BST281: Genomic Data Manipulation, Spring 2017

Monday 11: Transcriptional regulatory motif discovery

Start planning for final projects - sign up group + topic no later than next Monday.

 30 minute group presentations + code+data package submissions, plus 2+ page individual writeups.

 Leave time for iteration - ~2 paragraph project proposals typically require iteration.

Regulatory factor binding site discovery from primary sequence is difficult.

 Combinatorial among factors, sites, locations, and distances (cis- versus trans-).

When identifying from primary sequence, look for overabundant (enriched) motifs in regulatory regions.

A "motif" itself, which is generally a multiple sequence alignment, can be represented in one of several ways:

 Consensus sequence: independently most common base at each position.

 Degenerate consensus sequences, which use special nucleotides when ambiguous (or regular expressions).

 Position weight matrices (PWMs): probabilities rather than counts at each position.

 Position-specific scoring matrices (PSSMs): log-ratios of probabilities relative to background.

 Logos are the bit scores of these PSSMs that capture how different they are from background, easy to scan.

Learning motifs from sequence is an optimization problem - neither best motif nor best sequences known.

 Gibbs sampling iteratively re-computes one sequence at a time.

 Expectation Maximization (EM) iteratively recomputes all sequences simultaneously.

Chromatin immunoprecipitation (ChIP) pulls down DNA sequences bound by a factor of interest.

 Followed by microarray binding (ChIP-chip) or sequencing (ChIP-seq) to see what's bound.

 Does not differentiate functional from nonfunctional binding!

Best motif finders integrate data types: enrichment, binding, conservation, expression.

miRNAs an additional small RNA mechanism of sequence-based post-transcriptional degradation / inhibition.

miRNA genes found in unusual places, discovered by scanning genome for RNA structures.

miRNA targets often found in 3' UTRs, discovered by searching for seed complementarity, RNA structure.

Regulatory networks / circuits are very difficult to reconstruct globally, can be modeled using:

 Bayes networks - probabilistic graphical models of causes (regulators) and effects (targets).

 Biophysical models, e.g. differential equations.

 Linear models, in which amount of target is a linear function of amounts of regulators (positive or negative).

# Textbooks

Regulatory sequences: Pevsner, Chapter 10 p307-345

# Literature

[An improved map of conserved regulatory sites for Saccharomyces cerevisiae. MacIsaac, BMC Bioinformatics 2006](https://www.ncbi.nlm.nih.gov/pubmed/16522208)

[A user's guide to the encyclopedia of DNA elements (ENCODE). ENCODE, PLoS Biology 2011](https://www.ncbi.nlm.nih.gov/pubmed/21526222)

[Inferring direct DNA binding from ChIP-seq. Bailey, NAR 2012](https://www.ncbi.nlm.nih.gov/pubmed/22610855)

[P-value-based regulatory motif discovery using positional weight matrices. Hartmann, Genome Research 2013](https://www.ncbi.nlm.nih.gov/pubmed/22990209)