基于支持向量机的增强子预测器

刘振宇 元培学院 1700017853

2020.1.2

摘要：增强子（enhancer）在真核生物的转录调控中发挥着重要的作用。作为很多转录因子和转录激活蛋白的靶点，增强子可以通过与启动子之间的相互作用，特异性调节下游基因的表达。但由于增强子缺少特异的序列标签，且难以通过测序手段大规模鉴定，增强子的预测一直是转录调节领域面临的问题之一。本文的工作设计了一个利用随机森林进行变量选择，基于支持向量机进行分类的增强子预测器。利用多种表观遗传修饰以及相关的酶，染色质开放程度等信息，对特定基因组区域是否具有增强子活性进行预测。模型在人慢性髓系白血病细胞系 K562 中取得了很好的预测效果。

背景介绍

真核生物的基因表达受到一系列精细复杂的调控，这一调控过程涉及到转录、翻译、以及翻译后修饰等多个不同层面。目前的模型认为，真核生物的基因转录受到染色质中一系列顺式作用因子（如启动子、增强子、绝缘子等）和染色质蛋白（如组蛋白、转录因子、表观修饰酶）的共同调节。在各种顺式作用因子中，增强子起到了相对重要的作用：它可以通过三维基因组的特定结构，如拓扑相关结构域（topological associated domain，TAD）[1]以及 DNA 环（loop）与基因的启动子形成相互作用，并在相关的蛋白作用下启动基因转录。增强子区域序列的突变可以显著影响所调控的基因的表达。

早期对于增强子功能的研究主要是依赖实验的方法，如 DNA 序列的凝胶迁移率实验（EMSA）等。近年来全基因组的表观遗传学观测技术显示，激活的增强子上往往会富集特定的表观遗传修饰，如组蛋白 H3 的 27 号赖氨酸的乙酰化（H3K27ac）[2]。这为我们识别和研究增强子的位置和功能提供了新的思路。

在本文的工作中，我设计并训练了一个用于预测基因组上增强子的机器学习模型。运用近几年产生的基因组表观遗传学、特定染色质修饰酶、转录因子的 ChIP-Seq 数据，以及全基因组的染色质开放性（ATAC-Seq）数据，应用随机森林进行变量选择，利用选择后的特征对一个特定基因组区域是否是增强子进行预测。在人慢性髓系白血病细胞系 K562 中的测试显示，预测器达到了很高的预测准确率。

数据集选择与准备

1. 数据下载

已经有成熟的数据库收录了人类基因组上的增强子区域，这些数据库有些收录的是基于实验证据证实的增强子，如 VISTA Enhancer Browser[3]，也有一些是基于计算的预测结果，如 Broad ChromHMM [4]。由于 VISTA Enhancer Browser 所包含的增强子序列较少（全基因组只有 1339 个，第 22 染色体上只有 12 个），可能会影响训练的效果，我使用了 Broad ChromHMM 对 K562 细胞基因组的注释作为金标准，指导进行模型的训练。注释数据从 UCSC table browser 中下载（https://genome.ucsc.edu/cgi-bin/hgTables），下载文件为 bed 格式，具体参数见图 1。

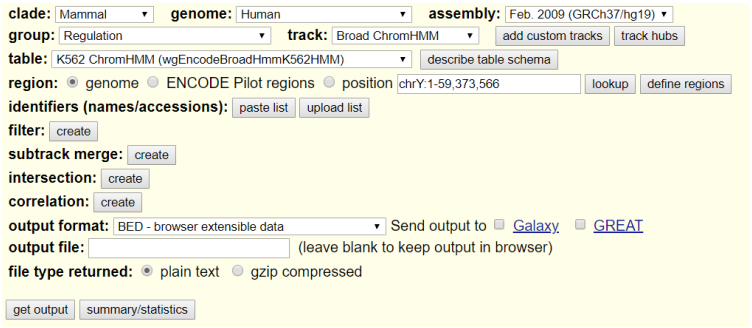

图 1 “金标准”K562 Broad ChromHMM 的下载

工作中使用的所有的 ChIP-Seq 的数据全部来自于 ENCODE 的 Broad Histone 数据集（http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeBroadHistone），数据集中收录了多种细胞的 ChIP-Seq 数据。我下载了基于 K562 细胞系的 21 组数据，其中包括 7 种表观修饰相关的酶、1 个转录因子、12 种表观遗传修饰和 K562 细胞系的对照组（input）。下载的文件均为 bigwig 格式，用于下一步计算基因组区域上的信号强度。

除了上述的表观遗传组，相关修饰酶和转录因子之外，染色质开放性也被认为是活跃增强子的特征之一。因此，我下载了已发表的工作中对 K562 细胞系的 ATAC-Seq 数据[5](GSE70482)。下载的数据也以 bigwig 格式储存，包含两次独立的重复实验。

表 1. 使用数据信息及来源

数据名称	组织	实验方法	来源	格式	链接
Enhancer 注释	K562	HMM	Broad ChromHMM	bed	https://genome.ucsc.edu/cgi-bin/hgTables
表观遗传修饰		ChIP-Seq	Broad Histone	bigwig	http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeBroadHistone
表观修饰酶					
转录因子					
染色质开放性		ATAC-Seq	GSE70482[5]	bigwig	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE70482

2. 阳性数据集筛选，阴性数据集生成

ChromHMM 对增强子的区域包含两种注释：Strong_Enhancer 和 Weak_Enhancer，为了获得更高的置信度，我只选择了标记为 Strong_Enhancer 的区域，作为阳性数据集，即真实的增强子。同时，考虑到传统定义上的增强子一般距离转录起始位点（TSS）距离较远，TSS 附近的区域一般注释为启动子，因此我又对每一个基因组区域距离最近 TSS 的位置进行了筛选。使用 Bioconductor 的 TxDb.Hsapiens.UCSC.hg19.knownGene 作为参考注释将每一个区域注释到最近的 TSS 上，只有距离 TSS 大于 3000bp 的才被认为是真正的增强子。最终得到的增强子数据集包含 33393 个增强子。

作为对照，需要产生非增强子的阴性数据集。我采用随机生成的策略产生阴性对照组。为了排除所选择的区域长度，染色体分布等因素的影响，我在随机生成时加上了诸多限制，使得产生的阴性数据集的宽度，染色体分布与阳性数据集相同，从而把二者的差异限制在所使用的特征上（图 2）。具体的实现命令为：

```
bedtools shuffle -i TrueEnhancer.bed -g hg19.genome.size -excl exclude.bed -chrom > FalseEnhancer.bed
```

其中 TrueEnhancer.bed 为筛选后的增强子区域，exclude.bed 为随机抽样时排除的区域，包括之前的增强子区域和 TSS 上下游 3kb 的区域。这样得到的阴性数据集，每条染色体上的区域数量（图 2.A），宽度分布（图 2.B）都与阳性数据集相同，而且二者之间没有重叠区域（数据未展示）。

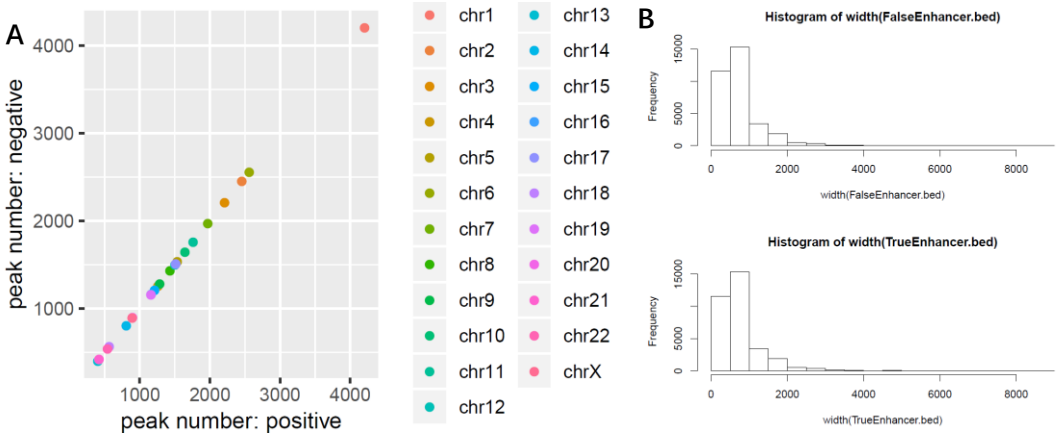

图 2. 阴性，阳性数据集的基因组分布与宽度

3. 特征的值的计算

得到了阳性和阴性的增强子区域后，在这些区域上分别计算各种特征的信号强度，作为每一个区域上该特征的值。由于每一个特征都有自己明确的生物学意义，而且计算出的信号强度不可能出现缺失值，因此直接使用计算出的信号强度作为输入的特征。这样既可以保证分类的准确性，也保证了模型的可解释性。模型选用了 21 个不同的特征（表 2），包含了染色质区域的组蛋白修饰、组蛋白修饰酶、转录因子以及染色质开放程度的特征。

表 2 模型所使用的所有特征

组蛋白修饰		组蛋白修饰酶	转录因子	染色质开放程度
H3K27ac	H3K79me2	P300	Ctcf	ATAC
H3K4me1	H3K36me3	Hdac2a		
H3K4me2	H3K9me1	Hdac1		
H3K9ac	H4K20me1	Ezh2		
H3K4me3	H3K9me3	Hdac6a		
H2az	H3K27me3	Suz12		
		Cbp		

在计算 ChIP-Seq 得到的信号强度时，由于 Broad Histone 提供了空白对照的数据（input），因此对于每一组 ChIP-Seq 的数据，都直接减去的对应区域的对照值，作为这一个区域的信号强度。

对于 ATAC 的数据，下载到的是两组独立重复实验的结果。两组数据的信号强度的算数平均值被用作这一区域的信号强度。

4. 分割数据集

至此，我们已经得到了阳性和阴性增强子区域，以及每个区域上的 21 个特征的值。由于最终的目的是预测 22 号染色体上的所有增强子，模型训练时没有使用 22 号染色体的数据，而是将其留作最后的预测使用。其他染色体上的阴性，阳性增强子被均分为训练集和测试集。

特征选择

1. 随机森林进行变量选择

模型使用随机森林进行变量选择，即使用所有的 21 个特征，使用随机森林的方法构建一个分类器，用以预测所给区域是不是一个增强子。基于得到的随机森林，可以计算每一个决策树的基尼不纯度（Gini impurity），进而取平均值得到随机森林的平均不纯度。从 21 个变量中删除一个变量，利用剩下 20 个变量构建的随机森林，其平均基尼不纯度会下降，下降的程度反映了删除变量对于整个随机森林分类器的贡献程度，即变量的重要性[6]。

在计算得到了 21 个变量的平均不纯度减小（mean decrease impurity）之后，我们把它当做变量的重要性进行变量选择。重要性值大于 100 的变量被保留（图 3）。

2. 所选变量的生物学意义

从变量选择的结果来看，对最终预测结果预测最大的几个特征都是组蛋白的表观遗传修饰，如 H3K27ac，H3K4me1，H3K4me2，H3K9ac 等，这些都是之前报道过的与增强子功能相关的表观遗传修饰[7]。尤其是 H3K27ac 和 H3K4me1，前者被认为是活化的增强子的特异性标记，而后者则帮助建立了增强子的“待发”（poised）状态[2]。同样的，H3K4 的二甲基化和三甲基化，以及 H3K79me2 都是活跃转录的标记，而 H3K4me3 在富集在启动子的区域作用更加明显。同样的，H3K36me3 也是之前报道过的活跃转录的基因的标记[8]。

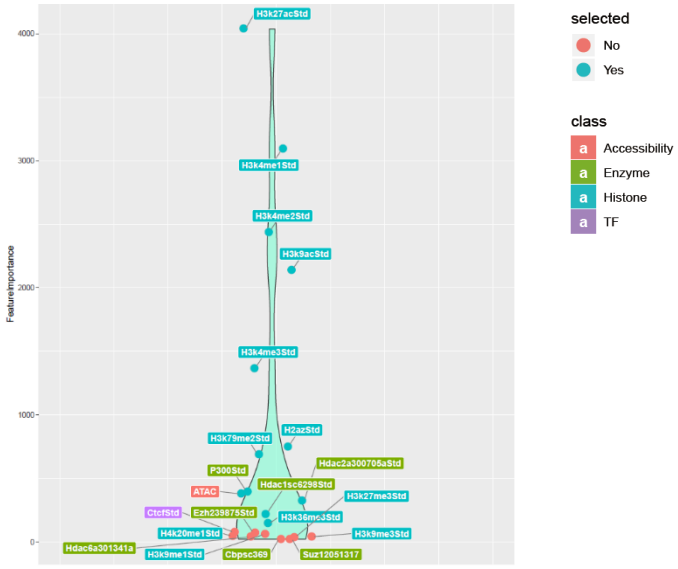

图 3. 变量选择

表 3. 通过变量选择的特征

组蛋白修饰		组蛋白修饰酶	染色质开放程度
H3K27ac	H3K4me3	P300	ATAC
H3K4me1	H2az	Hdac2a	
H3K4me2	H3K79me2	Hdac1	
H3K9ac	H3K36me3		

被选出的特征中还有三个组蛋白修饰的酶，从选择结果来看，所选出的三个酶都起到调节组蛋白乙酰化的作用。P300 是哺乳动物中常见的组蛋白乙酰基转移酶（Histone acetyltransferase，HAT），以乙酰辅酶 A 为底物，介导组蛋白的乙酰化。而 Hdac2a，Hdac1 都是组蛋白去乙酰酶家族（Histone deacetylase，HDAC）的酶，催化组蛋白的去乙酰化。如前文所说，组蛋白的乙酰化，尤其是 H3K27ac 在增强子功能中起到决定作用，因此 HAT 和 HDAC 在模型中发挥重要作用符合生物学意义。而同属于 HAT 的 CBP 在模型中重要性不强，可能与 P300 与 CBP 选择性表达有关。

同样意料之中的是，染色质开放程度在模型中发挥了重要的作用。目前的理论认为，增强子的激活可能是一个或几个先锋因子（pioneer factor）结合，促进包装紧密的染色质开放，同时伴随组蛋白修饰的过程。因此染色质的开放程度可以用于预测增强子功能，符合生物学常识。

模型训练

1. 利用拆分的训练集和测试集进行训练

至此，我们已经得到了阳性和阴性的增强子区域，以及这些区域上筛选过的特征的值。有这些数据，我们就可以进行模型的训练。我们采用采用支持向量机（Supporting Vector Machine，SVM）作为分类模型。为了得到最好的模型效果，我们首先测试了两种不同的正则化手段（C-SVM 和 Nu-SVM），以及四种不同的核函数（线性、多项式、放射状和 Sigmoid 核函数），在默认参数下的训练结果可以看出（表 4），使用 C-Classification 和放射状（radial）核函数可以得到最好的预测结果。

表 4. 不同正则化方法和核函数下错误预测个数

	Linear	Polynomial	Radial	Sigmoid
C-Classification	364	702	306	1171
Nu-Classification	1419	1106	1277	1447

表 5. 模型在测试集上的表现

Model1	Truth	
	Pos	Neg
True	16323	221
False	103	16205

Model2	Truth	
	Pos	Neg
True	16312	177
False	114	16249

使用 C-Classification 和 radial 核函数，以及默认的超参数，即：

$$gamma = 1/data\ dimension = 0.0833 \ ; \ \text{cost} = \ 1$$

我们可以得到最简单的分类器（模型 1）。使用模型 1 对测试集进行预测，得到的结果准确率达到了 99.01%（表 5. 左侧），灵敏度（Sensitivity）为 0.9937，特异性（Specificity）为 0.9865。说明模型 1 已经能够很好地对测试集中的数据进行分类。

之后，我们又对模型中的超参数进行了进一步的调整。参数调整的过程是使用“e1071”包中自带的 tune() 函数实现的，它可以在所给的参数向量中依次进行尝试，对每一组参数和所给的训练集进行 10 折交叉检验，找出最终准确率最高的一组参数。由于参数列长度较大、训练集规模较大、同时 10 折交叉检验的计算复杂度较高，这一步的计算量非常惊人。为了减小计算，我们在调试参数之前对训练集进行了缩小采样，随机选取 500 个数据产生一个小的集合进行调参。最终结果显示，在参数组合：

$$gamma = 0.01\ ;\ \text{cost} = \ 100$$

下，模型对训练集有最好的表现。使用上述优化后的参数，我们重新训练了一个新的预测器（模型 2），并再次对测试集进行了预测。结果显示，优化后的模型准确率为 99.11%，相比未优化的模型上升了仅仅 0.1%，灵敏度为 0.9931，特异性为 0.9892。可以看出，优化参数之后的模型相比于原始模型，性能有只有微弱的提升，但是优化后模型的 gamma 值远小于模型 1，所以模型 2 需要更多的支持向量，复杂度较高。因此我们在之后的预测中仍然主要使用模型 1，因为二者的预测性能并没有显著的差异。

2. 采用 10 折交叉检验进行学习

虽然我们在数据准备过程中尽可能保证了阴性与阳性增强子，训练集与测试集的平行性。但是采用随机抽样的方法可能会造成训练集和测试集之间存在无法预料的隐性误差。因此，对训练集进行交叉检验是更合理的学习策略。

但是由于 10 折交叉检验计算量较大，加之 SVM 本身复杂度高、数据集庞大，工作中未能完成这一策略的计算。但是相关的代码已经写好并调试过了。可以想到的是，10 折交叉检验可能会给模型带来更好的性能，但这种性能的提升很大可能是非常局限的。因为之前的模型已经达到了很高的准确率，在不改变模型搭建的基础上，仅仅通过交叉检验不太可能对模型 1 有显著的性能提升。

模型预测与性能评估

1. 对 22 号染色体上增强子的预测。

经过训练与测试，我们已经得到了一个预测效果很好的增强子预测器（模型 1），之后我们可以利用这个模型对人 22 号染色体上的增强子进行预测。22 号染色体上所有待预测区域的特征值已经在之前与训练集和测试集一同计算过了，直接输入模型就可以进行预测。预测结果与 ChromHMM 的金标准进行比较（表 6），准确率为 97.69%，灵敏度 0.9908，特异性 0.9648。可以看出，得到的模型可以较好地区分数据集中的增强子区域和非增强子区域。

表 6 模型对 22 号染色体增强子的预测

Chr22	Truth	
	Pos	Neg
True	536	20
False	5	521

表 7 模型对 H1 细胞增强子的预测

H1 cell line	Truth	
	Pos	Neg
True	4806	168
False	13270	17908

2. 对不同细胞系基因组上增强子的预测

目前我们训练的模型只是针对人慢性髓系白血病细胞系 K562，所使用的金标准为对 K562 细胞系的注释，用来计算特征的数据也全部来自于对此细胞系的实验。鉴于模型对于 K562 细胞系取得了不错的效果，我们不禁思考：对于另一个不同的人细胞系，模型的预测效果又会怎么样呢？

我们选择人胚胎干细胞 H1 细胞系进行了尝试，同样是使用 Broad ChromHMM 作为金标准，采用前述

的方法选择训练集和测试集，对 H1 细胞系全基因组上的增强子进行了预测。可以从预测结果可以看出，预测的准确率仅有 62.83%，而且错误的预测出现了显著的偏向性：假阴性的预测个数远远多于假阳性（表 7），假阴性的数量达到了 13270 个，接近假阳性的 80 倍。这种现象可能是 K562 细胞与 H1 细胞的分化程度有关，与取样策略以及增强子的分布也有关，这一问题在之后的讨论部分有进一步的分析。

可使用命令行参数的程序

至此，对于模型的训练和调试，以及性能测试已经基本结束了。但是，之前各个步骤的工作在不同目录下一步一步完成，对应的脚本也封装在不同的文件中，使用起来有诸多不便。为了使用的方便与快捷，我把数据读入，数据预处理，筛选后特征的计算，模型预测以及输出结果整合成了一个完整的程序（EnhancerPredictor.R）。该程序可以接收命令行参数，可以在安装了 R 的系统中使用命令行的方式直接调用。

EnhancerPredictor 的输入文件为基因组的区域，即 bed 格式文件。输出结果为含有附加列的 bed 文件。每一个行为一个基因组区域，顺序与输入的 bed 文件相同，附加列中是各个特征值的计算结果，以及对该区域是否为增强子的预测结果。

调用 EnhancerPredictor 时有两个参数，第一个参数是输入的 bed 文件，第二个，第二个是输出文件。具体调用方法为：

```
Rscript EnhancerPredictor.R input.bed output.bed
```

需要注意的是，调用程序前必须要修改工作目录为程序所在的位置，而且必须保证配置文件（.\Config.RData）和数据文件夹（.\data\）的正确相对位置。图 4 展示了正确调用的方式与运行结果，为了展示的方便，bed 文件使用 excel 打开，实际文件均为纯文本文件。

A

chr1	8519414	8520213
chr1	20315614	20316413
chr1	25296614	25297213
chr1	26738414	26740413
chr1	28834814	28837413
chr1	29490414	29491413
chr1	57015813	57017412
chr1	82444213	82445212
chr1	91488213	91488412
chr1	144965444	144965643
chr1	158072777	158073576
chr1	169082777	169082976
chr1	36338	36537
chr1	780738	781137
chr1	948338	949337
chr1	958938	960337
chr1	960338	960737

B

```
D:\Final>Rscript EnhancerPredictor.R Performance\ChromHMM.Enh.H1.bed test1.txt
grid
Warning message:
'DelayedArray' dims'
stats4
BiocGenerics
parallel

GenomeInfoDb
Warning message:
'GenomeInfoDb' R3.5.2
Warning message:
'e1071' R3.5.3
[1] "Calculating feature value: H3k27acStd ..."
[1] "Calculating feature value: H3k4me1Std ..."
[1] "Calculating feature value: H3k4me2Std ..."
[1] "Calculating feature value: H3k9acStd ..."
[1] "Calculating feature value: H3k4me3Std ..."
[1] "Calculating feature value: H2azStd ..."
[1] "Calculating feature value: H3k79me2Std ..."
[1] "Calculating feature value: P300Std ..."
[1] "Calculating feature value: Hdac2a300705aStd ..."
[1] "Calculating feature value: Hdac1sc6298Std ..."
[1] "Calculating feature value: H3k36me3Std ..."
[1] "Calculating feature value: ATAC ..."
[1] "Normalizing Signal Density..."
[1] "Classification..."
[1] "Generating output file..."
[1] "Done."
```

C

seqnames	start	end	width	strand	H3k27acSt	H3k4me1S	H3k4me2S	H3k9acStd	H3k4me3S	H2azStd	H3k79me2	P300Std	Hdac2a30	Hdac1sc6	H3k36me3	ATAC	prediction
chr1	8519415	8520213	799	*	-0.51174	2.270839	-0.93377	2.843705	0.057722	1.687059	4.365607	7.149287	3.267434	3.943579	0.653166	0.000108	0
chr1	20315615	20316413	799	*	-0.19469	-0.05442	-0.45512	1.796045	0.521201	2.478899	-0.05577	6.163955	1.086608	3.738473	0.379775	0.000272	0
chr1	25296615	25297213	599	*	-0.20674	9.757596	6.447412	1.85202	1.340367	13.66918	0.294758	7.064908	4.306511	8.09576	0.119132	0.001213	0
chr1	26738415	26740413	1999	*	0.023512	10.16414	-0.21011	0.662651	0.413327	2.791536	-0.19684	8.728604	7.385393	2.384972	-0.92076	0.000403	0
chr1	28834815	28837413	2599	*	142.3548	18.82192	130.5316	196.3463	126.9227	-0.74536	145.1744	23.59521	9.769819	15.47032	45.06867	0.001352	1
chr1	29490415	29491413	999	*	0.015175	6.811532	0.057057	3.188789	2.898819	-0.01421	41.90915	11.76521	0.626386	2.689329	12.41301	0.000258	0
chr1	57015814	57017412	1599	*	-0.81711	-0.53944	-1.23174	-0.01699	-0.83822	0.637924	-0.99034	5.625016	1.0403	2.176785	-0.4794	0.00064	0
chr1	82444214	82445212	999	*	-0.91391	-0.48549	-0.9029	-0.90498	-0.04505	-0.31431	-0.5025	5.69958	1.343343	1.247848	-0.49049	0.000172	0
chr1	91488214	91488412	199	*	57.67638	19.05809	38.76201	52.96703	20.41709	68.1998	24.2607	9.986733	2.825327	12.14111	2.2601	0.002089	1
chr1	1.45E+08	1.45E+08	199	*	14.14231	30.41749	3.249447	12.71879	7.220301	4.304523	85.19276	31.53246	12.10432	16.38754	18.89045	0.00543	1

图 4. EnhancerPredictor 的输入（A），调用（B）与输出（C）

至此，我们完成了一个简单的增强子预测程序（EnhancerPredictor），并对它的表现进行了测试与评估。图 5 展示了整个工作的大致流程与原理。

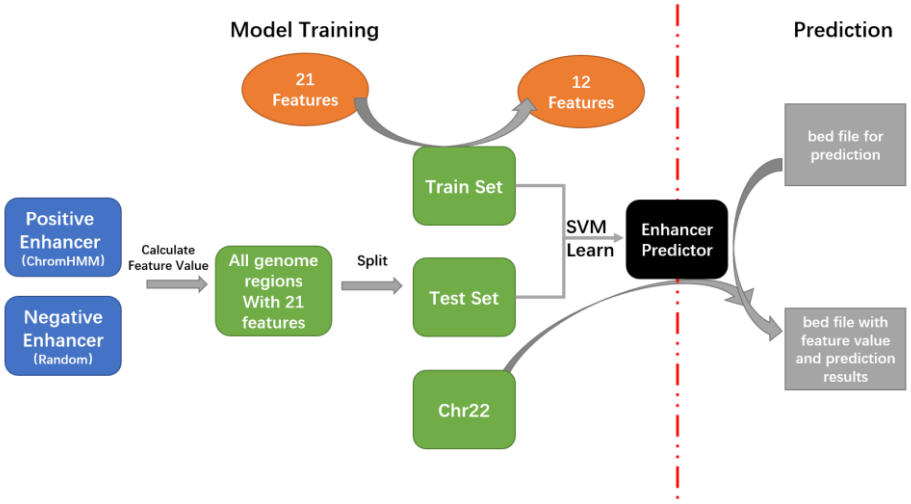

图 5. 项目的工作流程与原理

问题与讨论

1. 模型对其他细胞系的表现较差

在“模型预测与性能评估”部分，曾经提到了此模型对于人胚胎干细胞 H1 细胞系的表现较差，准确率只有 62.83%，说明模型对细胞品系表现出了很强的特异性。即基于 K562 细胞系的数据并不能用来预测其他细胞系的增强子。进一步观察预测结果可以发现，造成准确率低的原因主要是出现了大量的假阴性结果，而假阳性结果则相对较少。

我们首先考虑出现数量较多的假阴性错误。假阴性意味着即很多在 H1 细胞系中表现出增强子活性的区域被预测成了非增强子。出现这种错误的原因有二，一是分类器本身对于边界值的处理存在模糊，二是这些增强子在 K562 细胞系中确实没有活性，因此使用 K562 细胞系的数据不能很好区分这些区域和真正的非增强子。为了进一步研究产生这种现象的原因，我们使用 IGV（Integrated Genome Browser）对这些假阴性区域上的组蛋白修饰情况进行了观察（图 5）。可以看出，这些区域在 H1 细胞系中确实有较强的 H3K27ac 修饰和 H3K4me1 修饰，这与它们被标记为增强子的金标准是符合的。然而在 K562 细胞系中，这些区域往往没有上述的两种修饰，因此使用 K562 细胞的数据来预测这些区域，会得到阴性的预测结果。综上，模型预测出现的大量假阴性结果主要是 H1 细胞相对于 K562 细胞特意的增强子。这也说明了模型仅仅针对 K562 细胞系有效。

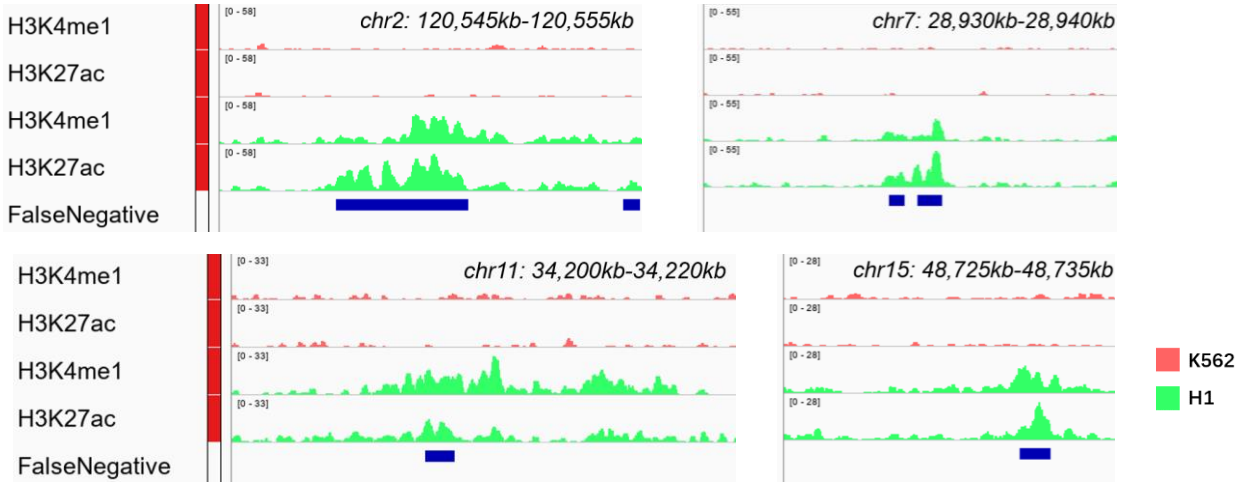

图 6. 假阴性区域上两种细胞系的染色质修饰

之后我们考虑假阳性错误数量较少的原因。假阳性错误主要是指 H1 细胞中并不是增强子的区域被模型预测为有活性增强子。由之前的讨论可知，模型可以较为准确地识别 K562 细胞系中的增强子，因此我们可以认为这些假阳性区域大部分确实是 K562 细胞中的增强子，但是在 H1 细胞中不具有增强子活性。由于增强子在基因组上的比例相对较小，我们采用随机选择阴性区域的策略产生的阴性集合中，包含的 K562 细胞的增强子的区域自然会相对较小。但这也与阴性集合的大小有关，当随机选取的区域不断增加时，阴性集合中的 K562 细胞系内的增强子数量也会逐渐增加。

综上，对于 H1 细胞系的预测中，假阴性的区域主要是 H1 细胞系特异的增强子，而假阳性的区域主要是阴性集合包含的 K562 细胞系的增强子。二者的数量悬殊主要是由于阴性集合中包含 K562 细胞系的增强子数量较少，而不是模型带来的系统误差。

2. 预测结果的生物意义问题

基于 ChromHMM 的金标准，我们的模型得到了不错的预测结果。但是需要注意的是，这种看似较好的预测结果并不一定具有真实的生物学意义。首先，ChromHMM 使用的信息也是组蛋白修饰和相关的催化酶的 ChIP-Seq，与本模型使用的数据集存在较高的重合度，只是二者所用的学习模型不同，ChromHMM 主要是基于隐马尔可夫模型进行预测，而我的模型则是使用支持向量机。因此，本模型的预测结果与 ChromHMM 的金标准重合度高，这是可以预见的。

但需要注意的是，无论是基于 HMM 还是 SVM，都完全没有生物学功能的验证，只是根据基因组上的修饰情况对基因组功能进行预测。因此，二者预测出的增强子都无法保证是真正有生物学功能的增强子，这些区域完全有可能是富集了特定表观遗传修饰，但并无真正功能的区域，其生物功能的验证必须通过遗传学手段进行验证。这也是大数据方法分析生物过程的普遍问题。

当然，我们也可以使用有实验证据证实的增强子区域作为模型中的阳性增强子进行训练，如之前提到的 Vista Enhancer Browser 数据库中数据。但此方法面临最大的问题来自数据量过小。正如前文提到的，Vista Enhancer Browser 在人类 hg19 基因组上仅有 1339 个增强子注释，具体到问题中的 22 号染色体上，只有 12 个。这样的数据量显然不够用来进行模型的训练与测试。因此采用较大的数据量进行预测，再用实验方法检验预测结果似乎是更有效的办法。

除此之外，我们还可以考虑通过引入其他层面数据的方法来增强预测结果的可信度，例如使用三维基因组的数据帮助进行预测。由于活跃开放的增强子往往与其调控的启动子之前存在空间上的相互作用，使用三维基因组的数据，如 HiC 等，可以帮助预测真正参与这种相互作用的增强子，以及它们参与的增强子-启动子相互作用[9]。

致谢

感谢课程的助教学长对本工作相关问题的解答。感谢北京大学-北极星-高性能计算平台提供的计算支持。

成员分工

本项目所有的工作，从最初的数据查询到最终的报告撰写均由我一人独立完成。

数据位置

模型中共使用了 K562 细胞系的 21 组 ChIP-Seq 数据以及 2 组 ATAC 数据，还有 H1 细胞系的两个 ChIP-Seq 数据。上述数据均以 bigwig 格式储存，共计 25 个 bigwig 文件。都可以在本组（第七组）教学服务器上找到，具体目录为：/home/mibg7_pkuhpc/lzy/Final/data/

同时，在目录/home/mibg7_pkuhpc/lzy/Final/ 下有完整的项目代码和数据。

参考文献

[1] Dixon, J.R., S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J.S. Liu, and B. Ren. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 485:376-380.

[2] Creyghton, M.P., A.W. Cheng, G.G. Welstead, T. Kooistra, B.W. Carey, E.J. Steine, J. Hanna, M.A. Lodato, G.M. Frampton, P.A. Sharp, L.A. Boyer, R.A. Young, and R. Jaenisch. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A. 107:21931-21936.

[3] Visel, A., S. Minovitsky, I. Dubchak, and L.A. Pennacchio. 2007. VISTA Enhancer Browser--a database of tissue-specific human enhancers. Nucleic Acids Res. 35:D88-92

[4] Ernst, J., and M. Kellis. 2012. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 9:215-216.

[5] Schmidl, C., A.F. Rendeiro, N.C. Sheffield, and C. Bock. 2015. ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. Nat Methods. 12:963-965.

[6] Robin Genuer, Jean-Michel Poggi, Christine Tuleau-Malot. Variable selection using Random Forests. Pattern Recognition Letters, Elsevier, 2010, 31 (14), pp.2225-2236.

[7] Calo, E., and J. Wysocka. 2013. Modification of enhancer chromatin: what, how, and why? Mol Cell. 49:825-837.

[8] Sims, R.J., 3rd, and D. Reinberg. 2009. Processing the H3K36me3 signature. Nat Genet. 41:270-271.

[9] Ron, G., Y. Globerson, D. Moran, and T. Kaplan. 2017. Promoter-enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains. Nat Commun. 8:2237.