第 4 卷◆第 1 期◆版本 1.0◆2026 年
文章类型：论文|刊号（ISSN）：2972-4236(P) / 2972-4244(O)

# 基于计算机算法的生物序列分析研究

刘珂伶
西华大学
DOI:10.32629/acair.v4i1.19357

[摘　要] 生物序列承载着生命体遗传信息,其分析是生物信息学的核心课题。本文以计算机算法为支撑,系统探讨生物序列分析的基础理论、核心算法及应用拓展。首先阐述生物序列类型、特征与算法分类,明确性能评价指标；随后聚焦序列比对、特征提取与预测三大核心方向,剖析各类算法的原理、优势及优化路径；最后针对海量数据处理、智能算法融合等实际需求,提出算法改进策略。研究旨在为生物序列分析提供算法层面的理论参考,助力基因与蛋白质功能研究、疾病诊断等领域的技术突破,推动生物信息学向精准化、高效化发展。
[关键词] 生物序列；计算机算法；序列比对；特征提取；智能融合
**中图分类号**：F224−39　**文献标识码**：A

## Research on Biosequence Analysis Based on Computer Algorithm

Keling Liu
Xihua University, Chengdu, Sichuan

[Abstract] Biological sequences carry genetic information of living organisms, and their analysis is a core subject in bioinformatics. Supported by computer algorithms, this paper systematically explores the fundamental theories, core algorithms, and application extensions of biological sequence analysis. First, it elucidates the types, characteristics, and algorithmic classifications of biological sequences, clarifying performance evaluation metrics. Subsequently, it focuses on three major core directions—sequence alignment, feature extraction, and prediction—analyzing the principles, advantages, and optimization paths of various algorithms. Finally, it proposes algorithm improvement strategies to address practical needs such as massive data processing and intelligent algorithm integration. The study aims to provide theoretical references at the algorithmic level for biological sequence analysis, facilitating technological breakthroughs in areas such as gene and protein function research and disease diagnosis, and promoting the development of bioinformatics toward precision and efficiency.
[Key words] biological sequence; computer algorithm; sequence alignment; feature extraction; intelligent fusion

## 引言

基因测序技术不断发展,生物序列数据急速增长,传统分析方法无法适应精准高效的研究需求。计算机算法的参与给生物序列分析带来新方案,成了生物数据和生命规律之间的纽带。生物序列分析包含比对、特征获取、功能预测等诸多方面,其成果直接支持基因编辑、药物研发这些前沿领域取得发展[1]。文章站在算法角度,由基本理论过渡到实际应用逐步展开探讨,整理当前算法的关键逻辑及其不足,探索优化与融合路径,从而为应对生物序列分析中海量数据处理、算法稳定等难点供应参考,推动跨学科技术在生命科学中的深度应用。

## 1 生物序列与算法基础

1.1生物序列基本类型与特征。生物序列包含核酸序列和蛋白质序列两类,是遗传信息的主要承载者。核酸序列又分为DNA序列和RNA序列。DNA序列含有腺嘌呤、鸟嘌呤、胞嘧啶以及胸腺嘧啶这四种碱基,其呈双螺旋状,其中蕴含着生物体所有的遗传信息。RNA序列大多为单链形式,它在基因表达过程中起到传递遗传信息、催化化学反应等作用。蛋白质序列由20种不同氨基酸按照一定顺序串联起来形成,该序列的结构决定了蛋白质的空间形态及其生物学功能。生物序列具备特异性、保守性以及多变性的特点,特异性保证了各个物种的独特性,保守性体现出物种之间存在的进化联系,而多变性则是推动生物进化的关键因素[2]。这些特性给算法的设计提供了重要依据,也是生物序列分析得以开展的基本前提所在。

1.2生物序列分析核心算法分类。生物序列分析的核心算法

可依功能划分成四类,各类算法会塑造独特的逻辑模型以应对不同的分析需求。序列比对算法属于基本核心类型,它用来识别序列之间的相同点和不同点,并为进化分析以及功能推断提供支持。特征获取算法关注序列所包含的重要信息,并通过数学建模把序列数据转换成可被量化的特征。预测算法依照所得特征去推测基因结构、蛋白质功能之类的未明信息[3]。数据处理算法专门针对大量序列数据执行清理、去噪操作,并实现高效保存的目的。按照算法原理可区分为传统算法和智能算法,传统算法逻辑缜密,具有较强的可解释性,而智能算法有着更强的非线性拟合能力和自适应能力,两者相互配合来共同支撑生物序列分析。

1.3算法性能评价指标。算法性能评价要形成全面的指标体系,兼顾准确性、效率和稳定性这三大核心需求。其中,准确性指标包含正确率、召回率和F1值,它们用来度量算法输出结果与真实值的符合程度,是序列比对、功能预测算法的关键评价标准。效率指标覆盖时间复杂度和空间复杂度,直接体现算法处理大量数据的能力,非常适合大规模生物序列分析的场合。而稳定性指标通过对不同数据集执行重复检测来达成,用以考量算法在数据变动时输出是否一致[4]。可解释性属于重要的附加指标,其会影响算法在生命科学研究中的合适性。合理选择评价指标,可以客观显示算法的优劣之处,为算法优化与应用提供科学依据。

## 2 生物序列比对算法研究

2.1全局序列比对算法。全局序列比对算法期望针对两条或者更多条序列执行全长比对,找到整体相似度最高的一种合适的匹配形式,比较适合那些长度差不多、进化关系密切的序列分析。经典算法有Needleman-Wunsch算法,其依靠动态规划的原理,通过形成得分矩阵,确立好匹配、错配以及插入删除的惩罚规则,去考察所有的可能比对途径,并从中挑选出最佳答案。这个算法逻辑很严谨,得出的结果也比较可信,不过它的时空复杂度比较高,很难用在长序列的分析当中[5]。后来出现的一些优化算法,它们简化了得分矩阵的计算步骤,采用了一些启发式的规则,在保留正确性的情况下提升了效率。全局比对算法被广泛应用于创建物种进化树、识别同源基因等方面,是生物序列比对的基础算法之一。

2.2局部序列比对算法。局部序列比对算法重点找出序列里相似度较高的局部片段,并非全部序列都要实施比对,这很适合包含局部保守区域、长度差别较大的序列分析。Smith-Waterman算法属于局部比对的关键算法,它依据动态规划优化了得分规则,如果比对得分低于零就停止计算,只保存局部理想片段,这样就能极大简化计算复杂度。此算法可有效地识别出序列中的功能结构域、保守位点,在基因片段分析、蛋白质功能位点预测方面表现出色[6]。当前的优化方向主要集中在并行计算和硬件加强上,以进一步加强算法应对大规模数据的能力。局部比对算法弥补了全局比对存在的局限,是生物序列分析中最广泛使用的算法之一,其凭借自身特点在诸多场景下发挥着重要作用。

2.3多序列比对算法优化方向。多序列比对算法用于同时比对三条及以上序列,挖掘序列间的共性保守区域与特异性差异,是进化分析、蛋白质家族分类的核心工具。传统算法多基于渐进比对策略,通过先比对相似性高的序列,逐步加入差异序列构建比对结果,但易受初始比对结果影响,存在局部最优问题[7]。当前优化方向主要集中于三点：一方面,采用智能优化算法,如遗传算法、粒子群算法等,打破局部最优局限；另一方面,融入序列结构信息,结合蛋白质空间构象改善比对精度；此外,优化并行计算框架,满足海量多序列比对需求。优化之后的算法既保留效率,又显著提升了比对结果的可靠性和稳定性。

## 3 生物序列特征提取与预测算法

3.1序列特征提取算法设计。生物序列特征获取是关联原始序列数据和预测模型的重要部分,重点在于把杂乱的碱基或者氨基酸序列变成可量化且具备生物学意义的特征向量。算法设计要兼顾序列的局部特征和全局特征,局部特征获取着重于单个碱基、氨基酸的构成及其邻近关系,比如k-mer频率法,通过统计不同长度片段的出现频率来形成特征；而全局特征获取则着眼于序列的整体结构,像理化性质编码法,就是把氨基酸的疏水性、亲水性等理化特性转换成数值特征[8]。特征获取算法要协调好维度与有效性,防止陷入维度灾难,规避特征重复。当前的算法大多凭借特征筛选和融合手段,优化后续预测模型的准确度与效率,从而为生物序列的深入剖析形成基础。

3.2基因结构预测算法。基因结构预测算法旨在从基因组序列中识别基因的编码区、非编码区、启动子等功能区域,明确基因的边界与结构,是基因功能研究的前提。算法主要分为两类：依靠序列相似性的预测算法会对比已知基因序列来识别同源区域,并据此推断基因结构,这种算法可信度较高,不过它依赖的是已有的数据；而像隐马尔可夫模型这样的依托统计模型的预测算法,则通过学习基因序列的统计特性来形成概率模型以做到预测,适合用于分析未知的基因序列[9]。此类算法碰到的主要难点在于区分编码区与非编码区之间那条较为模糊的界限,当下着重于整合多种来源的数据,从而优化预测的准确性,进而给基因克隆、基因编辑等相关研究给予技术上的支持。

3.3蛋白质结构与功能预测算法。蛋白质结构与功能预测属于生物序列分析的关键目标之列,算法依照蛋白质序列信息来推断其空间构象及生物学功能。结构预测算法存在三种类型,即同源建模、折叠识别以及从头预测。同源建模凭借已知结构的同源蛋白质,具有较高的效率和准确性,也最为普遍；从头预测无需同源序列,而是遵照物理化学原理去模拟蛋白质折叠过程,适合那些没有同源结构的蛋白质[10]。功能预测算法大多依靠序列相似性和结构特征,通过对比已具备功能的蛋白质序列,来推断目标蛋白质的功能。当前的算法正在渐渐融入深度学习技术,从而改善结构预测的精准度并加强功能注释的全面性,有益于药物靶点筛选、疾病机制探究等方向。

## 4 算法应用拓展与改进

4.1海量生物序列数据处理算法。海量生物序列数据正在爆发式增长,这给数据处理算法的高效性与扩展性带来了严格要求。传统算法应对大规模数据时,往往会陷入计算效率低、内存

损耗大等情况。当下核心解决办法包含数据压缩算法和并行计算算法。数据压缩算法通过剔除多余信息，优化编码来削减数据存储和传送的成本，而且保留关键信息不遗失；并行计算算法依靠多核CPU、GPU等硬件结构，把大规模数据分解成子任务并行处理，极大地改善了计算效率[11]。采用分布式计算框架之后，达成了海量序列数据的分布式存储与协作处理，给基因组学、转录组学等大规模研究赋予了高效的数据处理支持。

4.2智能算法在序列分析中的融合应用。智能算法与传统生物序列分析算法相融合，这已成为提高分析精度和效率的关键趋向。深度学习、加强学习这些智能算法有着很强的非线性拟合能力和特征学习能力，可以有效地找出生物序列里潜藏的规律。卷积神经网络被用来做序列特征的获取的时候，能够自己识别出复杂的序列模式；循环神经网络合适用来处理序列的时序依赖关系，从而改善预测模型的性能[12]。这种融合不是简单的叠加，而是通过优化算法框架，达成传统算法逻辑的严谨性和智能算法自适应能力相互补益的效果。这种融合形式在序列比对、蛋白质结构预测等情形下已经得到了证实，明显改善了分析成果，扩展了生物序列分析的应用范围。

4.3算法稳定性与效率改进策略。算法的稳定性与效率会直接影响它在实际场景中的合适程度。对于当前算法存在的不足之处，要从大量角度制订优化方案。就稳定性而言，可以利用正则化技术，并改善参数设定，从而减小算法对数据噪音和异常值的敏感度，增强其在不同数据集里的输出一致性；通过多场景的验证以及鲁棒性评估，来优化算法的适配性能。至于效率方面的优化，除开依靠硬件加强和并行计算之外，还可以借助算法的精简，采用启发式规则，以此来削减时间和空间上的复杂程度。而且，把算法结合在一起并执行模块化设计，就能够达成各种算法之间长处相互弥补的目的，既守住稳定性又加强效率。优化之后的算法会更符合复杂生物序列分析的需求，促使这项技术真正得以投入应用[13]。

## 5 结语

本文就依靠计算机算法的生物序列分析执行系统研究，理清生物序列和算法的基本情况，详细分析序列比对、特征获取与预测这些核心算法，而且探究算法的应用拓展和优化途径。研究表明，优化传统算法并融入智能算法，对于改善生物序列分析的准确率、效率和稳定性十分关键[14]。当下算法还存在一定局限性，比如缺乏可解释性等情况。未来要加大跨学科融合力度，把生命科学和计算机技术的新动态应用到算法设计当中去。本文的研究成果能够给生物序列分析提供理论层面的支持，有益于相关领域的技术得以实现，促使生物信息学朝着优质的方向向前迈进。

## [参考文献]

[1]穆学琛.高维非结构化生物序列数据的深度学习表征及降维算法[D].吉林大学,2025.

[2]王艳,冀松,刘静.基于KMP算法的生物序列模式自动识别应用研究[J].电脑知识与技术,2024,20(36):42-44.

[3]李飞.面向生物功能序列识别的多模型整合算法研究[D].吉林大学,2024.

[4]郭育洲,周小安,林洋.基于神经网络的生物序列分类探析[J].数字技术与应用,2024,42(11):152-156.

[5]卫泽刚,陈旭,张小丹,等.基于Edlib的启发式生物序列聚类算法[J].宝鸡文理学院学报(自然科学版),2024,44(03):50-55.

[6]Jumper J,Evans R,Pritzel A,et al.Highly accurate protein structure prediction with AlphaFold[J].Nature,2021,596(7873):583-589.

[7]Baek M,DiMaio F,Anishchenko I,et al.Accurate prediction of protein structures and interactions using a three-track neural network[J].Science,2021,373(6557):871-876.

[8]Lin Z,Akin H, Rao R,et al.Evolutionary-scale prediction of atomic-level protein structure with a language model[J].Science,2023,379(6637):1123-1130.

[9]Rives A, Meier J, Sercu T, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences[J]. Proceedings of the National Academy of Sciences,2021,118(15):e2016239118.

[10] Elnaggar A, Heinzinger M, Dallago C, et al. ProtTrans: Toward Cracking the Language of Life’s Code Through Self-Supervised Deep Learning and High Performance Computing[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2022,44(10):7112-7127.

[11]Ji Y, Zhou Z, Liu H, Davuluri R V. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome[J].Bioinformatics,2021,37(15):2112-2120.

[12]Dalla-Torre H,Gonzalez L,et al.The Nucleotide Transformer: Building and Evaluating Foundation Models for DNA[J]. Nature Methods,2025,22:287-297.

[13]Gligorijević V,Renfrew P D,Kosciolek T,et al. Structure-based protein function prediction using graph convolutional networks[J].Nature Communications,2021,12:3168.

[14]Mirdita M, Schütze K, Moriwaki Y,et al.ColabFold: making protein folding accessible to all[J]. Nature Methods, 2022, 19(6):679-682.

## 作者简介：

刘珂伶(1995--),女,汉族,四川广安人,硕士研究生,助教,人工智能,生物信息。