Exploring the H2H genes in 3D view

Mingming Zhao, Jianguo Zhou, Zifeng Wu, Wenyu Peng, Wei Zhou,
Yu Liang
Software School of Yunnan University, Kunming, China.

Abstract. It is estimated to contain hundreds of head-to-head (H2H) gene in the eukaryotic genomes, often with two genes in an H2H pair been prone to co-express and o-function. Therefore, h2h plays a crucial role in human disease control and there have been many studies on H2H gene and its related bidirectional promoters. Recent chromosome conformation capture techniques, such as Hi-C, and ChIA-PET have provided us with new opportunities to study H2H in 3D view. This paper proposes a powerful machine learning algorithm LightGBM to predict h2h cluster. Two sets of features, namely protein features and sequence features, are extracted. Then these two sets of features are used to train a classifier to predict h2h cluster. Experimental results show that this method can effectively predict h2h cluster. Our results show a large fraction of long-range H2H (TSS >1 kb) exist and help us to understand the H2H at 3D level.

1. Introduction

Bidirectional gene organization is a common architectural feature in the eukaryotic genome [1]. As shown in figure1, a "head-to-head" (h2h) gene pair is defined as a genomic locus in which two adjacent genes are transcribed from opposite DNA strands and the region between two transcription start sites (TSS), also known as the 'two-way promoter', is typically less than 1000 bp. [1, 2]. there are many studies on the h2h genes, hamsters, etc. [3], which are related to bidirectional promoter characteristics [4], disease association [5,6], evolutionary conservation [7,8] and Functional association genes [9, 10, 11], to name a new. For instance, It has been found that the distance between the nearest neighbor TSSs in the relative chain is bimodal, and many bidirectionally paired transcripts are co-expressed, providing a unique regulatory mechanism for bidirectional alignment of a large number of mammalian genes [2]. Analysis of genome-wide transcriptional regulation of H2H genes reveal that they are linear and spatial interactions [12], as well as bidirectional genes are often co-expressed, and bidirectional gene pairs associated with similar functions appear to have a stronger expression correlation [13].
Figure 1. Two-Dimensional Head to Head (H2H) gene organization.

Figure 2. Three-dimensional H2H gene organization. A: The distance between two genes in two-dimensional space is greater than 1 kb, not h2h. B: But may be h2h in three-dimensional space.

Recent chromosome conformation capture techniques, such as Hi-C [14], and ChIA-PET [15] have provided us with new opportunities to study H2H in 3D view. As figure 2 shows, although the distance between two adjacent genes is larger than 1 kb in linear level, they are really quite close due to the spatial folding of the genome. Therefore, it is worth to research the H2H in 3D space.

We wanted to explore how the sequence and protein characteristics affect the H2H in three-dimensional space and predict h2h cluster. Therefore a LightGBM algorithm is introduced to predict the H2H genes pairs. First, we extracted two characteristics of h2h cluster (i.e. sequence features and protein features), and then combined these two features to train a classification model to predict h2h cluster. Finally, using the trained model to input sequence features and protein features of specific sites to identify the h2h cluster gene. Our

Figure 3. Flow chart of the method for predicting h2h cluster gene

Table 1. H2h Cluster Statistics.

| Cell line | Number of h2h gene pairs in H2H cluster | H2H cluster data total |
|-----------|----------------------------------------|------------------------|
| K562      | 16234                                  | 5940                   |
work demonstrates that the machine learning model has a very good effect on the prediction of h2h cluster based on sequence and protein features.

An overview of our method shows in Fig. 3. It contains five steps: (1) Collect data. We used the 5 resolution hic data, the human genome data hg19, with histone modification and binding protein. (2) Processing datasets. We use 5k hic data to find spatially interacting points through fithic, and find a h2h cluster in a series of screening searches (see 2.2. for more details). (3) Feature extraction. We extracted the sequence features and protein features from different protein data and sequence information across k562 cell lines in the region using the locus information corresponding to the h2h cluster gene (4) Model training. We used a total of four different models(lightgbm, xgboost, lstm and restnet18) to use the features obtained in step three to make predictions, and demonstrate that lightgbm has better performance. (5) Predicting. We use those two-part features to predict h2h cluster gene in our model. To verify the performance of the classifier, we use 3 different evaluation indicators, they are auc, acc, and f1-score.

2. Materials and Methods

2.1. datasets

In this work, the dataset we used was derived from hic data for k562 at 5k resolution, and human genome-wide data for hg19. And we use histone modification and binding protein ChIP-seq datasets from ENCODE Project, including three cell lines. This includes protein peaks, concentration data, and gene sequence data in the h2h gene region. In K562, a total of 16234 pairs of known h2h, after removal of the repeat, the total number of known h2h genes was 5940, as shown in Table 1.

2.2. Definition of h2h cluster

We find the spatial interaction points by fit-hic. According to q-value<0.05, the two sites obtained are made into 5k loop data according to the needs of our 5k data, according to the human genome. The genes, determine which position they belong to, find them, and only the genes on one side of the ring form a h2h according to the combination of positive and negative, and the genes on both sides of the ring are also combined in two, also composed H2h, so that all the genes at the position of the loop are a h2h cluster. Then the non-h2h gene is the gene in all human genomes except the h2h cluster gene, which is the unh2h.

2.3. Feature extraction

2.3.1. Data Analysis. This study used a gene-annotated sequence based on the human genome hg19. It is not clear which regions of a chromosome are useful for predicting the h2h gene. In order to predict the h2h gene, it is important to find the regions of the feature selection and extraction features on one chromosome. Based on the analysis of the genes and protein sequences of h2h in previous work, we also studied its sequence and protein characteristics. Therefore, we extracted their sequence characteristics and protein characteristics based on the locus information of the h2h gene.

In general, data characteristics require both quality and quantity. However, due to the small number of h2h genes, the number is defective. In the k562 cell line, we extracted multiple sets of protein features, and by analyzing multiple sets of proteins at different concentrations and densities, derived multiple sets of features based on protein concentration and peak.

2.3.2. Extraction of protein features. First, we mapped the locus information of the h2h gene on the chromosome to each protein, and included multiple sets of concentration and peak data in this region of the h2h gene, and thus we extracted the following features under each protein: the average signal value, maximum signal value, minimum signal value, variance of signal value, standard deviation of signal value, average peak , maximum peak , minimum peak, etc., the calculation formulas of these features are shown in Table 2.
**Table 2.** Formula for calculating protein features.

| Feature name            | Calculation formula                                      |
|-------------------------|----------------------------------------------------------|
| average signal value    | $\frac{1}{n} \sum_{i=1}^{n} SignalValue_i$             |
| maximum signal value    | $\max \{SignalValue\}$                                   |
| minimum signal value    | $\min \{SignalValue\}$                                   |
| variance of signal value| $\frac{1}{n} \sum_{i=1}^{n} SignalValue_i$, $S^2 = \frac{1}{n} \sum_{i=1}^{n} (M - SignalValue_i)^2$ |
| standard deviation of signal value | $\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\text{SignalValue}_i - \mu)^2}$ |
| average peak            | $\frac{1}{n} \sum_{i=1}^{n} \text{Peak}_i$             |
| maximum peak            | $\max \{\text{Peak}\}$                                  |
| minimum peak            | $\min \{\text{Peak}\}$                                  |

The above is just the extraction information of one protein. For other proteins, the extraction method we adopt is the same, adopting the method of parallel extraction.

![Figure 4. Extracting histone map](image)

The extracted histone map is shown in Figure 4. For the other chromosomes, we take the same processing steps and extract the information of the protein according to the position information of the h2h cluster. All the chromosomes of the same cell line share a set of proteins. Thus, the 1761-dimensional feature was extracted in the K562 cell line.

### 2.4 Sequence feature vector extraction based on information entropy

Information entropy is a concept in information theory. First, understand the concept of information volume. The amount of information is measured by the amount of information. The amount of information we receive is related to the specific event. The size of the information is related to the probability of the machine event, and the smaller the probability of the occurrence of the event, the greater the amount of information generated. For example, if the sun rises from the east, this is common sense, so there is very little information. If there is an earthquake in a certain place, the
amount of information brought by such a small probability event is very large. The amount of information measures the information brought about by a specific event, and the entropy is the expectation of the amount of information that can be generated before the result comes out - consider all possible values of the random variable, that is, all possible events the expectation of the amount of information brought. The formula for calculating information entropy can be defined as:

$$H(X) = \sum_{i=1}^{n} P(X_i) \log P(X_i)$$

Thus, 16-dimensional features were extracted in the K562 cell line.

2.5. Model

In this paper, we use the lightgbm model in machine learning as a predictive algorithm for classification, based on protein features and sequence characteristics to predict whether it is h2h gene. LightGBM is a fast, distributed, high performance decision tree based gradient Boosting framework developed by Microsoft. The main advantages are: faster training efficiency, low memory usage, better accuracy, support for parallel learning, and the ability to process large-scale data. Compared with common machine learning algorithms, the speed can be very fast. Based on the Histogram algorithm, LightGBM is further optimized. First, it discards the level-wise decision tree growth strategy used by most GBDT tools, and uses a leaf-wise algorithm with depth limitations. Level-wise data can split the same layer of leaves at the same time, easy to multi-thread optimization, control model complexity, and not easy to overfit. But in fact ,Level-wise is an inefficient algorithm because it treats the leaves of the same layer indiscriminately, which brings a lot of unnecessary overhead, because in fact many leaves have lower split gains so that there is no need to search. And split it. Leaf-wise is a more efficient strategy. When it splits, it finds the leaf with the highest split gain from all the current leaves, then splits and cycles. Therefore, compared with Level-wise, Leaf-wise can reduce more errors and get better precision when the number of splits is the same. The disadvantage of Leaf-wise is that it may grow a deeper decision tree and produce over-fitting. Therefore, LightGBM adds a maximum depth limit above Leaf-wise to prevent over-fitting while ensuring high efficiency.

![LightGBM Model Diagram](image)

**Figure 5.** lightgbm model diagram.

Another optimization of LightGBM is the Histogram (histogram) for differential acceleration. An easily observed phenomenon is that a histogram of a leaf can be obtained by the histogram of his father node and his brother's histogram. Usually constructing a histogram requires traversing all the data on the leaf, but the histogram is only necessary to traverse the k buckets of the histogram. Using
this method, LightGBM can obtain a histogram of its sibling leaves at a very small cost after constructing a histogram of a leaf, which can be doubled in speed. The basic idea of the histogram algorithm is to discretize successive floating-point eigenvalues into k integers and construct a histogram of width k. When traversing the data, the statistic is accumulated in the histogram based on the discretized value as an index. When the data is traversed once, the histogram accumulates the required statistic, and then according to the discrete values of the histogram, traverse to find the optimal segmentation point.

In our study, the prediction of the h2h gene is a two-class problem. A large number of features of proteins and sequence are extracted in the corresponding locus regions, and the above four models are used to implement classification training, and the lightgbm algorithm is relatively best. In order to get the best performance of the model, we obtained the current optimal parameters by tuning.

2.6. Evaluation indicators

First, we used the sequence and protein features of the h2h gene region of k562 in the lightgbm model. We used AUC values, F1-score, Accuracy (ACC) as our evaluation indicators, and evaluated the results. AUC (Area under the Curve of ROC) is the area under the ROC curve and is the criterion for judging the pros and cons of the two-class prediction model. The AUC value gives the average performance value of the classifier, and the AUC value can be used to evaluate the pros and cons of the classification problem of the two classification problem. The F1 score (F1-score) is a measure of the classification problem. It is the harmonic mean of the accuracy rate and the recall rate, with a maximum of 1 and a minimum of 0.

\[
AUC = \frac{\sum_{i=1}^{m} rank_i - \frac{M(1+M)}{2}}{M \times N}
\]

Formula explanation:

1. In order to obtain a combination, the score value of the positive sample is larger than the negative sample. If all the positive sample score values are greater than the negative sample, then the first and any combined score values are larger, we take its rank value. It is n, but M-1 in n-1 is a combination of a positive example and a positive example. This is not in the statistical range (for the convenience of calculation, we take n groups, and the corresponding non-conformities have M), so we want to reduce Off, then the same in the second place n-1, there will be M-1 is not satisfied, and so on, so get the following formula \( M \times (M + 1) / 2 \), we can verify the positive sample Under the assumption that the score is greater than the negative sample, the value of AUC is 1.

2. according to the above explanation, it is not difficult to conclude that the value of rank represents the number of combinations that can generate a large before and after the score, but here contains (positive, positive), so subtract this group (That is, the number of positive examples that follow it), you can get the above formula.

In addition, it is important to note that in the case of equal scores, the same rank should be assigned to samples of equal scores (regardless of whether this equal score is between samples of the same type or different classes, this needs to be handled in this way). The specific operation is to average the ranks of all the samples with the same score. Then use the above formula.

TP: The correct number of positives
FN: Missing report, no number of correct matches found
FP: False positive, no match is incorrect
TN: the number of non-matches that are correctly rejected
F1-score:

\[
F1 = \frac{2TP}{2TP + FP + FN}
\]

ACC:
2.7. Results of the model

As can be seen from Figure 6, from the acc, lightgbm is higher than xgboost, lstm, and restnet18 by 0.0007-0.1044. From auc, lightgbm is significantly higher than the other three models, which is 0.0693-0.2094 higher than the other three models. From F1_score, lightgbm is much higher than lstm and restnet18, but slightly less than lightgbm, so in general, lightgbm has better performance.

2.8. The importance of lightgbm features

In order to verify whether the sequence and protein features in the h2h gene region are helpful for predicting the h2h gene, the lightgbm model was used to sort according to the weight values of each feature, as shown in Figure 7.

Thus, sequence features are more important for predicting h2h cluster, and some protein features are also important. This shows that the features of protein and sequence we are looking for are somewhat useful for predicting the h2h cluster gene.

3. Conclusion

In this work, we explored the h2h gene in three-dimensional space from h2h in two-dimensional space. We found these h2h genes in three-dimensional space. Because there are more than one pair of
h2h genes in three-dimensional space, we named them. For the h2h cluster, and to explore the sequence characteristics and protein characteristics of these h2h cluster genes, using machine learning methods to achieve the two classification, to determine whether it can correctly predict the h2h cluster gene, we found four models for comparison, and used 3 evaluation indicators, finally, in general, lightgbm has better performance.

Acknowledgments
This work was supported in part by the National Natural Science Foundation of China under Grant 61762089, Grant 91631305, Grant 61863036, Grant 61862067 and Grant 81560380 and in part by the Science and Technology Innovation Team Project of Yunnan Province under Grant 2017HC012.

References
[1] Adachi N, Lieber MR: Bidirectional gene organization: a common architectural feature of the human genome. Cell 2002, 109 (7):807-809.
[2] Trinklein, N. D. An Abundance of Bidirectional Promoters in the Human Genome [J]. Genome Research, 2003, 14 (1):62-66.
[3] Yang M, Elnitski L. Orthology-driven mapping of bidirectional promoters in human and mouse genomes [J]. BMC Bioinformatics, 2014, 15(17 Supplement): S1.
[4] Lin,J.M, Collins,P.J., Trinklein,N.D. et al. (2007) Transcription factor binding and modified histones in human bidirectional promoters. Genome Res., 17, 818–827.
[5] Yang,M.Q., Koehly,L.M. and Elnitski,L.L. (2007) Comprehensive annotation of bidirectional promoters identifies co-regulation among breast and ovarian cancer genes. PLoS Comput. Biol., 3, e72.
[6] Tu J, Li X, Wang J. 2019. Characterization of bidirectional gene pairs in The Cancer Genome Atlas (TCGA) dataset. PeerJ 7: e7107 DOI 10.7717/peerj.7107.
[7] Yang,M.Q. and Elnitski,L.L. (2008) Prediction-based approaches to characterize bidirectional promoters in the mammalian genome. BMC Genomics, 9 (Suppl. 1), S2.
[8] Franck,E., Hulsen,T., Huynen,M.A. et al. (2008) Evolution of closely linked gene pairs in vertebrate genomes. Mol. Biol. Evol., 25, 1909–1921.
[9] Li,Y.Y., Yu,H., Guo,Z.M. et al. (2006) Systematic analysis of head-to-head gene organization: evolutionary conservation and potential biological relevance. PLoS Comput. Biol., 2, e74.
[10] Liu B, Chen J, Shen B. Genome-wide analysis of the transcription factor binding preference of human bi-directional promoters and functional annotation of related gene pairs[J]. BMC Systems Biology, 2011, 5).
[11] Hao P, Yu Y, Zhang X, et al. The contribution of cis-regulatory elements to head-to-head genes’ co-expression pattern [J]. SCIENCE IN CHINA SERIES C LIFE SCIENCES - ENGLISH EDITION-, 2009.
[12] Chen Y, Li Y, Wei J, et al. Transcriptional regulation and spatial interactions of head-to-head genes[J]. BMC Genomics, 2014, 15 (1).
[13] Wang Q, Wan L, Li D, et al. Searching for bidirectional promoters in Arabidopsis thaliana[J]. BMC Bioinformatics, 2009, 10 (1 Supplement):0-0.
[14] Lieberman-Aiden E, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009; 326:289–293.
[15] Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH, et al. An oestrogen-receptor-alpha-bound human chromatin interactome. Nature. 2009; 462: 58–64.