Enhancer Predictions Ensemble Methods - I

Anurag Sethi
TECH

Setting the problem

We have 46 sets of predicted enhancer activity in mouse E11.5. Each set contains:
120 regions for forebrain/any
120 regions for heart/any
whole genome predictions for heart/forebrain

Experimental results from Len for 70 regions (39 H3K27ac peaks in forebrain and 31 H3K27ac peaks in heart).

AUC for ROC/PR show that a number of methods may outperform H3K27ac-based peak calling. But no method consistently comes on top for heart/forebrain across different metrics/assays!

Immediate Goal: Develop an ensemble method to predict active enhancers for 3rd round of Enhancer Validation experiments.

Ensemble strategies

Split the VISTA database into 2 halves. People who train with the VISTA database can use one half to train their model.

Ensemble method of choice uses 2nd half while training its model (bagging/boosting/stacked generalization) - supervised methods.

Alternatively, use ensemble-based methods to get average score/rank order the enhancer predictions from different groups (unsupervised) and assess which ensemble-based method works best on the 70 experimentally known values.

Ensemble strategies - Attempted so far

Mean score (mean probability of activity)

Weighted mean score (removes correlations) - can also try another weighting scheme based on accuracy in experimental assay.

Rank order - Borda count

Rank order - MC-based method

To be attempted
Rank order - Kemenization, EM-based method, Spectral method.

Borda Count

Input - lists of ranking of “N” candidate regions.

Candidate region ranked # 1 on a list get N-1 votes, region ranked #2 on a list gets N-2 votes, and so on….

Vote counting used to identify winner (or ordering of candidate regions).

Many applications of this method including rank sports teams in NCAA.

Markov Chain Method

The basis is to build a Markov chain that codes for the probability of candidate A winning over candidate B.

All candidates belong to set S.
Each method provides a ranking of subset of elements in S - τ. $\tau(i) < \tau(j)$ => candidate i ranked above candidate j

If the current state is candidate i, then the next state is chosen as follows: first pick a candidate j uniformly from S. If $\tau(j) < \tau(i)$ for the majority of the lists $\tau \in R$ that ranked both i and j, then go to j, else stay in i.

Left eigenvector of this Markov Chain matrix is the stationary distribution of an MC simulation and represents the final ranking of all candidates.

Dwork, et al.,10th intl conf on the WWW, 2001

Some preliminary results for heart

Method	AUC (ROC)	AUC (PR)
Average	0.707	0.461
WeightedAverage	0.424	0.220
BordaRank	0.685	0.456
rankMC4	0.707	0.468