Integration of Genomic and Epigenomic Features to Predict Meiotic Recombination Hotspots in Human and Mouse

Min Wu
School of Computer Engineering
Nanyang Technological University
Singapore, 639798
wumin@ntu.edu.sg

Chee Keong Kwoh
School of Computer Engineering
Nanyang Technological University
Singapore, 639798
asckkwoh@ntu.edu.sg

Teresa M Przytycka
Computational Biology Branch
NCBI, NLM, National Institutes of Health
Bethesda, MD 20894, USA
przytyck@ncbi.nlm.nih.gov

Jing Li
Department of Electrical Engineering & Computer Science
Case Western Reserve University
Cleveland, OH 44106, USA
jingli@cwru.edu

Jie Zheng*
School of Computer Engineering
Nanyang Technological University
Singapore, 639798
*zhengjie@ntu.edu.sg

ABSTRACT

The regulatory mechanism of meiotic recombination hotspots is a fundamental problem in biology, with broad impacts on areas ranging from disease study to evolution. Recently, many genomic and epigenomic features have been associated with recombination hotspots, but none of them can explain hotspots consistently. It is highly desirable to integrate the different features into a predictive model, and study the relation of the features with hotspots and themselves with a systems approach. Moreover, due to rapid and dynamic evolution of recombination hotspots, regulatory mechanisms of hotspots that are evolutionarily conserved among species remain unclear.

We propose a machine learning approach that encodes genomic and epigenomic features into a support vector machine (SVM). Trained on known hotspots and coldspots in human and mouse genomes, the model is able to predict hotspots based on the features with good performance in both species. Moreover, the model reports a ranking of feature importance, uncovering the interactions of the features with hotspots and among themselves. Applying the method to large-scale data, we identified evolutionarily conserved patterns of *trans*-regulators and feature importance between human and mouse hotspots. This is the first attempt to build a predictive model to identify evolutionarily conserved mechanisms for recombination hotspots by integrating both genomic and epigenomic features.

Keywords

Recombination hotspots; epigenetics; histone modifications; SVM; comparative genomics.

1. INTRODUCTION

Meiotic recombination is the process that homologous chromosomes exchange arms during the generation of sex cells. As a fundamental cellular process, meiotic recombination plays key roles in sexual reproduction, birth-defect diseases, disease gene mapping, molecular evolution, etc. It is observed that meiotic recombination events tend to occur preferentially within short regions of a few kilo bases long, called "recombination hotspots". Despite its importance, the regulatory mechanism of recombination hotspots (e.g. the locations and intensities of hotspots) remains unclear. Development of new Bioinformatics methods to help elucidate the molecular architecture behind the control of recombination hotspots are highly desirable and would make broad impacts on biomedical research.

To understand genetic factors responsible for the recombination hotspots, various genomic features have been analyzed. Prominent sequence features include GC content, repetitive elements, sequence motifs, among others [16, 13]. Relations with genes (e.g. gene fraction, exon count, enrichment of GO terms) are also important features [7, 30]. The feature analysis is usually conducted using the following methods. First, the occurrences of each feature are counted in hotspots and coldspots, and features with large differences in occurrence are associated with recombination hotspots [20, 31]. Second, multiple linear regression model is used to predict recombination rate from genomic variables [13]. Third, machine learning approaches (support vector machine and random forest) have been applied to predict hotspots, but using limited features of DNA sequences (e.g., codon composition) [32, 14].

The above approaches compare hotspots and coldspots, but ignore the inheritable variation in recombination rates and sequence polymorphisms among individuals. Considering recombination rate as a phenotype, then genome-wide association studies (GWAS) can be carried out to identify associated genetic factors responsible for variations of recombination hotspots. This strategy has been applied successfully using pedigree-based methods [17, 5] or LD-based methods [31]. The genomic loci identified using these GWAS methods can be extended to detect proximal genomic features.

Meiotic recombination is profoundly connected with evolution. Like mutation, recombination is an important process to generate genetic diversity which is indispensible for natural selection. On the other hand, the evolutionary mechanism of recombination hotspots themselves remains elusive with several challenging questions. Due to biased gene conversion, a recombination hotspot tends to kill itself; nevertheless, many recombination hotspots in extant human populations have existed for thousands of generations. This is called *hotspot paradox* which remains open for more than a decade [4]. Another puzzle about evolution is the lack of conservation in locations of recombination hotspots between human and chimpanzee genomes, despite over 99% sequence identity between the two species [29]. Although this can be explained by the rapid evolution of hotspots [21], an important question remains: "What are the evolutionarily conserved mechanisms for regulating recombination hotspots among closely related species?" In this paper, we aim to address this question by comparative genomics approach to feature analysis between human and mouse.

The discovery of PRDM9 protein as a *trans*-regulator of recombination hotspots represents a major breakthrough in this field [3, 23, 21]. PRDM9 is a DNA-binding zinc finger protein whose binding motif is enriched in hotspots. Its zinc finger binding array evolves rapidly, which partially explains the hotspot paradox and the lack of location conservation between human and chimpanzee hotspots [21]. Following the discovery, multiple studies confirmed the role of PRDM9 as a regulator of hotspots [31, 25, 30]. However, PRDM9 is unlikely the only *trans*-regulator of hotspots, and it is desirable to identify other regulators like PRDM9. To this end, authors of this paper developed an approach to predicting *trans*-regulators of recombination hotspots in mouse genome [30]. Basically, we searched for transcription factors (TFs) with binding motifs enriched in hotspots against coldspots. Our analysis confirmed that PRDM9 is a major *trans*-regulator of hotspots; moreover, a top list of TFs is reported as putative regulators. Interestingly, epigenetic functions (especially histone modifications) are enriched in these candidate regulators as shown in our gene-ontology (GO) term analysis. This observation is consistent with the long-standing hypothesis that epigenetic control plays an important role in the regulation of recombination hotspots [1].

In addition to sequence-based features (e.g. GC content) reviewed above, there is increasing new evidence for the importance of epigenetic features in regulating meiotic recombination hotspots. For instance, DNA methylation tends to inhibit the formation of crossover; the occurrence of double strand breaks (DSBs) is enriched in open chromatin (see the review [1] and references therein). Histone modification is an important category of epigenetic features associated with recombination that is under intense research [25, 18]. The aforementioned PRDM9 protein is a meiosis-specific histone methyl-transferase that trimethylates H3 at K4 [12]. In a recent paper, feature selection was used to analyze the association of chromatin features (H3K4me3, H4K36me3, and nucleosome occupancy) and GC content with hotspots in yeast [11]. But this paper did not attempt to predict hotspots from these features, nor did the authors compare different species.

Due to the large amount and complexity of data involved, the research of meiotic recombination hotspots calls for new development of bioinformatics methods. A challenging problem is how to integrate the genomic and epigenomic features associated with recombination hotspots into a predictive model. Moreover, it is highly desirable to uncover regulatory mechanisms that are evolutionarily conserved among species.

In this paper, we developed a machine learning approach to integrate genomic and epigenomic features into an SVM-based classifer, which can be used to predict hotspots. The pipeline of our method is shown in Figure 1. We first identify a list of transcription factors (TFs) as putative *trans*-regulators of recombination hotspots, using an approach that we previously developed [30]. Then, we combine the binding site occurrences of the TFs with GC content and histone modification signals as features in an SVM model for classification. The SVM model is trained on known hotspots and coldspots in human and mouse genomes, and is used to predict if a short genomic region is a hotspot. Moreover, the training of SVM model automatically ranks the importance of features, which sheds light on the relations of the features with hotspots and among themselves. Last but not least, the orthologous features between human and mouse are compared to uncover the evolutionarily conserved mechanisms for recombination hotspots.

Applying this integrative method to large-scale real data of human and mouse, we achieved good predictive performance and obtained biologically interesting results. For instance, the AUC (area under the ROC curve) of chromosome-wide and genome-wide prediction of human hotspots are higher than 80% using Gaussian kernel. Orthologous TFs we identified as *trans*-regulators are correlated in the binding preference to hotspots between human and mouse. Moreover, the ranks of feature importance reported by our SVM model are significantly correlated between the human and mouse genomes. To our best knowledge, the identification of such *conserved cross-species* patterns of recombination hotspots with both genetic and epigenetic factors is new in the field. Our method can be applied to understand the molecular regulatory machinery as well as the evolutionary mechanisms of meiotic recombination hotspots. Moreover, the machine learning method can be used to predict hotspots in many other species for which genetic maps are not yet available.

2. METHODS

2.1 Predicting trans-regulators for recombination hotspots

We applied the framework in our preliminary study [30] to predict *trans*-regulators for recombination hotspots. For better readability, we briefly describe this framework in sub-

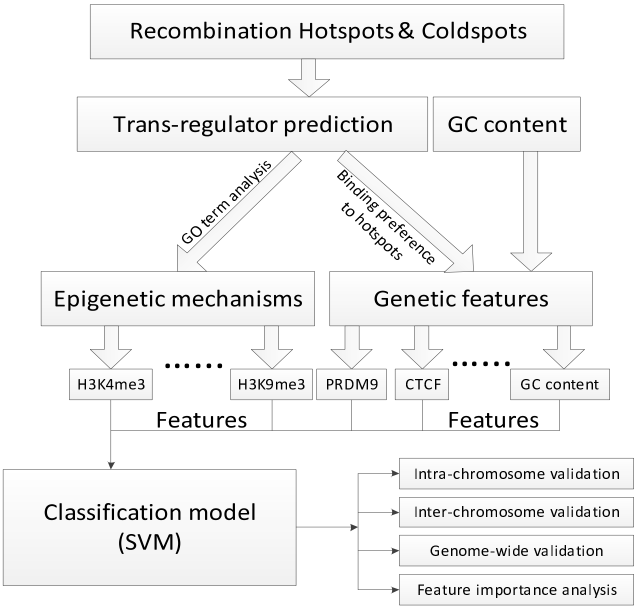
Recombination Hotspots & Coldspots
Trans-regulator prediction
GC content
GO term analysis
Binding preference to hotspots
Epigenetic mechanisms
Genetic features
H3K4me3
......
H3K9me3
PRDM9
CTCF
......
GC content
Features
Features
Classification model (SVM)
Intra-chromosome validation
Inter-chromosome validation
Genome-wide validation
Feature importance analysis

Figure 1: The framework for discovering the mechanisms of recombination hotspots and their computational prediction.

sections 2.1 and 2.2. In our previous work [30], we applied this framework to mouse hotspots inferred from ChIP-seq data and here we also apply to historical hotspots of human derived from SNPs.

For those proteins with known binding motifs available, we employed the software tool FIMO [8] to scan for their occurrences in both hotspots and coldspots. FIMO takes two files as inputs, namely, a file containing query motifs and the other file containing DNA sequences. Particularly, each query motif is represented as a position-specific frequency matrix and the sequence database consists of known hotspots and randomly generated coldspots (see more details about coldspots in [30]). FIMO computes a log-likelihood ratio score for each occurrence of the motif in the given sequences and converts this score to p-value and q-value (correcting the multiple-testing issue) to show the statistical significance of this occurrence. Finally, for a motif, an occurrence in a hotspot with low p-value or q-value indicates that the protein with this motif tends to bind to the hotspot.

In our preliminary study [30], we analyzed the FIMO results and defined odds ratio scores for TFs in equation 1 to represent the preference of a TF to bind to hotspots. Here, $h(g)$ and $c(g)$ are the number of hotspots and coldspots covered by the gene g respectively. Those TFs with high odds ratio scores are predicted as candidate *trans*-regulators of recombination hotspots.

$$O_{hc}(g) = \frac{h(g)}{c(g)} \cdot \frac{N - c(g)}{N - h(g)} \tag{1}$$

2.2 GO term analysis

Given a gene g, $T(g)$ is the set of GO terms annotating this gene. We define the similarity between a term t and a gene g, $S(t, g)$, in equation 2 and subsequently define the similarity between t and a set of genes G, $S(t, G)$, in equation 3.

$$S(t,g) = \frac{1}{|T(g)|} \sum_{t' \in T(g)} sim(t,t') \tag{2}$$

$$S(t,G) = \frac{1}{|G|} \sum_{g \in G} S(t,g) \tag{3}$$

Here, $sim(t, t')$ in equation 2 is the semantic similarity between GO terms t and t', which is calculated using the method in [28].

Let HG denote the set of genes with high odds ratio scores. Now, $S(t, G)$ and $S(t, HG)$ can be utilized to show the term t's enrichment in the whole gene group G and HG, respectively. Therefore, the gap score for the term t, $gap(t)$, defined in equation 4, can be used to discriminate t's enrichment in HG and G. For example, a large gap indicates that t is enriched in the genes with high odds ratio scores while not enriched in the whole gene group G.

$$gap(t) = \frac{S(t,HG) - S(t,G)}{S(t,G)} \tag{4}$$

2.3 Classification model for recombination hotspot

In this section, we collect different features to represent a sequence and then build classification models to predict whether it is a recombination hotspot or not.

Based on the GO term analysis for our predicted *trans*-regulators, epigenetic terms (especially histone modifications) are found to be enriched in these *trans*-regulators, providing insights into the epigenetic regulatory mechanism of recombination hotspots. Therefore, a sequence with high signal of histone modifications would be more likely to be recombination hotspots than those with low signal. In this paper, each hotspot or coldspot is divided into small bins and each bin is assigned with signal values for histone modifications based on our collected data. Thus, we can assign a hotspot (coldspot) a signal score which is the average signal value over all the bins of this hotspot (coldspot). Various histone modification data can be considered as different features for us to predict hotspots and coldspots.

Recently, PRDM9 has been identified as a *trans*-regulator of recombination hotspots in both human and mouse [3, 21, 23] and its binding site contains a 13-mer motif which is enriched in human hotspots, i.e., covering about 41% of human hotspots [22]. Therefore, the binding preference of PRDM9 to hotspots may help us to predict whether a sequence is a hotspot or coldspot based on the binding affinity of PRDM9 to the sequence. Meanwhile, our predicted *tran*-regulators such as MYC, SP1, CTCF and so on, have the similar binding preference to hotspots as PRDM9. As such, we will take the binding information of these *trans*-regulators as another subset of features for predicting recombination hotspots.

In addition, many other genomic features are related to recombination hotspots. For example, hotspots are generally accompanied with local increases of GC contents [25, 26]. Sequences with the potential to form non-B DNA structures are associated with the recombination activity [6]. In this paper, we also integrate the GC contents of the given sequences to identify their propensity to be hotspots.

After collecting the above features, each sequence will be associated with a vector (f_1, f_2, ..., f_l) where f_i (i = 1, 2, ..., l) is the i^{th} feature value, e.g., the average signal of histone modifications, binding information of *trans*-regulators or the percentage of GC content. Subsequently, we will apply support vector machines (SVM) to classify the given sequences to be hotspots or coldspots. SVM [27] is a state-of-the-art classification technique in machine learning and it has been proven to be one of the best classifiers in many application domains such as text categorization, image recognition, protein function prediction [2], and so on. It will find a maximum-margin hyperplane between the instances of the two classes. For non-linearly separable classification problems, SVM will use kernel functions to map the feature vectors into a higher dimensional feature space to obtain non-linear boundaries.

3. RESULTS

3.1 Data

We downloaded the mouse recombination hotspots from [25], which were detected by chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq). There are 9,874 hotspots in mouse genome and the average hotspot width is 3414.08 bases. Human recombination hotspots here are collected in a manner different from mouse hotspots. They are computationally inferred from SNP data as the peaks of the recombination rate profile estimated by the LDhat package [20]. As such, we finally collected 39,551 human hotspots with width less than 6k bases. DNA sequences for mouse (version: MGSCv37) and human (version: GRCh37) were downloaded from NCBI.

The binding motifs of TFs were collected from JASPAR [24] and TRANSFAC [19]. After processing, we obtained 158 human binding motifs and 148 mouse binding motifs respectively. The histone modification data were downloaded from UCSC genome browser. In addition, the data for GO term analysis were downloaded from http://www.geneontology.org.

3.2 GO term conservation for trans-regulators

Here, we apply the method in subsection 2.1 to human and mouse TFs and then compare the top list of TFs between the two species. Human TFs with odds ratio score at least 1.2 are selected similarly as in [30] to form the gene set HG. As such, HG consists of 29 TFs in human and 40 TFs in mouse respectively. Table 2 shows the GO terms enriched in human HG (the GO term enrichment in mouse HG genes is similar and not shown here). We can find that epigenetic terms, such as histone acetylation, histone methylation (i.e., H3K4 and H3K9 methylation), histone deacetylation and chromatin modification, are enriched in both human and mouse HG TFs. Our observation on GO term conservation shows the epigenetic regulatory mechanism for recombination hotspots is conserved between human and mouse.

For the above human and mouse HG TFs, their intersection identified 11 ortholog pairs as shown in Table 1. For these orthologs, Table 3 shows their GO term enrichment based on the *gap* scores. In Table 3, those epigenetic terms are more enriched with higher *gap* scores in those conserved orthologs than in all those HG TFs. Take the term GO:0016568 (chromatin modification) as an example. In Table 3, it is ranked as the first with the highest gap score 0.307, much higher than its gap score 0.119 in human HG. Meanwhile, the term GO:0007126 (meiosis), which is directly related to meiotic recombination hotspots, is enriched in human orthologs while not enriched in human HG TFs. As such, the orthologs in both human and mouse as shown in Table 3 are more associated with recombination hotspots than those *trans*-regulators predicted in individual species.

Table 1: Orthologs in human and mouse HG TFs.

Orthologs	Odds ratio in human	Odds ratio in mouse
MYC	1.489	1.689
USF1	1.397	1.619
PRDM9	1.386	1.630
TP53	1.275	1.230
CTCF	1.247	1.336
PAX5	1.238	1.289
SP1	1.232	1.430
ZIC2	1.228	1.205
JUN	1.225	1.263
ESR1	1.207	1.286
ARNT	1.203	1.516

We will next utilize the binding information of these 11 proteins in Table 1 as features. Meanwhile, there are 10 different kinds of histone modifications under 3 different cell lines for human. Thus, we have 30 features from histone modifications for human. Similarly, we have 20 features from histone modifications for mouse. Together with the feature of GC content, we finally have 42 features for human and 32 for mouse to build the classification model.

3.3 Cross-validation for individual chromosomes

Both linear SVM and SVM with Gaussian kernels (RBF kernel) have been used in our experiments (using SVM^{light} software [15]). Two evaluation metrics are used to measure the performance of SVM for predicting hotspots and coldspots, i.e., AUC and Accuracy. AUC is the area under the Receiver Operating Characteristics (ROC) curve, which is a graphical plot of the sensitivity vs. 1-specificity for a classifier as the decision threshold varies. Accuracy is the fraction of instances that are correctly predicted, i.e., (TP + TN) / N where TP is the number of true positives (correctly predicted as positives), TN is number of true negatives and N is the total number of instances for prediction.

Figures 2 and 3 show the cross-validation results of SVM on the hotspots and coldspots from individual chromosomes of human and mouse (i.e., intra-chromosome validation), respectively. For example, there are 5,906 instances including hotspots and coldspots on human chromosome 1. By conducting a 5-fold cross-validation on human chromosome 1, the AUC and accuracy of SVM with RBF kernels are 0.866 and 0.779 respectively and those of SVM with linear kernel

Table 2: GO terms enriched in human HG(with top 15 gap scores).

Rank	GO terms	GO term descriptions	gap
1	GO:0051573	negative regulation of histone H3-K9 methylation	0.153
2	GO:0031060	regulation of histone methylation	0.146
3	GO:0035065	regulation of histone acetylation	0.144
4	GO:0006337	nucleosome disassembly	0.143
5	GO:0016573	histone acetylation	0.136
6	GO:0051574	positive regulation of histone H3-K9 methylation	0.134
7	GO:0051571	positive regulation of histone H3-K4 methylation	0.132
8	GO:0045947	negative regulation of translational initiation	0.131
9	GO:0031065	positive regulation of histone deacetylation	0.125
10	GO:0006334	nucleosome assembly	0.124
11	GO:0016568	chromatin modification	0.119
12	GO:0035066	positive regulation of histone acetylation	0.118
13	GO:0042986	positive regulation of amyloid precursor protein biosynthetic process	0.117
14	GO:0006338	chromatin remodeling	0.111
15	GO:0032348	negative regulation of aldosterone biosynthetic process	0.11

Table 3: GO terms enriched in the orthologs as shown in Table 1, based on the GO annotations for human genome (with top 15 gap scores).

Rank	GO terms	GO term descriptions	gap
1	GO:0016568	chromatin modification	0.307
2	GO:0051574	positive regulation of histone H3-K9 methylation	0.302
3	GO:0016573	histone acetylation	0.3
4	GO:0051573	negative regulation of histone H3-K9 methylation	0.3
5	GO:0051571	positive regulation of histone H3-K4 methylation	0.297
6	GO:0006338	chromatin remodeling	0.294
7	GO:0031060	regulation of histone methylation	0.288
8	GO:0035066	positive regulation of histone acetylation	0.278
9	GO:0031065	positive regulation of histone deacetylation	0.273
10	GO:0035065	regulation of histone acetylation	0.273
11	GO:0007126	meiosis	0.262
12	GO:0045799	positive regulation of chromatin assembly or disassembly	0.26
13	GO:0016584	nucleosome positioning	0.254
14	GO:0006334	nucleosome assembly	0.245
15	GO:0006337	nucleosome disassembly	0.239

are 0.748 and 0.688 respectively. Table 4 shows the average AUC and accuracy of SVM for intra-chromosome validation.

From Figure 2 and Table 4, we can observe that the kernel trick are quite effective for predicting human hotspots. For example, SVM with RBF kernel has the AUC 0.812, which is 9.1% higher than the AUC of linear kernel (0.721). Similarly, the accuracy of SVM on human hotspots can also be improved by the RBF kernel. This indicates that the features for human hotspots have complicated associations with each other, in which case sophisticated boundaries generated by the RBF kernel could perform better than linear kernel. However on mouse chromosomes, linear kernel achieves higher AUC and accuracy than RBF kernel. Possible reasons could be that we have more features for human (i.e., 30 histone modification features for human vs. 20 for mouse) and the relationship among human features may be more complicated than those of mouse. As such, the kernel tricks are more effective for predicting human hotspots than for mouse. In addition, the good performance of SVM (e.g., the AUC of SVM with RBF kernel on human chromosomes 1 and 15 can achieve up to 0.86 and the average AUC of RBF kernel is 0.812 over all the human chromosomes), demonstrates that the features we used here, including histone modification data, binding information of *trans*-regulators and GC contents, are indeed highly related with recombination hotspots. It can be reasonably expected that the performance of SVM (or other classifiers) can be further improved by integrating more genetic and epigenetic features in the future.

3.4 Inter-chromosome validation

In the previous subsection, we showed the cross-validation results for each individual chromosome. In fact, those are intra-chromosome validation results—training the SVM on some instances in a chromosome and testing on some other instances in the same chromosome. In this subsection, we will show the results of inter-chromosome validation, i.e., training on the instances in one chromosome and testing on the instances in all the other chromosomes.

Figures 4 and 5 show the inter-chromosome validation results for human and mouse, respectively. For example, when training SVM with RBF kernel on human chromosome 1 and

Table 4: AUC and accuracy of SVM for various kinds of validations.

		Intra-chromosome validation		Inter-chromosome validation		Genome-wide validation	
		AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
Human	Linear	0.721	0.655	0.640	0.583	0.680	0.621
	RBF	0.812	0.735	0.657	0.597	0.815	0.736
Mouse	Linear	0.720	0.653	0.703	0.643	0.739	0.650
	RBF	0.699	0.642	0.686	0.633	0.722	0.664

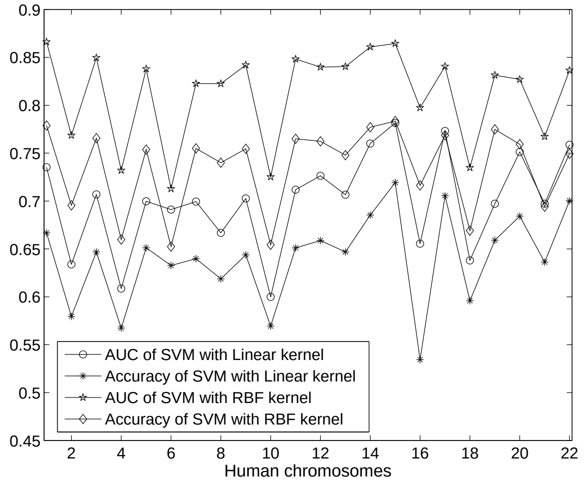

Figure 2: Intra-chromosome validation: AUC and accuracy of SVM on each chromosome of human.

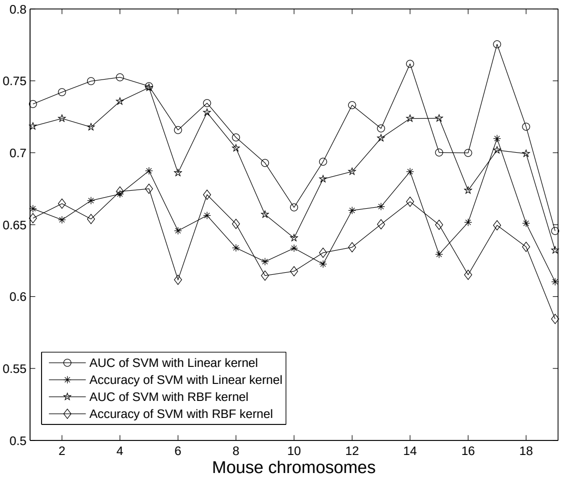

Figure 3: Intra-chromosome validation: AUC and accuracy of SVM on each chromosome of mouse.

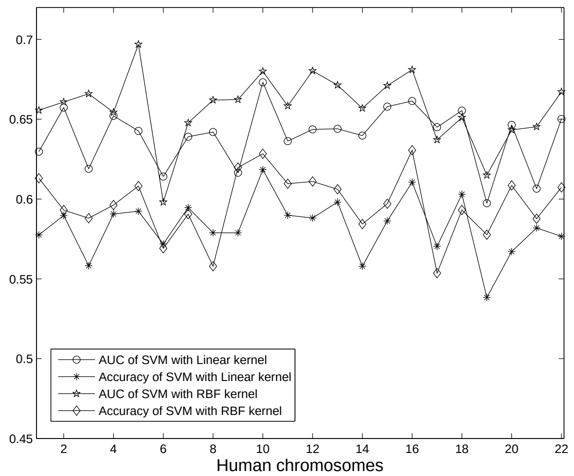

Figure 4: Inter-chromosome validation: AUC and accuracy of SVM for human.

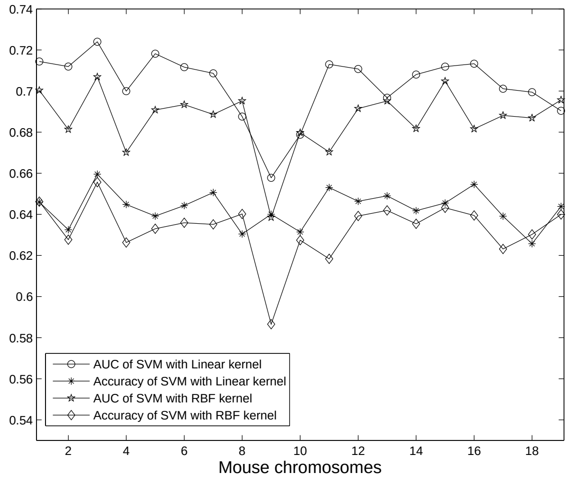

Figure 5: Inter-chromosome validation: AUC and accuracy of SVM for mouse.

testing on all the other chromosomes, the average AUC and accuracy are 0.656 and 0.613 respectively as shown in Figure 4. Table 4 also includes the average inter-chromosome validation results over all the chromosomes. It is interesting that intra-chromosome validation results for human are better than its inter-chromosome validation results while for mouse they are quite similar. A possible reason could be that the feature distributions (or even the mechanisms) for hotspots on different chromosomes may be different for human while they are more consistent for mouse. Meanwhile, SVM with RBF kernel still achieve better performance than linear kernel for predicting human hotspots. We speculate that those features for human recombination hotspots have more complicated interactions and exploring those relationships would be our future studies.

3.5 Genome-wide validation

We also conducted the genome-wide validation for both human and mouse. Genome-wide validation means that we randomly select the training instances from all the chromosomes (i.e., genome-wide selection of training samples) and then test on the remaining instances. Table 4 also shows the average AUC and accuracy of SVM for genome-wide validation.

As discussed previously, the mechanisms for hotspots on different human chromosomes may be different. Therefore, we can expect that the genome-wide validation results will be a tradeoff between intra- and inter-chromosome validation results. The accuracy and AUC of SVM with linear kernel in the genome-wide validation also confirm the above hypothesis. However, RBF kernel still achieves a high performance in human genome-wide validation, which once again indicates the complicated interactions among human features. Meanwhile, mouse hotspots have consistent mechanisms over different chromosomes. Hence, the performance of the genome-wide validation is better than those of intra- and inter-chromosome validation as shown in Table 4.

3.6 Feature importance analysis

After training linear SVM, the absolute values of the feature weights show the importance of these features [10], i.e., the larger $|w_j|$ is, the more important role the j^{th} feature plays in predicting hotspots. Figure 6 shows the importance of common features between human and mouse. We find that the trained models in human and mouse (i.e., the correlation of the feature importance) are highly correlated. For example, the Pearson correlation coefficient between the two feature importance vectors is 0.673 with a low p-value 0.003. This demonstrates that a common feature between human and mouse play similar roles in predicting hotspots, indicating that the major determinants for recombination hotspots are evolutionarily conserved.

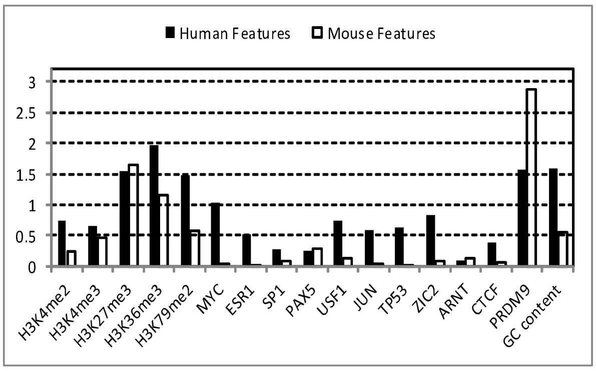

Figure 6: Importance scores of 17 common features between human and mouse.

In Figure 6, the top 5 important features for both human and mouse are H3K27me3, H3K36me3, H3K79me2, the binding information of PRDM9 and GC contents. For GC contents, our results are consistent with the fact that they are closely related to recombination hotspots [25, 26]. PRDM9, the major *trans*-regulator for recombination hotspots, plays an important role for predicting hotspots in our experiments, as expected. Furthermore, it is a novel discovery that the histone modifications have such a high impact for the prediction. This discovery once again confirms the epigenetic regulatory mechanisms for recombination hotspots.

In a recent study, H3K4me3 was observed to be enriched in mouse hotspots [25], i.e., 94% of mouse hotspots overlap with peaks of H3K4me3 signals. However, most of these enriched regions in hotspots are not transcription promoters which have even higher H3K4me3 signals. As such, H3K4me3 signals are believed to be insufficient for predicting hotspots [18]. Interestingly, the importance of H3K4me3 in Figure 6 is moderate, which is consistent with this point [18].

4. DISCUSSIONS AND CONCLUSIONS

In this paper, we first predict transcription factors with high binding preference to recombination hotspots as their *trans*-regulators. Subsequent GO term analysis provides insights into the epigenetic regulatory mechanisms for recombination hotspots. Inspired by this discovery, we integrate epigenetic and genetic data as the features of sequences and build classification models (SVM with linear and RBF kernels) to computationally predict recombination hotspots. With the limited number of features used, the intra- and inter-chromosome validations show good performance, demonstrating our collected epigenetic and genetic features are highly associated with recombination hotspots. In addition, the absolute values of the feature weights learned by linear SVM indicate the relative importance of these features. The vectors of feature importance for human and mouse are highly correlated with respect to Pearson correlation. That is, a common feature between human and mouse play similar roles in predicting hotspots, showing the existence of evolutionarily conserved mechanisms for recombination hotspots.

In the future, an importance issue is to investigate the associations among various features with respect to recombination hotspots. For example, mouse PRDM9 DNA-binding specificity determines the sites of H3K4me3 [9]. This would help us to better understand the inherent mechanisms for recombination hotspots. We are also going to explore more features for both human and mouse, e.g., DNA quadruplex data and more epigenetic data (e.g., DNA methylation data). As more data are integrated, our classification model is expected to be more powerful and accurate to predict recombination hotspots. In addition, the difference in regulatory mechanisms of recombination hotspots between human and mouse as suggested by our results would be an interesting topic worth investigation.

5. REFERENCES

[1] P. Barthes, J. Buard, and B. de Massy. Epigenetic factors and regulation of meiotic recombination in mammals. *Epigenetics and Human Reproduction*, 2011.

[2] Z. Barutçuoglu, R. E. Schapire, and O. G. Troyanskaya. Hierarchical multi-label prediction of gene function. *Bioinformatics*, 22(7):830–836, 2006.

[3] F. Baudat and *et al.* Prdm9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science*, 327(5967):836–840, 2010.

[4] A. Boulton, R. S. Myers, and R. J. Redfield. The hotspot conversion paradox and the evolution of meiotic recombination. *Proc. Natl Acad. Sci. USA*, 94(15):8058–8063, 1997.

[5] R. Chowdhury, P. R. J. Bois, E. Feingold, S. L. Sherman, and V. G. Cheung. Genetic analysis of variation in human meiotic recombination. *PLoS Genet*, 5(9):e1000648, 09 2009.

[6] N. Chuzhanova, J. M. Chen, and *et al.* Gene conversion causing human inherited disease: evidence for involvement of non-b-dna-forming sequences and recombination-promoting motifs in dna breakage and repair. *Hum Mutat*, 30(8):1189–98, 2009.

[7] K. A. Frazer, D. G. Ballinger, D. R. Cox, and *et al.* A second generation human haplotype map of over 3.1 million snps. *Nature*, 449:851–861, 2007.

[8] C. E. Grant, T. L. Bailey, and W. S. Noble. Fimo: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011.

[9] C. Grey, P. Barthes, G. C.-L. Friec, F. Langa, F. Baudat, and B. de Massy. Mouse prdm9 dna-binding specificity determines sites of histone h3 lysine 4 trimethylation for initiation of meiotic recombination. *PLoS Biol*, 9(10):e1001176, 10 2011.

[10] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.

[11] L. Hansen, N.-K. Kim, L. Marino-Ramirez, and D. Landsman. Analysis of biological features associated with meiotic recombination hot and cold spots in saccharomyces cerevisiae. *PLoS ONE*, 6(12):e29711, 12 2011.

[12] K. Hayashi, K. Yoshida, and Y. Matsui. A histone h3 methyltransferase controls epigenetic events required for meiotic prophase. *Nature*, 438(7066):374–8, 2005.

[13] M. I. Jensen-Seaman, T. S. Furey, B. A. Payseur, and *et al.* Comparative recombination rates in the rat, mouse, and human genomes. *Genome Research*, 14:528–538, 2004.

[14] P. Jiang, H. Wu, J. Wei, F. Sang, X. Sun, and Z. Lu. Rf-dymhc: detecting the yeast meiotic recombination hotspots and coldspots by random forest model using gapped dinucleotide composition features. *Nucleic Acids Research*, 35(Web-Server-Issue):47–51, 2007.

[15] T. Joachims. *Making Large-Scale SVM Learning Practical.* Advances in Kernel Methods: Support Vector Machines, 1998.

[16] A. Kong, D. F. Gudbjartsson, J. Sainz, and *et al.* A high-resolution recombination map of the human genome. *Nature Genetics*, 31:241–247, 2002.

[17] A. Kong, G. Thorleifsson, H. Stefansson, G. Masson, A. Helgason, D. F. Gudbjartsson, G. M. Jonsdottir, S. A. Gudjonsson, S. Sverrisson, T. Thorlacius, and et al. Sequence variants in the rnf212 gene associate with genome-wide recombination rate. *Science*, 319(5868):1398–1401, 2008.

[18] M. Lichten and B. de Massy. The impressionistic landscape of meiotic recombination. *Cell*, 147:267–270, 2011.

[19] V. Matys and *et al.* Transfac: transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, 31(1):374–378, 2003.

[20] S. Myers, L. Bottolo, C. Freeman, G. McVean, and P. Donnelly. A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310(5746):321–324, 2005.

[21] S. Myers, R. Bowden, A. Tumian, R. E. Bontrop, C. Freeman, T. S. MacFie, G. McVean, and P. Donnelly. Drive against hotspot motifs in primates implicates the prdm9 gene in meiotic recombination. *Science*, 327(5967):876–879, 2010.

[22] S. Myers, C. Freeman, A. Auton, P. Donnelly, and G. McVean. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nature Genetics*, 40(9):1124–1129, 2008.

[23] E. D. Parvanov, P. M. Petkov, and K. Paigen. Prdm9 controls activation of mammalian recombination hotspots. *Science*, 327(5967):835, 2010.

[24] E. Portales-Casamar, S. Thongjuea, and *et al.* Jaspar 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 38(Database-Issue):105–110, 2010.

[25] F. Smagulova, I. Gregoretti, K. Brick, P. Khil, R. Camerini-Otero, and G. Petukhova. Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. *Nature*, 472(7343):375–378, 2011.

[26] C. C. A. Spencer, P. Deloukas, and *et al.* The influence of recombination on human genetic diversity. *PLoS Genetics*, 2(9):e148, 09 2006.

[27] V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer-Verlag New York, Inc. New York, NY, USA, 1995.

[28] J. Wang, Z. Du, R. Payattakool, P. Yu, and C. Chen. A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23(10):1274–1281, 2007.

[29] W. Winckler, S. R. Myers, and *et al.* Comparison of fine-scale recombination rates in humans and chimpanzees. *Science*, 308:107–111, 2005.

[30] M. Wu, C. K. Kwoh, T. M. Przytycka, J. Li, and J. Zheng. Prediction of trans-regulators of recombination hotspots in mouse genome. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 57–62, 2011.

[31] J. Zheng, P. P. Khil, R. D. Camerini-Otero, and T. M. Przytycka. Detecting sequence polymorphisms associated with meiotic recombination hotspots in the human genome. *Genome Biology*, 11(R103):1–15, 2010.

[32] T. Zhou, J. Weng, X. Sun, and Z. Lu. Support vector machine for classification of meiotic recombination hotspots and coldspots in saccharomyces cerevisiae based on codon composition. *BMC Bioinformatics*, 7(223), 2006.