Motif analysis

Stockholm, November 8 2018

Jakub Orzechowski Westholm

Long-term bioinformatics support

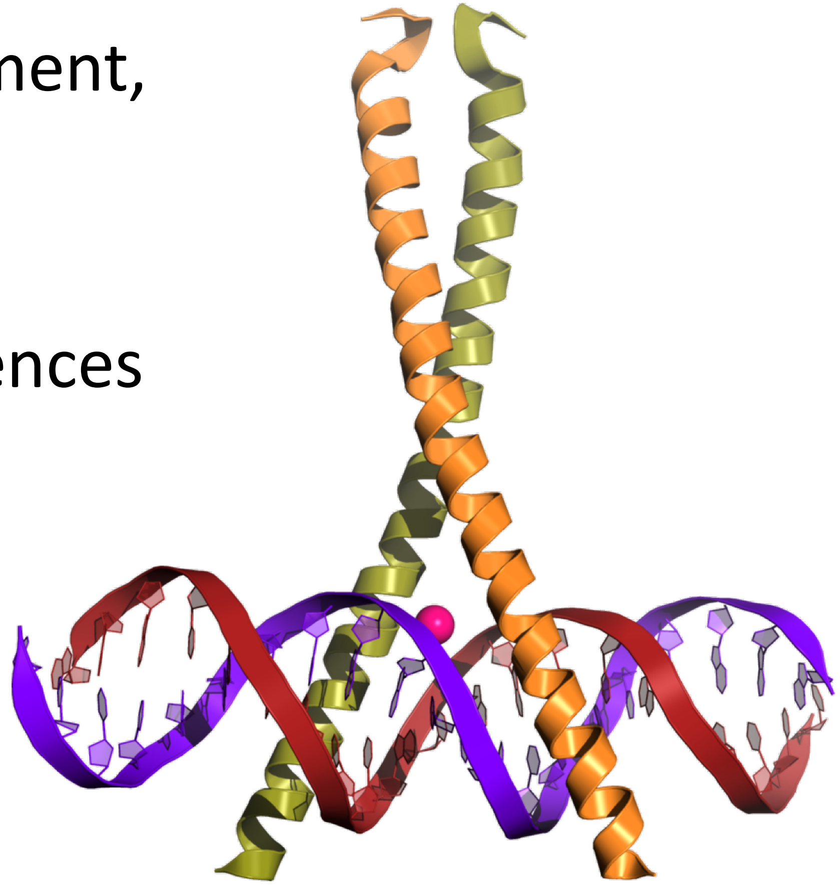
NBIS, SciLifeLab, Stockholm University

The problem

From a transcription factor (TF) ChIP-seq experiment, <u>find the DNA sequences recognized by the TF</u>.

In this context: Motif = a set of nucleotide sequences

Typically 4-20 bp



This lecture

- What is a motif? How is it represented?
- *De-novo* motif discovery: What the problem is, principles behind the programs
- Examples of motif discovery programs
- Practical considerations: data size, how to handle repeats etc.

How can DNA sequence motifs be represented?

1. As a *sequence* of nucleotides, e.g. CTGGAG
2. As a *regular expression*, taking into account ambiguity e.g. [C or G][C or T]GG[G or A]G
3. As a *matrix,* based on nucleotide frequency in each position

Pos	1	2	3	4	5	6
A	0	1	0	0	5	0
C	5	4	0	0	0	1
G	4	0	10	10	4	9
T	1	5	0	0	1	0

4. More complicated representations, taking dependencies between positions into account (HMMs, dinucleotide matrices, deep learning networks etc.)

Position weight matrices

- A position weight matrix (PWM) is based on nucleotide frequencies in a set of aligned sequences.
- The frequencies are converted to probabilities, and then to log-likelihoods given a background model.

Pos	1	2	3	4	5	6
A	0	1	0	0	5	0
C	5	4	0	0	0	1
G	4	0	10	10	4	9
T	1	5	0	0	1	0

Position *frequency* matrix

count nucleotides in each position

Pos	1	2	3	4	5	6
A	0.0	0.1	0.0	0.0	0.5	0.0
C	0.5	0.4	0.0	0.0	0.0	0.1
G	0.4	0.0	1.0	1.0	0.4	0.9
T	0.1	0.5	0.0	0.0	0.1	0.0

Position *probability* matrix

divide by total nr of sequences

Pos	1	2	3	4	5	6
A	-Inf	-1.32	-Inf	-Inf	1.0	-Inf
C	1.0	0.68	-Inf	-Inf	-Inf	-1.32
G	0.68	-Inf	2.0	2.0	0.68	1.85
T	-1.32	1.0	-Inf	-Inf	-1.32	-Inf

Position *weight* matrix

divide by background freq, and log-transform $-\log(m_{n,p}/b_n)$

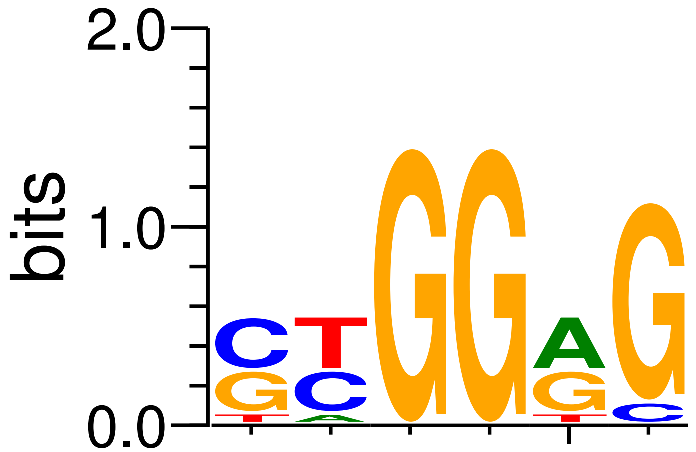
- We might need to add a pseudo count to the frequency matrix, to avoid –Inf.

(Stormo et al. Nucleic Acids Research 1982)

Sequence logos

- Sequence logos are used to visualize PWMs.
- Nucleotide frequency and information content for each position can be represented.

Pos	1	2	3	4	5	6
A	0	1	0	0	0	0
C	4	4	0	0	5	1
G	5	5	10	10	4	9
T	1	0	0	0	1	0



Height: 2 – entropy = $2-\sum_{i=1}^{n} \mathrm{P}(x_i)\log_b \mathrm{P}(x_i),$

Databases with TF binding site motifs

- JASPAR (http://jaspar.genereg.net). Good, curated, free, data base with around 1500 motifs from all kinds of species.
- Transfac (http://genexplain.com/transfac/, http://gene-regulation.com/pub/databases.html). Good, curated, not free, data base with around 2800 motifs from all kinds of species.
    - Older version is free for academic use.
- Other databases
    - ChIPBase http://rna.sysu.edu.cn/chipbase/
    - HOCOMOCO (human only) http://hocomoco11.autosome.ru
    - footprintDB (combining several databases) http://floresta.eead.csic.es/footprintdb/index.php

Scanning the genome with a PWM

- Every sequence can be scored on how well it matches the PWM, by adding up the scores for each position:

Pos	1	2	3	4	5	6
A	-Inf	-1.32	-Inf	-Inf	1.0	-Inf
C	1.0	0.68	-Inf	-Inf	-Inf	-1.32
G	0.68	-Inf	2.0	2.0	0.68	1.85
T	-1.32	1.0	-Inf	-Inf	-1.32	-Inf

GAGGGC → 0.68 -1.32 + 2.0 +2.0 + 0.68 -1.32 = 2.72
CTGGGG → 1.0 + 1.0 + 2.0 + 2.0 + 1.0 + 1.85 = 8.85
CTGAGG → 1.0 + 1.0 - Inf + 2.0 + 1.0 + 1.85 = - Inf

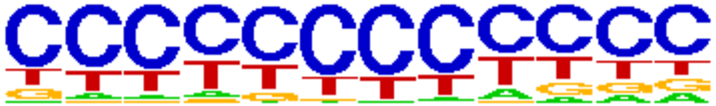
- The score represents the log likelihood of the sequence being a motif compared to bg
- High scores → likely strong TF binding → long time spent on DNA by TF
- Useful to have a cutoff on what we consider is a match. Setting cutoff can be tricky!

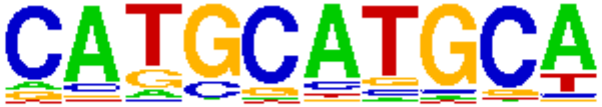
Limitations of position weight matrices

- In 90% of tested cases, matrix based models perform as well as more complex models (Weirauch et al. Nature Biotech. 2013).
- But PWMs can be inaccurate if there is
  - Dependencies between nucleotides
  - Variable spacing between sequences

De-novo motif finding

- Given a set of transcription factor binding sites (e.g. from ChIP-seq), are any motifs enriched?
- Some kind of background model is needed
  - A set of background sequences
    - Regions nearby the peaks (e.g. 2 Kbp away), with similar GC content
  - Nucleotide (or dinucleotide) frequencies
  - A bad background model will give strange and misleading results!





Motif finding methods

- We need methods to search the space of possible motifs
- We also need a way to score motif candidates (e.g. enrichment, complexity)
- Optimal results are not guaranteed.

MEME

MEME
Multiple Em for Motif Elicitation

- Method:
  - Starts with a guess, M, of what the motif might be. It then produces estimates, L, of where motif is located.
  - Given L, the motif M is updated. Then L is updated with a new motif and so on, until the motif M doesn’t change much.
  - When the motif search has converged, the resulting motif is scored (based on enrichment and information content).
  - To finds more motifs, all occurrences of the motif are then removed from the input sequences, and the algortim is the re-run with a new start guess.
- Output
  - A set of PWMs, with scores and p-values
- Pros: Old, widely used method. Often works well.
- Cons: Slow, has trouble handling large inputs (>500 peaks)

DREME

DREME

Discriminative Regular Expression Motif Elicitation

- Method:
  - Look at all 3-8mers to find the most enriched sequences (Fisher test)
  - Iteratively, try to make these more general with search
    - CTGGGG
    - → CTGG[G or A]G
    - → C[C or T]GG[G or A]G
    - → [C or G][C or T]GG[G or A]G
  - Convert this to PWM
- Output: PWMs, with p-values
- Pros: Very fast, good performance
- Cons: Restricted to short sequences (up to 8 bp). Does not take nucleotide frequency into account.

(Bailey, Bioinformatics 2011)

Homer

- Method
  - Looks at all 8,10 and 12-mers to find the most enriched.
  - The most enriched sequences are then converted to weight matrices are refined.
- Output
  - A set of PWMs, with info on e-values and which known motif it’s similar to.
  - If any known motifs are enriched in the given regions.
- Pros
  - Nice output, includes matching to known motifs
  - Quite fast
  - Usually works well
- Cons
  - The documentation is not good
  - It’s a bit hard to install, need to install genomes too.

Practical considerations

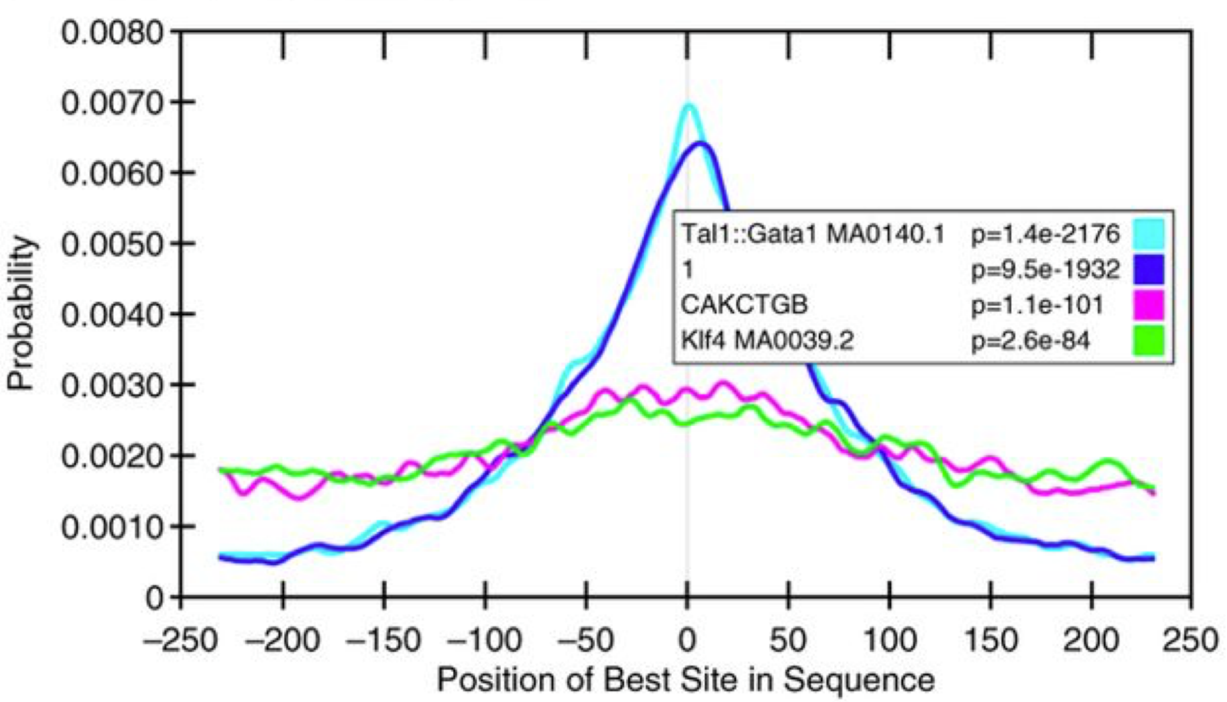
- Less information content → harder problem
  - Short motifs are harder to find
  - Degenerate motifs are harder to find
- Which peaks to use?
  - Some methods will have problems handling tens of thousands of peaks.
  - Also, many weak peaks don’t provide useful information
  - → often only the top 500 etc. peaks are used.
- Repeats (e.g. low complexity repeats) can throw the motif finding methods off. → Work on repeat masked sequences!

How well do these methods work?

- There is no good benchmarking study on motif finding in ChIP-seq data, but usually finding the main motif is not that difficult
  - ChIP-seq gives short regions to look in
  - The top ChIP-seq peaks are typically very enriched for the motif of interest.
- There might also be co-factor motifs. These are harder to find.
- Compare this to analysis on promoters of co-regulated genes:
  - We have very long promoters to search for motifs
  - We have don’t have as clear enrichment of the motifs.

Further analysis

- PhyloGibbs – incorporating sequence conservation in the motif finding.
- Ensemble methods – combining the results from several motif finding programs
- TomTom – Comparison of a new motif to a database of known motifs
- Centrimo – Motif location.



Todays exercise

- Takes sets of peaks from ENCODE
    - ChIP-seq against CTCF (human and mouse data sets)
    - ChIP-seq against REST, from previous lab
- Try a few different motif finders
    - DREME
    - MEME
    - Centrimo
    - HOMER
- Try a motif comparison tool, Tomtom