Predicting 3D genome folding from DNA sequence

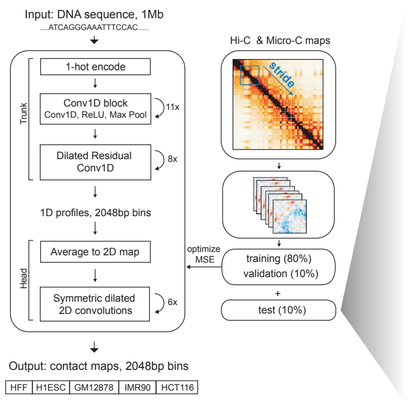
Geoff Fudenberg*, David R Kelley*, Katherine S Pollard
geoff.fudenberg@gladstone.ucsf.edu, drk@calicolabs.com, katherine.pollard@gladstone.ucsf.edu
@gfudenberg, @drklly

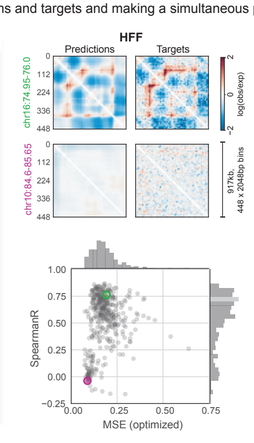
Abstract. In interphase, the human genome sequence folds in three dimensions into a rich variety of locus-specific contact patterns. The high-resolution views offered by genome-wide chromosome conformation capture techniques (e.g.Hi-C and Micro-C) have advanced our understanding of the proteins and sequences driving 3D genome folding, including the interplay between CTCF and cohesin and their roles in development and disease. Still, predicting the consequences of perturbing any individual CTCF site, or other regulatory element, on local genome folding remains a challenge. While disruptions of single bases can alter genome folding, in other cases genome folding is surprisingly resilient to large-scale deletions and structural variants. Convolutional neural networks (CNNs) have emerged as powerful tools for modelling genomic data as a function of DNA sequence, directly learning DNA sequence features from the data. CNNs now make state-of-the-art predictions for transcription factor binding, DNA accessibility, and transcription. Here we present Akita, a CNN that accurately predicts genome folding from DNA sequence alone. Representations learned by Akita underscore the importance of CTCF and reveal a complex grammar underlying genome folding. Akita enables rapid in silico predictions for sequence mutagenesis, genome folding across species, and genetic variants. In the future, we envision that end-to-end sequence-to-genome-folding approaches that build upon Akita will advance our ability to design functional screens, model enhancer-promoter interactions, prioritize causal variants in association studies, and predict the impacts of rare and de novo variants.

Trained models & open-source code available at: https://github.com/calico/basenji/tree/master/manuscripts/akita.

Akita makes locus-specific predictions for 3D genome folding from DNA sequence.

Akita consists of a 'trunk,' based on the Basenji architecture (Kelley et al. 2018), followed by a 'head' to transform to 2D maps of genome folding. The trunk involves: (i) input 1Mb of 1-hot encoded DNA; (ii) 1D convolution trunk, where each block performs a max pool operation between adjacent positions to iteratively reduce to a bin size of 2048 bp; (iii) dilated residual 1D convolutions to propagate local information across the sequence. The 'head' involves: (i) forming 2D maps from the 1D vectors by averaging each pair of vectors at positions (i, j); (ii) symmetric dilated residual 2D convolutions; (iii) dense layer with linear activation to predict log(observed/expected) chromosome contact maps, with one separate output per dataset. We considered 2048bp binned maps, as high-quality Hi-C and Micro-C datasets ascertain genome folding at this resolution with tractable technical variance. We compared upper triangular regions of maps cropped by 32 bins on each side, making symmetric predictions for 448x448 bin (~917kb) maps. We trained our model on regions of the genome obtained by striding along Hi-C maps, using an 80/10/10 training/validation/test split. We trained Akita with five of the highest-quality Hi-C and Micro-C datasets as targets, focusing on the locus-specific patterns evident in log(observed/expected) maps, minimizing the mean squared error (MSE) between predictions and targets and making a simultaneous prediction for each of these five maps.



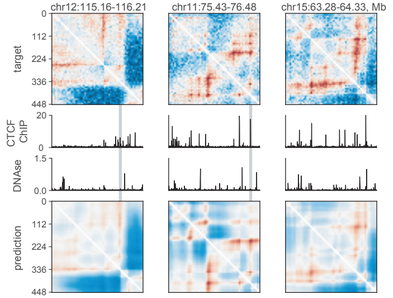


Predicted and experimental log(observed/expected) contact frequency for two representative regions in the test set for Human Foreskin Fibroblast (HFF) Micro-C (Krietenstein et al. 2019).

Quantification for the held-out test set: mean-squared error (MSE), which we optimize in model training, versus Spearman R, both calculated per region for each pair of targets and predictions for HFF Micro-C. Green and purple circles show regions above. Note correlations display a bimodal shape: regions with few locus-specific features have low MSE and low Spearman R.

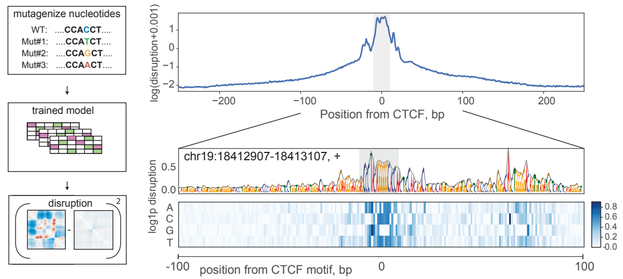
Predicted patterns often aligned with CTCF ChIP-seq peaks

Akita predicted more prominent patterns in regions with greater CTCF binding and DNAse hypersensitivity. Salient predicted patterns often also aligned with CTCF ChIP-seq peaks (*right,* from the ENCODE data portal, Davis et al., 2018). However, CTCF motifs are too prevalent to connect DNA sequence to genome folding at the bin level. Fortunately, Akita enabled us to directly quantify nucleotide influences via in silico mutagenesis; while training Akita was computationally intensive, effects of sequence changes could be predicted in seconds. Akita predicted greatly diminished locus-specific patterns after mutating all CTCF motifs . In this extreme scenario, Akita predicted some patterns would persist, and these often aligned with DNase hypersensitive sites that lacked evidence of strong CTCF binding.



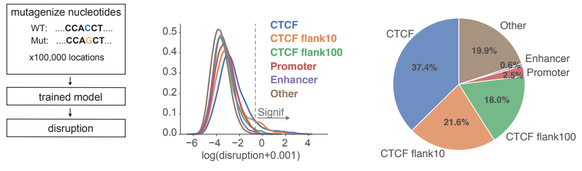
Akita extracts informative base pair level features of genome folding.

Given the substantial predicted impact of mutagenizing whole CTCF motifs on genome folding, we sought to quantify the predicted contact map disruptions for mutations to individual nucleotides. We performed in silico saturation mutagenesis of 500-bp regions centered at strong CTCF motifs (JASPAR p-value < 1e-6, Khan et al., 2018). Predicted disruptions were largest for nucleotides around the motif, but remained high relative to background in the flanking regions, slowly decaying with increasing distance. Profile shows the average score for each position after taking the maximum across alternative alleles. CTCF motif position is indicated in grey. Zoom-in (*below*) shows an motif with high disruption scores in flanking regions. Heatmap shows scores for each possible nucleotide substitution. Nucleotide letter heights are drawn proportional to the max across three possible substitutions per position.



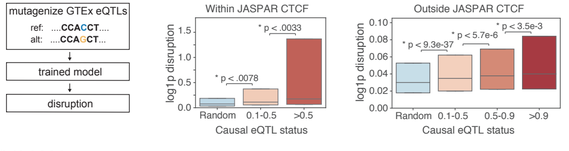
Genome-wide mutagenesis reveals contributions beyond core CTCF motifs.

To quantify the influence of nucleotides within and near CTCF motifs relative to other genomic features we generated 100,000 random single nucleotide mutations uniformly spaced across the test set. Mutations that altered CTCF motifs and their flanking regions displayed many of the highest predicted disruption scores, which likely reflect influences on CTCF binding or function, either directly or via cofactors. To visualize this we split significant mutations (threshold inidicated by vertical line) into annotation categories, excluding previous categories in the hierarchy. Categories are considered hierarchically counter clockwise, starting from those that influence CTCF motifs (CTCF, Flank10, Flank100, Promoter, Enhancer, Other). Flank10 and Flank100 represent nucleotides falling within 10 or 100bp of a CTCF motif. This conservative categorization provides strong evidence for the contribution of nucleotides beyond canonical CTCF sites for genome folding.



Akita enables systematic GTEx eQTL mutagenesis

To gain insight into how genome folding influences gene expression, we studied a set of fine-mapped likely causal eQTLs from GTEx whole blood samples. Using Akita, we calculated the predicted disruption to local 3D folding for eQTLs at varying causal posterior probability thresholds. We observed significantly larger predicted disruptions for single nucleotide variants (SNPs) with higher causal eQTL status, both for SNPs overlapping and outside of CTCF motifs. Plots show results for 1,906 SNPs with causal posterior probability >0.9, 1,844 SNPs from 0.5 to 0.9, and 16,064 SNPs from 0.1 to 0.5 and 9,298 random set of control SNPs with significant genome-wide marginal association with gene expression.



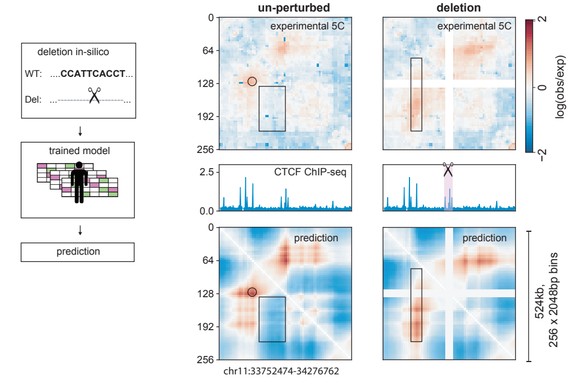
References

For full list of references and methods please see bioRxiv doi: https://doi.org/10.1101/800060

Davis, C. A. et al. Nucleic Acids Res. 46, D794–D801 (2018).
Hnisz, D. et al. Science 351, 1454–1458 (2016).
Kaaij, L. et al. Cell 178, 1437–1451.e14 (2019).
Kelley, D. R. et al. Genome Res. 28, 739–750 (2018).
Khan, A. et al. Nucleic Acids Res. 46, D260–D266 (2018).
Kraft, K. et al. Nat. Cell Biol. 21, 305–310 (2019)
Krietenstein, N. et al. Mol. Cell (2020)
Wang, G., et al. bioRxiv 501114 (2019)

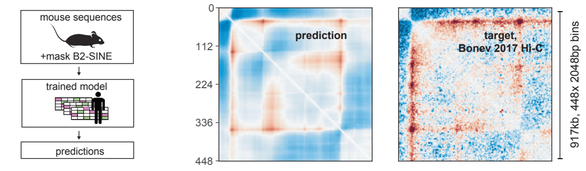
Predicting a genetically engineered deletion

At the Lmo2 locus in HEK293T cells, two domains are separated by a boundary positioned at a cluster of three CTCF-bound sites on chromosome 11 (Hnisz et al., 2016). In wild-type cells (left), this region displays a peak at the boundary (circle) between two ~130kb domains that are relatively insulated from each other (rectangle), separated by a boundary that overlaps a cluster of three CTCF-bound sites. In cells where this boundary has been deleted (right), the two domains merge and display a flare of enriched contact frequency (thin rectangle). Middle: CTCF profiles for HEK293T27. Bottom: Computational predictions for WT (left) and deletion (right) of the boundary, using the HFF output from our human-trained model, showing similar changes.



Predicting mouse genome folding

Given similar overall human and mouse genome folding, we reasoned the mouse genome could provide evolutionarily perturbed sequences to further test Akita. Using mouse DNA sequences as input, we compared predictions from our human-trained model (hESC output) with mESC Hi-C data. These cross-species predictions rrecapitulated some aspects of mouse genome folding. Intriguingly, poorer predictions had more B2 SINE elements, which dramatically expanded in murid lineages and carry CTCF sites. Mutagenizing B2 SINE elements to ablate their CTCF sites improved our predictions for mouse genome folding (median Spearman R 0.55 vs 0.50). This suggests that the mouse genome specifically mitigates the influence of these elements, consistent with recent experimental observations (Kaaji et al., 2019)



Predicting a genetically engineered inversion

To further test the generality of our approach, we trained a model with the same architecture using mouse Hi-C as target data and mouse DNA sequence as input. This model was both more predictive than the human-trained model on held-out test regions of the mouse genome and was not improved by mutating B2 SINE elements. Together this indicated it correctly learned to mitigate the impact of CTCF sites inside of these elements. Using this model, we then considered its ability to predict an engineered inversion in the mouse genome. At the Eph4A locus in limb buds, two domains are separated by a boundary with a prominent downstream flare (*WT below*). Upon inversion of ~622kb encompassing this boundary and a downstream enhancer, the orientation of the flare flips (Kraft et al., 2019). We found a similar predicted change in silico. This result illustrates the generality of our approach, for both a new organismal context (mouse instead of human) and class of structural variant (inversion instead of deletion).