Machine learning for genomic discovery

Anshul Kundaje

Twitter: @anshulkundaje

Website: http://anshul.kundaje.net
Decoding regulatory DNA

Active gene

Repressed gene

Motif syntax: rules of
- Composition
- Affinity
- Arrangement
- Spacing
- Orientation
=> cooperativity

Regulatory control elements

Protein

https://www.broadinstitute.org/news/1504
Mapping biochemical markers of regulatory activity

Active gene

Repressed gene

Protein

99 % Non-coding

1.5 % Protein Coding

https://www.broadinstitute.org/news/1504
100s of Cell-Types/Tissues

Machine learning models

Decoding regulatory DNA sequence

Predicting functional genetic variant & mutations
BPNet: Mapping DNA sequence to base-pair resolution profiles

Ziga Avsec et al. 2021 Nature Genetics
BPNet predicts reg. profiles from sequence with unprecedented accuracy

TF ChIP-exo

Oct4
- Obs
- Pred

Sox2
- Obs
- Pred

Nanog
- Obs
- Pred

Klf4
- Obs
- Pred

TF ChIP-seq

Obs

Pred

DNase-seq / ATAC-seq / pseudo-bulk scATAC-seq

Obs

Pred
Opening up the blackbox
DeepLIFT: Inferring predictive nucleotides at individual binding events

Avanti Shrikumar

Shrikumar et al. ICML 2017
Lundberg et al. NeurIPS 2017
TF-MoDISCO: Cluster and consolidate predictive subsequences into contribution weight matrix (CWM) motifs

Insight: conv. filter contributions are integrated at the nucleotide level

CODE: https://github.com/kundajelab/tfmodisco
Complex repertoire of motifs due to cooperative binding

50 motifs for 4 TFs!

Subtle differences in Nanog motifs
Subtle low affinity patterns with helical periodicity flanking Nanog motif

10 bp periodic binding of homeobox TFs to nucleosome DNA from recent *in vitro* NCAP-SELEX data (Zhu et al. Nature 2018)
Soft syntax: helical spacing preference between Nanog motifs in the genome

Motif pairwise distance

Motif pairwise distance

Motif pairwise distance
Using the model as an “oracle” to perform large-scale *in-silico* experiments
Deciphering syntax dependent TF cooperativity with synthetic designed sequences

![Graph showing TF cooperativity and footprint height](image)

- **Nanog**
- **Oct4-Sox2**

Motif distance = 157

Fold-change in footprint height:
- **Oct4-Sox2 → Nanog**
- **Nanog → Oct4-Sox2**
Deciphering syntax dependent TF cooperativity with *in-silico* genome editing

Footprint strength of target TF

Distance between motifs (bp)
Using the model to design CRISPR experiments to validate discoveries
Model-driven prioritization of functional genetic variation
Predicting and interpreting variants influencing multiple layers of regulatory activity

**Predicted SPI1 TF ChIP-seq**

- **ref=C**
- **alt=G**

**Predicted DNase-seq**

- **ref=C**
- **alt=G**

**Predicted H3K27ac ChIP-seq**

- **ref=C**
- **alt=G**

Model interpretation predicts sequence drivers
Prioritizing putative causal variants in disease-associated loci

rs1237999 in the PICALM locus for Alzheimer’s disease GWAS disrupts an oligodendrocyte-specific FOS enhancer
Summary & Outlook

• Predictive blackbox models + interpretation frameworks
  • Prediction, de-noising & Imputation
  • Biological discovery of causal phenomena
  • Hypothesis generation and optimized experimental design

• Important to be transparent about the limits, blind spots, biases & pitfalls of each model

• What do we need?
  • Large-scale, harmonized ML-ready observational and perturbational data
  • Decentralized, scalable, affordable compute resources
  • Unified ecosystem: Compute ⇔ Data Portals ⇔ Model Zoos ⇔ Literature Mining
  • New user-interfaces to models for interactive discovery, search and design
  • Incentivizes collaborative efforts & diverse contributions
Democratizing ML for genomics: [http://kipoi.org/](http://kipoi.org/)

Kipoi (pronounce: kípi; from the Greek κήποι: gardens) is an API and a repository of ready-to-use trained models for regulatory genomics. It currently contains 1709 different models, covering canonical predictive tasks in transcriptional and post-transcriptional gene regulation. Kipoi’s API is implemented as a python package ([github.com/kipoi/kipoi](https://github.com/kipoi/kipoi)) and it is also accessible from the command line or R.

**Numbers**

- # of models: 1709
- # of model groups: 16
- # of contributors: 6
- # of model groups supporting postprocessing:
  - Variant effect prediction: 11/16

**Model groups by tag**

- Easy installation of dependencies
- Few lines of code to use models to predict
- Exactly reproduce analyses
- Trivial to compare models
- Retrain models
- Fine tune models
- Combine models
- Contribute models

Avsec et al. 2019 Nature Biotech
Kundaje lab

Collaborators

Funding

1R01HG009674
1U01HG009431
1U24HG009446
R01ES02500902
1DP2OD022870