**非典型 DNA 位点G4探测方法综述**

Review：Detection methods for Non-canonical DNA structures G4

生物信息 张佳胜 指导老师 张勇 刘晓兰 侯爽

摘要部分草稿：

四链DNA结构，又被称为G4，通常是由一段鸟嘌呤富集的序列通过特殊的拓扑结构形成。计算研究发现，基因组上的G4 motif位点超过140w个。进一步的实验验证，在体内存在大量G4结构，并且它们富集在一些关键性的区域，承担着极其重要的功能。为了探究G4结构的位点以及功能相关性，目前迫切地需要鉴定基因组上存在的G4结构，找出可能的G4结构位点。即使G4具有一定的motif结构，但是它的序列组成却很多样化，这使得纯计算机式的序列检索无法完全准确地鉴定出基因组上的G4结构。随着研究的进展，逐步发展出了以二代测序技术为核心的G4 ChIP-seq以及G4-seq技术，能够高通量的识别基因组层面的G4位点，而结合二代测序技术和计算机预测方法的策略也成为潜在G4结构预测的可靠手段。

本课题的核心是G4结构探测方法，兼顾G4结构的位点分布特征和潜在功能性研究。我将从以下几个方面展开：G4的基本特性调研；已知算法的测试；核小体位点复合算法的构建；比较和验证新模型的优劣性；G4结构位点的分布特点与G4结构的功能的推断。

正文部分草稿：

1. G4的结构和特点

绝大部分的基因组上的DNA都表现为右手双螺旋结构，我们通常称之为典型性DNA(B-form DNA, B-DNA)。但是实际上，DNA的二级结构复杂多样，也存在着部分特殊的非典型DNA结构。非典型的DNA结构包括了Z-DNA、十字DNA、三链体DNA以及四链体DNA等。再过去的二十年中，越来越多的研究证据表明，非典型DNA在体内与双链DNA结构存在着竞争的关系并且承担着一部分重要的生物学功能。正是因为这些特定的性质，使得它成为基因组结构研究的一个重要方向，受到越来越多的关注。

G4(G-quadruplexes)，又被称为鸟嘌呤四链体，是非典型的DNA结构之一。它是一种单链鸟嘌呤富集的DNA序列，能够在特定情况下折叠成四链DNA结构。这种结构产生于G四平方面的自我堆叠，这些G四平方面通常由四个鸟嘌呤通过环Hoogsten氢键连接，排列而成。在这些G四平方面的中心，存在一个一价阳离子（通常是K+或者Na+）来稳定结构，而这些四平方面之间，由受挤压的单链loop序列连接。G4s存在多种空间拓扑结构，这取决于G4s堆叠G四平方面的数量、中间相互连接loop的长度、loop的序列组成、折叠时核酸链的方向以及中心通道阳离子的性质。这些特殊的空间拓扑结构使得G4s的热稳定性通常较高，进而阻碍DNA双链的杂交以及其他分子的交流。

G4结构存在于多种生物体内，其中人类基因组上已发现有超过716,310条G4s。G4s在基因组上的位点并不随机，而且通常在近缘物种之间高度保守。这种G富集的单链序列具备一种相同的序列模式，我们称之为典型G4s，典型的G4s 基序如右所示：G≥3N1-7G≥3N1-7G≥3N1-7G≥3。而部分G4序列它们的loop长度可能大于7或者鸟嘌呤runs数目大于4，我们称之为非典型的G4序列。通过这些计算机检索这些G4 motif，科学家们在人类基因组上发现了超过140w个潜在的G4结构位点，这意味着G4的探测手段以及预测技术仍需要改进。

通过研究已知的G4s以及G4 motifs，科学家们发现，这种结构与部分关键性的基因区域有关，尤其是富集在TSS上游的启动子区域，以及DNA链的端粒等区域。这种基因组上的不平均性揭示了某种潜在的重要功能，这可能意味着G4结构与转录、转录因子、基因组稳定性，端粒作用以及癌症之间具备一些相关性。随着研究的进展，越来越多的直接或者间接证据表明，G4s与DNA的复制，复制起始，转录，基因表达，基因组稳定性，表观遗传等都具有一定的相关性。

由此，我们迫切地需要获得G4s在基因组上的确切分布和位点，方便更深入更准确地了解G4结构在体内的作用。本研究旨在结合测序技术和计算机预测的手段，来获取更准确的G4结构的位点，并由此分析G4结构可能的作用。

1. 核小体位点与G4的关联

二代测序技术介绍

核小体定位与G4结构位点的关联性

1. MNase-seq和ATAC-seq的比较
2. G4的motif预测、统计以及分类策略(G4 Hunter&pqsfinder)

全基因组的G4 motif预测

G4 motif 的分布特点

G4 motif的分类策略以及验证数据集

1. 构建核小体信号到G4的预测模型

鉴于之前的结论，我们从核小体定位与G4结构位点的关联性出发，构建一个结合核小体定位信号的预测模型。由于G4的是与否是一个离散型的标签，而核小体定位信号则是连续型的数据，经过比较，我们考虑采用逻辑回归来作为二者的关系映射函数。

构建核小体信号到G4标签的预测算法

1. G4的双重预测模型

构建双重预测模型

模型的训练与测试

模型的验证

1. 比较与验证双重模型的优劣性

比较G4 Hunter、pqsfinder以及双重模型的各项预测指标

1. \*利用模型进行G4 ChIP-seq中G4 peak的正负链标记以及纯化
2. \*G4与转录相关,G4 motif的分布

TSS附近的G4 ChIP-seq信号分布特征

TSS附近的G4 motif profile

1. 总结和讨论

技术方法：

MNase-seq: 采用限制性外切酶将染色质上不受保护的区域(染色质开放区域)切除，留下核小体上缠绕的DNA序列，然后针对这部分序列进行扩增和测序。

ATAC-seq: 通过Tn5转座酶随机插入染色质开放区域，将不受染色体保护的染色质区域(染色质开放区域)切割下来，然后针对这部分序列进行扩增和测序。通常用于定位染色质开放区域。

ChIP-seq: ChIP-seq技术，全名染色质免疫共沉淀技术。主要是针对细胞内特定结构或者特定序列的DNA，采用蛋白质结合捕获的手段，特异性地富集沉淀目的蛋白结合的DNA片段，然后对这些DNA片段进行纯化和文库构建，最后进行高通量测序。目前，G4的主要高通量探测手段就是G4 ChIP-seq，采用G4的选择性探针(配体或者抗体)去识别基因组上的G4 motif序列或者G4结构。

G4-seq: 一种结合了二代测序技术和聚合酶停滞模型的新型高通量测序技术。由于G4空间拓扑结构的特异性和稳定性，单链上的G4结构会导致聚合酶链式反应中的聚合酶停滞，进而影响到碱基的召唤，最终导致聚合酶链式反应出错。研究人员比较了G4影响导致聚合酶停滞的序列和没有G4影响下的正常序列的测序质量和碱基含量，发现在G4位点（聚合酶停滞）的碱基错配概率极高，测序质量极低。于是，研究人员通过体外给定G4稳定条件来产生一条G4结构阻碍聚合酶的高错配序列，与正常条件下完成聚合酶链式反应的低错配序列比较，找出这部分碱基错配的位点，由此推断出序列上特定的G4结构位点。G4-seq的一般流程如下：在Na+条件下对DNA序列进行准确的测序和比对，获得比对结果read1，然后在K+ & PDS（促进G4稳定）条件下重新测序原模板，获得比对结果read2，由于G4诱导聚合酶停滞改变了从G4结构开始的测序读数，导致read2的测序质量下降从而识别G4位点。

nucleoATAC: Tn5转座酶会插入染色质开放区域进行切割，当Tn5转座酶的两端并不恰好包含一个开放区域，而是包含了一个核小体的时候，可以认为切割下来的这个片段中能够定位到核小体。那么，我们通过筛选ATAC-seq文库中的序列长度，当序列长度大于146bp(核小体的序列长度)时，通过mapping后获取序列的覆盖信息，核小体位点处的序列会高度覆盖，从而获取了核小体定位。

正则表达匹配: 计算机的正则表达匹配是指一段能够指代某一种模式的不连续的字符串，常常用于共性模式序列的检索。在计算生物学上，我们根据实验观察以及经验，归纳出G4序列具备的一些共性——由长度至少为1nt的loop连接起来的四个鸟嘌呤runs，我们称之为G4 motif。这种motif的序列正则表达形式如右所示：G≥3N1-7G≥3N1-7G≥3N1-7G≥3。我们利用计算语言，设计出对应的正则表达式，然后针对全基因组的序列进行检索，找出符合特征的序列以及位点，它们就是潜在的G4结构位点。

滑动窗口&打分矩阵:

Pqsfinder:

G4Hunter:

机器学习: 传统的计算方法大多基于研究经验以及专家知识，往往依靠人力总结有限数量的已观测数据的共性，来归纳总体数据的特性。这无法完美地描绘G4多样性构想的完整图像，可能存在被遗漏的G4结构的信息。近二十年来，硬件技术的提升使得机器学习的方法逐渐热门，它们依靠数据本身驱动预测，通过大量的数据作为基础，来避免过多的人为假设以及人为失误。机器学习的方法通常采集一批有效的实验数据，其中包含了一定量的阳性样本和阴性样本，通过设计回馈模型，让计算模型从数据内部学习特征，然后使用这些特征去预测新的数据。为了使得预测的准确率有所保障，极其学习模型往往设计一个训练数据集和一个测试数据集，在学习特征的过程中，只有模型同时在训练数据集和测试数据集上都表现良好的时候，我们才认为这个学习器初步完成。

逻辑回归算法: 逻辑回归(logistic regression)，也有说法称为对数几率回归，是一种广义的线性回归分析模型。虽然被称为回归模型，但是逻辑回归一般用于处理分类的问题。这种算法的本质是将线性回归的结果映射到一个跃迁曲线上，在设定好阈值T后，这个曲线要求在x>T时y≈1，而x=T时y≈0.5，x>T时y≈0，在满足这些条件后它还应当是单调可微且任意阶可导的。一个简单的例子就是sigmoid函数，

这个函数具备较好的跃迁性，它在远离阈值T=0的地方函数值会很快地接近1或者0。之后只需要将z=*WTX*(此处T代表转置)的线性回归表达式带入，获得逻辑回归函数式。后续通过代价函数以及极大似然等策略拟合函数，就能够得出用于分类的逻辑回归模型。

数据来源：

引用文献：

[] Bochman ML, Paeschke K, Zakian VA. DNA secondary structures: stability and function of G-quadruplex structures [J]. Nat Rev Genet. 2012 Nov; 13(11):770-80.

[] Rhodes D, Lipps HJ. G-quadruplexes and their regulatory roles in biology [J]. Nucleic Acids Res. 2015 Oct 15; 43(18):8627-37.

[] Mukherjee AK, Sharma S, Chowdhury S. Non-duplex G-Quadruplex Structures Emerge as Mediators of Epigenetic Modifications [J]. Trends Genet. 2019 Feb; 35(2):129-144.

[] Hänsel-Hertsch R, Di Antonio M, Balasubramanian S. DNA G-quadruplexes in the human genome: detection, functions and therapeutic potential [J]. Nat Rev Mol Cell Biol. 2017 May; 18(5):279-284.

[] Chambers VS, Marsico G, Boutell JM, et al. High-throughput sequencing of DNA G-quadruplex structures in the human genome [J]. Nat Biotechnol. 2015 Aug; 33(8):877-81.

[] Kwok CK, Merrick CJ. G-Quadruplexes: Prediction, Characterization, and Biological Application [J]. Trends Biotechnol. 2017 Oct; 35(10):997-1013.

[] Hänsel-Hertsch R, Beraldi D, Lensing SV, et al. G-quadruplex structures mark human regulatory chromatin [J]. Nat Genet. 2016 Oct; 48(10):1267-72.

[] Puig Lombardi E, Londoño-Vallejo A. A guide to computational methods for G-quadruplex prediction [J]. Nucleic Acids Res. 2020 Jan 10; 48(1):1-15.

[] Brázda V, Kolomazník J, Lýsek J, et al. G4Hunter web application: a web server for G-quadruplex prediction [J]. Bioinformatics. 2019 Sep 15; 35(18):3493-3495.

[] Brazda V, Kolomaznik J, Mergny JL, et al. G4Killer web application: a tool to design G-quadruplex mutations [J]. Bioinformatics. 2020 May 1; 36(10):3246-3247.

[] Hon J, Martínek T, Zendulka J, et al. pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R [J]. Bioinformatics. 2017 Nov 1; 33(21):3373-3379.

[] Hu S, Chen X, Zhang Y, et al. CAM: A quality control pipeline for MNase-seq data. PLoS One. 2017 Aug 7; 12(8):e0182771.

[] Meyer CA, Liu XS. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. Nat Rev Genet. 2014 Nov;15(11):709-21.

[] Alicia N. Schep, 1 Jason D. Buenrostro, 1 Sarah K. Denny, et al. Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. Genome Res. 2015 Aug 27; 25: 1757-1770.

[] 周志华. 机器学习. 北京: 清华大学出版社, 2016: 57-59

[] Pedregosa, F. and Varoquaux, G. and Gramfort, et al. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[] Kim N. The Interplay between G-quadruplex and Transcription. Curr Med Chem. 2019;26(16):2898-2917.

[] Hänsel-Hertsch R, Simeone A, Shea A, et al. Landscape of G-quadruplex DNA structural regions in breast cancer. Nat Genet. 2020 Sep;52(9):878-883.

[] Bedrat A, Lacroix L, Mergny JL. Re-evaluation of G-quadruplex propensity with G4Hunter. Nucleic Acids Res. 2016 Feb 29;44(4):1746-59.

[] Marsico G, Chambers VS, Sahakyan AB, et al. Whole genome experimental maps of DNA G-quadruplexes in multiple species. Nucleic Acids Res. 2019 May 7;47(8):3862-3874.

[] Awadasseid A, Ma X, Wu Y, et al. G-quadruplex stabilization via small-molecules as a potential anti-cancer strategy. Biomed Pharmacother. 2021 Apr 5;139:111550.

[] Kosiol N, Juranek S, Brossart P, et al. G-quadruplexes: a promising target for cancer therapy. Mol Cancer. 2021 Feb 25;20(1):40.

[] Shahsavar K, Hosseini M, Shokri E, et al. New insight into G-quadruplexes; diagnosis application in cancer. Anal Biochem. 2021 May 1;620:114149.

[] Zheng KW, Zhang JY, He YD, et al. Detection of genomic G-quadruplexes in living cells using a small artificial protein. Nucleic Acids Res. 2020 Nov 18;48(20):11706-11720.

[] Weisz K, Jana J. Thermodynamic Stability of G-Quadruplexes: Impact of Sequence and Environment. Chembiochem. 2021 Apr 12.