Interpreting noncoding variants

BMI/CS 776
www.biostat.wisc.edu/bmi776/
Spring 2016
Anthony Gitter
getter@biostat.wisc.edu
Goals for lecture

Key concepts

• Mechanisms disrupted by noncoding variants
• Epigenetic data and transcriptional regulation
• Deep learning to predict epigenetic impact of noncoding variants
GWAS output

• GWAS provides list of SNPs associated with phenotype

• SNP in coding region
  – Link between the protein and the disease?

• SNP in noncoding region
  – What genes are affected?
Noncoding variants common in GWAS

- Meta-analysis of GWAS for 21 autoimmune diseases
  - Rheumatoid arthritis, lupus, multiple sclerosis, etc.
- Method to prioritize candidate causal SNPs
- 90% of causal variants are noncoding

Farh Nature 2015
Almost all single nucleotide variants in cancer are noncoding

However, very few of these are driver mutations.
Ways a noncoding variant can be functional

- Disrupt DNA sequence motifs
  - Promoters, enhancers
- Disrupt miRNA binding
- Mutations in introns affect splicing
- Indirect effects from the above changes

Examples in Ward and Kellis *Nature Biotechnology* 2012
Variants altering motifs

*Khurana Nature Reviews Genetics 2016*
Variants affect proximal and distal regulators

- Enhancers are regulatory elements 10-100 kb from target gene

Khurana Nature Reviews Genetics 2016
Mapping regulatory elements genome-wide

- Genome-wide motif scanning is imprecise
- ChIP-seq measures one transcription factor (TF) at a time
- Epigenetic data suggests where some TF binds

Shlyueva Nature Reviews Genetics 2014
DNase I hypersensitivity

- Regulatory proteins bind accessible DNA
- DNase I enzyme cuts open chromatin regions that are not protected by nucleosomes

Wang *PLoS ONE* 2012
Histone modifications

• Mark particular regulatory configurations

  Chromatin as accessibility barrier

  Closed ↔ Open or accessible

  Active enhancer

  Active promoter

  Enhancer

  Core promoter

  DNA-binding proteins: TFs, CTCF, repressors and polymerases

  H3K4me1, H3K4me3, H3K27ac, H3K27me3

• H3 (protein) K27 (amino acid) ac (modification)
Large-scale epigenetic maps

- Epigenomes are condition-specific
- Roadmap Epigenomics Consortium and ENCODE surveyed over 100 types of cells and tissues
Evidence used to prioritize noncoding variants

Interpreting GWAS signals using functional and comparative genomics datasets

Dissect associated haplotype using functional genomics

Dissect associated haplotype using regulatory genomics

Dissect associated haplotype using comparative genomics

Chromatin state annotations

Motifs altered by variants

Mammalian constraint

Ward and Kellis Nature Biotechnology 2012
Visualizing evidence

Data supporting chr11:5246957 (rs33914668)

Score: 2a
Likely to affect binding

Summary of evidence

Genes
Epigenetic annotations
Conservation
Affected motifs

Boyle Genome Research 2012
Combined Annotation–Dependent Depletion (CADD)

• Example of an algorithm that integrates multiple types of evidence into a single score
  – Conservation
  – Epigenetic information
  – Protein function scores for coding variants

• Train support vector machine on simulated and observed variants

• Variants present in simulation but not observed are likely deleterious

Kircher *Nature Genetics* 2014
Prioritizing variants with epigenetics summary

+ Disrupted regulatory elements one of the best understood effects of noncoding SNPs
+ Make use of extensive epigenetic datasets
+ Similar strategies have actually worked
  • rs1421085 in *FTO* region and obesity
  • Claussnitzer *New England Journal of Medicine* 2015

- Epigenetic data at a genomic position is often in the presence of the reference allele
  • Don’t have measurements for the SNP allele
DeepSEA

• Given:
  – A sequence variant and surrounding sequence context

• Do:
  – Predict TF binding, DNase hypersensitivity, and histone modifications in multiple cell and tissue types
  – Predict variant functionality

Zhou and Troyanskaya *Nature Methods* 2015
Classifier input and output

• Output
  – 200 bp windows of genome
  – Label 1 if window contains peak
  – Label for each epigenetic data type
    • Multiple types of epigenetic features
    • Multiple types of cells and tissues

• Input: 1000 bp DNA sequence centered at window

\[ x_i = \begin{array}{ccccccc}
  \text{index} & 1 & \ldots & 401 & 402 & 403 & \ldots & 1000 \\
  A & 0 & \ldots & 1 & 0 & 0 & \ldots & 0 \\
  C & 0 & \ldots & 0 & 0 & 0 & \ldots & 1 \\
  G & 1 & \ldots & 0 & 1 & 1 & \ldots & 0 \\
  T & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\
\end{array} \]
Desired properties for epigenomic classifier

• Learn preferences of DNA-binding proteins
  – Locally: “motifs”
  – Sequence context: “cis-regulatory modules”

• Support nonlinear decision boundaries

• Multiple, related prediction tasks

Neph *Nature* 2012

Roadmap Epigenomics Consortium *Nature* 2015
Perceptron

- Inspired by neuron

- Simple binary classifier
  - Linear decision boundary

\[ o = \begin{cases} 
1 & \text{if } w_0 + \sum_{i=1}^{n} w_i x_i > 0 \\
0 & \text{otherwise}
\end{cases} \]

\[ x_{1,A} \cdots x_{1000,T} \]
Neural networks

• Single perceptron not useful in practice

• Neural network combines layers of perceptrons
• Learn “hidden” features
• Complex decision boundary
• Train with backpropagation
  – Stanford’s CS231n materials
  – Andrej Karpathy’s gentle introduction
  – CS 760 slides
Activation function

• What makes the neuron “fire”?
  – Step function
    \[ f(x) = \begin{cases} 
      0 & \text{if } x < 0 \\
      1 & \text{if } x \geq 0 
    \end{cases} \]
  – Sigmoid function
    \[ f(x) = \frac{1}{1 + e^{-x}} \]
  – Rectified linear unit (ReLU)
    \[ f(x) = \max(0, x) \]
First hidden layer

• First hidden layer scans input sequence
• Activation function fires if motif is recognized

Motif width (window size) \( s = 6 \)

Sequence length \( L \)

\[
x = \begin{bmatrix}
A & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
C & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\
G & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
T & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
\end{bmatrix}
\]
First hidden layer

- Multiple hidden nodes to recognize different motifs at a particular position
- Check for motif at each position in sequence

A hidden node with its own weight vector

$D$ motifs (hidden layer depth)

$W = L - s + 1$ starting positions
First layer problems

• We already have a *lot* of parameters
  – Each hidden node has its own weight vector

• We’re attempting to learn different motifs at each starting position
Convolutional layers

- Input sequence and hidden layer as matrices
- Share parameters for all hidden nodes in a row
  - Search for same motif at different starting positions
Pooling layers

• Account for sequence context

• Multiple motif matches in a cis-regulatory module

• Search for patterns at a higher spatial scale
  – Fire if motif detected anywhere within a window
Pooling layers

- Take max over window of 4 hidden nodes

\[ D \times \left( \frac{W}{4} \right) \]

\[ D \times W \]
Subsequent hidden layers

- Next convolutional hidden layer on top of pooling layer

\[ D' \] is new number of patterns
\[ s' \] is new window size
\[ W' = (W / 4) - s' + 1 \]

Once again, shared weight vector for all nodes in a row
Full DeepSEA neural network

- Multitask output makes simultaneous prediction for each type of epigenetic data
- ReLU activations

![Diagram of DeepSEA neural network with layers: input sequence, convolutional layers, pooling layers, fully connected layer, and 919 classes.](image)
Evaluating predictions

- Calculate area under receiver operating characteristic curve (AUC)
- True Positive Rate versus False Positive Rate

\[ TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \]

\[ FPR = \frac{FP}{N} = \frac{FP}{FP + TN} \]
Predicting epigenetic annotations

- Compute median AUC for three types of classes

Zhou and Troyanskaya *Nature Methods* 2015
Predicting allele-specific DNase hypersensitivity

• Allelic imbalance
  – Heterozygous locus
  – Observe one allele much more frequently in DNase-seq data

• Color indicates true allele bias
  – Blue = reference
  – Red = alternative

Zhou and Troyanskaya Nature Methods 2015
Predicting functional variants

• Can predict epigenetic signal for any novel variant (SNP, insertion, deletion)

• Define novel features to classify variant functionality
  – Difference in probability of signal for reference and alternative allele

• Train on SNPs annotated as regulatory variants in GWAS and eQTL databases
Predicting functional variants

Zhou and Troyanskaya Nature Methods 2015
DeepSEA summary

- Ability to predict how unseen variants affect regulatory elements
- Accounts for sequence context of motif
- Multitask learning to improve hidden layer representations

- Does not extend to new types of cells and tissues
- Use of AUC for evaluation has been questioned
Deep learning in genomics

• DeepSEA
  – http://deepsea.princeton.edu/

• DeepBind
  – Alipanahi *Nature Biotechnology* 2015
  – http://tools.genes.toronto.edu/deepbind/

• Basset
  – Kelley bioRxiv 2016 http://dx.doi.org/10.1101/028399
  – https://github.com/davek44/Basset
  – Best documentation