Planning for ENCODE Encyclopedia

Anurag Sethi
P2-TECH

Enhancer Predictions for ENCODE Encyclopedia/Round 3 Enhancer Validation

Goal: To make enhancer predictions for ENCODE Encyclopedia and validate these predictions using Len's assay

Enhancer prediction method has to work for both human and mouse ENCODE cell-lines/tissues.

Focusing at the moment on Mouse

Summary of Ren group's ENCODE efforts

Histone mod ChIP-seq:

1. H3K4me2
2. H3K4me3
3. K3K27ac
4. H3K9ac

Narrow marks

5. H3K4me1
6. H3K27me3
7. H3K9me3
8. H3K36me3

Broad marks

(also methlyC-seq & RNA-seq)

In these tissues:

	Developmental Stages							
	e10.5	e11.5	e12.5	e13.5	e14.5	e15.5	e16.5	P0
forebrain	Underway	Done	Planned	Done	Done	Done	Done	Done
midbrain	Underway	Done	Planned	Done	Done	Done	Done	Done
hindbrain	Underway	Done	Planned	Done	Done	Done	Done	Done
neural tube	Underway	Done	Planned	Done	Done	Done	Problematic due to bone formation	
limb	Underway	Done	Planned	Done	Done	Done		
cranioface	Underway	Done	Planned	Done	Done	Done		
heart	Underway	Done	Planned	Done	Done	Done	Done	Done
liver	Underway	Done	Planned	Done	Done	Done	Done	Done
intestine	Too early for these tissues				Done	Done	Done	Done
kidney					Done	Done	Done	Done
lung					Done	Done	Done	Done
stomach					Done	Done	Done	Done

Slide from David Gorkin (Bing Ren lab)

Round 2 Enhancer Prediction Methods

Method	Features	Training Data	Performance
Beer 1-3	H3K27ac and/or P300	Unsupervised	Better than baseline for forebrain (heart mixed)
Beer 4-6	Sequence	H3K27ac and/or P300 peaks	Good for heart
Brown	Methylation, DNase, TF, histone, CAGE	VISTA	Better for heart
Ensembl	ChromHMM + SegWay	Unsupervised	N/A
Gerstein	Histone	Unsupervised	Better for forebrain
Hardison	TF occupancy conservation	Unsupervised	N/A
Keles	Histone, TF, DNase, Sequence	VISTA	Better for heart
Kellis1	Histone, TF	VISTA	Better than baseline
Kellis2	Sequence	VISTA	Better than baseline
Valouev	Histone, DNase	VISTA	Better for heart
Yuan1-2	Histone, TF motifs	VISTA	Better for heart
Yuan3-4	Histone	VISTA	Better for heart
Weng	Histone	VISTA	Better for forebrain
Kingslay	Histone + TF	VISTA + P300	Better for heart
Wang	Histone	Unsupervised	Better for heart

Some notes about Overall performance

A number of theoretical prediction methods outperformed H3K27ac peaks.

H3K27ac was the minimal dataset required for making good predictions in forebrain (Caution: 39 observations).

For heart enhancer predictions, methods that used sequence and/or DNase information did better than methods that used just H3K27ac datasets (Caution: 31 observations with just 8-14 positives).

Composition of VISTA database

Tissue	Number of positives
heart	204
forebrain	376
midbrain	313
hindbrain	277
neural tube	202
limb	232
craniofacе	-
liver	8
intestine	-
kidney	-
lung	-
stomach	-

Note: most of these enhancers were validated experimentally at E11.5 stage.

Suggestions for ENCODE Encyclopedia/ Enhancer Validation Round 3

We can only use VISTA enhancers for training for 8 tissues - **not all** 16 tissues (especially the later embryonic stages).

VISTA positives are providing valuable information about enhancers positive in the transgenic assays (especially heart) - in my opinion, not using this information will be bad for the encyclopedia.

So suggestion is to let everyone predict enhancers for 8 tissues but use unsupervised methods alone for predictions in the other 8 tissues (as well as H3K27ac peaks).

Then, we use the 70 new experiments as a cross-validation dataset to come up with best ensemble-based method for predicting enhancers for the whole genome in these 16 tissues.

This method can be used for Encyclopedia as well as Round 3 of Enhancer Validation.

Ensemble Methods for Enhancer Prediction

Ensemble learning methods:

- Supervised methods (require separate learning data for ensemble training fr baseline method training) - Boosting, Bagging, and Stacking.
- Unsupervised methods - Merging scores from different methods - Currently truing out a few unsupervised methods (using 70 new experimental results a crossvalidation dataset).

Ensemble methods combine many weak-learners (better than baseline) to create a strong-learner (very good model). Criteria:
Weak-learners have to be diverse (we have this) - Machine learning methods focus on how to make a diverse set of weak learners (bagging) and how to create strong-learner from it (boosting, stacking, unsupervised methods, etc).

Methods to combine probability from different prediction methods

Average Score:

$$\overline{p(j)} = \sum_i p_i(j)$$

i - different methods

j - different enhancer candidates

Correlation Weighted Average Score:

$$\overline{p(j)} = \sum_i w_i p_i(j)$$

where $$w_i = \frac{\sum_k C_{ik}^{-1}}{\sum_k \sum_i C_{ik}^{-1}}$$

Methods to combine rankings from different prediction methods

The Condorcet candidate or Condorcet winner of an election is the candidate who, when compared with every other candidate, is preferred by more voters.

Optimizing this is NP-hard - algorithm is O(N!) where N is number of candidates.

So, approximate methods exist in lieu of this:

Borda Rank - Candidate ranked 1 gets N-1 votes, candidate ranked 2 gets N-2 votes, and so on.

Markov Chain (similar in spirit to PageRank kind of methods) - create a Markov Chain based on the comparison of pairs of candidates across different lists and then calculate the steady state distribution of such a Markov chain - this steady state distribution gives the ranking of different methods.

More methods might be tested.

Performance of Ensemble Models (Unsupervised)

Section 1 - Forebrain predictions - active in any tissue

Method	AUROC	AUPR
Average	0.698	0.838
Weighted Average	0.657	0.753
Borda Rank	0.719	0.856
Markov Chain	0.725	0.861
Best Methods	0.667 (Gerstein)	0.842 (Beer2)

Weak learners chosen based on performance on cross-validation set

Performance of Ensemble Models (Unsupervised)

Section 2 - Forebrain predictions - active in forebrain

Method	AUROC	AUPR
Average	0.666	0.736
Weighted Average	0.592	0.595
Borda Rank	0.697	0.792
Markov Chain	0.713	0.800
Best Methods	0.737 (Beer1)	0.741 (Beer3)

Weak learners chosen based on performance on cross-validation set

Performance of Ensemble Models (Unsupervised)

Section 3 - Heart predictions - active in any tissue

Method	AUROC	AUPR
Average	0.870	0.835
Weighted Average	0.576	0.609
Borda Rank	0.899	0.870
Markov Chain	0.887	0.847
Best Methods	0.840 (Keles8)	0.842 (Keles7)

Weak learners chosen based on performance on cross-validation set

Performance of Ensemble Models (Unsupervised)

Section 4 - Heart predictions - active in heart

Method	AUROC	AUPR
Average	0.647	0.336
Weighted Average	0.658	0.335
Borda Rank	0.685	0.496
Markov Chain	685	0.458
Best Methods	0.704 (Yuan3)	0.489 (Valouev4)

Weak learners chosen based on performance on cross-validation set

Performance of Ensemble Models (Unsupervised)

Even unsupervised models sometimes outperform the best baseline model, while at other times, it gives nearly comparable performance.

Borda Rank and Markov Chain methods tend to perform the best so far - though we are in the process of testing a few more methods.