# Active Learning For DBSI: DNA Binding Site Identifier

Max Horowitz-Gelb

December 11, 2016

## Abstract

DBSI is a structure based method for predicting the positions of protein interaction sites associated with DNA binding [3]. Here I present a method for applying active learning to the training of DBSI. This method optimizes the training of DBSI in a way that considers the batch style form of labelled data collection necessary when creating a training set. This method shows slight improvements in efficiency in comparison to naive methods.

## Introduction

For area under an ROC curve, DBSI has been shown to achieve 88%, a high degree of separability. The score was achieved by training DBSI on a set of 263 unique proteins. The model as a result of this training is now accessible to anyone on a public server [2]. The quality of this model could be improved further by training with more labelled data.

## Methods

[1]

## Results

## Conclusion

## References

[1] Brinker K. Incorporating diversity in active learning with support vector machines. In *In Proceedings of the 20th International Conference on Machine Learning*, pages 59–66. AAAI Press, 2003.

[2] Sukamar S, Zhu X, Ericksen SS, and Mitchell JC. Dbsi server: Dna binding site identifier. *Bioinformatics*, 2016.

[3] Zhu X, Ericksen E., and Mitchell JC. Dbsi: Dna-binding site identifier. *Nucleic Acids Research*, 2013.