利用机器学习预测 RBP 结合位点方法简述

姓名：张若弛　班级：自 36　学号：2013011551

一、 背景介绍

近期的研究发现，RNA-binding proteins(RBPs)是 RNA 的转录后调控的重要影响因子。所以如果能预测出 RBP 的结合位点,生成相应的 motif 从而对其结合喜好有进一步的认识，是研究后转录基因调控机制的基础。

研究发现，RBPs 含有许多已知的 motif，譬如锌指、RRM 等，这也为使用机器学习方法，学习进而预测 RBP 的结合位点提供了理论基础。而 CLIP-seq（紫外脚链棉衣沉淀及高通量测序）技术的应用和发展提高了识别 RBP 的结合位点的效率和准确率，从而提供了大量的实验数据，即提供了机器学习所必须的数据基础。同时由于 CLIP-seq 技术的一些局限性，使得许多的 RBP 结合位点没能被识别，或者由于噪声影响，数据中有一定的错误数据。综上所述，使用机器学习的方法来预测 RBPs 的结合位点是有着良好的基础，也是十分有意义的。

使用机器学习的方法研究生物问题一般可以具体落实到以下几个方面:①数据预处理与数据选择（数据的编码）②模型的选取与搭建 ③分析结果。以下我就介绍一下我阅读的文献中，各种机器学习的方法是如何编码 CLIP-seq 得到的数据以及使用了什么模型进行学习以及预测的。

二、 方法介绍

其实编码 DNA 数据和选取模型这两者之间本来就有着比较强的联系。因为如果编码方式足够合理科学，只要使用简单的神经网络模型就能取得很好的学习预测效果；对应的，如

果模型本身具有足够强的泛化和学习能力，数据只需要简单地处理即可以输入模型进行学习。这里介绍的几种方法有些选择了“复杂编码模式简单模型”的形式，有些选择了“输入简单的数据使用泛化能力强的模型进行处理”的形式。

首先在数据的预处理和选择上。原始的数据为 RBP 已经实验检测到的结合位点及周边的序列。不同的方法选择了不同的方式从中获取更多的数据。MEMERIS 从序列中计算了 RNA 序列的亲和性（碱基错配的概率，可以理解为较笼统的二级结构信息）作为下一步处理的数据。RNAcontext 进一步区分了不同的二级结构形式（譬如 loop,hairpin,buldge 等），通过预测统计各种二级结构的数目。CapR 进一步考虑不同结构的位置相关性，即对序列位置和二级结构的联合分布进行建模。Graphprot 这个方法是我觉得比较有意思的，它将预测得到的二级结构信息和序列一起编码为一个超图（hypergraph）。最后是课堂展示那篇论文的 Zhang Sai 的方法（以下用其所用的模型 DBN 指代），它综合了序列、二级结构信息以及第一次引入了三级结构信息，并使用限制性玻尔兹曼机进行预学习，将其输出的向量作为数据。

小结一下，不同的方法所使用的数据包括：序列本身、序列的二级结构、序列的三级结构。不同方法选取了一种或者所有的数据类型，并通过图形化或者无监督学习的方法对数据进行了预处理和编码。那么这些方法又使用了哪些不同的模型呢？

除了常见的支持向量机和神经网络之外，这里着重介绍一下 Graphprot 和 DBN 的方法。正如前面提到的 Graphprot 将信息编码为超图后使用了一个基于图像核函数的 SVM 进行学习。即首先将编码得到的超图(包含了序列和二级结构信息)通过一个图像核函数提取特征，然后使用一个标准的 SVM 进行学习（如果想计算具体的亲和度则使用 SVR）。而 DBN 即深度置信网络，它的一个优点在于，由于其结构使得它可以同时学习不同格式的数据。序列信息和及其对应的二级结构信息被分别输入一个基本的限制性玻尔兹曼机（2000 维，1000 维）进行无监督的学习并输出向量，这两组向量和序列对应的三级结构信息（向量，529 维），

一起被输入上一层的限制性玻尔兹曼机输出向量（3000 维），这输出的向量和一个 label 变量一起输入最顶层的网络进行学习。同时这个模型的另一个有点在于，该模型可以直接得到 RBPs 结合位点的 motif。由于深度置信网络是一个生成模型，在固定 label 为 1 的情况下，在最顶层使用基于平均场的吉布斯采样（mean-field based Gibbs sampling，这个不知道翻译是否准确），即可逐层向下输出一组向量（序列及其二级结构等信息）。但这个还不是最终的 motif，其中还包括了大量的“白噪声”。通过固定 label 为 0 的情况下，重复上述操作输出一组向量，两组向量之差才是真正表征 motif 的信息，通过适当处理即可得到 RBPs 结合位点的 motif。

三、 讨论与总结

在原始数据仅仅是序列的情况下，研究该领域的学者创新的通过预测二级结构、三级结构的方式从序列中提取了更多的信息，同时采用了不同的编码方式，譬如利用无监督学习或者图形化的手段；机器学习的模型上也有着各自的特点，都是与其数据编码的方式相匹配，达到了很好的学习的效果。我比较好奇对于 Graphprot 的方法，如果下一层的模型使用卷积神经网络而不是 graph-kernal 的 SVM 效果会怎样，因为直观感觉卷积神经网络在边界变化上有着更高的敏感性（卷积运算强化边界），希望有空可以尝试做一下。

总之，在阅读文献的过程中，我对不同的 RBPs 结合位点的预测方法有了了解，更重要的是能大致感受到研究人员在科研过程中思路和逻辑的变化，这对我自己相关的科研也有很大的启发。

参考文献

[1] D. Maticzka, S. Lange et al., “GraphProt: modeling binding preferences of RNAbinding proteins,” Genome Biology, vol. 15, no. 1, p. R17, 2014.

[2] R. Pudimat, E.-G. Schukat-Talamazzini et al., “A multiple-feature framework formodelling and predicting transcription factor binding sites,” Bioinformatics, vol. 21,no. 14, pp. 3082–3088, 2005.

[3] H. Kazan, D. Ray et al., “RNAcontext: a new method for learning the sequenceand structure binding preferences of RNA-binding proteins,” PLoS ComputationalBiology, vol. 6, no. 7, p. e1000832, 2010.

[4] M. Hiller, R. Pudimat et al., “Using RNA secondary structures to guide sequence motif finding towards single-stranded regions,” Nucleic Acids Research, vol. 34, no. 17,p. e117, 2006.

[5] T. Fukunaga, H. Ozaki et al., “CapR: revealing structural specificities of RNAbinding protein target recognition using CLIP-seq data,” Genome Biology, vol. 15,no. 1, p. R16, 2014.

[6] Sai Zhang, Jingtian Zhou et al., “A deep learning framework for modeling structural features of RNA-binding protein targets”, Nucleic Acids Research