计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190900065

使用深层卷积神经网络提高 Hi-C 数据分辨率

程 哲 白 茜 张 浩 王世普 梁 宇
云南大学软件学院 昆明 650000
(chengzhe_xg@foxmail.com)

摘 要 Hi-C 技术是一种测量整个基因组中所有成对交互的频率的技术,已成为研究基因组 3D 结构最流行的工具之一。通常情况下,基于 Hi-C 数据的研究需要测序大量的染色体数据,而测序深度较低的 Hi-C 数据虽然成本较低,但不足以提供充足的生物学信息给后续研究。由于 Hi-C 数据包含了类似的子模式,且一定区域内具有数据连续性,因此可以被预测。文中探究了基于卷积神经网络模型的改进方法,该模型以更大的范围预测核心的 Hi-C 数值,并扩展卷积神经网络的深度和感受野,通过 1/16 的原始测序读数,预测出 Hi-C 数据的原始测序读数。实验结果以皮尔森相关系数和斯皮尔曼相关系数衡量,并使用 Fit-Hi-C 分析明显的相互作用对,以及通过调用 ChromHMM 标记的染色质状态区域进行染色质状态分析。实验结果表明,预测结果不仅在数值分布规律上接近,而且在位点互作信息和染色质状态等方面也比低分辨率 Hi-C 数据更加可靠。

关键词:Hi-C 技术;超分辨率;卷积神经网络;生物信息学;深度学习
中图法分类号 TP391

Improving Hi-C Data Resolution with Deep Convolutional Neural Networks

CHENG Zhe, BAI Qian, ZHANG Hao, WANG Shi-pu and LIANG Yu
School of Software, Yunnan University, Kunming 650000, China

Abstract Hi-C technology measures the frequency of all paired-interaction in the entire genome. It has become one of the most popular tools for studying the 3D structure of genomes. In general, Hi-C data-based studies require sequencing of a large number of Chromosome data, while Hi-C data with lower sequencing depth, although less expensive, is not sufficient to provide sufficient biological information for subsequent studies. Since the Hi-C data contains similar sub-patterns and has data continuity within a certain area, it can be predicted. This paper explored an improved method based on convolutional neural network model. It predicts the core Hi-C values in a larger range and extends the depth and receptive field of the convolutional neural network, predicts the original sequencing reading of Hi-C by 1/16 of the original sequencing readings. The experimental results were measured by the Pearson correlation coefficient and the Spearman correlation coefficient, and the apparent interaction pairs were analyzed using Fit-Hi-C, and the state analyses of 12 chrom HMM-marked chromatin with ChromHMM were called. The experimental results show that the prediction results are not only close to the numerical distribution, but also more reliable than the low-resolution Hi-C data in terms of site interaction information and chromatin state.

Keywords Hi-C technology, Super-resolution, Convolutional neural network, Bioinformatics, Deep learning

1 引言

全面了解基因组结构和功能之间的关系是一项重要但极其困难的任务,新开发的生物化学方法(如 3C,4C,5C,Hi-C 和 ChIA-PET)已被用于探索物理相互作用频率,这些物理相互作用频率被定义为一对染色体基因座在大细胞群体中相互作用的概率。但是这种方法极大程度地依赖于化学实验,从而导致实验复杂性和实验的消耗非常大。此外,目前的技术仅限于测量高阶染色体组织,同时实现更高的分辨率。对 3D 基因组组织和功能的详细和可理解的描述将需要开发能够在不同物理尺度上揭示这种复杂层级组织的计算技术。

高通量染色体构象捕获(Hi-C)技术[1]是一种测量整个基因组中所有成对交互的频率的技术,可以更清晰地观察染色体的空间构象[2,3]。Hi-C 技术促进了人类对染色体功能和基因表达调节的新认识,同时启发了对染色体在细胞发育和疾病发生过程中的基因调控机制的新认识[4]。Hi-C 技术的快速发展,极大地提升了人们对染色体 3D 结构的认识,促进了一系列重大发现:从染色质环[5]、拓扑相关结构域(TADs)[6]、A/B 隔室[1],一直到完整的多染色体 3D 结构。染色体不同层次结构的有序性,是细胞正常活动的基础,因此这些 3D 结构的紊乱,是多种疾病发生的重要原因。

Hi-C 数据通常表示为 $n\times n$ 的邻接矩阵,其中基因组被分成 n 个相等的区间,并且矩阵的每个网格内的值表示这个区间内的读数的数目。根据测序深度的不同,这些网格尺寸

本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61762089,91631305,61863036)
This work was supported by the National Natural Science Foundation of China (61762089,91631305,61863036).
通信作者:梁宇(yuliang@ynu.edu.cn)

区间通常为[1 kb，1 mb]。Hi-C 相互作用矩阵的网格大小也称为“分辨率”，是 Hi-C 数据分析中最重要的参数之一，因为它直接影响基于 Hi-C 数据的分析结果，如预测增强子-启动子相互作用或识别 TAD 边界。测序深度是决定 Hi-C 数据分辨率的最关键因素，测序深度越高，互作信息越丰富，可用的分辨率就越高（箱尺寸越小）。

高分辨率的 Hi-C 数据测序成本巨大，需要数百万哺乳动物细胞，并且涉及大量的测序成本，因为分辨率的线性增加需要测序读数呈指数增加[4]，所以大部分常用 Hi-C 的分辨率比较低（一般为 25 kb 或 40 kb）。这些较低分辨率的 Hi-C 数据集可被用来识别大规模的基因组模式（A/B 区室或 TAD 域），但不能用来识别更加精细的结构（例如增强子-启动子相互作用）。因此，高分辨率 Hi-C 数据的产生成为了主流的研究热点，尤其在低成本 HI-C 数据的基础上，生成更加精细的高分辨率 Hi-C 交互矩阵，研究意义更大。

近年来，深度学习在多个领域中取得了巨大的成功，特别是卷积神经网络（ConvNet），在计算机视觉和自然语言处理等方面取得了重大的突破[7]。在生物信息学研究中，ConvNet 已经被成功地用于 DNA 序列预测、DNA 甲基化[8]或基因表达模式潜在功能的挖掘。

受 ConvNet 优秀表现的启发，本文提出一种改进的卷积神经网络模型方案，如图 1 所示。该模型的目的在于通过低分辨率（低测序读数）Hi-C 数据预测出具有更好品质的高分辨率 Hi-C 相互作用矩阵，并且该模型预测出的 Hi-C 数据具有与高分辨率数据相似的生物学信息和模式。总而言之，本文提出的卷积神经网络模型提供了一种更好的产生高分辨率 Hi-C 相互作用矩阵数据的方式，为染色质 3D 结构和相互作用的研究提供了更好的资源和基础。

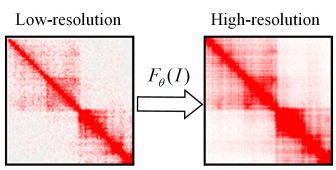


图 1　高分辨率 HI-C 数据生成示意图

Fig. 1　Schematic diagram of high-resolution HI-C data generation

本文第 2 节介绍相关工作；第 3 节阐述了网络结构和相应的算法；第 4 节详细介绍了本文实验的相关环境和实验结果；最后总结全文。

2　相关工作

高分辨率 Hi-C 数据生成的原理主要是通过提取低分辨率 Hi-C 数据中的关键特征信息，预测生成高分辨率 Hi-C 数据。基于 bicubic，Lanczos[9]，或者高斯平滑的方法可以快速得到相应效果，但这些线性方法过于简化了问题，导致产生的结果过于平滑，与真实 Hi-C 数据生物意义差距较大。

更优秀的方法旨在通过学习的方式获得低分辨率和高分辨率图像之间的复杂映射关系。许多基于学习的方法将低分辨率作为训练数据，高分辨率作为已知目标，邻域嵌入方法通过在低维特征中找到相似的低分辨率切片训练数据，并组合相应的高分辨率切片进行重建，产生高分辨率图像[10-11]。这类回归问题也可以使用随机森林解决[12]，因此 Dai 等开发了基于稀疏编码的显式训练方法[13]，在训练期间学习了许多切片的回归量，并使用 KNN 为给定的低分辨率切片选择最合适的回归量。

在计算机视觉研究方面，深度卷积神经网络表现出了优异的性能。SRCNN[14]首先将深度学习应用在超分辨率重建工作上，该方法仅使用了 3 个卷积层，就可以达到远超传统方法的效果。这 3 个卷积的结构分别代表了对数据学习的 3 个步骤：模式提取和特征表示、特征非线性映射、利用预测结果重建染色体的 Hi-C 数据。在对 Hi-C 数据超分辨率问题的研究中，由于 Hi-C 邻接矩阵包括可重复模式[5,6]特征，因此可以通过机器学习算法从这些模式中学习揭示低分辨率 Hi-C 数据中不明显的模式，然后通过数据重建的算法探索更深层次的生物潜在信息。为了重建图像之间的复杂映射，Zhang 等[15]首次将深度学习的方法应用到 Hi-C 数据推断高分辨率作用矩阵上，提出了 HiCPlus 模型。HiCPlus 模型的主要原理是基于周围染色质相互作用矩阵分布预测 Hi-C 接触矩阵中的数值。HiCPlus 的成功，体现了深度学习方法在 Hi-C 数据超分辨率问题上的优越前景。但是 HiCPlus[15]模型仍有待改进，它的图像裁切方式使得训练数据过小，网络深度不足，不能很好地完成更加复杂的映射。虽然文中提到在预测具体位点 Hi-C 数值时周围数据的采样范围并不影响预测结果，但是在整条染色体的大范围 Hi-C 数据的预测问题上，输入数据大小仍有着较大的影响，因此可以考虑改进邻接矩阵的裁切方法，并使用更深的网络结构。

3　方法

3.1　数据预处理和 Hi-C 矩阵生成

本文实验数据集来源于 Rao 等[5]提供的高分辨率 Hi-C 配对末端读数 GSE63525，该数据集被定位到 8 种不同细胞类型的细胞系中。实验中主要测试了 GSE63525 数据集中的 GM12878 和 K562 细胞系。对于这 2 种细胞系 Hi-C 数据，首先对所有的配对末端读数进行遍历，统计每条染色体对应的配对末端读数，同时将 Hi-C 的最大值限制为 0～100，因为过大的 Hi-C 值对识别模式并没有较大帮助，反而会影响网络的训练。然后把这些数据根据位点信息填入相应的 Hi-C 矩阵框内，从而产生真正的高分辨率 Hi-C 邻接矩阵。对于低分辨率数据，本文通过在高分辨率数据下采样获得，下采样率为 1/16，这意味着只需要随机采样所有读数的 1/16，再通过类似的双端读数匹配方式生成低清 Hi-C 数据。通过下采样测序读数产生的数据分辨率与高清数据具有相同的大小和分辨率，但是包含的信息仅有原数据的 1/16，常常以更低的分辨率表示，这与对真实测序低分辨率的数据包含的信息极其类似。

在训练前期准备中，由于单个染色体的 Hi-C 矩阵过于庞大，难以使用卷积神经网络进行处理，因此实验前需要将 10k 分辨率 Hi-C 矩阵进行切分，将每个子矩阵增大到 100×100，横纵步长均为 25，每个子矩阵被视作一个样本，同时将低清样本和高清样本对应起来。由于 Hi-C 数据大部分集中在 1Mb 的范围内，在这个范围之外的数据过于稀疏，将额外数据投入训练没有益处，因此本文只研究两个基因位点之间的基因组距离低于 2 Mb 的数据。本文对 GM12878 细胞系中的 1－17 号染色体的低分辨率和高分辨率 Hi-C 数据进行裁剪和随机下采样，获得 17 332 对训练数据。

3.2 网络模型结构

结合 Hi-C 数据高分辨率的特定需求，本文使用更深的卷积网络结构，如图 2 所示。整个网络模型包含 10 个卷积层，其中输入层卷积核大小为 9×9×64，步长为 1；中间 8 个卷积层卷积核大小均为 3×3×64，步长为 1；输出层使用 9×9×1，步长为 1 的卷积核将维度降为 100×100×1。模型采用零填充的方式来保证卷积输出尺寸与原尺寸相等，同时为了提升网络模型的非线性表达能力，在模型中添加了 RELU 激活函数。由于网络是完全卷积的，因此可以处理任意尺寸的 Hi-C 数据矩阵。

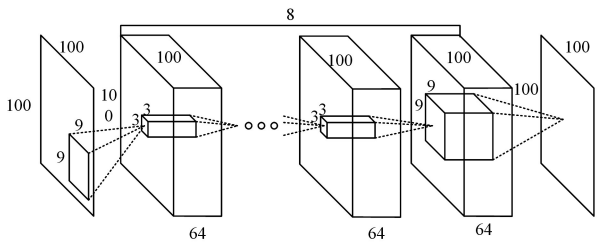


图 2 卷积神经网络模型架构

Fig. 2 Convolutional neural network model architecture

该卷积神经网络模型的目标是通过提取低分辨率输入数据 I^{LR} 潜在的特征，最终准确预测出高分辨率 Hi-C 矩阵。网络损失函数 l^{SR} 的定义对整个训练过程至关重要。均方误差(Mean-Square Error，MSE)是机器学习领域最常用的损失函数之一，计算方式如下：

$$l_{MSE}^{SR}=\frac{1}{r^2WH}\sum_{x=1}^{rW}\sum_{y=1}^{rH}(I_{x,y}^{HR}-F_\theta(I^{LR})_{x,y})^2 \tag{1}$$

通过训练后的卷积神经网络预测高分辨率子矩阵 $F_\theta(I_n^{LR})$，需要再按照切分时的相同步长进行重组。由于该网络输出与输入尺寸相同(100×100)，为了避免边缘效应，重建 Hi-C 数据时只取核心的 1/4，再将 $F_\theta(I_n^{LR})$ 按照索引合并到对应的染色体 Hi-C 矩阵中。

4 实验与结果分析

4.1 训练细节与参数

为了便于实验结果的分析比较，本文采用与 HiCPlus 相同的 GM12878 细胞系的 1－17 号染色体作为训练集，18－22 号染色体作为测试数据集，通过随机下采样的方法共生成 17 332 对子矩阵训练样本。模型的训练在 Nvidia Quadro P5000 GPU(显存 16 GB)上完成。优化函数采用 Adam，其中 $\beta=0.9$，初始学习率为 1×10^{-4}，每隔 100 次迭代，学习率降低 1/10，每个批次读入 16 对高清-低清图像组合进行训练，训练过程可以在大约 300 个 epoch 收敛，且未出现过拟合。

4.2 效果评估

本文在 GM12878 细胞系的其余 Hi-C 数据集和 k562 细胞系的全部 Hi-C 数据上测试卷积神经网络。采用了皮尔森系数和斯皮尔曼相关系数对比网络表现，以度量生成 Hi-C 数据和高清数据之间的线性相关性。一定基因位点距离上的线性相关性强弱可以体现数据分布的相似性。由于 Hi-C 数据的特殊性，相对于像素具体数值差距较大的情况，对比线性相关性强弱更为可靠。如图 3 所示，首先在相同的细胞类型即 GM12878 细胞系上进行网络表现的评估，可以观察到该网络表现普遍优于 HiCPlus，其中随机选择了染色体 18 展示网络表现。图中展示的分别是低分辨率(下采样率 1/16)、HiCPlus 增强和该网络生成数据对比真实高分辨率 Hi-C 数据的皮尔森系数和斯皮尔曼系数。可以发现，不仅该网络优于 HiCPlus 方法，且二者生成的数据明显比低分辨率 HI-C 矩阵更好，证明超分辨率方法有效提升了 Hi-C 数据可可用性。

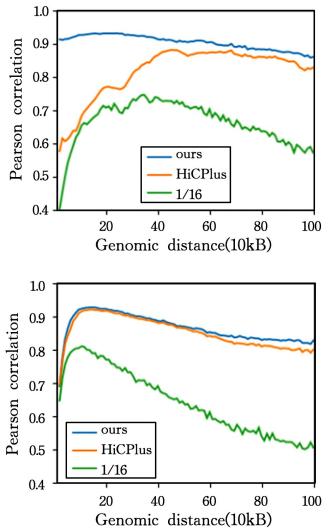


图 3 GM12878 chr18 皮尔森相关系数和斯皮尔曼相关系数对比

Fig. 3 Comparison of Pearson correlation coefficients and Spearman correlation coefficients of GM12878 chr18

其后进行了更多的测试来验证该模型的可移植性，在 K562 数据集上应用训练好的模型来增强低分辨率 Hi-C 数据，以验证训练完成的模型是否能用于增强其他细胞类型。随机选择了染色体 15 的对比结果，如图 4 所示。可以看出，我们的网络在跨细胞系中的表现依然优于 HiCPlus。

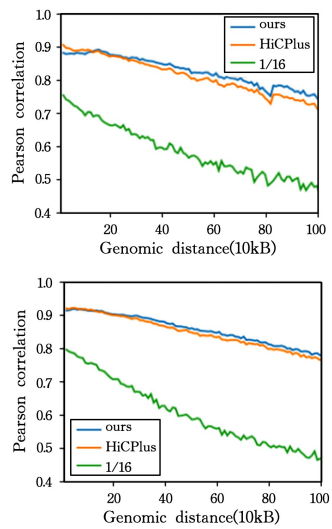


图 4 跨细胞系的 K562 chr15 皮尔森相关系数和斯皮尔曼相关系数对比

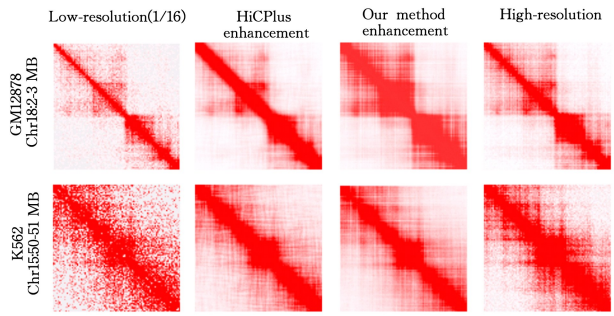
Fig. 4 Comparison of Pearson correlation coefficients and Spearman correlation coefficients of K562 chr15

4.3 在生物信息上的统计评估

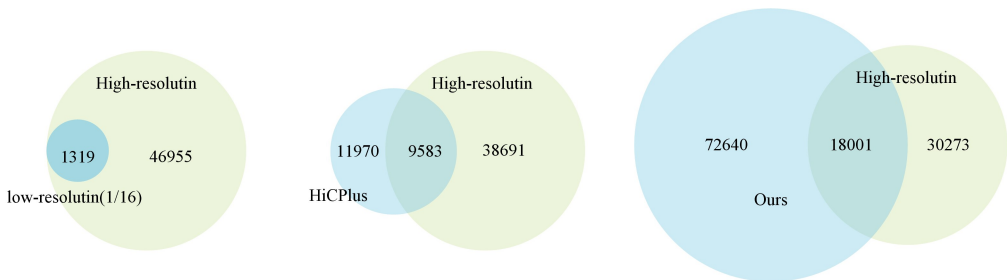
为了评估生成数据与真实高分辨率 Hi-C 数据的生物意义相似性，本文使用 Fit-Hi-C[16] 在低分辨率、HiCPlus 增强数据以及本文的预测结果和真实高分辨率的 Hi-C 数据上统计

了明显的相互作用对，这里取 $qvalue < 1 \times 10^{-6}$。对 GM12878 的 18 号染色体的评估结果如图 4 所示，取基因位点距离为 50 kb 到 2 Mb 范围内的相互作用对。在低分辨率 Hi-C 数据、HiCPlus 增强数据、本文的预测结果和真实高分辨率的 Hi-C 邻接矩阵 4 种数据集中分别检测出了 1 349，21 570，90 641，48 274 个显著的相互作用对；同时，低分辨率 Hi-C 数据、HiCPlus 增强数据和本文的预测结果 3 个数据集与高分辨率 Hi-C 矩阵之间分别有 1 349，9 583，18 001 个公共的相互作用对，可见该网络模型的表现依然优于已有方法。

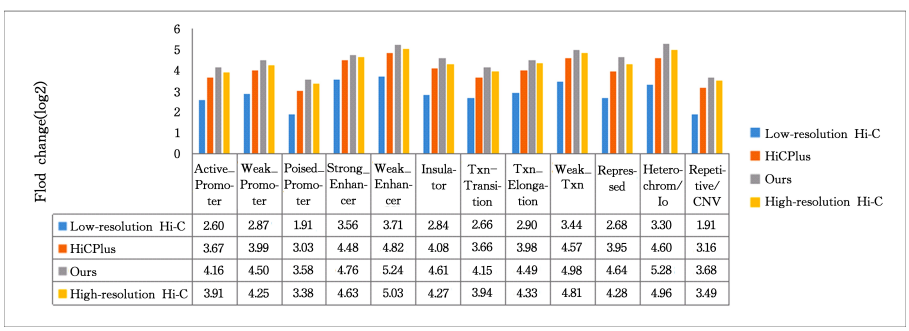
此外，本文还比较了 ChromHMM[19] 中标记的 12 种染色质状态的富集谱，分别在低分辨率(1/16)、HiCPlus 增强、本文的预测结果和真实高分辨率数据上统计检测到的显著相互作用对应的染色质状态。通过这种方式，能更好地观察 Hi-C 数据包含的染色质状态富集模式，从而分析生物信息相似性。从图 5 中可以看出，本文的预测结果可以更好地拟合真实高分辨率邻接矩阵的染色质状态。



(a)高分辨率 Hi-C 热图示例对比展示



(b)对 GM12878 染色体 18 的基于 Fit-Hi-C 的显著相互作用统计分析



	Active_Promoter	Weak_Promoter	Poised_Promoter	Strong_Enhancer	Weak_Enhancer	Insulator	Txn_Transition	Txn_Elongation	Weak_Txn	Repressed	Heterochrom/lo	Repetitive/CNV
Low-resolution Hi-C	2.60	2.87	1.91	3.56	3.71	2.84	2.66	2.90	3.44	2.68	3.30	1.91
HiCPlus	3.67	3.99	3.03	4.48	4.82	4.08	3.66	3.98	4.57	3.95	4.60	3.16
Ours	4.16	4.50	3.58	4.76	5.24	4.61	4.15	4.49	4.98	4.64	5.28	3.68
High-resolution Hi-C	3.91	4.25	3.38	4.63	5.03	4.27	3.94	4.33	4.81	4.28	4.96	3.49

(c)基于 ChromHMM 标记的染色质状态分析

图 5　对 Hi-C 数据高分辨率处理结果的热图分析以及生物信息学分析

Fig. 5　Heat map analysis and bioinformatics analysis of Hi-C data high-resolution processing results

结束语　本文探索了增强 Hi-C 超分辨率的深度学习算法，通过增大输入数据大小以及加深卷积网络的深度，在 Hi-C 数据裁切阶段使用 100×100 的更大范围来预测核心部分 Hi-C 数据，并且堆叠高达 10 层的卷积神经网络，使网络感受野更大；使用 1/10 随机采样训练数据，以更短的训练时间获得更精确的神经网络模型。该框架利用具有较少测序读数的低分辨率 Hi-C 数据构建具有更高质量和包含更充分生物信息的高清相互作用矩阵，这种预测可以跨组织/细胞类型。Hi-C 数据共有的局部模式可以通过网络有效地捕获，用于增强不同细胞类型的 Hi-C 矩阵。可以得出结论，与先进的 HiCPlus 相比，通过增大输入数据的尺寸，加深卷积神经网络的深度，可以有效增强算法的表现，达到较为优秀的性能。

参 考 文 献

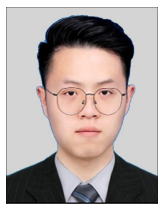
[1] LIEBERMAN-AIDEN E，VAN BERKUM N L，LOUISE V，et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome[J]. Science，2009，326：289-293.

[2] HU M，DENG K，QIN Z H，et al. Bayesian inference of spatial organizations of chromosomes[J]. PLoS computational biology，2013，9(1)：e1002893.

[3] VAROQUAUX N，FERHAT A，STAFFORD N W，et al. A statistical approach for inferring the 3D structure of the genome[J]. Bioinformatics，2014，30(12)：i26-i31.

[4] SCHMITT A D，HU M，JUNG I，et al. A compendium of chromatin contact maps reveals spatially active regions in the human

genome[J]. Cell Rep., 2016, 17: 2042-2059.
[5] 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[J]. Cell, 2014, 159(7): 1665-1680.
[6] DIXON J R, SIDDARTH S, YUE F, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions[J]. Nature, 2012, 485(7398): 376-380.
[7] HAYAT K. Multimedia super-resolution via deep learning: A survey[J]. Digital Signal Processing, 2018.
[8] WANG Y H, LIU T, XU D, et al. Predicting DNA Methylation State of CpG Dinucleotide Using Genome Topological Features and Deep Networks[J]. Scientific Reports, 2016, 6: 19598.
[9] DUCHON C E. Lanczos Filtering in One and Two Dimensions [J]. Journal of Applied Meteorology, 1979, 18(8): 1016-1022.
[10] FREEMAN W T, PASZTOR E C, OWEN T, et al. Learning Low-Level Vision [J]. International Journal of Computer Vision, 2000, 40: 2000.
[11] FREEMAN W T, JONES T R, PASZTOR E C. Example-based superresolution[J]. Computer Graphics and Applications, 2002, 22(2): 56-65.
[12] SCHULTER S, LEISTNER C, BLSCHOF H. Fast and accurate image upscaling with super-resolution forests[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015: 3791-3799.
[13] DAI D, TIMOFTE R, VAN GOOL L. Jointly optimized regressors for image super-resolution[C]// Eurographics. 2015: 8.
[14] DONG C, LOY C C, HE K, et al. Learning a Deep Convolutional Network for Image Super-Resolution[D]. Cham: Springer International Publishing, 2014: 184-199.
[15] ZHANG Y, AN L, XU J, et al. Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus [J]. Nature Communications, 2018, 9(1): 750.
[16] AY F, BAILEY T L, NOBLE W S. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts[J]. Genome Res., 2014, 24: 999-1011.
[17] ERNST J, KELLIS M. ChromHMM: automating chromatin-state discovery and characterization[J]. Nat. Methods, 2012, 9: 215-216.



CHENG Zhe, born in 1994, postgraduate. His main research interests include deep learning, computer vision and bioinformatics.

LIANG Yu, born in 1968, postgraduate, professor, Ph. D supervisor. His main research interests include computer networks, software-defined networks and cloud computing.

(上接第 69 页)

Journal of Computers, 2017, 40(6): 1440-1452.
[12] QU X M, YAO H Y, WANG Y G, et al. Research on Adaptive Control Strategy Based on Effective Green Light Time Utilization [J]. Transportation Research, 2015(1): 54-58.
[13] SUN D H, YANG C C, LIAO X Y, et al. Timing parameter estimation of intersection signals based on GPS data of public transportation [J]. Control and Decision, 2018, 33(4): 724-730.
[14] XIA X H. Urban traffic signal timing decision-making under interactive coordination reinforcement learning [J]. Computer Engineering and Applications, 2018, 54(11): 265-270.
[15] RONG H G, HUO S X, HU C H, et al. Collaborative filtering recommendation algorithm based on user similarity [J]. Journal of Communications, 2014, 35(2): 16-24.
[16] CHEN H Y, LIU C H, SUN B. A summary of the similarity measure of time series data mining [J]. Control and Decision, 2017, 32(1): 1-11.
[17] KONG X X, SU B C, WANG H Z, et al. Research on recommendation model and algorithm based on label weight scoring [J]. Journal of Computers, 2017, 40(6): 1440-1452.
[18] PAN Y T, HE F Z, YU H P. A social recommendation algorithm based on the implicit similarity of trust relationships [J]. Journal of Computers, 2018, 41(1): 65-81.

LUO Jia-lei, born in 1995, master. His main research interests include smart transportation.

MENG Li-min, born in 1963, Ph.D, professor, Ph. D supervisor, is member of China Computer Federation. Her main research interests include wireless communication and network, streaming media transmission and IoT communications.