96], DNABIND [102], DNA-Threader [5], and DBPPred [97] were obtained from [97].
doi:10.1371/journal.pone.0106691.t003

- DBPPred – random forest, with Gaussian naïve Bayes
  - GB outperformed decision tree, logistic regression, kNN, SVN (polynomial kernel), SVN (RBF kernel).
  - Trained on PDB594, tested on PDB186
  - Doi: 10.1371/journal.pone.0086703

# Caveats

- Works only on well-studied protein functional types
- Finding an optimal reduced alphabet set
  - Other computational tools for getting this set?
- Approach works for identifying a single type of protein
  - Assumption: all amino acid clusters characteristic of ‘protein class X’ being encompassed in the reduced alphabet space

# Applications

- Current anticancer chemotherapy drugs
  - DNA damaging agents
  - Gene regulators!
- Non-genotoxic DNA binding proteins
  - Prevent adverse effects of chemotherapy
  - Alter DNA plasticity and physico-chemical properties.
- Classification of new proteins
- Single class differentiation -> multi-class differentiation

# Questions?


WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY AREN'T THERE DINOSAUR GHOSTS

# Extensions to current approach

- Reduced alphabet set
  - Learn that through unsupervised clustering analyses
- Prediction of co-interacting proteins
  - Learn about new functional attributes of proteins
  - Learn about new protein classes?
- Multi-class separation using functionally characteristic alphabet sets

# SVM with different feature spaces

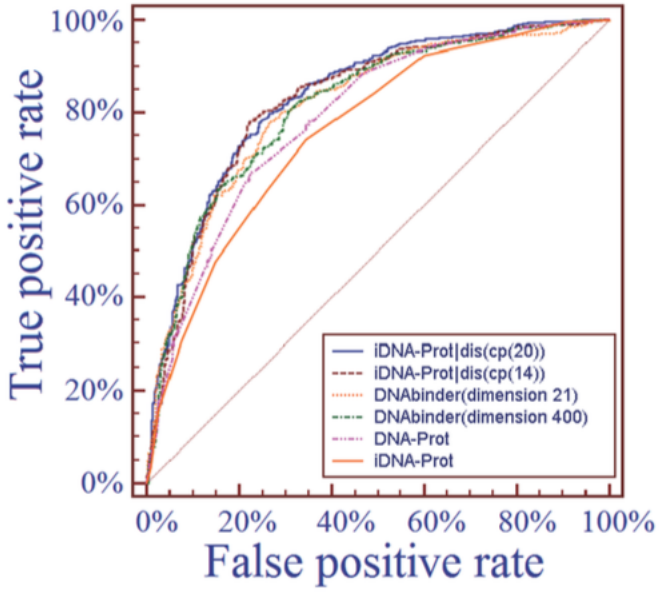



**Figure 4. The ROC (receiver operating characteristic) curves obtained by different methods on the benchmark dataset using the jackknife tests.** The areas under the ROC curves or AUC are 0.834, 0.826, 0.814, 0.815, 0.789 and 0.761 for iDNA-Prot|dis (cp(20)), iDNA-Prot|dis (cp(14)), DNAbinder (dimension 21), DNAbinder(dimension 400), DNA-Prot and iDNA-Prot, respectively. See the main text for further explanation.
doi:10.1371/journal.pone.0106691.g004

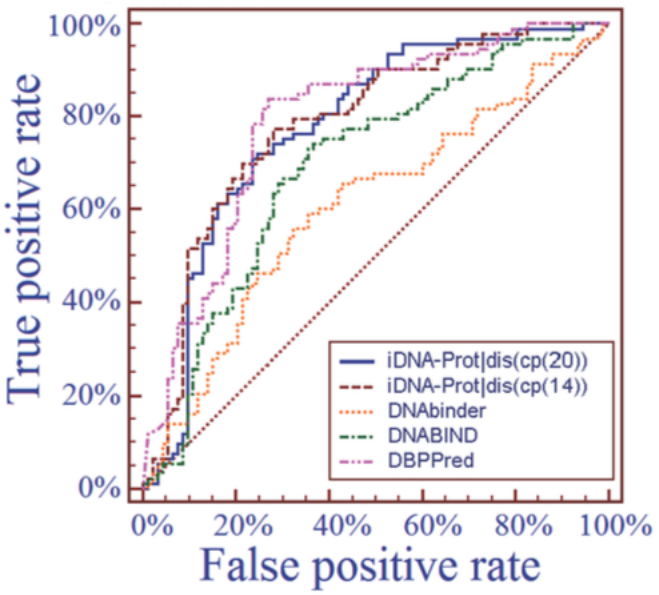



**Figure 5. The ROC (receiver operating characteristic) curves obtained by different methods on the independent dataset PDB186.** The areas under the ROC curves or AUC are 0.786, 0.779, 0.607, 0.694, and 0.791 for iDNA-Prot|dis(cp(20)), iDNA-Prot|dis(cp(14)), DNAbinder, DNABIND and DBPPred, respectively. See the main text for further explanation.
doi:10.1371/journal.pone.0106691.g005

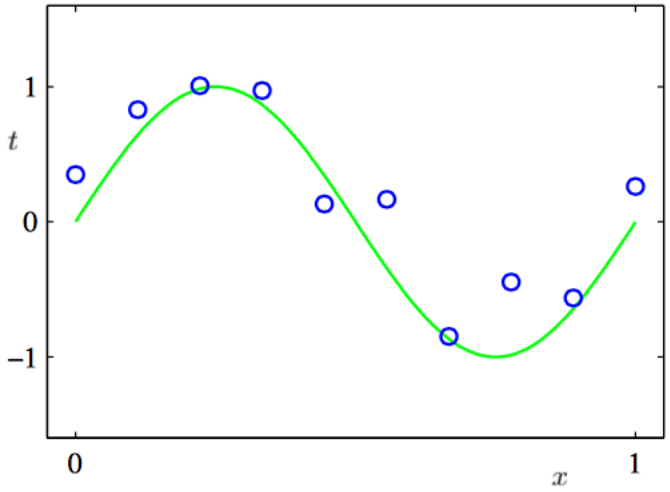
# Why study DNA binding proteins

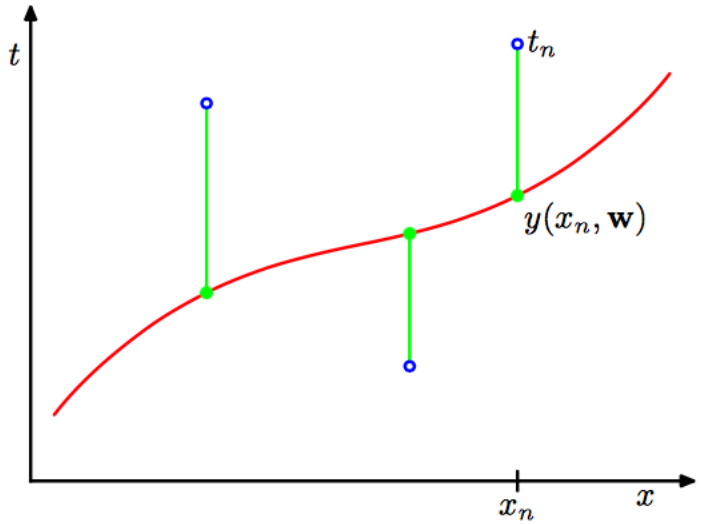
- 6-7% of eukaryotic proteins are DNA binding[1]
- Replication, transcription, DNA packaging
- Chemotherapy drugs: DNA damaging agents
- Identify non-genotoxic DNA binding proteins

CAP is a gene regulatory protein from *E.coli.* In the absence of the bound protein, this DNA helix is straight.[2]

1. Kumar et al. Identification of DNA binding proteins using SVMs and evolutionary profiles. 2007, BMC bioinformatics.
2. Molecular Biology of the Cell. 4th edition. Alberts B, Johnson A, Lewis J, et al. New York: Garland Science; 2002.

# Data points



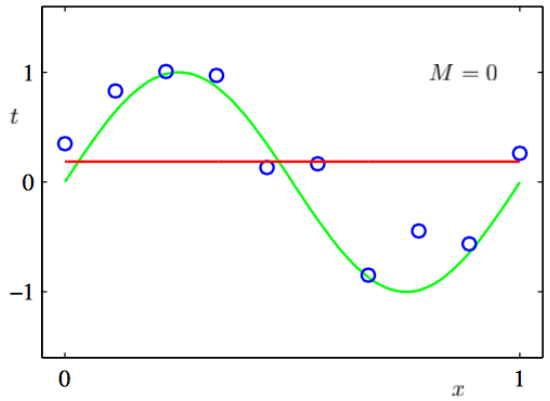



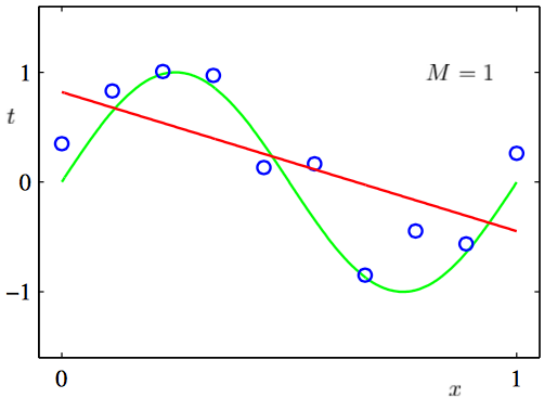

- Have a spread (density distribution)
- Optimal fit has an associated error rate
- Error calculated as ‘least squares’:

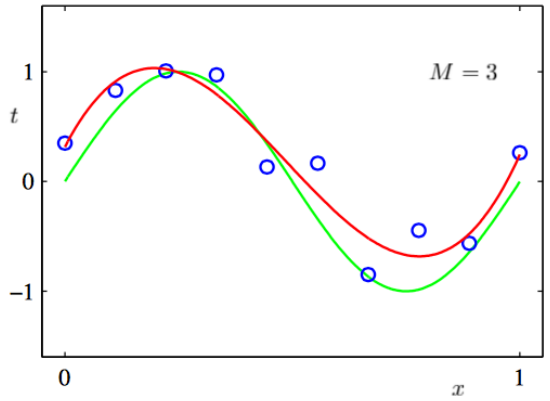
$$\tilde{E}(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2$$

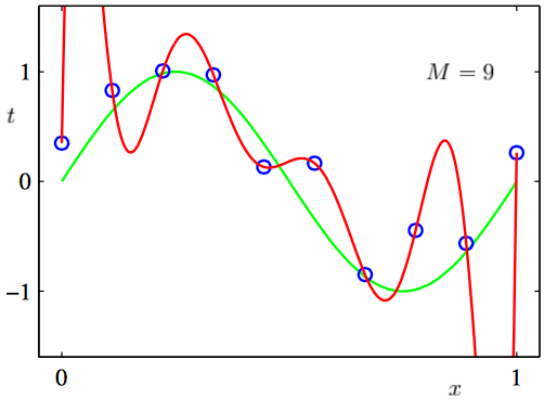
# Explaining the data points

- Find the best fit for my data, minimize error without getting carried away by the way it looks right now




$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} E(\boldsymbol{w})$$

# Preventing over fitting - regularization

- Penalize large coefficients

$$\tilde{E}(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{\lambda}{2}||\boldsymbol{w}||^2$$

Least Squares error Regularizer

- Best coefficients = least error

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} E(\boldsymbol{w})$$

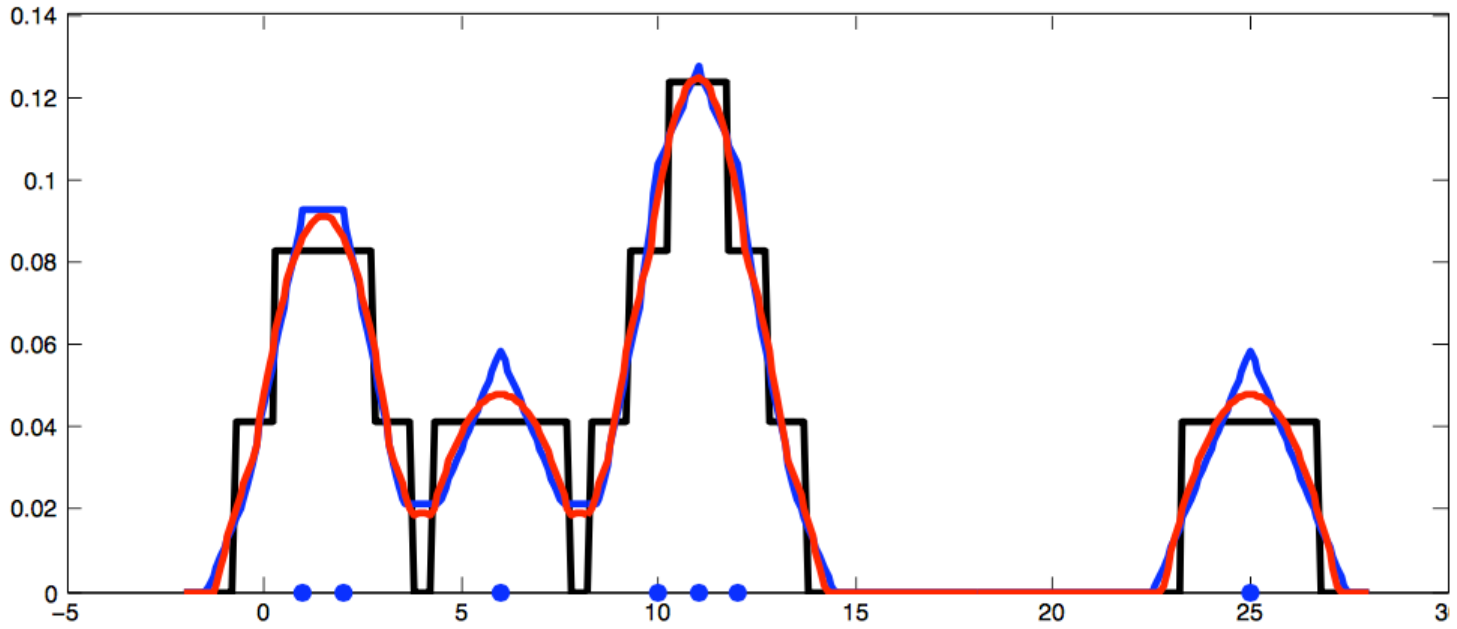
# Kernel Functions 

- Estimate distribution of data (density differences -> clusters)
- Estimate density by placing a tiny ‘bump’ around each datapoint – incorporate inherent variability in datapoint
- Kernel function k() determines shape of these bumps
- Commonly use (**R**adial) Gaussian **B**asis **F**unction (RBF)

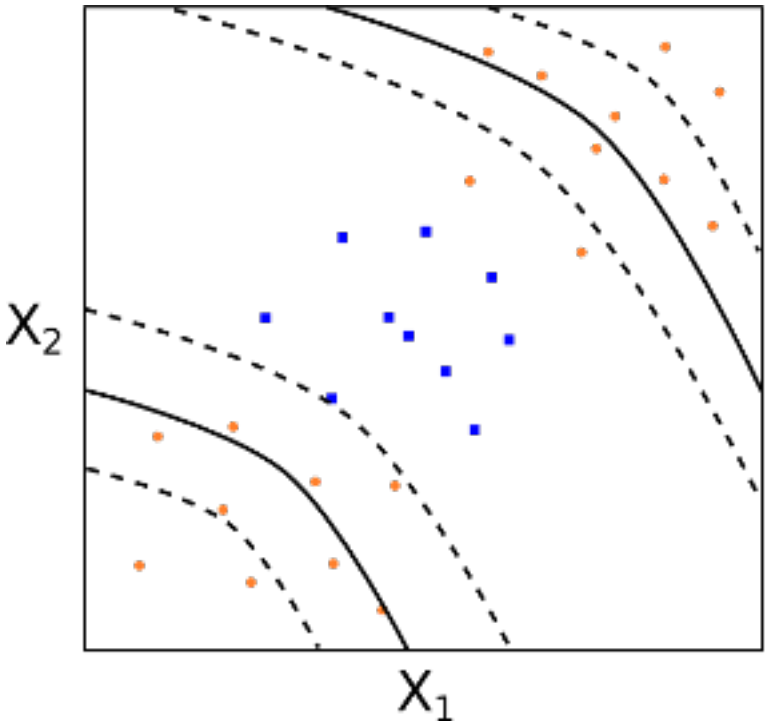
$$\phi_j(x) = \exp\{-\frac{(x-x_j)^2}{2s^2}\}$$

- Can now specify how ‘similar’ two points are, rather than construct an ‘exact comparison’ for every single pair of points.

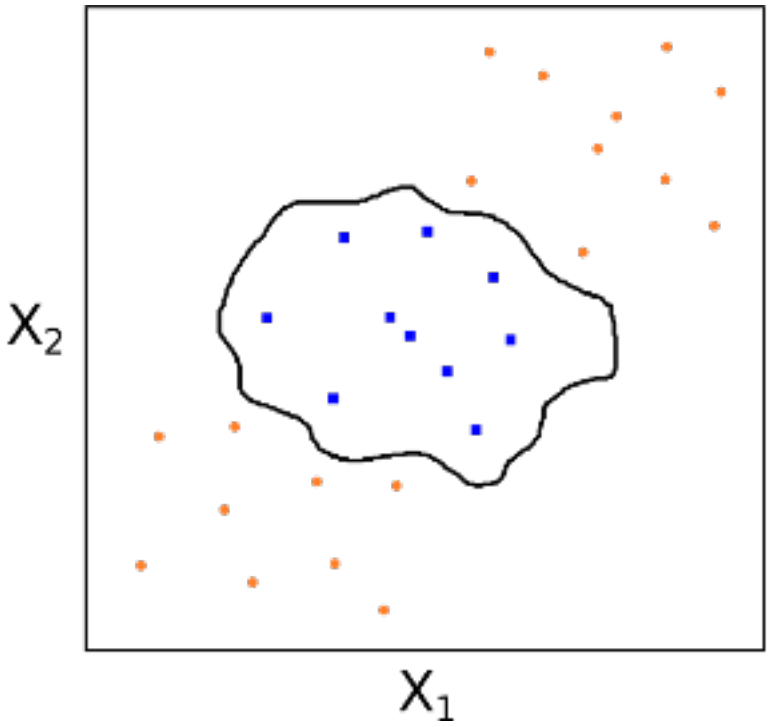
# Other types of Kernel Functions




**Rectangle**, Triangle, Epanechnikov



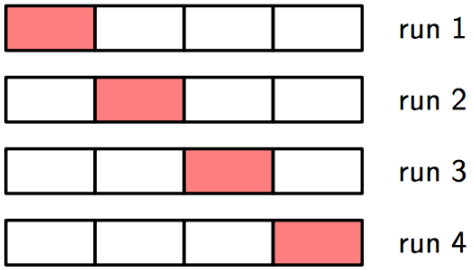

A d-degree polynomial kernel



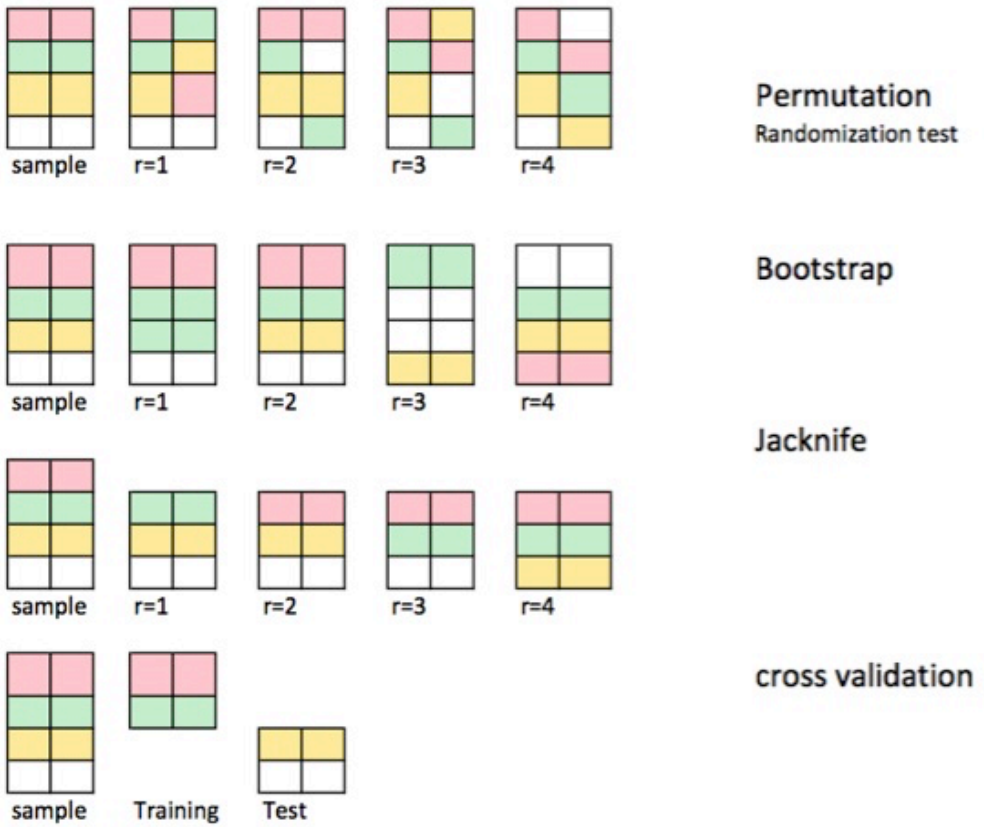

A radial kernel

# Cross validation - assessing performance

- Jackknifing
- Bootstrapping
- Leave-one-out
- Leave-k-out (subsampling)




LOO Cross Validation



# Regularization – assessing performance

- Train complex model but penalize it getting “too complex”
  - Sweet spot between error and regularization
  - Can use cross validation to find optimal regularization coefficient (λ)
- Prevent overfitting