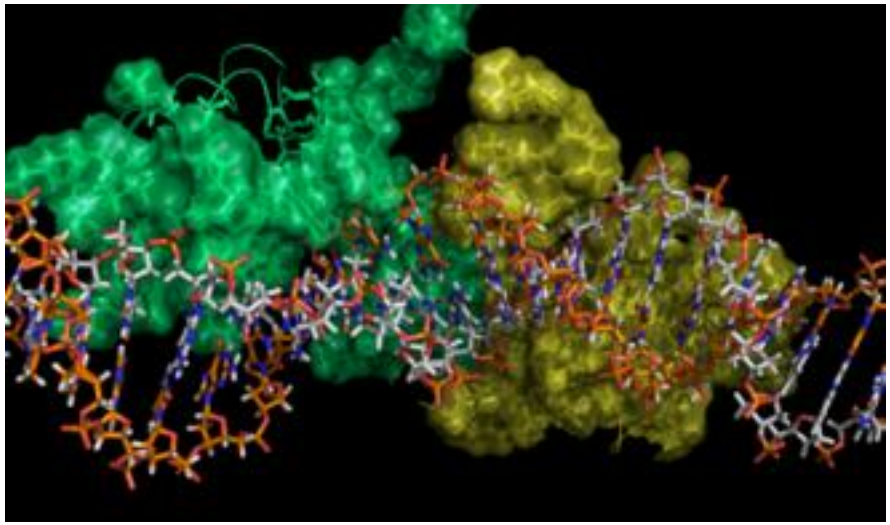
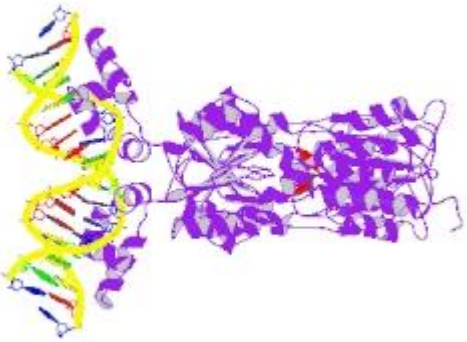
Machine Learning Study of DNA Binding by Transcription Factors from the LacI family

Gennady Fedonin

Transcription factors





- Specifically bind DNA to control transcription
- Contain one or more DNA-binding domains (DBDs).

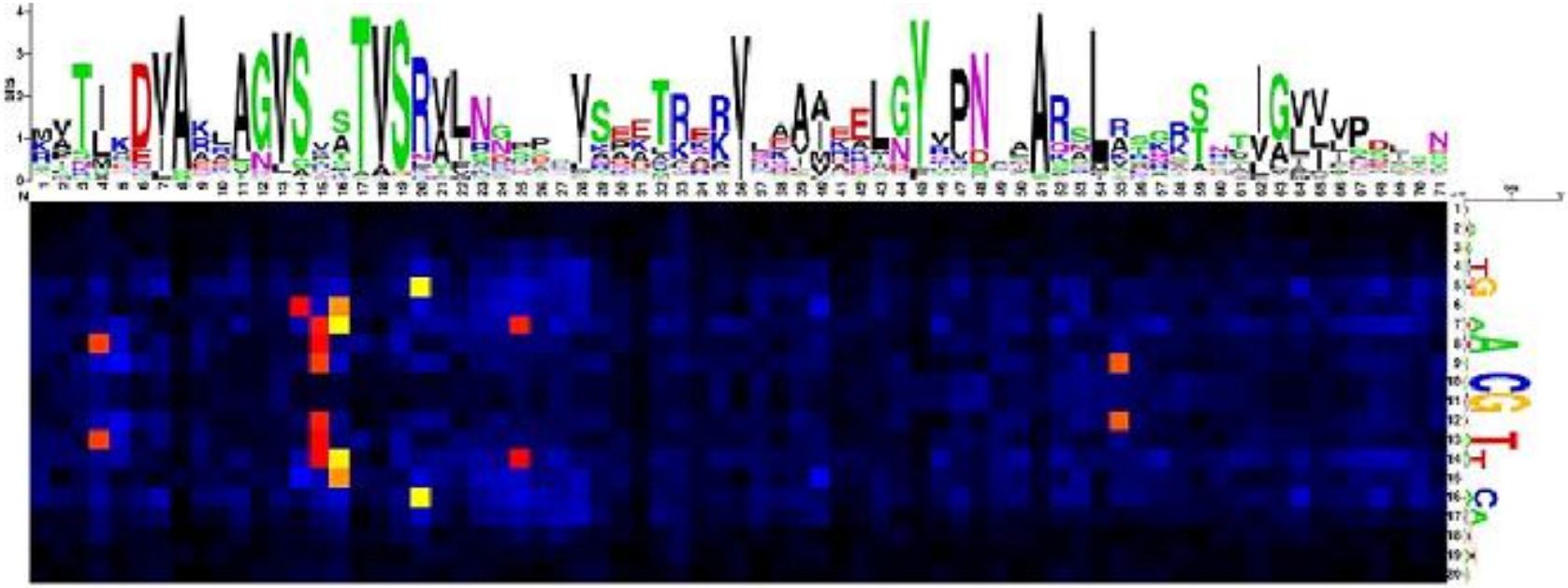
Data

- LacI-family bacterial transcription factor and their binding sites (LACI_DB).
- Boundaries of DNA-binding domain (HTH_LACI) defined by SMART
- Sequence alignment against the standard HTH_LACI domain alignment.

Data

- 1372 transcription factors (TF)
- 4484 TF-binding site pairs
- DNA binding domain length = 87 positions
- Binding site length = 20 positions (centrally symmetric)

Early studies



Correlations seem to be not limited to pairs of positions!

Problem statement

We have:

a sample of protein-site pairs.

The task:

to predict the probability density of nucleotides at site positions given a protein amino acid sequence (AAS).

Algorithms

Classifiers:

- Naïve Bayes (NB)
- Logistic Regression (LR)

Feature selection:

- Forward selection based on Mutual Information.
- Greedy forward selection with NB and LR.

Cross validation

Log-likelihood was defined as:

$$\log L = \frac{\sum_i w_i \sum_j \log P(n_{ij} \mid S_i)}{\sum_i w_i}$$

Cross validation

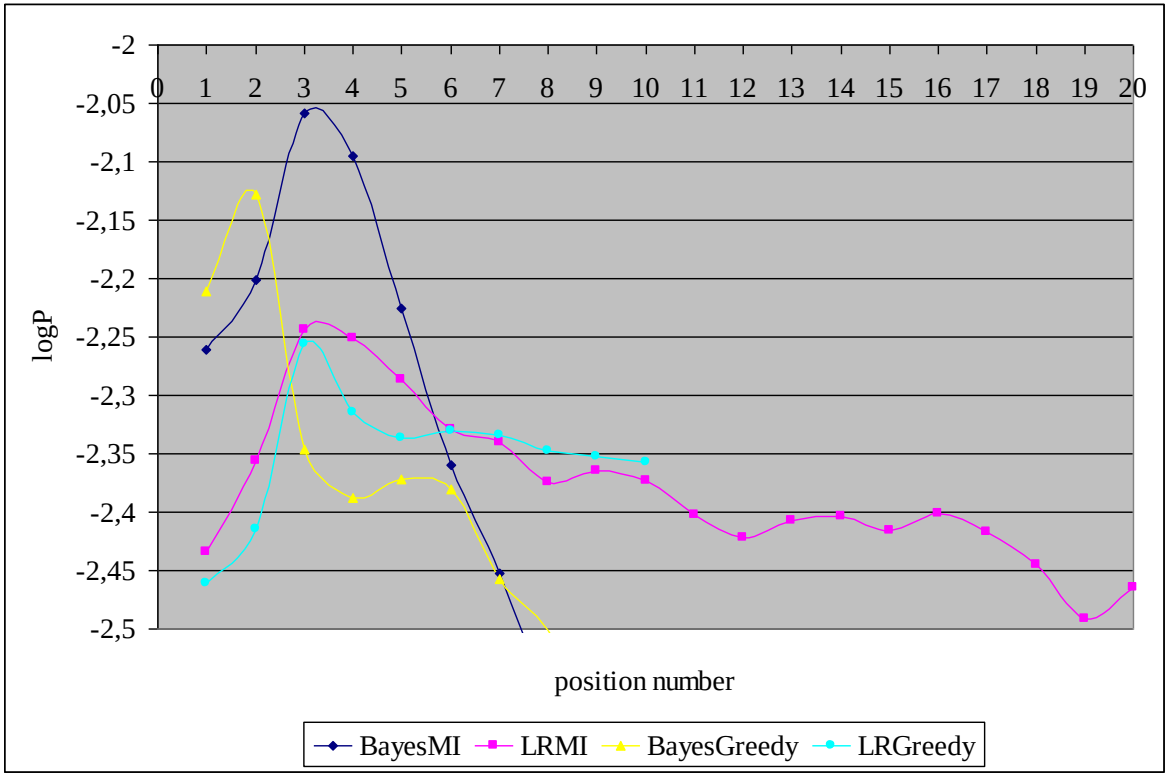
Care taken to avoid very similar TFs in training and testing.

10 fold CV on similarity clusters:

- Pairwise similarity (SV) for all AAS pairs was calculated.
- Graph with AASs as vertices and edges weighted with the SV was built.
- All edges with weight less than a fixed threshold were removed.

The similarity clusters were defined as maximal connected components.

Example.
Position 7. Prediction quality



Well-defined maxima are obtained when three positions are used (3 of 4 algorithms).

Position 7. Selected positions

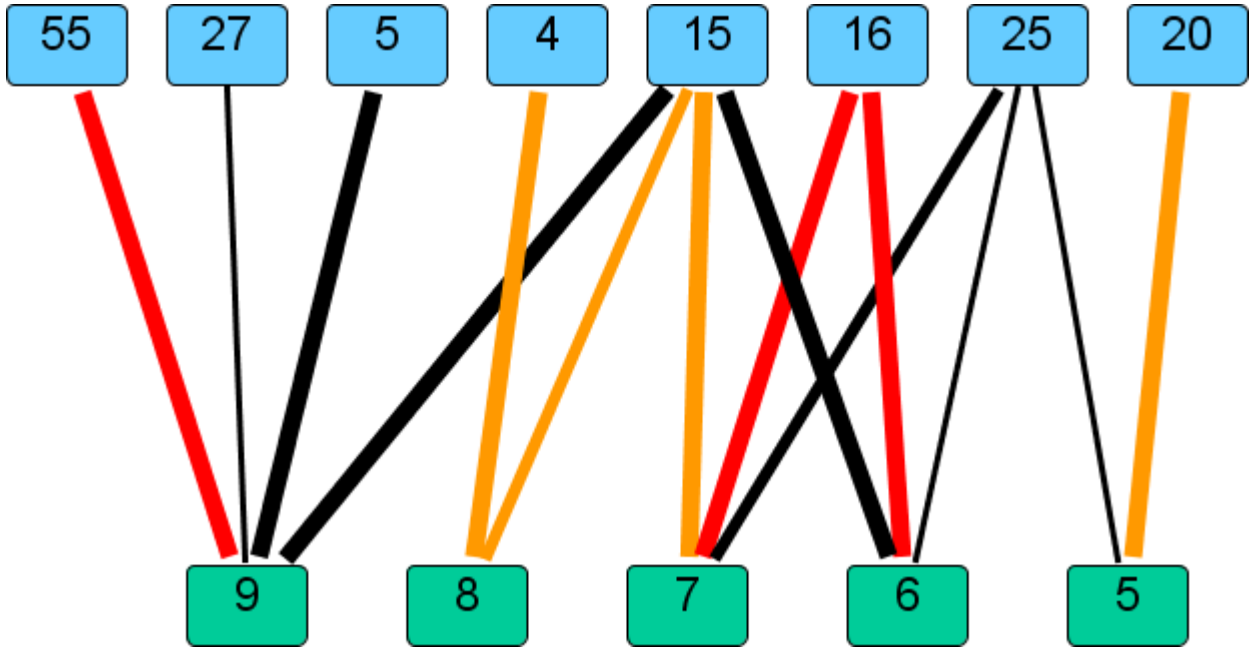
	MI						Bayes						LR					
	16	25	15	68	5	46	16	15	49	68	50	19	16	15	25	49	68	50
1	**1**	0	0	0	0	0	**1**	0	0	0	0	0	**1**	0	0	0	0	0
2	1	**0,96**	0,04	0	0	0	1	**0,97**	0	0,02	0	0	1	**0,69**	0,3	0	0	0
3	1	1	**1**	0	0	0	1	0,97	**0,71**	0,11	0,09	0,02	1	0,99	**0,99**	0	0	0
4	1	1	1	**0,84**	0,05	0,03	1	0,98	0,89	**0,59**	0,33	0,07	1	1	1	**0,56**	0,2	0,05
5	1	1	1	0,94	**0,25**	0,18	1	0,98	0,92	0,94	**0,75**	0,12	1	1	1	0,78	**0,57**	0,16
6	1	1	1	0,97	0,38	**0,46**	1	0,99	0,93	1	0,86	**0,64**	1	1	1	0,89	0,84	**0,42**

The MI-based search stably selects three positions 16, 25 and 15. The greedy LR stably selects the same three positions, whereas the greedy Bayes classifier selects wrong position at the third step.

Significance criteria for positions

- Selection stability.
- Existence of well-defined maxima on the log-likelihood plots.

Significant positions



On top – AAS positions, on bottom – site positions. Line width shows position significance. Red color means that connected positions form specific contacts in all three known protein-DNA complex structures (according WHATIF), orange – in two of three, black – in no one.

Conclusions

- Few key AAS positions are sufficient for the prediction of nucleotide densities. These positions form significantly correlated pairs with corresponding site alignment positions, having high mutual information values.
- Proposed method can be used to predict pairs of positions, which form specific contacts in protein-DNA complex structures.

Acknowledgements

- O. Laikova for data
- Y. Korostelev for the heat map slide
- M. Gelfand for advice and supervision
- Russian Fund of Basic Research for funding