Mini review

Predicting 3D chromatin interactions from DNA sequence using Deep Learning

Robert S. Piecyk a,1, Luca Schlegel a,1, Frank Johannes a,b,*

a Department of Molecular Life Sciences, Technical University of Munich, Freising, Germany
b TUM Institute for Advanced Study, Garching, Germany

ABSTRACT

Gene regulation in eukaryotes is profoundly shaped by the 3D organization of chromatin within the cell nucleus. Distal regulatory interactions between enhancers and their target genes are widespread and many causal loci underlying heritable agricultural or clinical traits have been mapped to distal cis-regulatory elements. Dissecting the sequence features that mediate such distal interactions is key to understanding their underlying biology. Deep Learning (DL) models coupled with genome-wide 3C-based sequencing data have emerged as powerful tools to infer the DNA sequence grammar underlying such distal interactions. In this review we show that most DL models have remarkably high prediction accuracy, which indicates that DNA sequence features are important determinants of chromatin looping. However, DL model training has so far been limited to a small set of human cell lines, raising questions about the generalization of these predictions to other tissue-types and species. Furthermore, we find that the model architecture seems less relevant for model performance than the training strategy and the data preparation step. Transfer learning, coupled with functionally curated interactions, appear to be the most promising approach to learn cell-type specific and possibly species-specific sequence features in future applications.

© 2022 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Contents

1. Introduction ..................................................................................................... 3440
2. Computational methods ................................................................................... 3441
   2.1. Preprocessing ................................................................................................. 3441
     2.1.1. Preparation of training data ................................................................. 3441
     2.1.2. Auxiliary data .......................................................................................... 3442
     2.1.3. Sequence embedding .............................................................................. 3442
     2.2. Deep Learning approaches ......................................................................... 3442
     2.3. Performance and training strategies ............................................................ 3443
     2.4. Biological insights and applications .............................................................. 3444
     2.4.1. Genomic determinants of looping .......................................................... 3444
     2.4.2. Screening of disrupting variants .............................................................. 3445
     2.4.3. Cell-type and species specific chromatin interactions ............................. 3445
3. Discussion ......................................................................................................... 3445

Author contributions............................................................................................ 3446

Corresponding author at: Department of Molecular Life Sciences, Technical University of Munich, Freising, Germany.
E-mail address: frank@johanneslab.org (F. Johannes).
1 These authors contributed equally to conceptualisation and writing of the paper. Robert S. Piecyk and Luca Schlegel are both shared first authors.

https://doi.org/10.1016/j.csbj.2022.06.047
2001-0370 © 2022 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
1. Introduction

The spatial organization of chromatin in the nucleus of animal and plant cells plays important roles in genome regulation. It has central functions in processes such as DNA replication and repair, the spatial- and temporal patterning of gene expression, and in the silencing of transposable element (TE). In the past two decades, chromosome conformation capture (3C) coupled with next-generation sequencing has emerged as a powerful method to interrogate 3D chromatin interactions in a high-throughput fashion [1]. Among these, Hi-C was the first developed method [2]. It is designed to capture all chromatin interactions at the genome-wide scale, albeit at low resolution (see Fig. 1,2, [2]). By contrast, more recent methods, such as chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) (see Fig. 1,3), generate high-resolution interaction maps, but are restricted to specific loci occupied by proteins that can be pulled down by ChIP (e.g. modified histones, transcription factors, and RNA polymerase II). Together, these techniques have generated unprecedented insights into the function of 3D chromatin organization of mammalian and more recently also in plant genomes. They have led to the systematic identification of Enhancer-Promoter Interactions (EPIs), Insulator Loops (e.g. CTCF-cohesin loop in human cells) and interactions mediated by specific transcription factors [6].

Numerous machine learning (ML) approaches have emerged in parallel to these technological developments [7]. Their general aim is to use 3C data as input to train sequence-based predictors of chromatin looping and to identify specific sequence features that may facilitate physical contacts between distal genomic regions. The most promising of such ML approaches are supervised Deep Learning (DL) methods. As in other areas of genome biology, DL methods provide the most accurate predictions, can handle large and complex amount of genomic data and automatically detect patterns or unanticipated genomic relationships [8]. DL methods thus provide a powerful framework for dissecting the causal determinants underlying 3D chromatin interactions and for providing testable hypotheses for experimental follow up.

Here we review the state-of-the-art in the use of DL methods to predict 3D chromatin interactions from DNA sequence (Table 1). We evaluate these models in terms of their objective, data preprocessing, architecture, training procedure and finally by their prediction performance. We find that data selection and preprocessing in combination with transfer learning appear to be more important for model performance than the choice of model architecture. Moreover, even though all these models perform very well on their test data, they have so far been mainly restricted to the analysis of specific human cell lines. Similar model-based approaches are thus urgently needed for a wider range of cell lines, tissues or species. This information could facilitate deeper insights into the evolutionary and developmental factors that impact chromatin looping biology. Moreover, experimental validation of model-based predictions is needed to assess the biological value of these approaches and for fine-tuning model architecture.

1.1. Sequencing methods

Chromosome conformation capture (3C) followed by high-throughput sequencing has emerged as a powerful experimental approach to probe chromatin interactions at the genome-wide scale. Hi-C and ChIA-PET are two variants of these measurement approaches, which have served as the main data input for most of the DL methods reviewed in Table 1. The general experimental workflows for Hi-C and ChIA-PET involve cross-linking and fragmentation of chromatin, the addition of biomarkers, ligation, purification and finally sequencing [44]. We detail these experimental steps in Fig. 2. Despite the popularity of Hi-C, there are a number of well-known limitations with this method. First, its
resolution is limited by the choice of the digestion enzyme [45], so that spatially close chromatin interactions may be missed in genomic regions where the distribution of enzymatic cut sites is sparse. Second, Hi-C often does not capture long range interactions and omits many simultaneous promoter-enhancer interactions [46]. Third, false positive interactions are often detected because of spurious cross-linking and ligation. The first two limitations could be viewed as potential opportunities for DL methods, because computationally predicted loops could, in principle, be generated in genomic regions where the measurement technology has failed. However, the third limitation is disadvantageous for model training, where clean true positive (and true negative) loop sets are necessary. Newer experimental approaches, including ChIA-PET, Micro-C or Promoter Capture Hi-C (PCHi-C), try bypass many of these limitations. Micro-C uses micrococcal nuclease (MNase) as replacement for restriction enzymes to archive a higher resolution for short-distance interactions [47]. PCHi-C introduces an additional individual biotinylated RNA purification step for promoter rich fragments, to reduce the amount of ligation products before PCR sequencing is applied [48]. Unlike Hi-C, ChIA-PET combines 3C with chromatin immunoprecipitation (ChIP) followed by sequencing [3]. It performs chromatin fragmentation by sonication and uses a smarter biolinker concept [49]. Additionally, the use of ChIP enriches for chromatin contacts that harbor specific transcription factors, or other binding proteins, such as CTCF, and thus ensures a higher rate of true functional chromatin interactions [49]. However, as a trade-off ChIA-PET has relatively low sensitivity and requires a large amount of material [5], which could increase measurement variation due to cellular heterogeneity. Still, ChIA-PET is frequently applied and has thus been employed in the training of several DL algorithms (see Table 1).

### 2. Computational methods

#### 2.1. Preprocessing

##### 2.1.1. Preparation of training data

Measured (i.e. observed) chromatin interactions from Hi-C or ChIP-PET are the starting point for all DL methods reviewed here. However, DL methods differ in the way they use this information at the input stage. We can broadly distinguish between classification- and regression-based methods.

| Method      | Objective          | Architecture              | Training data | Organism | Auxiliary data                                  | Input                        |
|-------------|--------------------|---------------------------|---------------|----------|------------------------------------------------|------------------------------|
| EPIANN [9]  | predict EPI        | CNN                       | Hi-C [10]     | human    | annotations [11, 12]                            | anchor 1: 3 kb anchor 2: 2 kb |
| TransEPI [13]| predict EPI        | CNN                       | Hi-C [10]     | human    | ChIP-seq [14] histone marks [22] DNase-seq [12] BINGI [15] annotations [16, 17] | DNA sequence: ~ 2.5 Mb + chromosome features |
| Rambutan [18]| predict 3D chromatin interaction | CNN                       | Hi-C [10]     | human    | DNease [12]                                     | anchor 1: 1 kb anchor 2: 1 kb |
| Akita [19]  | predict 3D chromatin interaction | CNN                       | Hi-C [10]     | human and mouse | -                                     | DNA sequence: ~ 1 Mb |
| DeepC [22]  | predict 3D chromatin interactions | CNN, transfer learning     | Hi-C [10, 23] | human and mouse | annotations, ChIP-seq [23, 24, 25] ATAC-seq [23, 25, 24] DNase-seq [26, 20] histone marks [12, 24] | DNA sequence: ~ 1 Mb + chromosome features |
| Orca [27]   | multiscalar 3D chromatin interaction | CNN, transfer learning     | Micro-C [20]  | human    | annotations [11, 12]                            | DNA sequence: from 1 to 265 Mb |
| EPiScNN [28]| predict EPI        | CNN, transfer learning     | Hi-C [10]     | human    | annotations [11, 12]                            | anchor 1: 3 kb anchor 2: 2 kb |
| EPiSHiLBert [29]| predict EPI        | CNN, transfer learning     | Hi-C [10]     | human    | annotations [11, 12]                            | anchor 1: 3 kb anchor 2: 2 kb |
| EPHIC [30]  | predict EPI        | Multilayer Perceptron transfer learning | Hi-C [10]     | human    | annotations [11, 12]                            | anchor 1: 3 kb anchor 2: 2 kb + chromosome features |
| ChiNN [31]  | predict 3D chromatin interaction | CNN, gradient tree boosting | Hi-C [10] | human    | ChIA-PET [32]                                   | anchor 1: 1 kb anchor 2: 1 kb |
| DeepTact [33]| predict EPI/PPi    | CNN + RNN                 | PChi-C [34]   | human    | ChIP-seq [14] FANTOM5 [35] DNase-seq [24] TSS [36] CHICAGO scores [37] | anchor 1: 1 kb anchor 2: 1 kb |
| DeepMILO [38]| predict 3D chromatin interaction | CNN + BLSTM               | ChIA-PET [39, 33] | human    | ChIP-seq [20]                                   | anchor 1: 4 kb anchor 2: 4 kb |
| SEPT [40]   | predict EPI        | CNN + LSTM transfer learning | Hi-C [10]     | human    | annotations [11, 12]                            | anchor 1: 3 kb anchor 2: 2 kb |
| SPEID [41]  | predict EPI        | CNN + LSTM transfer learning | Hi-C [10]     | human    | annotations [11, 12]                            | anchor 1: 3 kb anchor 2: 2 kb |
| EPIVAN [42] | predict EPI        | CNN + BiGRU               | Hi-C [10]     | human    | annotations [11, 12]                            | anchor 1: 3 kb anchor 2: 2 kb |
| EPI-DLMH [43]| predict EPI        | CNN + BiGRU               | Hi-C [10]     | human    | annotations [11, 12]                            | anchor 1: 3 kb anchor 2: 2 kb |

### Table 1

Deep Learning algorithms for 3D chromatin interactions, sorted by architecture. All models are based on Convolution Neural Networks or Recurrent Neural Networks.

[18,41,43,19,27,28,13,22,29,33,38,40,42,31,30] [11,12,14–17,20,21,23,25,26,32,34–37].
positives. The goal is to learn specific sequencing features within the anchors that may facilitate looping. The sizes of the input anchors can range from 2 kb to 3 kb, depending on the anchor type (see Table 1). In addition to these true positives, it is also necessary to supply true negative anchor pairs.

Most methods do this by “rewiring” the positive set of anchors in new ways; that is, they form in silico loops between anchors that have not been observed to interact in the original Hi-C or ChIP-PET data. These generated negatives are often matched for distances similar to those in the positive set, in order to avoid biases in model training (DeepMILO, DeepTACT, SEPT, SPEID). Alternatively, any distance information can be added as the auxiliary input in the training procedure itself (ChINN, Rambutan, TransEPI). Because of the substantial imbalance between positive and negative interactions, a data augmentation procedure is sometimes employed that adds arbitrary sub-sequences flanking the original anchors from the positive set (EPIANN, EPISCNN, EPISHilbert, SEPT, SPEID, EPI-VAN, EPI-DLMH, EPICH).

In contrast to classification-based approaches, regression-based methods (Akita, DeepC, Orca) take contact frequency maps as input, which are constructed at megabase (Mb) resolution. To this end, the genome is partitioned into non-overlapping virtual contigs, and interaction counts within ~1 Mb sub-sequences are taken to generate a two-dimensional frequency matrix. DNA sequences and corresponding frequency matrix are fed into the DL algorithm to predict factors affecting local interaction typologies (Akita, DeepC). Orca introduces, in addition to the sequence encoder, a multilevel cascading decoder to provide genome wide interaction on window sizes from 1 Mb to 265 Mb with resolution between 4 kb and 1024 kb.

2.1.2. Auxiliary data

Many classification-based methods further employ auxiliary data to filter anchor pairs prior to training. For example, methods such as EPIANN, TransEPI or SPEID (see Table 1), select anchors that map to annotated enhancers and promoters. Such filtering strategies reduce the genome-wide interaction landscape to a subset of functional regions, and (probably) reduce biases or measurement errors arising from the 3C assay itself. Clearly, such a priori filtering strategies are only sensible in applications involving high-quality genomes where sufficient epigenomic information is available and all regulatory elements are well annotated. This is certainly true for all DL methods to date, as they have been developed specifically for human genomic applications, and therefore benefit from the extensive ENCODE data resources [24].

Common auxiliary data modalities include RNA-seq (gene expression), DNase-seq (accessible regions), as well as ChIP-seq for various histone modifications, RNA polymerase, transcription factors or CTCF binding proteins. Using such functional data, anchors overlapping low-expressed transcripts from RNA-seq can be removed from the enhancer-promoter data, since they are less likely to be actively regulated by enhancers [13,40,50]. Similarly, anchors can be screened for chromatin accessible regions related to CTCF or RNA Pol II binding sites [31] to be enriched for active loops. Furthermore, the incorporation of functional data makes it possible to identify cell-type-specific chromatin interactions. This creates avenues for predicting and understanding looping biology underlying cell lineage determination.

As an alternative to a priori filtering, some methods integrate such auxiliary data directly into the model training, either in a pre-training step (DeepC, Orca, ChINN) or in the full training procedure (TransEPI, EPICH). Either way, it has been shown that the inclusion of auxiliary data into DL methods can improve algorithm performance (TransEPI, DeepC, Orca, ChINN, EPICH), and thus appears to be an important aspect of data preparation.

2.1.3. Sequence embedding

The nucleotide sequences used in DL model training need to first be converted into a machine readable format. One-hot encoding is a commonly used approach to represent sequences as binary vectors. The four nucleotides A, T, C and G are saved in four separate channels. A is represented as $[1, 0, 0, 0]^T$, T as $[0, 1, 0, 0]^T$, C as $[0, 0, 1, 0]^T$ and G as $[0, 0, 0, 1]^T$. The unknown nucleotide category N can be either removed (SPEID, EPIANN, SIMCNN), represented as $[0.25, 0.25, 0.25, 0.25]^T$ (ChINN), or saved as the fifth dimension in matrix representation (DeepMILO). Moreover, one-hot encoding can be easily extended to capture the spatial relationship between anchors by converting a two-dimensional one-hot matrix into a three-dimensional matrix-vector representation by applying a space-filling Hilbert curve [51,52] (see EPISHilbert).

Another group of methods uses instead a distributed representation of small DNA sequences, called k-mers. Dna2vec embedding [53] is based on a popular word2vec [54] natural language processing model. The algorithm takes distributed samples of variable-length k-mers to train a shallow two-layer neural network model. This aggregated model enables the user to perform a decomposition by k-mer length, followed by the selection of the best low-dimensional and high-quality vector representation used for sequence embedding. This increases computational efficiency as well as specificity in DL approaches (EPIVAN), and preserves the hidden information from the correlation between single small DNA sequences. It can be either trained on the whole genome or on a set of anchors.

2.2. Deep Learning approaches

We examined the model architectures of the DL methods in Table 1 in detail. The most frequent architecture uses a Convolutional neural network (CNN) in combination with a long short-term memory (LSTM) (see Fig. 2). CNNs, or simply convolutional networks (CNs), are a specific type of DL, which are based on artificial neural networks (ANNs) [55]. Models with a long short-term memory (LSTM) unit belong to the family of recurrent neural networks (RNNs) and are also a class of ANNs [56]. The family of LSTM units also comprises Gated Recurrent Units (GRU). It is a slightly simpler version with less parameters [57]. Both models often occur with additional connections in opposite directions within one layer [58]. In that case, they are called Bidirectional LSTM (Bi-LSTM) and Bidirectional-Gru (Bi-Gru). Due to the continuous refinement of 3C-based methods, especially Hi-C, ChIA-PET or other throughput techniques, large amounts of high-quality chromatin data is available. This data motivates the use of supervised versions of the introduced ANN models. If we consider Table 1, we notice that all models are using CNNs as core models and six of 16 additionally add an extra (Bi-)LSTM or Bi-Gru unit (DeepTact, DeepMILO, SEPT, SPEID, EPI-VAN and EPI-DLMH). This observation could be explained by the historical applications of CNNs and LSTMs. As its name indicates, CNNs use a convolution operator, also called kernel function, to search for specific patterns in discrete grid-based topologies or sequential data [59]. The concept was originally developed for image or sound classification and is related to the neuronal connections of the human brain [60]. Section 2.1 explained the conversion of genomic sequences into a matrix or vector-based representation. This indicates the connection between image recognition and the computation of chromatin interaction since the input data of both problems can be represented by a matrix or linear vector combination that contain unknown patterns and possibly long-term dependencies. LSTM units are designed to capture long-term dependencies, due to shared parameters that function as a memory unit. One could say, if CNNs catch mostly local patterns, LSTMs act as a global
observer. Therefore, LSTMs have been used mainly in time dependent problems, as they store crucial information over a long period of time. A typical application is speech recognition, where it can take a long time until a specific word or phrase is repeated [56]. A similar event occurs in promoter-enhancer interactions, which often span distances of up to several Mbs [61].

Even though the layer-based representation shown in that figure is helpful for visualizing the overall model structure, it is important to keep the actual neural-based architecture in mind. All layers contain nodes that are connected through edges. On edges, linear transformations are applied, which provide all the weights or parameters. On nodes, preselected non-linear transformations \( f = f(x) \) are necessary. The process of training is equivalent to minimizing a loss function \( L \) with respect to lambda, which contains the input values \( x \), the parameter set \( \lambda \) and all non-linear functions \( f(x) \). This representation is drastically reduced and on a highly abstract level. Many additional decisions are necessary to define a full CNN model with LSTM units. (C) After training, we end up with a set of optimal parameters \( \lambda \). This set of trained parameters in combination with the model architecture must be applied to a final test data for validation, before applying to completely new data.

2.3. Performance and training strategies

We sought to rank the prediction performance of the models reviewed in Table 1. The majority of the original studies provide values for the area under precision recall curve (AUPR) as statistical measurement of prediction accuracy. A higher AUPR value is indicative of better model performance. Since Akita, Orca and DeepC are regression based models, they did not published those values and are not included in the performance analysis. Nevertheless, since Akita and Orca are trained on the same Micro-C data set, it is possible to compare their Pearson correlation coefficient. On average, in Orca this correlation coefficient is ~ 7.4% higher for the H1 embryonic stem cells (H1-ESC) and ~ 6.4% higher for the human foreskin fibroblasts (HFF). Beside these, we exclude all algorithms that are not trained and tested on Hi-C [10] data, because mixed data modalities would render comparisons difficult.

The remaining models (EPIHilbert, EPIVAN, EPIsCNN, EPI-DLMH, EPIsCNN, EPIVAN, EPIHC and EPIANN), which provide AUPR values.

![Fig. 2. Training procedure of a typical CNN + LSTM model. (A) During input data preprocessing, chromatin sequences are translated using a one-hot encoding technique. (B) Schematic representation of a commonly used architecture in chromatin interaction detection. Two input anchors are managed separately by CNN blocks, which are defined by the number of convolutional and pooling layers and several hyperparameters. Once the architecture, the arrangement of the layers and hyperparameters are selected, we start with an initial unbiased parameter distribution \( \lambda \). The concept of ‘training’ refers to minimizing a certain loss function \( L \) with respect to lambda, which contains the input values \( x \), the parameter set \( \lambda \) and all non-linear functions \( f(x) \). This representation is drastically reduced and on a highly abstract level. Many additional decisions are necessary to define a full CNN model with LSTM units. (C) After training, we end up with a set of optimal parameters \( \lambda \). This set of trained parameters in combination with the model architecture must be applied to a final test data for validation, before applying to completely new data.](image)

![Fig. 3. Performance bar plot for models, which provided AUPR value for cell type specific training and testing. Light blue bars indicate a pure CNN model, dark blue refers to a CNN + RNN model. Transfer learning is represented by dotted bars. The yellow interval is defined by the minimum and maximum AUPR value.](image)
values, have been trained on a combination of six cell lines: K562 (mesoderm lineage cells from a leukemia patient), GM12878 (lymphoblastoid cells), Hela-S3 (ectoderm lineage cells from a cervical cancer patient), HUVEC (umbilical vein endothelial cells), IMR90 (fetal lung fibroblasts) and NHEK (epidermal keratinocytes) [10], and employ a rich set of auxiliary functional data [62].

We observe four essential training strategies:

- Training on a specific cell line
- Training on a combination of all cell lines
- Model-based learning
- Cellular transfer learning (data based)

Model-based training often introduces an attention layer, to merge the acquired feature knowledge of cell-line specific sub-models. Cellular transfer learning suggests two or more stages. First, all cell line data is used for training, and second, a specific cell-line is selected for another training procedure. The second cell-line specific training consumes little computational effort [28], since the model architecture remains the same and the primary trained parameters are used as initial weights. If cellular transfer learning has been applied, the process is repeated for all cell lines separately to obtain a total of six AUPR values. Training and test-sets originate from the same cell line. The mean value of these six AUPR values is plotted in Fig. 3 with the range of all values indicated by the yellow interval. We observe that models using transfer learning tend to perform comparatively well. This could be an indicator that cell-specific features are important predictors [63]. Interestingly, we also see that the more sophisticated CNN + LSTM architectures are not among the top performers.

There are a number of caveats with this simple side-by-side performance evaluation. First, AUPR values are just one of several statistical measurement tools. The imbalanced training data set contains approximate 20 times more true interactions. Since the AUPR value mainly composes true interactions [64], it is the most frequently used performance tool. Second, we also do not consider the computational effort and costs of the models, which is an essential factor for the efficiency.

2.4. Biological insights and applications

2.4.1. Genomic determinants of looping

DL models are undoubtedly powerful tools for predicting 3D chromatin looping based on DNA sequence. However, they also provide an intriguing framework for dissecting the underlying looping biology. A common approach in classification-based algorithms is to use in silico mutagenesis, whereby mutations are artificially introduced into specific anchor pairs, and then re-supplied to the trained DL model. The concept of this approach is presented in the Fig. 4. Mutations that lead to significant drops in the predicted interaction probability would suggest an important role in facilitating chromatin contacts, perhaps because they are located in crucial protein binding motifs. Algorithms such as DeepC, Akita, Orca and DeepMILO have extensively used this approach to assess the effect of specific deletions, single nucleotide polymorphisms (SNP) or structural variants. Akita, for instance, mutagenized a set of random regions within and near CTCF motifs. The results of this study showed the significant impact of SNPs on CTCF binding, either directly or by flanking cofactors. Additionally, a mouse-trained model from Akita was used to predict the effect of a 622 kb inversion at the enhancer locus Eph4A on 3D folding. The experimental studies observed that the inversion affected CTCF binding [65]. Using its predicted contact maps, Akita confirmed that an in silico inversion of the Eph4A locus lead to a loss of CTCF mediated insulator looping. However, in silico mutagenesis is a brute force approach. It is not optimal for probing the complete combinatorial mutation space within a given anchor sequence. This limitation may be crucial in situations where looping is facilitated by combinations of specific (and possible complex) motif sequences. As an alternative, DeepC employs a metric called saliency score [66], which quantifies the importance of every base pair and motif to the predicted interaction. This metric can be calculated as the scalar product of the model output gradient, with respect to the one-hot-encoded input sequence. Akita used this metric as well. They found saliency peaks at the CTCF motifs and active promoters, which quantify the importance of every base pair and motif to the predicted interaction. This metric can be used for eQTL mapping when the expression changes are caused by chromatin architecture, and thus, gene expression, which can be investigated by expression quantitative trait locus (eQTL) studies. To test this, they used the set of cell-type specific eQTLs located in open chromatin or CTCF sites. They compared the saliency scores of these eQTLs and mutagenized random regions. The significantly higher saliency score for eQTLs revealed that this metric can be used for eQTL mapping when the expression changes are caused by chromatin architecture perturbation. Orca successfully predicted the influence of several structural variants in six studies while comparing the model output with experimental chromatin capture data.

Class Activation Maps (CAM) [67] is another approach to quantify the influence of sequence features on enhancer-promoter interaction. It visualizes three-dimensional vectors with a heatmap...
matrix, which represents the interaction occurrence and their spatial relationship. These association maps may be used to highlight sequence patterns leading to the chromatin interaction (EPIsHilbert). Another proposed method to detect motifs responsible for chromatin looping is based on the output of the first convolutional layer. This procedure can extract the best matching subsequences for each kernel, with respect to the model architecture. SEPT used this strategy to compute a position frequency matrix (PFM) and then, compare the PFM-related features with known TF motifs from HOCOMOCO database [68]. They found a set of potentially important regulatory elements, which are involved in transcription and cell-cycle regulation that may determine their role in chromatin looping. These results revealed that SEPT has the ability to learn cell-type specific patterns crucial in genome folding, which explains its relevance in transfer learning approaches. Attention layers [69] can be used not only to merge cell-line specific submodels in transfer learning, but also to evaluate the impact of those features on prediction. EPIANN labeled each base in observed enhancer and promoter sequences with the corresponding marginal attention to highlight regulatory elements in observed enhancer and promoter sequences with the corresponding marginal attention to highlight regulatory elements. This information can be further investigated to interpret the key patterns in chromatin interactions.

2.4.2. Screening of disrupting variants

Building on the above-describe approach, several studies have tried to test or identify specific loop-disrupting genetic variants and their impact on cancerogenesis [39]. Such genetic variants could result in the suppression, or in some cases even the activation, of chromatin loops. In the latter case, enhancer elements could be erroneously brought in physical contacts with proto-oncogenes and thus promote cancer progression [70,71]. The DL prediction framework provides a means to systematically screen through available GWAS or re-sequencing databases to identify putative causal variants underlying differential looping.

Using such an approach, DeepMILO was tested on two known deletions in T-Cell Acute Lymphoblastic Leukemia (T-ALL) patients at anchors containing oncogenes TAL1 and LM02. The algorithm correctly predicted that these mutations should lead to insulator loop disruption. In addition, DeepMILO employed in silico mutagenesis and further predicted that a number of smaller mutations, just outside the CTCF binding motif, may also affect chromatin interactions. These latter predictions provide concrete hypotheses for experimental follow-up. Conceptually similar approaches were taken by TransEPI and ChiNN to identify putative loop-disrupting mutations in the context of neuronal disorder and chronic lymphocytic leukemia (CLL), respectively. Using chronic lymphocytic leukemia (CLL) patient data, ChiNN was able to construct a patient-specific chromatin interaction profile, suggesting that such predictions could serve as a per-clinical tool to predict CLL disease risk. Orca predicted high structural impact regions (10 bp) consistently with Chip-seq data and confirmed reduced genome interactions through disrupting motifs like POUSF1::SOX2 or AP-1. Overall, these results suggest local and global effects of disrupting variants.

Many eQTLs identified in human population data have been shown to act in trans, that is, the QTL affects the expression state at distal genes, rather than locally. One possible mechanism for these distal interactions is chromatin looping [72]. Indeed, application of Akita on a set of eQTLs from GTEx (Genotype-Tissue Expression, [73]) whole blood samples revealed significantly higher disruption for SNP with greater causal probability, within and outside of CTCF motifs. This indicates the impact of CTCF and non-CTCF variants on genome folding. Hence, eQTL data sets can be used in biological validation of predicted interactions to reduce false positives from 3C-based data.

2.4.3. Cell-type and species specific chromatin interactions

Cross-cell line prediction poses a big challenge due to the presence of cell-type specificity [63,10]. Our review revealed that general DL models trained on all cell lines together, tend to perform poorly in capturing cell-specific events due to the relative disproportion between shared and private chromatin interactions. Conversely, most DL methods that are trained on each cell line separately, failed to identify general and cell-line specific interactions simultaneously. Approaches based on transfer learning perform much better in this setting. EPIsHilbert hypothesized that the effectiveness of these approaches is determined by the numerous common sequence patterns among all cell lines. To evidence this, they calculated the overlapping ratio of chromatin interactions between different tissues. It indicated that there are more common than private interactions. This knowledge can be used to create a model that has the ability to predict chromatin interactions on novel cell lines. SEPT used a feature extractor and domain discriminator to learn EPIs-related features and recognize cell-line specific patterns at the same time. This provides an opportunity to create an universal model, which is able to predict chromatin interactions not only on previously trained cell lines, but also on novel cell lines using general EPI features. Similarities between mammalian genome folding may allow us to predict species-specific differences in genome folding. Recent studies showed that ChAHPS complexes lead to the disruption of insulator loops within mouse-specific B2 SINE elements [74,75]. To test this, Akita trained models on human and murine embryonic stem cell (ESC) Hi-C to show the impact of in silico mutation in these elements on CTCF binding. Comparison of these results confirmed that both models correctly predicted the disturbance of genome folding before and after mutagenesis of B2 SINE elements. This highlights the opportunity to use DL and transfer learning approaches in studies investigating species-specific regulatory strategies (see Fig. 4).

3. Discussion

Here, we have reviewed current supervised DL models for predicting 3D chromatin interactions from DNA sequence. We find that these methods have remarkably high prediction accuracy, which indicates that DNA sequence features are important determinants of chromatin looping. By examining the learned sequence features it is possible to uncover complex, combinatorial, sequence motifs that would otherwise be difficult to discover, even with elaborate experimental assays. Thus, DL models have the potential to provide novel insights into chromatin biology. Similarly, trained DL models can be used as a tool to identify loop disrupting genetic variants from population-level sequencing data. This type of information is highly relevant for understanding the genetic basis of regulatory variation underlying complex traits, both in a clinical as well as in an agricultural setting. Indeed, numerous genomewide association studies have identified causal loci in non-coding regions of genomes, many of which appear to act as eQTL for distal target genes [76,77]. DL models could be used to assess if these non-coding loci and their targets are likely to interact physically, and whether the type of genetic variants seen at the eQTL locus is expected to cause differential looping. Understanding the mode of action of disease-associated non-coding variants can facilitate insights into disease etiology and potentially lead to novel treatment targets in biomedical applications.

Despite such exciting prospects, our review also revealed a number of key limitations with current DL methods. All models to date have been trained and tested on human data of six cell lines from the ENCODE project [24], or a subset of those cell lines. While the use of a single, or a few, reference data set(s) enables comparisons of the algorithms among themselves, it reduces the generalizability.
of the trained models. This potential limitation is already apparent in DL models that trained on multiple cell lines and showed how sensitive the trained model can be to cell-type specific features (EPIsCNN, EPiSHilbert, EPiVAN, SEPT, TransEPi). It is therefore highly unlikely that current models readily extend to other in vivo tissues in humans, and much less across different species. Hence, a much broader range of training data sets is urgently needed.

From our perspective, a cross-species DL model would be highly interesting. For instance, it is well known that the biology underlying chromatin looping in mammals and plants differs fundamentally. Plants lack CTCF proteins and display several other key topological differences in the 3D organization of their genomes. In contrast to mammalian TADs structure, plant compartment domains tend to interact with each other at intra and inter-chromosome level [78,79]. Moreover, although mammalian loop domains are conserved through different species due to conservation of CTCF binding sites, plant compartments might differ for several plant species [80]. The molecular components underlying chromatin looping in plants is not fully resolved. Cross-species DL models would not only be able to identify conserved and divergence sequence determinants, but also reveal how the molecular mechanisms underlying chromatin looping have evolved. However, training the DL models reviewed here on other species may not be trivial. Many of the models rely heavily on auxiliary functional data (DNase-seq, RNA-seq, ChIP-seq, CTCF binding proteins, transcription factors, RNA polymerase) and high-quality genome annotations in the training procedure itself (see Table 1). While this type of data is useful for boosting model performance, it is not readily available in most non-human species. Thus, such auxiliary data would have to be generated first, or alternative training strategy would have been employed that rely less on such data sources. From a bioinformatics point of view, it may be tempting to focus on the development of the ANN structure. However, if we consider the performance metrics in Fig. 3, there is no clear indication that ANN structure has a significant impact, if it consists of some minimal criteria like two well performing CNN blocks. Zhuang et al. [28], for instance, shows that a simple CNN model has similar performance on the same training data as a CNN models coupled with a RNN. This even raises questions about the implementation of an LSTM unit altogether, even though there are decent biological interpretations of their function. Since the computational cost of LSTM or related units is very expensive compared to CNN blocks, they might fail in the long run due to lack of efficiency. On the other hand, the training process itself seems to have a very deep impact on prediction results. If we consider the AUPR values in Fig. 3, we observe that the most promising results are derived by training processes that use transfer learning strategies. This improvement in statistical measurements, can be explained biologically by cell line specific and general features along the chromatin. In mammalian cells, for example, it is known that CTCF bindings are specific [81]. It should also be clear that the use of DL models cannot fully replace experimental data in revealing chromatin interactions. Experimental validation of predicted interactions should go hand-in-hand with model building. To date, there has been relatively little effort to perform such validations.

Author contributions

All authors contributed to the conceptualisation and writing of the paper.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

F.J. acknowledges support from the Technical University of Munich Institute for Advanced Study, funded by the German Excellence Initiative and the European Seventh Framework Programme under grant agreement No. 291763. R.S.P. was supported by the SFB Sonderforschungsbereich924 of the Deutsche Forschungsgemeinschaft (DFG). L.S. was supported by the DFG.

References

[1] Annette Denker and Wouter de Laat. The second decade of 3c technologies: detailed insights into nuclear organization. Genes Develop, 30(12), 1357–1382, 2016.
[2] Erez Lieberman-Aiden, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R. Lajoie, Peter J. Sabo, Michael O. Dorschner, Richard Sandstrom, Bradley Bernstein, M.A. Bender, Mark Groudine, Andreas Gnarra, John Stamatoyannopoulos, Leonid A. Mirny, Eric S. Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science, 326(5950), 289–293, Oct 2009.
[3] Fullwood Melissa J, Ruam Yujin. ChiP-based methods for the identification of long-range chromatin interactions. J Cell Biochem 2009;107(1):30–9.
[4] Rongxin Fang, Miao Yu, Guoqiang Li, Sora Chee, Tristin Liu, Anthony D Schmitt, and Bing Ren. Mapping of long-range chromatin interactions by proximity ligation-assisted ChiP-seq. Cell, 26(12), 1345–1348, Nov 2016.
[5] Maxwell B Mumbach, Adam J Rubin, Ryan A Flynn, Chao Dai, Paul A Khavar, William J Greenleaf, and Howard Y Chang. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nature Methods, 13(11), 919–922, sep 2016.
[6] Karin M. Too many transcription factors: positive and negative interactions. New Biol 1990;2:126–31.
[7] Hang Xu, Zhang Shijue, Xi Yanfou, Płewczynski Dariusz, Li Mulin Jun. Exploring 3d chromatin contacts in gene regulation: The evolution of approaches for the identification of functional enhancer-promoter interaction. Comput Struct Biotechnol J 2020;18:558–70.
[8] Eras Góksen, Ayse Ziga, Capnue Julien, Thies Fabian J. Deep learning: new computational modelling techniques for genomics. Nat Rev Genet 2019;20(7):389–403.
[9] Mao Wénguang, Kostka Dennis, Chikina Maria. Modeling enhancer-promoter interactions with attention-based neural networks. EPANN:17;11.
[10] Suhas S.P. Rao, Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochklov, James T. Robinson, Adrian L. Sanborn, Ido Machol, Arina D. Omer, Eric S. Lander, and Erez Lieberman Aiden. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell, 159(7), 1665–1680, Dec 2014.
[11] A user’s guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol, 9(4), e1000406, apr 2011.
[12] Anshul Kundaje, Wouter Meuleman, Jason Ernest, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J. Ziller, Vireni Ammin, John W. Whitaker, Matthew D. Schultz, Luisa D. Ward, Abhishek Sarkar, Gerald Quon, Richard S. Sandstrom, Matthew L. Eaton, Yi-Chieh Wu, Andreas R. Pfenning, Xinchen Wang, Melina Claussnitzer, Vaping Liu, Cristian Coara, R. Alan Harris, Noah Shores, Charles B. Epstein, Elizabeth Gjoneska, Danny Leung, Wei Xie, R. David Hawkins, Ryan Lister, Chibo Hong, Philippe Gascard, Andrew J. Mungall, Richard Moore, Eric Chuaah, Angela Tan, Theresa S. Canfield, R. Scott Hansen, Rajendral Kaul, Peter J. Sabo, Muki S. Bansal, Annaicke Carles, Jesse R. Dixon, Kao-Fhow Fari, Soheil Feizi, Rosa Karlic, Ah-Ram Kim, Ashwininkumar Kulkarni, Daodeng Li, Rebecca Lowndon, GilNell Elliott, Tim R. Mercer, Shane J. Neph, Vitour Ogrucih, Paz Polak, Nisha Rajagopai, Pradip Ray, Richard C. Sallari, Kyle T. Siebenthal, Nicholas A. Sinnott-Armstrong, Michael Stevens, Robert E. Thurman, Jie Wu, Bo Zhang, Xin Zhou, Arthur E. Beaudet, Laurie A. Boyer, Philip L. De Jager, Peggy J. Farnham, Susan J. Fishel, David Haussler, Steven J.M. Jones, Wei Li, Marco A. Marra, Michael T. McManus, Shamil Sunyaev, James A. Thomson, Dea T. Tisty, Li-Huei Tsai, Wei Wang, Robert A. Waterland, Michael O. Zhang, Lisa H. Chadwick, Bradley E. Bernstein, Joseph F. Costello, Joseph R. Ecker, Martin Hirst, Alexander McLawhorn, Alejandro Milanovsievic, Bing Ren, John A. Stamatoyannopoulos, Ting Wang, and Manolis Kellis. Integrative analysis of 111 reference human epigenomes. Nature, 518(7539), 317–330, Feb 2015.
[13] Ken Chen, Huaying Zhao, and Yuedong Yang. Capturing large genomic contexts for accurately predicting enhancer-promoter interactions. bioRxiv, 2021.
[14] Carrie A Davis, Benjamin C Hitz, Cricket A Sloan, Esther T Chan, Jean M Jungreis, James Wright, Joel
Armstrong, If Barnes, Andrew Berry, Alexandr Bignell, Silvia Bonnallé, Jacqueline Christ, Fionia Cunningham, Tomás Di Domenico, Sarah Donaldson, Ian T Fiddes, Carlos García Girón, Jose Manuel González, Tiago Grego, Matthew Harms, Thibault Huet, Tobias Huog, Yi Cuzo, Julião Lagarde, Ferjal F. Martín, Laura Martínez, Shamika Mohanan, Paul Muir, Fabio C. Navarro, Anne Parker, Baikang Pei, Fernando Pozo, Magali Ruffier, Bianca M Schmitt, Eloise Stapleton, Martin Siebert, Else Sycheva, Barbara Uszczynska-Ratajczak, Jiantao Tong, Andrew Yates, Daniel Zerbino, Alexander Ziemek, John H. Zou, Toun H. Choudhary, Mark Gerstein, Roderic Guigó, T J P Hubbard, Manolis Kellis, Benedict Paten, Alexandre Reynold, Michael L Tress, and Paul Fleck. CENCODE reference for the human and mouse genomes. Nucleic Acids Res, 47(1), D766–D773, Oct 2018.

[17] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38(1):164–168. 

[18] Jacob Schreiber, Maxwell Libbrecht, Jeffrey Bilmes, and William Stafford Noble. Nucleotide sequence and dasei sensitivity are predictive of 3D chromatin architecture. bioRxiv, 2017.

[21] Daniel R, Stegle Oliver, Uwehndal Willem H, Frontini Mattia, Wallace Chris, Spivakov Mikhail, Fraser Peter, Martens Joost H, Kim Bowon, Sharif Nilofar, Janssen-Megen Eva M, Yaso Marie-Lauré, Linis Matthias, Kovacsics Alexander, Sara Clark-Lauren, Richard David, Datta Avik, Fleck Paul. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. Cell 2016;167(5):1369–84. e19.

[25] Robert Schlessinger, Clément Touzet, Irene Miguez, Davido, Ilka Hoof, Jorn Bornhold, Mette Yue, Yuen Chen, Xiaochang Zhao, Christian Schmidt, Takahito Suzuki, Evgenia Vtnia, Erik Arner, Eivind Valen, Kang Li, Lucía Schwarzfischer, Dagmar Clatz, Johanna Raathiel, Birte Lilje, Nicolai Rapan, Frederik Otten Bagoss, Mette Jørgensson, Peter Rensg Andersen, Nicolai Bertin, Owen Rackham, A. Maxwell Burroughs, J. Kenneth Baille, Yuli Ishizu, Yumi Shimizu, Erina Furuhata, Shiori Maeda, Yukta Negishi, Christopher J. Mungan, Terence F. Meehan, Timo Lassmann, Masayoshi Ito, Hideya Kawaji, Naoto Kondo, Jun Kawabe, Masahiro U. O. Dahn, Yuji Ando, Masahiro Yamaoka, Torben Heinek Jensen, Harukazu Suzuki, Yoshioe Hayashizaki, Ference Müller, Alistair R.R. Forrest, Piero Cancini, Michael Rehl, and Albin Sandelin. An atlas of active enhancers across human cells and tissue types. Nature (707/4933), 345–461, March 2013.

[24] Fiona Cunningham, M. Ridwan Amode, Daniel Barrett, Kathryn Beal, Konstantinos Bils, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Stephen Fitzgerald, Laurent Gil, Carlos García Girón, Leo Gordon, Thibault Houltrier, Sarah E. Hunt, Sophie H. Janacek, Nathan Johnson, Thomas Juddmann, Terrance A. Kehler, Stephen K. Keenan, Fergal J. Martin, Thomas Maurel, William McLaren, Daniel N. Murphy, Rishi Nag, Bert Overduin, Anne Parker, Mateus Patricio, Emily Perry, Miguel Pignatell, Harpreet Singh Rial, Danielle Stebbins, Kwan Yau, Yi Zhang, and Cédric Zeller. DeepMILO: a deep learning method that integrates DNase I data to capture the conformation of genomes. Cell, 2014;159(12):S11.

[29] Rui Tuan, Marta Funiciule, Alexander Martinez-Fundichely, Ekta Khurana. DeepMILO: a deep learning method that integrates DNase I data to capture the conformation of genomes. Cell, 2014;159(12):S11.

[32] Chiara Cittadini, David K. Parker, and Giacomo Cavalli. Multiscale 3D genome rewiring is associated with human disease. Nature Genet, 53(2):151–156, Feb 2021.

[34] Vicente-Molina Carmen, Jonathan Poot, Yee Whye Teh, and Jim R. Hughes. An integrated platform to systematically identify causal variants and genes for polygenic human traits. Oct 2019.

[35] Xiaoyun Li, Xueqiu Liu, Xueqiu Liu, Zongwu He, Jiaqiang Wang, and Hung-jeng Chen. NGS-MIB: a machine learning-based method for predicting chromatin interaction from DNA sequences. Nucleic Acids Res, 48(20), 2020.

[36] Jiang, Fang, Zhang Shao-Wu, Zhang Shihua. Prediction of enhancer-promoter interactions from DNA looping interactions in capture hi-c data. Genom, 17(1), Jun 2016.

[37] Elise Tian, Xiaoyun Li, and Riki Yamanaka. Enhancer-promoter interactions using the cross-cell type information and domain adversarial neural network. BMC Bioinform 2020;21(1).

[38] Shashank Singh, Yang Yang, Barnabás Póczos, and Jian Ma. Predicting enhancer-promoter interactions from cell- and tissue-specific regulatory SNPs on transcription factor binding from cell- and tissue-specific DNase footprints. Genome Res, 27(10), 1730–1742, sep 2017.

[39] Ahmad Ghaemmaghami, Omar Al Marzooqi, and Morgan F. Smart. Rostock, Steffen Schaible, Milan Martin, Ilya Rode, Yutaka Negishi, Christopher J. Mungan, Zhe Wu, and Alistair R.R. Forrest. HiP: a comprehensive technique to capture the conformation of genomes. Cell, 2015;163(7), 554–565.e7, May 2015.

[40] Jing Fang, Zhang Shao-Wu, Zhang Shihua. Prediction of enhancer-promoter interactions using the cross-cell type information and domain adversarial neural network. BMC Bioinform 2020;21(1).

[41] Shashank Singh, Yang Yang, Barnabás Póczos, and Jian Ma. Predicting enhancer-promoter interactions from cell- and tissue-specific regulatory SNPs on transcription factor binding from cell- and tissue-specific DNase footprints. Genome Res, 27(10), 1730–1742, Sep 2017.
