机器学习在生物信息学领域的应用与研究进展

□文 / 汤胜男、辛学刚

汤胜男　　华南理工大学医学院生物医学工程在读博士生。

华南理工大学教授，医学影像技术系主任，学科带头人，博士，博士生导师、博士后合作导师。牵头主持国家重点研发计划项目 1 项，主持国家自然科学基金重点项目、面上项目等 4 项，研究兴趣集中在医工结合的生物医学基础研究及应用研究方面。

辛学刚

人工智能是计算机科学领域的一个重要分支，机器学习是人工智能领域重要的组成成分，机器学习使计算机能够模拟人类的学习行为，自发地通过学习来获得知识和生活技能，也能在学习的过程中不断改善自身性能，从而实现自我改善。生物信息学是将数学和计算机科学应用于生物分子信息的索引、分类与分析等方面的一门交叉学科，目的是研究生命科学中各种生物信息所代表的生物学意义。由于生物信息学领域数据的特点和需求，人工智能领域特别是机器学习很多算法已经在该领域应用广泛，有力地推进了生物信息学的发展。本文就机器学习技术在生物信息学中的应用及研究进展作一综述。

一、引言

人工智能（Artificial Intelligence, AI）是计算机科学领域一个重要的分支，可概括为通过计算机程序来呈现人类智能活动规律的技术系统 [1,2]。机器学习（Machine Learning）是人工智能领域重要的组成部分，也是实现人工智能的一个重要途径。机器学习使计算机能够模拟人类的学习行为，自发地通过学习来获得知识和生活技能，也在学习的过程中不断改善自身性能，从而实现自我改善 [3]。为了达到上述目的，机器学习的主要研究内容就是从数据中学习特定任务、开发随着经验而改进的计算机算法。

生物信息学（Bioinformatics）是将数学和计算机科学应用于生物分子信息的索引、分类与分析等方面的一门交叉学科，目的是研究生命科学中各种生物信息所代表的生物学意义 [4,5]。生物信息学的研究随着基因组研究的发展而发展，通过分析和解读基因组相关信息，来理解生命科学中生长发育、分化、疾病发生发展等过程。生物信息学领域数据结构复杂、种类繁多，数据量增长迅速，并且生物数据来源具有多样性和复杂性。

对于繁杂的生物数据，一方面要解决海量数据的存储和管理问题，一方面要能够在尽量保证反映生物学真实意义数据的前提下，从数据中提取有效的信息。对于第二个问题，要求应用于生物信息学领域的方法，不仅要对生物数据进行建模，还要能够在同时具有生物学意义和价值的基础上有新的发现。这些问题对于计算机科学领域来说，是机会也是挑战，而人工智能技术作为计算机科学的重要部分，也已经在生物信息学领域有了广泛的应用，取得了很大的成功。机器学习作为实现人工智能的重要方法，无需显式编程即可处理机器的自动学习，主要内容是执行基于数据的预测，在生物信息学领域已经应用广泛 [6,7]。针对于机器学习中的监督式学习、无监督学习、半监督学习以及神经网络在生物信息学中的研究与应用简要介绍如下。

二、监督式学习

监督式学习（Supervised Learning）算法是指那些需要外部帮助的算法。算法输入的数据集为训练数据集和测试数据集，训练数据集含有需要预测或分类的输出变量。所有算法都从训练数据集中学习某种模式，并将其应用于测试数据集以进行预测或分类。

决策树

决策树（Decision Tree）是根据属性值来进行排序并且进行分组的树类型，主要用

[1]Charniak E.Introduction to artificial intelligence[M].Pearson Education India,1985.
[2]Haugeland J.Artificial intelligence:The very idea[M].MIT press,1989.
[3]Michie D,Spiegelhalter D J,Taylor C C.Machine learning[J].Neural and Statistical Classification,1994,13(1994):1-298.
[4]Bergeron B P.Bioinformatics computing[M].Prentice Hall Professional,2003.
[5]Baldi P,Brunak S,Bach F.Bioinformatics:the machine learning approach[M].MIT press,2001.
[6]Larranaga P,Calvo B,Santana R,et al.Machine learning in bioinformatics[J].Briefings in bioinformatics,2006,7(1):86-112.
[7]Libbrecht M W,Noble W S.Machine learning applications in genetics and genomics[J].Nature Reviews Genetics,2015,16(6):321-332.

于分类。每棵树都由节点和分支组成，每个节点表示要分类的组的属性，每个分支表示节点可以采用的值。变异检测是二代测序数据分析中的关键链接，包括将一个或者多个样本的 reads 比对到基因组、检测变异位点和鉴定出每个变异位点基因型等步骤，变异检测的准确性会影响数据的下游分析，从而影响分析结果。为了提高变异检测的准确性，Li Z 等人 [8] 研究了基于决策树的变异检测算法，并且在全基因组测序数据，全基因组捕获数据和低覆盖率数据中都进行了应用。miRNA 是短的、非编码 RNA，仅包含约 22 个核糖核苷酸，与平均长度为 1500 个核糖核苷酸的 mRNA 相比，它非常短，但是每个 miRNA 都能够与不同的 mRNA 形成复合物，从而调节这些 mRNA 的功能。这个过程可以说成 miRNA 与 mRNA 的相互作用，也可以说成 mRNA 是 miRNA 的靶标。因此，近年来与 miRNA 相关的研究逐渐增加，如基于现有积累的生物学数据开发基于决策树的计算模型，来推断 miRNA 与疾病之间的关联 [9] 和提高 miRNA 靶蛋白预测的准确性 [10] 等方面。细胞内部的蛋白质复合物网络可以协调许多生物学过程，这个网络被称为蛋白质互作网络。蛋白质互作网络含有比个别蛋白作用更多的系统信息，网络中形成的复合物可以执行众多细胞功能。为了解读生物体中给定生物学功能的分子机制，正确识别蛋白质之间相互作用的网络至关重要。蛋白质互作网络可以建模为图的形式，图中的边代表蛋白质之间的相互作用，子图代表蛋白质复合物。因此，需要同时考虑拓扑和生物学特征的算法，从而正确识别蛋白质互作网络内部具有变化的拓扑结构和具有生物学特征的蛋白质复合物。Sikandar A 等人 [11] 就提出了基于决策树从蛋白相互作用图来检测蛋白复合物的算法，并取得了较好的结果。但是决策树不适用于目标是预测连续属性值的估算任务，而且在相对较多和相对较少类别的情况下，决策树容易在分类问题中出错，除此之外，大多数决策树算法一次只检查一个字段，导致矩形分类框与决策空间中记录的实际分布不太吻合 [12]。

支持向量机

支持向量机（Support Vector Machine, SVM）是最近广泛使用的一种机器学习技术，按照边距计算的原理，在两个类别之间创建一个决策边界，使边距与类别之间的距离最大，从而使分类的误差最小。蛋白质的三维结构对于详细了解生物分子的功能至关重要，已知蛋白质序列的数量与其实验解析的三维结构之间存在巨大差距 [13]。就目前来说，蛋白质结构的成功预测是弥合这一差距的较为实用的方法，而蛋白质结构预测中主要的两个步骤就是对给定的蛋白质序列生成大量的结构模型，以及对这些结构模型进行排

[8]Li Z,Wang Y,Wang F.A study on fast calling variants from next-generation sequencing data using decision tree[J].BMC bioinformatics,2018,19(1):145.
[9]Chen X,Zhu C C,Yin J.Ensemble of decision tree reveals potential miRNA-disease associations[J].PLoS computational biology,2019,15(7):e1007209.
[10]Zhao B,Xue B.Improving prediction accuracy using decision-tree-based meta-strategy and multi-threshold sequential-voting exemplified by miRNA target prediction[J].Genomics,2017,109(3-4):227-232.
[11]Sikandar A,Anwar W,Bajwa U I,et al.Decision tree based approaches for detecting protein complex in protein protein interaction network(PPI)via link and sequence analysis[J].IEEE Access,2018,6:22108-22120.
[12]Kaur K A,Bhutani L.A review on classification using decision tree[J].International Journal of Computing and Technology,2015,2(02).
[13]Wong K C.Computational biology and bioinformatics:Gene regulation[M].CRC Press,2016.

列，从而选择最佳的模型。基于支持向量机的原理也开发了相应的解决上述预测蛋白质结构问题的方法 [14-16]。通过荧光激活细胞分选进行高通量筛选，是蛋白质工程和定向进化中的常见步骤，如果假阳性率或假阴性率高，则需要进行多轮富集。当前的荧光激活分选软件要求用户依靠经验来定义分类门，并仅限于二维的层面，当进行多轮富集时，软件也将无法预测富集所需的工作量。Yu J S 等人 [17] 开发了基于支持向量机算法的细胞分选方法，可以使用正控制和负控制群的机器学习来识别最佳排序门，也可以利用两个以上的维度来增强区分群体的能力。高通量技术的发展产生了大量的基因组和表观基因组数据，支持向量机的分类功能也扩大了其在癌症基因组学中的用途，在癌症基因组分类或分型中，可以用来发现新的生物标志物、新的药物靶点以及深层次的理解癌症诱导的基因等 [18]。尽管支持向量机在分类方面应用广泛，但是仍然存在不足，如要找到最佳模型，需要测试内核和模型参数的各种组合，在数据量大的前提下，需要花费很长的测试时间 [19]。

三、无监督学习

无监督学习（Unsupervised Learning）算法很少从新数据中学习特征，在应用于新的数据时，将采用以前学习的功能来识别数据的类别，主要用于聚类和特征约简。

聚类

聚类（Cluster）是一种无监督的学习技术 [20]，使用时会自动创建分组，将具有相似特征的数据放在同一个类群中。K-means 是无监督聚类中常用的一种聚类算法，其原理是先随机选取 K 个对象作为初始的聚类中心，然后计算其他对象与初始聚类中心点的距离，根据距离将每个对象分配给距离它最近的聚类中心点。每次分配一个样本，聚类中心点就会重新计算一次，不断重复这个过程直到满足某个终止条件。终止条件可以是没有（或最小数目）对象被重新分配给不同的聚类，没有（或最小数目）聚类中心再发生变化，误差平方和局部最小。

微阵列数据在多种实验条件监测下基因的表达谱中起着至关重要的作用，在生物信息学的研究中，微阵列数据分析的主要目标是共表达和连续模式的识别。由于 K-means

[14]Manavalan B,Lee J.SVMQA:Support-vector-machine-based protein single-model quality assessment[J].Bioinformatics,2017,33(16):2496-2503.
[15]Manavalan B,Shin T H,Lee G.PVP-SVM:sequence-based prediction of phage virion proteins using a support vector machine[J].Frontiers in microbiology,2018,9:476.
[16]Cogill S,Wang L.Support vector machine model of developmental brain gene expression data for prioritization of Autism risk gene candidates[J]. Bioinformatics,2016,32(23):3611-3618.
[17]Yu J S,Pertusi D A,Adeniran A V,et al.CellSort:a support vector machine tool for optimizing fluorescence-activated cell sorting and reducing experimental effort[J].Bioinformatics,2017,33(6):909-916.
[18] Huang S,Cai N,Pacheco P P,et al.Applications of support vector machine(SVM)learning in cancer genomics[J].Cancer Genomics-Proteomics,2018,15(1):41-51.
[19]Huang S,Cai N,Pacheco P P,et al.Applications of support vector machine(SVM)learning in cancer genomics[J].Cancer Genomics-Proteomics,2018,15(1):41-51.
[20] Kassambara A.Practical guide to cluster analysis in R:Unsupervised machine learning[M].STHDA,2017.

聚类算法可以通过比较基因表达谱或者样本表达谱来有效的分析微阵列数据 [21]，该算法在知识发现的领域应用的越来越广泛 [22]。但是 K-means 算法得到的聚类效果严重依赖于初始聚类中心的选择，如果没有选择好初始聚类中心，就会陷入局部最优解而不是全局最优解。目前也发现一些方法来解决上述问题，如引入 k-means++ 来做相应的缺陷优化 [23]，结合其他算法如结合遗传算法，来优化平方和误差等方面 [24]。

单细胞转录组测序的一个主要目的就是对细胞进行细致的分类，确定细胞类型，解决异质性的问题。因此，在单细胞转录组测序技术中应用的聚类方法也适用于不同的具有异质性生物学特点的类型和对象，其中所应用的聚类算法将相似的单细胞表达谱归为一簇，每个簇代表不同的细胞类型。如针对 10× 平台中单细胞转录组测序中等尺寸数据的聚类方法 [25]，针对单细胞转录组高维数据的聚类方法 [26]，针对超大单细胞转录组数据量的聚类方法 [27,28] 等。

降维

降维（Dimension Reduction）是一种对具有高维度特征数据的预处理方法，即减少大数据集的维数，保留高维度的数据最重要的一些特征，去除噪声和不重要的特征，从而提升数据的处理速度，在把信息丢失降到最低的同时，使结果更加容易理解。

主成分分析方法（Principal Component Analysis, PCA）是无监督学习特征约简中使用最广泛的的降维算法，PCA 的主要思想是将 n 维特征映射到 k 维上， k 维是全新的正交特征也被称为主成分。其中，第一个新坐标轴选择是原始数据中方差最大的方向，第二个新坐标轴选取是与第一个坐标轴正交的平面中使得方差最大的，第三个轴是与第 1、2 个轴正交的平面中方差最大的。依次类推，一共可以得到 n 个这样的坐标轴，而这 n 个坐标轴中，大部分方差都包含在前面 k 个坐标轴中，后面的坐标轴所含的方差几乎为 0。于是就可以只保留前面 k 个含有绝大部分方差的坐标轴而忽略剩下的坐标轴。事实上，这相当于只保留包含绝大部分方差的维度特征，而忽略包含方差几乎为 0 的特征维度，从而实现对数据特征的降维处理。

生物信息学数据的一个特点就是数据量非常大，在预处理的部分对数据进行降维处理，大大减少数据计算量的同时，也方便了数据的储存。在聚类步骤之前对数据进

[21]Trivedi N,Kanungo S.Performance enhancement of K-means clustering algorithm for gene expression data using entropy-based centroid selection[C]//2017 International Conference on Computing,Communication and Automation(ICCCA).IEEE,2017:143-148.
[22]Wiwie C,Baumbach J,Röttger R.Comparing the performance of biomedical clustering methods[J].Nature methods,2015,12(11):1033.
[23]Arthur D,Vassilvitskii S.k-means++:The advantages of careful seeding[R].Stanford,2006.
[24]Kapil S,Chawla M,Ansari M D.On K-means data clustering algorithm with genetic algorithm[C]//2016 Fourth International Conference on Parallel, Distributed and Grid Computing(PDGC).IEEE,2016:202-206.
[25]Freytag S,Tian L,Lönnstedt I,et al.Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data[J]. F1000Research,2018,7.
[26]Weber L M,Robinson M D.Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data[J].Cytometry Part A, 2016,89(12):1084-1096.
[27]Sinha D,Kumar A,Kumar H,et al.dropClust:efficient clustering of ultra-large scRNA-seq data[J].Nucleic acids research,2018,46(6):e36-e36.
[28]Abdel R A,Seoud A A,Mahmoud M A,et al.BIG-BIO:-big data hadoop-based analytic cluster framework for bioinformatics[C]//2017 International Conference on Informatics,Health&Technology(ICIHT).IEEE,2017:1-9.

行降维处理，也可以使得聚类结果更加快速、准确 [29]。PCA 也是种群遗传学的基本工具，结合输入样本的基因型和预定义的参考人群，使得 PCA 在低测序深度下也具有很高的准确性，还能矫正小参考群体引入的偏差 [30]。在蛋白质应用方面，PCA 常用于揭示蛋白质中最重要的运动，很多流行的分子动力学软件包也利用 PCA 来分析蛋白质轨迹 [31]。代谢组学中使用 PCA 先主观选择一些代谢物，再对 PCA 的因子负载进行假设检验来选择代谢物，以及结合代谢物富集分析等方法对代谢物进行生物学推断 [32]。尽管 PCA 的优点很多，但是它的输出局限在于要求数据的基本结构必须是线性的，而且高度相关的数据模式可能无法进行解析 [33]。

四、半监督学习

生物信息学领域中的数据大多数都不含标签，因为添加标签需要耗费大量时间和昂贵的人工输入。半监督学习（Semi-supervised Learning）使用已经标记的数据和未标记的数据来构造分类器。半监督学习的目的是使用未标记的实例，将未标记数据中的信息与已标记的显示分类信息相结合以提高分类性能，从而改善学习过程中的问题。因此，半监督学习的主要问题是如何从未标记的数据中挖掘需要的信息。

自我训练

自我训练（Self-training）是用于半监督学习的一种迭代方法，它利用现有训练数据得到的模型，先进行预测将标签分配给无标签的数据，然后选择一组新标记的置信度高的数据，并将其添加到训练集中以进行下一次迭代，直到数据集不发生变化为止，不发生变化包括所有的数据都被标注了标签，以及该模型找不到置信度高的预测结果两种情形。自我训练算法的性能在很大程度上取决于训练过程每次迭代时所选择的新标记的数据。

基因预测是基因组注释过程中最重要的步骤之一，用于基因验证的生物学实验方法昂贵、耗时且劳动强度很大。因此需要开发准确而且快速的方法来分析基因组序列，尤其是鉴定基因并确定其功能。Chan K L 等人 [34] 使用自我训练的算法和转录组测序数据开发了一种通用的基因预测流程，该流程可以应用于新的或者部分测序的植物基

[29]Allab K,Labiod L,Nadif M.Simultaneous semi-NMF and PCA for clustering[C]//2015 IEEE International Conference on Data Mining.IEEE,2015:679-684.

[30] Jørsboe E,Hanghøj K,Albrechtsen A.fastNGSadmix:admixture proportions and principal component analysis of a single NGS sample[J].Bioinformatics,2017,33(19):3148-3150.

[31]David C C,Jacobs D J.Principal component analysis:a method for determining the essential dynamics of proteins[M]//Protein dynamics.Humana Press,Totowa,NJ,2014:193-226.

[32]Yamamoto H,Fujimori T,Sato H,et al.Statistical hypothesis testing of factor loading in principal component analysis and its application to metabolite set enrichment analysis[J].BMC bioinformatics,2014,15(1):51.

[33]Lever J,Krzywinski M,Altman N.Points of significance:Principal component analysis[J].2017.

[34]Chan K L,Rosli R,Tatarinova T V,et al.Seqping:gene prediction pipeline for plant genomes using self-training gene models and transcriptomic data[J].BMC bioinformatics,2017,18(1):1-7.

因组。蛋白质修饰在调节各种生物过程中起着关键作用，如 Pupylation 在细菌中作为一种有利于细胞行使功能的翻译后修饰类型，通常调节原核细胞中的蛋白质功能。在 pupylation 过程中，原核生物类泛素化标记作用在功能上与泛素化相似，可以标记靶蛋白来促进蛋白酶体的降解。因此，准确定位修饰位点对于理解其潜在机制非常重要，识别修饰位点的实验方法耗时耗力且花费昂贵，所以基于蛋白质序列信息来预测相应修饰位点的计算方法得到发展 [35]。

基于图的半监督算法

基于图的半监督学习算法（Graph-based Semi-supervised Learning）用图形来描绘样本空间，用近邻点的位置来控制标记信息的传播。标签传播算法（Label Propagation Algorithm）是一种基于图的半监督学习算法，通过构造图结构（数据点为顶点，点之间的相似性为边）来寻找训练数据中有标签数据和无标签数据的关系。

寻找已知 miRNA 与疾病之间关联是有价值的，Chen X 等人 [36] 基于功能相似的 miRNA 可能与相似的疾病之间有联系反之亦然的前提，提出了 miRNA 与疾病关联预测的异质标签传播算法，并在食道肿瘤、乳腺肿瘤和淋巴瘤数据中得到了验证。确定致病基因的优先级就是在确定给定表型的潜在致病基因，该基因可用于揭示人类疾病的遗传基础和促进相应的药物开发。从基因组测量中对疾病相关的生物标志物进行排名和鉴定 [37]，对于理解常见疾病的遗传基础也具有重要的发展前景。Zhang Y 等人 [38] 利用此类数据集中存在的假阳性蛋白相互作用，结合标签传播算法，对候选疾病基因进行了优先排序并取得了较好的效果。

半监督支持向量机

标准的支持向量机是基于监督学习的，虽然可以有效地解决各种实际的问题，但是需要手工对大量的样本进行标记，以获得足够的训练样本，效率低而代价高。因此，根据实际需求开发了半监督的支持向量机。半监督支持向量机是基于聚类假设，通过探索未标记的数据来规范以及调整决策的边界。为了利用未标记的数据，半监督支持向量机在原来支持向量机的基础上，对未标记的数据点增加了两个限制。一个是假设未标记的点是属于类别 1 并且计算其错分率；一个是假设同样的点是属于类别 2 的并且计算其错分率，而目标函数是计算二者错分率中较小的一个。

生物学数据包含基因组、蛋白质组学、代谢、微阵列基因表达等信息，这些信息

[35]Ju Z,Gu H.Predicting pupylation sites in prokaryotic proteins using semi-supervised self-training support vector machine algorithm[J].Analytical biochemistry,2016,507:1-6.

[36]Chen X,Zhang D H,You Z H.A heterogeneous label propagation approach to explore the potential associations between miRNA and disease[J].Journal of translational medicine,2018,16(1):348.

[37]Stokes M E,Barmada M M,Kamboh M I,et al.The application of network label propagation to rank biomarkers in genome-wide Alzheimer's data[J].BMC genomics,2014,15(1):282.

[38]Zhang Y,Liu J,Liu X,et al.Prioritizing disease genes with an improved dual label propagation framework[J].BMC bioinformatics,2018,19(1):47.

可以分为序列和结构两个大类。由于蛋白质所含的氨基酸序列从几百到几千不等，对蛋白的氨基酸序列进行分类十分困难。Chaturvedi B 等人 [39] 提出了一种新的半监督支持向量机分类器，该分类器结合了标签数据集和非标签数据集，结果显示，该方法对于未标记数据集或只含有少量标记的数据集分类效果更好。在疾病分类上，以往的计算机辅助诊断系统需要用特定方式标记的大量采集的数据作为特征，但是要采集带有标签的患者记录并不容易。引入半监督学习可以在训练阶段更改每次迭代中标记数据的比例，从而更加适应分类器的参数，如在乳腺癌的诊断中，通过半监督学习方法可以更好地解释乳房 X 射线图像 [40]。绘制基因调控网络的拓扑图是系统生物学中的核心问题，控制基因表达的调控结构也是控制着随后的细胞行为，如发育、分化、体内稳态和对刺激的反应等，而这些网络的失调与肿瘤发生、发展有关。因此，用精确的计算方法从基因表达数据中推断基因调控网络十分必要 [41,42]。

五、神经网络

神经网络（Neural network），机器学习的一个重要组成部分，是由多个处理层组成的计算模型，可以用于学习具有抽象特征的数据。神经网络对于深度学习的构建发挥了重要的作用，深度学习通过使用反向传播算法，可以指示机器应该如何更改其内部参数来发现大数据集中的复杂结构，这些内部参数可以根据上一层的指示来计算每一层的指示 [43]。由于建立了重要的算法细节基础，深度学习成功应用于各大领域，这些基础大致可以分为两部分，即深度学习的架构的构建以及训练。深度学习的架构基本上为多层非线性的人工神经网络，根据输入数据的特征和研究目标提出了一些深度学习的架构。Min S 等人 [44] 将深度学习的架构分为了深度神经网络、卷积神经网络、递归神经网络和新兴架构（深度时空神经网络 DST-NNs、多维递归神经网络 MD-RNNs 和卷积自动编码器 CAEs）四类。

深度神经网络和递归神经网络可以应用于预测蛋白质的结构 [45-48]、检测远距同

[39]Chaturvedi B,Patil N.A novel semi-supervised approach for protein sequence classification[C]//2015 IEEE International Advance Computing Conference (IACC).IEEE,2015:1158-1162.

[40]Zemmal N,Azizi N,Dey N,et al.Adaptive semi supervised support vector machine semi supervised learning with features cooperation for breast cancer classification[J].Journal of Medical Imaging and Health Informatics,2016,6(1):53-62.

[41]Ang J C,Mirzal A,Haron H,et al.Supervised,unsupervised,and semi-supervised feature selection:a review on gene selection[J].IEEE/ACM transactions on computational biology and bioinformatics,2015,13(5):971-989.

[42]Maetschke S R,Madhamshettiwar P B,Davis M J,et al.Supervised,semi-supervised and unsupervised inference of gene regulatory networks[J].Briefings in bioinformatics,2014,15(2):195-211.

[43]LeCun,Y.,Bengio,Y.&Hinton,G.Deep learning.Nature 521,436-444,2015.

[44]Min S,Lee B,Yoon S.Deep learning in bioinformatics[J].Briefings in bioinformatics,2017,18(5):851-869.

[45]Yang Y,Heffernan R,Paliwal K,et al.Spider2:A package to predict secondary structure,accessible surface area,and main-chain torsional angles by deep neural networks[M]//Prediction of protein secondary structure.Humana Press,New York,NY,2017:55-63.

[46]Spencer M,Eickholt J,Cheng J.A deep learning network approach to ab initio protein secondary structure prediction[J].IEEE/ACM transactions on computational biology and bioinformatics,2014,12(1):103-112.

[47]Guo Y,Wang B,Li W,et al.Protein secondary structure prediction improved by recurrent neural networks integrated with two-dimensional convolutional neural networks[J].Journal of bioinformatics and computational biology,2018,16(05):1850021.

[48]Pan X,Rijnbeek P,Yan J,et al.Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks[J].BMC genomics,2018,19(1):511.

源的蛋白质结构 [49] 和评估蛋白质模型质量 [50] 等方面。DNA 和 RNA 结合蛋白在基因调控（包括转录和选择性剪接）中起着核心作用，明确DNA和RNA结合蛋白的序列，对于开发生物系统中调控过程的模型以及鉴定疾病的病因变异体至关重要 [51]。为此，基于神经网络的算法已经应用于蛋白质结合的序列特异性建模 [52-54]，且性能优于现有的传统学习方法。部分疾病，如间质性肺病，在计算机断层扫描出有几种异常的成像模式，这些模式的准确分类在疾病的范围和性质的准确临床决策中起着重要作用。Gao M 等人 [55] 提出的基于卷积神经网络的算法，将整个图像作为整体进行输入，虽然前提设置比以往方法困难，但显示了使用整体图像来预测疾病类型的能力。在各种尺度上分析大脑结构中的基因表达，对于理解基因如何调节大脑结构的发育十分关键。Zeng T 等人 [56] 在大量自然图像上训练深度卷积网络，从发育中小鼠大脑的原位杂交图像中提取特征，应用于基因表达模式注释并表现出了优异的性能。

尽管神经网络在生物信息学领域应用的非常广泛，但是大多数算法都是假设数据量足够大并且具有平衡性，而生物信息学中的很多数据并不满足这个前提条件的要求。生物数据采集过程复杂和造价昂贵的特点，限制了生物信息学数据集的规模，而且生物数据很难做到平衡，如与疾病相关的病例数据，疾病类型的数据肯定是少于正常类型数据，又因为伦理以及隐私等问题，很多病例数据并不能公开，从而导致可用的病例数据进一步不足。因此，衡量在数据量不足以及不平衡的情况下，算法的性能就显得尤为重要。目前已经有一些方法用于解决上述的问题，如对数据进行预处理、成本敏感型学习和进行算法修改等 [57,58]。

六、总结与展望

随着基因组研究的发展，极大地推动了生物信息学领域的发展，随之而来的是对生物学数据的处理问题。生物学数据的特点是数据量大而且包含一定的生物学意义，

[49]Nguyen S P,Shang Y,Xu D.DL-PRO:A novel deep learning method for protein model quality assessment[C]//2014 International Joint Conference on Neural Networks(IJCNN).IEEE,2014:2071-2078.
[50]Zheng W,Wuyun Q,Li Y,et al.Detecting distant-homology protein structures by aligning deep neural-network based contact maps[J].PLoS Computational Biology,2019,15(10).
[51]Alipanahi B,Delong A,Weirauch M T,et al.Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning[J].Nature biotechnology,2015,33(8):831-838.
[52]Alipanahi B,Delong A,Weirauch M T,et al.Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning[J].Nature biotechnology,2015,33(8):831-838.
[53]Zhou J,Troyanskaya O G.Predicting effects of noncoding variants with deep learning-based sequence model[J].Nature methods,2015,12(10):931-934.
[54]Zeng H,Edwards M D,Liu G,et al.Convolutional neural network architectures for predicting DNA-protein binding[J].Bioinformatics,2016,32(12):i121-i127.
[55]Gao M,Bagci U,Lu L,et al.Holistic classification of CT attenuation patterns for interstitial lung diseases via deep convolutional neural networks[J].Computer Methods in Biomechanics and Biomedical Engineering: Imaging&Visualization,2018,6(1):1-6.
[56]Zeng T,Li R,Mukkamala R,et al.Deep convolutional neural networks for annotating gene expression patterns in the mouse brain[J].BMC bioinformatics, 2015,16(1):147.
[57]He H,Garcia E A.Learning from imbalanced data[J].IEEE Transactions on knowledge and data engineering,2009,21(9):1263-1284.
[58]López V,Fernández A,García S,et al.An insight into classification with imbalanced data:Empirical results and current trends on using data intrinsic characteristics[J].Information sciences,2013,250:113-141.

结合生物学数据的特点，本文讨论了机器学习算法中监督式学习、无监督学习、半监督学习以及神经网络算法在生物信息学领域的应用。监督式学习算法是需要外部帮助的算法，在一定程度上也受到了这个条件的限制；无监督学习算法通过以前学习的经验来应对新的数据，这对新的数据要求比较限制，并且需要衡量指标来判断对新数据学习的情况；半监督学习应用的数据一部分有标记一部分无标记，适合病例数据类型集，但无标记部分的数据不容易质控；神经网络适合处理大数据，但是受限于神经网络应用的假设前提，有些不符合前提的生物学大数据在应用神经网络时可能不会达到预期效果。

为了使各种算法更好地应用于生物信息学领域，一方面要对各种算法的原理以及处理过程有详细的理解，让生物学背景的研究人员可以针对自己的问题，来寻找相应的解决方法，一方面要对所处理的生物学数据代表的生物学意义有一定的认识，让计算机领域的研究人员可以开发出更合适的处理方法。除此之外，未来人工智能技术在生物信息学领域的发展，除了对相应生物学问题的数据进行合理的分析处理，还要能够与实验过程的步骤进行智能结合。二十一世纪是生命科学的世纪，随着人工智能技术和生物学技术的快速发展，二者深度融和大放异彩指日可待！

查看内容精选