Predicting 3D genome folding from DNA sequence

Anonymous Authors[1]

Abstract

In interphase, the human genome sequence folds in three dimensions into a rich variety of locus-specific contact patterns. Here we present a deep convolutional neural network, Akita, 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.

Preprint available at: https://www.biorxiv.org/content/10.1101/800060v1. Trained models, open-source code, and documentation for Akita available at: https://github.com/calico/basenji/tree/master/manuscripts/akita.

1. Introduction

In interphase, the human genome sequence folds in three dimensions into a rich variety of locus-specific contact patterns. Recent research has 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 (Merkenschlager & Nora, 2016). 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 (Despang et al., 2019; Rodriguez-Carballo et al., 2017). 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 (Alipanahi et al., 2015; Zhou & Troyanskaya, 2015; Kelley et al., 2016).

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review at the ICML 2020 Workshop on Computational Biology (WCB). Do not distribute.

2. Main

Here we present Akita, a CNN that accurately predicts genome folding from DNA sequence alone. Akita takes 1Mb (2^{20} bp) of DNA sequence as input and predicts contact frequency maps for all pairs of 2kb (2048bp) bins within this region. Crucially, this allows Akita to predict the effect of mutating single base pairs. 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. The Akita architecture consists of a 'trunk' based on the Basenji (Kelley et al., 2018) architecture to obtain 1D representations of genomic sequence, followed by a 'head' to transform to 2D maps of genome folding (Fig. 1a).

Akita learned a predictive representation of genome folding from DNA sequence, as evaluated on the held-out test set (0.61 Pearson R), approaching the limit set by noise between experimental replicates. On a region-by-region basis, Akita captured the variety of patterns seen experimentally (Fig. 1b). In silico mutagenesis and inversions of CTCF motifs indicated that Akita learned an orientation-specific grammar of the CTCF sites most crucial for genome folding (Fig. 1c), consistent with experimental results.

Leveraging Akita's ability to make rapid predictions for single-base pair perturbations, we studied the influence of fine-mapped eQTLs from GTEx on genome folding. We calculated the predicted disruption to local 3D folding for eQTLs at varying causal posterior probability (PP) thresholds. We observed significantly larger predicted disruptions for single nucleotide variants (SNPs) with greater causal eQTL PP, both for SNPs overlapping and outside of CTCF motifs. Akita also displayed predictive utility for larger genetically engineered structural variants. At the Lmo2 locus in HEK293T cells (Hnisz et al., 2016), two domains are separated by a boundary positioned at a cluster of three CTCF-bound sites (Fig. 1d). In cells with a 25kb deletion encompassing this boundary, the two domains merge. Making the same deletion in silico recapitulated this effect in the predicted Hi-C map.

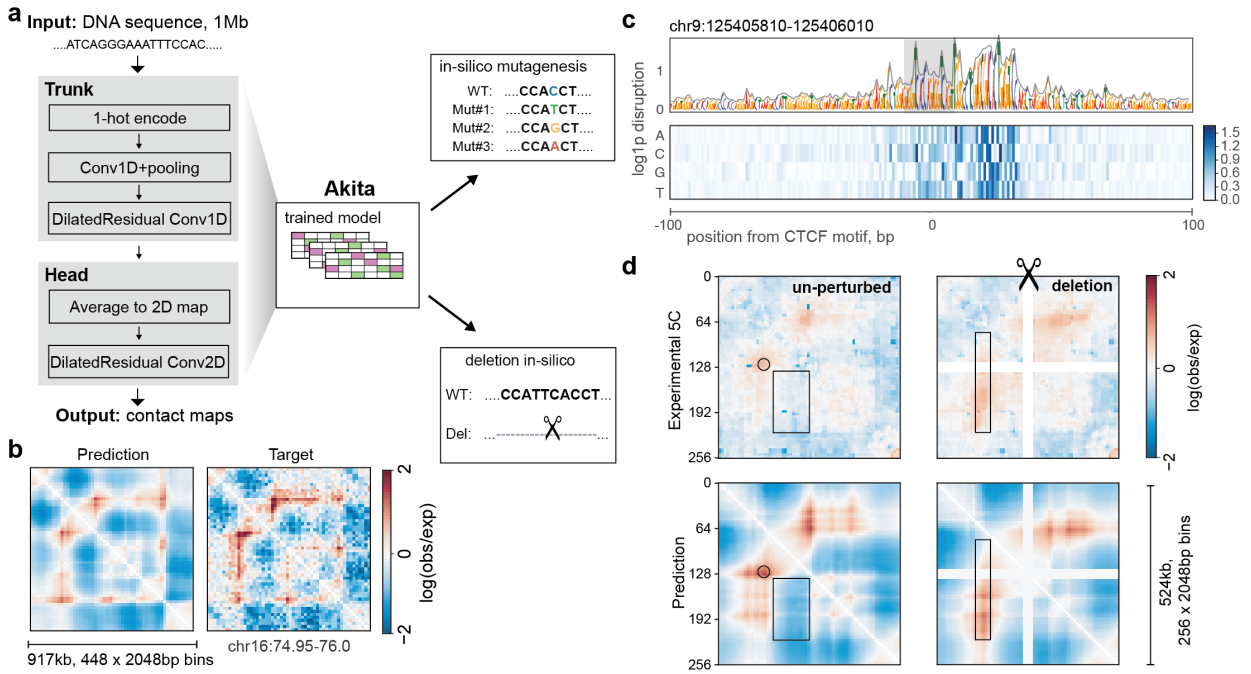

Figure 1. **a.** Akita consists of a 'trunk,' based on the Basenji architecture, followed by a 'head' to transform 1D DNA sequence into 2D maps of genome folding. **b.** Predicted and experimental (Krietenstein et al., 2020) log(observed/expected) contact frequency for a representative region in the test set. **c.** In silico mutagenesis around a strong CTCF motif revealed high disruption scores in flanking regions. Core motif shown in grey. Disruption computes the L2 norm of the difference between contact maps. Motif positions from JASPAR (Khan et al., 2018) **d.** Top: Experimental (Hnisz et al., 2016) 5C data in HEK293T cells for wild-type (*left*) and a CRISPR/Cas9-mediated deletion of a 25kb boundary region (*right*) at the Lmo2 locus. In wild-type cells (*left*), this region displays a peak at the boundary (circle) between two 130kb domains that are 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). Bottom: Computational predictions for WT (*left*) and deletion (*right*) of the boundary, showing similar changes.

3. Outlook

In the future, end-to-end sequence-to-genome-folding approaches 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.

References

Alipanahi, B. et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, 33(8):831–838, August 2015.

Despang, A. et al. Functional dissection of the Sox9-Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. *Nat. Genet.*, 51(8):1263–1271, August 2019.

Hnisz, D. et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, 351(6280): 1454–1458, March 2016.

Kelley, D. R. et al. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.*, 26(7):990–999, July 2016.

Kelley, D. R. et al. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.*, 28(5):739–750, May 2018.

Khan, A. et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.*, 46(D1):D260–D266, January 2018.

Krietenstein, N. et al. Ultrastructural details of mammalian chromosome architecture. *Mol. Cell*, March 2020.

Merkenschlager, M. and Nora, E. P. CTCF and cohesin in genome folding and transcriptional gene regulation. *Annu. Rev. Genomics Hum. Genet.*, 17:17–43, August 2016.

Rodriguez-Carballo, E. et al. The HoxD cluster is a dynamic and resilient TAD boundary controlling the segregation of antagonistic regulatory landscapes. *Genes Dev.*, 31 (22):2264–2281, November 2017.

Zhou, J. and Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods*, 12(10):931–934, October 2015.