DOI:10.12113/j.issn.1672-5565.201812007

基于 Hi-c 数据的酵母染色体三维结构重构

丰继华, 牟　锦, 郭亚茹
(云南民族大学 电气信息工程学院, 昆明 650500)

摘　要:通过染色体交互频率数据(Hi-c)来预测染色体三维空间结构是近年表观遗传研究热点。研究表明染色体三维空间结构在生物基因表达、调控等方面起到重要作用,对其进行三维重构是研究细胞代谢过程的基本途径。针对酵母 Hi-c 数据在不同染色体所呈现出的统计特征,拟合出每条染色体交互频率数据分布的数学模型, 然后利用梯度上升迭代算法预测并重构其三维结构,并给出模型评估指标。实验结果表明,模型具有较高可重复性和预测精确度。

关键词:Hi-c 数据;三维结构;梯度上升算法
中图分类号:S963.32+7　　**文献标志码**:A　　**文章编号**:1672-5565(2019)03-182-07

Three-dimensional structure reconstruction of yeast chromosome based on Hi-c data

FENG Jihua, MOU Jin, GUO Yaru
(*School of Electrical Information Engineering*, *Yunnan Minzu University*, *Kunming* 650500, *China*)

Abstract: The prediction of the three-dimensional spatial structure of chromosomes by chromosome interaction frequency data (Hi-c) has been a hot topic in epigenetic studies in recent years. It has been shown that the three-dimensional spatial structure of chromosome plays an important role in gene expression and regulation. Three-dimensional reconstruction is the basic way to study the process of cell metabolism. According to the statistical characteristics of yeast Hi-c data on different chromosomes, the mathematical model of the distribution of the cross-frequency data of each chromosome was fitted, and then the gradient ascending iterative algorithm was used to predict and reconstruct its three-dimensional structure. The evaluation index of the model was established. Experimental results showed that the model has high repeatability and prediction accuracy.

Keywords: Hi-c data; Three-dimensional structure; Gradient ascending algorithm

Dekker 等人于 2002 年就利用 3c 技术[1]用于重构酵母染色体的空间结构,由于受限于当时的技术条件,他们只进行了染色体区域内一对一的相互作用研究。近年来基于此技术发展出了 4C[2],5C[3],Hi-c 以及 TCC 技术[4],为染色体三维结构研究奠定了基础。4C 是将 DNA 利用连接酶连接成环状后用特异 PCR 引物进行反向 PCR,再对其产物进行分析得到染色体相互作用数据;5C 技术是在 3c 基础上增加了测序通量,主要用于研究染色体高频率的空间结构。而 Hi-c 技术[5]就是一种在 3C 基础上发展的高通量测序染色体全基因组的交互数据的生物信息学技术。随着基因组学的深入研究,人们发现染色体的相互作用数据在一定程度上可以反应基因组在三维空间的表达状况[6]。另一方面,相较于其他的 3C 衍生技术,利用 Hi-c 技术捕获到的全基因组交互数据为利用二维接触矩阵构建三维空间结构提供了可能。通过基因空间结构解析与传统转录数据相结合,研究人员可以更深入的阐释生物在基因调控以及表观遗传中的真实性状[7]。因此预测和重构染色体的三维结构对于后基因组时代研究具有指导意义。尽管目前部分染色体的组织结构可以通过电子显微镜进行研究。显微镜提供了单个细胞的信息,但分辨率相对较低;而染色体构象捕获能获得高分辨率的染色体三维信息,极大拓展了我们对基

收稿日期:2018-12-31;**修回日期**:2019-03-05.
基金项目:国家自然科学基金项目(No.31160234).
作者简介:丰继华,男,副教授,研究方向:生物信息学.E-mail:feng_jihua@126.com.

因组的全面认识。

文中使用 Duan 等人研究使用的酵母数据样本[8],根据酵母 16 条染色体 Hi-c 数据构建了数据统计分布模型,在此基础上利用梯度优化算法预测并绘制了酵母染色体的三维结构图。

1　理论与方法

利用 Hi-c 技术获得的基因组高通量染色体交互频率数据,通过特定数学模型预测基因组空间结构是重构染色体三维结构普遍采取的方法。其预测流程主要分为:Hi-c 数据归一化处理;Hi-c 数据转化为距离接触矩阵;染色体三维重构模型;和模型结果分析等。其中,Lieberman-Aiden 等人曾对染色体上两个片段的接触频率和基因线性以及空间距离做了开创性研究,发现染色体片段之间的接触频率值与两个片段之间空间的距离成反比关系,即空间越接近则接触频率值越大,空间距离越远接触频率越小[9],在此原理上提出了以下距离转换关系式:

$$D_{ij}=\begin{cases}F_{ij}^{-1}, F_{ij}>0\\ \infty, \text{其他}\end{cases} \tag{1}$$

式(1)中,D_{ij} 是表示酵母染色体上两个片段之间的通过转换的空间距离值,F_{ij} 表示酵母染色体片段间的接触频率值。

1.1　酵母染色体 Hi-c 数据分布拟合函数模型

首先,需要对根据酵母染色体交互数据建立统计分布模型,为此,分别对酵母 16 条染色体的 Hi-c 数据分布情况进行高斯拟合,对每条染色体的数据我们都分别与高斯 8 个线性组合核函数进行拟合,再最终选取出拟合指标 SSE,RMSE,R-square 最优的高斯核函数,最终选取核函数的拟合指标结果如表 1 所示。

表 1　16 条染色体拟合情况表

Table 1　Fitting of 16 chromosomes

染色体	Gaussian 核函数	SSE	RMSE	R-square
chr1	Gaussian7	0.014 60	0.013 59	0.999 9
chr2	Gaussian8	0.044 60	0.024 22	0.999 8
chr3	Gaussian6	0.083 67	0.031 94	0.999 5
chr4	Gaussian6	0.178 20	0.046 61	0.999 6
chr5	Gaussian8	0.063 94	0.029 01	0.999 7
chr6	Gaussian6	0.036 26	0.021 03	0.999 8
chr7	Gaussian8	0.013 94	0.013 54	0.999 9
chr8	Gaussian7	0.069 66	0.029 70	0.999 7
chr9	Gaussian5	0.114 70	0.036 73	0.999 4
chr10	Gaussian7	0.077 19	0.031 26	0.999 7
chr11	Gaussian6	0.079 78	0.031 19	0.999 7
chr12	Gaussian7	0.042 01	0.023 06	0.999 9
chr13	Gaussian8	0.101 10	0.03647	0.999 7
chr14	Gaussian7	0.062 22	0.028 06	0.999 8
chr15	Gaussian8	0.166 90	0.046 87	0.999 5
chr16	Gaussian7	0.064 58	0.028 59	0.999 8

在最终确定了每条染色体拟合出对应的高斯核函数后,绘制了 16 条染色体 Hi-c 数据分布的拟合曲线(见图 1)。

通过与酵母 16 条染色体交互数据分布拟合后,获得目标函数如下:

$$\sum_{n=1}^{m} a_n \exp\left(-\frac{1}{c_n^2}\left(D_i^s - D_i^2\right)\right) \tag{2}$$

式中,D_i^s 是对应于 D_i 的区域的欧几里德距离[10](i 表示染色体上片段数),是由三维结构 S 中两个染色体片段区域的(x,y,z)坐标计算得到,a_n 和 c_n 是目标函数中的拟合参数。各参数见表 2 所示。

1.2　染色体三维模型建立

在似然估计中,用 S 表示酵母染色体结构,D 表示从染色体交互数据导出的接触矩阵,似然函数 $P(D_i \mid S)$ 表示在结构 S 条件下 D 中数据点概率[11],在此,真实的 Hi-c 数据分布由拟合得到的组合高斯模型代替,因此 $P(D_i \mid S)$ 可以表示为:

$$P(D_i \mid S) \sim \sum_{n=1}^{m} a_n \exp\left(-\frac{1}{c_n^2}\left(D_i^s - D_i^2\right)\right) \tag{3}$$

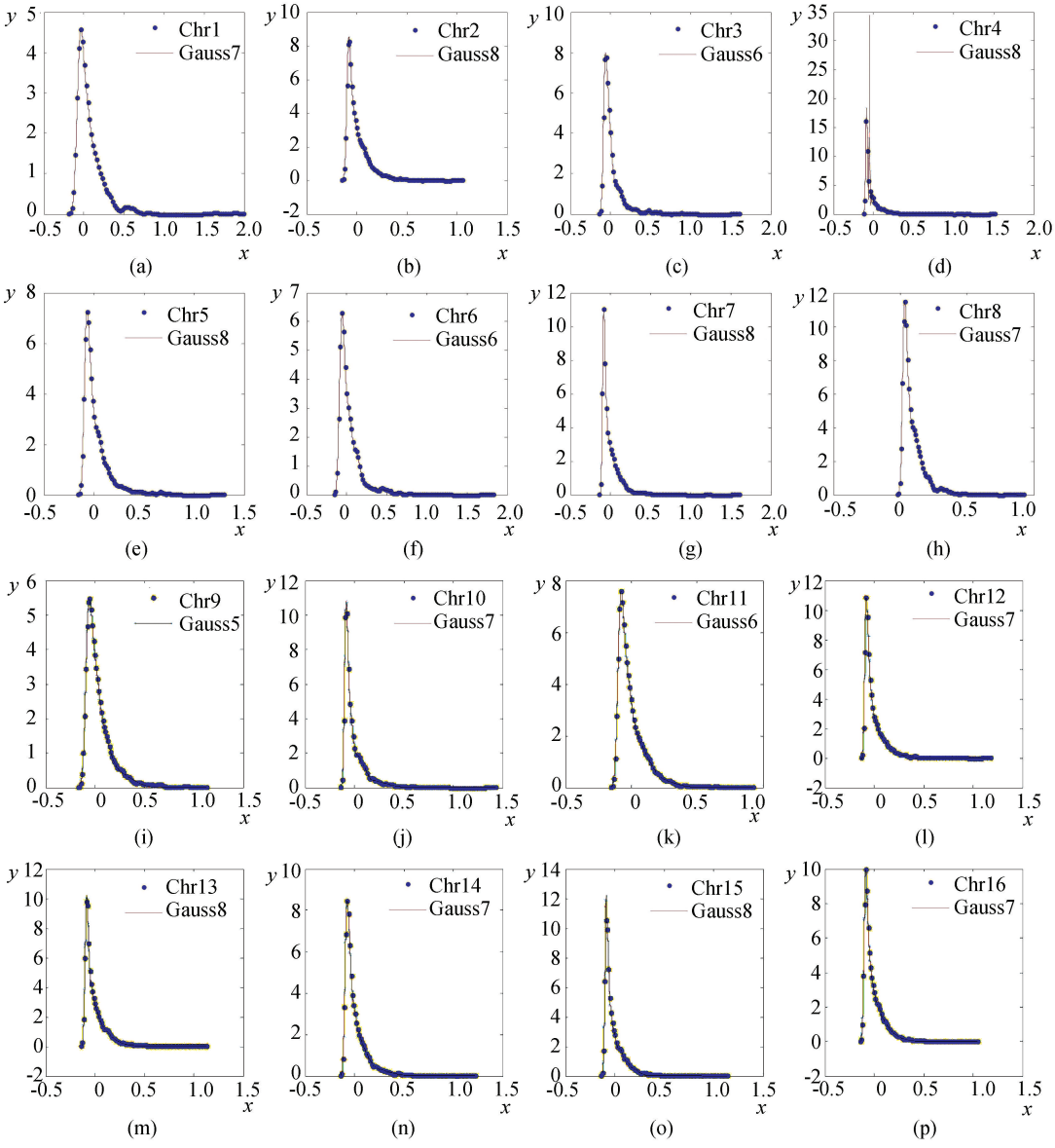

图 1 染色体 Hi-c 数据分布拟合情况图

Fig.1 Fitting map of distribution of chromosome Hi-c data

我们的目的是找到一个最大化似然函数的结构 S^*。式(3)中的目标函数仅依赖染色体结构中的 (x,y,z) 坐标。

1.3 梯度上升优化算法

利用梯度上升算法对式(3)进行迭代优化,直到算法收敛为止。具体过程:如果使用新的 (x,y,z) 坐标计算的似然函数值和前一步的差值小于一个阈值,就认为算法收敛[11]。

梯度上升迭代优化中。利用等式(3)计算偏导数,再根据偏导采用梯度上升优化算法对各坐标进行调整,并按下式更新似然概率:

$$S^{(t+1)} = S^{(t)} + \lambda^{(t)} \nabla L(S^{(t)}) \tag{4}$$

式(4)中 t 是迭代索引指标,$S^{(t)}$ 是迭代索引指标 t 的结构坐标,$\lambda^{(t)}$ 是在 t 处的学习速率,随着迭代的进行可能发生变化,$\nabla L(S^{(t)})$ 是结构中坐标的似然偏导数。式(5)表示的是在时间步长 t 处参数 S_i 的似然梯度。因此,式(4)中的随机梯度上升可以用式(6)表示。S_i 是 S 中的参数。

$$g_{t,i} = \nabla(S_i^{(t)}) \tag{5}$$

$$S_i^{(t+1)} = S_i^{(t)} + \lambda^{(t)} g_{t,i} \tag{6}$$

在 Ada Grad 的迭代规则中,根据式(7),在参数 S_i 的之前计算的梯度基础上,修正了每个参数 S_i 在每一时间步长上的学习速率。

$$S_i^{(t+1)} = S_i^{(t)} + \frac{\lambda}{\sqrt{G_{t,ii} + \varepsilon}} g_{t,i} \tag{7}$$

式中,G_t 是一个对角元素为 i 的对角矩阵,i 是 S_i 的梯度平方和,如式(8)所示。其中 $G_{t,ii}$ 是 G_t 中染色体片段 i 对应的值,而 ε 是一个平滑项,它是避免函数除以零(通常是 1×10^{-6})。

$$G_t = \sum_{i=1}^{t} (g_i g_i) \qquad (8)$$

Ada Grad 的主要优点之一是它不需要在每次迭代时手动调整学习速率。

表 2　酵母染色体 Hi-c 数据分布模型参数

Table 2　Parameters of yeast chromosome Hi-c data distribution model

染色体	n	1	2	3	4	5	6	7	8
	a_n	2.584 00	0.136 60	0.713 80	2.270 00	0.263 40	1.361 00	0.170 10	/
Chr1	b_n	−0.047 81	−0.004 417	−0.084 74	0.010 22	0.328 50	0.129 80	0.554 00	/
	c_n	0.046 68	0.028 39	0.041 40	0.077 79	0.071 33	0.139 50	0.118 90	/
	a_n	−1.329 00	6.907 00	−14.990 00	0.257 80	−0.022 28	0.871 40	1.801 00	17.990 00
Chr2	b_n	−0.103 50	−0.084 14	−0.039 26	0.004 12	−0.070 95	−0.051 57	0.000 291 3	−0.040 40
	c_n	0.098 32	0.024 67	0.026 49	0.019 29	0.004 019	0.310 90	0.108 80	0.028 28
	a_n	5.027 00	3.475 00	3.858 00	0.287 90	1.191 00	0.219 40	/	/
Chr3	b_n	−0.054 91	−0.014 60	0.079 87	0.036 27	0.076 01	0.236 20	/	/
	c_n	0.030 14	0.038 95	0.024 37	0.027 17	0.091 18	0.299 50	/	/
	a_n	18.010 00	33.550 00	0	1.506 00	0.745 60	2.258 00	/	/
chr4	b_n	−0.086 25	−0.050 79	−0.016 72	−0.001 73	0.092 01	−0.019 14	/	/
	c_n	0.015 07	0.005 481	0.000 656	0.054 69	0.158 10	0.009 90	/	/
	a_n	4.609 00	0.714 80	3.101 00	2.125 00	0.461 70	0.280 70	1.168 00	0.289 40
Chr5	b_n	−0.059 92	−0.025 85	−0.085 76	−0.011 04	0.044 20	0.071 33	0.071 49	0.215 50
	c_n	0.031 60	0.024 93	0.027 49	0.050 34	0.022 26	0.019 88	0.099 86	0.247 80
	a_n	5.670 00	36.310 00	3.134 00	−36.310 00	3.312 00	0.257 00	/	/
Chr6	b_n	−0.068 88	−0.009 226	0.044 94	−0.004 862	0.050 81	0.223 50	/	/
	c_n	0.040 54	0.064 13	0.044 30	0.069 90	0.094 92	0.365 30	/	/
	a_n	37.610 00	7.107 00	0.179 00	3.009 00	−33.120 00	0.711 10	1.009 00	0.814 60
Chr7	b_n	−0.090 28	−0.076 39	0.292 50	0.021 80	−0.090 00	0.029 31	0.063 57	0.126 40
	c_n	0.021 74	0.030 16	0.165 40	0.048 12	0.022 89	0.024 37	0.037 62	0.072 40
	a_n	3.663 00	2.570 00	2.392 00	2.418 00	0.563 50	5.042×10^{9}	0.200 30	/
Chr8	b_n	−0.065 26	−0.040 29	−0.083 21	0.033 38	0.174 20	0.089 01	0.351 40	/
	c_n	0.030 20	0.041 53	0.025 62	0.087 36	0.074 70	0.000 95	0.146 80	/
	a_n	3.196 00	1.875 00	2.607 00	1.302 00	0.498 00	/	/	/
Chr9	b_n	−0.044 98	0.008 176	−0.077 96	0.071 34	0.190 50	/	/	/
	c_n	0.042 16	0.050 92	0.034 90	0.095 13	0.222 30	/	/	/
	a_n	7.070 00	6.229 00	2.217 00	2.477 00	0.861 60	0.934 60	0.409 10	/
Chr10	b_n	−0.073 00	−0.091 79	−0.043 94	−0.019 28	0.036 15	0.080 43	0.154 20	/
	c_n	0.020 75	0.017 60	0.022 06	0.035 13	0.028 92	0.056 29	0.221 70	/
	a_n	1.032 00	3.394 00	3.199 00	4.540 00	0.618 40	0.372 60	/	/
Chr11	b_n	0.085 24	−0.062 56	−0.020 49	−0.086 44	0.044 88	0.166 50	/	/
	c_n	0.083 07	0.033 05	0.045 97	0.026 50	0.040 11	0.182 30	/	/
	a_n		5.079 00		0.334 30	−41.300 00	0	1.966×10^{11}	/
Chr12	b_n	−0.087 81	−0.076 83	−0.084 93	0.022 86	−0.096 29	−9.492 00	−5.480 00	/
	c_n	0.011 57	0.020 82	0.066 10	0.018 79	0.074 05	0.246 50	1.098 00	/
	a_n	15.050 00	−19.130 00	0.708 70	4.386 00	0.250 30	22.190 00	230.600 00	−15.990 00
Chr13	b_n	−0.089 07	−0.092 47	−0.026 57	−0.014 35	0.050 71	−0.089 06	−1.040 00	−0.106 80
	c_n	0.023 99	0.030 51	0.020 26	0.048 43	0.026 52	0.045 76	0.503 00	0.098 51
	a_n	4.384 00	5.024 00	2.214 00	−0.120 80	0.908 50	1.034 00	0.349 30	/
Chr14	b_n	−0.084 27	−0.061 02	−0.023 00	−0.130 50	0.018 52	0.073 63	0.183 90	/
	c_n	0.022 38	0.029 09	0.044 02	0.001 404	0.058 27	0.088 54	0.216 70	/
	a_n		6.840 00	3.419 00	2.256 00	0.597 10	2.458 00	−0.217 00	1 .135 00
Chr15	b_n	−0.095 73	−0.079 43	−0.057 06	−0.014 46	0.059 60	−0.088 90	0.165 00	0.086 58
	c_n	0.015 51	0.019 14	0.028 23	0.045 69	0.021 39	0.001 572	0.037 91	0.137 70
	a_n	4.423 00	4.334 00	0.840 00	8.403 00	2.760 00	0.442 50	−5.894 00	/
Chr16	b_n	−0.091 25	−0.066 14	−0.008 09	−0.348 20	−0.077 04	0.047 17	−0.145 30	/
	c_n	0.019 00	0.044 61	0.032 23	0.324 80	0.021 25	0.022 71	0.111 50	/

2 模型评价

为了评估染色体三维结构模型的准确性,我们使用 Pearson 相关系数(PCC)、Spearman 相关系数(SCC)这两个参数作为评价指标。假设两个模型的成对距离数据集,其中 $\{d_i,\ldots,d_n\}$ 有 n 个值,另一数据集 $\{D_i,\ldots,D_n\}$ 也含 n 个值,那么 DPCC、DSCC 可以使用以下公式来计算。

(1)距离 Pearson 相关系数(DPCC)

定义为:

$$\mathrm{DPCC}=\frac{\sum_{i=1}^{n}(d_i-\bar{d})(D_i-\bar{D})}{\sqrt{\sum_{i=1}^{n}(d_i-\bar{d})^2\sum_{i=1}^{n}(D_i-\bar{D})^2}} \quad (9)$$

其中,d_i 和 D_i 表示第 i 个距离样本值,n 是成对距离的个数,$\bar{d}$ 和 $\bar{D}$ 分别表示距离均值为:$\bar{d}=\frac{1}{n}\sum_{i=1}^{n}d_i$,$\bar{D}=\frac{1}{n}\sum_{i=1}^{n}D_i$。

(2)距离 Spearman 相关系数(DSCC)

定义为:

$$\mathrm{DSCC}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \quad (10)$$

DSCC 测量了两个三维结构距离剖面的相似性。DSCC 值在-1.0 和 1.0 之间变化,DSCC 值越高,这两个结构就越相似。

3 实验结果

根据酵母 16 条染色体的 Hi-c 数据建立特定的分布函数,在此基础上构建目标函数,然后利用梯度上升算法对每条染色体 Hi-c 数据目标函数进行迭代,迭代的最大次数为 2 000 次,而收敛阈值设置为 0.000 01,酵母 16 条染色体目标函数收敛曲线如图 2 所示。

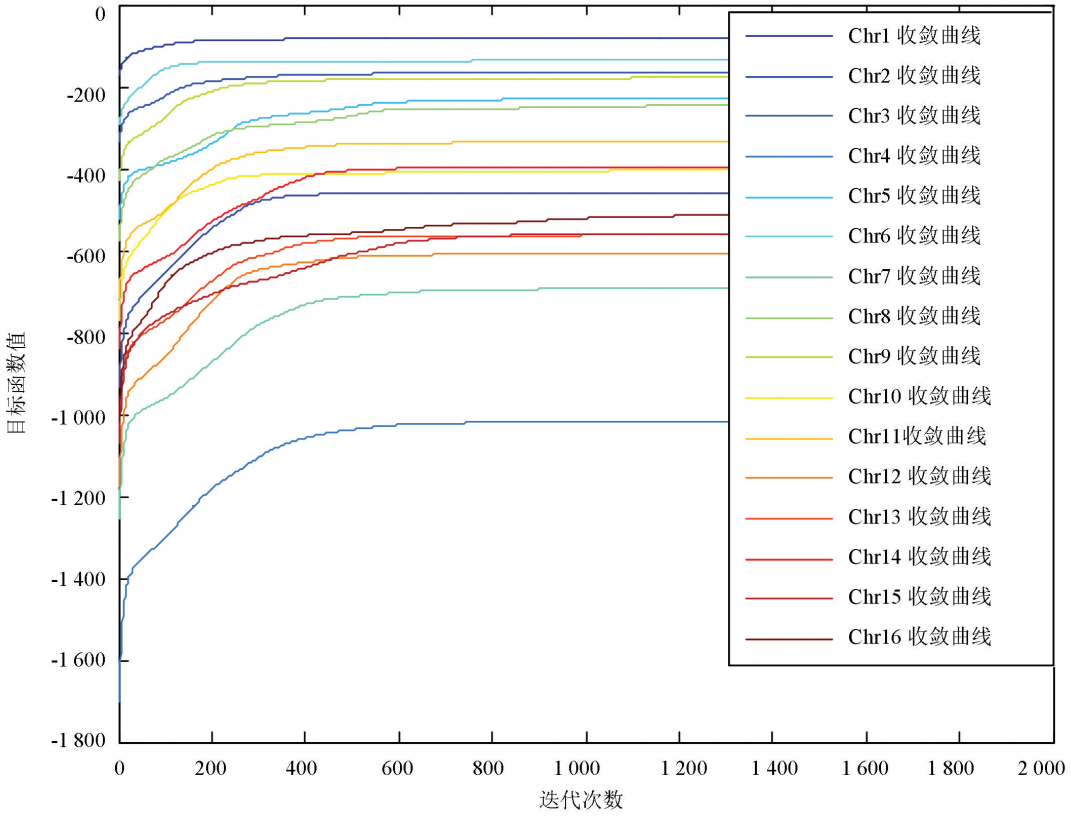

图 2 酵母 16 条染色体目标函数收敛曲线

Fig.2 Convergence curve of 16 chromosome objective functions in yeast

从图 2 中可以看出酵母 16 条染色体目标函数最终都达到收敛,说明该目标函数模型是有效可行的。文中使用梯度上升优化算法对酵母 16 条染色体数据进行三维空间结构重构,其模型的评价指标如表 3 所示。

表 3　16 条染色体结构 Spearman 和 Pearson 系数

Table 3　Spearman and Pearson coefficients of 16 chromosomal structures

系数	Chr1	Chr2	Chr3	Chr4	Chr5	Chr6	Chr7	Chr8	Chr9	Chr10	Chr11	Chr12	Chr13	Chr14	Chr15	Chr16
Spearman	0.961 5	0.997 1	0.972 9	0.998 0	0.994 0	0.965 2	0.997 9	0.995 7	0.992 4	0.995 0	0.996 3	0.996 9	0.996 8	0.997 1	0.998 0	0.998 1
Pearson	0.687 9	0.716 7	0.680 7	0.732 5	0.680 7	0.642 7	0.734 2	0.698 2	0.741 8	0.701 8	0.751 0	0.710 5	0.743 7	0.700 1	0.776 3	0.746 3

从表中可以看出,经过对每条染色体的 Hi-c 数据分布特征进行拟合出具体函数作为目标函数的模型的 Spearman 系数都达到了 0.95 以上,Pearson 系数平均值也能达到 0.71 以上,说明对每条染色体进行 Hi-c 数据分布特征进行拟合出不同目标函数的方法来预测其结构是有效可行的。通过不同目标函数模型预测出的染色体结构如图 3 所示。

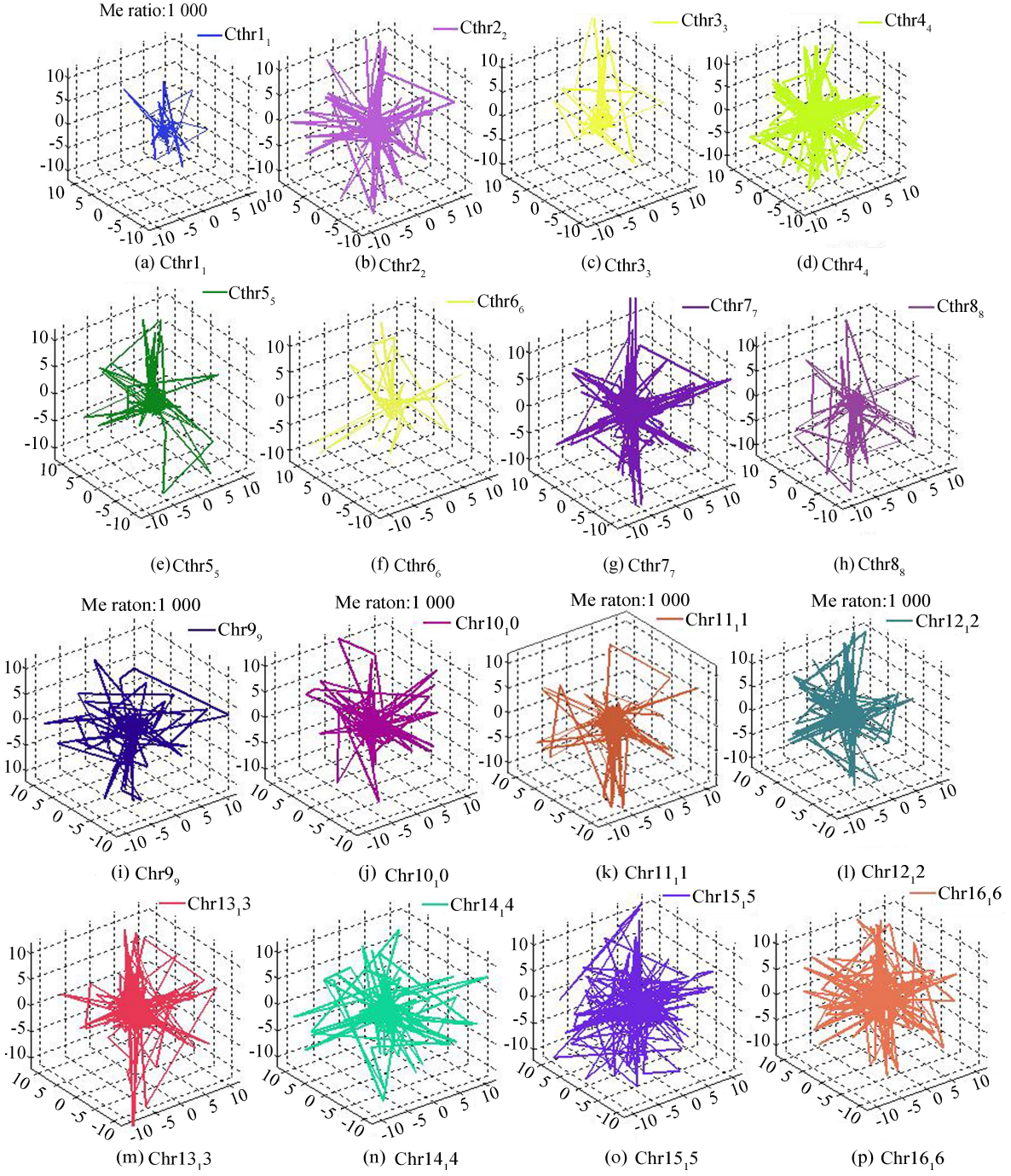

图 3　16 条染色体结构图

Fig.3　Charts of 16 chromosomal structures

4 结论与展望

目前,利用染色体 Hi-c 交互数据预测三维空间结构大多根据单一分布模型,并没有考虑每条染色体数据的具体分布情况,而本文通过分析酵母 16 条染色体 Hi-c 数据的实际分布,从而拟合出更真实的分布模型,在此基础上利用梯度优化算法预测出较准确的染色体三维结构。但是为了分析方便,在对酵母 16 条染色体 Hi-c 数据进行距离转换时使用了统一的参数,后续我们将会针对具体染色体不同数据对转换函数的参数进行优化,增强模型的自适应性,从而进一步提高模型预测的准确性。

参考文献(References)

[1] DEKKER J, RIPPE K, DEKKER M, et al. Capturing chromosome conformation[J]. Science, 2002, 295(5558): 1306-1311. DOI: 10.1126/science.1067799.

[2] ZHAO Z, TAVOOSIDANA G. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and inter-chromosomal interactions [J]. Nature Genetics, 2006, 38(11): 1341-7. DOI: 10.1038/ng1891.

[3] DOSTIE J, DEKKER J. Mapping networks of physical interactions between genomic elements using 5C technology[J]. Nature Protocols, 2007, 2(4): 988-1002. DOI: 10.1038/nprot.2007.116.

[4] KALHOR R, TJONG H, JAYATHILAKA N, et al. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling[J]. Nature Biotechnology, 2011, 30(1): 90-98. DOI: 10.1038/nbt.2057.

[5] NAGANO T, LUBLING Y, STEVENS T J, et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure[J]. Nature, 2013, 502(7469): 59. DOI: 10.1038/nature12593.

[6] ALT F W, ZHANG Y, MENG F L, et al. Mechanisms of programmed DNA lesions and genomic instability in the immune system[J]. Cell, 2013, 152(3): 417-429. DOI: 10.1016/j.cell.2013.01.007.

[7] DEKKER J. Gene regulation in the third dimension [J]. Science, 2008, 319(5871): 1793-1794. DOI: 10.1126/science.1152850.

[8] DUAN Z, ANDRONESCU M, SCHUTZ K, et al. A three-dimensional model of the yeast genome[J]. Nature, 2010, 465(7296): 363-367. DOI: 10.1007/978-3-642-20036-6_28.

[9] LIEBERMAN-AIDEN E, VAN BERKUM N L, WILLIAMS L, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome [J]. Science, 2009, 326(5950): 289-293. DOI: 10.1126/science.1181369.

[10] DEZA M M, DEZA E. Encyclopedia of distances [J]. Reference Reviews, 2009, 24(6): 1-583. DOI: 10.1007/978-3-642-00234-2_19.

[11] 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.