第5章　RNA排序预测

随着生命科学和计算机科学的迅猛发展，其专业技术相互结合形成一门新兴的学科，即生物信息学(Bioinformatics)。它通过综合利用生物学、计算机科学和信息技术揭示大量而复杂的生物数据所具有的生物学奥秘，是当今生命科学和自然科学的重大前沿领域之一。其研究重点主要体现在基因组学(Genomics)和蛋白质组学(Proteomics)两方面，也即从核酸和蛋白质序列出发，分析序列中表达的结构功能的生物信息。

本项挖掘任务从生物信息学和计算生物学的角度，分别对内部核糖体进入位点(Internal Ribosome Entry Site，IRES)和冷冻电镜图像蛋白质颗粒挑选进行数据挖掘与预测，也即分别进行生物领域内的文本和图像挖掘。

5.1　概　　述

蛋白质与RNA的相互作用广泛存在于RNA剪切、翻译、病毒的复制以及细胞中的其他生物学过程中，因此，探讨蛋白质与RNA相互作用并确定蛋白质中与RNA结合的氨基酸残基，对于理解蛋白质与RNA之间的相互作用机制具有重要的意义。为了在只给定一条蛋白质序列的情况下，判断它的哪些位点是RNA结合高发区，哪些是不容易发生

RNA 结合的位点，本项任务选取最近刚通过质粒筛选试验产生突破性进展的内部核糖体进入位点作为数据来源，通过机器学习的手段，结合蛋白质序列的特征提取，实现在氨基酸残基的水平上对蛋白质序列中的 RNA 结合位点进行预测，从而完成文本的数据挖掘工作。

通过研究蛋白质颗粒的微观三维结构，可以识别其具有的功能。蛋白质颗粒三维结构的获取如今主要有三种方法，分别是 X 射线衍射法、核磁共振法和三维电镜重构法。在冷冻电镜的三维重构法中，首先需要挑选出大量的二维投射样本，然后利用一定的三维重构技术重构出其三维空间结构。随着重构精度逐渐要求到原子级水平，待挑选的蛋白质颗粒也达到了上万乃至上百万的水平。在如此庞大数据量的挑战下，人工的挑选成为制约该技术发展的一个主要瓶颈。随着计算机视觉技术的发展，已经有许多模式识别算法被应用到蛋白质挑选的领域中，主要的方法便是基于模板匹配和基于特征学习。本项挖掘任务选取了已经被解析出蛋白质结构的 TRPV1 作为数据来源，基于上述特征选取、机器学习手段，进行蛋白质颗粒挑选，从而完成图像的数据挖掘工作。

5.2 研发现状

5.2.1 内部核糖体进入位点的数据挖掘研发现状

真核生物的大多数蛋白质合成采用依赖帽子结构的翻译起始方式。但一组缺乏帽子结构的 RNA 病毒蛋白质合成起始方式依赖于其 5′端非翻译区（Untranslated Region，UTR）翻译调控的顺式作用元件——内部核糖体进入位点。

它们能够在一些反式作用因子的辅助下，招募核糖体小亚基到病毒 mRNA 的翻译起始位点。目前，依赖 IRES 元件翻译起始的 RNA 病毒在哺乳动物、无脊椎动物及植物中均有发现，因此，对 RNA 病毒 IRES 元件的深入研究，不仅有助于阐明相关疾病的发生机理，而且能够为工业应用和疾病治疗提供借鉴意义。IRES 的结构如图 5.1 所示。

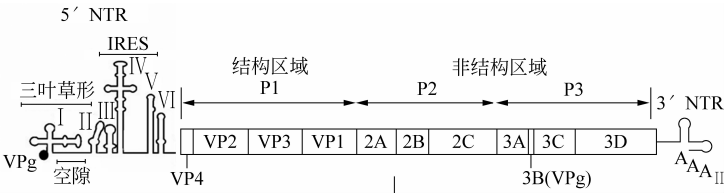


图 5.1　IRES 的结构示意图

IRES 是一段与众不同的 RNA 分子，能够募集真核生物核糖体到 mRNA 分子 5′UTR 上进行翻译起始。这个过程就是内部起始翻译。具备 IRES 元件的 RNA 病毒 5′端不含有帽子结构，GC 含量较高，具有复杂而且稳定的二级和三级结构，如稳定的颈环结构。但是，在一级和二级结构上还没有发现它们的共同特征。

目前，针对 IRES 结构是否在 RNA 序列中存在的预测性研究主要从实验与生物信息学两个方面入手。从实验上看，一种是通过 X 射线晶体衍射、核磁共振等方法得到蛋白质与 RNA 复合物的三维结构信息，从而发现基于三维结构信息进一步确定与蛋白质中和 RNA 相互作用的氨基酸残基；另一种是通过高通量的筛选和已有的序列进行对比，从而定性地筛选出包含有 IERS 的蛋白质 RNA 序列。实验的优点是结果可靠；缺点在于时间、人力和经费方面的花费较

大，并且在具体实施时，还面临着不少实际问题。例如，某些蛋白质-RNA 复合物结晶很难获得，某些序列合成后可能会不稳定。随着蛋白质结构数据的增多，研究人员开始尝试从生物信息学角度出发对这个问题进行研究与分析。从研究途径看，可以分为通过 RNA 结合结构域来判定、通过分子动力学模拟来判定以及通过统计分析或者机器学习方法来判定这三个方面。结构域方法即通过一些已有的蛋白质结构数据库（如 SCOP 等）进行检索，并确定蛋白质中 RNA 结合结构域所在的位置，从而大致确定该蛋白质与 RNA 的结合位点，该方法的缺陷在于仅适用于已测定了 RNA 结合结构域的蛋白质。此外，目前对 RNA 结合结构域的作用机制尚未完全阐明，存在着结构域中的氨基酸残基不与靶标 RNA 区域结合，而是结合到其他区域，甚至导致该蛋白质结合到另一个蛋白质上的情况。分子动力学模拟是另一种寻找 RNA 作用位点的方法，通过一些算法进行动力学（或热力学）的模拟，可以较为直观地观察到蛋白质与 RNA 的结合过程以及这个过程中一些能量和构象上的变化。该方法的缺陷在于模拟耗时较长，仅适用于小规模体系。此外，各种参数的设定也对模拟结果的准确性有影响。综上所述，如果从生物信息学角度出发对该问题进行研究，比较适合的途径是通过提取各种特征，利用机器学习方法构建模型来判别。

虽然在其他生物问题领域已经有了基于机器学习的预测方法，但是在 IRES 这一问题中，由于之前根据实验所积累的数据量较小（数据库 ires.org 中只累计了大约 100 条数据），如若采用上述手段，可能存在偏性。同时，有些特征需要从二维、三维结构数据中获取，或者是比较复杂、难以计算

的理化特征，限制了其应用。在本问题中，由于最近在 *Science* 上发表的一篇基于质粒构建高通量筛选的论文揭示了大量的包含 IRES 的 RNA 序列数据（约 10 000 条），从而得以进行大规模的数据模型构建，并充分考虑特征与二维、三维结构，进行文本挖掘的工作。

5.2.2 冷冻电镜图像蛋白质颗粒挑选研究现状

冷冻电镜三维重构是结构生物学研究中的一种较新的技术。它的基本技术路线为：利用快速冷冻技术对样品进行冷冻固定，然后利用冷冻电镜和低剂量成像技术对样品进行电子成像，利用高灵敏底片进行成像记录，利用高分辨扫描仪对底片进行数字化，对数字化的图像进行二维图像分析——选点、分类、校正和平均，最后完成样品的三维重构计算，如图 5.2 所示。

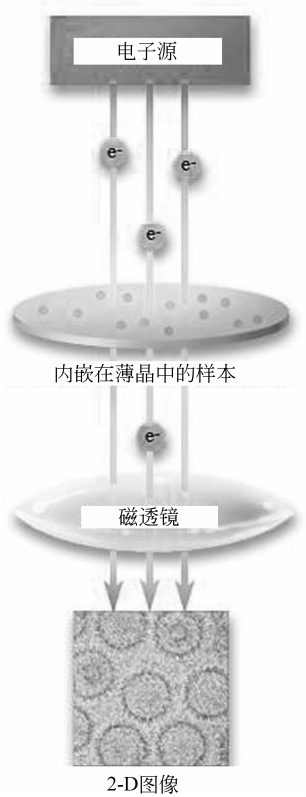


图 5.2 冷冻电镜二维成像流程图

和传统的 X 射线晶体学和核磁共振的技术相比，冷冻电镜具有可以直接获得分子的形貌信息和相位信息，能够解析

那些不适合应用X射线晶体学和核磁共振技术进行分析的样品等优点，已经成为一种公认的研究生物大分子、超分子复合体及亚细胞结构的有力手段。

然而，为了避免电子对样品的损伤，冷冻电镜必须在极低的电子剂量下成像，因而其图像信噪比非常低，如图5.3所示。从一副冷冻电镜照片中挑选有颗粒的部分和没有颗粒的部分，其灰度直方图的形状几乎没有差别。因此，剔除噪声，增加信噪比，提高颗粒图像挑选的精度，是冷冻电镜技术的关键问题。冷冻电镜图像中具有高质量的图像通常是随机的出现，并且只出现在一定的区域内，很难对其进行一定的控制。另外，由于低剂量的电子辐射使得图像的信噪比非常低，要提高信噪比，就必须采集更多的图像数据，通常需要10 000张才能满足分子分辨率的要求，而要获得原子分辨水平(大约4Å)的结构需要上百万张图像，若要人工处理这些数据几乎是不现实的。所以，挑选大量颗粒图像已经成

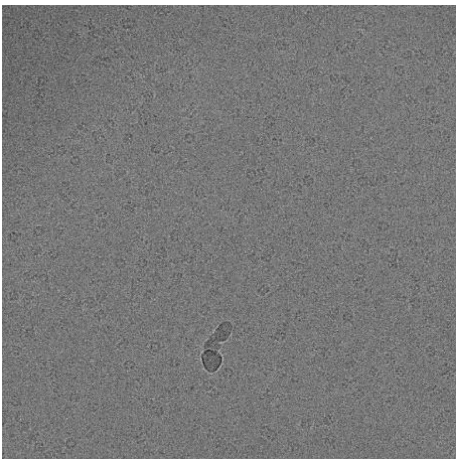


图5.3　一张典型的冷冻电镜二维成像图片

为冷冻电镜蛋白质颗粒三维重构的一个瓶颈，因此迫切需要发展自动化采集单粒子图像的方法。

如上所述，在冷冻电镜结构解析的工作中，目前有相当大的一部分时间和人力都消耗在重复性的工作中，例如电镜胶片的拍摄、电镜图像的预筛选、电镜图像粒子的挑选等。这些繁杂而重复性高的工作会大大减缓研究的进程，实际上已经成为阻碍当前高分辨率的冷冻电镜解析的主要瓶颈之一。为了解决这一问题，本项挖掘任务提出了基于机器学习的自动化对冷冻电镜二维图像的颗粒挑选，从而进行图像挖掘的工作。

5.3 工作设计与实现

5.3.1 基本的设计框架与实现思路

本工作将对生物信息学中两大分支性的数据分别进行数据挖掘的工作。

1. 内部核糖体进入位点的文本挖掘设计

1) 数据来源

序列的数据来源于已有的 IRES 数据库 www.ires.org，以及在 2016 年 1 月发表在 *Science* 上的 *Systematic Discovery of Cap-Independent Translation Sequences in Human and Viral Genomes*。其中，有五个经过实验的数据集，每一个数据集均包含人类或者病毒身上不同位置(5′UTR、3′UTR 以及转录区)中包含或者不包含 IRES 的基因序列。其中，最多数据的数据集为人类的 5′UTR 区域，包含 906 个正样本和 9031

个负样本。

2）数据预处理与数据清洗

RNA 的一级结构序列本身是一种文本，预处理时分为特征选取、增加特征并整合、降维三部分进行。由于数据来源格式较为规整，每条序列均已经分成了 174 的长度，故不需要进行数据的归一化处理（实际操作时将 A，T，C，G 转化成了 1，2，3，4 与向量）。

特征选取：采用 *k*-mer 的方式对文本进行特征提取，即选取文本中相邻的 *k* 个字符组成一组并沿着序列进行一维的平行移动从而遍历，统计该组的出现次数，实际操作时如图 5.4 所示。

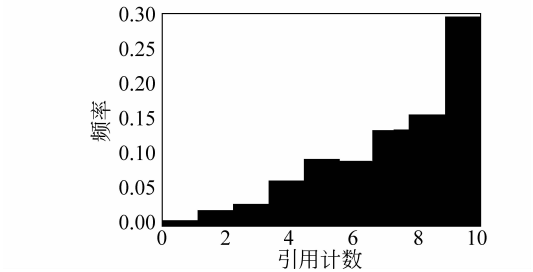
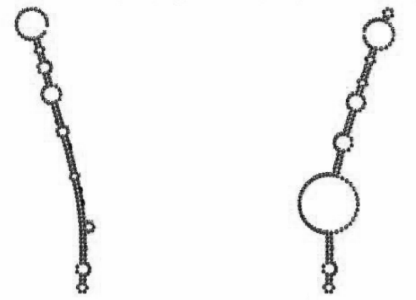


图 5.4　进行 *k*-mer 处理后的统计直方图

（带单引号的 1，2，3，4 分别代表 A，T，C，G，不带单引号的数字表示统计出现的次数）

增加特征并整合：由于只通过一维的序列特征提取过于单一，事实证明效果也不好，故增加二维的结构预测作为另一特征，加入到处理后的数据向量中。这里使用了 RNAfolder 软件，基于热力学统计规律，对 RNA 的序列进行二维

的结构预测，得到的结果如图 5.5 所示。



(a) MFE第二结构　　(b) 中心点第二结构

图 5.5　二维结构预测结果示意图

降维：采用主成分分析的方法，对特征进行特征抽取，将原特征进行线性变换，映射至低维空间中。

3）数据挖掘

采用基于机器学习的模型对经过预处理的数据进行学习，选择更优的模型并调参，从而达到更好的预测效果。由于 Python 的 Sklearn 包的便捷性，本挖掘任务将 KNN、SVM、随机森林、多层感知机、Logstic 回归和神经网络均进行了一定的尝试，同时利用 Kares 包和 MATLAB 实现了卷积神经网络。

4）数据可视化

利用 Python 的 Matplotlib 包和 MATLAB 的工具箱可以便捷地达到数据可视化的目的。同时，很多在线的可视化软件也值得使用。

2. 冷冻电镜图像蛋白质颗粒挑选的图像挖掘设计

1）数据来源

图像的数据来源于 2013 年发表在 *Nature* 上的 *Structure*

of the TRPV1 Ion Channel Determined by Electron Cryo-Microscopy 中已经解析出结构的 TRVP1 蛋白的冷冻电镜图像。其中，包含 60 张 3710×3710 的高分辨率图像，以及人工选取的颗粒点在图像中的坐标位置。

2）数据预处理与数据清洗

初步观察数据集中的图像，其信噪比非常低，同时包含可能会干扰结果的气泡现象。所以，对这一图像挖掘工作的预处理分为数据清洗（气泡探测）、数据归一化、特征提取三个步骤进行。

数据清洗（气泡探测）有两种方法。第一种，首先对图像进行高斯滤波，再进行中值滤波，然后通过 *k*-means 方法聚类出图像中气泡所在位置，如图 5.6 所示。第二种，对图像直接采用腐蚀的方法，进行形态学上的重建，如图 5.7 所示。之后，采用没有颗粒处的图像均值填补这些有气泡的图像区域。

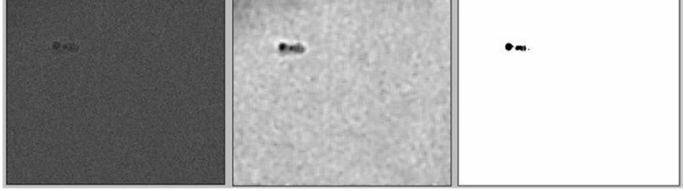


图 5.6　使用 *k*-means 检测气泡区域

数据归一化：对于图像矩阵中的每一个点 z，采用 $z:=\frac{z-\min}{\max-\min}$ 进行归一化，然后进行高斯滤波与中值滤波，最后进行直方图均衡化，如图 5.8 所示。