**电 子 科 技 大 学**

UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

**学士学位论文**

**BACHELOR THESIS**



**论文题目** **基于间隔序列寡核苷酸的必需基因预测算法研究**

**学 院** **生命科学与技术学院**

**专 业 生物技术（生物-信息复合培养实验班）**

**学 号 2016090202016**

**作者姓名 高鑫辰**

**指导教师 郭锋彪 教授**

摘 要

必需基因是生物体或细胞在一定条件下生存所必需的那些基因,它们所编码的功能被认为是生命的基础。必需基因不仅具有重要的生物学功能,而且在实际应用中特别受关注，因此必需基因的识别和预测任务十分重要。本文从大肠杆菌的基因序列出发，受到随机过程理论中的Markov链模型的启发提取并筛选出特征后，将支持向量机算法应用到必需基因预测问题上构建分类模型，经五折交叉验证后获得了0.85的平均AUC。

**关键词：**必需基因，机器学习，生物信息学。

ABSTRACT

Essential genes are those genes necessary for organisms or cells to survive under certain conditions. The functions they determine are considered to be the basis of life. Essential genes not only have important biological functions, but also receive attention in practical applications. So, the task of identifying and predicting essential genes is very important. In this paper, I collected the information of *E. coli*’s gene sequence, extracting and selecting the feature matrix motivated by the Markov chain model in the random process theory. Then I apply support vector machine (SVM) algorithm to the task of essential genes classification, and the classification model obtains an average AUC of 0.85 after 5 cross-validation.

**Keywords:** Bioinformatics, Essential gene, Machine learning

目 录

[摘 要 I](#_Toc42247946)

[ABSTRACT II](#_Toc42247947)

[目 录 III](#_Toc42247948)

[第一章 绪论 1](#_Toc42247949)

[1.1 研究背景 1](#_Toc42247950)

[1.2必需基因介绍 1](#_Toc42247951)

[1.3 必需基因研究现状 2](#_Toc42247952)

[1.4 论文的内容与研究安排 3](#_Toc42247953)

[第二章 数据与方法 5](#_Toc42247954)

[2.1 大肠杆菌必需基因数据集的获取和应用 5](#_Toc42247955)

[2.1.1 数据集的获取 5](#_Toc42247956)

[2.1.2 大肠杆菌数据集的整理 5](#_Toc42247957)

[2.2 基因序列特征的提取和筛选 6](#_Toc42247958)

[2.2.1 特征提取方法介绍 6](#_Toc42247959)

[2.2.2 特征筛选方法介绍 8](#_Toc42247960)

[2.2.3 具体实验过程 10](#_Toc42247961)

[2.3 分类模型的设计与构建 11](#_Toc42247962)

[2.3.1 机器学习与机器学习算法的评价指标介绍 11](#_Toc42247963)

[2.3.2 分类模型的设计与构建 14](#_Toc42247964)

[2.4 本章总结 16](#_Toc42247965)

[第三章 实验结果分析 17](#_Toc42247966)

[第四章 总结与展望 20](#_Toc42247967)

[4.1 总结 20](#_Toc42247968)

[4.2 未来展望 20](#_Toc42247969)

[致 谢 22](#_Toc42247970)

[参考文献 23](#_Toc42247971)

[外文资料原文 25](#_Toc42247972)

[外文资料译文 42](#_Toc42247973)

第一章 绪论

1.1 研究背景

在生命科学的组学时代，数据以多种形式呈现，代表了生物系统各个层面的信息，包括有关基因组，转录组，表观基因组，蛋白质组等[1]。随着有关人类，动物或微生物的大量信息的日益积累，研究人员开始研究海量数据集，并进一步阐明这些数据集在生物学中的基本含义。但是这其中有一个很大的问题就是生物数据规模非常大，由此产生了以下亟需解决的问题，例如：如何处理信息的复杂性；如何从差异性很大的资源中整合信息；合适采用何种原则或标准。没人会怀疑生物学数据会创造出巨大的价值。面对大数据，分析大型生物数据的工具和技术使我们能够将大量信息转化为对基本生物医学机制的更好理解，并可以进一步应用于转化医学或个性化医学。通常，大数据具有四个重要特征：数据量、处理数据的速度、数据源的可变性和数据质量的准确性。大数据的这四个标志需要以特殊的理论和技术来表征，但是，目前还没有令人满意的解决方案。

现在，由于高通量生物测序技术的飞速发展，越来越多的生物学家参与了大数据研究。比如人类基因组计划那样，利用了来自20个机构的专业知识，基础设施和人员，花了13年的时间，耗资30亿美元，确定了大约30亿个核苷酸的整个基因组结构[1]。现在的我们，可以在很短的时间内用较低的测序价格去对某个物种的基因组进行测序。但是在大数据井喷式增长的当下，我们如何通过理解含有大量噪声的大数据来解释生物学系统的基本机制是目前生命科学研究的新瓶颈。当今的生命科学需要更强大、更具表现力、可计算性、定量、准确和精确的方式来处理大数据。实际上，该领域的最新工作已经带来了显著的优势和机遇，这暗示了生物信息学和生物信息学家在未来生物和生物医学领域研究中的重要作用。

1.2 必需基因介绍

基因在生物学中是指DNA或RNA内编码产物的合成的核苷酸序列，简单来说，基因指导产生肽链和功能RNA，是生物体的基本遗传单位。在DNA层面上，基因主要由编码区和非编码区组成。真核生物的编码区是非连续的，分为外显子和内含子两部分，而在原核生物中，编码区是连续的。非编码区在基因的表达调控中发挥着重要作用，该区域有增强子、启动子等基因调控元件[2]。值得注意的是，在人类基因中非编码区的占比超过百分之九十。一个个体中所有基因的总和称为基因组，一个物种中所有等位基因的总和称为基因库。

根据基因功能的不同可以把基因分为结构基因、调节基因和操纵基因。结构基因是指能为多肽链编码的基因，调节基因是指用来控制其他基因表达的基因，操纵基因是那些本身并不进行转录，但对其临近的结构基因的转录起控制作用的基因[3]。

在基因中，有一类基因如果丧失功能会损害生存能力或导致严重丧失适应能力，这类基因就是必需基因[17]。必需基因是生物体或细胞在一定条件下生存所必需的那些基因，因此他们所编码的功能被认为是生命的基础[4]。最小基因组是由生物的必需基因所构成的，它是最小的可能基因组，该基因组在最有力的条件下足以维持正常的细胞生命形式。确定生物的最小基因集解决了一个概念上重要的问题：维持生命形式所需的基本功能是什么，因此最小基因集的概念在新兴领域合成生物学中起着关键作用[18]。

必需基因不仅具有重要的生物学功能，而且在实际应用中特别受关注。例如，通过生物的必需基因构建最小基因组进行合成生物学的研究，或者将最小基因组作为一个基准，向其中添加其他基因以获取能制造出特定产物的细胞；此外，由于必需基因被破坏所致的致死性，必需基因是药物的诱人靶标，即药物可以通过作用于病变细胞的必需基因来杀死病变的细胞，这对于药物靶标鉴定和包括癌症在内的疾病治疗有着重要意义；除此之外，必需基因的研究对于了解生命起源和相关的进化研究也很有帮助。总之，研究必需基因对于了解生命的本质以及广泛的应用场景都具有非凡的意义。

1.3 必需基因研究现状

由于必需基因对于生物体的至关重要的功能以及其在合成生物学和医疗领域光明的应用前景，这些基因引起了越来越多的关注。近年来，对必需基因进行确定的研究发展迅速，自从1995年有美国团队用引导诱变法鉴别出了6个枯草芽孢杆菌的必需基因[14]，以此为开端，其中许多小组利用了湿实验的方法，在原核生物中，有研究小组使用单基因敲除技术、转座子诱变和RNA技术来识别必需基因；关于真核生物，有研究小组已经通过实验技术为几种物种和细胞系鉴定了许多必需基因，采用的实验方法有RNA干扰、CRISPR-Cas9基因编辑技术和逆转录病毒基因陷阱[21]。其中CRISPR-Cas9因为其精准、抗干扰能力强等优点而备受青睐，值得注意的是，有三个研究团队利用CRISPR-Cas9和基因捕获技术大约坚定了人类癌细胞系中的2000个必需基因。他们的结果显示出高度的一致性，这进一步证实了必需基因集的准确性和鲁棒性。这些研究提供了对肿瘤特异性必需基因的深入分析和筛选肿瘤特异性必需基因的可行方法。这三个小组筛选的必需基因为维持单个人类肿瘤细胞类型的基本细胞活性提供了明确的定义，实际上，这些必需基因可被视为治疗癌症的靶标。

使用湿实验确定必需基因的方法具有成本高、耗费巨大且对实验条件要求很高等缺点。并且近些年来，随着高性能计算机的出现、计算机运算能力的不断提升、研究人员对生物体海量数据的获取以及生物信息学的迅猛发展，使得利用计算生物学方法研究必需基因成为了可替代的选择。这种方法大大降低湿实验的巨大耗费和时间的同时，提供了较高的必需基因预测准确率。在计算生物学方法构建预测模型时，主要使用了五种主要特征，这五种特征包括进化保守性、域信息、网络拓扑、序列组成和表达水平[5]。实际上，使用计算生物学方法来预测必需基因已经使用了数十年。第一种计算方法是1996年Mushegian等人基于比较基因组学开发的方法[19]。随后，Chen等人在2005年首先将机器学习方法引入预测蛋白质可分配性的领域，他们基于进化速率、重复率、蛋白质-蛋白质相互作用网络和基因表达相关网络中蛋白质的连通性分析了酿酒酵母的蛋白质可分配性[20]。Deng等使用神经网络和支持向量机，使用固有的基因组特征、上下文相关特征等作为四个分类器的输入数据，以训练模型来预测必需基因，这是首次研究跨物种必需基因预测模型的准确性[15]。最近，Guo等人结合使用了序列成分（SC）功能与SVM一起预测人类必需的基因，并提供了在线服务。上面提到的所有这些工作都依靠计算模型并做出了很好的预测[16]。

1.4 论文的内容与研究安排

本次课题研究主要使用大肠杆菌作为研究对象，主要任务是基于已经被确定的大肠杆菌必需基因和非必需基因的相关信息，理解并采用计算机编写出间隔序列寡核苷酸特征组成的代码，基于特征描述代码构建必需基因预测模型，获得85%（严格交叉验证）以上的AUC。以下是论文的主要研究内容：

1、在第一章主要介绍了本课题开展的研究背景、基因和必需基因的相关概念和知识、必需基因的研究现状（包括湿实验的确定方法和计算生物学的确定方法）以及论文内容的总体安排；

2、第二章主要介绍本课题采取的数据与方法，包括以下几个方面：（1）如何获取和整理使用的大肠杆菌必需基因和非必需基因数据集；（2）主要介绍基于基因的序列信息，同时考虑邻接和间隔关联的序列特征描述提取出序列特征信息，并对获得的高维特征矩阵进行特征的筛选以达到降维的目的；（3）主要介绍不同的机器学习算法的原理和评价指标，其中着重介绍本课题最终采用的支持向量机算法（SVM）和基于SVM的多种分类器模型的设计构建；

3、第三章主要介绍使用构建的分类器模型，在大肠杆菌的数据集上采用5-折交叉验证评估并对比不同核函数的效果；

4、第四章主要总结论文的主要内容和创新点，分析毕业设计过程中出现的问题与不足，并对未来的研究进行展望。

第二章 数据与方法

2.1 大肠杆菌必需基因数据集的获取和应用

2.1.1 数据集的获取

大肠杆菌染色体(PEC)数据库(http://www.shigen.nig.ac.jp/ecoli/pec/)专门提供了大肠杆菌的染色体数据[6]。旨在使大肠杆菌研究人员能够有效地访问功能基因组学研究中的信息。该数据库包含两种主要类型的数据：基因的必需性和大量的大肠杆菌遗传研究资源。必需性数据基于已发表的单基因必需性研究的数据集以及大缺失突变体的细胞生长研究。从PEC数据库下载得到的信息主要包括基因名、基因的起始位点和终止位点、基因的长度和基因必需性数据等信息。除这些数据外，PEC数据库还提供了具有全面接口的其他细菌基因组的同源基因和蛋白质结构信息的摘要。因此，PEC数据库是当代大肠杆菌研究人员的便捷实用平台。从PEC数据库下载得到的大肠杆菌基因信息的文件PECData.dat经整理后的格式如图2-1所示：

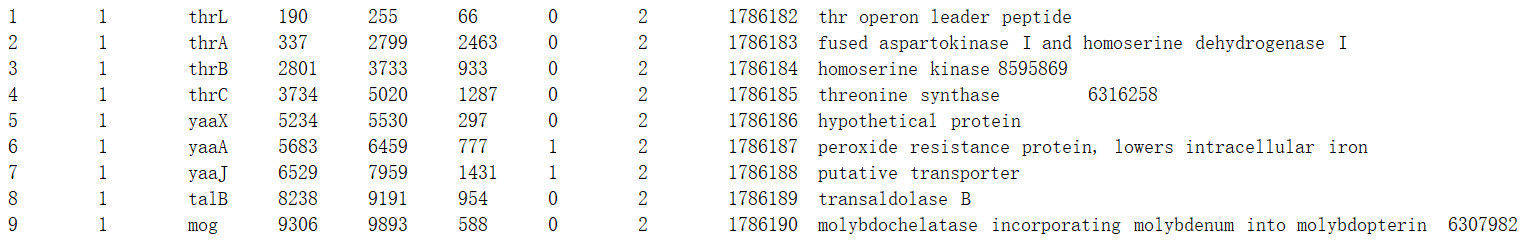


图2-1 大肠杆菌基因信息

其中每列的含义从左至右依次是Orf ID、Feature Type（其中1代表基因）、Orf、起始位点、终止位点、长度、方向（0代表正向，1代表反向）、类别信息（1：必需基因，2：非必需基因，3：未知）、PID、产物和PMID。

2.1.2 大肠杆菌数据集的整理

经过统计，在从PEC数据库中下载下来的大肠杆菌基因信息中，总共有4497条基因，其中4190条是非必需基因，302条为必需基因，5条所属类别未知。然后我们将该数据集中的基因信息与美国生物信息中心（NCBI）所提供的基因序列信息进行比对，对数据集中的数据进行筛选和清洗，即将数据集中的基因ID和从NCBI基因组信息库中获取的蛋白质序列的ID号进行比对，在两个数据库中均存在的基因最终被留下来。经过整理，最终筛选出4144条基因，其中287条为必需基因，3857条为非必需基因。同时，必需基因和非必需基因分别用1和-1作为正负样本的标记。最终，我们整理出两个数据文件，一个是大肠杆菌的基因序列信息文件PECsequence.txt，文件的内容是由4414条基因的名称和每个基因对应的用a、t、c、g四种核苷酸字符表示的核苷酸组成序列；另一个是大肠杆菌基因类别信息文件EcoliPEC\_mark.txt，文件的内容是每条基因的名称和其所属的正负样本类别，其中1代表该基因是必需基因，-1代表该基因是非必需基因。整理出的基因序列信息如图2-2所示，整理出的基因类别信息如图2-3所示。

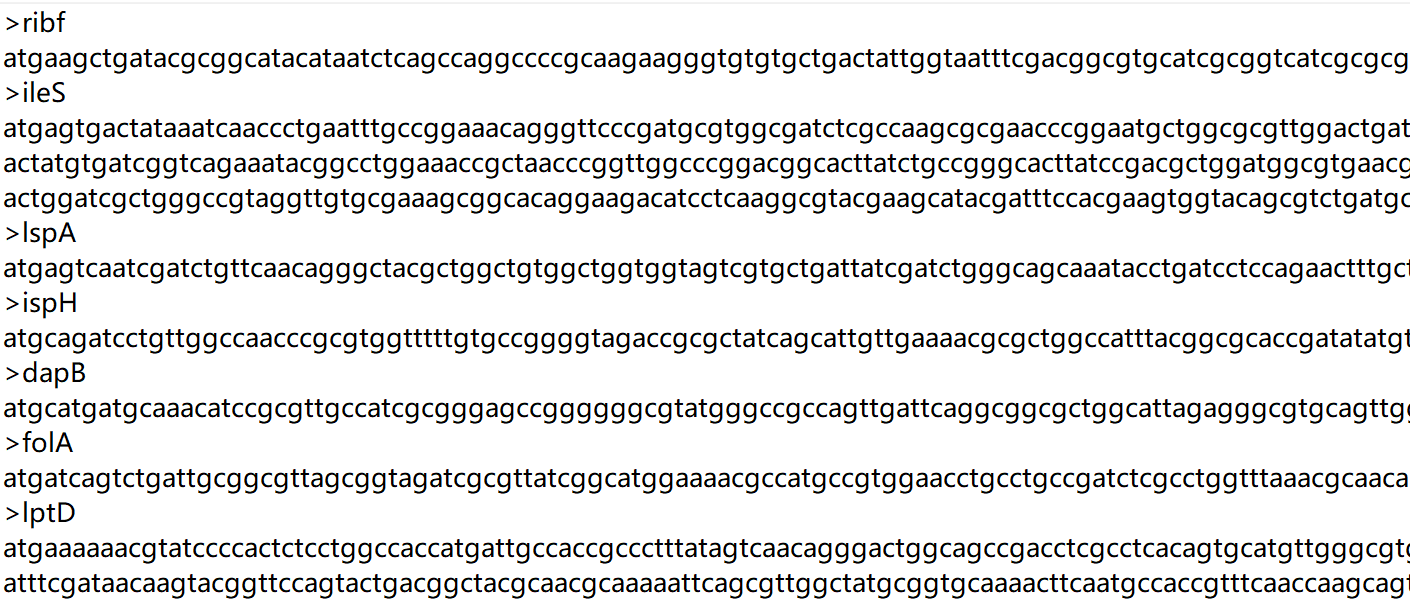


图2-2 大肠杆菌基因序列信息

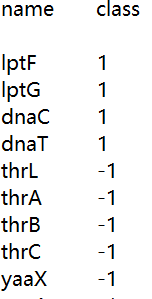


图2-3 大肠杆菌基因类别信息

2.2 基因序列特征的提取和筛选

2.2.1 特征提取方法介绍

基因是具有遗传效应的DNA片段，能够作为转录和翻译的模板，指导合成RNA和多肽链，在生物体的生长、发育和繁殖中起到非常重要的作用。DNA分子主要由脱氧核糖、含氮碱基和磷酸构成，不同的脱氧核糖分子之间的区别主要在于含氮碱基的不同。在DNA中，含氮碱基主要有四类：腺嘌呤（A）、鸟嘌呤（G）、胞嘧啶（C）、胸腺嘧啶（G）。DNA以双螺旋结构排列，每条链分为5’端和3’端，规定5’端到3’端为正向，且两条链之间因为氢键的存在缠绕在一起，四种核苷酸以碱基互补配对的方式存在，其中A与T配对，C与G配对。所以DNA序列可以表征为一个方向上的由A、T、C、G（或者a、t、c、g）四种字符组成的字符串。就像在计算机科学中的0、1二进制编码一样，A、T、C、G四种字符可以组成无穷多的字符串序列，其中蕴藏着无尽的生物遗传信息。在氨基酸的翻译过程中，每三个核苷酸决定了一个氨基酸，所以可以从基因DNA序列的起始点开始，每三个核苷酸编为一组，将在每组中位于第一位的核苷酸称为第一相位，位于中间位的核苷酸称为第二相位，位于第三位的核苷酸称为第三相位。

随机过程的Markov模型假设下一个字符的状态由当前几个连续的字符的状态相关。我们假设在DNA序列中间隔λ个字符的核苷酸之间也存在相关性，这种相关不是马尔科夫链的实际相关，而是在一级序列中虚拟的，只有三级结构中才真实存在的相关。

对于单核苷酸，我们提取DNA序列（由a、t、c、g四种字符表征）中三个密码子相位上，每个相位中a、t、c、g四种脱氧核糖核苷酸字符分别出现的频率作为特征，特征数目为3\*4^1=12个，计算方法如公式2-1所示：



其中的a、c、g、t对应于四种核苷酸的频率或个数，1、2、3表示三个相位。对于邻接双核苷酸，我们提取三个相位每一相位的16种相邻双核苷酸分别出现的频率作为描述变量。同时，我们准备多组变量用以描述具有间隔距离的两个核苷酸间的虚拟关联。间隔距离用λ表示，λ的最大取值则用L表示。由此我们一共可以得到16×3+16×3×L个变量来描述双核苷酸（邻接的和具有间隔距离虚拟的双核苷酸）的关联组成频率，用公式表述如公式2-2、2-3、2-4、2-5所示：









上式中aa等双字符表示16种双核苷酸组合中其中一种组合的频率，1、2、3还是表示相位。不管是邻接的还是间隔的双核苷酸都可以表示为等式右侧的变量组的形式。λ=0时，表示邻接的双核苷酸。当λ不为0时，实际计算的是间隔一段距离形成的虚拟的双核苷酸的频率。我们对于三核苷酸、四核苷酸、五核苷酸…进行类似处理。类似于马尔科夫模型，对于k核苷酸频率，间隔前面的核苷酸片段取k-1个字符，间隔后面的核苷酸只取1个字符，然后计算这个长度为k的虚拟核苷酸片段分别出现的频率。用通用的公式表示就是公式2-6所示：



公式3-6对应于一个特定多核苷酸组合X与一个特定λ间隔值的一组变量。X代表k-1个核苷酸所能形成的4k-1种组合。相位1、2或3由X的第一个字符所处的实际位置确定。λ则从0 一直取值到最大考虑值L。实际应用时，我们可以只用k=2或3或4等等的其中一种情况，但为了取得更好的预测结果会尽可能考虑联合k=0一直到k=M等各种情况下的变量。

2.2.2 特征筛选方法介绍

提取的特征中包含了大量无用的信息，需要从提取的特征中筛选出有价值的特征，达到减少特征数目和降维的目的，从而加快学习的速度。现介绍几种特征筛选的方法。一般通过特征提取得到的特征矩阵维度都很高，特征选择（Feature Selection）是模式识别应用程序中一种广泛使用的技术[7]。通过从原始特征空间中删除无关，嘈杂和多余的特征，特征选择减轻了过度拟合的问题并提高了模型的性能。还可以减少学习算法的时间和空间成本。更重要的是，我们可以通过分析功能的重要性来获得对数据的更深入的了解[7]。在分类任务中，基于特征选择算法与分类器的交互方式，可以将特征选择算法分为三类：Filter方法、Wrapper方法、Embedded方法[8]。Filter方法通过预定义的统计学指标（例如相关系数和互信息）对每个特征进行评估，这些标准与分类器无关；Wrapper方法将分类器视为黑匣子，旨在找到在训练数据上具有最小交叉验证误差的特征子集，Wrapper的示例包括顺序正向选择、遗传算法和模拟退火；Embedded方法通常包括两种方法，在某些方法中分类器的训练会从本质上选择特征的子集，而另一些方法根据分类器中的系数来估计特征的重要性。

2.2.2.1 递归特征消除

递归特征消除（RFE）是一种常见的Wrapper方法，其主要思想是不断的构建分类模型。根据参数（如线性模型对应的参数coefficients）对特征进行排序，消除最弱的一个或多个特征，然后在剩余的特征中重复这个过程，直到达到指定数量的特征为止。为了找到最佳数量的特征，交叉验证与RFE一起使用以对不同的特征子集进行评分，并选择最佳的特征评分集合[7]。

2.2.2.2 t检验

t检验可以用来判别两组数据之间的统计性差异有多显著，换句话说，t检验能够用来判断不同类别数据集之间的差异性是否是偶然因素导致的。t-score是两组之间的差异与组内的差异之间的比率。t-score越大，组之间的差异就越大；t-score越小，组之间的相似度越高。当运行t检验时，t值越大，结果可重复的可能性就越大。简单来说，一个较大的t-score说明两个样本之间是有显著差异的，一个较小的t-score说明两个样本之间是有相似的。关于t值“足够大”有多大，每个t值都有一个p值。p值是样本数据的结果偶然发生的概率。p值从0％到100％，它们通常写为小数，例如，p值为5％时写成0.05。较低的p值是好的，它们表明数据的差异性是偶然发生的概率很小，甚至可以认为不是偶然发生的。例如，p值为0.01，表示实验结果偶然发生的可能性只有1％。在大多数情况下，接受p值为0.05（5％）表示数据有效。t检验主要有三种方式：独立样本t检验比较两组的均值；配对样本t检验比较同一组在不同时间（例如相隔一年）的平均值；一个样本t检验可将单个组的平均值与已知平均值进行比较。在将t检验用于二分类任务的特征筛选时，可以根据数据集中数据类别的不同将数据集分为两组，然后用t检验计算每个特征在两组不同类别数据中差异的显著性，筛选出差异性较大的特征，从而得到维度较小并且效果更好的特征矩阵。

2.2.3 具体实验过程

python语言是近年来备受关注的高级程序语言，由于很多团队开发的基于python的科学包（例如numpy、pandas、scipy等）的存在，使python成为用于数据科学以及机器学习项目的重要工具。本课题的特征提取和特征筛选工作也都是由python语言完成的。

在第二章中，我已经介绍了整理出的两个数据集文件：PECDNAsequence.txt和EcoliPEC\_mark.txt：其中PECDNAsequence.txt文件中存放的是大肠杆菌每条基因的序列信息，EcoliPEC\_mark.txt文件中存放的是大肠杆菌每条基因的类别信息（1代表必需基因，-1代表非必需基因）。接下来依次介绍在我们获取的数据集上进行特征提取和特征筛选的实验过程。

2.3.3.1 特征提取

使用python语言读取基因的序列信息，根据3.1节所介绍的特征提取方法进行特征的提取。我定义了两个参数用来表征目前提取特征的情况，其中K代表核苷酸数目，λ代表引入的间隔大小。首先从单核苷酸开始，提取出12个频率作为特征，然后增加K的数目，每次增加1个核苷酸；对于每个K，λ的值从λ=1开始一直增加到λ=K；当进行核苷酸数目为K的特征提取获得特征矩阵后，把核苷酸数目从1到K-1的情况下的特征都整合到一个特征矩阵上，获得一个最终的特征矩阵。目前进行到K=7的情况，特征矩阵大小为(4144, 524172)，表示该特征矩阵有4144个样本，524172个特征值。

2.3.3.2 特征筛选

特征筛选我采用的是t检验的方法，首先将上一步提取的特征矩阵数据集划分为训练集和测试集，其中百分之八十的样本作为训练集，百分之二十的样本作为测试集，注意在测试集和训练集中正负样本的比例均要保持与原数据集相等。然后在训练集上进行特征的筛选。筛选的简要过程介绍如下：首先将训练集的正负样本分开，分别放入两个矩阵essential和non\_essential中；然后对每个特征在两个样本集中进行t检验，t检验由python的科学程序包scipy完成，计算出每个特征对应的p值，放入列表中；然后对列表中的p值进行升序排列。由3.2节的介绍可知p值越小，说明该特征的差异性越大，所以具有小p值的特征就是我们需要留下的特征。在这里值得注意的是，在计算完p值后，有两种用于特征筛选的策略，一种策略是按比例筛选特征，比如可以筛选出p值为前百分之一或者前千分之一小的特征。该策略的缺点是对于比例的选择没有一个确定的标准，并且当选定一个比例（例如百分之一）时，当核苷酸数目较小时筛选出的特征数目太少，而核苷酸数目较大时筛选出的特征数目又太多，对实验结果的影响十分不利。另外一种策略是预先设定一个p值（例如0.001），然后将所有p值小于预设的p值的特征全部归于具有显著性差异的特征，这种筛选策略的方法解决了第一种策略的一些缺陷和不足。

2.3 分类模型的设计与构建

2.3.1 机器学习与机器学习算法的评价指标介绍

2.3.1.1 机器学习算法介绍

目前人类社会进入大数据时代，各领域各行业的海量数据井喷式增长。如何从海量的数据中挖掘出有价值的信息造福人类是数据挖掘的主要任务。机器学习算法是人工智能（AI）的一种应用，它使系统能够自动学习并从经验中进行改进，而无需进行明确编程。机器学习专注于计算机程序的开发，该程序可以访问数据并自己学习。学习的过程始于观察或数据，例如示例、直接经验或指导，以便根据我们提供的示例寻找数据模式并在将来做出更好的决策。主要目的是允许计算机在没有人工干预或帮助的情况下自动学习，并相应地调整操作。举个简单的例子，人类可以根据生活经验来对未来进行预测，比如可以根据今天空气的湿度、昆虫的行为等来对明天的天气做出判断，而机器学习算法的目标就是让机器能够像人一样从已知的数据中学习到规律，然后根据学习到的规律对未知的数据进行预测和分析。机器学习可以分析大量数据。虽然它通常可以提供更快，更准确的预测效果，但它可能还需要更多的时间和资源来正确地对其进行训练。将机器学习与AI和认知技术相结合，可以使其在处理大量信息方面更加有效。

机器学习算法可以分为监督学习和无监督学习。在监督学习中，训练集本身已经包括输入和输出数据，例如在分类任务中，训练集本身已经包括了特征和每条样本所属的类别。监督学习的任务就是根据训练集提供的输入输出信息中学习得到从输入到输出的映射模式，然后对没有给出输出只有输入特征的检验集进行预测，得到每个样本的输出；在无监督学习中，训练集本身只包含输入的特征数据而不包含输出的信息，无监督学习的任务是根据训练集的特征，将数据划分为不同集团。举个例子，可以把训练集中每个样本表示为向量，每个特征作为向量的分量，根据向量之间的距离，把所有样本分为不同类别。很显然，无监督学习任务的难度要比监督学习的难度要大。本课题的数据集中提供了每个样本的类别信息，属于监督学习的范畴。并且由于数据集的输出是离散的，并且只有两种：1代表必需基因，-1代表非必需基因，所以本次基于间隔序列寡核苷酸的必需基因预测算法研究的课题是一个典型的二分类问题。现介绍几种常用的分类算法如下。

支持向量机（SVM）是一种监督式机器学习模型，该模型使用分类算法解决二分类问题[9]，在为SVM模型提供带有标签的训练数据集后，该模型便能够对文本进行分类。给定训练数据集，支持向量机在该分类学习的任务就是要找到一个最能分隔标签的超平面（在二维上，该平面只是一条线），这个超平面是决策的界限，把正负样本分隔开。显然这样的超平面有很多，SVM的任务就是找到最好的超平面，最好的超平面是使两个类别标签的边距最大化的一种方法，换句话说，该超平面与每个类别标签的最近数据点的距离最大，称之为最大间隔超平面[9]。在多数情况下，训练样本并不像之前的例子那样是线性可分的，训练集数据往往并不是线性可分的，这就需要针对不同问题选择核函数，常见的核函数有线性核、多项式核、（高斯）径向基函数核（RBF核）等。

朴素贝叶斯分类算法（NB）是以贝叶斯定理作为原理的分类器，该算法也是主要用于监督学习的分类学习任务。贝叶斯公式如公式2-7所示：



在分类问题中，A代表特征，B代表类别，则贝叶斯公式可以转换成如公式2-8所示：



通过给定的训练数据集，利用统计学知识不难计算先验概率P(类别)和样本相对于类别标签的条件概率P(特征|类别)，进而计算P(类别|特征)。朴素贝叶斯算法的原理简单，但是该算法假设数据集的特征之间相互独立，在使用上具有局限性。

决策树算法是一个决策支持工具，在分类学习任务中希望模型能从给定的数据训练集中学习到一个模式对测试集进行分类。决策树把分类任务看作一个决策的过程，通过一个树状的结构从根节点开始对每个特征进行决策，直到在叶节点得到分类结果。简单的来说，决策树对应一个树的数据结构，它包括根节点、内部节点和叶子结点。除了叶节点，每个内部节点代表一个属性上的“测试”（例如西瓜的色泽是绿色还是黄色），每个分支代表测试的结果（例如色泽是绿色），在每个结点都会进行这样的判断，直到叶节点。而每个叶节点代表一个类别标签，它是在计算所有属性后做出的决定（例如瓜是熟的）。从根结点到叶子结点的路径代表了分类的规则。决策树具有自然的“ if…then…else…”构造，使其易于适应程序结构。它们也非常适合分类问题，在这些问题中系统检查属性或特征以确定最终类别。决策树易于理解，是一个在数据探索很有效的非参数方法，但是当决策数过于复杂时，模型会出现过拟合，导致泛化能力较差，并且决策树不适合连续的变量，他在将变量归类到不同类别时会丢失信息。

2.3.1.2 机器学习算法的评价指标介绍

在机器学习领域，尤其是在二分类任务中，评价一个分类的机器学习算法的性能通常需要用到混淆矩阵（Confusion Matrix）。混淆矩阵涉及到以下四个参数：真正例（TP）、假正例（FP）、真负例（TN）、假负例（FN）[10]。其中，真正例（TP）表示实际为正样本，预测也为正样本的样例数；假正例（FP）表示实际为负样本却被预测为正样本的样例数；真负例（TN）表示实际为负样本，预测也为负样本的样例数；假负例（FN）表示实际为正样本却被预测为负样本的样例数[10]，混淆矩阵如图2-4所示：

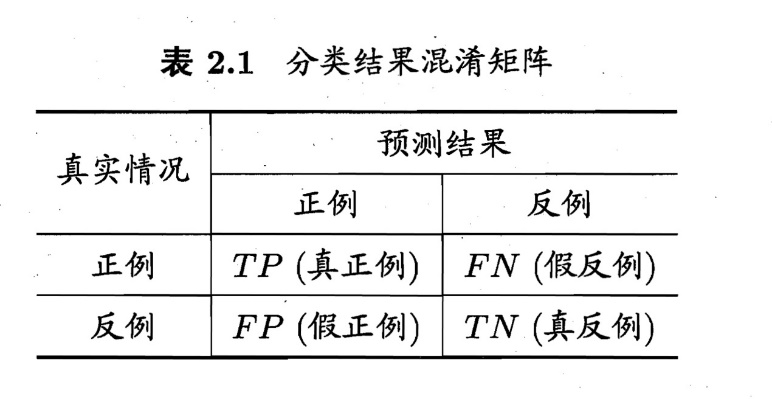


图2-4 混淆矩阵

根据混淆矩阵，可以得到几个常用的分类算法评价指标：准确率（Accuracy）、错误率（Error rate）、精度（Precision）、召回率（recall）以及ROC曲线和AUC等。准确率表示所有样本中预测正确的样本所占的比例，计算公式为。错误率和准确率相反，是所有样本中预测错误的样本所占的比例，错误率和准确率之和为1，计算公式为。精度表示在所有预测为正样本的样本中，实际为正样本的所占的比例，计算公式为。召回率表示在所有真正的正样本中，被预测为正样本的比例，计算公式为。ROC曲线和AUC是两个密不可分的概念，现在详细介绍一下这两个概念如下。

ROC曲线是评价分类的机器学习算法性能的有力工具，它是一个概率曲线，其纵轴是真阳性率TPR，表示所有正样本中被正确预测为正样本的比例，计算公式为。曲线的横轴是假阳性率FPR，表示所有负样本中被错误预测为正样本的比例[11]，计算公式为。ROC曲线就是以TPR为纵轴、以FPR为横轴的曲线。很显然，在ROC曲线的横轴上，较小的值表示较低的假阳性和较高的真阴性，在ROC曲线的纵轴上，值越大表示真阳性值越高而假阴性值越低[11]。AUC表示ROC曲线下的面积，对所有可能的分类阈值的效果进行综合衡量，是对模型性能的估计。换句话说，AUC的值表示一种可能性，它表示模型把一个正样本预测为正样本的概率比模型把一个负样本预测为正样本的概率大的可能性[11]，由此可见AUC的值越大，说明我们构建的分类模型效果越好。

2.3.2 分类模型的设计与构建

2.3.2.1 分类算法的选择

在之前介绍的几种常用分类算法中，朴素贝叶斯分类算法（NB）的结果依赖于先验概率的计算且该算法建立在这样的假设基础上：训练数据集的特征之间相互独立。而在本论文第二章介绍本课题的特征提取方法时，我假设DNA序列中间隔λ个字符的核苷酸之间存在相关性，所以没有选择朴素贝叶斯分类算法。决策树算法不适合连续的变量，而我提取的特征是间隔寡核苷酸序列在每个相位上出现的频率，并且特征数量非常多，所以不适合在本课题中使用决策树算法。SVM算法在生物信息学上使用十分广泛，并且在高维特征空间上的二分类任务上具有较好的区分能力，而必需基因预测的本质是典型的二分类问题。并且在前面的介绍中可以得知，本课题的特征空间维度很高，而在之前的研究中，SVM算法在必需基因预测上的表现很出色，所以本课题采用SVM作为分类模型的算法。

在之前对支持向量机的介绍中，我们简单介绍了在SVM中常用的几种核函数，其中最常用的核函数就是线性核和径向基函数核（RBF核），线性核常用于线性可分的分类学习任务中，需要调节的参数少，具有较快的学习速度，而径向基函数核在更复杂的情况中有出色的表现，但需要调节的参数很多。我查阅了相关文献,有研究表明，如果特征维数很大，样本数量很小，使用线性核作为核函数就可以获得出色的效果。在和同课题组的研究生师兄交流时得知，他在相关的研究中使用了径向基函数核作为SVM的核函数，结果训练速度很慢而且最终并未取得很好的效果。所以我在本课题使用线性核的SVM算法作为分类模型的算法，并且与师兄使用径向基函数核的结果进行比较。

在必需基因的识别领域，科研人员普遍使用AUC作为分类算法的评价指标，所以在本课题中，我们也采用了AUC作为最终评估分类模型预测必需基因效果的指标。

2.3.2.2 k折交叉验证

为了保证建立的模型是可靠的，我们采用k折交叉验证的方法，这种方法可以保证模型的误差较小，因为它可以确保原始数据集中的每个观察结果都有机会出现在训练和测试集中。如果输入数据有限，这是最好的方法之一。此方法遵循以下步骤：

1、将整个数据随机分成k份（k的值不应太小或太高，理想情况下，根据数据大小选择5到10）。越高的k值可以使模型的偏差就越小（但是可能会导致拟合过度）；而较小的k值的效果与随机将数据集分成训练集和测试集相似；

2、用其中的k-1份作为训练集来训练分类模型，用剩下的一份作为测试集来验证模型，并计算出AUC的值[12]；

3、重复第2步k次，直到这k份数据集每一份都作为测试集出现过一次，每次都将其余k-1份作为训练集并记录每次的AUC的值。最后求出k个AUC的平均值作为模型的最终评价指标[12]。

2.3.2.3 分类模型的构建

scikit-learn是针对python语言的机器学习库，它提供了一些常见的数据集（例如鸢尾花数据集等）以及包括各种特征提取、特征选择、分类、聚类等机器学习算法，是目前非常受欢迎的机器学习资源库[13]。scikit-learn的官方文档的链接为<https://scikit-learn.org/stable/>，网站内有详细的用户指南和示例。在之前我也介绍了本课题的代码实现主要是基于python语言，而scikit-learn这个库也是我构建模型的重要工具。

使用sklearn的StratifiedKFold方法对提取后的特征矩阵和类别标签进行分层分割，按照正负样本的比例对数据集进行分割。我选择的是5折交叉验证，所以将整个数据集按比例分成5份，每次取其中1份作为测试集，其余4份作为训练集[13]。然后按照之前介绍的方法在训练集上用t检验筛选特征，筛选的策略是选择标准的p值为1e-3，将在每个特征上计算得到的p值小于1e-3的全部归于具有显著性差异的特征。sklearn提供了线性支持向量分类的算法LinearSVC，直接使用该函数建立分类模型，并在训练数据集上进行训练，用训练好的分类模型对测试集进行预测并打分，然后计算出AUC。重复该过程五次，计算得到平均的AUC。最后，将结果用可视化的方法画出直观的曲线图，观察随着特征数目的增加，分类模型评价指标AUC的变化趋势。

2.4 本章总结

本章主要介绍了本次课题主要采用的数据和方法。2.1节介绍了大肠杆菌基因信息数据集的获取和整理，2.2节介绍了基于马尔科夫链模型的特征提取方法和特征递归消除、t检验两种筛选特征的方法，2.3节介绍分类模型的设计与构建，其中，对比了几种机器学习算法和机器学习算法的评价指标，以及分类模型构建的方法。

第三章 实验结果分析

至此，我已经详细介绍了本课题的特征提取方法、特征筛选方法以及分类器的设计与构建，接下来就是对分类效果的分析。我主要是通过核苷酸数目K的增加和间隔λ的增加来不断增加特征数目，然后提高分类器评价指标AUC的值的。

表3-1给出了我选择用不同的K值和λ值时，提取的特征矩阵大小以及采用t检验对特征进行筛选得到用于分类的数据集后，使用构建的分类模型在得到的数据集上进行5折交叉验证后再计算平均的AUC值；

表3-1 特征矩阵信息

|  |  |  |  |
| --- | --- | --- | --- |
| K | λ | 特征矩阵大小 | AUC |
| 2  2  3  3  3  4  4  4  4  5  5  5  5  5  6  6  6  6  6  6  7  7  7  7  7  7  7 | 1  2  1  2  3  1  2  3  4  1  2  3  4  5  1  2  3  4  5  6  1  2  3  4  5  6  7 | (4144, 108)  (4144, 156)  (4144, 492)  (4144, 732)  (4144, 972)  (4144, 2028)  (4144, 3036)  (4144, 4044)  (4144, 5052)  (4144, 8172)  (4144, 12252)  (4144, 16332)  (4144, 20412)  (4144, 24492)  (4144, 32748)  (4144, 49116)  (4144, 65484)  (4144, 81852)  (4144, 98220)  (4144, 114588)  (4144, 131052)  (4144, 196572)  (4144, 262092)  (4144, 327612)  (4144, 393132)  (4144, 458652)  (4144, 524172) | 0.6977128254798154  0.6977128254798154  0.7375505948641768  0.7679198198843957  0.7739188744328562  0.8131350136175813  0.8199856282860963  0.8256500879844306  0.8302752348089438  0.8292357261806241  0.8333275261531569  0.8339229607216181  0.8335864024315125  0.837734877905033  0.837460237799475  0.8411204917967705  0.8411608188496451  0.8437790415671932  0.8431362177415135  0.8457592517915875  0.840246636804572  0.8437367533103721  0.8450352616053969  0.8453701577101096  0.8478701873985255  0.8472195966116527  0.8498514465255281 |

将AUC随着K值和λ值的增加的变化趋势用曲线图表示出来，横轴表示特征的模式K-λ，K表示提取的核苷酸数目，λ表示引入的间隔，曲线图如图5-1所示：

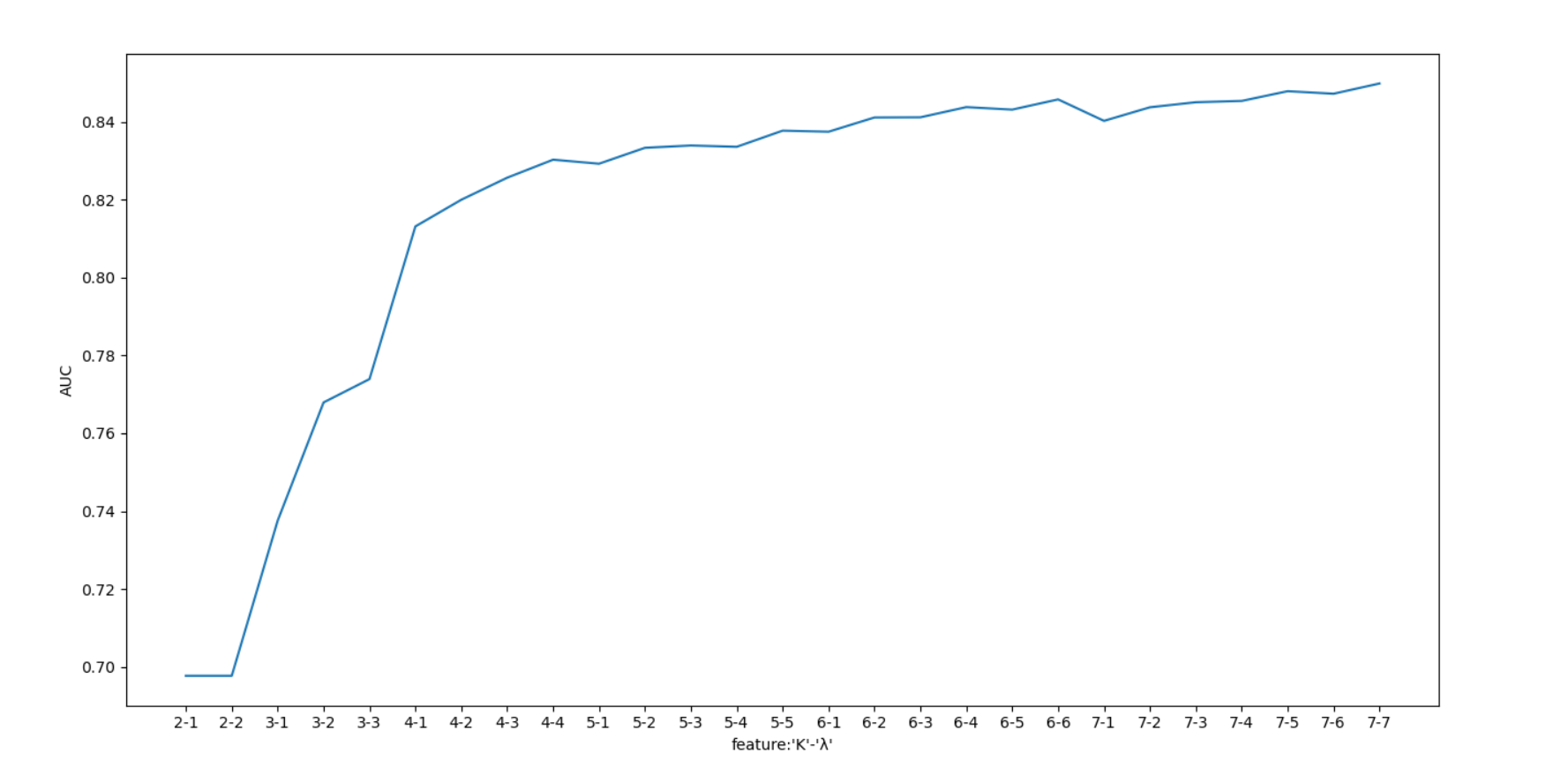


图3-1分类器分类效果随提取特征数目变化曲线

由图3-1可知，随着特征数目和间隔的增加，分类器对数据集进行五折交叉验证的AUC逐渐增大，但是增长速度逐渐变缓，并逐渐收敛于85%。AUC的值达到了毕业设计的预期。

我采用的是线性核的SVM分类算法作为分类器，在实验中我将以线性核为核函数的分类器与以径向基函数作为核函数的分类器了对比，在最终的评价指标上很接近，但是线性核相对于RBF核的算法在时间的花费上更少。在对比实验中，我选择以核苷酸数目K=6，间隔λ=4提取特征的模式上进行采用两种核函数的分类器的对比。特征矩阵的大小为4144\*81852，均采用t检验筛选特征，线性核的分类器进行五折交叉验证的平均AUC为0.8451355159606205，程序执行时间为428.4365087秒；将rbf核的分类器的惩罚系数设为1.0，gamma设为0.5，其进行五折交叉验证的平均AUC为0.8438623766761397，程序执行时间为534.7767515999999秒。由此可见，线性核的分类器在取得和RBF核的分类器相近的分类效果的同时，时间成本更低。所以在本课题中，在SVM的核函数的选择问题上，采用线性核是比采用径向基函数更好的选择。

在特征的筛选上，我们先后使用了t检验的方法和RFE（特征递归消除）的方法，我们发现t检验的时间成本要比RFE的时间成本小得多，并且两种筛选特征的方法在相同的分类器上有这相似的结果，所以采用t检验作为特征筛选的方法是更好的选择。

第四章 总结与展望

4.1 总结

必需基因是生物体或细胞在一定条件下生存所必需的那些基因,因此他们所编码的功能被认为是生命的基础。本文围绕必需基因这个主题，主要研究了必需基因的预测问题。全文首先介绍了生物信息学的发展，阐述了必需基因的相关概念以及它在生命科学研究领域的重要意义。并从现有的文献资料出发，总结概述了目前在必需基因预测领域的研究现状。然后介绍了大肠杆菌基因信息数据集的获取和整理。之后介绍了如何从大肠杆菌的基因序列信息中提取出特征矩阵以及如何使用t检验等方法从高维的特征矩阵中筛选出在不同类别的基因中具有显著性差异的特征。然后对机器学习和几种机器学习算法评价指标进行了介绍并且分析了几种机器学习算法的优缺点，最后选择支持向量机作为分类模型的核心算法。接下来采用五折交叉验证对该模型分类效果进行了评估，随着核苷酸序列的核苷酸数目和引入间隔的增加，AUC的值逐渐增大，达到了0.85，获得了较好的必需基因预测效果。在实验中我还对支持向量机算法的不同核函数做了对比，以及对不同特征筛选方法做了对比。本课题的主要创新点在于特征提取方法的不同，发展的特征表述方法可以推广到其他具有周期性的字符序列，有望成为一种普适性的字符序列特征表述形式。

4.2 未来展望

本课题采用机器学习方法支持向量机对大肠杆菌做了必需基因和非必需基因的分类，达到了0.85的AUC值。分类器的效果达到了毕业设计的预期，但目前仍然存在很多不足，还有许多地方需要改进：

1、特征提取方法的改进：

本课题提取的特征主要是基于基因的DNA序列信息展开的，仅仅利用了DNA的一维信息。实际上，还有很多信息没有利用，比如基因的三维结构、由密码子翻译得到的氨基酸序列信息和对应肽链的三维结构、与基因结合的转录因子以及转录因子结合位点等，在今后的研究中，如果能够把不同层次的信息等提取出来作对比，或者把不同层次的信息整合在一起综合考虑，实验的完整性和效果会更好。

2、分类算法的选择：

本课题在选择分类算法时仅仅从理论上对比了支持向量机和其他分类算法的优势和劣势并最终选择支持向量机。如果在实验中直观地对比几种分类算法的分类效果再进行分类算法的选择会更有说服力。除此之外，近些年来，利用深度学习方法构建深层神经网络在数据挖掘任务中越来越流行。虽然其对计算资源的消耗较大，但在大规模分类任务中等都表现出色，所以可以考虑将深度神经网络作为分类模型并将其与传统的机器学习算法的分类效果进行对比。

3、物种的选择：

本课题只在大肠杆菌上进行了并取得了较好的结果，但是能否将本课题所用的特征提取方法拓展到其他物种并未尝试，在今后的研究中，如果在不同物种之间进行对比会使研究更加完整。

致 谢

2020年注定是不平凡的一年，因为新冠疫情在全世界范围内的蔓延，我们这一届的毕业生只能在家中进行毕业设计，在线上和老师、同学交流。回顾本科期间，我在学习和生活上都承蒙了太多人的照顾和帮助，借此机会，我想对大家表达一下感激之情。

首先，我要感谢我的本科毕业设计指导老师——郭锋彪教授。郭老师是一位在科研工作上一丝不苟、认真严谨的科研人员；在课堂上是幽默风趣、知识渊博的老师；在生活上是关心学生、和蔼可亲的前辈和朋友。我在大学二年级的时候进入郭老师的课题组，在郭老师的指导下开展生物信息学相关课题的研究，老师十分耐心，经常与我在教研室和电话中交流沟通，在我大三和大四准备考研期间，郭老师也经常给予我鼓励和支持。我非常感谢郭老师在我本科期间给予我的帮助。

我还要感谢郭老师课题组的王淑轩师姐、高依舟师兄、华红丽师姐等人，感谢你们在我课题进行过程中给予我的耐心指导。同时感谢严茜同学、谭文祥同学、豆佳乐同学等人，感谢你们在我学习和生活上给我提供的帮助。感谢辅导员万永利老师在本科期间对我的关心和帮助。

祝愿所有的老师、同学和朋友们身体健康，前程似锦！

最后向全国的抗疫英雄们致敬！

参考文献

1. Li Y, Chen L. Big biological data: challenges and opportunities[J]. Genomics, Proteomics & Bioinformatics, 2014, 12(5): 187.
2. Shafee T, Lowe R. Eukaryotic and prokaryotic gene structure[J]. WikiJournal of Medicine, 2017, 4(1): 2.
3. Levitan I B, Kaczmarek L K. The neuron: cell and molecular biology[M]. Oxford University Press, USA, 2015.
4. Zhang Z, Ren Q. Why are essential genes essential?-The essentiality of Saccharomyces genes[J]. Microbial Cell, 2015, 2(8): 280.
5. Dong C, Jin Y T, Hua H L, et al. Comprehensive review of the identification of essential genes using computational methods: focusing on feature implementation and assessment[J]. Briefings in Bioinformatics, 2020, 21(1): 171-181.
6. Yamazaki Y, Niki H, Kato J. Profiling of Escherichia coli Chromosome database[M]//Microbial Gene Essentiality: Protocols and Bioinformatics. Humana Press, 2008: 385-389.
7. Yan K, Zhang D. Feature selection and analysis on correlated gas sensor data with recursive feature elimination[J]. Sensors and Actuators B: Chemical, 2015, 212: 353-363.
8. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics[J]. Bioinformatics, 2007, 23(19): 2507-2517.
9. Chang C C, Lin C J. LIBSVM: A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(3): 1-27.
10. Tharwat A. Classification assessment methods[J]. Applied Computing and Informatics, 2018.
11. Fawcett T. An introduction to ROC analysis[J]. Pattern recognition letters, 2006, 27(8): 861-874.
12. Airola A, Pahikkala T, Waegeman W, et al. An experimental comparison of cross-validation techniques for estimating the area under the ROC curve[J]. Computational Statistics & Data Analysis, 2011, 55(4): 1828-1844.
13. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python[J]. the Journal of Machine Learning Research, 2011, 12: 2825-2830.
14. Itaya M. An estimation of minimal genome size required for life[J]. FEBS letters, 1995, 362(3): 257-260.
15. Deng J, Deng L, Su S, et al. Investigating the predictability of essential genes across distantly related organisms using an integrative approach[J]. Nucleic Acids Research, 2011, 39(3): 795-807.
16. Guo F B, Dong C, Hua H L, et al. Accurate prediction of human essential genes using only nucleotide composition and association information[J]. Bioinformatics, 2017, 33(12): 1758-1764
17. Ji Y, Zhang B, Van S F, et al. Identification of critical staphylococcal genes using conditional phenotypes generated by antisense RNA[J]. Science, 2001, 293(5538): 2266-2269.
18. Mobegi F M , Aldert Z , De J M I , et al. Advances and perspectives in computational prediction of microbial gene essentiality[J]. Briefings in Functional Genomics(2):2.
19. Mushegian A R, Koonin E V. A minimal gene set for cellular life derived by comparison of complete bacterial genomes[J]. Proceedings of the National Academy of Sciences, 1996, 93(19): 10268-10273.
20. Chen Y, Xu D. Understanding protein dispensability through machine-learning analysis of high-throughput data[J]. Bioinformatics, 2005, 21(5): 575-581.
21. Hart T, Chandrashekhar M, Aregger M, et al. High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities[J]. Cell, 2015, 163(6): 1515-1526.

外文资料原文

**Comprehensive review of the identification of essential genes using computational methods: focusing on feature implementation and assessment**

**Chuan Dong∗, Yan-Ting Jin∗, Hong-Li Hua, Qing-Feng Wen, Sen Luo, Wen-Xin Zheng and Feng-Biao Guo**

**Abstract**

Essential genes have attracted increasing attention in recent years due to the important functions of these genes in organisms. Among the methods used to identify the essential genes, accurate and efficient computational methods can make up for the deficiencies of expensive and time-consuming experimental technologies. In this review, we have collected researches on essential gene predictions in prokaryotes and eukaryotes and summarized the five predominant types of features used in these studies. The five types of features include evolutionary conservation, domain information, network topology, sequence component and expression level. We have described how to implement the useful forms of these features and evaluated their performance based on the data of *Escherichia coli* MG1655, *Bacillus subtilis* 168 and human. The prerequisite and applicable range of these features is described. In addition, we have investigated the techniques used to weight features in various models. To facilitate researchers in the field, two available online tools, which are accessible for free and can be directly used to predict gene essentiality in prokaryotes and humans, were referred. This article provides a simple guide for the identification of essential genes in prokaryotes and eukaryotes.

**Key words:** essential genes; computational methods; feature assessment; online tool

**Introduction**

The most important genes on the whole-scale genome of an organism that are in charge of fundamental life activities, such as survival or reproduction, are known as essential genes. Because of the vital function of essential genes for living organisms, these genes have attracted growing attentions. In prokaryotes, many research groups have utilized experimentalmethods, such as transposon mutagenesis and RNA interference, to identify these genes. Studies on bacterial essential genes are beneficial to comprehensively understand the essence of life, to reconstruct a minimal gene set artificially and to screen out potential drug targets to treat pathogenic diseases. Regarding eukaryotes, numerous essential genes have been identified for several species and cell lines. Recently, several studies identifying human essential genes were consecutively published. These genes were determined by experimental technologies, such as RNA interference, CRISPR-Cas9 or retroviral gene trap. Of note, the essential genes identified in human cancer cell lines can be used to identify therapeutic targets for corresponding diseases. To date, 48 bacterial essential gene sets and 26 eukaryotic essential gene sets determined by experimental methods have been deposited into the DEG database.

To better understand gene essentiality factors and provide more essentiality information, many research groups have adopted more cost-effective and time-efficient approaches of bioinformatics to predict essential genes. In fact, using computational biology methods to predict essential genes has been used for decades. The first computational approach was developed based on comparative genomics by Mushegian *et al*. in 1996. Subsequently, Chen *et al*. first introduced themachine learning method into the field of predicting protein dispensability in 2005. They analyzed *Saccharomyces cerevisiae*’s protein dispensability based on the combination of evolutionary rate, duplication rate, protein’s connectivity in protein–protein interaction network (PPI network) and gene-expression correlation network using neural network and support vector machine (SVM). Deng *et al.* used intrinsic genomic features, context-dependent features and others as the input data of four classifiers to train models to predict essential genes, which investigated the accuracy of models in distantly related organisms for the first time. In 2012, Yuan *et al.* applied the machine learning method to predict essential genes for mouse, which was the first time to predict essential genes for higher eukaryote. Recently, Guo *et al.* used sequence component (SC) features combined with SVM to predict essential genes for human and provided an online service. All of these works mentioned above drew upon computational models and made good predictions.

In general, computationalmethods, as alternative approaches, already have a wide range of research applications in the field of identifying essential genes. In this article, we have performed a comprehensive review of the prediction of essential genes from the perspective of various types of biological features and modeling approaches, where modeling approaches meant machine learning technique or empiric formula used to integrate

feature variables. For more particular knowledge of how various features contribute to predicting essential genes or proteins in computational methods, we have made a series of tests through choosing some representative features from each type to perform the prediction instead of only summarizing or elaborating published works. According to these tests and summarizations, selecting the proper features is necessary and helpful for predicting essential genes.

**Results**

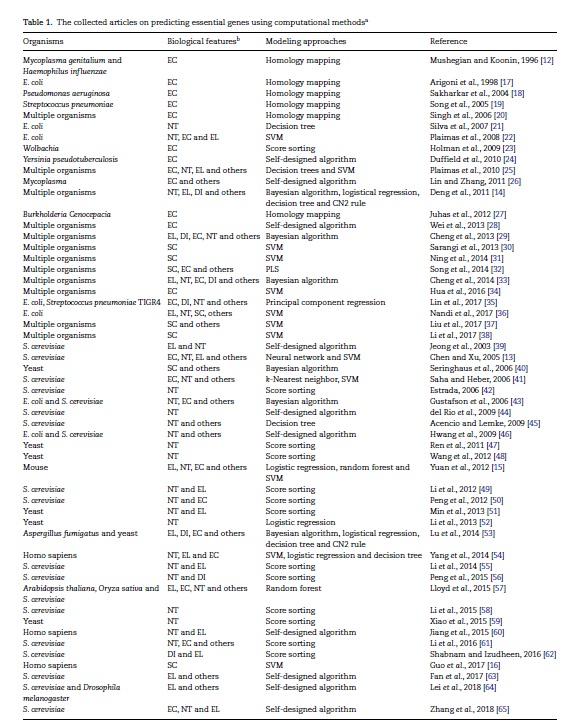
**Feature summarization**

To understand the computational methods used in identifying essential genes, we have displayed related studies that were collected manually in Table 1. Specially, most of the works, but not all, were listed in this table. Based on the information exhibited in Table 1, we summarized five typical biological features that include evolutionary conservation (EC), domain information (DI), network topology (NT), SC and expression level (EL) to perform a thorough inquiry and discussion. The five groups of features demonstrate a high capacity in differentiating essential and non-essential genes. The common and efficient biological features could be summarized as the following categories.

**Evolutionary conservation**

Persistent genes are genes that are conserved in certain bacterial genomes, even if they are not omnipresent. These genes are usually expressed at high levels, located on the leading DNA strand and code for proteins that either perform essential functions or are involved in maintaining and repairing indispensable functions. Thus, certain persistent genes could be presented as essential in the long term of evolutionary progress

for a species. The ‘knockout-rate’ prediction provides further evidence that essential genes are more stringent than nonessential genes when facing negative selection. Mushegian and Koonin have noted that numerous essential genes could be conserved across organisms. With the advent of highthroughput sequencing technology, it is easy to obtain the whole-genome sequences for a target organism. Therefore, the method of comparative genomics has been widely applied in distinguishing essential genes from non-essential genes. Several research groups have effectively exploited genome comparisons to identify essential genes for exploring potential drug targets. Because of the characteristics of highly conserved essential sequences across multiple organisms, this type of feature contributes to discovering broad-spectrum antibacterial targets. In the 54 articles retrieved, characteristics of EC were applied in 27 studies. Remarkably, some of these studies adopted EC as their major identifying feature.



**Domain information**

Compared with EC of an entire gene or protein, protein domainrepresents the specific conservation of the local sequences. A domain may exist in various proteins, and a protein may contain several domains. As independent and stable units, diverse domains can perform their functions independently. Deng *et al*. first utilized this type of feature to predict essential genes. Among their selected types of features, DI was the strongest for discriminating essential and non-essential genes. Since then, as a helpful characteristic, DI has appeared in various approaches. As measured by Wang *et al.*, protein domain types and their frequencies are two important elements in exploring the relationship between protein domains and essential proteins. The proteins with more domain types are more likely to be essential because essential proteins are more conserved and have independent biological features. Furthermore, proteins tend to be essential if they contain domains that infrequently emerge in other proteins because the function of these proteins cannot be substituted easily. Seven articles among the collected 54 research articles used DI to distinguish essential genes from non-essential genes, individually or in combination with other features. Both score sorting, which was calculated by self-designed formulas, and machine learning methods were adopted to identify essential genes when using this feature.

**Network topology**

Similar with the feature of EC, topologies in PPI network are popular in the recognition of essential proteins by computational methods. Zhang *et al.* generalized a number of works about predicting essential genes based on network topological features and machine learning methods in their review article. The commonly used features for describing topologies of PPI networks are degree centrality, betweenness centrality, closeness centrality and clustering coefficient. Furthermore, some researchers have also extracted other features using graph theory. In 2001, Jeong *et al.* explored and elucidated a positive correlation between gene lethality and degree centrality by Pearson’s linear correlation coefficient. Additionally, essential genes with greater degrees tend to be network hubs and have more neighbors. Therefore, most studies considered these features in the prediction of lethal genes or proteins and noted that the topologies of PPI networks have important roles in developing essential gene-predicting models. In total, 32 works predicting essential genes were performed through the characteristics of network topologies, amajority of the 54 chosen papers. In addition to the commonly used PPI networks, several other types of biological networks, such as metabolic networks, gene co-expression networks, transcriptional networks and their integrated networks, were utilized for mining effective features related to essential genes or proteins. Briefly, this is an important and often used type of feature for studying gene essentiality/

**Sequence component**

An SC usually refers to a DNA or protein sequence that can be characterized as a group of different oligonucleotides or amino acid composition forms such as *K*-mer. This feature extraction method can transform the short or long amino acid sequences or base sequences into an occurrence frequency of *K-*mer in the target sequence, or it can be characterized as a regular representation of a given sequence such as the Z curve method. Supposedly, the important difference between essential and non-essential genes is that there are distinct sequence compositions between them. Therefore, these two types of genes have diverse evolutionary rates and conservation and should perform different functions. With the development of high-throughput sequencing technology, genome sequence data provide insights into predicting essential genes using SC features. It is a newly proposed feature extraction method in recent years. All of the features are extracted directly from nucleotide sequences or amino acid sequences. Although only eight works reported this feature among our collected 54 articles, nearly all works uniquely adopted this type of feature to predict essential genes and acquired good results. Sarangi *et al.* first introduced Chou’s pseudo amino acid composition and various physic-chemical features into the identification of essential genes. Similarly, our group analyzed the SC of essential and non-essential genes based on primary sequences (nucleotide or protein sequences). Subsequently, a linear method called Z-curve used partial least squares (ZUPLS) was provided to predict bacterial essential genes based on Z-curve theory and other sequence-based features combined with partial least squares (PLS) by Song *et al*. Recently, an online prediction tool called predictor of human essential genes (Pheg) was proposed to predict human essential genes using only nucleotide composition features, which were designed based on the *λ*-interval Z curve method. The sequence-based method is general and universal for sequenced genomes and can achieve great predictions.

**Expression level**

The EL of essential genes represents a narrower range than that of non-essential genes, and essential genes are more likely to be expressed at higher levels. Although no article in Table 1 used EL alone to identify essential genes, 20 studies integrated this type of feature with others to make predictions. Several studies note that interacting proteins in PPI networks are more likely to be co-expressed as these proteins are often present in the same pathway. Accordingly, most of these methods calculated the topologies in a gene co-expression network, combining the PPI network and gene expression profile. Similarly, other approaches calculated the average expression value as the feature value for each gene because of the relatively higher EL of essential genes.

**Feature implementation and assessment**

In the section above, we have summarized five commonly used and representative types of features for predicting essential genes. We have not only described these generalizations and provided relevant literatures but also assessed these features using standard data sets of essential genes in this review, which is helpful for understanding and utilizing these features.

**Data resources and evaluation method**

To assess the effectiveness of each class of independent features summarized above, we used three representative data sets as a benchmark to predict essential genes. The gene essentiality data of Gram-negative bacterium *Escherichia coli* K-12 MG1655 (*E. coli*) was collected from the Profiling of *E. coli* Chromosome (https://shigen.nig.ac.jp/ecoli/pec/) database. This data set contains 287 essential genes and 3857 non-essential genes. The gene essentiality data of Gram-positive bacterium *Bacillus* *subtilis* 168 (*BAS*) was downloaded from the DEG database. It contains 271 essential genes and 3904 non-essential genes. The benchmark data set of humans was obtained from the research article of Guo *et al.,* and 1516 essential genes and 10 499 non-essential genes were collected for this species. The data quality of these three species may be more reliable among the species with known essential gene information. To build the prediction model, we assigned a Boolean value for each gene to mark its essentiality. A gene was marked as 1 if it presented in the essential gene data set; otherwise, it was marked as −1. DNA and protein sequences of all genes were downloaded from the file transfer protocol (ftp) site of National Center for Biotechnology Information (NCBI). DI was obtained from the Pfam database, and the domains of each gene were searched by the hidden Markov model algorithm. The PPI network data were downloaded from the STRING database. The gene expression data were downloaded from Gene Expression Omnibus (GEO) for *E. coli* and *BAS* (GSE56133 for *E. coli* and GSE45933 for *BAS*) and from The Cancer Genome Atlas (TCGA) for human.

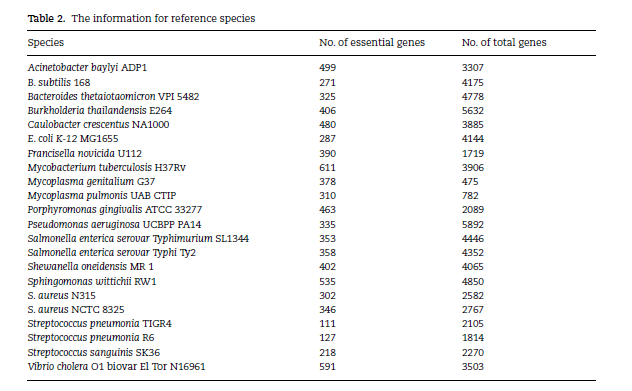
SVM is the most popular machine learning method in predicting essential genes among our collected 54 papers. To make a preliminary assessment for each type of features, SVM was selected to train the model and test the feature set when it was needed. The area under the receiver operating characteristic (ROC) curve (AUC) was used to evaluate the predictions. In this review, we are aiming to provide a comprehensive review of the prediction of essential genes by computational methods from several aspects instead of discussing the prediction results. Therefore, we selected the proper models based on the papers listed in Table 1 to assess each specific characteristic. Some brief descriptions on the performance of various types of features were provided in the following sections.

**Implementation and assessment of EC**

For the bacterial species, Cheng *et al.* have noted that the transferability of essentiality annotations between species is significantly affected by the quality of the training data set. Considering this point and our previous works, we selected 21 species (all species in Table 2 except for the query species *E.* *coli* or *BAS*) from the DEG database as the reference species. The essential gene data sets and protein sequences of these genes were downloaded from the DEG database and NCBI, respectively. We assessed the number of orthologousessential sequences between the target species and its closest species in our selected organisms. The closest species for *E. coli* and *BAS* was determined by composition vector (CV) algorithm. Using the Basic Local Alignment Search Tool for Protein (BLASTp) program, we found that *E. coli* and its closest species *Salmonella enterica serovar Typhimurium* SL1344 (*STS*) shares 253 common orthologous-essential genes and 3204 orthologous genes. Additionally, 286 of 287 essential genes of *E. coli* have orthologous sequences in *STS*. In the same way, it was found that there are 206 orthologous-essential genes and 1575 orthologous genes between *BAS* and its closest species *Staphylococcus aureus* NCTC 8325 (*STN*), and 248 of 271 essential genes of *BAS* have orthologous sequences in *STN*. These data showed high proportion of essential homology between two closely related species. In addition, we adopted the algorithm of Gene Essentiality Prediction Tool based on Orthology and Phylogeny (Geptop) (1) to predict the essential genes.

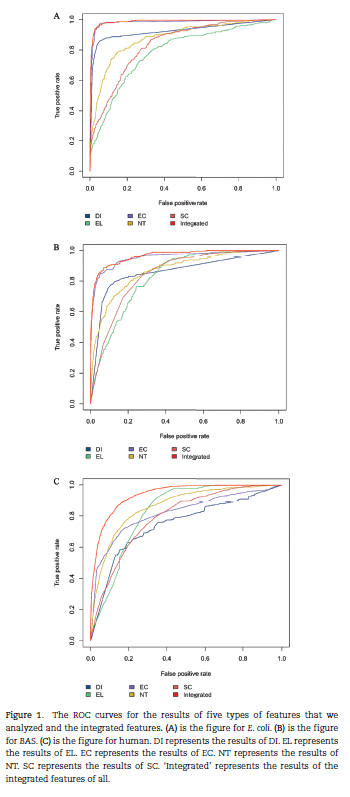


In which, *Eij* is the mark of essentiality for the optimal orthologous gene of query gene *i* in *j*th reference species; *Dj* is the evolutionary distance between the target species and the *j*th reference species; *N* is the total number of reference species, which is equal to 21 in this work; and *Si* is the essentiality score for gene *i* in query species. The expected essentiality scores of target genes were compared with the observed essentiality labels, and an AUC value of 0.9829 and 0.9543 was acquired, respectively, for *E. coli* (Figure 1A) and *BAS* (Figure 1B), which was better than the result obtained from homology mapping only. Obviously, the feature based on essential persistence of genes is more suitable for predicting essential genes than that just based on persistence of genes. For the higher eukaryote such as human, there is not enough number of eukaryotic reference species with validated essential genes. Therefore, we downloaded orthologs between human and 19 other species from HCOP (http:// www.genenames.org/cgi-bin/hcop). If a gene has ortholog in a species, then ‘1’ will be marked; otherwise, ‘0’. Finally, we obtained 19 features by this method. These 19 features combined with the essentiality of all human genes were used as the input data of SVM. An AUC value of 0.8186 was acquired through 10-fold cross-validation (Figure 1C). Although the method of homology mapping would be restricted by the number of conserved genes and the evolutionary distances between various organisms, themethod is still quite effective and reliable for identifying essential genes in closely related organisms. It is noteworthy that this method has reliability and generalization. Undeniably, homology mapping is an efficient characteristic for exploring essential genes if there is enough number of reference species. However, the method requires the use of genomes with known gene essentialities as a reference for alignment. If this information is not available, the performance of the homology mapping feature will be discounted.



**Implementation and assessment of DI**

Deng et al., who were the first to exploit the domain feature and the first to perform the prediction in distantly related organisms, reported that DI provided the greatest contribution to the prediction of essential genes among their refined 13 features. In this part, DI of each protein was acquired from a database of protein family alignments (P-famA) entries. Proteins with no domain annotation were excluded. Accordingly, only 3637, 2623 and 9832 genes for E. coli, BAS and human, respectively, were retained to evaluate this type of feature. We used the essential domain prediction (EDP) model proposed by Lu et al. to identify essential domains. To objectively evaluate this feature, the closest species with known gene essentialities of the target species was selected as the reference species to calculate the domain essentiality scores using the EDP model. For a protein in the target species, its protein essentiality score is measured by the max domain score, in which the domain belongs to this protein. Therefore, STS and STN were used as the reference species for calculating domain essentiality scores for E. coli and BAS, respectively. As for human, mouse was selected as the reference species because of no other higher eukaryote having known gene essentiality information. The gene essentialities of mouse were downloaded from OGEE database. AUC values of 0.9249, 0.8592 and 0.7442 were acquired by comparing the essentiality scores and the essentiality labels in the benchmark data set of *E. coli*, *BAS* and human (Figure 1), respectively. Although these results were acquired only for the proteins with domain annotation, the performance of this feature was still great for the former two target species. These results were the best in identifying essential genes using DI. Even so, the complete DI of research species is still required for the better predictions. The poor performance in human might be caused by relatively far evolutionary distance between reference (mouse) and query species (human). It indicates that reference species should be closely related with the target species if we want to obtain excellent performance. Moreover, we found that 208 of 287 essential proteins in *E. coli* did not share any common domain with non-essential proteins. This phenomenon indicated that some domains are related to gene essentiality and are unique for several proteins to perform specific and important functions. Therefore, it should be noted that the performance of the domain feature would be poor if the method of *n*-fold cross-validation was performed in a genome. This feature will only be effective and meaningful when cross-species validation is performed because there are various domains in a genome and some domains are unique, especially for essential genes. Additionally, the gene essentialities in reference species should be known. Therefore, only by making a whole-genome statistic can we acquire an overall profile of domain distribution.

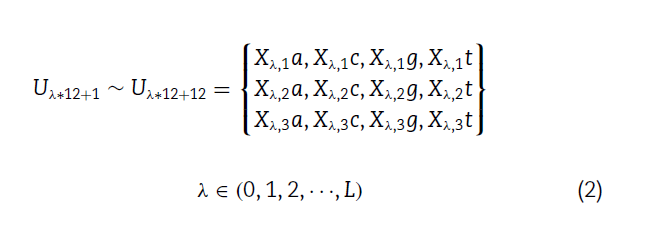


**Implementation and assessment of NT**

In this part, we have downloaded PPI data from the STRING database. The cut-off value for the interaction score between two proteins was set at a default value of 0.4. For some proteins absent from this network, we will set the class corresponding average value of each topological feature to them. The following four commonly used and easily available network topologies, degree centrality, betweenness centrality, closeness centrality and clustering coefficient, were chosen to train and validate the prediction model. These features combinedwith the essentiality of all target species’ genes were used as the input data of SVM. AUC values of 0.8780, 0.8700 and 0.8619 were acquired through 10-fold cross-validation for *E. coli*, *BAS* and human (Figure 1), respectively. The precision of 0.7875 for essential genes and precision of 0.8792 for non-essential genes for *E. coli* were obtained. Silva *et al.* previously acquired average precision value of 0.811 for essential genes and 0.818 for non-essential genes using the features of NT for *E. coli*. Although these results were not the best when using network topological features, it still illustrated that network topologies are truly associated with gene essentiality. In our collected 32 papers related to network topological features, approximately half of them report the utilization of them in conjunction with other types of features. The information of PPI should be as complete as possible. Performance of the network topologies relies on the detailed form of each topology. Optimizing their characterization is important and may significantly improve the effects.

**Implementation and assessment of SC**

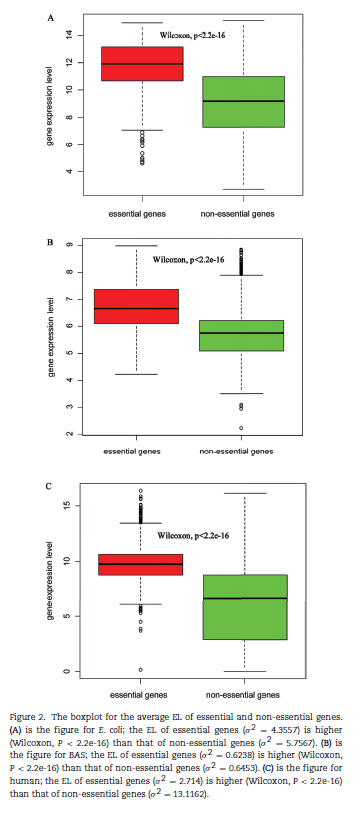
The four articles listed in Table 1 utilized different methods to calculate SC features (see Sequence component section). In this article, we used another formula constructed on the basis of phased oligonucleotides (2) to describe the gene sequences.



In which, *Uλ*∗12+*i* represents the (λ ∗ 12 + i)th feature; *Xλ*,*j* represents *k*-nucleotide, whose first base is located at the *j*th(*j* = 1, 2, 3) phase of a codon, *λ* is the internal length between *X* and the separated *a*, *c*, *g*, or *t*. For fixed *k* value, *X* has 4kcombinedforms; *n* represents the last base in the virtual oligonucleotides *Xλ*, *jn* (*n* = *a, c, g, t*). We used *fXλ*,*jn* to represent the frequency ofoligonucleotides *Xλ*,*jn* in a DNA sequence. Specially, this formuladescribes the frequency of four bases in three phaseswhen *k* = 0 and *λ* = 0. Thus, we could get 4k+1 ∗ 3 ∗ *(λ* + 1*)* features for each *k* (*k* = 0, 1, 2, 3, · · · ). To consider more possible nucleotide compositions,we set the parameters of k = 1, 2, · · ·, 7 and λ = 0, 1, · · ·, 6 to extract features and combined all of these features with the essentiality of *E. coli* genes as the input data of the SVM. AUC values of 0.8314, 0.8502 and 0.8084 were acquired through 10-fold cross-validation for *E. coli*, *BAS* and human (Figure 1), respectively. Nandi *et al.* previously used sequence features to predict essential genes of *E. coli* and acquired AUC value of 0.862*.* Notably, we did not perform feature selection in this partas we only wanted to make a basic evaluation of SC features in gene essentiality. Better predictions could be obtained if featureselection was performed or more features were derived from longer oligonucleotides. Although we did not use the characteristics proposed in other papers, the predictions were not so far from the results by them. An obvious advantage of sequence composition is that it is very convenient to obtain features. Thesefeatures can be directly derived from DNA sequences. However, an inadequacy is that it takes long runtime due to the large number of features.

**Implementation and assessment of EL**

Although noworks focused on researching essential genes using only EL, this feature has been integrated within other features in several studies. In this part, we attempted to explore the performance of expression characteristic for gene essentiality prediction. For the expression data obtained from the GEO database and TCGA database, we calculated the average EL for each gene of *E. coli*, *BAS* and human as the feature value. The feature values combined with the essentiality labels of all genes for each species were used to calculate the AUC values. Finally, AUC values yielded by them were 0.7839 for *E. coli*, 0.8157 for *BAS* and 0.8272 for human (Figure 1). While these results are not particularly spectacular, it reflects the difference between essential genes and non-essential genes to a certain degree. This feature is usually used through being in combination with other types of features. Additionally, we have tried to analyze the relationship between gene EL and the gene essentialities from another angle. We drew boxplots (Figure 2) to reflect the distribution of gene expressions between essential and nonessential genes. Although there are a few outliers in essential genes or non-essential genes, it is clear that the EL of essential genes is higher than that of non-essential genes. Additionally, the influence of outliers can be ignored when this feature is combined with other features to perform prediction. Nevertheless, this feature is often used by being integrated into co-expression network together with PPI network. It must be noted that the selected expression profile data should be tested under the rich medium but not the conditional growth environment. In addition, it should be combined with other features because of its relatively poor performance.



**Implementation and assessment of all features**

In addition to the features analyzed above, we integrated all of these features together to make prediction for the three benchmark species. Zero was set for the missing values in domain feature because no domain annotation was found for those proteins. The missing values in network topologies or EL were set as the average value of each class of genes in each feature. Finally, AUC values of 0.9874, 0.9655 and 0.9375 were yielded through 10-fold cross-validation of SVM for *E. coli*, *BAS* and human (Figure 1), respectively. All of the predictions were improved even for *E. coli* and *BAS*, which have obtained quite satisfactory results using single type of features. In addition, the advantage of integrated features was reflected the most in human. These results indicate that the integrated method could greatly improve the results of prediction compare to those based on single type of feature only. It could be widely applied in essential gene prediction in the future.

**Modeling approaches integrating features in identifying essential genes**

The modeling approaches of integrating features in identifying essential genes can be categorized as empiric formulas and machine learning methods. The empirical formulas are always used when there is some prior knowledge about reference species or feature distribution and the features are homogeneous. For instance, Geptop uses evolutionary distance to weight each feature of homology mapping against essential genes in various reference species. UDoNC is a method that utilizes the domain features and their topological properties in protein-protein interaction to identify essential genes. These empirical formulas are efficient in categorizing gene essentiality. Alternatively, machine learning methods, the algorithms that can address large and complex data sets rapidly and automatically, are also applied in the field of predicting essential genes. These methods can usually be divided into three types: supervised learning algorithms, unsupervised learning algorithms and semi-supervised learning algorithms. Supervised learning methods such as SVM, decision tree, Naive Bayes and logistic regression are commonly used to identify essential genes. The input data of these approaches consist of the desired output values and a series of attribute vectors. Compared with an empiric formula, machine learning methods can integrate more types of features together to develop the prediction classifier and filter the useless features using feature selection. However, the predictions could be different using different machine learningmethods even for the same organism with the same features. Therefore, some researchers choose the approach that can produce the best results. For example, Lloyd *et al.* selected the best model and outputs of the random forest method. Others combine the outputs of several independent classifiers. Deng *et al.* chose to integrate the predictions of four diverse classifiers using an unweighted average approach. Remarkably, the performance of machine learning methods can be greatly influenced by the feature selection or various types of feature combinations. Continual optimization of parameters and training features is needed to improve the performance of classifiers.

**Online services for predicting essential genes**

With the development of computationalmethods, several online essential gene prediction servers have been provided for users. Geptop, a gene essentiality prediction tool for sequenced bacterial genomes, was designed based on orthology and phylogeny. Orthology is estimated by the reciprocal best hit. Phylogeny is represented by the evolutionary distance between query and reference species, which is calculated through the CV algorithm. Researchers can upload the DNA or protein sequences in the fasta format and then they can receive the prediction results including the predicted essentiality of genes or proteins and the essentiality probability. Users can use this tool online or obtain its stand-alone version at http://cefg.uestc.edu.cn/geptop/ for free. Pheg, which can be used to predict gene essentiality in humans,was developed based on *λ*-interval Z-curve theory. It uses only nucleotide composition and association information as training data of the machine learning method. Pheg can predict essentiality for anonymous gene sequences from humans. Users can input the DNA sequences in the fasta format in the input box. This web server can be freely accessed at <http://cefg.uestc.edu.cn/Pheg>.

**Conclusion**

Identifying essential genes using computational methods has received increasing attention because of the increasing volumes of data and the improvement of machine learning algorithms. Various types of characteristics have been extracted fromexperimental data and many modeling approaches have been used to predict essential genes in silico. In this review, we have collected articles on essential gene prediction through computational algorithms and analyzed the efficiency of each type of features.

For different characteristics, the performance of homologymapping depends on the essentiality-known reference species and the evolutionary distance between them. There are more orthologous-essential genes between two relatively closely related species, such as *E. coli* and *STS*. In contrast, there are fewer orthologous-essential genes between two distantly related species, such as *E. coli* and *Mycoplasma pulmonis* UAB CTIP. It’s a good choice to select the homology mapping method such as Geptop to predict essential genes in prokaryotes because there is enough number of reference species with known essential genes and the results of prediction by it are well. DI would lead to better results if the DI is obtained fromclosely related species. Selecting large number of species with known essential genes as the references to calculate domain essentiality scores is a new idea for exploring. However, lack of domain annotation in some proteins is a problem that should be addressed. This feature is unsuitable for eukaryotes yet due to the fact that no suitable species can be used as reference species. Network topologies ref lect the pivotal role of essential genes in biological networks. This feature was used in most articles we have collected. Their performance is efficient both in prokaryotes and eukaryotes. The characterizations of NT can also be further developed. High-quality and sufficient network data help to acquire good results. EL is a type of feature that is not often used alone. The performance of it is not as good as the others, and there are a certain number of missing values in EL data. Integrating multiple networks and gene expression data is a commonly used approach to improve prediction performance. For SC, the greatest advantage is that it is convenient to extract features from sequences and there will be no missing values. On the contrary, its disadvantage is that too many features will cause poor operation efficiency. Feature selection is a good choice to remove the redundant information and improve the results of prediction for SC feature. This type of features is applicable to eukaryotes and prokaryotes. In addition, homology mapping, DI and SC are more suitable for identifying essential genes across organisms if close species have gene essentiality data.

It is still a challenge to explore prediction models for distantly related species using commonly efficient features. Discovering new features could creatively improve the predictions. Inaddition, we should adopt more effective forms to improve the features found previously, select more significant features, remove the multicollinearity among features and optimize training models in the following study. Integrating these features together and introducing deep learning methods into this field may lead to unexpected surprises. Another important and interesting issue is to investigate the computational predictability of essential genes across three domains, particularly between organisms belonging to different domains but with similar genomic G+C content.

外文资料译文

使用计算方法全面鉴定必需基因：专注于功能实现和评估

摘 要

由于必需基因在生物体中的重要功能，近年来它们已引起越来越多的关注。在用于鉴定必需基因的方法中，准确而有效的计算方法可以弥补昂贵且费时的实验技术的不足。在这篇综述中，我们收集了原核生物和真核生物中必需基因预测的研究，并总结了这些研究中使用的五种主要特征。这五种特征包括进化保守性，域信息，网络拓扑，序列组成和表达水平。我们已经描述了如何实现这些特征的有用形式，并根据大肠杆菌MG1655，枯草芽孢杆菌168和人类。描述了这些功能的前提条件和适用范围。此外，我们还研究了用于各种型号的重量特征。为了方便该领域的研究人员，参考了两个可用的在线工具，这些工具可免费使用，并可直接用于预测原核生物和人类的基因必需性。本文为鉴定原核生物和真核生物中的必需基因提供了简单指南。

关键字：必需基因；计算方法；在线工具。

介绍

生物体整个基因组中负责基本生命活动（例如生存或繁殖）的最重要基因被称为必需基因。由于必需基因对于生物体的至关重要的功能，这些基因引起了越来越多的关注。在原核生物中，许多研究小组利用了实验方法，例如转座子诱变和RNA干扰来识别这些基因。对细菌必需基因的研究有利于全面理解生命的本质，人为地重建最小的基因集，并筛选出治疗病原体疾病的潜在药物靶标。关于真核生物，已经为几种物种和细胞系鉴定了许多必需基因。最近，相继发表了一些鉴定人类必需基因的研究。这些基因是通过实验技术确定的，例如RNA干扰，CRISPR-Cas9或逆转录病毒基因陷阱。值得注意的是，在人类癌细胞系中鉴定出的必需基因可用于鉴定相应疾病的治疗靶标。迄今为止，有48种细菌必需基因集和26种通过实验方法确定的真核必需基因集已保存到DEG数据库中。

为了更好地理解基因必要性因素并提供更多必要性信息，许多研究小组采用了更具成本效益和时间效率的生物信息学方法来预测必需基因。实际上，使用计算生物学方法来预测必需基因已经使用了数十年。第一种计算方法是基于Mushegian等人的比较基因组学开发的。1996年。随后，Chen等人。首先在2005年将机器学习方法引入预测蛋白质可分配性的领域。他们基于进化速率，重复率，蛋白质-蛋白质相互作用网络（PPI网络）和基因表达相关网络中蛋白质的连通性，分析了酿酒酵母的蛋白质可分配性。使用神经网络和支持向量机（SVM）。邓等。使用固有的基因组特征，上下文相关特征等作为四个分类器的输入数据，以训练模型来预测必需基因，这是首次研究远距离生物中模型的准确性。2012年，袁等人。运用机器学习方法预测小鼠必需基因，这是第一次预测高等真核生物必需基因。最近，郭等人。结合使用了序列成分（SC）功能与SVM一起预测人类必需的基因，并提供了在线服务。上面提到的所有这些工作都依靠计算模型并做出了很好的预测。

通常，作为替代方法，计算方法已经在鉴定必需基因领域中具有广泛的研究应用。在本文中，我们从各种生物学特征和建模方法的角度对基本基因的预测进行了全面的回顾，其中建模方法意味着机器学习技术或用于集成特征变量的经验公式。为了更详细地了解各种特征如何在计算方法中预测必需的基因或蛋白质，我们通过从每种类型中选择一些具有代表性的特征来进行预测，而不是仅对已发表的著作进行汇总或精心制作，进行了一系列测试。根据这些测试和总结，选择合适的特征对于预测必需基因是必要的且有帮助的。

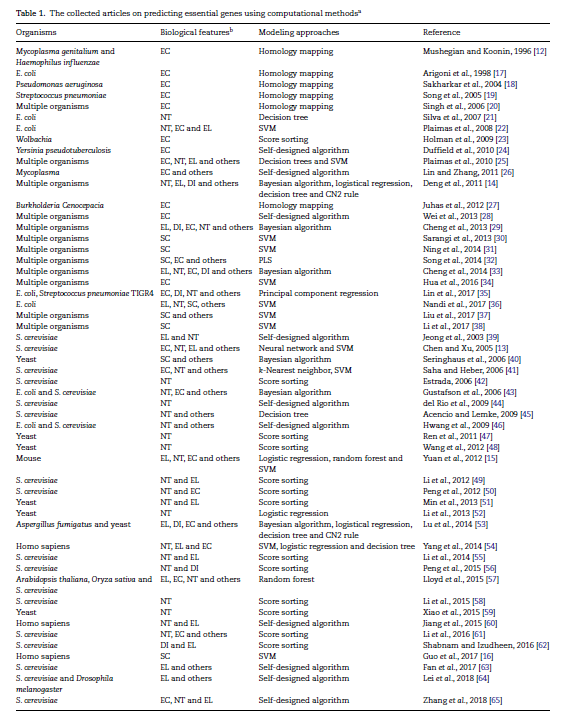
结果

特征汇总

为了了解用于鉴定必需基因的计算方法，我们在表1中显示了手动收集的相关研究。特别是，此表中列出了大部分（但不是全部）工作。根据表1中显示的信息，我们总结了五个典型的生物学特征，包括进化保守性（EC），域信息（DI），网络拓扑（NT），SC和表达水平（EL），以进行全面的研究和讨论。这五组特征证明了区分必需基因和非必需基因的高能力。共同和有效的生物学特征可归纳为以下几类。

进化保守性

持久基因是某些细菌基因组中保守的基因，即使它们不存在。这些基因通常以高水平表达，位于前导DNA链上并编码执行基本功能或参与维持和修复必不可少功能的蛋白质。因此，某些持久性基因可能在一个物种的长期进化进程中被认为是必不可少的。“敲除率”预测提供了进一步的证据，证明当面对否定选择时，必需基因比非必需基因更严格。Mushegian和Koonin指出，许多必需基因在整个生物体中都可以保守。随着高通量测序技术的出现，很容易获得目标生物的全基因组序列。因此，比较基因组学的方法已被广泛应用于区分必需基因和非必需基因。几个研究小组已经有效地利用基因组比较来鉴定用于探索潜在药物靶标的必需基因。由于跨多个生物的高度保守的基本序列的特征，这种类型的特征有助于发现广谱抗菌靶标。在检索到的54篇文章中，EC的特征被应用于27项研究中。值得注意的是，其中一些研究采用EC作为其主要识别特征。



域信息

与整个基因或蛋白质的EC相比，蛋白质结构域代表了局部序列的特定保守性。域可以存在于各种蛋白质中，并且蛋白质可以包含多个域。作为独立且稳定的单位，不同的域可以独立执行其功能。Deng等首先利用这种类型的特征来预测必需基因。在他们选择的特征类型中，DI最能区分必需基因和非必需基因。从那以后，作为一种有用的特性，DI已经以各种方式出现。根据Wang等人的测量，蛋白质结构域类型及其频率是探索蛋白质结构域与必需蛋白质之间关系的两个重要元素。具有更多结构域类型的蛋白质更有可能是必需的，因为必需蛋白质更保守且具有独立的生物学特征。此外，如果蛋白质包含不经常出现在其他蛋白质中的结构域，则这些蛋白质往往是必不可少的，因为这些蛋白质的功能无法轻易取代。收集的54篇研究文章中有7篇文章使用DI分别将必需基因与非必需基因区分开，或与其他特征结合使用。通过使用自行设计的公式计算得出的分数排序，以及使用机器学习方法来使用此功能时都可以识别必需的基因。

网络拓扑结构

与EC的功能类似，PPI网络中的拓扑在通过计算方法识别必需蛋白质中很流行。张等。在他们的综述文章中，根据网络拓扑特征和机器学习方法，推广了许多有关预测必需基因的著作。描述PPI网络拓扑的常用功能是度中心性，中间性中心，紧密性中心性和聚类系数。此外，一些研究人员还使用图论提取了其他特征。2001年，Jeong等人。通过皮尔森线性相关系数，探索并阐明了基因致死率与中心度之间的正相关关系。此外，程度较高的必需基因往往是网络中心，并具有更多的邻居。因此，大多数研究在预测致死基因或蛋白质时考虑了这些特征，并指出PPI网络的拓扑结构在开发基本的基因预测模型中具有重要作用。通过网络拓扑的特征，总共发表了32篇预测必需基因的著作，其中54篇论文占绝大多数。除了常用的PPI网络之外，还利用了其他几种生物网络，例如代谢网络，基因共表达网络，转录网络及其整合网络，来挖掘与必需基因或蛋白质相关的有效特征。简而言之，这是研究基因本质的一种重要且经常使用的特征。

序列成分

SC通常是指可以表征为一组不同的寡核苷酸或氨基酸组成形式（例如K-mer）的DNA或蛋白质序列。这种特征提取方法可以将短或长氨基酸序列或碱基序列转化为目标序列中Kmer的出现频率，或者可以将其表征为常规给定序列的表示，例如Z曲线方法。假设必需基因和非必需基因之间的重要区别是存在不同的序列组成它们之间。因此，这两种基因具有不同的进化速率和保守性，应执行不同的功能。随着高通量测序技术的发展，基因组序列数据提供了使用SC功能预测必需基因的见识。这是近年来新提出的特征提取方法。所有特征均直接从核苷酸序列或氨基酸序列中提取。尽管在我们收集的54篇文章中只有八篇作品报道了此功能，但几乎所有作品都独特地采用了这种功能来预测必需基因并获得了良好的结果。Sarangi等。首先将周的伪氨基酸组成和各种理化特性引入了必需基因的鉴定中。同样，我们的小组根据一级序列（核苷酸或蛋白质序列）分析了必需和非必需基因的SC。随后，称为Z曲线的线性方法使用了部分Song等人提供了基于Z曲线理论和其他基于序列的特征与偏最小二乘（PLS）相结合的最小二乘（ZUPLS）来预测细菌必需基因。最近，一种在线预测工具称为提出了人类必需基因（Pheg）仅使用核苷酸组成特征来预测人类必需基因，这些特征是基于λ间隔Z曲线法设计的。基于序列的方法对于已测序的基因组是通用且通用的，并且可以实现很好的预测。

表达水平

必需基因的EL表示的范围比非必需基因的EL窄，并且必需基因的表达水平更高。尽管表1中没有文章单独使用EL来鉴定必需基因，但20项研究将这种类型的特征与其他特征结合在一起进行了预测。几项研究指出，PPI网络中的相互作用蛋白更可能共表达，因为这些蛋白通常以同一途径存在。因此，这些方法大多数都结合了PPI网络和基因表达谱，在基因共表达网络中计算了拓扑。同样，由于必需基因的EL相对较高，其他方法也将平均表达值作为每个基因的特征值。

功能实施和评估

在上面的部分中，我们总结了五种常用的具有代表性的特征类型，可用于预测必需基因。在本综述中，我们不仅描述了这些概括并提供了相关文献，还使用必需基因的标准数据集评估了这些特征，这有助于理解和利用这些特征。

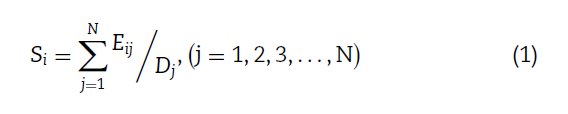
数据资源及评价方法

为了评估以上概述的每类独立特征的有效性，我们使用了三个代表性数据集作为预测必需基因的基准。革兰氏阴性细菌K-12 MG1655（E. coli）的基因必需性数据是从E.coli Chromosome（Prof.）的分析（https://shigen.nig.ac.jp/ecoli/pec/）数据库中收集的。该数据集包含287个必需基因和3857个非必需基因。从DEG数据库下载了革兰氏阳性枯草芽孢杆菌168（BAS）的基因必需性数据。它包含271个必需基因和3904个非必需基因。人类的基准数据集是从Guo等人的研究文章中获得的，该物种收集了1516个必需基因和10499个非必需基因。在已知基本基因信息的物种中，这三个物种的数据质量可能更可靠。为了构建预测模型，我们为每个基因分配了一个布尔值以标记其必要性。如果某个基因出现在基本基因数据集中，则将其标记为1；否则，标记为-1。所有基因的DNA和蛋白质序列均从美国国家生物技术信息中心（NCBI）的文件传输协议（ftp）网站下载。从Pfam数据库获得DI，然后通过隐马尔可夫模型算法搜索每个基因的域。从STRING数据库下载了PPI网络数据。基因表达数据可从E. coli和BAS的Gene Expression Omnibus（GEO）（对于E. coli的GSE56133和BAS的GSE45933）和人类的Cancer Genome Atlas（TCGA）下载。

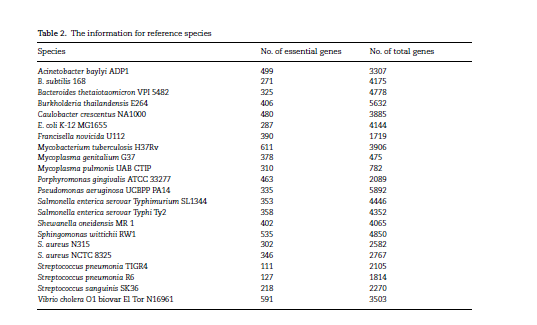
在我们收集的54篇论文中，SVM是预测必需基因最流行的机器学习方法。为了对每种类型的功能进行初步评估，选择了SVM来训练模型并在需要时测试功能集。接收器工作特性（ROC）曲线（AUC）下的面积用于评估预测。在这篇综述中，我们旨在从几个方面而不是讨论预测结果，通过计算方法对必需基因的预测提供全面的综述。因此，我们根据论文选择了合适的模型，表1中列出的评估每个特定特性。以下各节提供了有关各种功能的性能的一些简短说明。

EC的实施和评估

对于细菌种类，Cheng等。已经注意到，必要性注释在物种之间的可传递性受到训练数据集质量的显着影响。考虑到这一点和我们以前的工作，我们从DEG数据库中选择了21种（表2中的所有种，除了查询种E. coli或BAS外）作为参考种。这些基因的必需基因数据集和蛋白质序列分别从DEG数据库和NCBI下载[11]。我们评估了所选物种中目标物种与其最接近物种之间直系同源序列的数量。大肠杆菌和BAS的最接近物种是通过成分向量（CV）算法确定的[87]。使用蛋白质的基本局部比对搜索工具（BLASTp）程序，我们发现大肠杆菌及其最接近的物种小肠沙门氏菌血清鼠伤寒沙门氏菌SL1344（STS）共有253个直系同源必需基因和3204直系同源基因。另外，大肠杆菌的287个必需基因中的286个在STS中具有直系同源序列。同样，发现BAS与其最接近的金黄色葡萄球菌NCTC 8325（STN）之间有206个直系同源基因和1575个直系同源基因，而BAS的271个必需基因中有248个在STN中具有直系同源序列。这些数据表明两个紧密相关物种之间的基本同源性比例很高。此外，我们采用了基于正交和系统发育的基因本质预测工具算法（Geptop）来预测必需基因。

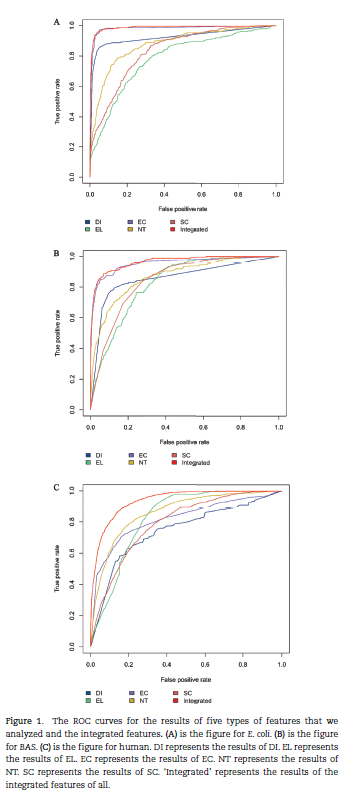


其中，Eij是第j个参考物种中查询基因i的最佳直系同源基因必不可少的标记；Dj是目标物种与第j个参考物种之间的进化距离；N是参考物种的总数，在本工作中等于21；Si是查询物种中基因i的必要性得分。将目标基因的预期必要性得分与观察到的必要性标签进行比较，对于大肠杆菌（图1A）和BAS（图1B）分别获得0.9829和0.9543的AUC值，这比获得的结果仅来自同源性映射。显然，基于基因持久性的特征比仅基于基因持久性的特征更适合于预测必需基因。对于诸如人类的高级真核生物，没有足够数量的具有经过验证的必需基因的真核参考物种。因此，我们从HCOP（http://www.genenames.org/cgi-bin/hcop）下载了人类与19个其他物种之间的直系同源物。如果某个基因在一个物种中具有直系同源物，则将标记为“1”；否则为“0”。最后，通过这种方法获得了19个特征。这19个特征与所有人类基因的本质相结合，被用作SVM的输入数据。通过10倍交叉验证获得的AUC值为0.8186（图1C）。虽然同源性作图的方法将受到保守基因数量和各种生物之间进化距离的限制，该方法在鉴定密切相关生物中的必需基因方面仍然相当有效和可靠[14，40]。值得注意的是，该方法具有可靠性和通用性。不可否认，如果有足够多的参考物种，同源性作图是探索必需基因的有效特征。然而，该方法需要使用具有已知基因必需性的基因组作为比对的参考。如果此信息不可用，同源性映射功能的性能将降低。



DI的实施与评估

Deng等人率先利用该域特征，并率先在远距离相关的生物体中进行预测，他报告说，DI在其完善的13个特征中对基本基因的预测贡献最大。在这从蛋白质家族比对（P-famA）条目的数据库中获取每种蛋白质的DI。排除没有结构域注释的蛋白质。因此，仅保留了分别针对大肠杆菌，BAS和人的3637、2623和9832基因来评估这种类型的特征。我们使用了Lu等人提出的基本域预测（EDP）模型。确定必要的领域。为了客观地评估此特征，选择了具有已知目标物种基因重要性的最接近物种作为参考物种，以使用EDP模型计算域必要性得分。对于目标物种中的蛋白质，其蛋白质必需性得分是通过最大结构域得分（该结构域属于该蛋白质）来衡量的。因此，STS和STN分别用作计算大肠杆菌和BAS的域必要性得分的参考物种。对于人类，由于没有其他具有已知基因必需性信息的高级真核生物，因此选择小鼠作为参考物种。小鼠的基因必需性可从OGEE数据库下载[90]。通过比较AUC值为0.9249、0.8592和0.7442大肠杆菌，BAS和人的基准数据集中的必要性得分和必要性标签（图1）。尽管这些结果仅针对具有域注释的蛋白质获得，但对于前两个目标物种而言，此功能的性能仍然很高。这些结果是使用DI鉴定必需基因的最佳方法。即使这样，仍然需要完整的研究物种DI才能获得更好的预测。在人类中表现不佳的原因可能是参考（鼠标）和查询物种（人类）之间的进化距离相对较远。这表明如果我们想要获得出色的性能，参考物种应该与目标物种紧密相关。此外，我们发现大肠杆菌中287种必需蛋白中的208种与非必需蛋白没有任何共同结构域。这种现象表明某些域与基因必需性有关，并且对于几种蛋白质执行特定而重要的功能而言是唯一的。因此，应该注意的是，如果在基因组中执行n倍交叉验证的方法，则域特征的性能将很差。此功能仅在进行跨物种验证时才有效且有意义，因为基因组中存在多个域，并且某些域是唯一的，尤其是对于必需基因而言。另外，应该知道参考物种中的基因必需性。因此，只有通过进行全基因组统计，我们才能获得域分布的总体概况。

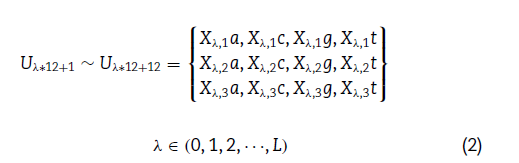


NT的实施与评估

在这一部分中，我们从STRING数据库中下载了PPI数据。两种蛋白质之间相互作用评分的临界值设置为默认值0.4。对于此网络中缺少的某些蛋白质，我们将为其设置每个拓扑特征的类别对应平均值。选择以下四种常用且易于获得的网络拓扑结构：度中心性，中间性中心性，紧密性中心性和聚类系数，以训练和验证预测模型。这些特征与所有目标物种基因的必要性相结合，被用作SVM的输入数据。通过对大肠杆菌，BAS和人的10倍交叉验证分别获得了0.8780、0.8700和0.8619的AUC值（图1）。大肠杆菌的必需基因的精确度为0.7875，非必需基因的精确度为0.8792。Silva等。以前，利用大肠杆菌的NT特性，必需基因的平均精度值对于必需基因为0.811，对于非必需基因为0.818。尽管在使用网络拓扑功能时这些结果并不是最好的，但它仍说明网络拓扑确实与基因本质相关。在我们收集的与网络拓扑特征相关的32篇论文中，大约有一半报告了它们与其他类型的特征结合使用的情况。PPI的信息应尽可能完整。

SC的实施和评估

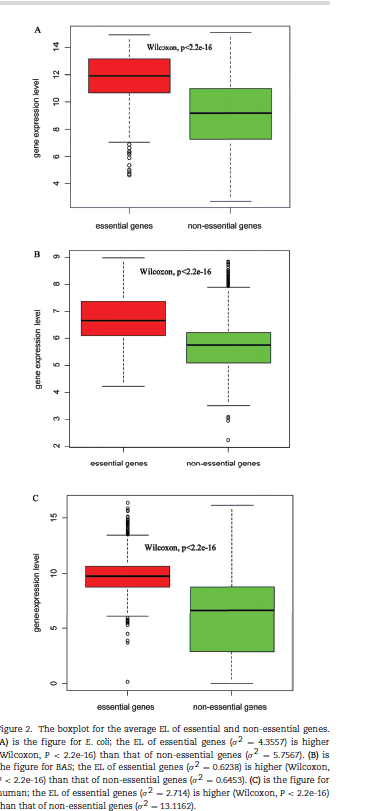
表1中列出的四篇文章使用了不同的方法来计算SC特征（请参见“序列组件”部分）。在本文中，我们使用了基于阶段性寡核苷酸（2）构造的另一个公式来描述基因序列。



其中，Uλ\*12+3表示第（λ\*12+i）个特征；Xλ,j代表k核苷酸，其第一个碱基位于密码子的第j（j=1，2，3）相位，λ是X与分离的a，c，g或t之间的内部长度。对于固定的k值，X具有4k的组合形式。n代表虚拟寡核苷酸Xλ,jn中的最后一个碱基（n=a，c，g，t）。我们使用fXλ,jn表示DNA序列中寡核苷酸Xλ,jn的频率。特别地，该公式描述了当k=0和λ=0时三个相位的四个碱基的频率。因此，对于每个k（k=0、1、2、3……），我们可以获得4k+1\*3\*（λ+1）个特征。为了考虑更多可能的核苷酸组成，我们设置参数k=1、2，…，7和λ=0、1，…，6，以提取特征并将所有这些特征与E的本质结合起来。大肠杆菌基因作为SVM的输入数据。通过对大肠杆菌、BAS和人的10倍交叉验证分别获得了0.8314、0.8502和0.8084的AUC值。先前利用序列特征预测大肠杆菌的必需基因，获得的AUC值为0.862。值得注意的是，我们在这一部分中没有执行特征选择，因为我们只想对基因本质中的SC特征进行基本评估。如果执行特征选择或从更长的寡核苷酸衍生出更多特征，则可以获得更好的预测。尽管我们没有使用其他论文中提出的特征，但是预测结果与它们的结果相差不远。序列组成的一个明显优点是获得特征非常方便。这些特征可以直接源自DNA序列。然而，一个不足之处是由于功能众多，因此运行时间较长。

EL的实施和评估

尽管没有工作仅专注于使用EL来研究必需基因，但是在一些研究中，此功能已集成到其他功能中。在这一部分中，我们尝试探索表达特征对基因本质预测的性能。对于从GEO数据库和TCGA数据库获得的表达数据，我们计算了大肠杆菌，BAS和人的每个基因的平均EL作为特征值。将特征值与每个物种的所有基因的必需性标签组合在一起，以计算AUC值。最后，它们产生的AUC值对于大肠杆菌为0.7839，对于BAS为0.8157，对于人为0.8272（图1）。虽然这些结果不是特别引人注目，它在一定程度上反映了必需基因和非必需基因之间的差异。通常将此功能与其他类型的功能结合使用。此外，我们尝试从另一个角度分析基因EL与基因本质之间的关系。我们绘制了箱形图（图2）以反映必需基因和非必需基因之间基因表达的分布。尽管必需基因或非必需基因有一些异常值，但很明显必需基因的EL高于非必需基因的EL。此外，当此功能与其他功能结合进行预测时，可以忽略异常值的影响。但是，此功能通常与PPI网络一起集成到共表达网络中使用。必须注意的是，所选表达谱数据应在丰富培养基而不是条件生长环境下进行测试。此外，由于性能相对较差，应将其与其他功能结合使用。



实施和评估所有功能

除了上面分析的功能外，我们还将所有这些功能集成在一起，可以对三个基准物种进行预测。域特征中的缺失值设置为零，因为未发现这些蛋白质的域注释。将网络拓扑或EL中的缺失值设置为每个特征中每个基因类别的平均值。最后，通过对大肠杆菌、BAS和人的SVM进行10倍交叉验证，得出的AUC值为0.9874、0.9655和0.9375（图1）。甚至对于大肠杆菌和BAS，所有的预测都得到了改善，使用单一类型的特征已经获得了相当令人满意的结果。此外，集成功能的优势在人类中得到了最大体现。这些结果表明，与仅基于单一特征类型的方法相比，该集成方法可以大大改善预测结果。将来可广泛应用于必需基因的预测。

整合特征以识别必需基因的建模方法

整合特征以识别必需基因的建模方法可以归类为经验公式和机器学习方法。当对参考种类或特征分布有一些先验知识并且特征是同质的时，总是使用经验公式。例如，Geptop利用进化距离来加权针对各种参考物种中必需基因的同源性作图的每个特征。UDoNC是一种利用蛋白质-蛋白质相互作用中的域特征及其拓扑特性来识别必需基因的方法。这些经验公式可有效地对基因本质进行分类。另外，机器学习方法（可以快速，自动处理大型和复杂数据集的算法）也被应用于预测必需基因的领域。这些方法通常可以分为三种类型：监督学习算法，非监督学习算法和半监督学习算法。有监督的学习方法（例如SVM，决策树，朴素贝叶斯和逻辑回归）通常用于识别必需基因。这些方法的输入数据由所需的输出值和一系列属性向量组成。与经验公式相比，机器学习方法可以将更多类型的特征集成在一起，以开发预测分类器并使用特征选择过滤无用的特征。但是，即使对于具有相同特征的同一生物，使用不同的机器学习方法的预测也可能不同。因此，一些研究人员选择了可以产生最佳结果的方法。例如，劳埃德（Lloyd）等人。选择了最佳模型和随机森林方法的输出。其他组合了几个独立分类器的输出。邓等选择使用非加权平均方法来整合四个不同分类器的预测。值得注意的是，机器学习方法的性能会受到特征选择或各种类型的特征组合的极大影响。需要持续优化参数和训练功能以提高分类器的性能。

必需基因预测的在线工具

随着计算方法的发展，已经为用户提供了几种在线必需基因预测服务器。Geptop是用于测序细菌基因组的基因必要性预测工具，是基于正畸和系统发育设计的。矫形器是根据最佳匹配来估计的。系统发育由查询物种和参考物种之间的进化距离表示，这是通过CV算法计算得出的。研究人员可以上载Fasta格式的DNA或蛋白质序列，然后获得预测结果，包括预测的基因或蛋白质的必需性以及必需性概率。用户可以在线使用此工具，也可以从http://cefg.uestc.edu.cn/geptop/免费获得其独立版本。Pheg是基于λ区间Z曲线理论而开发的，可用于预测人类的基因必需性。它仅使用核苷酸组成和关联信息作为机器学习方法的训练数据。Pheg可以预测人类匿名基因序列的必需性。用户可以在输入框中以fasta格式输入DNA序列。可以从http://cefg.uestc.edu.cn/Pheg免费访问此Web服务器。

结论

对于不同的特性，同源性映射的性能取决于已知的必要性参考物种及其之间的进化距离。在两个相对密切相关的物种（例如大肠杆菌和STS）之间存在更多的直系同源基因。相反，在两个遥远相关的物种（如大肠杆菌和肺炎支原体）UAB CTIP之间，直系同源的必需基因较少。选择诸如Geptop之类的同源性作图方法来预测原核生物中的必需基因是一个很好的选择，因为有足够数量的已知必需基因的参考物种，并且由此得出的预测结果也很好。如果DI是从近缘物种获得的，那么DI将带来更好的结果。选择大量具有已知必需基因的物种作为计算域必需性得分的参考是探索的新思路。但是，某些蛋白质中缺少域注释是一个应解决的问题。该特征不适用于真核生物，但是由于没有合适的物种可以用作参考物种的事实。网络拓扑反映了必需基因在生物网络中的关键作用。我们收集的大多数文章都使用了此功能。它们的性能在原核生物和真核生物中都是有效的。NT的表征也可以进一步发展。高质量和充足的网络数据有助于获得良好的结果。EL是一种不经常单独使用的功能。它的性能不如其他，并且EL数据中存在一定数量的缺失值。集成多个网络和基因表达数据是提高预测性能的常用方法。对于SC，最大的优点是从序列中提取特征很方便，并且不会丢失任何值。相反，其缺点是过多的功能会导致运行效率低下。特征选择是删除冗余信息并改善SC特征的预测结果的好选择。这种类型的功能适用于真核生物和原核生物。此外，如果近缘物种具有基因必需性数据，则同源性作图，DI和SC更适合跨生物体鉴定必需基因。

使用通常有效的特征来探索远缘物种的预测模型仍然是一个挑战。发现新功能可以创造性地改善预测。此外，在接下来的研究中，我们应该采用更有效的形式来改进先前发现的特征，选择更多重要的特征，消除特征之间的多重共线性并优化训练模型。将这些功能集成在一起并在该领域引入深度学习方法可能会导致意外的意外。另一个重要而有趣的问题是研究跨三个域的基本基因的计算可预测性，尤其是在属于不同域但基因组G+C含量相似的生物之间。