



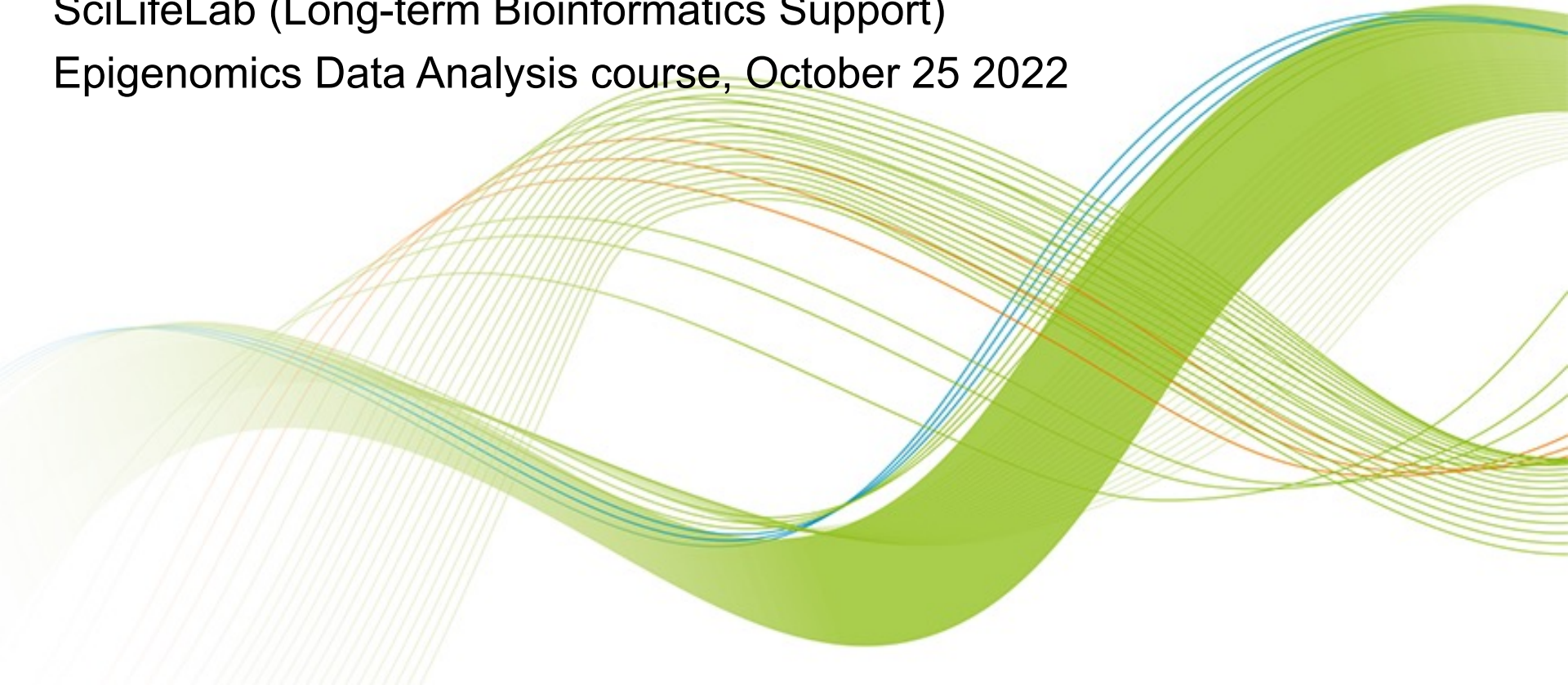
SciLifeLab

Motif analysis

Jakub Orzechowski Westholm

SciLifeLab (Long-term Bioinformatics Support)

Epigenomics Data Analysis course, October 25 2022



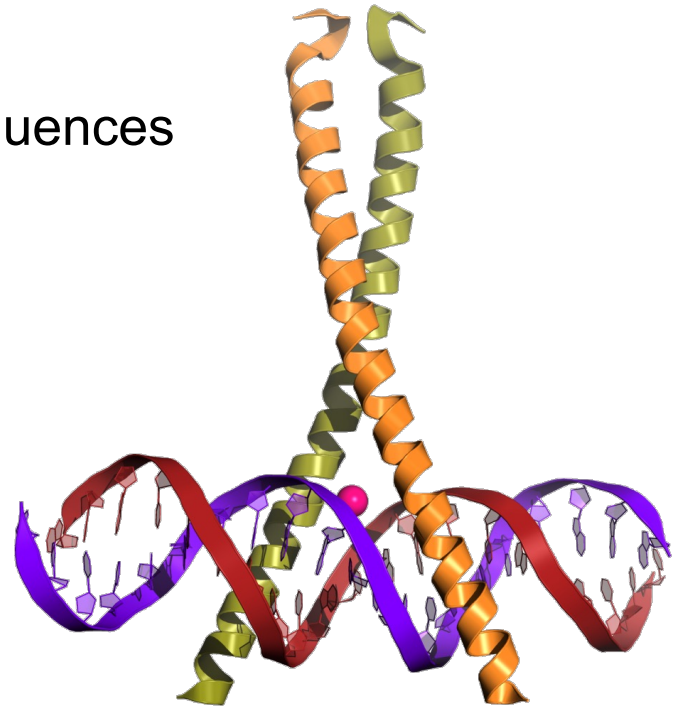
The problem

From a set of genomic regions, find sort DNA sequences (motifs) enriched in the regions. These sequences correspond to (candidate) TF binding sites.

The genomic regions we search through could come from e.g. ATAC-seq (differentially accessible regions etc.), or transcription factor (TF) ChIP-seq experiments,

In this context: Motif = a set of nucleotide sequences

Typically 4-20 bp



-
- 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 DNA sequence motifs be represented' SciLifeLab

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

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

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
A	0	1	0	0	5
C	5	4	0	0	0
G	4	0	10	10	4
T	1	5	0	0	1

Position *frequency* matrix

count nucleotides in each position

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

Position *probability* matrix

divide by total nr of sequences

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

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.

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
A	0	1	0	0	0
C	4	4	0	0	5
G	5	5	10	10	4
T	1	0	0	0	1



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

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

- 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
A	-Inf	-1.32	-Inf	-Inf	1.0
C	1.0	0.68	-Inf	-Inf	-Inf
G	0.68	-Inf	2.0	2.0	0.68
T	-1.32	1.0	-Inf	-Inf	-1.32

GAGGG $\rightarrow 0.68 - 1.32 + 2.0 + 2.0 + 0.68 = 4.04$

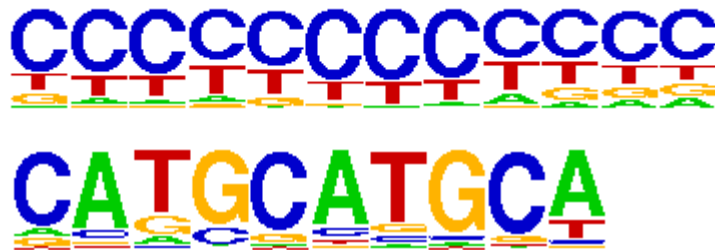
CTGGG $\rightarrow 1.0 + 1.0 + 2.0 + 2.0 + 1.0 = 7$

CTGAG $\rightarrow 1.0 + 1.0 - \text{Inf} + 2.0 + 1.0 = -\text{Inf}$

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

-
- 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

- 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!



-
- 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.

- 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 find more motifs, all occurrences of the motif are then removed from the input sequences, and the algorithm is 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)

- Method:
 - Look at all 3-8mers to find the most enriched sequences (Fisher test)
 - Iteratively, try to make these more general with search
 - CTGGG
 - → CTGG[G or A]
 - → C[C or T]GG[G or A]
 - → [C or G][C or T]GG[G or A]
 - 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.

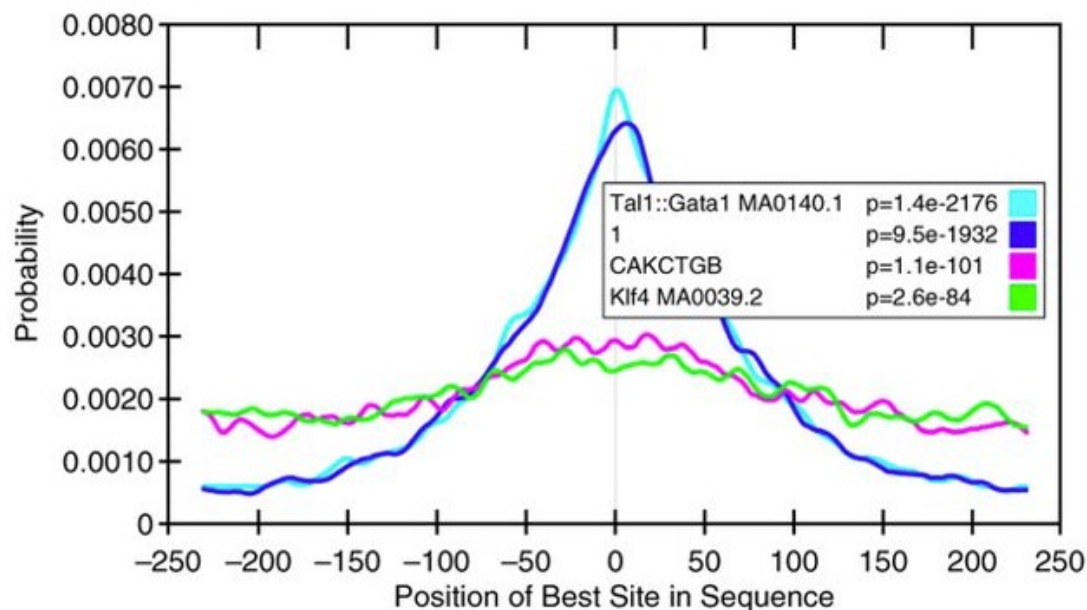
- Method
 - Looks at all 8,10 and 12-mers to find the most enriched.
 - The most enriched sequences are then converted to weight matrices and are refined.
- Output
 - A set of PWMs, with info on e-values and which known motif it's similar to.
 - If any known motifs are enriched in the given regions.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



- 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
 - → for ChIP-seq, often only the top 500 etc. peaks are used.
 - → for ATAC-seq this is not as clear
- Repeats (e.g. low complexity repeats) can throw the motif finding methods off. → Work on repeat masked sequences!

- 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 of 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.

- 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.



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