Multi-levels 3D Chromatin Interactions Prediction Using Epigenomic Profiles

Ziad Al Bkhetan[1,2,5,6] and Dariusz Plewczynski[1,3,4,7(✉)]

[1] Centre of New Technologies, Warsaw University, Warsaw, Poland
{z.albkhetan,d.plewczynski}@cent.uw.edu.pl
[2] Biology Department, Warsaw University, Warsaw, Poland
[3] Faculty of Pharmacy, Medical University of Warsaw, Warsaw, Poland
[4] Centre for Innovative Research, Medical University of Bialystok, Białystok, Poland
[5] Department of Computing and Information Systems, The University of Melbourne, Parkville, VIC, Australia
[6] Centre for Neural Engineering, The University of Melbourne, Parkville, VIC, Australia
[7] Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland

Abstract. Identification of the higher-order genome organization has become a critical issue for better understanding of how one dimensional genomic information is being translated into biological functions. In this study, we present a supervised approach based on Random Forest classifier to predict genome-wide three-dimensional chromatin interactions in human cell lines using 1D epigenomics profiles. At the first level of our in silico procedure we build a large collection of machine learning predictors, each one targets single topologically associating domain (TAD). The results are collected and genome-wide prediction is performed at the second level of multi-scale statistical learning model. Initial tests show promising results confirming the previously reported studies. Results were compared with Hi-C and ChIA-PET experimental data to evaluate the quality of the predictors. The system achieved 0.9 for the area under ROC curve, and 0.86–0.89 for accuracy, sensitivity and specificity.

Keywords: 3D genome organization · 3D chromatin interactions · Epigenomics · Hi-C · ChIA-PET · CTCF motifs · Statistical learning · Physical interactions · Chromatin looping · Topologically associated domains · Chromatin contact domains · Random Forest

1 Introduction

The DNA of the mammalian cell is well organized together with histone proteins forming nucleosomes. It is packed into 10 nm chromatin fiber, that allows approximately two meters of human DNA [1] to be fitted in 6–10 μm diameter nucleus [2]. The biological function of a genome is influenced by major factors: one dimensional genomic data such as the DNA sequence of coding (genes) and non-coding (regulatory motifs) regions, further modifications of chromatin described as epigenomics; and

M. Kryszkiewicz et al. (Eds.): ISMIS 2017, LNAI 10352, pp. 19–28, 2017.
DOI: 10.1007/978-3-319-60438-1_2

finally the three-dimensional structure of chromatin at the scale of loops, domains, chromosomal territories and compartments. In this contribution, we focus on the association between 1D epigenomic information and 3D interactions. Epigenetics sequence annotation could be categorized into DNA methylation and the post-translational modifications of the histone proteins (i.e., histone modifications) [4]. Epigenetics play a major role in gene activity regulation either by up-regulation, or down-regulation of the associated genes [5], while the genome 3D organization can actively organize the physical interaction of DNA regulatory elements (e.g. enhancers, promoters, etc.). The three-dimensional structure is defined by Chromatin Interactions (CI), similarly like in the case of proteins their contact maps determine their spatial conformation [6]. Detailed analysis of chromatin interactions that define the three-dimensional structure of the genome is still an open question that requires further investigations.

Hi-C experiments can detect chromatin interactions mediated by any protein but it requires high sequencing depth to obtain accurate results [3]. Chromatin Interaction Analysis by Paired-End Tag sequencing (ChIA-PET) experiments are used to detect subset of interactions mediated by specific proteins, such as CTCF and RNAPII [7, 8]. Those and some other methodologies for chromosomes conformation capture (3C) experiments are modified and specifically tailored to detect either local or global spatial interactions at unprecedented resolution. Yet, they are always affected by noise introducing false positive interactions, or by unavoidable systemic biases. Furthermore, many genome-wide high-resolution experiments (reaching 1 kb genomic size of interacting chromatin segments) were recently introduced to find these interactions such as in situ Hi-C [9], and new long-read ChIA-PET method using CTCF immunoprecipitation [8] that our study focused on. Each of them has its advantages and disadvantages, together with other newly introduced chromosome conformation capture (3C-based) approaches. The long-read ChIA-PET experimental data were accompanied by computational simulations [10] bridging the structural and functional interpretation of chromatin contacts and allowing biologists to visualize their data using specifically tailored bioinformatics web server [11]. Following the development of experimental and computational methods, there is a need for machine learning algorithms that can either predict the interactions at the higher spatial scales using accurate ChIP-seq experiments, or allow denoising of 3C-type of experimental data by identifying false interactions. One of such methods, EpiTensor [3]: unsupervised learning method introduced recently to predict the chromatin interactions, and detect high-resolution interactions at the genomic scale of 200 bp within the topologically associating domains (TADs) with the high accuracy. Yet, this method is yet not very effective in prediction of the physical interactions mediated by specific protein factors (like CTCF), and promoter-exon and exon-exon interactions. The Epitensor method together with earlier studies by various authors proved the possibility of predicting 3D chromatin interactions by analyzing 1D epigenomic data collected for the same cell line. But, does the epigenetics pattern differ across the whole collection of TADs in the same cell type? Does it differ between the chromosomes in the same cell type? Does it follow the same pattern at different levels (domains, chromosomes, compartments, whole-genomes, cell types) among different individuals? Translating these questions into more precise context: do the high resolution Hi-C and ChIA-PET interactions have

the same epigenetics profiles in the interacting segments across domains, chromosomes, compartments, cell types, and finally individuals at the scale of whole populations? These are open questions of great importance that we are trying to approach by this preliminary study. Previous studies were able to predict genome-wide chromatin interactions based on epigenomic data [3, 12]. In our study, we applied supervised learning approach to predict chromatin interactions using 1D epigenomic profiles and the spatial distance between the segments involved in these interactions. The study aims to build multi-level machine learning predictors starting from specifically tailored statistical models trained separately on each level of topologically associating domains, chromosomes and compartments, then the whole genome level and each cell type meta-predictors. In addition, we aim at evaluating the in silico predictors trained on one specific cell type and tested on different one to see if the epigenomics profiles follow the same pattern in different cell types. In this preliminary study, we report interactions at the TADs level because intra-domain interactions are dominating in typical Hi-C or ChIA-PET experiments [3, 8, 9, 12]. It allows also to reduce the computational time needed to train the statistical models at the whole genome scale.

The data used in this study was obtained from high-throughput experiments conducted on the following cell lines: Human lymphoblast GM12878, Human lung fibroblasts IMR90, and Human embryonic stem cells H1.

2 Pipeline Description

2.1 Features Encoding

Two different features encoding schemes were applied to represent the genomic segments:

1. **Density Coverage Encoding:** The density coverage of all histone modifications was mapped to the genome using 1 kb resolution, the percentage of each histone modification was assigned to the related segment in the genome.
2. **Peaks Information Encoding:** MACS2 peaks calling method [13] was invoked for each histone modification assay, then the peak height and the distance between the peak summit and the center of the segment were assigned to each 1 kb genomic segment.

2.2 Data Preparation

Features encoding was applied for all genomic segments with the same resolution (1 kb or 5 kb). Both encoding methods were used in different experiments. All possible pairs formed from the genomic segments in the same TAD were mapped to interactions confirmed by either Hi-C, or ChIA-PET according to the applied experiment. These interactions were used to train the predictor then evaluate its performance. In the first phase we built one predictor for each single TAD, so all pairs belong to the same TAD were split into training and testing datasets. During data partitioning phase, all pairs related to any segment exist in one dataset either training or testing, but not in both.

2.3 Building Random Forest Classifier

We chose Random Forests Classifier to accomplish the prediction task, as it improves the generalization accuracies by using group of decision trees working together. We aim in the next steps at trying different classifiers to assess their performance in such problems. The Random Forest classifier was trained on the training dataset for each TAD, then evaluated on the testing dataset for the same TAD. Moreover, we tested predictors on the same TADs but different cell lines to assess if the epigenomics follow the same pattern in different cell types. Figure 1 illustrates the whole pipeline used in the study.

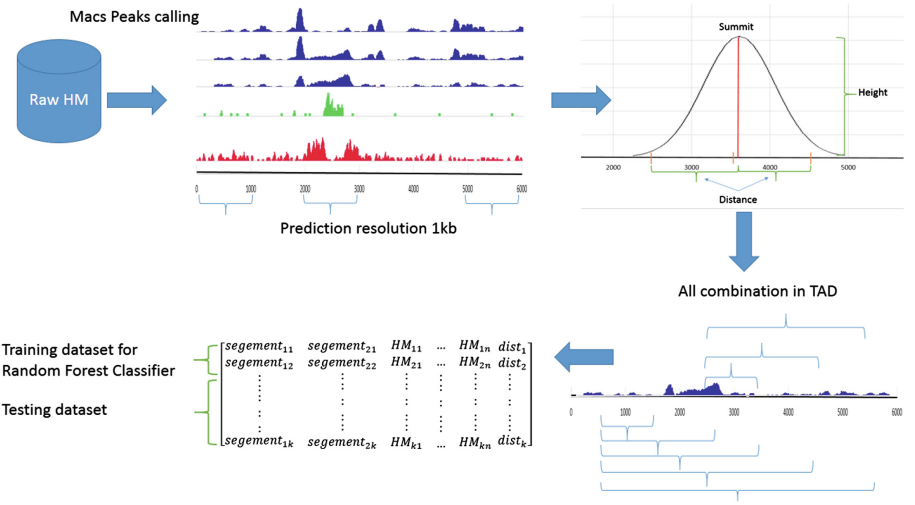

Fig. 1. The whole workflow when using peaks encoding. 1- MACS2 is invoked for all histone modifications involved in the test. 2- Peaks are mapped to the genome at 1 kb resolution. 3- Peaks height and the distance between the peak summit and the center of the segment are used to represent each segment. 4- For each TAD, all possible combinations of segments are formed. 5- These pairs were split into training and testing datasets (80–20%).

3 Results

The initial results obtained using above two features encoding methods for interacting segments representation showed the better efficiency is when using peaks information. That is why we suggest using this encoding scheme for the final genome-wide predictor. Nevertheless, we report here the features ranking for both encoding schemes, to identify which features are present in both cases, and which ones are unique. First, density coverage feature encoding was applied on IMR90 and H1 hESC cell lines, using sixteen (16) histone modification assays in addition to DNA accessibility assay. The distance between each pair anchors was also included in the features. All epigenomic data was mapped to the human genome version hg19 using 1 kb resolution. Hi-C interactions were used for training and testing datasets as true interaction

representing spatial proximity. IMR90 in situ Hi-C interactions were obtained from reference [9], while Hi-C interactions for H1 hESC were obtained from reference [14]. We present results of this encoding scheme for selected domains (TADs) from chromosome 1. The domain-level predictors trained on H1 hESC cell line were tested on the same domains in IMR90 cell line, and vice versa. Figure 2 illustrates the most important features according to the Random Forest classifier. The distance between the pair's segments was the most important feature, which was expected as all Hi-C heatmaps show that the interactions are denser close to the diagonal. The rankings of features importance at the same TADs was almost identical in these two cell lines. This suggests that statistical learning trained on one cell type could predict the interactions in another cell type.

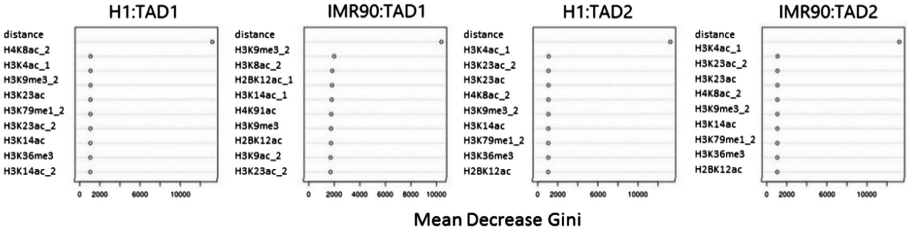

Fig. 2. Features importance in IMR90 and H1 hECS, using density coverage encoding scheme. TAD1 - chr1:43691938-46036881, TAD2 - chr1:3826747-5976883. Y-axis represent the histone modifications, the numeric suffix for each feature could be interpreted as ***1***: the maximum value of left and right segments, ***2***: refers to the minimum, ***none***: the multiplication of the left and right anchors values. X-axis refers to the mean decrease gini which represents the importance of the variable in data partitioning.

We calculated the confusion matrix, accuracy, precision, sensitivity and specificity for all predictors. Moreover, the ROC curve and the area under it were used to describe the performance of the predictors. Table 1 shows the area under the ROC curve obtained for five (5) different samples TADs located in Chromosome 1. Their locations are as follows: TAD1 - chr1:43691938-46036881, TAD2 - chr1:3826747-5976883, TAD3 - chr1:8975291-11021795, TAD4 - chr1:49037987-50814828, TAD5 - chr1:14421136-16068594. These results confirm that the epigenetics profiles have the same pattern among two tested different cell lines. The predictor trained on H1 cell line could detect IMR90 in situ Hi-C interactions with approximately the same accuracy as the predictor trained on IMR90. The same is also true for the predictor trained on IMR90. Figure 3 illustrates the ROC curve for the selected examples. The average accuracy for IMR90 cell line was 0.607 while for H1 hECS was 0.778. Figure 4 illustrates the distribution of the interactions according to the distance between the anchors. Secondly, we used peaks information to encode the epigenetic features. MACS2 peaks calling method was applied using the same parameters as in the reference [15]. We applied this scheme on IMR90, and GM12878 cell lines using in situ Hi-C data at 5 kb resolution from reference [9].

Table 1. The area under ROC curve for density coverage encoding predictors, columns headers represent training and testing cell types, the left side is the cell line where the predictor was trained, while the right side is the cell line where the predictor was tested.

TAD	H1 - H1	IMR90 -IMR90	H1 - IMR90	IMR90 - H1
TAD1	0.8582383	0.7997983	0.7990944	0.8580672
TAD2	0.820146	0.7555849	0.7551435	0.8205423
TAD3	0.8537698	0.7675929	0.7592675	0.8539412
TAD4	0.8936046	0.6441319	0.6456873	0.8941119
TAD5	0.8661579	0.7845425	0.7899033	0.8663949

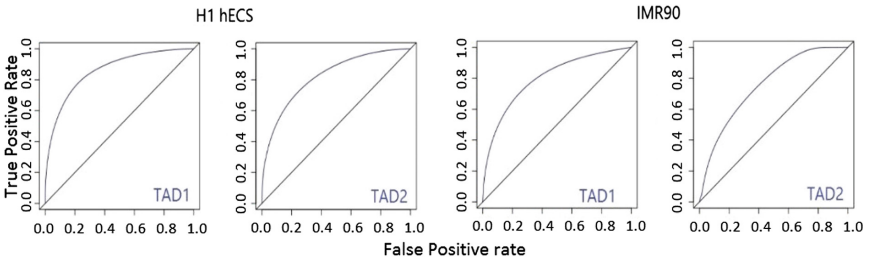

Fig. 3. ROC Curves for two different predictors applied on IMR90 and H1 hECS cell lines. TAD1 - chr1:43691938-46036881, TAD2 - chr1:3826747-5976883.

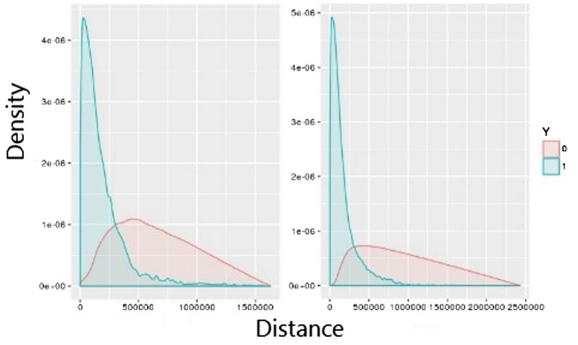

Fig. 4. Interaction density profiles as the function of the genomic distance between the pairs of interacting segments selected from GM12878 cell line. Y = 1 encodes the interaction. X-axis represents the genomic distance while Y-axis represents the density.

Figure 5 illustrates the important features according to the predictor for both cell lines GM12878, and IMR90. The average of the area under ROC curve obtained by these predictors was 0.9. Figure 6 illustrates the ROC curves for some samples from all chromosomes. Finally, the third in silico experiment used the predictor trained on GM12878 cell line with peaks encoding scheme and *in situ* Hi-C interactions to predict the ChIA-PET interactions for the same cell line. *In situ* Hi-C interactions represent

spatial proximity between chromatin segments, whereas long-read ChIA-PET interactions are believed to be true physical interactions mediated by CTCF proteins. In general, number of *in situ* Hi-C loops reported in [9] is lower than ChIA-PET interactions recognized as statistically significant with calling score larger or equal to four paired-end reads for each interaction [8]. This discrepancy is caused by the better quality of loop calling in the case of CTCF ChIA-PET experiments, whereas unspecific *in situ* Hi-C uses non-trivial and complicated method to call loops from interaction heatmaps. Surprisingly, *in situ* Hi-C based machine learning predictor could recover in average about 66% of ChIA-PET strong interactions, resolving partially the above-described discrepancy. In the selected 150 TADs our method predicted approximately 53,416 physical interactions from the whole set of 80,112 CTCF interactions as reported in [8].

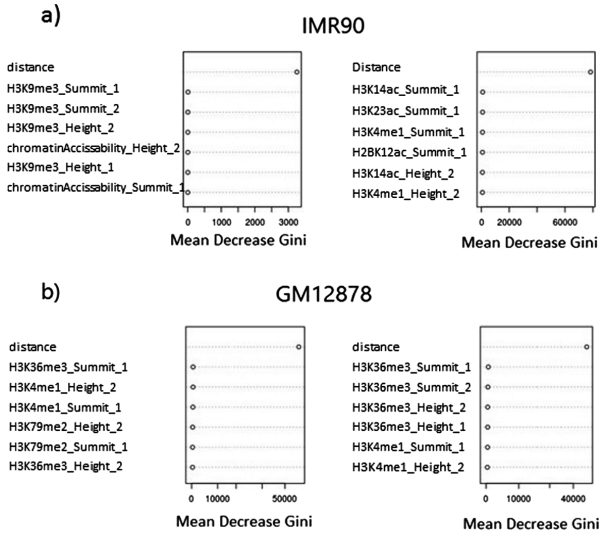

Fig. 5. Features importance in GM12878, IMR90 cell lines.

4 Data Resources

All Histone modification experiments were obtained from two on-line public resources as follows: H2BK12ac, H3K14ac, H3K18ac, H3K23ac, H3K27ac, H3K27me3, H3K36me3, H3K4ac, H3K4me1, H3K4me2, H3K4me3, H3K79me1, H3K9ac, H3K9me3, H4K8ac, H3K91ac, and DNase-seq for IMR90 and hESC H1 were downloaded from Roadmap Epigenomics project (NCBI). H3K04me1, H3K4me2, H3k4me3, H3K9ac, H3K9me3, H3K27ac, H3K27me3, H3K36me3, H3K79me2, H4K20me1 for GM12878 were downloaded from ENCODE Project.

In situ Hi-C Interactions Data for IMR90, and GM12878 were obtained from NCBI using accession number GSE63525 [9], then we normalized them using KR normalization vector obtained from the same study. Raw data for Hi-C interactions in hESC

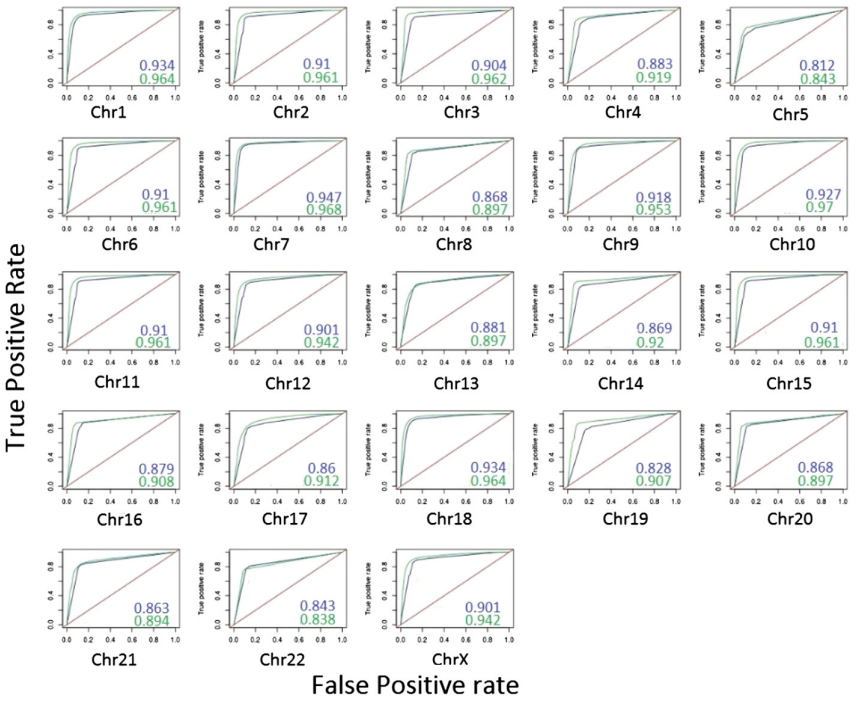

Fig. 6. ROC Curves and the area under ROC curves for selected domains from GM12878 (blue) and IMR90 (green) cell lines chromosomes. (Color figure online)

H1 were obtained from NCBI using the accession number GSE43070 [14], then we created the heatmap using 5 kb resolution, and normalized it using the Vanilla Coverage Normalization method. At the end, liftover tool was used to transform the coordinates from hg18 to hg19. p-values for Hi-C interactions were calculated based on the normalized value of the interaction strength, and we considered the most important interactions as real interactions when p-value ≤ 0.05 while the rest considered as noise. Topologically associating domains coordinates were obtained from NCBI using accession number GSE63525 [9]. We used chromatin contact domains (CCDs) as domains genomic coordinates as found in GM12878 cell line using long-read ChIA-PET [8]. The genomic localization of three-dimensional domains are typically similar for all cell types, because TADs are highly conserved among different cell lines and organisms for mammals [16]. High-resolution ChIA-PET interactions were obtained from NCBI using the accession number GSE1806 [8].

5 Summary

The initial results obtained in this study confirmed the validity of using epigenomics profiles to predict chromatin interactions, the machine learning predictors performed very well for interactions prediction. Different evaluation methods were used to assess

the performance of the predictors. The average accuracy, specificity, sensitivity and the area under ROC curve achieved by the predictors were above 0.85. TADs level predictors could identify the interaction for the same cell line and other cell lines. The tests showed that Peaks information coding could describe the interactions better than density coverage information, so in the successor studies we will focus on this method to build better predictors. More tests should be done to confirm many different cases. For TADs level predictors we will train the predictors on ChIA-PET interactions and check how many interactions we can recover using these predictors. Training a predictor on one cell line then test it on different individuals for the same cell type will be covered in the next tests. Multi-level predictors are also planned by assessing two approaches:

1. *Universal machine learning predictor* for each chromosome, trained on subset of the interactions from all TADs in the chromosome, then tested on the whole chromosome, and other chromosomes. Build one predictor for all cell lines trained on subset data from all chromosomes, and test it on the same and other cell lines.
2. Combining all predictors trained on the TADs in each chromosome to predict the interactions for each chromosome, then combine the chromosomes predictors to obtain cell line level predictor.

Acknowledgements. This work was supported by grants from the Polish National Science Centre (2014/15/B/ST6/05082), the European Cooperation in Science and Technology action (COST BM1405) and 1U54DK107967-01 "Nucleome Positioning System for Spatiotemporal Genome Organization and Regulation" grant within 4DNucleome NIH program.

References

1. McGraw-Hill Encyclopedia of Science and Technology. McGraw-Hill Education, New York (1997)
2. Bruce, A., Alexander, J., Julian, L., Martin, R., Keith, R., Peter, W.: Molecular Biology of the Cell. Garland Science, New York (2002)
3. Yun, Z., Zhao, C., Kai, Z., Mengchi, W., David, M., John, W.W., Bo, D., Nan, L., Lina, Z., Wei, W.: Constructing 3D interaction maps from 1D epigenomes. Nature Commun. **7** (2016). Article no. 10812. doi:10.1038/ncomms10812
4. Zubek, J., Stitzel, M.L., Ucar, D., Plewczynski, D.M.: Computational inference of H3K4me3 and H3K27ac domain length. PeerJ **4**(1), e1750 (2016)
5. Tyson DeAngelis, J., Farrington, W.J., Tollefsbol, T.O.: An overview of epigenetic assays. Mol. Biotechnol. **38**(2), 179–183 (2008)
6. Pietal, M.J., Bujnicki, J.M., Kozlowski, L.P.: GDFuzz3D: a method for protein 3D structure reconstruction from contact maps, based on a non-Euclidean distance function. Bioinformatics **31**(21), 3499–3505 (2014)
7. Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Han, X., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H., Chew, E.G.Y., Huang, Y.H., Welboren, W.-J., Han, Y., Ooi, H.-S., Pramila, N., Wansa, S.: An oestrogen-receptor-α-bound human chromatin interactome. Nature **462**(7269), 58–64 (2009)

8. Tang, Z., Luo, O.J., Li, X., Zheng, M., Zhu, J.J., Szalaj, P., Trzaskoma, P., Magalska, A., Wlodarczyk, J., Ruszczycki, B., Michalski, P., Piecuch, E.: CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. Cell **163**(7), 1611–1627 (2015)
9. Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T.: A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell **159**(7), 1665–1680 (2014)
10. Przemysław, S., Zhonghui, T., Paul, M., Michal, J.P., Oscar, J.L., Michał, S., Xingwang, L., Kamen, R., Yijun, R., Dariusz, P.: An integrated 3-dimensional genome modeling engine for data-driven simulation of spatial genome organization. Genome Res. **26**(12), 1697–1709 (2016)
11. Szalaj, P., Michalski, P.J., Wróblewski, P., Tang, Z., Kadlof, M., Mazzocco, G., Ruan, Y., Plewczynski, D.: 3D-GNOME: an integrated web service for structural modeling of the 3D genome. Nucleic Acids Res. **44**(1), 288–293 (2016)
12. He, C., Li, G., Nadhir, D.M., Chen, Y., Wang, X., Zhang, M.Q.: Advances in computational ChIA-PET data analysis. Quant. Biol. **4**(3), 217–225 (2016)
13. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., Liu, X.S.: Model-based analysis of ChIP-seq (MACS). Genome Biol. **9**, R137 (2008)
14. Jin, F., Li, Y., Dixon, J.R., Selvaraj, S., Ye, Z., Lee, A.Y., Yen, C.-A., Schmitt, A.D., Espinoza, C.A., Ren, B.: A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature **503**, 290–294 (2012)
15. Feng, J., Liu, T., Zhang, Y.: Using MACS to identify peaks from ChiP-seq data. Curr. Protoc. Bioinform. **14**, 1–14 (2011)
16. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., Ren, B.: Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature **485**(7398), 376–380 (2012)

Springer

http://www.springer.com/978-3-319-60437-4

Foundations of Intelligent Systems
23rd International Symposium, ISMIS 2017, Warsaw, Poland, June 26-29, 2017, Proceedings
Kryszkiewicz, M.; Appice, A.; Slezak, D.; Rybinski, H.; Skowron, A.; Ras, Z. (Eds.)

2017, XXIX, 747 p. 182 illus., Softcover
ISBN: 978-3-319-60437-4