Diving into DNA-Protein Interaction with Machine Learning

Jun Wang

Outline

- Background
- Introduction
  - Features
  - Models
  - Evaluation
- Current progress
  - Data
  - Naïve bayes model
- Next step

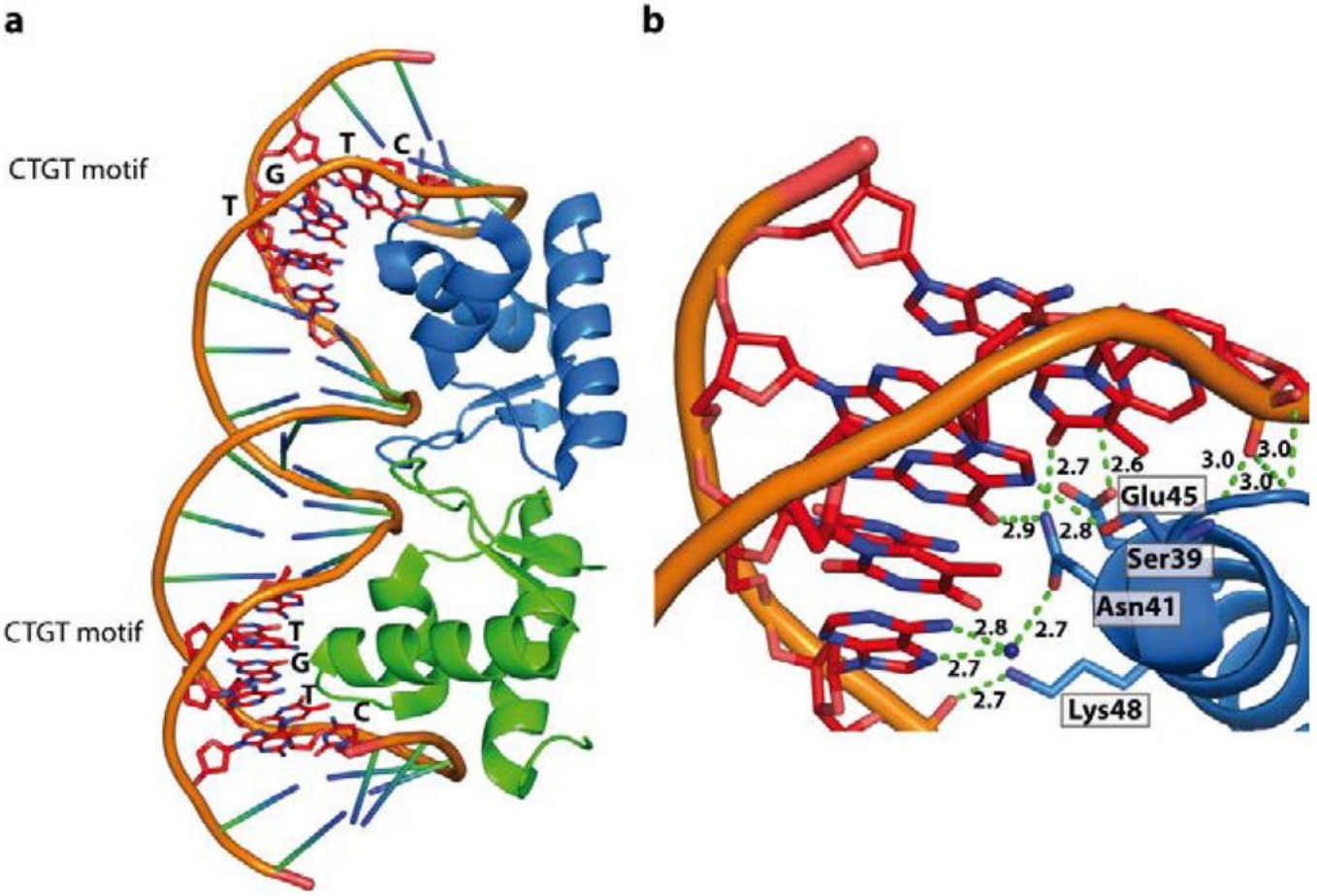
Background

- DNA-protein interaction is an important component for gene regulation system. Finding the interaction sites on protein and DNA molecules would be a critical step for drug design and researches on the metabolic network.
- Methods for investigating the binding sites:
  - Experimental identification in the lab
  - Analysis from 3-d structure of the molecules
  - Inference from some other data like residue sequence

Considering the disadvantages of the first two methods, more and more researchers focused to investigate some methods of the third way.

Ofran, Y, et al. Bioinformatics (2007)

Background



Interaction between protein and DNA

Background

- Features
  - Sequence-Based Features
  - Structural-Based Features
  - Physical and Chemical Features
- Algorithms:
  - Support vector machines
  - Artificial neural networks (ANN)
  - Bayesian learning
  - Random forest
  - Decision tree

Si, J, et al. *International Journal of Molecular Sciences* (2015)
Chu, W. Y., et al. *Nucleic Acids Research* (2009)
Tjong, H., et al. *Nucleic Acids Research* (2007)
Yan, C. Identification of interface residues involved in protein-protein and protein-DNA interactions from sequence using machine learning approaches. (2005)

Introduction: Features

- Considerable Features
  - Sequenced-based features: Sequence of residues
  - Structural-based features: Structural neighbors
  - Physical and Chemical Features: Electrostatic potential
- Feature Selection:
  - Information gain
  - Filter method
  - Wrapper method

Introduction: Models and Evaluation

- Models
  - Naïve Bayes model
  - Support Vector Machine
  - Artificial Neural Network
- Evaluation
  - Leave-one-out cross validation within each model
  - T-test on average performance between different models

Current Progress: Data

- Data component:
  - Residue sequences for a chain in different proteins
  - List of residues which are binding to DNA molecules on the corresponding chains of proteins
  - Structural information and physical properties
- Data source: NPIDB
  - PDB file (protein data bank)
  - Interaction file
  - Reference

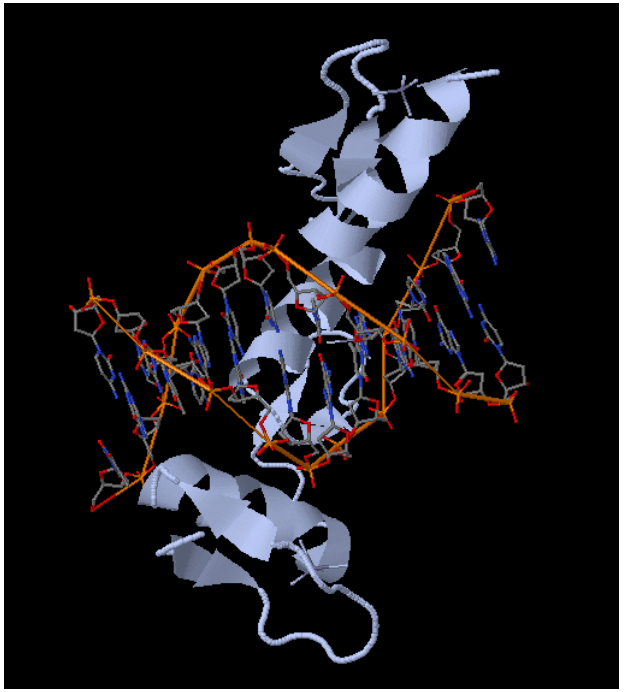
Current Progress: Data

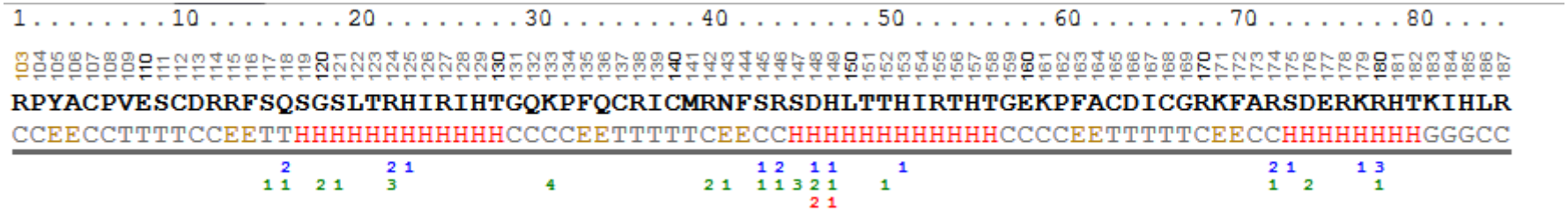
- Data preprocess
  - Residue sequence:

    Extract the sequence of residues from PDB files and convert them into a list of residue names with single letter in fasta form.

    Example:

    The first four residues extracted from 1a1h.pdb file is ” R P Y A” which represents four residues.





Current Progress: Data

- Data preprocess
  - Label for each residue:

  Download the interaction data from the database named NPIDB. We say 3.7 Angstrom is the cutoff distance to determine whether the atoms on protein and DNA has interaction and mark the label for this residue as 1, otherwise, mark the label for this residue as 0.

  Example:

  The labels of the corresponding residues in the previous example is "0 0 0 0".

Nucleic atom	dist	Protein atom
DG2:B.N7/1	2.67	ARG180:A.NH2/1
DG2:B.O6/1	2.72	ARG180:A.NH1/1
DC3:B.N4/1	3.34	ARG180:A.NH1/1
DG4:B.OP1/1	2.84	HIS153:A.ND1/1
DG4:B.N7/1	2.91	ARG174:A.NH1/1
DG4:B.O6/1	2.86	ARG174:A.NH2/1
DG6:B.OP2/1	2.77	SER145:A.OG/1
DG6:B.N7/1	2.71	HIS149:A.NE2/1
DG7:B.OP1/1	2.91	HIS125:A.ND1/1
DG7:B.N7/1	2.86	ARG146:A.NH1/1
DG7:B.O6/1	2.77	ARG146:A.NH2/1
DG8:B.N7/1	2.93	ARG124:A.NH2/1
DG8:B.O6/1	3.01	ARG124:A.NH1/1
DA10:B.N7/1	2.83	GLN118:A.NE2/1
DA10:B.N6/1	3.02	GLN118:A.OE1/1
DC55:C.N4/1	3.17	ASP148:A.OD2/1
DC56:C.OP2/1	2.54	SER175:A.OG/1
DC57:C.OP2/1	2.83	LYS179:A.NZ/1

Current Progress: Naïve Bayes Model

- Input features:

Take the 8 neighboring residues and the target residue itself (9 residues in total) as 9 features for training and predicating the labels for the target residue.

<s>	<s>	<s>	<s>	X_4	X_5	X_6	X_7	X_8
-1	-1	-1	-1	1	0	0	0	1

- Training process:

We use a threshold value θ to determine whether the label of this residue is 1 or 0:

$$\theta = \frac{P(c=1)}{P(c=0)}$$

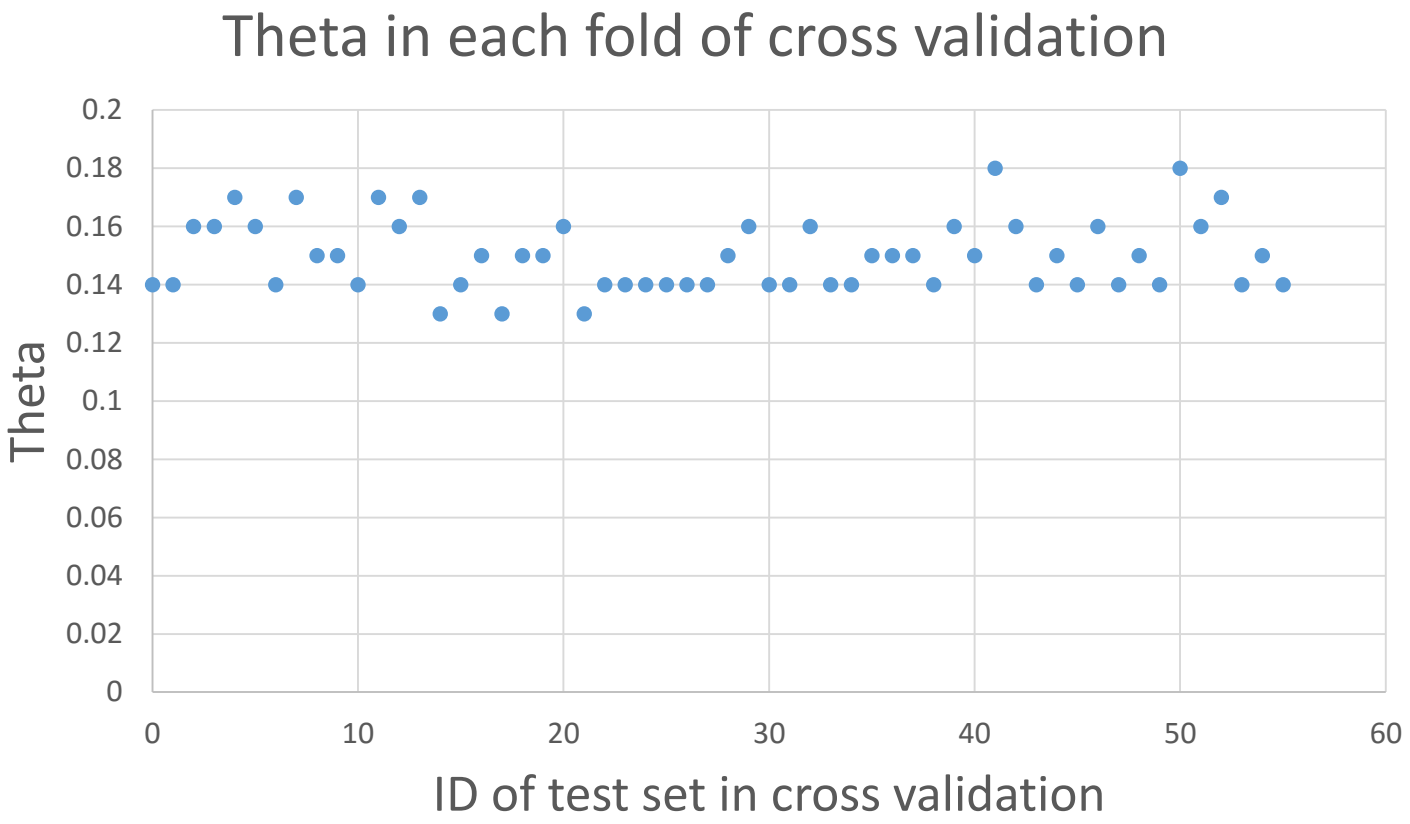
If the ratio of P(c=1) to P(c=0) is larger than θ, we will set the label to 1, otherwise, set the label to 0.

When training the model, we initialize θ as 0.01 and increase it until 1 with stepsize = 0.01. After the iteration on θ, take the value with the highest correlation coefficient as the training result.

<s>	<s>	<s>	<s>	X_4	X_5	X_6	X_7	X_8
-1	-1	-1	-1	1	0	0	0	1

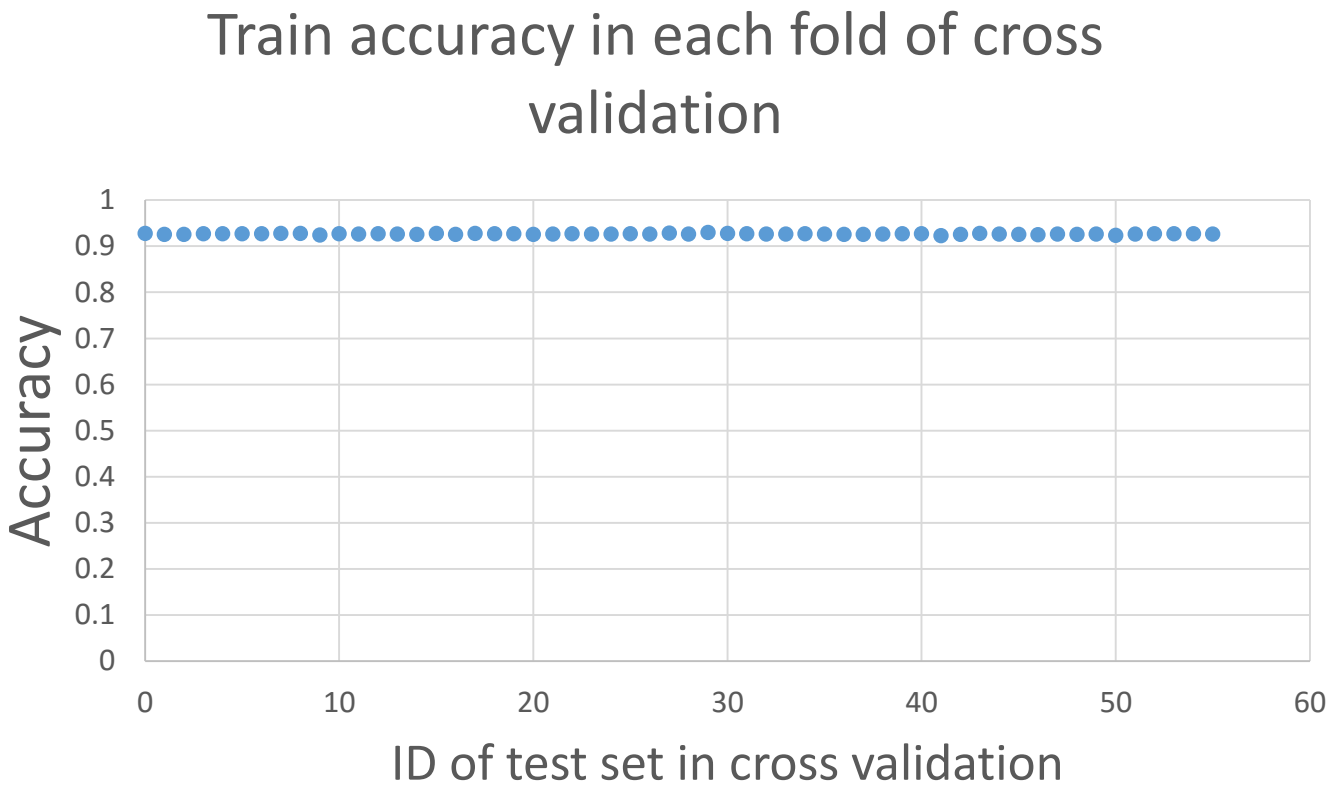
X1	X2	X3	X4	X_4	</s>	</s>	</s>	</s>
0	1	0	0	1	-1	-1	-1	-1

Current Progress: Naïve Bayes Model



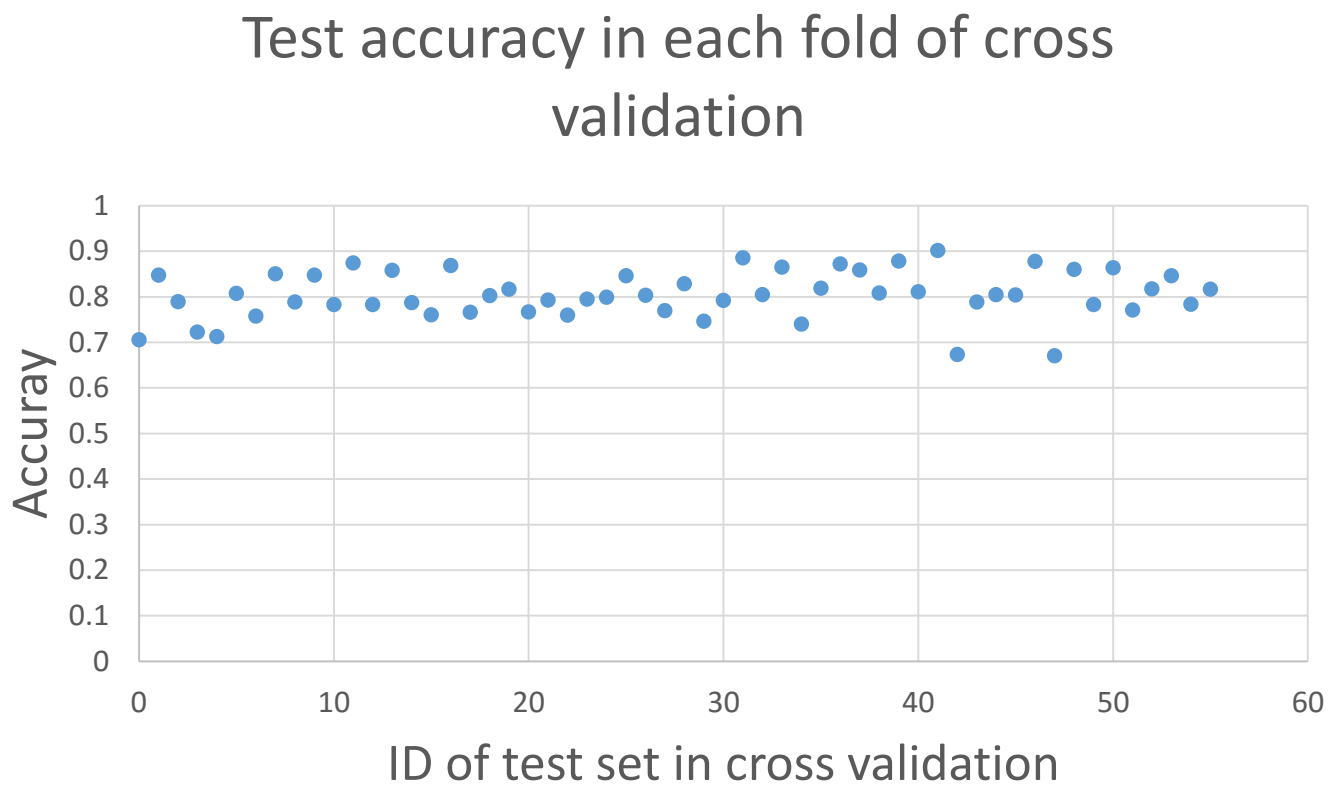
Theta value

Current Progress: Naïve Bayes Model



Traing accuracy

Current Progress: Naïve Bayes Model



Test accuracy

Current Progress: Naïve Bayes Model

- Implementation Result:

  The average test accuracy is: 80.4%, the standard deviation is 0.052.

- Comparison with the reference:

  The average test accuracy on the reference is around 77% when using the 9 neighboring residues as the features

- Causes for the difference between these two results:
  - Using different standards to classify the residues into two classes: binding residues and non-binding residues
  - Taking the residue sequences from different chains when analyzing the same proteins.
  - Arithmetic calculation details might also affect the result slightly.

Next step

- Implementation of other models:
  - Support Vector Machine
  - Artificial Neural Network
- Add more features
  - Considering the electrostatic potential for residues
  - Structural Neighbors
- Filtering the features
  - Information gain
  - Filter method
  - Wrapper method

Thank you!