Summary of method

Bobcat Bioinformatians, Ohio University

Please contact Yichao Li, yl079811@ohio.edu, and Lonnie Welch, welch@ohio.edu if you have any questions.

A good machine learning model is primarily depending on the number of training examples and the relevance of features. Thus, we used a non-parametric algorithm, random forest, to train the models because we want to focus on increasing the training samples and selecting better features.

Table 1. Features in the final model

DNA Motif (120+40)		Shape.HelT(2)		Shape.Roll(2)		Shape.MGW(2)		Shape.ProT(2)		Open Chromatin (2)		Expression PC (3)		Nearest EXP	TF regulators EXP		Engineered features		Label
9.90909	11.7273	32.2203	0.16	50.8698	0.3	88.0262	0.44	110.693	0.6	108.981	0.54491	-5.95571	18.146	32.83	14.7	7.79	0.035	2.428028	0
0	0	50.996	0.25	51.7692	0.3	89.9228	0.45	53.2472	0.3	2.80704	0.01404	-5.95571	18.146	0	14.7	7.79	0.445	0.62858	0
0	0	51.7833	0.26	51.6979	0.3	86.2778	0.43	55.2131	0.3	44.1414	0.22071	69.9107	9.0224	0.015	10.72	22.46	8.6	0.717418	0
0	0	56.5515	0.28	46.722	0.2	77.9534	0.39	52.1668	0.3	69.1549	0.34577	69.9107	9.0224	5.045	3.7	4.38	5.34	0.749065	0
0	0	46.3174	0.23	44.4924	0.2	77.3652	0.39	83.6261	0.4	123.11	0.61555	-2.94641	-19.69	0	390.97	318.01	0.05	0.420277	0
0	0	0	0	0	0	0	0	0	0	0	0	-3.62664	-36.81	0	14.7	7.79	9.335	31.13	0
0	0	53.6317	0.27	46.332	0.2	77.0487	0.39	62.862	0.3	187.184	0.93592	-3.62664	-36.81	0	390.97	318.01	38.275	9.33	0
0	10.9091	48.772	0.24	48.0832	0.2	81.3489	0.41	71.5114	0.4	95.0321	0.47516	-6.27161	-2.07	3.7	390.97	318.01	37.56	17.395	0
0	0	52.7229	0.26	49.9124	0.2	83.3367	0.42	52.2135	0.3	0	0	-3.62664	-36.81	390.97	14.7	7.79	16.085	28.515	0
0	0	47.1393	0.24	51.4341	0.3	88.734	0.44	72.6849	0.4	1043.31	5.21657	-2.94641	-19.69	14.7	14.7	7.79	0.709204	0.586809	1

As shown in Table 1, we have used 5 TF binding motif databases and 1 epigenetically derived motif database. We used a random forest feature importance ranking method to select top 120 motifs from TF-motifs and top 40 motifs from Epigenetic-motifs. We also add the sum and mean signals of 4 types of DNA shape and Dnase fold-enrichment signal using bigWigAverageOverBed.

The PCA algorithm is applied to the 14 cell types gene expression data. And the first 3 PCs (average value since two replicates per cell type) were included in the model. We also calculated the nearest gene expression level and the given TF regulators expression level.

Finally, we engineered 10 features to add some non-linear relationship; they are: mean motif scores multiply mean signal of dnase, mean signal of 4 types of DNA shape, first 3 PCs, nearest gene expression level, and average expression level of given TF-regulators.

Highlights. We found our model is particularly good on CTCF, EGR1 and SPI1.