Diving into DNA-Protein Interaction with Machine Learning

comp135

Jun Wang

Introduction

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 (Vert et al, 2013). Nowadays, there are many different ways to investigate the binding sites, including experimental identification from lab, analysis from 3-d structure of the molecules, or inferrence from some known data such as amino acid sequences(Ofran et al, 2007).

However, there are different disadvantages for these methods. For example, experimental identification is very expensive and relatively slow in practice. Especially, when dealing with a large amount of proteins and DNA molecules, it is not a good choice in fact. Analysis with the 3-d structure is a good way to determine the DNA-protein binding residues. But until now, it is still not easy to obtain the accurate 3-d structure of molecules for most proteins. Thus, inferring the DNA-protein binding sites from the protein sequences is a relatively more promising way, especially with the development of the machine learning theory.

Dataset

A small datasets composed by 56 double-stranded DNA binding protein complexes used in previous researches (Yan, 2005), will be extracted from the Nucleic Acid Database (NDB) (Berman et al, 1992). This dataset is derived from 427 DNA-protein complexes with a resolution better than $3\AA$. After grouping these 427 complexes into homologous families, take a representative protein with best resolution from each group to form our dataset(Jones et al, 2003).

Experiments

Models Selection

Considering that support vector machine, naiive bayes' algorithm, and the neural network are all conventional algorithms for supervised learning for classification in machine learning, it is rational to implement these three algorithms for our investigation. Especially, these three algorithms has already been applied in many different areas in computational biology. For example, support vector machine has been used for inferring the DNA-protein binding sites (Niu et al, 2014), naiive bayes classifier has been implemented for identification of DNA-protein interfaces(Yan, 2005), and the neural networks are also used for determining the binding sites between RNA and protein molcules (Terribilini et al, 2006). Thus, support vector mahine (SVM), naiive bayes classifier, and the neural networks will be implemented for inferring the binding sites of the DNA-protein complexes.

Classess

Identification of DNA-protein binding sites can be formulated as a sequence labelling problem (Vert et al, 2013). To be specific, since the experiment only concerns about whether the residue in the protein sequence can be bound to DNA molecules, it will be a binary classification for all the residues in all the protein molecules: Yes or No. For all the residues within Yes class, it means that these residue are probabily the binding sites for DNA-protein complexes.

Feature selection

Previous research has indicated that it is rational to inferring the DNA-protein binding sites according to the protein sequence alone (Ofran et al, 2007), so it is reasonable to inferring the binding sites from the protein sequence in our dataset. In practice, all the residues will be considered individually for machine learning. For example, in Naiive Bayes Classifier, the neighbouring residues near the target residues will be checked for calculating the probability of whether the target residue can be the binding sites to the DNA molecules.

Model assessment

In this project, a 10-fold cross validation will be applied for assessing the accuracy of the models.

Model improvements

In previous researches, some researcher also indicated that it is rational to utilize the physical and chemical property for determining the binding sites of DNA-protein complexes as well. For example, Jones proposed to use electrostatic potentials to predict the DNA-binding sites(Jones et al, 2003), and some other researchers trained the neural network with the physical and chemical property to determine which residues are the binding sites in the DNA-protein complexes(Keil et al, 2004). Thus, it is reasonable, to combine the amino acid sequences and some other properties like electrostatic potentials together to improve the accuracy of predicting the DNA-protein binding sites.

Reference

[1] Vert, J. (2013). Machine Learning in Computational Biology (the frequentist approach), (September).
[2] Ofran, Y., Mysore, V., & Rost, B. (2007). Prediction of DNA-binding residues from sequence. Bioinformatics, 23(13), 347–353.
[3] Yan, C. (2005). Identification of interface residues involved in protein-protein and protein-DNA interactions from sequence using machine learning approaches.
[4] Jones, S., Shanahan, H. P., Berman, H. M., & Thornton, J. M. (2003). Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins.

Nucleic Acids Research, 31(24), 7189–7198.

[5] Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., … Schneider, B. (1992). The nucleic-acid database - a comprehensive relational database of 3-dimensional structures of nucleic-acids. Biophysical Journal, 63(3), 751–759.

[6] Terribilini, M., Lee, J., Yan, C., Jernigan, R. L., Honavar, V., & Dobbs, D. (2006). Prediction of RNA binding sites in proteins from amino acid sequence Prediction of RNA binding sites in proteins from amino acid sequence. Bioinformatics, 1450–1462.

[7] Niu, X.-H., Hu, X.-H., Shi, F., & Xia, J.-B. (2014). Predicting DNA binding proteins using support vector machine with hybrid fractal features. Journal of Theoretical Biology, 343, 186–92.

[8] Keil, M., Exnep, T. E., & Brickmann, J. (2004). Pattern recognition strategies for molecular surfaces: III. Binding site prediction with a neural network. Journal of Computational Chemistry, 25(6), 779–789.