Predicting 3D genome folding from DNA sequence

Geoff Fudenberg*, David R Kelle}*1, Katherine S Pollard
geoff.fudenberg@gladstone.ucsf.edu, dk@calico.com, katherine.pollard@gladstone.ucsf.edu
@gladstone, @dkrily

Abstract.

In the mouse genome, the human genome sequence folds three-dimensional 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 enhancers 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 these loci 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 modeling 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 encode the importance of CTCF and reveal a complex grammar underlying genome folding. Akita enables rapid in silico predictions for sequence perturbations, genome folding across species, and genetic variants. In the future, we envision that end-to-end sequence-prediction 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.

Akitas makes locus-specific predictions for 3D genome folding from DNA sequence. Akita consists of a 3D (box) architecture, based on the ResNet architecture (Kathy et al., 2019), that is trained to fold 2D maps of genome contact. 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 sequences; (iii) 2D convolution trunk, which performs 6x6x6x6x6x6 convolution in a single pass; (iv) classification of the 2D convolution to predict target 3D maps; (v) dense layer with linear activation to predict log(1+observed/expected) contact maps, with one output per genome region. We trained Akita for each genome region by slicing along Hi-C maps, using an 80% training/validation/test split. We trained Akita for 448x448 input maps, minimizing the mean squared error (MSE) between predictions and targets and making a simultaneous prediction for each of these five maps.

Akitas [pseudo] maps are converted to contact maps, which are then used to visualize changes in genome folding. Akitas makes predictions that can be used to visualize changes in genome folding. Akitas makes predictions that can be used to visualize changes in genome folding. Akita extracts informative base pair level features of genome folding. Given the substantial predicted impact of mutating whole CTCF motifs on genome folding, we sought to quantify the predicted context-dependent disruptions for mutations at individual nucleotides. We performed in silico mutagenesis of 500-bp regions centered at the 120 CTCF sites with the most disruptive disruptions in this experimental neighborhood. Predictive disruptions were largest for nucleotides around the motif, but remained high relative to background in the neighborhood. Disruptions were direction-specific, with more nucleotides showing an increase in predicted nucleosome substitution after taking the maximum across alternative alleles. CTCF motif positions are indicated in grey. Zoom-in (below) shows an increase in high disruption around CTCF motifs. Repeating this analysis for each nucleotide in the genome enabled rapid in silico predictions for sequence mutagenesis, genome folding across species, and genetic variants. In the future, we envision that end-to-end sequence-prediction 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.

Akitas are trained to fold 2D maps of DNA sequence, and predict the impacts of rare and de novo variants. 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.

Published by the California Life Sciences Institute. Copyright 2019 California Life Sciences Institute. All rights reserved.

References.

For a list of references and methods please see below.

https://doi.org/10.1371/journal.pgen.1007938.s001

Predicting a genetically engineered deletion

At the locus of interest in Hi-C 2017 Gladiolus U08 data, two domains are separated by a boundary predicted at a cluster of three CTCF sites (112 probe sets). 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 extends over a cluster of three CTCF ChIP-seq sites. In cells making this boundary, tail extented (right), the two domain margin and display a flare of enriched contact frequency (this rectangle). Middle: CTCF profile for Hi-C2017 Gladiolus U08, CTD Topez prediction for WT (left), and deletion (right) of the boundary, using the HIF output from our human-trained model, showing similar change.

.. figure:: Images/003.png

    Predicting a genetically engineered deletion

| CTCF flank10 | CTCF flank100 | Promoter | Enhancer | Other |
|-------------|--------------|----------|----------|-------|
| 0.6%        | 0.2%         | 21.6%    | 0.6%     | 0.7%  |

Gene regulation.

We quantified whether our predictions could recapitulate historical results. Using mouse predicted disruption for local Hi-C folding, we compared these predictions to published disrupting effects from human-chimp sequence alignment. Given the substantial predicted impact of mutating whole CTCF motifs on genome folding, we sought to quantify the predicted context-dependent disruptions for mutations at individual nucleotides. We performed in silico mutagenesis of 500-bp regions centered at the 120 CTCF sites with the most disruptive disruptions in this experimental neighborhood. Predictive disruptions were largest for nucleotides around the motif, but remained high relative to background in the neighborhood. Disruptions were direction-specific, with more nucleotides showing an increase in predicted nucleosome substitution after taking the maximum across alternative alleles. CTCF motif positions are indicated in grey. Zoom-in (below) shows an increase in high disruption around CTCF motifs. Repeating this analysis for each nucleotide in the genome enabled rapid in silico predictions for sequence mutagenesis, genome folding across species, and genetic variants. In the future, we envision that end-to-end sequence-prediction 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.