QB 伯晓晨李昊陈河兵团队合作提出从三维染色质结构中高效检测拷贝数变异的机器学习方法

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/29882.html

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

QB
伯晓晨
李昊陈河兵团队合作提出从三维染色质结构中高效检测拷贝数变异的机器学习方法。论文标题：Effectiveness of machine learning at modeling the relationship between Hi-C data and copy number variation

期刊：Quantitative Biology

作者：Yuyang Wang, Yu Sun, Zeyu Liu, Bijia Chen, Hebing Chen, Chao Ren, Xuanwei Lin, Pengzhen Hu, Peiheng Jia, Xiang Xu, Kang Xu, Ximeng Liu, Hao Li, Xiaochen Bo

发表时间：06 July 2024

DOI：https://doi.org/10.1002/qub2.52

微信链接：点击此处阅读微信文章

拷贝数变异（CNV）指长度1kb以上的基因组大片段的拷贝数增加或者减少，这种变化可能会使得基因表达模式异常，导致癌症与遗传性疾病的发生。Hi-C能够捕获染色质水平的基因组互作信息，两个基因组位点之间的空间距离可以通过两个基因组位点之间近端连接产生的读数估计。疾病相关的结构变异与三维基因组结构紧密关联，肿瘤基因组中可通过拷贝数变异的方式改变增强子、启动子和其他转录调控元件之间的相互作用，进而影响区域内相关基因的表达。通过Hi-C数据识别CNV有助于我们理解这些变异如何影响参与基因及调控因子表达水平的基因组片段之间三维交互作用。然而，由于染色质三维拓扑结构的复杂性以及Hi-C数据的稀疏性，使得从Hi-C数据中精确计算CNV存在较大的挑战性。

近期，军事医学研究院伯晓晨，李昊和陈河兵课题组在Quantitative Biology期刊发表了一篇题目名Effectiveness of machine learning at modeling the relationship between Hi-C data and copy number variation的文章，提出了从Hi-C数据中高效预测拷贝数变异的机器学习算法，揭示了三维染色质拓扑结构与CNV之间的复杂关系。

Received: 12 September 2023 | Accepted: 6 March 2024
DOI: 10.1002/qub2.52

RESEARCH ARTICLE

Quantitative Biology
Open Access

Effectiveness of machine learning at modeling the relationship between Hi-C data and copy number variation

Yuyang Wang[1,2] | Yu Sun[1] | Zeyu Liu[3] | Bijia Chen[1] | Hebing Chen[1] | Chao Ren[1] | Xuanwei Lin[2] | Pengzhen Hu[4] | Peiheng Jia[5] | Xiang Xu[1] | Kang Xu[6] | Ximeng Liu[2] | Hao Li[1,3] | Xiaochen Bo[1]

全文概要

如图1所示，研究团队开发了用于精准预测Hi-C数据中的拷贝数变异机器学习算法。首先，他们选用了经典线性模型，为不同的染色质学习特定的回归权重系数来计算Hi-C读数与CNV之间的数值关系。通过使用多种线性与非线性降维算法统一不同染色体间的维度，从而探究线性模型在跨染色体上的预测性能。另外，研究团队使用图卷积神经网络（GCN）模型对三维染色质的拓扑结构进行建模，从而有效捕获染色质空间结构特征并预测拷贝数变异，以了解染色质空间相互作用背后的机制。同时，团队进行了一系列Hi-C数据扰动实验，以评估基于图神经网络模型捕获与CNV相关的染色体结构关键特征的能力。

爱科学
iikx.com

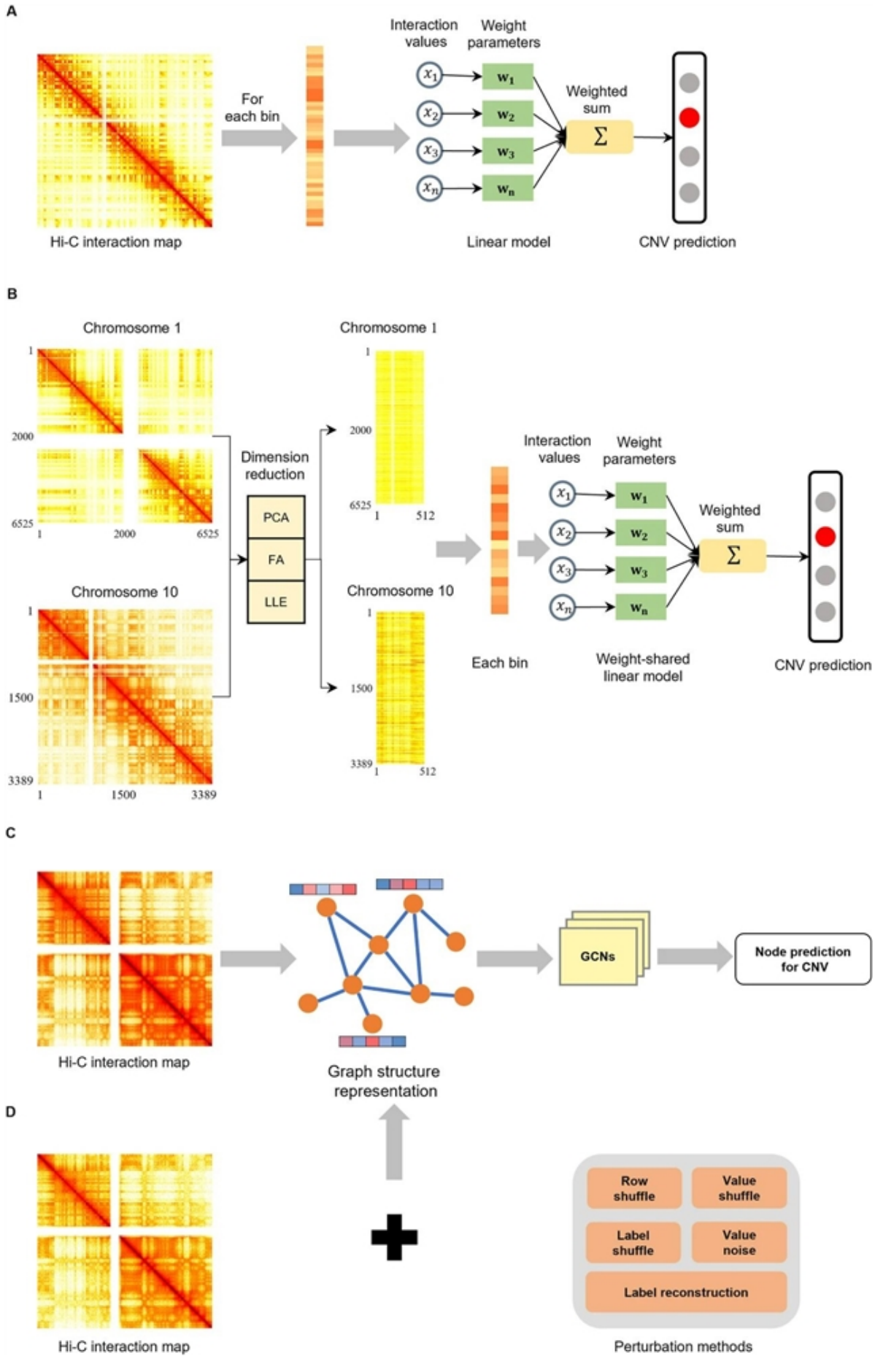

A
Interaction values
Weight parameters
For each bin
Weighted sum
Hi-C interaction map
Linear model
CNV prediction
B
Chromosome 1
Chromosome 1
Dimension reduction
PCA
FA
LLE
Chromosome 10
Chromosome 10
Interaction values
Weight parameters
Weighted sum
Each bin
Weight-shared linear model
CNV prediction
C
Hi-C interaction map
Graph structure representation
GCNs
Node prediction for CNV
D
Hi-C interaction map
Row shuffle
Value shuffle
Label shuffle
Value noise
Label reconstruction
Perturbation methods

图1 从Hi-C中预测拷贝数变异的机器学习方法研究框架

(A) 基于线性模型的CNV和Hi-C读数关系建模方法，并为不同染色体构建特定的回归系数。(B) 使用线性与非线性的降维方法统一不同染色体的尺寸，训练和评估权重共享的线性模型在不同染色体上的预测性能。(C) 以图结构的形式对Hi-C数据进行建模，并使用图卷积神经网络模型探究三维染色质拓扑结构与CNV间的关系。(D) 设计一系列扰动实验，用于评估图卷积神经网络在CNV预测任务中的鲁棒性。

线性模型在Hi-C数据中预测CNV的有效性

通过经典线性模型对K562、U226和RPMI8226三种细胞系的Hi-C数据进行建模，预测不同基因组片段中的CNV。结果表明，每条染色体的Hi-C读数一维信号和CNV之间表现出明显的线性模式：在K562、U226和RPMI8226三种细胞系上的预测准率达97.64%、98.80%和98.67% 。

线性模型在跨染色体上的CNV预测性能

为了研究不同染色体中的Hi-C读数一维信号和CNV之间是否存在相似的线性模式，使用一个权重参数共享的线性模型拟合细胞系中不同的染色体一维读数信号，并评估CNV预测准确性。由于线性模型的计算过程依赖于染色体长度，因此需要采用降维方法进行数据预处理，统一不同染色体的尺寸。通过主成分分析（PCA）降维处理后，权重参数共享的线性模型在三种细胞系中的CNV预测准确率分别为39.83%、68.94%和51.46%，准确率明显低于使用独立线性模型的预测效果（97.99%、95.20%和97.64%）。结果表明，每个染色体中均存在特定的Hi-C一维读数信号与CNV线性模式，难以泛化到不同的染色体中。

图卷积神经网络在CNV预测任务中的有效性

团队进一步探究了不同染色体空间结构特征与CNV之间的关系，将Hi-C数据使用图的方式进行建模，并基于图卷积神经网络模型从捕获的染色质空间结构特征推断CNV。首先，通过Node2Vec算法构造图的节点特征后，使用一个权重参数共享的GCN模型在三种细胞系的不同染色体中进行预测，预测准确率达97.18%、94.87%和95.81%，结果表明与CNV相关的染色质空间结构特征在不同染色体上具有高度相似的模式。

图卷积神经网络模型在不同细胞系中的迁移预测性能

基于前述研究结果，团队采用迁移学习策略评估不同细胞系中染色质空间结构特征与CNV之间的关联模式差异性。具体地，将GCN模型在RPMI8226细胞系中进行预训练后，仅使用U226细胞系中2%的CNV标签进行微调，模型在U226细胞系中的CNV预测率达93.21%，且仅需几个训练轮次便可以实现模型快速收敛。然而，如果在不使用少量CNV标签信息微调的情况下直接进行预测，模型准确率仅为69.31%，存在较大的预测性能差距。该结果表明，仅使用少量关于染色体结构的信息可以高效调整不同细胞系之间染色质空间结构特征与CNV之间的关联模式差异性。

基于数据扰动的图卷积神经网络模型性能评估

团队设计了一系列Hi-C数据扰动方法，验证GCN模型在CNV预测任务中的鲁棒性，并评估染色质结构与CNV标签的关联程度。具体地，设计了包括：CNV标签重构、CNV标签乱序、Hi-C行读数信号交换、Hi-C读数数值交换以及Hi-C读数高斯加噪，5种扰动方式。实验结果表示，GCN

模型对于Hi-C数据高斯加噪具有较强的鲁棒性。同时，两种CNV标签扰动方法对于模型预测准确率的影响最大，这也证明了GCN模型在训练和预测过程中并非为CNV标签驱动的。

未来应用前景

这一研究深入了解不同机器学习方法在Hi-C数据预测CNV方面的优势和局限性，还从三维染色质拓扑结构角度为CNV检测提供了新的技术手段。研究团队表示，这些发现可以加深了解CNV与染色质结构之间复杂关系的认识，及如何影响转录调控元件之间的三维交互，以支持未来的三维基因组研究。

QB期刊介绍

Quantitative Biology （QB）期刊是由清华大学、北京大学、高教出版社联合创办的全英文学术期刊。QB主要刊登生物信息学、计算生物学、系统生物学、理论生物学和合成生物学的最新研究成果和前沿进展，并为生命科学与计算机、数学、物理等交叉研究领域打造一个学术水平高、可读性强、具有全球影响力的交叉学科期刊品牌。

《前沿》系列英文学术期刊

由教育部主管、高等教育出版社主办的《前沿》（Frontiers）系列英文学术期刊，于2006年正式创刊，以网络版和印刷版向全球发行。系列期刊包括基础科学、生命科学、工程技术和人文社会科学四个主题，是我国覆盖学科最广泛的英文学术期刊群，其中12种被SCI收录，其他也被AHCI、Ei、MEDLINE或相应学科国际权威检索系统收录，具有一定的国际学术影响力。系列期刊采用在线优先出版方式，保证文章以最快速度发表。

中国学术前沿期刊网

http://journal.hep.com.cn



来源：Quantitative Biology

更多 科学进展 请访问 https://www.iikx.com/news/progress/

本文版权归原作者所有，请勿用于商业用途，爱科学iikx.com转发