ENGINE: an Enhancer-Gene Interaction dEtection method using robust feature extraction.

Part2: Tuning and feature selection

Lou Shaoke

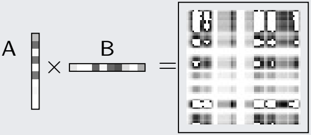
Department of Molecular Biophysics and Biochemistry

loushaoke@gmail.com

November 18, 2015

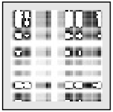
Flowchart

Flowchart



408 positive set:K562 ChIA-PET intersect with MIT mix-membership prediction
408 negative set:MCF7 specific ChIA-PET interactions

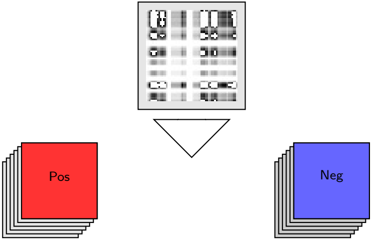
Data transformation



Flowchart

SURF: Speeded Up Robust Features, merits:

- Scale and image rotation invariant detectors and descriptors.
- blob detection
- ...

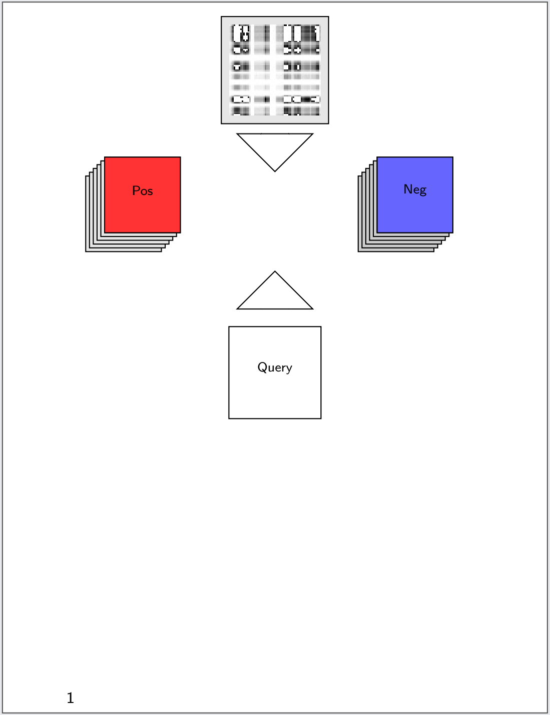


Flowchart

Feature S_i in $N_{j,k}$ matrix (feature sets), and recognition matrix

$$R_{i,j} = \begin{cases} 1, & \text{if } s_i = n_j \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

The enrichment score:
$ES(i) = -\sum log(\frac{\sum_j R_{i,j}}{N}) - log(\frac{\sum_j \sum_k 1\{s_i = n_j\}}{\sum_j \sum_k 1})$.
The relative enrichment score
$RS = ES(positive) - ES(negative)$.
The lower of RS, the better!
use RandomForest to do classification.



Flowchart

Feature S_i in $N_{j,k}$ matrix (feature sets), and recognition matrix

$$R_{i,j} = \begin{cases} 1, & \text{if } s_i = n_j \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

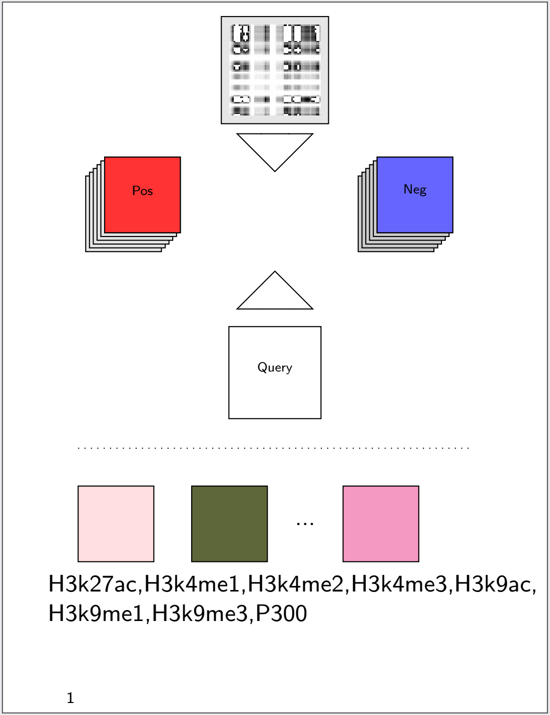
The enrichment score:

$$ES(i) = -\sum log(\frac{\sum_j R_{i,j}}{N}) - log(\frac{\sum_j \sum_k 1\{s_i=n_j\}}{\sum_j \sum_k 1}).$$

The relative enrichment score $RS = ES(positive) - ES(negative)$.

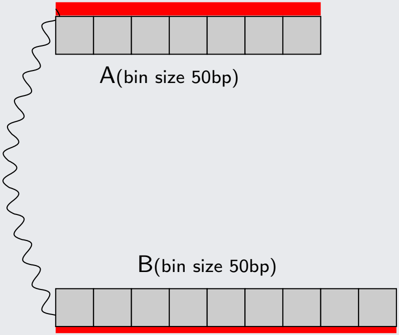
The lower of RS, the better!

use RandomForest to do classification.

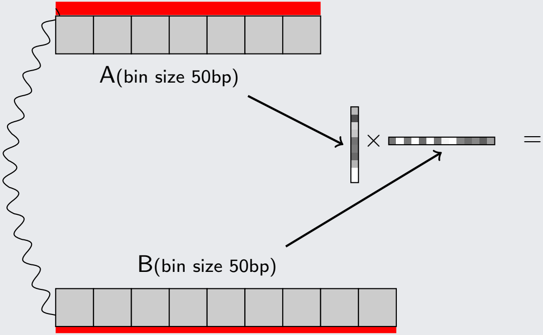


H3k27ac,H3k4me1,H3k4me2,H3k4me3,H3k9ac,
H3k9me1,H3k9me3,P300

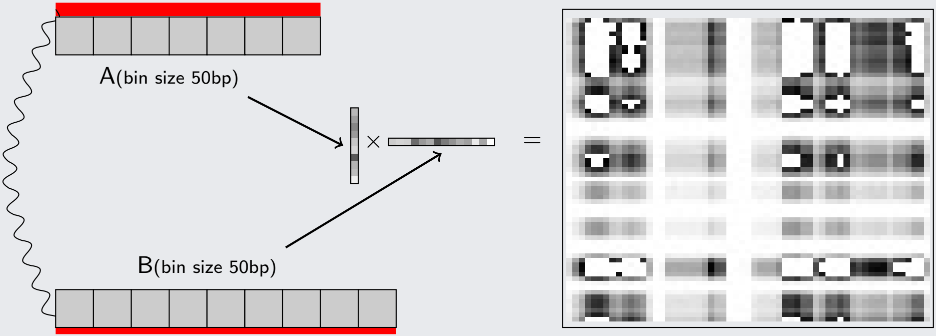
Psuedo Image transformation



Psuedo Image transformation



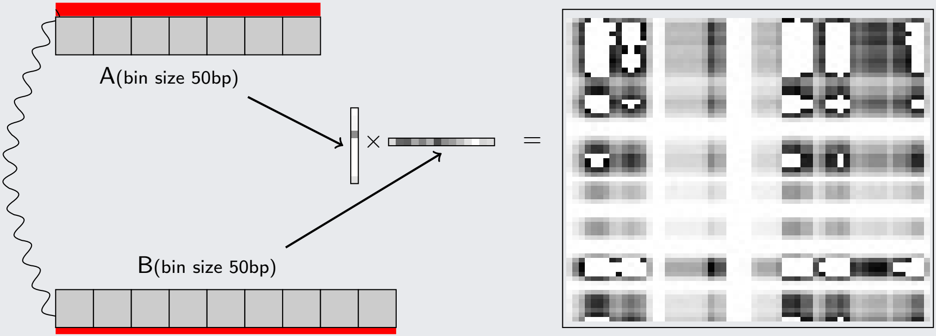
Psuedo Image transformation



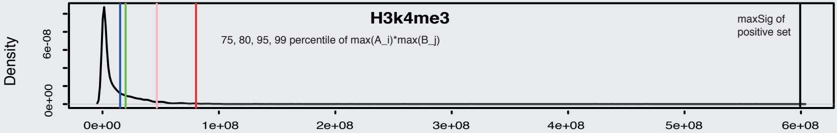
A(bin size 50bp)

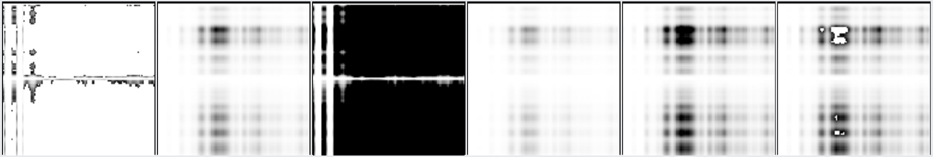
B(bin size 50bp)

Psuedo Image transformation



The range of signal is in [min(A)*min(B), max(A)*max(B)], then convert to grayscale psuedo image: integer in [0, 255].





1.5max(A,B) F

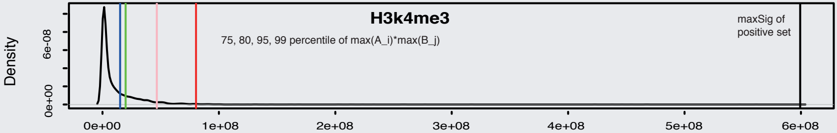
$maxA$, $maxB$

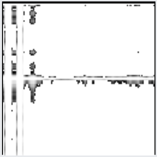
1.5$max(A, B)C$

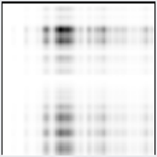
99 perc

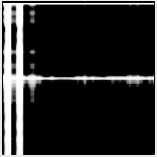
75 perc C

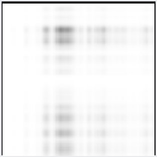
75 perc F

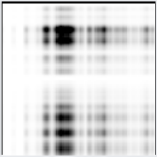


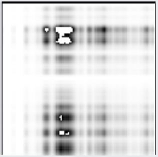

1.5max(A,B) F


$maxA$, $maxB$

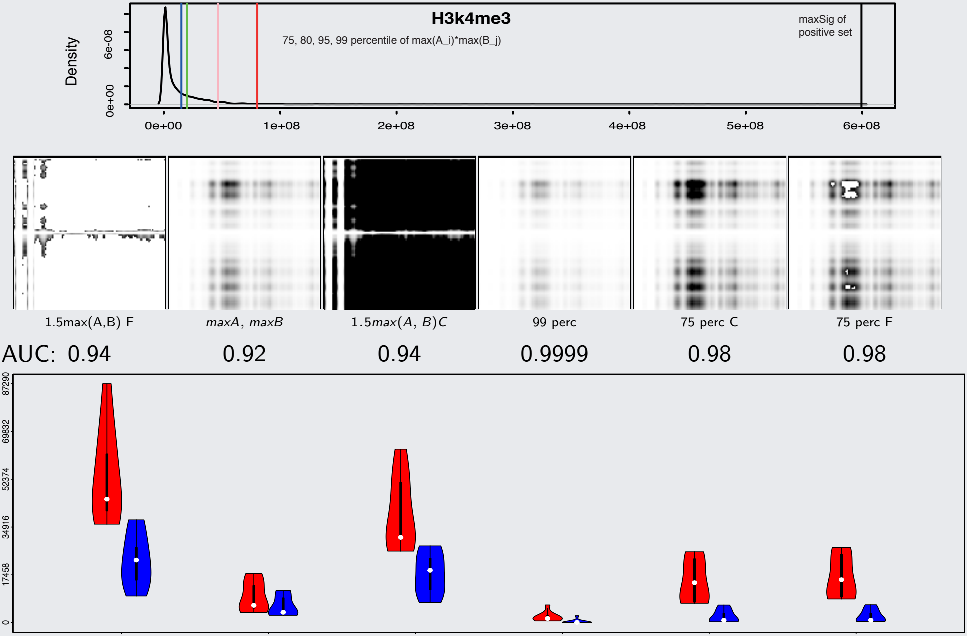

1.5$max(A, B)C$

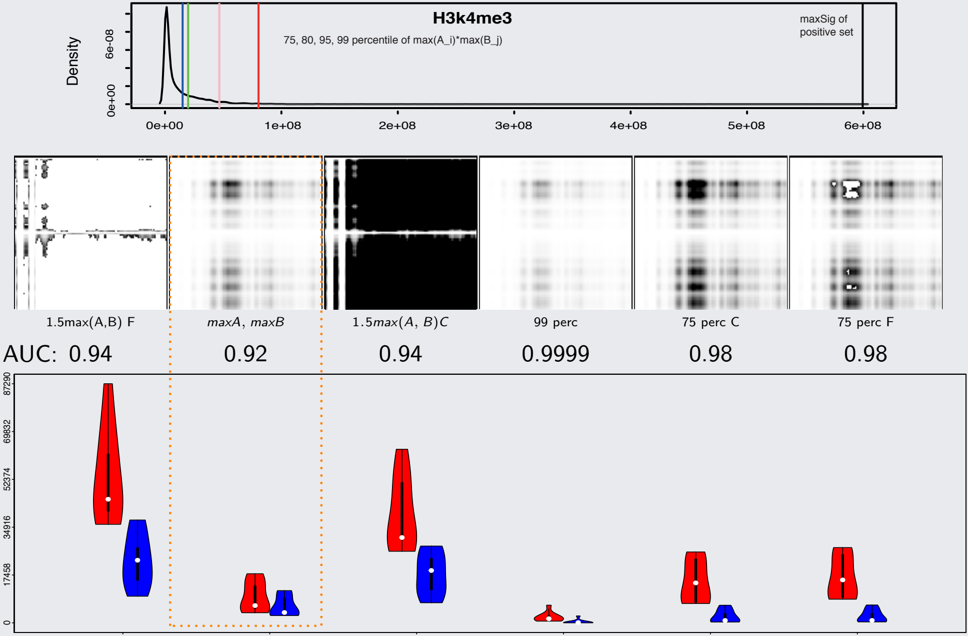

99 perc


75 perc C


75 perc F

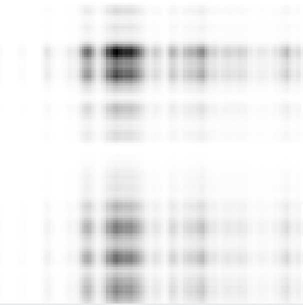
AUC: 0.94 0.92 0.94 0.9999 0.98 0.98


H3k4me3
75, 80, 95, 99 percentile of max(A_i)*max(B_j)
maxSig of
positive set
Density
0e+00
6e-08
0e+00
1e+08
2e+08
3e+08
4e+08
5e+08
6e+08
1.5max(A,B) F
maxA, maxB
1.5max(A, B)C
99 perc
75 perc C
75 perc F
AUC: 0.94
0.92
0.94
0.9999
0.98
0.98
0
17458
34916
52374
69832
87290



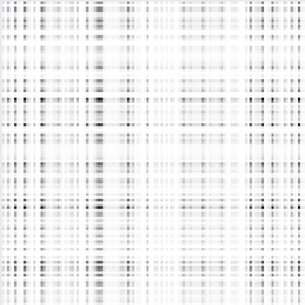
Heterogeneity; saturation affect feature detection; positive set have relative high signal

Additional negative dataset test



Orignal negative dataset

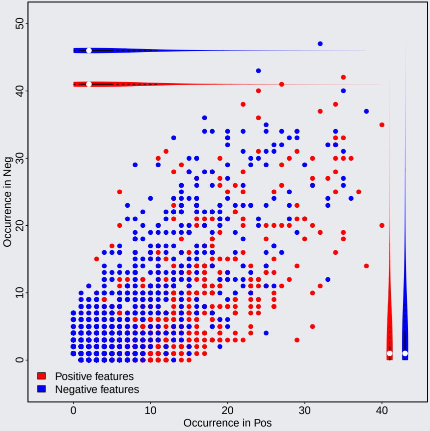
Random shift region



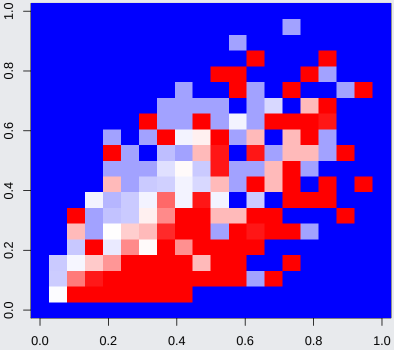
Random signal

AUC	0.92	0.93	0.93

Feature selection

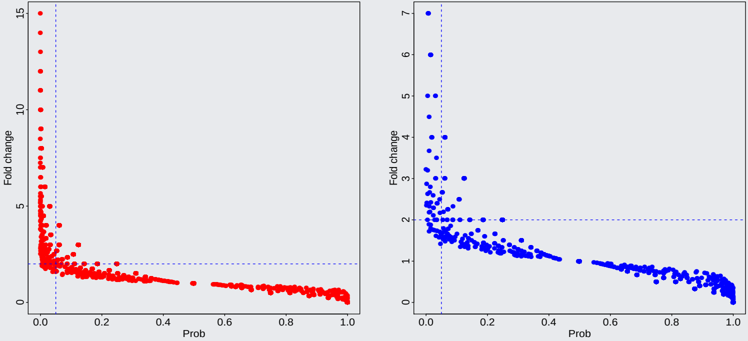


Feature distribution



$P(pos|neg)$ conditional density

Feature selection

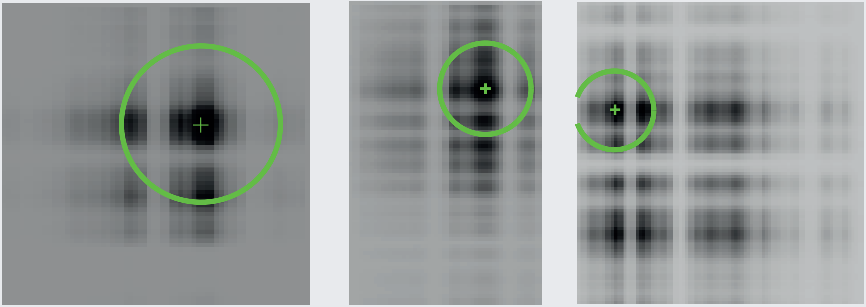


pvalue$(= \sum(dhyper(pos_hit : total_hit, \#pos_sample, \#neg_sample, total_hit))) < 0.05$ and FC$>$2,
#pos_features in each marker:

H3k27ac	H3k4me1	H3k4me2	H3k4me3	H3k9ac	H3k9me1	H3k9me3	P300	nCpG
395	835	742	462	400	1427	2110	672	1228

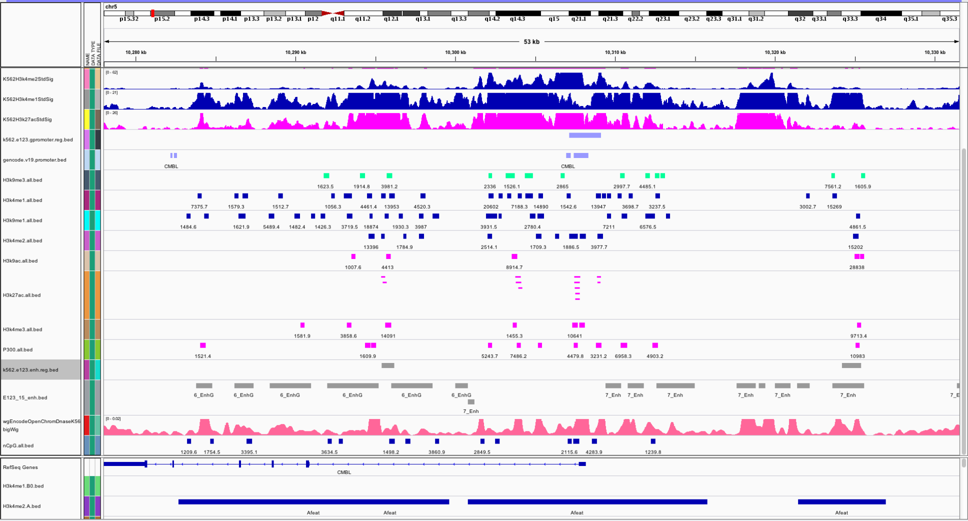
More #sig_features $\neq$ high importance;

Feature visualization

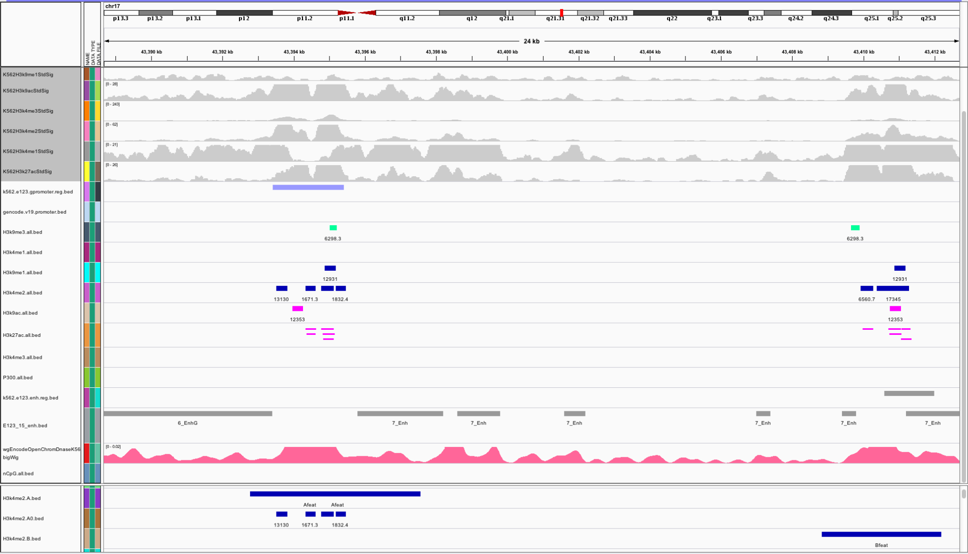


Example for top H3K27ac features

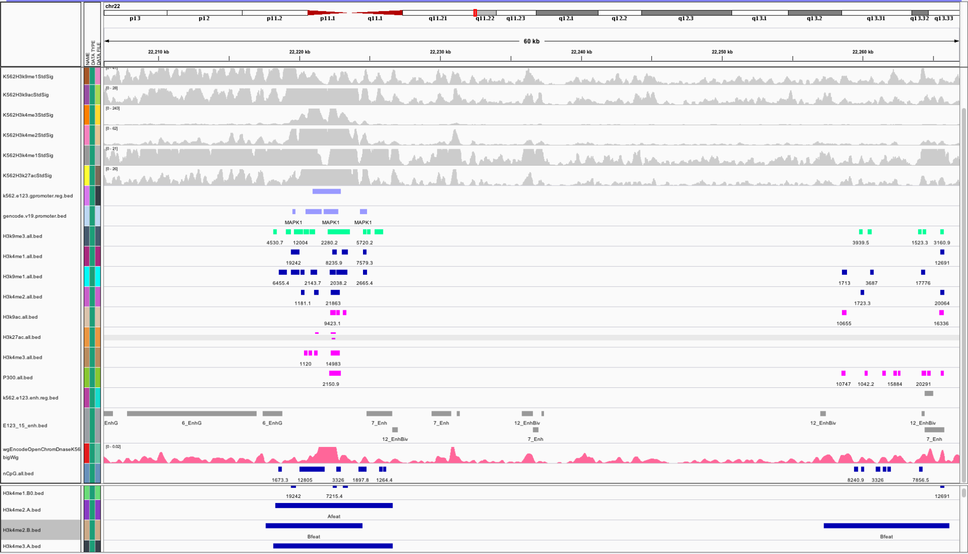
Feature visualization



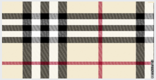
Feature visualization

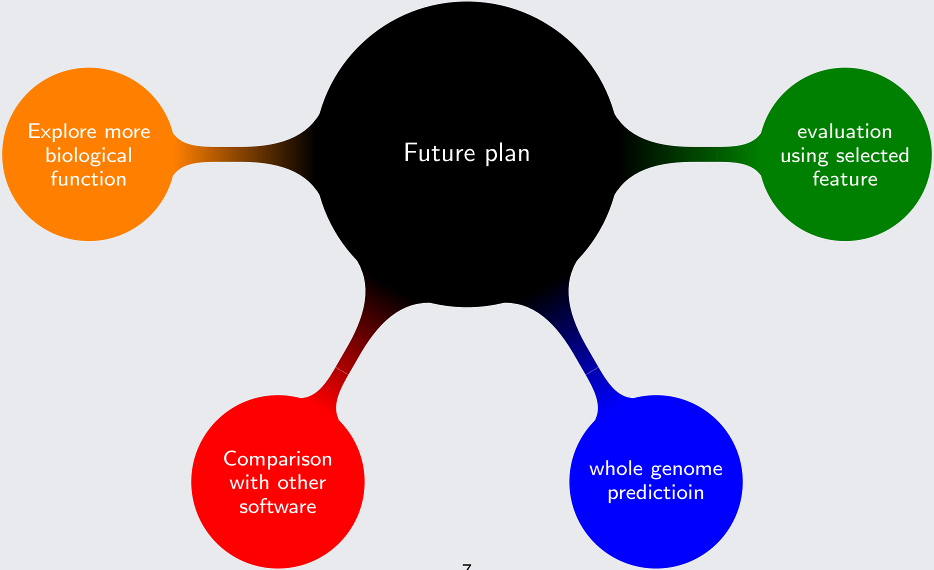


Feature visualization



Pattern





Future plan

Explore more biological function

evaluation using selected feature

Comparison with other software

whole genome predictioin