PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2018, 45(11): 1093~1105
www.pibb.ac.cn

三维基因组数据分析方法进展 *

张祥林　方　欢　汪小我 **
(清华大学自动化系，合成与系统生物学研究中心，生物信息学教育部重点实验室，
北京信息科学与技术国家研究中心生物信息学研究部，北京 100084)

摘要　对染色质三维结构的探究逐渐成为了解基因组功能与基因调控关系的必要手段．近年来，随着高通量染色体构象捕获(Hi-C)等技术的发展和高通量测序成本的降低，全基因组交互作用的数据量快速增长，交互作用图谱分辨率不断提高．这给三维基因组学发展带来机遇的同时，也给计算建模带来了挑战．当前，三维基因组数据的分析方法覆盖面广，包括了数据前期处理、标准化、可视化、特征提取、三维建模等环节，但是如何从中选择高效、准确的方法却成为制约研究者们开展研究的一项关键因素．本文根据这些方法的适用场景、原理及特点进行系统地归纳，并重点关注了针对新技术或新需求的分析方法，以期促进这一领域中信息学方法的应用和开发，助力三维基因组学的研究．

关键词　Hi-C，三维基因组学，基因组结构，生物信息学
学科分类号　Q5，Q6，Q7
DOI: 10.16476/j.pibb.2018.0101

随着人类基因组百科全书计划的发展，调控元件在人类基因组上的一维线性分布越来越清晰．然而，想要更深层次地理解基因的调控关系，必须要获知调控元件和基因在细胞核中的三维空间关系．近年来，作为研究基因组功能和基因调控关系的重要途径，三维基因组学得到了迅速的发展，被誉为基因组学的第三次浪潮[1].

目前，研究三维基因组的技术可分为两大类：显微成像技术和生物化学技术．以荧光原位杂交技术 (fluorescence *in situ* hybridization，FISH)[2]为代表的显微成像技术通过对两个或多个 DNA 位点的荧光染色确定位点间的位置和空间距离．显微成像技术具有单细胞水平观察物理位置和距离的优点，但普遍存在分辨率低、通量小的局限．以染色体构象捕获类技术为代表的生物化学技术则通过消化和重连接物理上接近的染色质片段确定不同位点的空间接近性．其中，3C (chromosome conformation capture)技术[3]用于检测单点对单点的交互作用，4C (circularized chromosome conformation capture 或 chromosome conformation capture-on-chip)技术[4-5]用于检测单点对多点的交互作用，5C (carbon-copy chromosome conformation capture)技术[6]用于检测多点对多点的交互作用，ChIA-PET (chromatin interaction analysis by paired-end tag sequencing) 技术[7]用于检测全基因组特定蛋白质介导的交互作用，高通量染色体构象捕获 (high-throughput chromosome conformation capture，Hi-C)技术[8]用于检测全基因组无偏的交互作用.有关这类技术的详细介绍可参见相关文献[9−10]．其中 Hi-C 和 ChIA-PET 技术已经成为这类技术的重要代表[1]，目前 Hi-C 技术文库制备步骤相对简单、所需细胞数较少，有着更广泛的应用．有关 ChIA-PET 数据计算方法可参考我们发表的相关综述[11]，本文重点介绍与 Hi-C 数据分析相关的计算方法．

1　三维基因组的研究现状

在细胞分裂、分化和衰老等生理过程中，染色

* 国家自然科学基金(31371341, 61773230, 61721003)和清华大学自主科研项目(20141081175)资助.
** 通讯联系人.
Tel: 010-62794294-808, E-mail: xwwang@tsinghua.edu.cn
收稿日期：2018-04-04，接受日期：2018-06-22

质状态及空间结构随时间动态变化．那么， DNA 分子如何通过层层折叠摆放在细胞核中且保证基因的精确调控成为了三维基因组研究关注的问题．从大尺度上来说，几十年前人们就观察到分裂间期的细胞核中不同染色体占据了核内不同的空间位置，即染色体疆域(chromosome territory，CT) 现象[12]．然而从染色质疆域到核小体之间的高级结构却是随着近年来染色体构象捕获技术的发展得到了系统性探究．2009 年，随着 Hi-C 技术的发明，研究者们发现染色体可划分为两种类别的区室(A/B compartment)，分别对应开放和关闭两种染色质状态[8]，且具有同类区室交互作用强，异类区室交互作用弱的特点．随着 Hi-C 分辨率的提高，人们进一步发现，在哺乳动物基因组中，约 1Mb 的基因组区域形成了拓扑关联结构域(topological associated domain，TAD)[13]，这些结构域具有域内交互作用强，域间交互作用弱的特点．拓扑关联结构域细分为染色质环(chromatin loop)，例如增强子 - 启动子环，与基因的调控密切相关．以哺乳动物基因组为例，细胞核内形成了从染色质环到拓扑关联结构域，到 A/B 区室，再到染色体疆域的层级结构(图 1)．

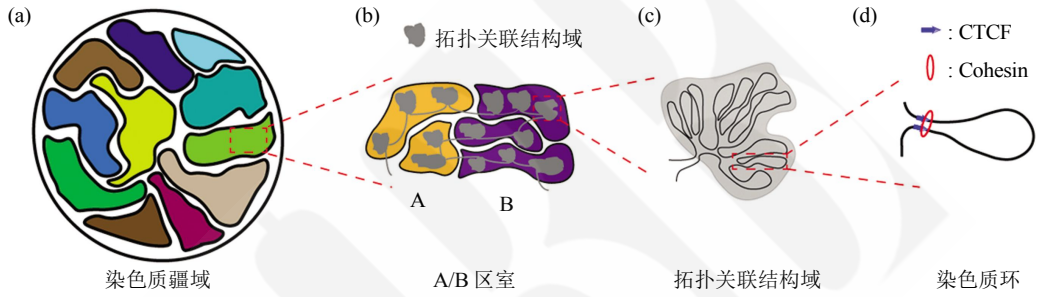

Fig. 1　Hierarchical levels of chromatin architecture

图 1　染色质层级结构

利用 Hi-C 等相关实验技术，研究人员以发育[14-17]、衰老[18-19]、疾病[20-21]等生理过程为模型，增进了对染色质层级结构的发生机制及其与基因组功能关系的理解．这其中，中国科学家取得了一系列重要成果．例如，清华大学颉伟组[15]、中国科学院刘江组[16]独立发现在小鼠胚胎发育早期，包括拓扑关联结构域和 A/B 区室在内的染色质高级结构有缓慢建立的过程；北京大学李程组[21]发现在多发性骨髓瘤中拷贝数变异的断点与拓扑关联结构域的边界重合；上海交通大学邵志峰和 Daniel 组[22]发现昆虫与哺乳动物之间具有拓扑关联结构域的保守特性；复旦大学文波组[23]发现核基质蛋白 HNRNPU 在三维结构中起重要作用；香港中文大学钟思林组与山东农业大学李平华组[24]发现大型植物的三维基因组由局部的 A/B 区室决定；华中农业大学张献龙与杨庆勇组[25]解析了在多倍体棉花中三维结构的变化．与此同时，国内外实验室在鉴别染色质层级结构的影响因素上也取得了重要进展，发现了 CTCF、cohesin 等蛋白质对拓扑结构域、染色质环结构形成的重要影响[26-30]．

近几年来，一系列 Hi-C 的衍生技术满足了不同的需求，例如 *in situ* Hi-C 通过减少随机连接的影响提高了信噪比[31]、Capture Hi-C 提高某些特定区域的分辨率[32]、单细胞 Hi-C 丰富了对单细胞层面染色体三维结构动态变化的理解[33-35]．尽管 Hi-C 及其变种技术已经被广泛地应用，但是解读 Hi-C 数据仍然依赖于系统、全面的生物信息学方法的进步．

2　Hi-C 的实验原理与数据特点

Hi-C 实验的基本流程分为五大步骤：a．将染色质交互作用通过甲醛交联固定；b．使用限制性内切酶切割基因组，并用生物素标记切割末端；c．使用 DNA 连接酶对切割末端连接制造嵌合分子；d．纯化和打断 DNA 嵌合分子，并筛选出带有生物素标记的 DNA 片段；e．对 DNA 文库进行双端测序[8]．对测序得到的嵌合分子序列比对可得到其两端在基因组上的来源，再通过统计基因组上任意两段区域间的嵌合分子数，构造染色质交互作用矩阵．其中，区域的长度称为交互作用图谱的分

辨率，以人的基因组(约 3×10^9 bp)为例，10 kb 分辨率下的全基因组交互作用图谱矩阵大小约为 300 000×300 000．尽管没有一个统一的分辨率选择标准，但是有相关研究建议，应使 >80%的单位分辨率区域有至少 1 000 个嵌合分子[31]．

随着测序成本的降低，Hi-C 数据的数据量越来越大，分辨率越来越高，现有实验利用 4.9×10^9 的交互作用数获得了人基因组最高分辨率为 1 kb 的交互作用图谱[31]．虽然分辨率的提高为研究三维基因组的精细结构带来了机遇，但是也给计算资源与分析方法带来了挑战．由于 DNA 分子具有多分子聚合物特性，线性距离越远的两个位点间随机交互作用越弱，染色体内交互作用矩阵因而呈现出偏离对角线越远强度越小的趋势，这种不协调的交互作用模式不利于对染色质层级结构的特征提取．此外，由于实验技术和基因组本身性质的影响，Hi-C 实验获得的交互作用矩阵受到诸如酶切片段长度不均、嵌合分子 GC 含量不一致等因素的影响，不同区域的交互作用强弱不适合直接进行比较．

为了从庞大的交互作用矩阵中提取出有价值的关键信息，一般需要对 Hi-C 数据进行建模分析．数据建模方法主要分为两大类，分别是数据驱动的方法和染色质物理结构驱动的方法．数据驱动的方法一般以交互作用数据为研究对象，基于统计模型进行拟合，挖掘有效信息．染色质物理结构驱动的方法主要是从物理结构或能量角度仿真重现交互作用图谱，其中包含了从细胞群体角度、染色质物理结构仿真角度等出发的一系列方法．这些方法为认识三维基因组结构提供了强大的支撑．接下来，我们将从 Hi-C 数据处理的基本流程出发对三维基因组研究中的信息学方法进行系统地归纳．

3　三维基因组数据分析方法

传统 Hi-C 实验数据的处理流程可分为测序数据定位、过滤、binning、标准化、可视化、层级结构鉴别、三维建模等部分(图 2)．测序数据定位到基因组的方法主要分为两种，一种是酶切位点截断后匹配，另一种是从较短序列开始，重复迭代匹配，直到唯一地匹配到基因组或者到达测序的最大长度．过滤步骤主要是去除自连接、扩增重复、非特异连接、未连接的 dangling 等片段．接下来，将有效交互片段按照一定分辨率通过 binning 可得到原始的交互作用图谱，进一步经过标准化方法校正可去除系统偏差．通常 Hi-C 数据前期处理的软件将读段比对、过滤、binning、标准化等步骤打包在一起供用户使用，表 1 列出了主要的 Hi-C 数据处理集成软件．为了更直观地查看不同基因组区域的交互作用及整合其他基因组信息，用户通常需要对交互作用可视化，国内方面，清华大学张奇伟组[36]、中国科学院张治华组[37]、北京大学李程组[38]分别开发了可实现基因组交互作用和物理结构可视化的软件，与可视化相关的其他软件可参照相关文献[39-40]．本文重点针对标准化、层级结构鉴别、三维建模等方法进行系统总结和归纳．

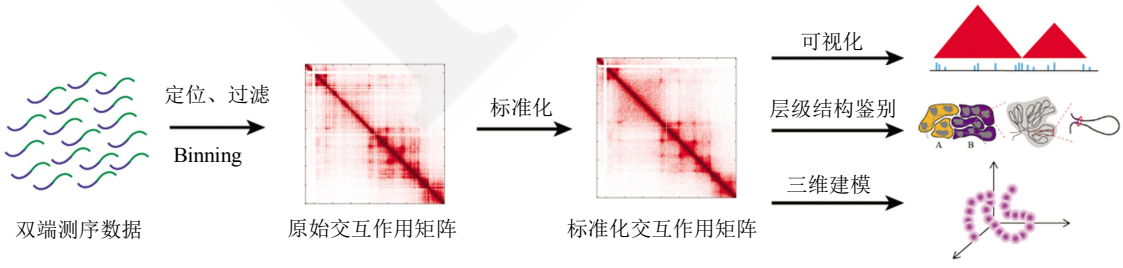

Fig. 2　The processing workflow of Hi-C data

图 2　Hi-C 数据处理流程

3.1　标准化方法

经过读段定位、过滤、binning 得到的原始交互作用矩阵受到诸如酶切片段长度、GC 含量等系统偏差的影响．这些偏差使不同基因组区域间交互作用强度的可比性下降，对后续的特征提取等分析产生不利影响．为了减少这种偏差，需要进行标准化的校正．常用的标准化方法可分为两大类：确定因子校正和矩阵平衡(matrix balancing)．

Table 1 Integration software of Hi-C data processing

表 1 Hi-C 数据处理集成软件

软件	读段定位	过滤	Binning	标准化	特点	实现语言
HiCUP[41]	√	√			针对读段定位和过滤	Perl，R
HIPPIE[42]	√	√			鉴别显著的增强子 - 目标基因交互作用	R，Python，Perl
HOMER[43]		√	√	√	可提供显著交互作用鉴别	Java，R，Perl
HiFive[44]		√	√	√	处理 Hi-C 或 5C 数据	Python
HiC-Pro[45]	√	√	√	√	可用于特异等位片段的分析	Python
HiCdat[46]	√	√	√	√	可实现可视化、样本比较、区室鉴别等功能	C++,R
TADbit[47]	√	√	√	√	提供了 TAD 的鉴别和三维模型构建	Python
Hiclib[48]	√	√	√	√	读段迭代定位，矩阵平衡，矩阵分解	Python
HiCapp[49]	√	√	√	√	包含 HiCcorrector、HiCUP、caICB	Perl，R
Juicer[50]	√	√	√	√	包含 HiCCUPS 和 Arrowhead 等特征提取模块	Java
HiC-bench[51]	√	√	√	√	整合多种特征鉴别和标注的功能	R，C++，Python
NucProcess[34]	√	√	√		单细胞 Hi-C 数据处理，配合三维建模软件 NucDynamics 使用	Python

确定因子校正的基本思路是确定系统偏差的来源，然后构建概率模型计算期望的交互作用用于校正原始交互作用矩阵．Yaffe 等[52]考虑酶切位点距离、嵌合分子 GC 含量、序列特异性三种系统误差来源构造联合概率修正模型，通过交互作用数据学习最大似然的参数，进而估计期望交互作用．但该方法需要估计的参数数量多，在酶切片段分辨率下计算最大似然导致较高的计算成本．HiCNorm 改进这一方法，使用基于泊松回归的方法计算修正模型，减少模型参数，在更低分辨率的交互作用上估计参数，提高了校正效果和计算效率[53].

矩阵平衡方法不考虑具体的系统偏差来源，假设基因组上任意两个区域应具有相同可见性(equal visibility)[48]，即有效交互作用数之和相等．按照这一假设，hiclib 通过迭代算法获得真实矩阵和偏差．Rao 等[31, 54]采用了矩阵平衡中 KR 方法提高了收敛速度．为应对高分辨率下交互作用矩阵过大的问题，HiCorrector 方法将交互作用矩阵分块并行处理，以实现矩阵平衡的目标[55].

此外，为适应单细胞 Hi-C 和癌症细胞 Hi-C 数据中的新特点，研究人员提出了针对性标准化方法．其中，针对单细胞 Hi-C 数据中多零特点，scHiCNorm 方法使用零膨胀模型(zero-inflated model)和栅栏模型(hurdle model)刻画单细胞的交互作用，对三种系统偏差进行校正[56]．而在癌症细胞系或组织中，通常会出现拷贝数的变异，对单条染色体交互作用使用矩阵平衡方法无法去除染色体之间因拷贝数变异而引起的偏差，caICB 方法根据不同染色体中交互作用与基因组线性距离关系一致的特点，在单条染色体矩阵平衡基础上对交互作用进行新的标准化，使得不同染色体之间具有可比性[49]. 当然，全基因组水平上的矩阵平衡方法也可以消除拷贝数变异的影响．

尽管两类标准化方法被广泛地采用，但两种方法有着不同的假设，在应用中应根据自身需求进行选择．其中，在确定因子校正方法中，有可能忽略到其他系统偏差，例如癌症基因组中的拷贝数变异，而此时采用矩阵平衡算法有助于去除这一偏差(如果需要保留这一偏差，确定因子校正方法更适合); 在矩阵平衡方法中，有些基因组区域由于自身性质不一定具有相同的“可见性”，例如超强增强子区域本身有更多的远程交互作用，利用确定因子校正方法则更有利于保留这种特征[57]．因此，需要根据研究目的和对象选择合适的标准化方法．

3.2 层级结构鉴别

三维基因组层级结构的鉴别包括区室、拓扑关联结构域、染色质交互作用的鉴别(图 3)．通过层级结构鉴别可将模式复杂的交互作用矩阵转化为容易解读的特征信号，既便于样本间的比较，也便于与其他生物特征关联分析．

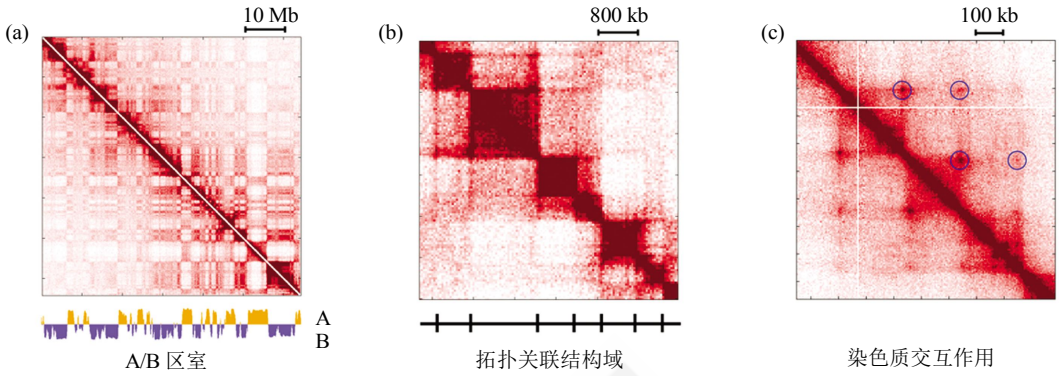

Fig. 3　The identification of hierarchical architecture from Hi-C interaction

图 3　基于 Hi-C 交互作用的层级结构鉴别

3.2.1　区室的鉴别

在较低分辨率的交互作用图谱中呈现深浅交替的格子状(plaid)交互作用模式(图 3a)，这一特点提示存在两种特性的染色质状态，即 A/B 区室．格子状交互模式反映了两种染色质状态在空间中的接近性．对 A/B 区室分析发现，A 区室与基因表达高、GC 含量高、染色质开放性强等活跃特征相关；反之，B 区室与不活跃特征相关，且 A/B 区室具有细胞特异性．通过对区室的分析，反映了三维结构层面的染色质开放性特征．

A/B 区室鉴别的步骤包括交互作用矩阵去除距离偏差、计算相关系数矩阵、主成分分析，最终获得呈现双峰分布的第一主成分，进而根据基因密度的高低分为 A 区室(开放)和 B 区室(关闭)[8]．CscoreTool 改进传统计算方式，将问题转化为求解最大似然，改善传统方法中计算速度慢、内存占用大的问题[58]．随着对三维基因组认识的发展，涌现出多种计算方法可计算类似 A/B 区室的特征．

迭代修正和特征分解(iterative correction and eigenvector decomposition，ICE)方法将经过矩阵平衡的全基因组交互作用矩阵进行特征值分解，前三个特征值对应的特征向量分别表征染色质的开放性、着丝粒位置、端粒位置信息[48]．而北京大学李程组[59]利用 Markov 过程模拟蛋白质在染色体上随机游走可得到蛋白质分子的平衡分布．这一特征与染色体开放性有更强的相关性．借助这一特征，研究者提出了在分化过程中结构蛋白基因先于其他基因改变空间结构和开放性的模型．同济大学张勇组[60]通过计算对数比例将基因组分为富集和缺乏功能元件的区域．此外，对染色体间交互作用矩阵聚类，可以避免染色体内基因组线性距离的影响，获得交互作用图谱相似的染色质区域．将低分辨率(Mb)的全基因组染色体间(trans)交互作用进行 K-均值聚类可得到高度活性区域、低度活性的着丝粒附近区域、低度活性的着丝粒远端区域 3 种区域[52]．利用更高分辨率(100 kb) 的染色体间交互作用矩阵进行聚类，可获得更加精确的子区室(sub-compartment)[31]．子区室将传统的 A 区室分为 A1 和 A2 2 种；B 区室分为 B1、B2、B3、B4 4 种，不同子类富集特异的生物信号．对染色体间交互作用聚类为我们研究染色体间交互作用提供了一种策略．

由于距离因素的影响，在高分辨率交互作用矩阵中远距离的交互作用更加稀疏．为避免稀疏的远端交互作用产生的偏差，计算 A/B 区室通常选择低分辨率的交互作用矩阵或者平滑过的高分辨率交互作用矩阵．此外，多个研究中发现，样本间除了区室出现翻转现象，区室间交互作用的程度也会有变化[14-15, 23]，例如在我们参与的早期胚胎发育的研究中，A/B 区室有逐渐建立的过程．因此，评价区室间交互作用程度变化也逐渐成为区室研究的重要内容．

3.2.2　拓扑关联结构域的鉴别

TAD 是染色质层级结构中非常重要的一部分，被认为是复制时间调控的稳定单元[61]，其边界通常富集 CTCF 蛋白和管家基因(housekeeping gene)，具有物种保守性．TAD 内包含子 TAD，形成 TAD 层面的层级结构，一般认为子 TAD 具有细胞特异

性．TAD 结构的破坏往往会引起基因调控的紊乱，从而引发疾病[62-64]，因此对 TAD 的鉴别有助于了解局部染色质区域的空间关系，提供调控元件与基因的潜在关系．目前针对 TAD 的鉴别方法非常丰富，根据实现的步骤可分为两类，一类在二维交互作用矩阵中提取一维特征进行分割，另一类基于交互作用矩阵直接分割．

Table 2 Methods of identifying TADs

表 2 TAD 的鉴别方法

方法	层级结构	单个 TAD 或边界的可信度衡量指标	特点
DI-HMM[13]			基于一维信号
Insulation Score[65]		√	基于一维信号
TopDom[66]		√	增加了统计检验，提高准确度
Armatus[67]			有交互作用距离对目标函数影响的参数
HiCseg[68]			借助图像分割理论，最大似然
Arrowhead[31]		√	高分辨率交互作用中鉴别更小的结构域(185 kb)
TADtree[69]	√		借助 TAD 内交互作用随距离变化的特点
TADbit[47]		√	BIC 惩罚的最大似然
HiTAD[70]	√		在改进的 DI-HMM 结果空间上优化 TAD 组合
GMAP[71]	√	√	利用混合高斯构建交互作用的后验
IC-Finder[72]	√	√	基于改进的分级聚类
ClusterTAD[73]	√		使用多种非监督机器学习方法
MrTADFinder[74]			转化为在网络中寻找模块
Laplacian[75]	√		转化为图模型，使用谱聚类
CaTCH[77]	√	√	可获得从间隔到子 TAD 层面的层级结构

第一类方法的代表包括 DI-HMM(directionality index-hidden markova model)[13]和基于阻隔系数(insulation score)[65-66]的方法．DI-HMM 方法首先采用方向指数(directionality index)来表征一个染色体区域与上下游交互作用的偏差，当这个偏差出现符号跳转时，意味着可能出现 TAD 的边界．在方向指数中利用隐马尔可夫模型可以推断出 TAD 的具体位置．阻隔系数用来反映基因组上一段区域对交互作用的阻隔效果．在 TAD 边界附近，阻隔系数会达到局部最大．通过算法确定这些局部极值位置即可以确定 TAD 的边界，进而确定 TAD 位置．然而，局部极值的位置易受参数选择的影响而偏离实际 TAD 边界，TopDom 方法则在获得局部极值后，将局部极值区域拓展，进一步通过统计检验的办法确定 TAD 的精确位置，提高了 TAD 鉴定的准确性[66]．这类方法得到的 TAD 不具有层级结构．

第二类鉴别方法既有根据 TAD 结构特点构造目标函数，通过优化来鉴别 TAD 结构[47, 67-71]，也有借助聚类方法鉴别 TAD 结构[72-75]．例如，HiTAD 方法基于最优化交互作用分割的原则，在改进的 DI-HMM 鉴别出的结构域为底层结构域基础上，构造优化函数搜索最优 TAD 分割方案，同样模式应用在子 TAD 的鉴别，从而获得层级的 TAD[70]．借助此方法，研究者定义了几种与复制时间、转录相关的层级 TAD 变化模式．目前，鉴别 TAD 的层级结构，已成为这类方法的发展方向，这也符合高分辨率交互作用图谱的实际特点．此外，部分 TAD 的鉴别给出衡量单个 TAD 或边界可信度的指标，这些指标为研究者提供了 TAD 或边界强度的定量表示，对研究一些特殊过程中 TAD 强度变化有着重要作用[15, 28]．通过 TAD 鉴别算法获得 TAD 的位置后，还可通过计算 TAD 内部和 TAD 间交互作用的相对比例衡量 TAD 的改变．研究人员利用这一方法在细胞衰老、干细胞分化、体细胞重编程等过程中发现 TAD 层面染色质交互作用的改变及其与相关生理过程的关系[17-18, 76]．表 2 收集了常

用的 TAD 鉴别算法.

虽然大量的 TAD 鉴别算法为用户提供了庞大的选择余地，但是不同算法侧重点不同，适用场景也有所差异，需要根据自身需要进行选择(表 2). 此外，某些算法的参数选择依赖经验或对数据的了解，更有一些参数的不同选择反映不同生物层面上的侧重. 因此，使用算法过程中应重视参数的选择，可通过变化参数的方式考察参数的效果. 目前，已有一些研究针对其中一部分算法的效果进行系统的评估[78-79].

3.2.3 染色质交互作用的鉴别

在 Hi-C 实验获得的交互作用矩阵中，大多数的交互作用并不是染色质环信号. 从交互作用矩阵中获取显著交互作用，对于认识染色质的结构、基因的调控具有重要的意义. 鉴别显著交互作用的基本思路是对交互作用矩阵元素进行统计建模，构建背景模型，从而识别显著交互作用(表 3). 按照鉴别的内容分为显著交互作用的鉴别和显著差异交互作用的鉴别.

Table 3 Methods of identifying significant interactions

表 3 显著交互作用鉴别方法

算法或软件	模型	针对性	特点
HIPPIE[42]	负二项分布	显著交互作用	在酶切片段分辨率下，鉴别近距离交互作用
Fit-HiC[80]	二项分布	显著交互作用	对交互作用与距离的关系两次拟合
GOTHiC[81]	二项分布	显著交互作用	利用覆盖率估计期望的交互作用
HiC-DC[82]	零截尾负二项分布	显著交互作用	模型考虑了 Hi-C 数据的多零和高散度的特性
HOMER[43]	二项分布	显著或差异交互作用	多功能集成软件
HMRFBayes[83]& FastHiC[84]	负二项分布、隐 Markov 随机场	显著交互作用	考虑相邻交互作用的影响
HiCCUPS[31]	泊松过程	显著交互作用	去除 TAD 结构的影响
PSYCHIC[85]	对数正态分布	显著交互作用	基于 TAD 结构构建背景模型
CHiCAGO[86]	负二项方分布、泊松分布	显著交互作用	针对 Capture Hi-C 实验数据
Dynamic Interactions[13]	二项分布	差异交互作用	利用生物学重复
HiBrowse[87] & DiffHiC[88]	负二项分布	差异交互作用	借助 edgeR，利用生物学重复
FIND[89]	空间泊松过程	差异交互作用	考虑相邻交互作用的关联性

显著交互作用的鉴别可分为两种，一种是考虑系统偏差或基因组线性距离的影响，直接对单独的交互作用元素建模[42-43, 80-82]. 例如，Fit-HiC 通过对交互作用与距离关系的两次建模，估计了随机聚合物的交互作用受距离的影响，利用二项分布对交互作用建模，给出交互作用的显著性[80]. 另一种是考虑相邻交互作用或其他层级结构的影响[31, 83-85]，这种方法通常适用于更高分辨率的交互作用，在一定程度上可以增强结果的准确性. 例如，PSYCHIC 方法在 TAD 结构域的基础上建立背景模型，用于鉴别增强子与启动子之间的显著交互作用[85]. 此外，针对 Capture Hi-C 数据，CHiCAGO[86]利用负二项随机变量与泊松随机变量结合的方法对其交互作用建模，其中负二项随机变量期望看作距离的函数，实现对布朗碰撞的刻画；泊松随机变量期望看作与片段性质相关的函数，实现对技术误差的刻画.

在鉴别显著差异交互作用方面，最初有研究利用二项分布对相同距离的交互作用建模，确定差异交互作用的 P 值，再通过组合来自不同样本的生物学重复建立背景，从而确定差异交互作用的错误发现率 (false discovery rate, FDR) [13]. 随后，HiBrowse[87]和 DiffHiC[88]借鉴了基因差异表达鉴别软件 edgeR 利用了生物学重复鉴定显著差异的交互作用. 最近，清华大学张奇伟组发表的 FIND 方法考虑高分辨率交互作用矩阵中相邻交互作用关联的特性，用空间泊松过程鉴别差异交互作用[89]，提高了鉴别结果的信噪比. 显著差异交互作用的鉴别是对显著交互作用鉴别的拓展.

在稀疏性和高散度的交互作用图谱中鉴别显著的交互作用仍然具有很高的挑战性. 最近相关研究比较了几种鉴别方法，发现其结果在生物学重复中

的重复性不高[79]，表明这类方法仍然有很大的改进空间.

3.3 三维建模

Hi-C 数据的一项重要应用是三维建模. 三维建模有助于重现染色质的物理结构，加深对染色质结构的理解，辅助科学研究. 在构建三维模型中有两个重要问题需要考虑，第一个问题是如何利用交互作用图谱，通常可以将交互作用频率按照确定或非确定的函数关系与空间距离进行转化，也可以将交互作用的频率作为权重加权在优化问题的目标函数中; 第二个问题是构建平均的三维模型还是三维模型的集合，理论上大量细胞的 Hi-C 数据是多个单细胞结构的叠加，构建三维模型集合更加符合理论要求. 三维建模的方法可分为优化模型和物理模型[90]，但目前常见方法以优化模型为主.

在优化模型中，如果考虑交互作用的不确定性，可以利用统计模型对交互作用的数据进行建模，然后构建三维坐标在内的后验分布，进而求解符合后验的空间结构，这类方法包括 MCMC5C[91]、BACH [92]、 BACH-MIX [92]、 PASTIS [93]、 HAS [94]、PRAM[95]等. 这些模型一般假设交互作用服从正态分布[91]或泊松分布[92–95]，通过幂率关系将交互作用和距离互相转化，既可获得平均模型[91–93]也可获得模型的集合[91–92, 95].

优化模型中如果不考虑交互作用的不确定性，两步走的策略被广泛地使用. 两步走的方法指，先将交互作用频率转化为距离，然后利用距离重构三维位置. 这类方法包括了 AutoChrom3D [96]、ChromSDE [97]、 ShRec3D [98]、 ShRec3D+ [99]、Chromosome3D[100]、LorDG[101]等方法. 这些方法重点关注了交互作用频率与距离的转化关系. 例如，AutoChrom3D 对交互数据归一化处理后使用两个线性变化进行距离转化[96]; ShRec3D 借用 Floyd-Warshall 算法计算最短距离[98]. 在距离到坐标的转化中，LorDG 方法将优化目标从平方误差替代为洛伦兹函数，从而减少弱交互作用不一致对优化的影响[101]. 为解决优化模型中构建高分辨三维模型的困难，“两阶段算法”和 miniMDS 方法被提出[102–103]. 两阶段算法利用 ChromSDE 方法获得单个染色体结构，然后利用基因组幂率关系组合多个染色体构建全基因组的三维结构[102]. MiniMDS 方法则是根据交互作用将基因组分为多层结构，在高分辨率的局部区域和低分辨率的整体区域分别进行三维构建，然后将高分辨率的局部三维结果低分辨处理后经过旋转、镜像、移动后与低分辨率的三维结果匹配，最终获得整体的三维结构[103].

虽然两步走策略被广泛采用，但是交互作用频率与最终的三维结构优化并没有紧密结合起来，建立基于交互作用的优化过程是一种新的选择[104–107]. 其中，GEM 方法借助流型的思想，建立基于交互作用频率和构象能量的优化过程，将 Hi-C 数据空间的约束转化到三维空间[104].

随着单细胞 Hi-C 技术的发展，单细胞数据的三维构建也受到关注. 单细胞 Hi-C 数据虽然有分辨率低、高稀疏性的特点，但理论上讲，数据来自一个细胞核，一定是协调一致的. 贝叶斯结构推断(ISD)方法[108]和基于流型的优化 (MBO)方法[109]都在单细胞 Hi-C 数据的三维建模上有所应用. NucDynamics 方法基于染色质物理模型，建立结构的能量函数，通过模拟退火求解单细胞三维坐标[34].

目前，三维建模的方法为理解三维结构特点和驱动力提供了新的手段. 染色体间的交互作用稀疏、尺度大，非常不利于与传统一维信号结合发掘信息，而利用三维建模的方法，研究发现，基因密度高、活跃的区域倾向分布在核内，而基因密度低、与核纤层(lamina)有关的区域倾向分布在核表层的特点[110]. 借助群体理论结合物理结构约束将集成的交互作用图谱拆分为单个细胞的三维结构模型，发掘出单细胞水平上着丝粒聚集、染色体间存在稳定聚合物、调控元件之间波动连接等特征[111–113]. 借助三维建模的方法有助于探究传统分析方法中不易处理的问题.

4 展 望

尽管针对同一种分析功能开发了多种算法，但是这些算法都有对应的假设和侧重点，为增强对这些算法性能的了解，采用全面的衡量体系进行评价也变得至关重要. 在一些研究中，同时采用两种或两种以上的方法对数据进行分析也成为保证分析结果准确性的手段之一[15–16, 21]. 然而各个软件或算法开发语言、文件处理格式不同，尝试多种算法无疑增加了研究者们的负担，因而统一文件格式、同一平台下整合多种算法也变得越来越重要.

近年来，除了上述传统分析方法外，基于 Hi-C 数据的研究也取得了新的突破. 最近，天津大学唐继军组与宾夕法尼亚州立大学岳峰组[114]合

作，利用深度卷积神经网络增强了 Hi-C 交互作用分辨率，这一研究既提示我们三维结构的形成机制有规律可循，也为提高 Hi-C 交互作用分辨率提供了新方法．此外，Hi-C 数据还可与其他信息结合进行信息挖掘．例如复旦大学田卫东组[115]利用 Hi-C 数据结合人和 45 种脊椎动物的系统发育相关性，预测了远端调控元件的目标基因；中国科学院上海生命科学研究院韩敬东组[116]和复旦大学钟扬组[117]独立借助 Hi-C 数据探究了 Alu 的远程交互作用特性；北京放射医学研究所伯晓晨组[118]借助高分辨率 Hi-C 数据发现了空间上临近的转录因子结合聚集点；中国科学院基因组所张治华组[119]利用中等数据量 Hi-C 数据结合核小体定位信息预测高分辨率交互作用．在国际方面，结合 Hi-C 等交互作用数据与单核苷酸多态性鉴别潜在致病基因或调控关系也取得了一系列成果[120−122]；Hi-C 数据中反映的染色体内比染色体间交互作用强、随距离增加交互作用强度递减的特性还可用于基因组组装[123−124]、拷贝数变异与染色体异位的鉴别[125−126]．包含 Hi-C 在内的三维基因组数据已经成为基因组功能研究的重要资源，目前已有整合三维基因组数据的数据库供研究者使用(如 3DGD[127]、3CDB[128]、CCSI[129])．

当今，通过对基因组编辑产生扰动来探究三维结构性质已经成为可能[26, 62, 64]．随着酵母基因组的合成成功[130]，合成生物学发展进入新的时代．在未来，基因组三维结构将逐渐成为基因组改造或合成中的重要关注因素[131]，这也使三维基因组学与合成生物学联系更加紧密．此外，随着多国大规模人群基因组计划的启动，整合单核苷酸多态性、表型、三维基因组结构关系也将有助于高通量筛选疾病的潜在致病基因，助力疾病研究．在此背景下，三维基因组数据分析方法也将不断进步，以适应新的需求．

参 考 文 献

[1] 李国亮, 阮一骏, 谷瑞升, 等. 起航三维基因组学研究. 科学通报, 2014, **59**(13): 1165−1172
Li G L, Ruan Y J, Gu R S, *et al*. Chinese Science Bulletin (Chinese Version), 2014, **59**(13): 1165−1172

[2] Langer-Safer P R, Levine M, Ward D C. Immunological method for mapping genes on Drosophila polytene chromosomes. Proc Natl Acad Sci USA, 1982, **79**(14): 4381−4385

[3] Dekker J, Rippe K, Dekker M, *et al*. Capturing chromosome conformation. Science, 2002, **295**(5558): 1306−1311

[4] Zhao Z, Tavoosidana G, Sjolinder M, *et al*. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nature Genetics, 2006, **38**(11): 1341−1347

[5] Simonis M, Klous P, Splinter E, *et al*. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nature Genetics, 2006, **38**(11): 1348−1354

[6] Dostie J, Richmond T A, Arnaout R A, *et al*. Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Research, 2006, **16**(10): 1299−1309

[7] Fullwood M J, Liu M H, Pan Y F, *et al*. An oestrogen-receptor-alpha-bound human chromatin interactome. Nature, 2009, **462**(7269): 58−64

[8] Lieberman-Aiden E, Van Berkum N L, Williams L, *et al*. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science, 2009, **326** (5950): 289−293

[9] Ramani V, Shendure J, Duan Z. Understanding spatial genome organization: methods and insights. Genomics, Proteomics & Bioinformatics, 2016, **14**(1): 7−20

[10] Sati S, Cavalli G. Chromosome conformation capture technologies and their impact in understanding genome function. Chromosoma, 2017, **126**(1): 33−44

[11] He C, Li G, Nadhir D M, *et al*. Advances in computational ChIA-PET data analysis. Quantitative Biology, 2016, **4** (3): 217−225

[12] Cremer T, Cremer M. Chromosome territories. Cold Spring Harbor Perspectives in Biology, 2010, **2**(3): a003889

[13] Dixon J R, Selvaraj S, Yue F, *et al*. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature, 2012, **485**(7398): 376−380

[14] Bonev B, Mendelson Cohen N, Szabo Q, *et al*. Multiscale 3D genome rewiring during mouse neural development. Cell, 2017, **171**(3): 557−572 e524

[15] Du Z, Zheng H, Huang B, *et al*. Allelic reprogramming of 3D chromatin architecture during early mammalian development. Nature, 2017, **547**(7662): 232−235

[16] Ke Y, Xu Y, Chen X, *et al*. 3D Chromatin structures of mature gametes and structural reprogramming during mammalian embryogenesis. Cell, 2017, **170**(2): 367−381 e320

[17] Dixon J R, Jung I, Selvaraj S, *et al*. Chromatin architecture reorganization during stem cell differentiation. Nature, 2015, **518**(7539): 331−336

[18] Chandra T, Ewels P A, Schoenfelder S, *et al*. Global reorganization of the nuclear landscape in senescent cells. Cell Reports, 2015, **10**(4): 471−483

[19] Criscione S W, De Cecco M, Siranosian B, *et al*. Reorganization of chromosome architecture in replicative cellular senescence. Science Advances, 2016, **2**(2): e1500882

[20] Mccord R P, Nazario-Toole A, Zhang H, *et al*. Correlated alterations in genome organization, histone methylation, and DNA-lamin A/C interactions in Hutchinson-Gilford progeria

syndrome. Genome Research, 2013, **23**(2): 260-269

[21] Wu P, Li T, Li R, *et al*. 3D genome of multiple myeloma reveals spatial genome disorganization associated with copy number variations. Nature Communications, 2017, **8**(1): 1937

[22] Wang Q, Sun Q, Czajkowsky D M, *et al*. Sub-kb Hi-C in D. melanogaster reveals conserved characteristics of TADs between insect and mammalian cells. Nature Communications, 2018, **9**(1): 188

[23] Fan H, Lv P, Huo X, *et al*. The nuclear matrix protein HNRNPU maintains 3D genome architecture globally in mouse hepatocytes. Genome Research, 2018, **28**(2): 192-202

[24] Dong P, Tu X, Chu P Y, *et al*. 3D Chromatin architecture of large plant genomes determined by local A/B compartments. Molecular Plant, 2017, **10**(12): 1497-1509

[25] Wang M, Wang P, Lin M, *et al*. Evolutionary dynamics of 3D genome architecture following polyploidization in cotton. Nature Plants, 2018, **4**(2): 90-97

[26] Guo Y, Xu Q, Canzio D, *et al*. CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. Cell, 2015, **162**(4): 900-910

[27] Tang Z, Luo O J, Li X, *et al*. CTCF-Mediated human 3D genome architecture reveals chromatin topology for transcription. Cell, 2015, **163**(7): 1611-1627

[28] Nora E P, Goloborodko A, Valton A L, *et al*. Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. Cell, 2017, **169**(5): 930-944 e922

[29] Schwarzer W, Abdennur N, Goloborodko A, *et al*. Two independent modes of chromatin organization revealed by cohesin removal. Nature, 2017, **551**(7678): 51-56

[30] Rao S S P, Huang S C, Glenn St Hilaire B, *et al*. Cohesin loss eliminates all loop domains. Cell, 2017, **171**(2): 305-320 e324

[31] Rao S S, Huntley M H, Durand N C, *et al*. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell, 2014, **159**(7): 1665-1680

[32] Mifsud B, Tavares-Cadete F, Young A N, *et al*. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nature Genetics, 2015, **47**(6): 598-606

[33] Nagano T, Lubling Y, Stevens T J, *et al*. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature, 2013, **502**(7469): 59-64

[34] Stevens T J, Lando D, Basu S, *et al*. 3D structures of individual mammalian genomes studied by single-cell Hi-C. Nature, 2017, **544**(7648): 59-64

[35] Nagano T, Lubling Y, Varnai C, *et al*. Cell-cycle dynamics of chromosomal organization at single-cell resolution. Nature, 2017, **547**(7661): 61-67

[36] Tang B, Li F, Li J, *et al*. Delta: a new web-based 3D genome visualization and analysis platform. Bioinformatics, 2018, **34**(8): 1409-1410

[37] Li R, Liu Y, Li T, *et al*. 3Disease Browser: a web server for integrating 3D genome and disease-associated chromosome rearrangement data. Scientific Reports, 2016, **6**: 34651

[38] Djekidel M N, Wang M, Zhang M Q, *et al*. HiC-3Dviewer: a new tool to visualize Hi-C data in 3D space. Quantitative Biology, 2016, **5**(2): 183-190

[39] Yardimci G G, Noble W S. Software tools for visualizing Hi-C data. Genome Biology, 2017, **18**(1): 26

[40] Yang D, Jang I, Choi J, *et al*. 3DIV: a 3D-genome interaction viewer and database. Nucleic Acids Research, 2018, **46** (D1): D52-D57

[41] Wingett S, Ewels P, Furlan-Magaril M, *et al*. HiCUP: pipeline for mapping and processing Hi-C data. F1000 Research, 2015, **4**: 1310 (DOI:10.12688/f1000research.7334.1)

[42] Hwang Y C, Lin C F, Valladares O, *et al*. HIPPIE: a high-throughput identification pipeline for promoter interacting enhancer elements. Bioinformatics, 2015, **31**(8): 1290-1292

[43] Heinz S, Benner C, Spann N, *et al*. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. Mol Cell, 2010, **38**(4): 576-589

[44] Sauria M E, Phillips-Cremins J E, Corces V G, *et al*. HiFive: a tool suite for easy and efficient HiC and 5C data analysis. Genome Biology, 2015, **16**: 237

[45] Servant N, Varoquaux N, Lajoie B R, *et al*. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biology, 2015, **16**: 259

[46] Schmid M W, Grob S, Grossniklaus U. HiCdat: a fast and easy-to-use Hi-C data analysis tool. BMC Bioinformatics, 2015, **16**: 277

[47] Serra F, Bau D, Goodstadt M, *et al*. Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. Plos Computational Biology, 2017, **13**(7): e1005665

[48] Imakaev M, Fudenberg G, Mccord R P, *et al*. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nature Methods, 2012, **9**(10): 999-1003

[49] Wu H J, Michor F. A computational strategy to adjust for copy number in tumor Hi-C data. Bioinformatics, 2016, **32**(24): 3695-3701

[50] Durand N C, Shamim M S, Machol I, *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Systems, 2016, **3**(1): 95-98

[51] Lazaris C, Kelly S, Ntziachristos P, *et al*. HiC-bench: comprehensive and reproducible Hi-C data analysis designed for parameter exploration and benchmarking. BMC Genomics, 2017, **18**(1): 22

[52] Yaffe E, Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nature Genetics, 2011, **43**(11): 1059-1065

[53] Hu M, Deng K, Selvaraj S, *et al*. HiCNorm: removing biases in Hi-C data *via* Poisson regression. Bioinformatics, 2012, **28**(23): 3131-3133

[54] Knight P A, Ruiz D. A fast algorithm for matrix balancing. IMA Journal of Numerical Analysis, 2012, **33**(3): 1029-1047

[55] Li W, Gong K, Li Q, *et al*. Hi-Corrector: a fast, scalable and memory-efficient package for normalizing large-scale Hi-C data. Bioinformatics, 2015, **31**(6): 960−962

[56] Liu T, Wang Z. scHiCNorm: a software package to eliminate systematic biases in single-cell Hi-C data. Bioinformatics, 2018, **34**(6): 1046−1047

[57] Schmitt A D, Hu M, Jung I, *et al*. A compendium of chromatin contact maps reveals spatially active regions in the human genome. Cell Reports, 2016, **17**(8): 2042−2059

[58] Zheng X, Zheng Y. CscoreTool: fast Hi-C compartment analysis at high resolution. Bioinformatics, 2018, **34**(9): 1568−1570

[59] Wang Y, Fan C, Zheng Y, *et al*. Dynamic chromatin accessibility modeled by Markov process of randomly-moving molecules in the 3D genome. Nucleic Acids Research, 2017, **45**(10): e85

[60] Liu L, Zhang Y, Feng J, *et al*. GeSICA: genome segmentation from intra-chromosomal associations. BMC Genomics, 2012, **13**: 164

[61] Pope B D, Ryba T, Dileep V, *et al*. Topologically associating domains are stable units of replication-timing regulation. Nature, 2014, **515**(7527): 402−405

[62] Lupianez D G, Kraft K, Heinrich V, *et al*. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell, 2015, **161**(5): 1012−1025

[63] Valton A L, Dekker J. TAD disruption as oncogenic driver. Current Opinion in Genetics & Development, 2016, **36**: 34−40

[64] Franke M, Ibrahim D M, Andrey G, *et al*. Formation of new chromatin domains determines pathogenicity of genomic duplications. Nature, 2016, **538**(7624): 265−269

[65] Crane E, Bian Q, Mccord R P, *et al*. Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature, 2015, **523**(7559): 240−244

[66] Shin H, Shi Y, Dai C, *et al*. TopDom: an efficient and deterministic method for identifying topological domains in genomes. Nucleic Acids Research, 2016, **44**(7): e70

[67] Filippova D, Patro R, Duggal G, *et al*. Identification of alternative topological domains in chromatin. Algorithms for Molecular Biology, 2014, **9**: 14

[68] Levy-Leduc C, Delattre M, Mary-Huard T, *et al*. Two-dimensional segmentation for analyzing Hi-C data. Bioinformatics, 2014, **30**(17): i386−392

[69] Weinreb C, Raphael B J. Identification of hierarchical chromatin domains. Bioinformatics, 2016, **32**(11): 1601−1609

[70] Wang X T, Cui W, Peng C. HiTAD: detecting the structural and functional hierarchies of topologically associating domains from chromatin interactions. Nucleic Acids Research, 2017, **45**(19): e163

[71] Yu W, He B, Tan K. Identifying topologically associating domains and subdomains by Gaussian mixture model and proportion test. Nature Communications, 2017, **8**(1): 535

[72] Haddad N, Vaillant C, Jost D. IC-Finder: inferring robustly the hierarchical organization of chromatin folding. Nucleic Acids Research, 2017, **45**(10): e81

[73] Oluwadare O, Cheng J. ClusterTAD: an unsupervised machine learning approach to detecting topologically associated domains of chromosomes from Hi-C data. BMC Bioinformatics, 2017, **18**(1): 480

[74] Yan K K, Lou S, Gerstein M. MrTADFinder: a network modularity based approach to identify topologically associating domains in multiple resolutions. Plos Computational Biology, 2017, **13**(7): e1005647

[75] Chen J, Hero A O, 3rd, Rajapakse I. Spectral identification of topological domains. Bioinformatics, 2016, **32**(14): 2151−2158

[76] Krijger P H, Di Stefano B, De Wit E, *et al*. Cell-of-origin-specific 3D genome structure acquired during somatic cell reprogramming. Cell Stem Cell, 2016, **18**(5): 597−610

[77] Zhan Y, Mariani L, Barozzi I, *et al*. Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes. Genome Research, 2017, **27**(3): 479−490

[78] Dali R, Blanchette M. A critical assessment of topologically associating domain prediction tools. Nucleic Acids Research, 2017, **45**(6): 2994−3005

[79] Forcato M, Nicoletti C, Pal K, *et al*. Comparison of computational methods for Hi-C data analysis. Nature Methods, 2017, **14**(7): 679−685

[80] Ay F, Bailey T L, Noble W S. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Research, 2014, **24**(6): 999−1011

[81] Mifsud B, Martincorena I, Darbo E, *et al*. GOTHiC, a probabilistic model to resolve complex biases and to identify real interactions in Hi-C data. Plos One, 2017, **12**(4): e0174744

[82] Carty M, Zamparo L, Sahin M, *et al*. An integrated model for detecting significant chromatin interactions from high-resolution Hi-C data. Nature Communications, 2017, **8**: 15454

[83] Xu Z, Zhang G, Jin F, *et al*. A hidden Markov random field-based Bayesian method for the detection of long-range chromosomal interactions in Hi-C data. Bioinformatics, 2016, **32**(5): 650−656

[84] Xu Z, Zhang G, Wu C, *et al*. FastHiC: a fast and accurate algorithm to detect long-range chromosomal interactions from Hi-C data. Bioinformatics, 2016, **32**(17): 2692−2695

[85] Ron G, Globerson Y, Moran D, *et al*. Promoter-enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains. Nature Communications, 2017, **8**(1): 2237

[86] Cairns J, Freire-Pritchett P, Wingett S W, *et al*. CHiCAGO: robust detection of DNA looping interactions in capture Hi-C data. Genome Biology, 2016, **17**(1): 127

[87] Paulsen J, Sandve G K, Gundersen S, *et al*. HiBrowse: multi-purpose statistical analysis of genome- wide chromatin 3D organization. Bioinformatics, 2014, **30**(11): 1620−1622

[88] Lun A T, Smyth G K. diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. BMC Bioinformatics, 2015, **16**: 258

[89] Djekidel M N, Chen Y, Zhang M Q. FIND: difFerential chromatin INteractions detection using a spatial Poisson process. Genome Research, 2018, **28**(3): 412−422

[90] 彭城, 李国亮, 张红雨, 等. 染色质三维结构重建及其生物学意义. 中国科学: 生命科学, 2014, **44**(8): 794-802
Peng C, Li G L, Zhang H Y, *et al*. Scientia Sinica Vitae, 2014, **44**(8): 794-802

[91] Rousseau M, Fraser J, Ferraiuolo M A, *et al*. Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. BMC Bioinformatics, 2011, **12**: 414

[92] Hu M, Deng K, Qin Z, *et al*. Bayesian inference of spatial organizations of chromosomes. Plos Computational Biology, 2013, **9**(1): e1002893

[93] Varoquaux N, Ay F, Noble W S, *et al*. A statistical approach for inferring the 3D structure of the genome. Bioinformatics, 2014, **30**(12): i26-i33

[94] Zou C, Zhang Y, Ouyang Z. HSA: integrating multi-track Hi-C data for genome-scale reconstruction of 3D chromatin structure. Genome Biology, 2016, **17**: 40

[95] Park J, Lin S. A random effect model for reconstruction of spatial chromatin structure. Biometrics, 2017, **73**(1): 52-62

[96] Peng C, Fu L Y, Dong P F, *et al*. The sequencing bias relaxed characteristics of Hi-C derived data and implications for chromatin 3D modeling. Nucleic Acids Research, 2013, **41**(19): e183

[97] Zhang Z, Li G, Toh K C, *et al*. 3D chromosome modeling with semi-definite programming and Hi-C data. Journal of Computational Biology, 2013, **20**(11): 831-846

[98] Lesne A, Riposo J, Roger P, *et al*. 3D genome reconstruction from chromosomal contacts. Nature Methods, 2014, **11**(11): 1141-1143

[99] Li J, Zhang W, Li X. 3D genome reconstruction with ShRec3D+ and Hi-C data. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2018, **15**(2): 460-468

[100]Adhikari B, Trieu T, Cheng J. Chromosome3D: reconstructing three-dimensional chromosomal structures from Hi-C interaction frequency data using distance geometry simulated annealing. BMC Genomics, 2016, **17**(1): 886

[101]Trieu T, Cheng J. 3D genome structure modeling by Lorentzian objective function. Nucleic Acids Research, 2017, **45**(3): 1049-1058

[102]Segal M R, Bengtsson H L. Reconstruction of 3D genome architecture *via* a two-stage algorithm. BMC Bioinformatics, 2015, **16**: 373

[103]Rieber L, Mahony S. miniMDS: 3D structural inference from high-resolution Hi-C data. Bioinformatics, 2017, **33**(14): i261-i266

[104]Zhu G, Deng W, Hu H, *et al*. Reconstructing spatial organizations of chromosomes through manifold learning. Nucleic Acids Research, 2018, **46**(8): e50

[105]Caudai C, Salerno E, Zoppe M, *et al*. Inferring 3D chromatin structure using a multiscale approach based on quaternions. BMC Bioinformatics, 2015, **16**: 234

[106]Paulsen J, Sekelja M, Oldenburg A R, *et al*. Chrom3D: three-dimensional genome modeling from Hi-C and nuclear lamin-genome contacts. Genome Biology, 2017, **18**(1): 21

[107]Trieu T, Cheng J. MOGEN: a tool for reconstructing 3D models of genomes from chromosomal conformation capturing data. Bioinformatics, 2016, **32**(9): 1286-1292

[108]Carstens S, Nilges M, Habeck M. Inferential structure determination of chromosomes from single-cell Hi-C data. Plos Computational Biology, 2016, **12**(12): e1005292

[109]Paulsen J, Gramstad O, Collas P. Manifold based optimization for single-cell 3D genome reconstruction. Plos Computational Biology, 2015, **11**(8): e1004396

[110]Di Stefano M, Paulsen J, Lien T G, *et al*. Hi-C-constrained physical models of human chromosomes recover functionally-related properties of genome organization. Scientific Reports, 2016, **6**: 35985

[111]Tjong H, Li W, Kalhor R, *et al*. Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. Proc Natl Acad Sci USA, 2016, **113**(12): E1663-1672

[112]Dai C, Li W, Tjong H, *et al*. Mining 3D genome structure populations identifies major factors governing the stability of regulatory communities. Nature Communications, 2016, **7**: 11549

[113]Giorgetti L, Galupa R, Nora E P, *et al*. Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. Cell, 2014, **157**(4): 950-963

[114]Zhang Y, An L, Xu J, *et al*. Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus. Nature Communications, 2018, **9**(1): 750

[115]Lu Y, Zhou Y, Tian W. Combining Hi-C data with phylogenetic correlation to predict the target genes of distal regulatory elements in human genome. Nucleic Acids Research, 2013, **41**(22): 10391-10402

[116]Su M, Han D, Boyd-Kirkup J, *et al*. Evolution of Alu elements toward enhancers. Cell Reports, 2014, **7**(2): 376-385

[117]Gu Z, Jin K, Crabbe M J, *et al*. Enrichment analysis of Alu elements with different spatial chromatin proximity in the human genome. Protein & Cell, 2016, **7**(4): 250-266

[118]Chen H, Jiang S, Zhang Z, *et al*. Exploring spatially adjacent TFBS-clustered regions with Hi-C data. Bioinformatics, 2017, **33**(17): 2611-2614

[119]Zhang H, Li F, Jia Y, *et al*. Characteristic arrangement of nucleosomes is predictive of chromatin interactions at kilobase resolution. Nucleic Acids Research, 2017, **45**(22): 12739-12751

[120]Grubert F, Zaugg J B, Kasowski M, *et al*. Genetic control of chromatin states in humans involves local and distal chromosomal interactions. Cell, 2015, **162**(5): 1051-1065

[121]Javierre B M, Burren O S, Wilder S P, *et al*. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. Cell, 2016, **167**(5): 1369-1384 e1319

[122]Di Iulio J, Bartha I, Wong E H M, *et al*. The human noncoding genome defined by genetic diversity. Nature Genetics, 2018, **50**(3): 333-337

[123]陶婧芬, 谢婷, 郑觉非, 等. 基于染色质交互数据的基因组组装方法. 生物技术通报, 2015, **31**(11): 43-50
Tao J F, Xie T, Zhang J F, *et al*. Biotechnology Bulletin, 2015,

31(11): 43-50

[124]Flot J F, Marie-Nelly H, Koszul R. Contact genomics: scaffolding and phasing (meta)genomes using chromosome 3D physical signatures. FEBS Letters, 2015, **589**(20 Pt A): 2966-2974

[125]Harewood L, Kishore K, Eldridge M D, *et al*. Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. Genome Biology, 2017, **18**(1): 125

[126]Chakraborty A, Ay F. Identification of copy number variations and translocations in cancer cells from Hi-C data. Bioinformatics, 2017, **34**(2): 338-345

[127]Li C, Dong X, Fan H, *et al*. The 3DGD: a database of genome 3D structure. Bioinformatics, 2014, **30**(11): 1640-1642

[128]Yun X, Xia L, Tang B, *et al*. 3CDB: a manually curated database of chromosome conformation capture data. Database: The Journal of Biological Databases and Curation, 2016, **2016**: baw044

[129]Xie X, Ma W, Songyang Z, *et al*. CCSI: a database providing chromatin-chromatin spatial interaction information. Database: The Journal of Biological Databases and Curation, 2016, **2016**: bav124

[130]Richardson S M, Mitchell L A, Stracquadanio G, *et al*. Design of a synthetic yeast genome. Science, 2017, **355**(6329): 1040-1044

[131]Mercy G, Mozziconacci J, Scolari V F, *et al* 3D organization of synthetic and scrambled chromosomes. Science, 2017, **355**(6329): eaaf4597

The Advancement of Analysis Methods of Chromosome Conformation Capture Data*

ZHANG Xiang-Lin, FANG Huan, WANG Xiao-Wo**

(*Department of Automation, Center for Synthetic and Systems Biology, Tsinghua University; Ministry of Education Key Laboratory of Bioinformatics; Bioinformatics Division, BNRIST, Beijing* 100084, China)

Abstract The investigation about chromatin 3D structure is becoming one indispensable way in studying genome functions and gene regulation. In past several years, thanks to the development of chromatin conformation capture technology and decreasing cost of high throughput sequencing, the amount of whole-genome interaction data increases rapidly with the ascending resolutions. This not only brought the chances for interpreting 3D genome, but also challenged the modeling methods. Nowadays, methods of analyzing these data covered a wide range, including pre-processing, normalization, visualization, features extraction and 3D modeling; however, choosing efficient and precise computational methods becomes an obstacle limiting the study of 3D genome. In this paper, we sum up these methods according to their suitable conditions, principles and characters and focus on the methods for new technologies and requirements in order to promote the application and development of these methods, assisting the investigation of 3D genome.

Key words Hi-C, three-dimensional genomics, genome architecture, bioinformatics

DOI: 10.16476/j.pibb.2018.0101

* This work was supported by grants from The National Natural Science Foundation of China (31371341, 61773230, 61721003) and Tsinghua University Initiative Scientific Research Program (20141081175).

**Corresponding author.

Tel: 86-10-62794294-808, E-mail: xwwang@tsinghua.edu.cn

Received: April 4, 2018 Accepted: June 22, 2018