文章编号: 1001-4543(2017)03-0227-04 DOI: 10.19570/j.cnki.jsspu.2017.03.011

蛋白质与 RNA 相互作用的预测和研究

王 彤, 薛建新, 杜 奕

(上海第二工业大学计算机与信息工程学院, 上海 201209)

摘 要: 确定蛋白质与 RNA 是否发生作用非常重要, 因为它广泛存在于生物学过程中, 在生物体细胞活动中起到至关重要的作用。特别是近几年随着蛋白质结构数据的增多, 如果仍用传统的物理化学方法去测定会非常困难, 找到能自动预测蛋白质与 RNA 的相互作用的方法迫在眉睫。首先采用 PsePSSM 算法表达蛋白质序列, 编码后的蛋白质特征向量维数很高; 接着采用 GPP 流形学习方法对其进行维数约简, 约简后的特征向量输入 SVM 分类器训练, 训练好的分类器预测未知的蛋白质与 RNA 是否相互作用; 最后, 采用 Jackknife 测试方法检验预测准确率, 测试结果表明, 上述方法是十分有效的, 为蛋白质与 RNA 是否相互作用的研究提供一条新的思路。

关键词: 蛋白质与 RNA 相互作用; 维数约简; 预测

中图分类号: TP 391; Q 617 **文献标志码**: A

0 引 言

蛋白质与 RNA 的相互作用很重要, 可以体现出蛋白质的功能。同时, 细胞内各种重要的生理过程, 以蛋白质与 RNA 的相互作用为基础, 这些生理过程包括信号的转导等。蛋白质与 RNA 的相互作用在蛋白合成 [1]、病毒复制 [2-3] 和转录调控 [4] 等方面都有广泛的应用。因此, 在生物信息学中, 蛋白质与 RNA 的相互作用的研究占有很重要的地位。

如今, 许多蛋白质与 RNA 复合物的三维结构被测出。如果采用传统的实验方法测定这些生物数据会带来很多问题, 如成本高、耗时长等。因此, 提出采用机器学习算法来预测蛋白质与 RNA 是否相互作用, 可以有效地解决传统实验方法带来的问题。Liu 等 [5] 采用氨基酸序列和结构描述子编码蛋白质序列, 然后采用随机森林算法来预测蛋白质与 RNA 是否相互作用。Kumar 等 [6] 采用氨基酸组成和进化信息编码蛋白质序列, 然后采用支持向量机 (Support Vector Machines, SVM) 来区分蛋白质与 RNA 是否相互作用, 但不能预测未知蛋白质与 RNA 是否相互作用。

本文提出的方法解决了上述的问题。首先, 采用伪特定位点记分矩阵 (Pseudo Position-Specific Scoring Matrix, PsePSSM)[7] 序列编码方法来表示蛋白质与 RNA 序列对。将这种特征提取方法引入到蛋白质与 RNA 相互作用的预测问题中, 能显著提高预测准确率。因为采用这种特征提取方法编码的蛋白质与 RNA 序列对, 包含了蛋白质序列的相似度和进化信息。它尽可能多地保留了蛋白质序列的原始信息, 但同时它可能会导致小样本问题。为了解决这个问题, 本文采用新的降维 (Dimensionality Reduction, DR) 算法从原始高维向量中提取关键特征向量。它有 2 个主要的降维技术: 过滤和包裹。过滤技术与分类算法无关, 与包裹方法相比, 其优点是计算简单快速, 所以很容易应用到非常高维数据集, 如本文的研究中所使用的特征向量。然后, 基于降维后的低维特征向量, SVM 分类器预测蛋白质与 RNA 的相互作用。实验结果表明, 本文的方法是非常有效的, 为蛋白质与 RNA 是否相互作用的研究提供一条新的思路。

1 材料和方法

1.1 数据集

建立一个数据集。首先, 本文检索了 603 个

收稿日期: 2017-04-10

通信作者: 王 彤 (1981–), 女, 河北保定人, 副教授, 博士, 主要研究方向为数据挖掘算法及其应用。
E-mail: wangtong@sspu.edu.cn。

基金项目: 国家自然科学基金 (No.61672022, No.61502296), 上海市自然科学基金 (15ZR1417000) 资助

RNA 结合蛋白复合物, 以 X 射线结晶分析的分辨率 ⩽0.35 nm 进行过滤。去除序列同源性上分别大于 25%的蛋白质和 RNA 链, 得到了 365 个非冗余蛋白 RNA 链。对于负样本集合, 从 UniProt 数据库中收集了 200 个不与 RNA 相互作用的蛋白质序列。

给定一个待查询蛋白质与 RNA 序列对 P, 预测它是否相互作用, 需要做的第 1 件重要的事情是采用适当的编码方法来表达它, 序列编码方法 PsePSSM 就是其中一种。

1.2 PsePSSM 序列编码方法

蛋白质由 20 种氨基酸组成, 每条蛋白质序列用字符串表示, 首先将蛋白质字符序列离散成数值序列。本文采用 PsePSSM 编码方式。下面介绍 PsePSSM [7] 方法。

依据参考文献 [7], PsePSSM 矩阵可以表示为

$$\boldsymbol{P}^{\xi}_{\text{Pse-PSSM}} = [\bar{M}_1 \ \bar{M}_2 \ \cdots \ \bar{M}_{20} \ G^{\xi}_1 \ G^{\xi}_2 \ \cdots \ G^{\xi}_{20}]^{\mathrm{T}} \quad (1)$$
$$\xi = 0, 1, 2, \text{or } L-1$$

式中,

$$\bar{M}_j = \frac{1}{L}\sum_{i=1}^{L} M_{i\to j}, \quad j = 1, 2, \cdots, 20 \quad (2)$$

L 为某蛋白质 P 的长度, 式 (2) 中的分数 $M_{i\to j}$ 表示该蛋白质序列 P 的第 i-th 个位置的氨基酸突变成第 j 种氨基酸的得分。序号 $1, 2, \cdots, 20$ 用来表示 20 种氨基酸中的一种 (按照字母顺序表排列)。利用 PSI-BLAST 程序搜索 Swiss-Prot 得到 $M_{i\to j}$[8]。其中用于 PSI-BLAST 的参数为: 3 次循环, E 值为 0.001。根据 PSSM 的定义, 用如下 $L\times 20$ 的分数矩阵表示蛋白质序列 P [7]。

$$P_{\text{PSSM}} = \begin{bmatrix} M_{1\to 1} & M_{1\to 2} & \cdots & M_{1\to 20} \\ M_{2\to 1} & M_{2\to 2} & \cdots & M_{2\to 20} \\ \vdots & \vdots & \vdots & \vdots \\ M_{i\to 1} & M_{i\to 2} & \cdots & M_{i\to 20} \\ \vdots & \vdots & \vdots & \vdots \\ M_{L\to 1} & M_{L\to 2} & & M_{L\to 20} \end{bmatrix} \quad (3)$$

式 (1) 中的分量 $G^{\xi}_1, G^{\xi}_2, \cdots, G^{\xi}_{20}$ 被定义为

$$G^{\xi}_j = \frac{1}{L-\xi}\sum_{i=1}^{L-\xi}[M_{i\to j} - M_{(i+\xi)\to j}]^2 \quad (4)$$
$$j = 1, 2, \cdots, 20; \xi < L$$

式中: G^{ξ}_j 是相关因子; ξ 为氨基酸的间隔距离; G^{i}_j 为相隔 $i-1$ 个残基的氨基酸残基之间的 PSSM 得分情况。ξ 取值须小于 50, 因为本文采用的数据库中蛋白质序列最短长度为 50。当 $\xi = 0$ 时, G^{ξ}_j 为空, 式 (1) 为前 20 维向量。

根据式 (1), 1 条蛋白质序列表示为 1 个高维向量包含 1 个 20 维的向量 ($\xi = 0$) 和 49 个 40 维的向量 ($\xi = 1, 2, \cdots$, 或 49)。这 49 个 40 维的向量中, $\boldsymbol{P}^{\xi}_{\text{Pse-PSSM}}$ 中前 20 维的向量都是相同的, 去掉重复的向量, 保留 1 个 20 维向量。得到 1 个 1 000($20 + 49\times 20$) 维的向量 $\boldsymbol{P}^{\xi}_{\text{Pse-PSSM}}$。1 000 维的高维特征向量会使预测系统复杂化。这里, 引入 GPP 方法来解决这个难题。

1.3 GPP 降维算法

几何保留投影 (Geometry Preserving Projections, GPP)[9] 是一种线性降维算法。GPP 的思想是保留局部的信息, 通过捕捉特征空间类间的几何属性和类内的几何性质来实现降维。对于关于 GPP 的概念更详细的描述, 参见文献 [9]。下面, 简要介绍一下 GPP。

数据集 $\boldsymbol{X} = [\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_N]$ 是由 m- 维的实数空间 R^m 内给出的, 数据集包含 C 个类别 $[\Phi_1, \Phi_2, \cdots, \Phi_C]$, 并且每个数据点 $\vec{x}_i (i = 1, 2, \cdots, N)$ 分别属于某一个类别。算法将原始数据 $\boldsymbol{X}$ 通过投影矩阵投影到低维空间 $R^d (d < m)$。目标是找到最优的投影矩阵 $\boldsymbol{B}$:

$$\boldsymbol{Y} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{X} \quad (5)$$

为了得到最优的投影矩阵 $\boldsymbol{B}$, 需最小化如下的目标函数:

$$\sum_{i=1}^{N}\left\|\boldsymbol{B}^{\mathrm{T}}\vec{x}_i - \sum_j w_{ij}\boldsymbol{B}^{\mathrm{T}}\vec{x}_j\right\|^2 - \theta\sum_{i=1}^{N}\sum_{j=1}^{N}\left\|\boldsymbol{B}^{\mathrm{T}}\vec{x}_i - \boldsymbol{B}^{\mathrm{T}}\vec{x}_j\right\|^2 s_{ij} \quad (6)$$

式中, $w_{ij} (i, j = 1, 2, \cdots, N)$ 为系数矩阵, $s_{ij} (i, j = 1, 2, \cdots, N)$ 为相似度矩阵。θ 为尺度因子, 其范围

是 $[0,1]$。

想要最小化目标函数, 即等价于最小化其被减数和最大化其减数:

(1) 最大化其减数, 即

$$\max \sum_{i=1}^{N}\sum_{j=1}^{N}\left\|\boldsymbol{B}^{\mathrm{T}}\vec{x}_i-\boldsymbol{B}^{\mathrm{T}}\vec{x}_j\right\|^2 s_{ij}$$

满足条件:

$$s_{ij}=\begin{cases}0, & \vec{x}_i和\vec{x}_j属于不同的类别\\ 1, & 否则\end{cases} \tag{7}$$

(2) 最小化其被减数, 即

$$\min \sum_{i=1}^{N}\left\|\boldsymbol{B}^{\mathrm{T}}\vec{x}_i-\sum_j w_{ij}\boldsymbol{B}^{\mathrm{T}}\vec{x}_j\right\|^2$$

如果 $\vec{x}_i$ 和 $\vec{x}_j$ 属于不同的类别, $w_{ij}=0$, 并且 $\sum_j w_{ij}=1$。

算法计算的样本点都标记过类别信息, 不包含未标记的样本子集

$$\boldsymbol{X}_{\mathrm{Unlabel}}=[\vec{x}_{L+1},\vec{x}_{L+2},\cdots,\vec{x}_N]$$
$$w_{ij}=0, e_{ij}=0(i,j=L+1,L+2,\cdots,N)$$

式 (6) 可以简化为:

$$\min \mathrm{tr}\left\{\boldsymbol{B}^{\mathrm{T}}\boldsymbol{X}(M-\theta L)\boldsymbol{X}^{\mathrm{T}}\boldsymbol{B}\right\} \tag{8}$$

为了能唯一地确定矩阵 $\boldsymbol{B}$, 施加约束 $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}=I$, 也就是, $\boldsymbol{B}$ 的列向量之间是正交的。现在, 目标函数可以写成如下的形式:

$$\left.\begin{aligned}&\arg\min_{U}\mathrm{tr}\left\{\boldsymbol{B}^{\mathrm{T}}\boldsymbol{X}(M-\theta L)\boldsymbol{X}^{\mathrm{T}}\boldsymbol{B}\right\}\\ &\text{s.t. }\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}=I\end{aligned}\right\} \tag{9}$$

上述问题转化为特征值求解问题:

$$\boldsymbol{X}(M-\theta L)\boldsymbol{X}^{\mathrm{T}}\vec{\beta}=\lambda\vec{\beta} \tag{10}$$

设列向量 $\vec{\beta}_1,\vec{\beta}_2,\cdots,\vec{\beta}_d$ 为式 (10) 的解, 其根据相应的特征值排序, $\lambda_1<\lambda_2<\cdots<\lambda_d$。这样, 最优的转换矩阵可以写成:

$$\boldsymbol{B}=\left[\vec{\beta}_1,\vec{\beta}_2,\cdots,\vec{\beta}_d\right] \tag{11}$$

上述的 GPP 算法能有效地避免小样本问题, 因为它不同于以往的算法, GPP 算法没有矩阵的逆运算, 避免了受到奇异值问题的困扰。

2　结果与讨论

用 GPP 算法针对 1 000-D 特征向量降维, 最终得到 70-D 特征向量。分别输入 SVM 和 K 近邻 (K Nearest Neighbor, KNN) 分类器进行训练, 训练好的分类器用来预测蛋白质与 RNA 是否相互作用。

采用 Jackknife 测试本文提出方法的准确率 [10]。为便于比较, 没有采用 GPP 算法得出的结果也列在表 1 中。从表 1 可以看出, 采用 GPP 降维算法和 SVM 分类器后的 Jackknife 测试可以获得超过 98%的准确率, 这比没有采用 GPP 算法得到的准确率高约 5%。实验结果表明, 通过降维冗余信息被去掉了, 同时原始数据中有用的信息被保留了下来。所以预测系统得到了简化, 同时分类准确率还提高了。原始向量由 1 000 维降到了 70 维, 小于样本数, 小样本问题也得到了解决。

表 1　采用不同的方法预测蛋白质与 RNA 的相互作用的 Jackknife 准确率

Tab. 1　The Jackknife success rates for protein-RNA interaction prediction by different methods

方法	输入形式	准确率/%
KNN	原始 1 000 维向量	91.3
SVM	原始 1 000 维向量	93.5
GPP & KNN	GPP 约简后的 70 维向量	96.7
GPP & SVM	GPP 约简后的 70 维向量	98.1

此外, 需调整 KNN 分类器中的最近邻数 K, K 的取值大小会影响分类的性能。基于不同 K 下的预测准确率如图 1 所示。可以看出, 当 $K=1$ 时采用 GPP 算法得出的预测准确率为最大值, 不采用 GPP 算法时, 预测准确率最大值也在 $K=1$ 时取得。

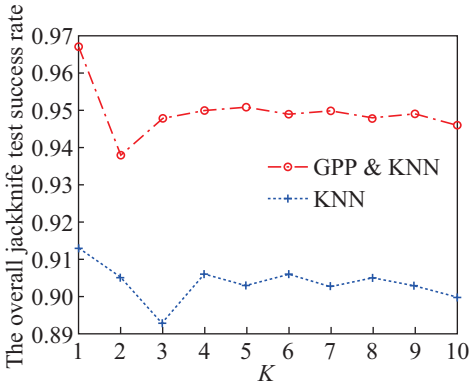

图 1　采用 KNN 方法当 K 取不同值时的 Jackknife 预测准确率比较结果

Fig. 1　The comparison results of Jackknife prediction success rates obtained by KNN algorithms with different K

3 结 语

本文所提出的方法在预测蛋白质与 RNA 相互作用方面是非常有效的, 现有的预测主要集中在寻找最佳的分类方案, 笔者则是从另外一个角度考虑简化生物系统的复杂性。本文应用 GPP 算法从高维空间中提取关键信息, 同时还解决了小样本问题, 基于降维后的特征向量利用 SVM 来预测蛋白质与 RNA 是否相互作用。结果表明, 该方法降低了预测系统的复杂性, 解决了小样本问题, 同时还提高了预测的准确率。

参考文献:

[1] BEAUDOIN M E, POIREL V J, KRUSHEL L A. Regulating amyloid precursor protein synthesis through an intemal ribosomal entry site [J]. Nucleic Acids Res, 2008, 36(21): 6835-6847.

[2] NEWCOMB L L, KUO R L, YE Q, et al. Interaction of the influenza a virus nucleocapsid protein with the viral RNA polymerase potentiates unpfimed viral RNA replication [J]. J Virol, 2009, 83(1): 29-36.

[3] YU Z, SANCHEZ-VELAR N, CATRINA I E, et al. The cellular HI V-1 Rev cofactor hRIP is required for viral replication [J]. Proc Natl Acad Sci USA, 2005, 102(11): 4027-4032.

[4] ABDELMOHSEN K, KUWANO Y, KIM H H, et al. Post-transcriptional gene regulation by RNA-binding proteins during oxidative stress: Implications for cellular senescence [J]. Biol Chem, 2008, 389(3): 243-255.

[5] LIU Z P, MIAO H. Prediction of protein-RNA interactions using sequence and structure descriptors [J]. Ifac Papersonline, 2016, 48(28): 28-34.

[6] KUMAR M, GROMIHA M M, RAGHAVA G P. SVM based prediction of RNA-binding proteins using binding residues and evolutionary information [J]. J Mol Recognit, 2011, 24(2): 303-313.

[7] CHOU K C, SHEN H B. MemType-2L: A web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM [J]. Biochemical and Biophysical Research Communications, 2007, 360(2): 339-345.

[8] 王彤, 薛建新, 孔亮亮. 细菌性病原体内病毒蛋白的预测和研究 [J]. 上海第二工业大学学报, 2016, 33(3): 231-235.

[9] ZHANG T H, LI X L, TAO D C, et al. Multimodal biometrics using geometry preserving projections [J]. Pattern Recognition, 2008, 41(3): 805-813.

[10] 王彤, 薛建新, 谭文安. 利用半监督降维算法预测蛋白质亚细胞位置 [J]. 上海第二工业大学学报, 2015, 32(3): 260-265.

The Prediction and Research of RNA-Protein Interactions

WANG Tong, XUE Jianxin, DU Yi

(School of Computer and Information Engineering, Shanghai Polytechnic University, Shanghai 201209, China)

Abstract: It is very important to determine whether RNA and protein interacts or not. Because it is widely present in the biological process and plays a vital role in the biological cell activity. Especially, in recent years, with the increase of protein structure data, it is very difficult to determine the interaction between protein and RNA with traditional physical and chemical methods. It is imminent to finding a way to predict the interaction between proteins and RNA. Firstly, the PsePSSM algorithm was used to express the protein sequence. The feature vector dimension of the encoded protein was very high. Then the GPP manifold learning method was used to reduce the dimension of the protein. The reduced feature vector was input into the SVM classifier, and the trained classifier predictors were used to predict whether the unknown protein interacted with the RNA. Finally, the Jackknife method was used to test the accuracy of prediction. The results showed that the method was very effective. It can provide a new way to study the interaction between protein and RNA.

Keywords: interaction between protein and RNA; dimensionality reduction; prediction