Biophysics 生物物理学, 2018, 6(1), 1-6
Published Online February 2018 in Hans. http://www.hanspub.org/journal/biphy
https://doi.org/10.12677/biphy.2018.61001

Hans汉斯

Prediction of Nucleosome Positioning Sequence for Yeast Genome

Shisai Hu, Yuxiang Chen, Ying Zhang, Jun Lv*

College of Science, Inner Mongolia University of Technology, Hohhot Inner Mongolia
Email: *lujun@imut.edu.cn

Received: Mar. 30th, 2018; accepted: Apr. 16th, 2018; published: Apr. 23rd, 2018

Abstract

Nucleosome is a basic unit of chromatin structure. Its location and distribution on the entire DNA sequence play a key role in the regulation of gene expression in eukaryotes. The prediction of nucleosome positioning with machine learning method has become a hot topic in recent years. Taken the 6-mer component of DNA sequence as the parameter, we used the increment of diversity feature selection technique proposed by us to select eight 6-mers as the classification characteristics. Furthermore, the total accuracy of the 10 fold cross validation is 98.2% using the support vector machine algorithm. The results show that the specific distribution of the k-mer component in the nucleosomal and linker sequences is the main factor that affected nucleosome positioning in yeast.

Keywords

Nucleosome Positioning Sequence, Increment of Diversity, Feature Selection Technology

酵母基因组核小体定位序列预测

胡世赛，陈宇翔，张　颖，吕　军*

内蒙古工业大学理学院，内蒙古　呼和浩特
Email: *lujun@imut.edu.cn

收稿日期：2018年3月30日；录用日期：2018年4月16日；发布日期：2018年4月23日

摘　要

核小体是染色质结构的基本单位，其在整条DNA序列上的定位分布情况，对于真核生物的基因表达调控

*通讯作者。

文章引用: 胡世赛, 陈宇翔, 张颖, 吕军. 酵母基因组核小体定位序列预测[J]. 生物物理学, 2018, 6(1): 1-6.
DOI: 10.12677/biphy.2018.61001

起关键作用。用机器学习方法预测核小体定位成为近年来的研究热点。以DNA序列6-mer组分为参数，采用我们提出的多样性增量特征选择技术，筛选出8个6-mer作为分类特征。进一步，采用支持向量机算法，10折交叉检验的总精度达到98.2%。结果表明，核小体定位序列和连接序列核苷k-mer组分的特异化分布，是影响酵母核小体定位的主要因素。

关键词

核小体定位序列，多样性增量，特征选择技术

1. 引言

核小体是真核生物特有的 DNA 包装方式，是染色质结构的基本组成单位。约 147bp 的 DNA 序列紧密缠绕在组蛋白 8 聚体上约 1.65 圈，形成核小体核心颗粒(nucleosome core particle)，这个核心颗粒阻断了大部分蛋白质分子与 DNA 的接触[1]。核小体核心颗粒之间由长度不等的连接 DNA(linker DNA)相连[2]。核小体和连接 DNA 在基因组上的精确定位，对调控基因的表达过程起到关键作用[3] [4]。

随着对表观遗传学研究热度的提升，基于 DNA 序列信息或序列依赖的各种性质的核小体定位预测模型不断被提出[5]-[10]。核小体在基因组中的分布要受到多种因素的影响，Segal 组的研究表明[5]，DNA 的序列组成是基因组中核小体组织的主要决定因素。Chen 等通过计算序列依赖的二核苷变形能[8]，发现对于酵母基因组，150 bp 长的核小体定位序列与连接序列的变形能存在显著差异，基于此，他们以 98.1% 的精度分类了酵母核小体定位序列与连接序列。如果我们仔细分析酵母基因组核小体定位序列与连接序列的序列组成，就不难发现变形能的这个差异，来源于二者序列组成的差异。最近，Awazu 提出一个多元回归模型[9]，以 3 核苷(或其互补)频次为参数，在 Chen 等使用的数据集上[8]，声称分类精度达到 100%，但 Awazu 的方法参数有 33 个之多[9]，且这些参数的取值强依赖性于数据集。2016 年，Teif 综述了现有的核小体数据资源和在线预测工具[10]。该文总结了近些年分别从生物信息学，物理学和生物学角度研究核小体定位的近百篇论文，为今后的核小体定位研究提供了一个手册。

采用实验手段测定基因组核小体定位情况耗时耗力，如果能预先由生物信息学方法给出一个基因组核小体定位预测的图谱，可能为实验的实施提供预设的靶标，进而节约实验成本。此外，基于生物信息学方法的核小体定位预测，还能发现隐藏在碱基序列中的规律性信息，为解开生命奥秘提供新的线索。因此，在分子生物学研究中，生物信息学手段是必要的。本文，在 Chen 等使用的酵母核小体定位数据集上[8]，我们基于 DNA 序列的 6 核苷组成，采用多样性增量特征选择技术，选出仅 8 个 6-mer 为分类参数，使用支持向量机算法进行分类，获得较高的分类精度。本文的创新点在于，采用了我们研究组新近发展的特征选择方法，在保证预测精度的同时，极大地压缩了模型的输入参数个数，使得该模型具有简单易用，参数少，精度高，鲁棒性好的特点。

2. 材料与方法

2.1. 酵母核小体定位序列和连接序列数据集

实验确定的酵母核小体定位数据取自 Lee 等的实验结果[11]。这个实验结果中给出 4 bp 分辨率的

1,206,683 个 DNA 片段，采用“套索模型(lasso model)”对每个片段分配一个核小体形成能力得分，得分的高低反映了核小体形成的倾向性高低。Chen 等从其中选出 5000 个长度为 150 bp 得分最高的片段作为核小体定位序列[8]，和 5000 个长度为 150 bp 得分最低的片段为连接序列。然后他们使用 CD-HIT 软件过滤相似序列[12]，最后得到序列相似性低于 80%的 3620 个样本，其中 1880 个正样本，即核小体定位序列，1740 个为负样本，即连接序列。用具有过高序列相似性的数据集检验预测模型，可能给出虚高的性能评价结果，本文，我们基于 Chen 等的结果，并进一步采用 BLASTClust (https://toolkit.tuebingen.mpg.de/blastclust)程序，使数据集中的序列相似性低于 30%。最后我们得到 2377 个样本，其中 1334 个核小体定位序列和 1043 个连接序列，序列长度均为 150 bp。

2.2. 方法

2.2.1. 以序列 K-Mer 组分为特征

由 Segal 组的研究结论可知[5]，序列组成是基因组中核小体组织的主要决定因素。那么，基于序列的碱基分布来预测核小体定位序列成为当然的想法。当 k 取值较大时，基于序列 k-mer 组分分布可能重构整个基因组序列，因此，一条序列的 k-mer 分布可以作为该序列的充分表示。我们以 150 bp 长序列中所含的 6-mer 频次作为序列特征。k = 6 对于长度 150 bp 的序列而言，应该是一个充分表示了。

显然，当 k = 6 时，提取的特征总数将达到 4096 个。如果这些特征均输入分类器用于分类，严重的过拟合现象将不可避免。因此，应用特征选择技术实现降维是必要的步骤。在应对小样本或高维数据的模式识别问题时，特征选择是许多机器学习方法的重要组成部分。特征选择技术的主要目标是，从整体特征集中发现一个能够有效描述数据的特征子集。因此，对于从序列提取到的 4096 个 6-mer 频次特征，我们要应用特征选择技术对其进行筛选，选择有效分类信息，使模型具有更好的鲁棒性。

2.2.2. 多样性增量特征选择技术

最近，我们研究组发展了一种新的特征选择技术，称为多样性增量特征选择(feature selection based on incremental of diversity, FSID) [13]。在此之前，我们主要将多样性增量方法结合二次判别分析直接用于模式分类[14] [15]。近些年，众多新的特征选择技术得以发展。2015 年，Drotár 等比较了 10 个目前最好的特征选择技术[16]，结果表明，信息增益(information gain, IG)方法导致最好的稳定性[17]，而最小冗余最大相关(minimum Redundancy Maximum Relevance, mRMR)方法导致最高的预测精度[18]。FSID 方法类似于 IG 方法，但不同之处在于，FSID 方法无需用特征在序列中出现的相对频次去近似特征出现的概率，而直接使用特征的绝对频次。这样的做法对于小样本问题优势明显，不存在理论困难[13]。因为我们知道，当对一个随机事件仅作有限几次观测后，就以观测的频率代替概率会存在很大风险。下面我们对 FSID 方法做一简要描述。

给定一个两分类问题 C_i (i = 1,2)，特征 X 出现在类别 C_1 的频次记为 n，除特征 X 以外的其它特征出现在类别 C_1 的频次记为 $\bar{n}$。则特征 X 在 C_1 类别中的多样性量定义为

$$D(X,C_1)=(n+\bar{n})\log_2(n+\bar{n})-n\log_2(n)-\bar{n}\log_2(\bar{n}) \tag{1}$$

类似地，特征 X 出现在类别 C_2 的频次记为 m，其它特征出现在类别 C_2 的频次记为 $\bar{m}$。按照与(1)式同样的方式，可以分别定义特征 X 在 C_2 类别中的多样性量 $D(X,C_2)$，以及在混合系统 $C_1 + C_2$ 中的多样性量 $D(X,C_1 + C_2)$为

$$D(X,C_2)=(m+\bar{m})\log_2(m+\bar{m})-m\log_2(m)-\bar{m}\log_2(\bar{m}) \tag{2}$$

和

$$D(X,C_1+C_2)=(n+m+\bar{n}+\bar{m})\log_2(n+m+\bar{n}+\bar{m})-(n+m)\log_2(n+m)-(\bar{n}+\bar{m})\log_2(\bar{n}+\bar{m}) \tag{3}$$

特征 X 在 C_1 和 C_2 类别之间的多样性增量(increment of diversity, ID)定义为

$$ID(X)=D(X,C_1+C_2)-D(X,C_1)-D(X,C_2) \tag{4}$$

由以上多样性增量的定义可以看出，当给定样本集时，某个特征 X 在 C_1 和 C_2 两个类别中出现的频次差异越大，$ID(X)$的值越大，而频次差异越小，$ID(X)$的值越小。如果特征 X 是类别无关的，那么一般地，特征 X 在两个类别中出现的频次应几乎无差别。因此，$ID(X)$就可作为特征 X 是否与类别相关的度量。也就是说，如果 $ID(X) > ID(Y)$，表明特征 X 与类别相关性要强于特征 Y。当这种类别相关性的强度达到我们的预期(ID_0)时，即当 $ID(X) > ID_0$，特征 X 被选择，ID_0 为特征选择阈值。阈值 ID_0 是以选出的特征使得预测结果总精度最大化来确定的。这个特征选择方案被我们称为多样性增量特征选择技术(FSID)。

2.2.3. 预测算法

预测算法采用支持向量机，算法实现采用 R 语言“e1071”包中的 svm 函数完成[19]。核函数采用径向核函数，参数 c 和 γ 分别采用默认值。

2.2.4. 分类性能评价指标

分类性能采用如下 4 个指标度量，分别是敏感性(Sensitivity, Sn)，特异性(Specificity, Sp)，总精度(Accuracy, ACC)和马氏相关系数(Matthews correlation coefficient, MCC)。这些量定义如下

$$\mathrm{Sn}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},\mathrm{Sp}=\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}},\mathrm{ACC}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{TN}+\mathrm{FP}},\quad \mathrm{MCC}=\frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FN}\times\mathrm{FP}}{\sqrt{(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TN}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FN})}} \tag{5}$$

这里，TP 是被正确预测的核小体定位序列数，FN 是被错误预测的核小体定位序列数，TN 是被正确预测的连接序列数，FP 是被错误预测的连接序列数。

上面 4 个分类性能评价指标分别表明了一个预报器的四个不同方面的性能。Sn 是在全体正样本中能够被正确预测为正样本的频率，它用来衡量一个预报器识别正样本的能力。类似地，Sp 是用来衡量一个预报器识别负样本的能力。ACC 测度正确识别全部样本的能力。MCC 是预测性能的一个最佳平衡测度。MCC 的取值范围是[−1,+1]。MCC = 0 表明预报器实际执行了一个随机猜测，也即它的预测结果与样本的真实分类标签不相关。MCC = ±1 表明预报器是完美的。同时给出一个预报器的 4 个性能指标，可以较全面地反映出预报器的输出性能。

3. 结果与讨论

3.1. FSID 特征选择结果

基于核小体序列中 6-mer 频数信息，采用 FSID 方法对酵母核小体定位序列和连接序列进行分类预测特征选择。将全部数据样本随机分割为 10 份，保证每份中正负集样本数之比大致与正负总样本数之比相当。合并其中的 9 份样本作为训练样本。在训练样本中，统计所有特征出现的频次，采用 FSID 方法选择 ID 值大于阈值 ID_0 的特征，然后轮换训练样本。如果 10 次轮换某一特征均被选出，则特征被最终选择。将最终选择出的特征送到 SVM 中进行核小体定位序列预测。本文中当阈值 $ID_0 = 615$ 时，预测精度达到

Table 1. Prediction results in 10 fold cross-validation
表 1. 10 折交叉检验预测结果

Method	Accuracy	Sn	Sp	MCC	TP	FP	FN	TN
Our method	98.2%	99.1%	97.7%	0.963	1322	31	12	1012
Chen's method [8]	98.1%	98.2%	98.0%	0.963	-	-	-	-
iNuc-PhysChem [21]	96.7%	97.2%	94.3%	0.936	-	-	-	-

最大，此时选取出 8 个特征参数分别是 AAAAAA、AAAAAT、ATATAT、ATTTTT、TAAAAA、TATATA、TTTTTA 和 TTTTTT。

可以看到，选出的 8 个特征均为 poly(dA:dT)。进一步的分析表明，这些特征均来自连接序列。或者说在酵母基因组中，核小体之间的连接序列中普遍地存在着 poly(dA:dT)片段，而核小体定位序列中则罕有。Poly(dA:dT)序列是刚性的，不利于核小体的形成[20]。这可能是 Liu 等和 Chen 等可以成功地基于序列依赖的二核苷弯曲能方法[7] [8]，来预测酵母核小体定位的原因所在。

3.2. SVM 预测结果

采用 10-fold 交叉检验，酵母核小体定位序列预测结果列于表 1。由表 1 结果可见，我们的模型对酵母核小体定位序列的预测敏感性(Sn)是 99.1%，特异性(Sp)为 97.7%，总精度(ACC)为 98.2%，马氏相关系数(MCC)值达到 0.963。我们的结果与 Chen 等基于序列依赖的二核苷变形能模型的结果，具有相当的精度[8]。但 Chen 等数学模型更为复杂[8]，且其使用的数据集的序列相似性更高。在 2012 年，Chen 等还曾提出一个基于氨基酸物化性质的酵母核小体预测模型[21]，该模型经特征选择后保留了 884 个特征作为输入，在与文献同样的数据集上进行 10 折交叉检验[8]，其预测结果的 4 个性能指标均低于我们的模型，如表 1 所示。

Awazu 提出一个多元回归模型[9]，以 3 核苷(或其互补)频次为参数，在 Chen 等使用的数据集上[8]，声称分类精度达到 100%，但 Awazu 的方法需拟合的参数有 33 个之多[9]，且这些参数的取值强依赖性于数据集，预测精度极不稳定，作者将该方法用于其他物种的核小体预测，发现获得的精度极低[9]。

从以上的分析看出，我们给出的模型具有参数少且精度高的特点，更少的参数将使得模型具有更高的鲁棒性和泛化能力。这个良好的性能得益于我们采用了高效的特征选择技术 FSID。

4. 结论

本文，我们采用多样性增量特征选择技术 FSID，对以核苷六联体(6-mer)为参数的特征集进行筛选，筛选出 8 个 poly(dA:dT)特征对酵母核小体定位序列进行分类预测，在序列相似性低于 30%的数据集上，10 折交叉检验获得 98.2%的高精度结果。模型具有数学方法简单，使用参数少，预测精度高的优点。模型给出的结果还表明，酵母核小体之间的连接序列的主要序列组成特点是，存在普遍的 poly(dA:dT)片段。这些片段具有很强的刚性，不易弯曲，难于形成核小体结构。

尽管由 FSID 方法所选择的特征与类别之间显著相关，但 FSID 方法中并未考虑特征与特征之间的相关性。如何将减少特征之间相关性的算法也融合在现有的 FSID 模型中，是今后研究中需要解决的问题。

基金项目

本项目由内蒙古自治区自然科学基金项目(2015MS0331 和 2016MS0306)资助。

参考文献

[1] Richmond, T.J. and Davey, C.A. (2003) The Structure of DNA in the Nucleosome Core. *Nature*, **423**, 145-150. https://doi.org/10.1038/nature01595

[2] Mavrich, T.N., Ioshikhes, I.P., Venters, B.J., Jiang, C., Tomsho, L.P., Qi, J., Schuster, S.C., Albert, I. and Pugh, B.F. (2008) A Barrier Nucleosome Model for Statistical Positioning of Nucleosomes throughout the Yeast Genome. *Genome Research*, **18**, 1073-1083. https://doi.org/10.1101/gr.078261.108

[3] Clapier, C.R. and Cairns, B.R. (2009) The Biology of Chromatin Remodeling Complexes. *Annual Review of Biochemistry*, **78**, 273-304. https://doi.org/10.1146/annurev.biochem.77.062706.153223

[4] Rando, O.J. and Ahmad, K. (2007) Rules and Regulation in the Primary Structure of Chromatin. *Current Opinion in Cell Biology*, **19**, 250-256. https://doi.org/10.1016/j.ceb.2007.04.006

[5] Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., Gossett, A.J., Tillo, D., Field, Y., LeProust, E.M., Hughes, T.R., Lieb, J.D., Widom, J. and Segal, E. (2009) The DNA-Encoded Nucleosome Organization of a Eukaryotic Genome. *Nature*, **458**, 362-366. https://doi.org/10.1038/nature07667

[6] Wu, J., Zhang, Y. and Mu, Z. (2014) Predicting Nucleosome Positioning Based on Geometrically Transformed Tsallis Entropy. *PLoS One*, **9**, e109395. https://doi.org/10.1371/journal.pone.0109395

[7] Liu, G., Xing, Y., Zhao, H., Wang, J., Shang, Y. and Cai, L. (2016) A Deformation Energy-Based Model for Predicting Nucleosome Dyads and Occupancy. *Scientific Reports*, **6**, 24133. https://doi.org/10.1038/srep24133

[8] Chen, W., Feng, P., Ding, H., Lin, H. and Chou, K.C. (2016) Using Deformation Energy to Analyze Nucleosome Positioning in Genomes. *Genomics*, **107**, 69-75. https://doi.org/10.1016/j.ygeno.2015.12.005

[9] Awazu, A. (2017) Prediction of Nucleosome Positioning by the Incorporation of Frequencies and Distributions of Three Different Nucleotide Segment Lengths into a General Pseudo k-Tuple Nucleotide Composition. *Bioinformatics*, **33**, 42-48. https://doi.org/10.1093/bioinformatics/btw562

[10] Teif, V.B. (2016) Nucleosome Positioning: Resources and Tools Online. *Briefings in Bioinformatics*, **17**, 745-757. https://doi.org/10.1093/bib/bbv086

[11] Lee, W., Tillo, D., Bray, N., Morse, R.H., Davis, R.W., Hughes, T.R. and Nislow, C. (2007) A High-Resolution Atlas of Nucleosome Occupancy in Yeast. *Nature Genetics*, **39**, 1235-1244. https://doi.org/10.1038/ng2117

[12] Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W. (2012) CD-HIT: Accelerated for Clustering the Next-Generation Sequencing Data. *Bioinformatics*, **28**, 3150-3152. https://doi.org/10.1093/bioinformatics/bts565

[13] Yang, S.Q., Hu, S.S., Zhang, Y. and Lv, J. (2017) Application of Feature Selection Technology Based on Incremental of Diversity in Prediction of Flexible Regions from Protein Sequences. *Letters in Organic Chemistry*, **14**, 621-624. https://doi.org/10.2174/1570178614666170221145333

[14] Lu, J. and Luo, L.F. (2008) Prediction for Human Transcription Start Site Using Diversity Measure with Quadratic Discriminant. *Bioinformation*, **2**, 316-321. https://doi.org/10.6026/97320630002316

[15] Lu, J., Luo, L.F., Zhang, L.R., Chen, W. and Zhang, Y. (2010) Increment of Diversity with Quadratic Discriminant Analysis—An Efficient Tool for Sequence Pattern Recognition in Bioinformatics. *Open Access Bioinformatics*, **2**, 89-96. https://doi.org/10.2147/OAB.S10782

[16] Drotár, P., Gazda, J. and Smékal, Z. (2015) An Experimental Comparison of Feature Selection Methods on Two-Class Biomedical Datasets. *Computers in Biology and Medicine*, **66**, 1-10. https://doi.org/10.1016/j.compbiomed.2015.08.010

[17] Yu, L. and Liu, H. (2003) Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. In: Fawcett, T. and Mishra, N., Eds., *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, The AAAI Press, California, 856-863.

[18] Peng, H., Long, F. and Ding, C. (2005) Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**, 1226-1238. https://doi.org/10.1109/TPAMI.2005.159

[19] Chang, C.C. and Lin, C.J. (2011) LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, **2**, 1-27. https://doi.org/10.1145/1961189.1961199

[20] Suter, B., Schnappauf, G. and Thoma, F. (2000) Poly(dA.dT) Sequences Exist as Rigid DNA Structures in Nucleosome-Free Yeast Promoters *in Vivo*. *Nucleic Acids Research*, **28**, 4083-4089. https://doi.org/10.1093/nar/28.21.4083

[21] Chen, W., Lin, H., Feng, P.M., Ding, C., Zuo, Y.C. and Chou, K.C. (2012) iNuc-PhysChem: A Sequence-Based Predictor for Identifying Nucleosomes via Physicochemical Properties. *PLoS ONE*, **7**, e47843. https://doi.org/10.1371/journal.pone.0047843

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2330-1686，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：biphy@hanspub.org