Nara Women's University

Development of SVM based Prediction System for Metalbinding Sites in Protein

Masami Nakazawa*, Masami Takata*, Kiyonobu Yokota+, Tamotsu Noguchi +, Masakazu Sekijima+, and Kazuki Joe*

* Nara Women's University

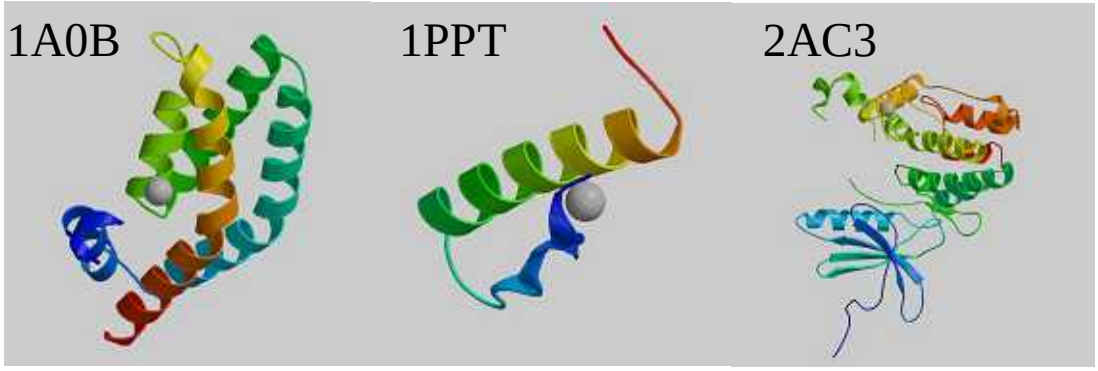
+ National Institute of Advanced Industrial Science and Technology (AIST)

Outline

- Background: Metalbinding in Protein
- Development of Prediction System
- Experiment
- Conclusion and Future works

Background: Metal protein

- Metal protein = Protein + Metal ion
- Activity of metal ion
  - **Folding**
  - **Cofactor**

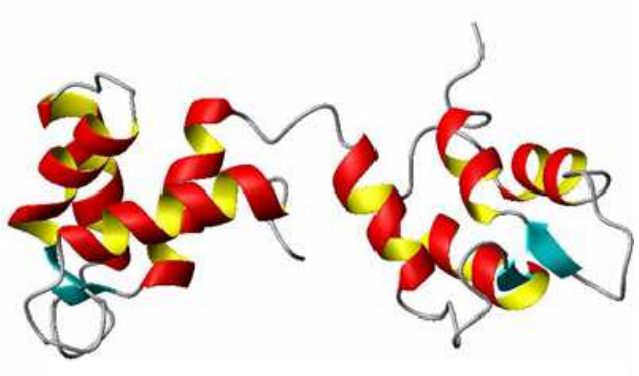


Example: Folding (Calmodulin)

low — Concentration of Ca ion — high

- Ca ion is not included
- Inactive enzymes cannot bind

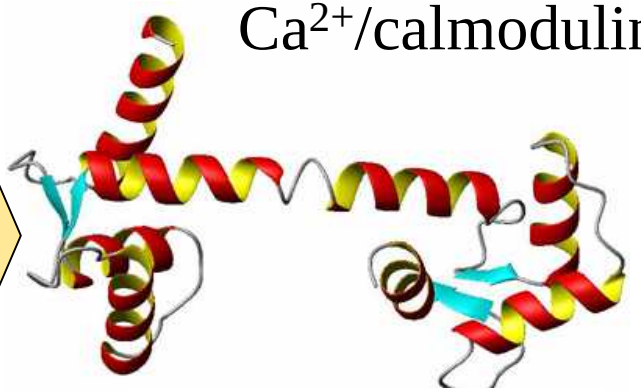
- Ca ion is included
- Inactive enzymes can bind



Not bind with Ca ion (1DMO)

Structural change

Ca^{2+}/calmodulin

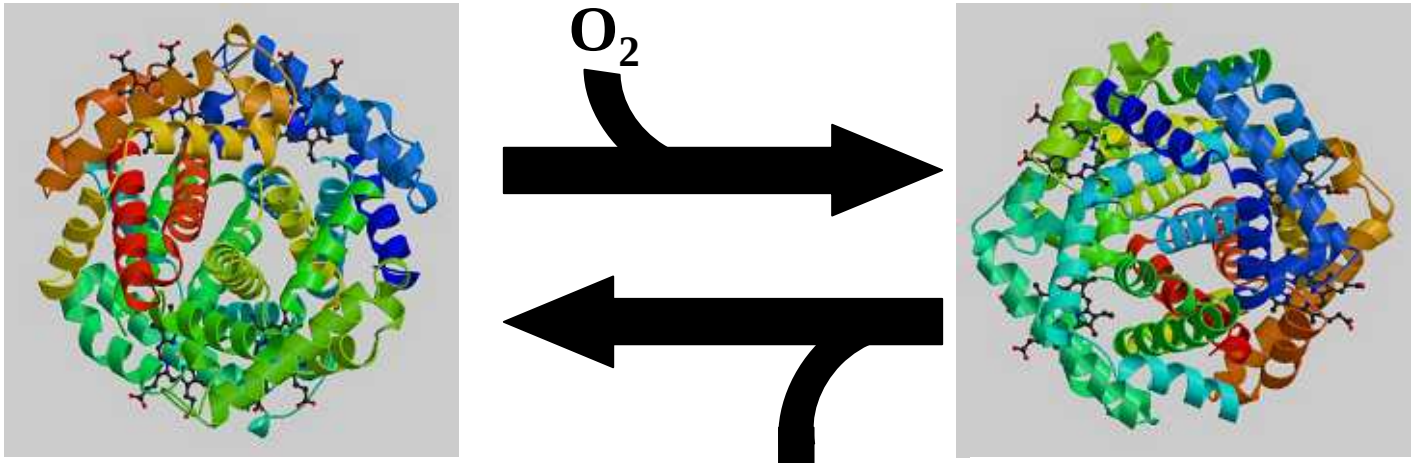


Bind with Ca ion (3CLN)

Example: Metal enzyme (hemoglobin)

- Hemoglobin: delivery of oxygen

low ← Partial pressure of oxygen → high



deoxyhemoglobin

oxyhemoglobin

Analysis approach for three-dimensional structure of proteins

- Analysis of three-dimensional protein structure
  - ✧ NMR (Nuclear Magnetic Resonance)
  - ✧ X-ray crystallography

huge monetary and time costs

- Computational approach
  - ✧ Experiment setting is easy
  - ✧ Reduce turn around time

Problems of computerized analysis

- No data of Three-dimensional protein structure
- No data of Potential function of metal ion

↓

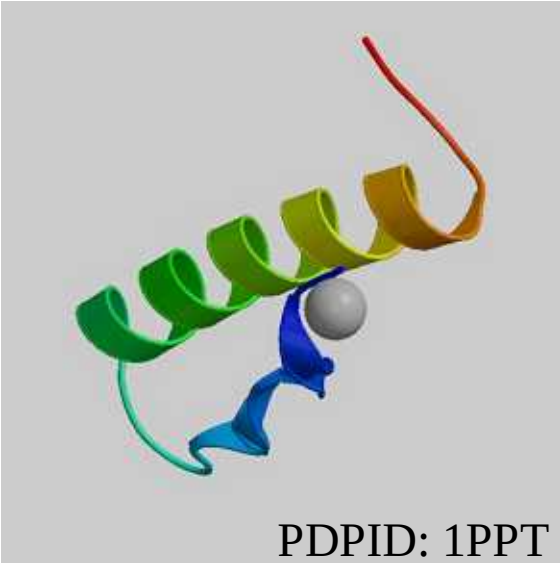
Computer simulation is hard

↓

A new software approach is required!!

Our approach

- From three-dimension to one dimension



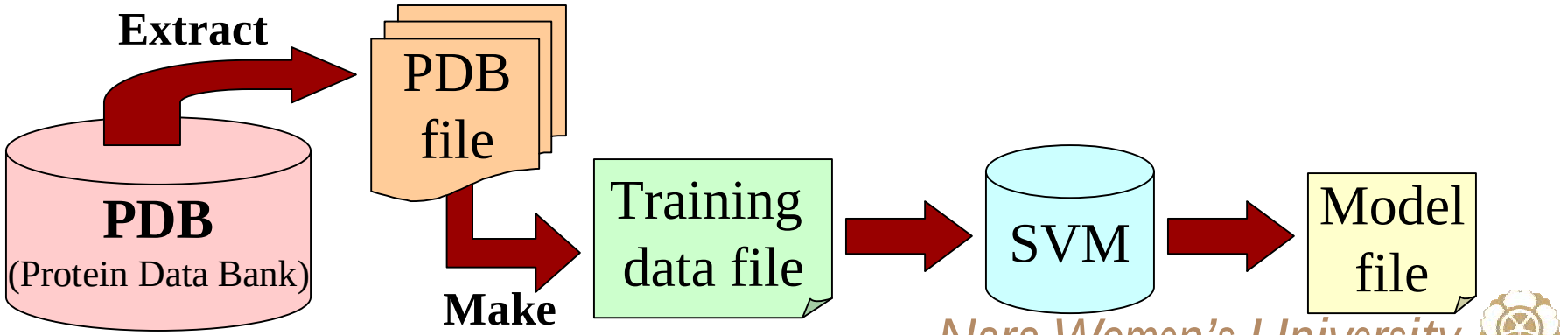
Three-dimensional structure

ARG – CYS – THR – HIS –
TYP – ALA – GLY – SER –
PRO – GLN – GLN – LEU –
CYS – ARG – PRO – MET –
PRO – HIS – ARG – LEU –
GLN – CYS – TYP – SER

Amino acid sequence
(One-dimensional structure)

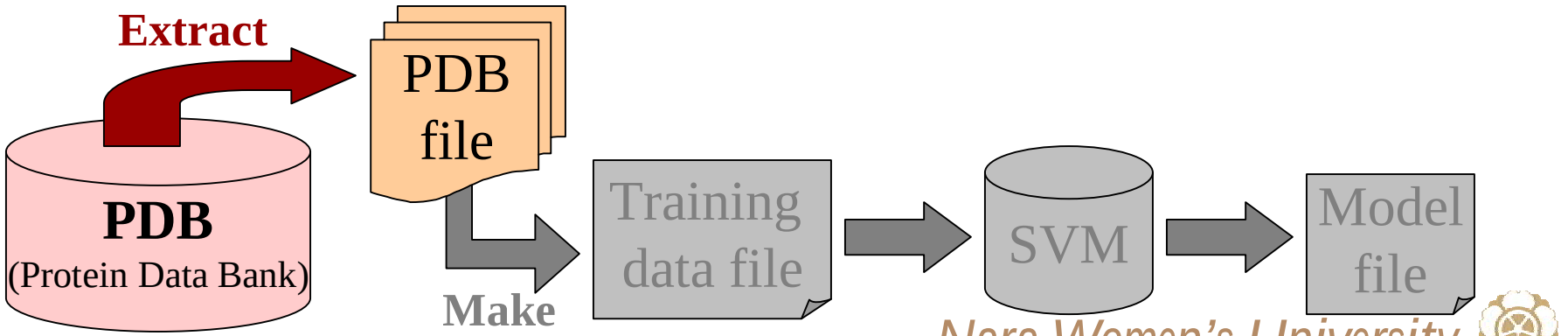
Procedure to develop a prediction system for metalbinding site

1. Extract a data file of metal protein from PDB (Protein Data Bank)
2. Make a training data file
3. Train and make a model file by using SVM (Support Vector Machine)



Procedure to develop prediction system for metalbinding site

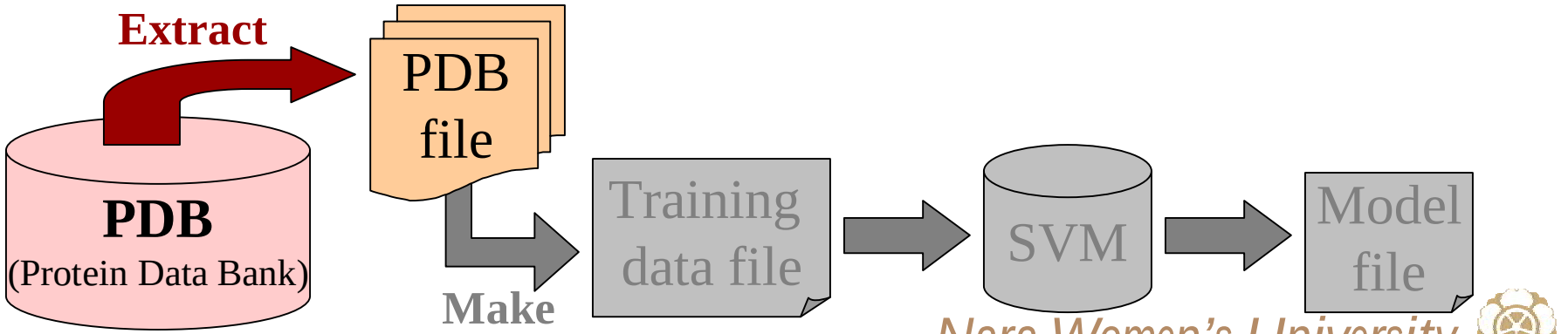
1. **Extract a PDB file of metal protein**
2. Make training data file
3. Train and make model file by using SVM (Support Vector Machine)



Extracting condition for PDB files

- Monomeric Protein
- X-ray crystallography
- No mutation
- Natural amino-acid
- Metal ion

Extracting conditions



Extracting condition (1/5)

- **Monomeric Protein**
- X-ray crystallography
- No mutation
- Only natural amino acid
- Metal ion

➢One chain
➢Not bind with nucleic acid

Extracting condition (2/5)

- Monomeric protein
- **X-ray crystallography**
- No mutation
- Only natural amino acid
- Metal ion

➢Use proteins that are determined by X-ray crystallography

➢Do not use proteins that are determined by NMR (Include several conformations)

Extracting condition (3/5)

- Monomeric protein
- X-ray crystallography
- **No mutation**
- Only natural amino acid
- Metal ion

➢Proteins that have mutation are not used.
(These proteins may be different from the wild types.)

Extracting condition (4/5)

- Monomeric protein
- X-ray crystallography
- No mutation
- **Only natural amino acid**
- Metal ion

➢Modified proteins are not used.
(These proteins are categorized as the non-wild type)

Extracting condition (5/5)

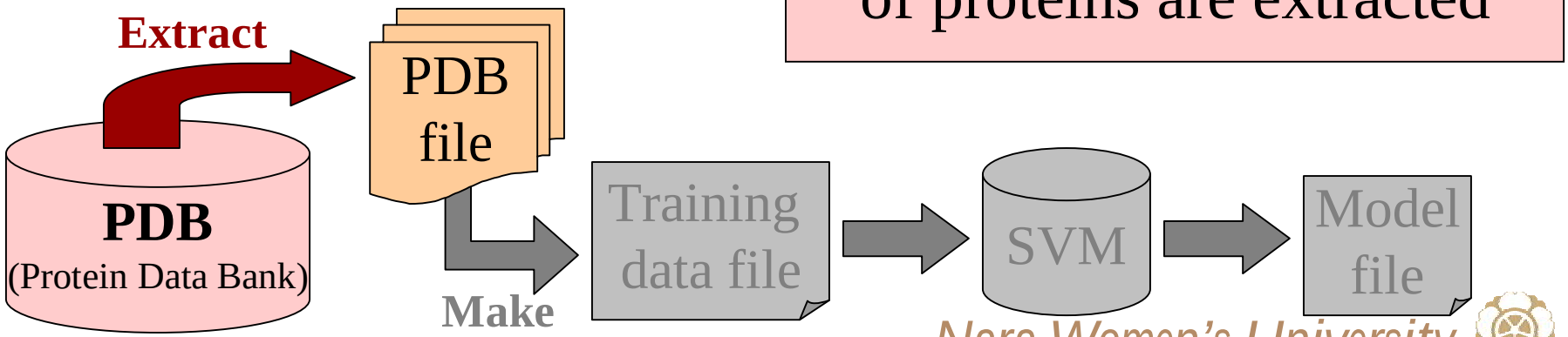
- Monomeric protein
- X-ray crystallography
- No mutation
- Only natural amino acid
- **Metal ion**

  - PDB file that has the description about metal ion in the HET line.
  - PDB file that has some molecules without water molecules or metal ions is not used.
    (The molecules may affect the protein structure)

Extracting condition for PDB files

- Monomeric Protein
- X-ray crystallography
- No mutation
- Natural amino-acid
- Metal ion

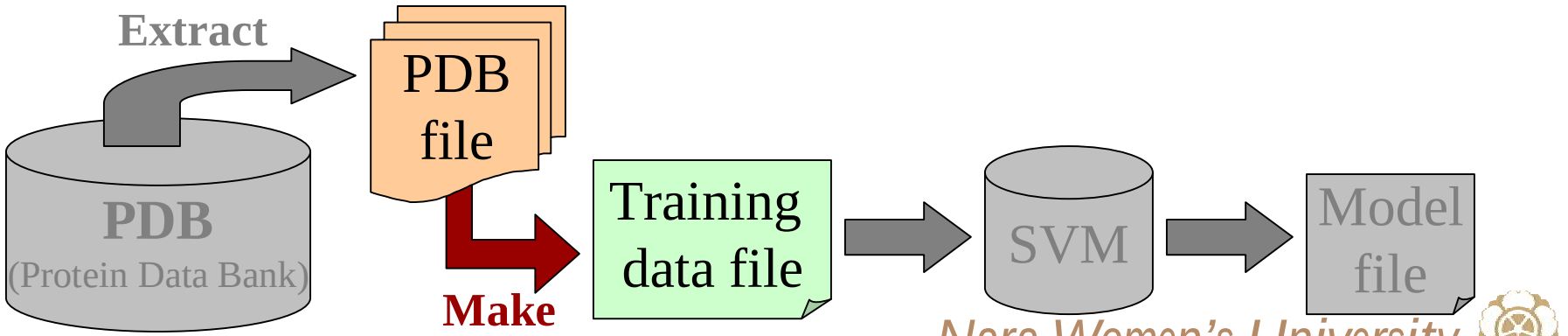
Metal proteins of which metal ion affect three-dimensional structures of proteins are extracted



Nara Women's University

Procedure to develop a prediction system for metalbinding sites

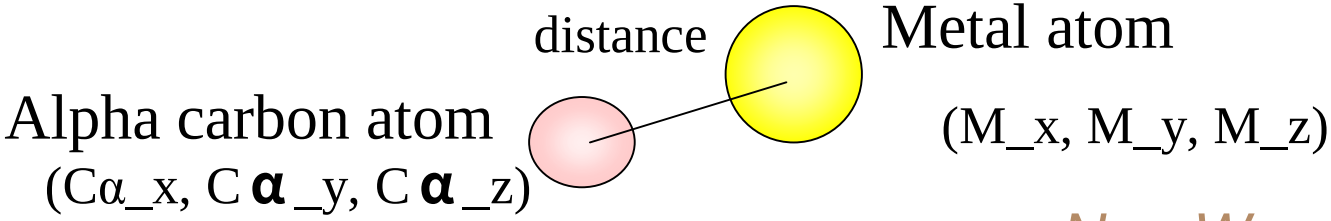
1. Extract PDB file of metal protein from PDB(Protein Data Bank)
2. **Make training data file**
3. Train and make model file by using SVM (Support vector Machine)



Making training data files(1/3)

1. Pick up coordinates of a alpha carbon atom and metal atoms
2. Calculate the distances between the metal atom and each of the alpha carbon atom

$$\text{distance}=\sqrt{(\text{M\_x}-\text{C}\boldsymbol{\alpha}\text{\_x})^2+(\text{M\_y}-\text{C}\boldsymbol{\alpha}\text{\_y})^2+(\text{M\_z}-\text{C}\boldsymbol{\alpha}\text{\_z})^2}$$

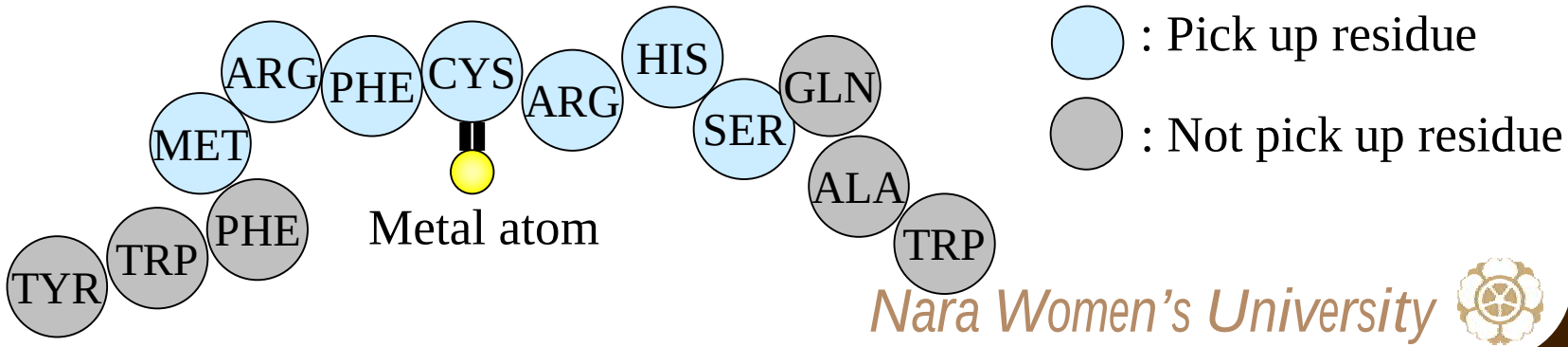


Making training data file (2)

3. Set the connection distance

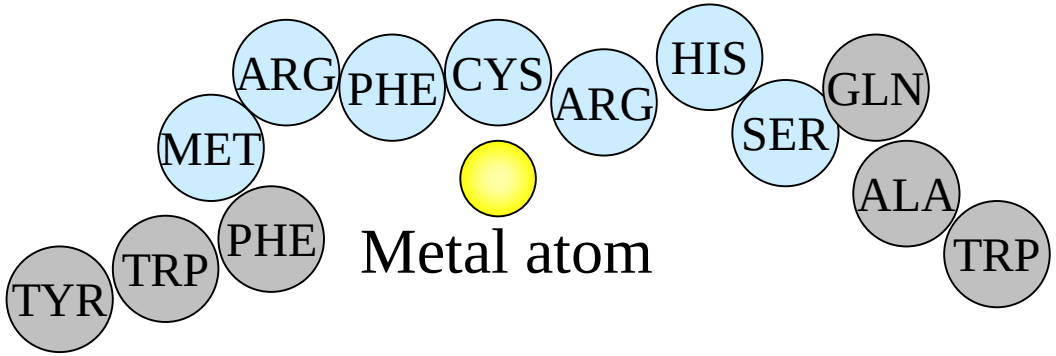
   Ex: If the distance is within 5 Å, the residue is binding

4. Pick up residue satisfying the above connection conditions and six residues abutting back and forth



Making training data file (3)

5. Count the number of seven amino acids by the kind of the amino acids



ALA:0, ARG:2, ASN:0, ASP:0, CYS:1, GLN:0, GLU:0, GLY:0, HIS:1, ILE:0, LEU:0, LYS:0, MET:1, PHE:1,PRO:0, SER:1, THR:0, TRP:0, TYR:0, VAL:0, ASX:0, GLX:0

Making training data file (4)

6. Arrange the number of residues in alphabetic order of amino acids (feature vector)

ALA:0, ARG:2, ASN:0, ASP:0, CYS:1, GLN:0, GLU:0,
GLY:0, HIS:1, ILE:0, LEU:0, LYS:0, MET:1, PHE:1, PRO:0,
SER:1, THR:0, TRP:0, TYR:0, VAL:0, ASX:0, GLX:0

Training data

1:0, 2:2, 3:0, 4:0, 5:1, 6:0, 7:0, 8:0, 9:1, 10:0,11:0,12:0,
13:1, 14:1,15:0, 16:1, 17:0, 18:0, 19:0, 20:0, 21:0, 22:0

Making training data file (5)

7. Output the formatted data to a training data file

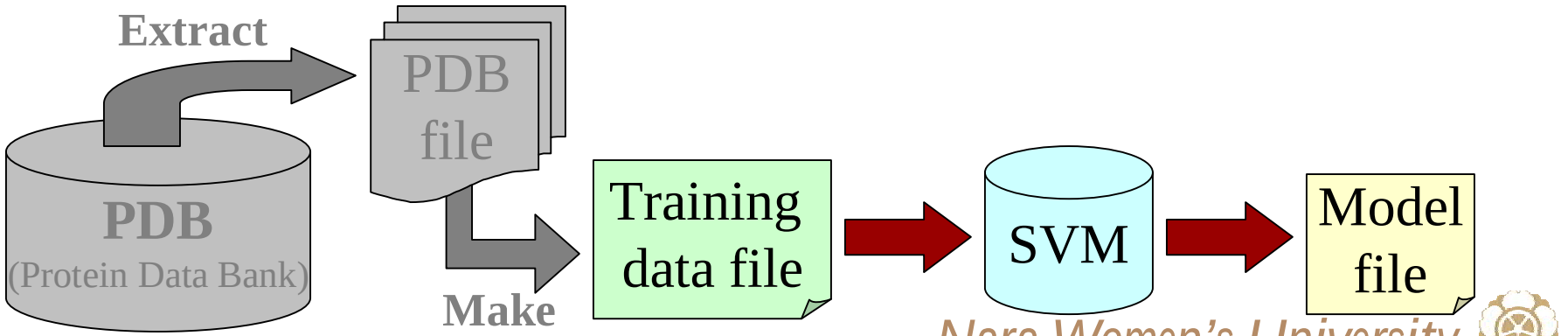
1:0, 2:2, 3:0, 4:0, 5:1, 6:0, 7:0, 8:0, 9:1, 10:0,11:0,12:0,
13:1, 14:1,15:0, 16:1, 17:0, 18:0, 19:0, 20:0, 21:0, 22:0

Training data file

1:0 2:2 3:0 4:0 5:1 6:0 7:0 8:0 9:1 10:0 …21:0 22:0
1:0 2:1 3:0 4:1 5:1 6:0 7:1 8:0 9:0 10:0 …21:0 22:1
1:0 2:1 3:0 4:0 5:2 6:0 7:0 8:1 9:1 10:1 …21:0 22:0

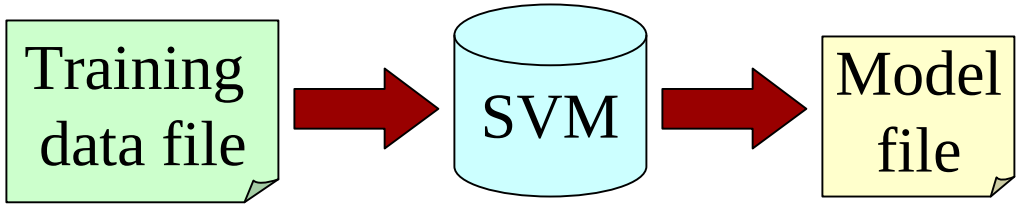
Procedure to develop a prediction system for metalbinding site

1. Extract a PDB file of metal protein
2. Make a training data file
3. **Train and make a model file by using SVM(Support vector Machine)**



Making model file

- Train
  - Use LIBSVM (A Library for Support Vector Machine)
    - High generalization capability
  - Make a model file for prediction
    - One-class SVM : binding data
    - C-SVC : binding and not binding data

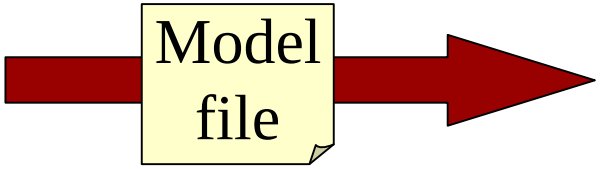


Prediction

- Predict
  - Use LIBSVM
  - Use the model file
  - Predict whether a metal ion is connected to a new amino-acid sequence

New amino acid sequence

ARG – CYS – THR – HIS –
TYP – ALA – GLY – SER –
⋮
GLN – CYS – TYP – SER

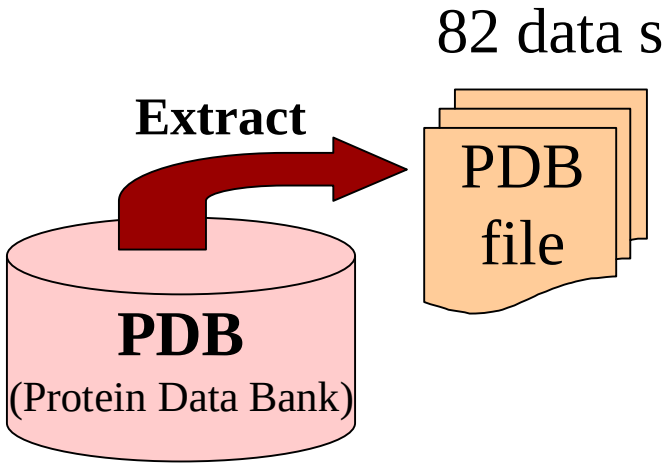


Get binding information

Zn: 23 residue
Zn: 35 residue
Zn: 78 residue

Data set for experiment

- Zn
  - ✧ Exists in the living cell
  - ✧ Most important metal ion to participate in maintenance of life function
    - – Composition of living body material, metabolism



According to the extraction condition,
82 PDB files about protein binding with Zn ion were extracted.

Recognition Rate

- Change the number of amino acid residues picked up
  - Little influence on the recognition rate
- Change the condition of connection distance
  - The stricter the condition is, the higher the recognition rate is

Recognition rate of Zn

	Zn_7	Zn_9	Zn_11
4.0 Å	85.28	85.29	85.09
5.0 Å	81.94	81.89	81.09
6.0 Å	80.82	81.48	81.86

condition of connection distance

The number of amino-acid picked up

Binding prediction accuracy

- Change the number of amino acid residues picked up
  - The more residues we pick up, the lower prediction accuracy is.
- Change the condition of connection distance
  - The stricter the definition is, the lower the prediction accuracy is.

Prediction accuracy of this system

condition of connection distance

	Zn_7	Zn_9	Zn_11
4.0 Å	54.14	53.61	48.63
5.0 Å	73.38	71.74	66.89
6.0 Å	81.15	80.12	76.45

The number of amino-acid picked up

Discussion (Recognition rate)

- Recognition rate

The stricter the condition is, the higher the recognition rate is.

→ Slack condition:
The unconnected data are used in training

Recognition rate of Zn

condition of connection distance

	Zn_7	Zn_9	Zn_11
4.0 Å	85.28	85.29	85.09
5.0 Å	81.94	81.89	81.09
6.0 Å	80.82	81.48	81.86

The number of amino-acid picked up

Discussion (Prediction accuracy)

- Prediction accuracy：

The more residue we pick up, the lower accuracy we obtain.

→ It is hard to get the feature

The stricter the definition is, the lower the accuracy is.

→ Training data is too sparse

Connection distance and the number of data

Connection distance（Å）	**4.0**	5.0	6.0
The number of data	**23**	153	318

Generalization capability

- Experiment of changing the residues in training and prediction

The number of residues in prediction

connection distance is 4 Å

	7	9	11
7	56.13	49.43	46.46
9	54.46	48.44	45.95
11	55.11	53.08	53.16

connection distance is 6 Å

	7	9	11
7	81.11	74.33	64.56
9	79.07	79.73	72.23
11	81.64	81.03	77.73

The number of residues in training

Conclusions

- Approach by a new system

Prediction of metalbinding sites in amino-acid sequence

  - ✧ Extracting datasets of metal protein from the PDB
  - ✧ Training and prediction using SVM

- Experiments

  - ✧ Recognition rate of binding with Zn ion exceeded 80%
  - ✧ Prediction accuracy of binding with Zn ion is about 70%

Future work

- Considering orientation of a side chain and distance
- Balance of sensitivity and specificity
- Apply to Cu, Ni, Fe, Mn, Co, etc. ions