DOI:10.12113/202008004

利用改进的 3DMax 算法重构染色体 3D 结构

刘立伟, 么会丽
(大连交通大学 理学院,辽宁 大连 116028)

摘　要: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,促使染色体和基因组三维结构建模的研究取得了很大进展,已经提出了几种从染色体构象捕捉数据中构建单个染色体或整个基因组结构的方法。文中通过对在 Hi-C 数据基础上对染色体三维结构重建的相关文献进行分析,总结了重建染色体三维空间结构的经典算法 3DMax 的原理,并且提出了一种新的随机梯度上升算法:XNadam,是 Nadam 优化方法的一个变体,将其应用于 3DMax 算法中,以便提高 3DMax 算法的性能,从而用于预测染色体三维结构。

关键词:Hi-C;染色体三维结构;梯度上升;三维基因组
中图分类号:Q93　　**文献标志码**:A　　**文章编号**:1672-5565(2021)04-254-06

An improved 3DMax algorithm to reconstruct three-dimensional structure of chromosomes

LIU Liwei, YAO Huili
(*School of Science*, *Dalian Jiaotong University*, *Dalian* 116028, *Liaoning*, *China*)

Abstract: In recent years, with the development of high-throughput chromosome conformation capture (Hi-C) technology and the reduction of high-throughput sequencing cost, the data volume of whole-genome interaction has increased rapidly, and the resolution of interaction maps keeps improving. Great progress has been made in the research of 3D structure modeling of chromosomes and genomes. Several methods have been proposed to construct chromosome structures from chromosome conformation capture data. Based on the Hi-C data, this paper analyzes the relevant literature of 3D structure reconstruction of chromosomes and summarizes the principle of 3DMax, which is a classical algorithm to construct the 3D structure of a chromosome. In this paper, a new gradient ascent optimization algorithm called XNadam is proposed, which is a variant of Nadam optimization method. When XNadam is applied to 3DMax algorithm, the performance of 3DMax algorithm can be improved, which can be used to predict the three-dimensional structure of a chromosome.

Keywords: Hi-C; 3D chromosome structure; Gradient ascent; 3D genome

染色体构象捕获技术,特别是 Hi-C 技术的发展,使基因组空间构象的分析和研究成为生物信息学和计算生物学的重要课题。在高通量下一代测序技术的帮助下,Hi-C 技术可以生成全基因组范围的、大规模的染色体内和染色体间相互作用数据,能够详细描述基因组内的空间相互作用。这些数据可以用来重建染色体的三维结构,用于研究 DNA 复制、基因调控、基因组相互作用、基因组折叠和基因组功能[1-9]。目前,根据构建染色体三维模型原理上的不同,可将结构模型分为两大类:即基于概率约束的预测模型和基于距离约束的预测模型。由于细胞染色体的三维结构是动态的,我们可以用一些概率分布来描述它,从而将染色体三维结构的构建问题转化为概率模型的构建问题。实现这一过程的方法称为基于概率约束的预测模型。其中有一些方法是通过两步来进行染色体基因组三维结构建模,即

收稿日期:2020-08-08;**修回日期**:2020-09-17.
基金项目:中国博士后科学基金项目(No.2018M631782);辽宁省教育厅项目(No. JDL2019032);辽宁省自然科学基金项目(No. 201800278).
作者简介:刘立伟,男,副教授、硕导,研究方向:生物信息学.E-mail: liutree80@163.com.

将 Hi-C 数据中片段对之间的相互作用频率(IF)转换为片段对之间的距离,然后通过对距离值进行优化,推断出最能满足距离的 3D 结构,即求解出染色体片段的三维空间坐标,实现这两步过程的方法称为基于距离约束的模型。在对距离值进行优化的过程中,常用梯度迭代优化算法优化目标函数,目前现有的随机梯度优化算法有 SGD、Momentum、Nesterov、Adagrad、Adadelta、Adam、Adamax、Nadam[10-16],本文根据目前存在的随机梯度优化算法的优点和缺点,提出了 XNadam 算法。

1　3DMax 算法原理简介

3DMax 算法[17]是由 Oluwadare 等人提出,它使用最大似然方法从 Hi-C 数据中推断染色体的三维结构。3DMax 比现有的大多数方法都要快,它只依赖于通过最小二乘残差来优化预测模型的结构坐标。3DMax 算法也是基于距离约束的预测模型,该算法首先通过初始化一组染色体空间坐标,计算出染色体片段之间的欧氏距离,然后通过转换函数将输入的接触频率矩阵转换成的距离矩阵,并假设距离矩阵中的每一个数据点独立并服从正态分布,从而构建一个基于正态分布的对数似然目标函数,再采用梯度上升算法迭代优化目标函数直到算法收敛,在迭代过程中同步更新染色体三维坐标。在 3DMax 算法中,采用梯度上升算法对目标函数进行迭代优化的过程中,计算出距离斯皮尔曼相关系数(DSCC),并选出最大 DSCC 系数对应的转换参数,最优转换参数下的染色体片段坐标,是我们推断出最能满足距离的可视化染色体结构。3DMax 算法还有一个变体,称为 3DMax1,它在输入噪声时对输入接触矩阵进行额外的预处理和滤波。

在 3DMax1 算法中,Oluwadare 等人,使用的随机梯度上升算法是自适应梯度算法(AdaGrad)[18]。自适应梯度算法(AdaGrad)是一种基于梯度的优化方法,它可以使学习速率适应每个参数,可以对低频的参数做较大的更新,对高频的做较小的更新,也因此,对于稀疏的数据它的表现很好。它的缺点是分母会不断积累,内存需求较大,这样学习率就会收缩并最终会变得非常小,甚至趋于零。

2　XNadam 算法原理简介

除了 3DMax1 算法中应用到的自适应梯度算法(AdaGrad)以外,目前现有的随机梯度优化算法还有 SGD、Momentum、Nesterov、RMSprop、Adadelta、Adam、Adamx、Nadam。本文提出了一种新的随机梯度上升算法 XNadam,是 Nadam 的一个变体。Nadam(Nesterov-acceler Adaptive Moment Estimation)是将 Adam 与 Nesterov 算法结合在一起,具备二者的优势。它对学习率的约束将更强,使得此算法在某些问题上效果更好。Nadam 的优点主要在于经过偏置矫正后,每一次迭代学习率都有个确定范围,为不同的参数计算不同的自适应学习率,使得参数比较平稳,同时内存需求较小,也适应于大多非凸优化,适应于大数据和高维空间。由于加入了 Nesterov 项,则在梯度更新时做一个校正,避免前进太快,同时提高灵敏;XNadam 作为 Nadam 的一个变体,原理跟 Nadam 类似。区别是 XNadam 采用正弦的相关方式进行学习率的衰减,同时利用正弦方式的衰减率对一阶矩估计进行修正,所以 XNadam 具有 Nadam 算法的所有优点。

3　XNadam 算法

XNadam 算法

要求:初始学习率 α_0,最小学习率 $\alpha_{\min}$

要求:矩估计的指数衰减速率 β_2,β_1,取值区间均为[0,1)内

要求:用于数值稳定的小常数 δ(默认 10^{-6})

要求:初始化一阶和二阶矩变量 n=0,m=0,初始化时间步长 t=0

计算梯度:$g_{t+1} = \nabla f(S_t)$

对梯度更新:$g\hat{}_{t+1} = \dfrac{g_{t+1}}{1 - \prod_{i=1}^{t+1}\beta_1^i}$

更新有偏二阶矩估计:$n_{t+1} = \beta_2 \cdot n_t + (1-\beta_2)\cdot g_{t+1}^2$

修正二阶矩的偏差:$n\hat{}_{t+1} = \dfrac{n_{t+1}}{1-\beta_2^{t+1}}$

更新有偏一阶矩估计:$m_{t+1} = \beta_1^{(t+1)}\cdot m_t + (1-\beta_1^{(t+1)})\cdot g_{t+1}$ 衰减速率:$\beta_1^{(t)} = \beta_1 \cdot 0.5 \cdot (1 + \sin(\dfrac{t\pi}{T}))\ T = \dfrac{t(t+3)}{2}$

对一阶矩修正:$m\hat{}_{t+1} = \dfrac{m_{t+1}}{1 - \prod_{i=1}^{t+2}\beta_1^i}$

继续修正一阶矩:$\bar{m}_{t+1} = \beta_1^{(t+2)}\cdot \hat{m}_{t+1} + (1-\beta_1^{(t+1)})\cdot \hat{g}_{t+1}$

学习率:$\alpha_t = \alpha_0 \cdot [(1-\alpha_{\min})\cdot 0.5\cdot(1+\sin(\dfrac{t\pi}{T})) + \alpha_{\min}]$

计算更新:$\Delta S_{t+1} = \alpha_{t+1}\dfrac{\bar{m}_{t+1}}{\sqrt{\hat{n}_{t+1}+\varepsilon}}$

应用更新:$S_{t+1} = S_t + \Delta S_{t+1}$

4 模型相似性和准确性的测量

为了测试该算法,利用 3DMaxAdaGrad,3DMaxNadam 和 3DMaxXNadam 算法对酵母染色体三维结构进行重构,进而比较他们的性能。3DMaxAdaGrad、3DMaxNadam、3DMaxXNadam 是 3DMax 算法的一个变体,其中 3DMaxAdaGrad 是结合了极大似然算法和 AdaGrad 算法预测模型,而 3DMaxNadam 是结合了极大似然和 Nadam 优化算法的预测模型,3DMaxXNadam 是结合了极大似然和 XNadam 优化算法的预测模型。在比较的过程中,3DMaxXNadam 算法中 $\beta_1 = 0.6$, $\beta_2 = 0.999\ 9$, $\alpha_{\min} = 0.000\ 1$,只有学习率是变化的。

本文采用距离均方根误差(Distance Root Mean Squared Error ,DRMSE)、距离斯皮尔曼相关系数(Distance Spearman Correlation Coefficient ,DSCC)、距离皮尔森相关系数(Distance Pearson Correlation Coefficient,DPCC)[19-22]来量化结构的相似性,从而衡量预测方法的性能。

在本研究中输入的归一化接触频率值所选择的数据集是酵母的 Hi-C 数据,酵母 Hi-C 数据用 Yaffe 和 Tanay 等人[23]提出的技术进行了归一化。3DMaxAdaGrad, 3DMaxNadam 和 3DMaxXNadam 算法对 1-12 号染色体三维空间结构重构的 DSCC、DPCC 和 DRMSE 三个评价指标结果(见图 1)。

图 1　3DMaxAdaGrad,3DMaxnadam 和 3DMaxnadam 算法重构染色体三维结构的评估结果

Fig.1　Evaluation results for 3DMaxAdaGrad, 3DMaxNadam, and 3DMaxXNadam

图 1 中的柱形图表示了 3DMaxAdaGrad, 3DMaxNadam 和 3DMaxXNnadam 三种算法重构染色体三维结构的 DSCC、DPCC 和 DRMSE 评估结果。第一列黄色柱形图代表三种算法对酵母 1-12 号染色体三维结构重构的 DSCC 结果,观察图形可知,对于 1 号、2 号染色体,3DMaxAdaGrad 算法重建染色体三维结构的 DSCC 值最大,其次是 3DMaxXNnadam 算法的 DSCC 值,其中 3DmaxNadam 算法的 DSCC 值最小;对于 8 号染色体,3DmaxNadam 算法重构染色体三维结构的 DSCC 值最大,其次是 3DmaxXNadam 算法的 DSCC 值较大,最小的是 3DMaxAdaGrad 的 DSCC 值;除了上述三条染色体,其余的染色体都是 3DmaxXNadam 算法重构染色体三维结构的 DSCC 值最大。

第二列绿色柱形图代表三种算法对酵母 1-12 号染色体三维结构重构的 DPCC 结果,从图可以看出,除了 7 号染色体,其余 11 条染色体中,3DmaxXNadam 算法重构染色体三维结构的 DPCC 值最大,其次是 3DmaxNadam 算法的 DPCC 值,最小的是 3DMaxAdaGrad 算法的 DPCC 值。

第三列红色柱形图代表三种算法对酵母 1-12 号染色体重构的三维结构 DRMSE 结果,从图可以看出,除了 7 号染色体,其余 11 条染色体中,3DmaxXNadam 算法重构染色体三维结构的 DRMSE

值最小,其次是 3DmaxNadam 算法的 DRMSE 值,最大的是 3DMaxAdaGrad 算法重构染色体三维结构的 DRMSE 值。

根据上述实验结果可知,3DmaxXNadam 算法的性能更好。

5 染色体三维结构可视化

通过算法比较,3DmaxXNadam 算法对染色体三维结构进行重构具有较高的准确度。在此,使用该算法对酵母 1-12 号染色体三维空间结构进行重构,得了酵母 1-12 号染色体的三维效果图(见图 2)。

6 结 论

本研究在 Nadam 优化算法的基础上发展了一种新的梯度上升优化算法(XNadam),并引入最大似然算法分别和 AdaGrad、Nadam、XNadam 算法相结合得到 3DMaxAdaGrad, 3DMaxNadam 和 3DMaxXNadam 算法。利用 3DMaxAdaGrad, 3DMaxNadam 和 3DMaxXNadam 算法对酵母染色体三维结构进行重构,并且计算了生成的染色体优化结构的 DSCC、DPCC 和 DRMSE 的值,从而衡量预测方法的性能。在真实数据集上的实验结果表明,3DmaxXNadam 算法可以更有效地从 Hi-C 接触矩阵中重建染色体模型,同时与其他方法相比,它速度更快,对内存的要求也较低。这个结论说明新的梯度上升优化算法(XNadam)在优化对数似然目标函数上,它的表现比大多数其他优化方法更好。

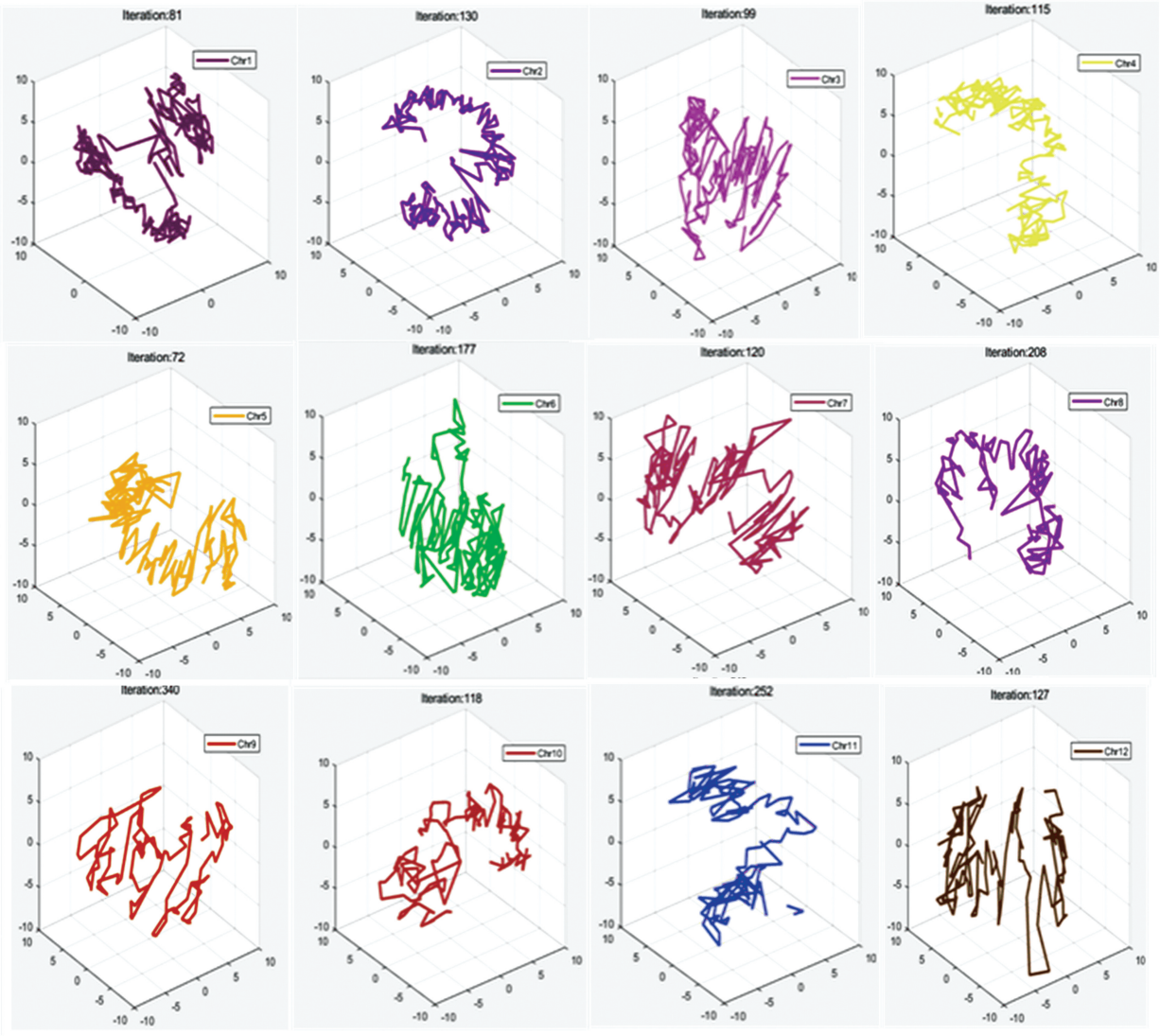

图 2 染色体的三维空间结构重构效果图

Fig.2 Sketch of 3D structure reconstruction of chromosomes

参考文献(References)

[1]BAÙ D, MARTI-RENOM M A. Genome structure determination via 3C-based data integration by the integrative modeling platform[J]. Methods,2012,58(3):300-306. DOI:10.1007/s10577-010-9167-2.

[2]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[J]. BMC Bioinformatics,2011,12(1):414. DOI:10.1186/1471-2105-12-414.

[3]TRUSSART M, SERRA F, BAÙ D, et al. Assessing the limits of restraint-based 3D modeling of genomes and genomic domains[J]. Nucleic Acids Research, 2015, 43(7):3465-3477.DOI:10.1093/nar/gkv221.

[4]ZHANG Zhizhuo, LI Guoliang, TOH K C, et al. 3D chromosome modeling with semi-definite programming and Hi-C data[J]. Journal of Computational Biology,2013,20:831-846.DOI:10.1089/cmb.2013.0076.

[5]HU M, DENG K, QIN Z, et al. Bayesian inference of spatial organizations of chromosomes[J]. PLoS Computational Biology,2013,9(1): e1002893. DOI:10.1371/journal.pcbi.1002893.

[6]ZOU C, ZHANG Y, OUYANG Z. HSA: Integrating multitrack hi-C data for genomescale reconstruction of 3D chromatin structure[J]. Genome Biology,2016,17(1):40. DOI:10.1186/s13059-016-0896-1.

[7]TRIEU T,CHENG J. Large-scale reconstruction of 3D structures of human chromosomes from chromosomal contact data [J]. Nucleic Acids Research,2014,42(7): 52. DOI:10.1093/nar/gkt1411.

[8]NOWOTNY J, AHMED S, XU L, et al. Iterative reconstruction of three-dimensional models of human chromosomes from chromosomal contact data[J]. BMC Bioinformatics, 2015,16(1):1. DOI:10.1186/s12859-015-0772-0.

[9]LESNE A, RIPOSO J, ROGER P, et al. 3D genome reconstruction from chromosomal contacts[J]. Nature Methods, 2015,11(11):1141-1143. DOI:10.1038/nmeth.3104.

[10]BORIS T P. Some methods of speeding up the convergence of iteration methods[J]. USSR Computational Mathematics and Mathematical Physics, 1964, 4(5):1-17. DOI:10.1016/0041-5553(64).

[11]NESTEROV Y. Introductory lectures on convex optimization: A basic course, volume 87[M]. Springer Science & Business Media,2013. DOI:10.1007/978-1-4419-1153-7_1171.

[12]NESTEROV Y. A method of solving a convex programming problem with convergence rate O(1/k2)[J]. Soviet Mathematics Doklady, 1983, 27(2), 372-376. DOI:10.3969/j.issn.0372-2112.2015.09.025.

[13]GRAVES A, MOHAMED A R, HINTON G. Speech recognition with deep recurrent neural networks[C]// IEEE International Conference on Acoustics, Speech and Signal Processing.Vancouver:IEEE, 2013. DOI:10.1109/ICASSP.2013.6638947.

[14]ZEILER M D. ADADELTA: An adaptive learning rate method[Z]. ArXiv preprint arXiv:1212.5701,2012.

[15]DIEDERIK K, JIMMY B. Adam: A method for stochastic optimization[Z]. ArXivPreprint ArXiv:1412.6980,2014.

[16]ILYA L, FRANK H. SGDR: Stochastic gradient descent with warm restarts[Z]. ArXivPreprint ArXiv:1608.03983, 2016.

[17]OLUWADARE O, ZHANG Y, CHENG J. A maximum likelihood algorithm for reconstructing 3D structures of human chromosomes from chromosomal contact data[J]. BMC Genomics,2018,19(1):161. DOI:10.1186/s12864-018-4546-8.

[18]DUCHI J, HAZAN E, SINGER Y. Adaptive subgradient methods for online learning and stochastic optimization[J]. Journal of Machine Learning Research, 2011, 12(7): 2121-2159. DOI:10.1109/TNN.2011.2146788.

[19]DUAN Z,ANDRONESCU M, SCHUTZ K, et al. A three-dimensional model of the yeast genome[J]. Nature,2010, 465(7296):363-367. DOI:10.1038/nature08973.

[20]VAROQUAUX N, AY F, NOBLE W S, et al. A statistical approach for inferring the 3D structure of the genome[J]. Bioinformatics, 2014, 30(12): 26-33. DOI:10.1093/bioinformatics/btu268.

[21]TRIEU T, CHENG J. MOGEN: A tool for reconstructing 3D models of genomes from chromosomal conformation capturing data[J]. Bioinformatics,2016,32(9):1286-1292. DOI:10.1093/bioinformatics/btv754.

[22]WANG S, XU J, ZENG J. Inferential modeling of 3D chromatin structure[J].Nucleic Acids Research,2015,43(8):e54.DOI:10.1093/narlgkv100.

[23]YAFFE E, TANAY A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture[J]. Nature Genetics, 2011,43(11):1059-1065. DOI:10.1038/ng.947.