Ecole Doctorale Sciences de la Vie et de la Santé

Clustering Transcription Factor Binding Motifs

CASTRO-MONDRAGÓN Jaime Abraham

PhD student 2nd Year

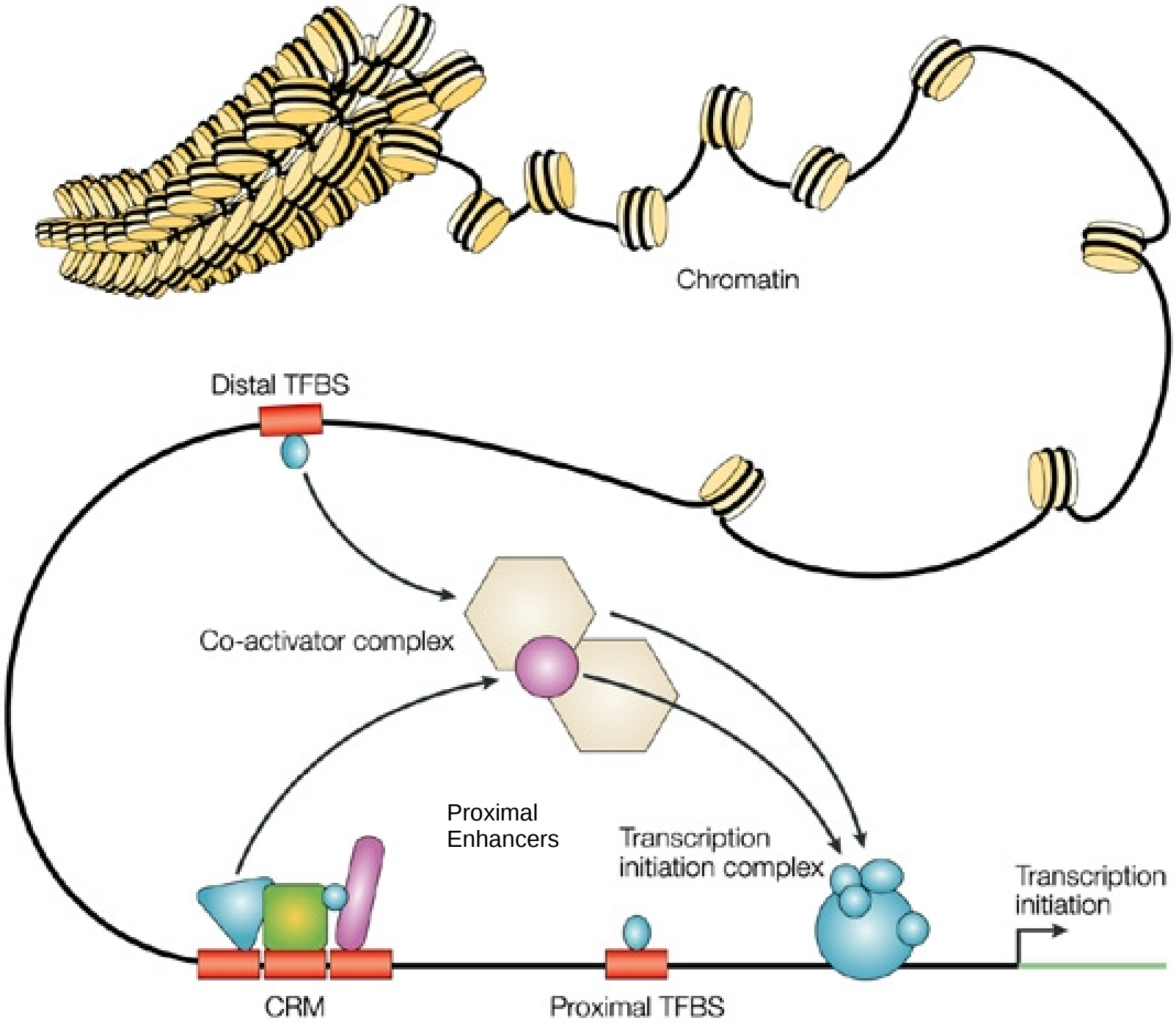
Lab. Technological Advances for Genomics and Clinics (TAGC)

Supervisor: Jacques VAN HELDEN

Aix-Marseille Université

Marseille, France

Introduction – Transcriptional regulation



Wasserman W, Sandelin A.(2004) **Applied bioinformatics for the identification of regulatory elements**. Nat. Rev. Genetics

Introduction – TFBS experimental detection

Experimental Methods

Table I: Comparison of *in vitro*- and *in vivo*-based methods to characterize TF binding specificities

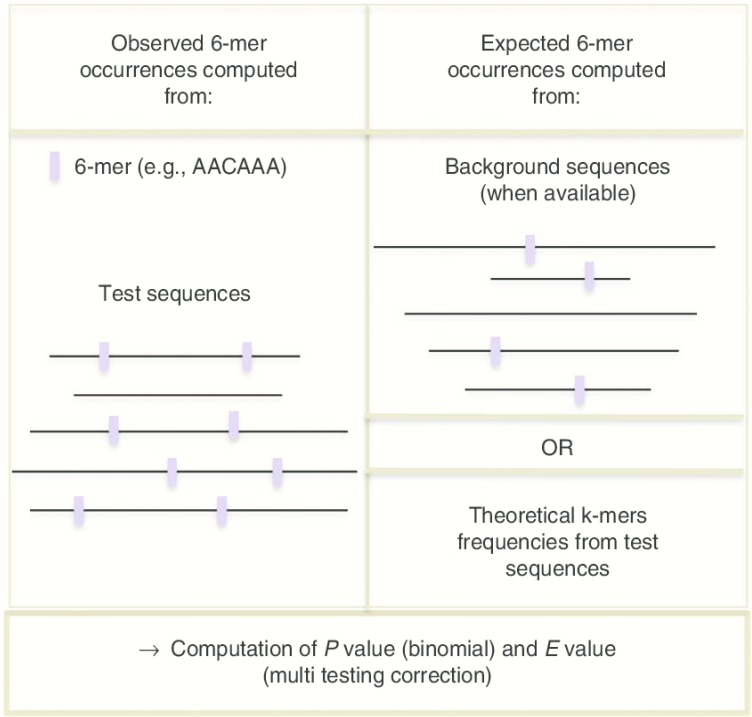
Method	Synonyms	Throughput (DNA sequence space)	Materials needed[c]	Data type		Resolution
In vitro approaches						
Selection of target	SELEX, CASTing	>200 000 sites	mg of P	+	Consensus site	few high affinity binding sites
Selection of target coupled to NGS	HT-SELEX, Bind-n-Seq	>200 000 sites	mg of P	++	PWM[a], relative K_D	Nucleotide resolution feasible
Protein binding microarray	PBM, CSI	up to I milion sites	mg of P	++	PWM[a], relative K_D	Nucleotide resolution feasible
DNA immunoprecipitation	DIP-chip	all genomic sites	µg of P	+	PWM[a]	between 100 and 500 bp
Mechanical trapping	MITOMI	1000 to 100 sites	ng of P	+++(+)	Absolute K_D (k_{on}, k_{off})	Nucleotide resolution
Gel shift	EMSA	around 10 sites	mg of P	++++	Absolute K_D, k_{on}, k_{off}	few binding sites only
Surface plasma resonance	BIAcore	up to 100 site	µg of P	++++	Absolute K_D, k_{on}, k_{off}	few binding sites only
In vivo approaches						
ChIP coupled to microarray	ChIP–chip	all genomic sites	ng of D	+	PWM[a,b]	between 100 and 500 bp
ChIP coupled to NGS	ChIP-seq	all genomic sites	ng of D	+	PWM[a,b]	between 100 and 500 bp
TF mediated DNA methylation profiling	DamID	all genomic sites	ng of D	+	PWM[a]	between 100 and 500 bp
Reverse ChIP	PICh	one genomic site	*	–		
DNaseI sensitivity profiling coupled to NGS	DNaseI-seq	all genomic sites	ng of D	+	PWM[a]	Nucleotide resolution feasible

1.- Geertz M, Maerkl J. (2010) **Experimental strategies for studying transcription factor-DNA binding specificities.** Briefings in functional genomics

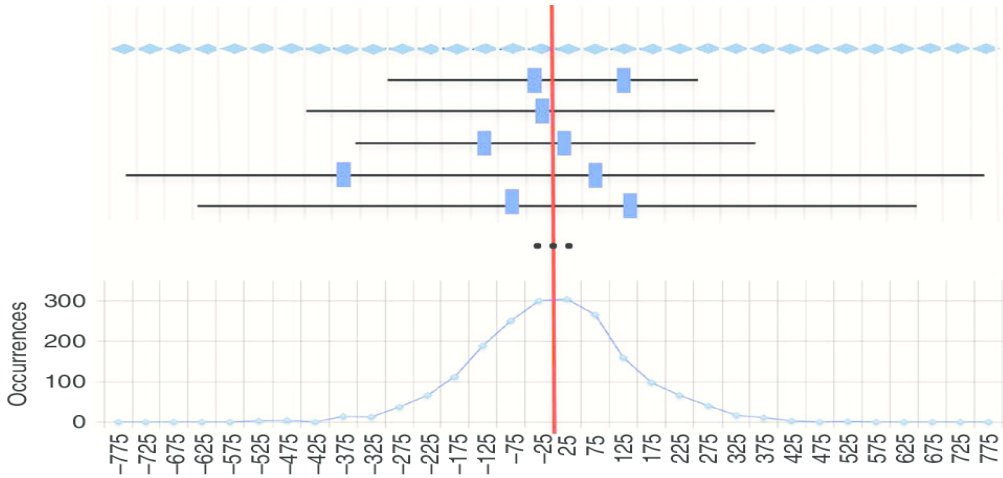
Introduction – TFBS computational detection

Different approaches

Overrepresentation (Observed vs Expected)



Positional bias



Different tools

MEME
Multiple Em for Motif Elicitation

RSAT

1.- Thomas-Chollier M, et al. (2012) A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. Nature Protocols.

Introduction – TFBS representation

Sox2 Binding sites

1	G	C	C	C	T	C	A	T	T	G	T	T	A	T	G	C
2	A	A	A	C	T	C	T	T	T	G	T	T	T	G	G	A
3	T	T	C	A	C	C	A	T	T	G	T	T	C	T	A	G
4	G	A	C	T	C	T	A	T	T	G	T	C	T	C	T	G
5	G	A	T	A	T	C	T	T	T	G	T	T	T	C	T	T
6	T	G	C	A	C	C	T	T	T	G	T	T	A	T	G	C
7	A	A	T	T	C	C	A	T	T	G	T	T	A	T	G	A
8	A	A	A	C	T	C	T	T	T	G	T	T	T	G	G	A
9	A	T	G	G	A	C	A	T	T	G	T	A	A	T	G	C
10	A	G	G	C	C	T	T	T	T	G	T	C	C	T	G	G
11	T	G	T	G	C	T	T	T	T	G	T	N	N	N	N	N
12	C	T	C	A	A	C	T	T	T	G	T	A	A	T	T	T
13	G	C	A	G	C	C	A	T	T	G	T	G	A	T	G	C
14	C	A	C	C	C	C	T	T	T	G	T	T	A	T	G	C
15	T	T	T	T	C	T	A	T	T	G	T	T	T	T	T	A
16	A	A	A	G	G	C	A	T	T	G	T	G	T	T	T	C

Position Specific Scoring Matrix

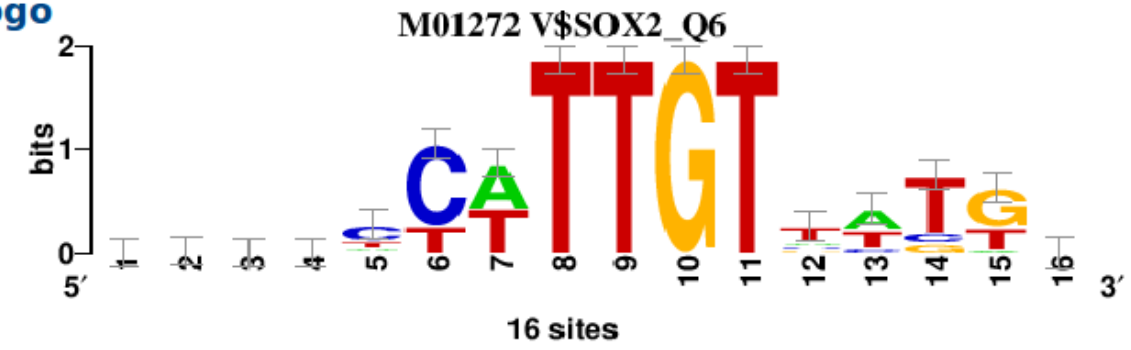
A	6	7	4	4	2	0	8	0	0	0	0	2	7	0	2	4
C	2	2	6	5	9	12	0	0	0	0	0	2	2	2	0	6
G	4	3	2	4	1	0	0	0	0	16	0	2	0	2	9	3
T	4	4	4	3	4	4	8	16	16	0	16	9	6	11	5	2

Consensus

d	w	h	v	y	Y	w	T	T	G	T	t	w	T	k	m

Y = C or T
W = A or T

(Logo



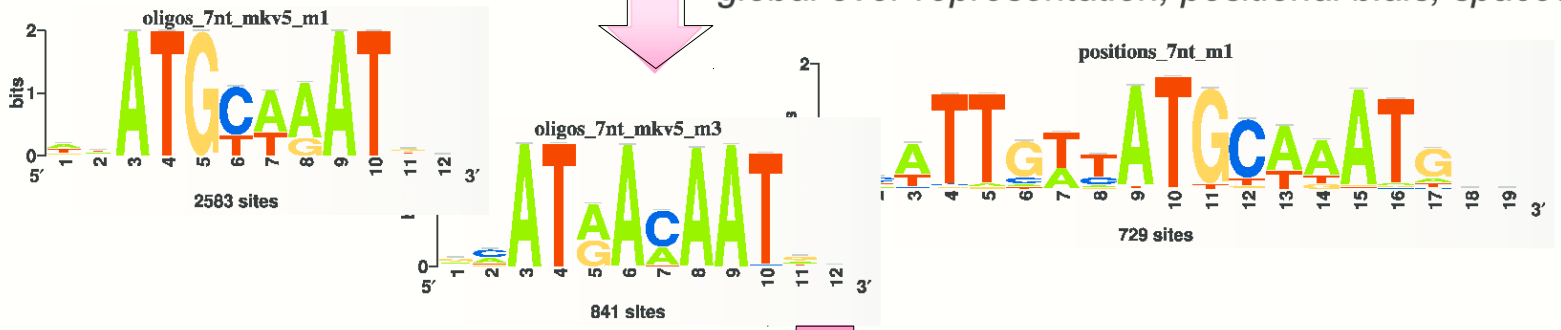
Introduction – In silico motif discovery

```
>mm9_chr1_3473041_3473370_+
ctgtctctctatcttgcttataaaggat
ctctttgtattggaaattggttgtttggg
tatatcctgtgcctaatttgcatatgga
```

Sequence dataset

***de novo* motif discovery**

global over-representation, positional biais, spaced motifs



Comparison with collections of motifs

various metrics to calculate motif similarity

DataBases with redundant motifs

HOmo sapiens
COmprehensive
MOdel
COllection

Combine several approaches + tools to discover motifs.

Pros: complementarity (motifs found by one program but not by the others).

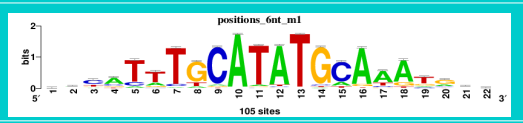
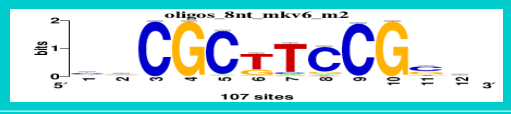
Cons: redundancy (same motifs found by several programs).

1.- Adapted from Thomas-Chollier et al. (2011) RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. Nucleic Acids Research.

Introduction – Where does redundancy comes from?

In Motif discovery algorithms:

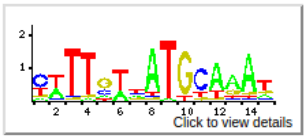
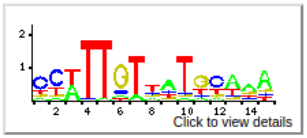
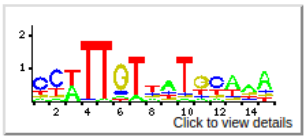
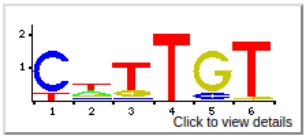
- Complementarity between different algorithms.
- One algorithm can be set with different parameters (e.g. different k-mer size, different background model, etc).

local-words	oligo-analysis	position-analysis
		
		
		
		
		
		
		

Introduction – Where does redundancy comes from?

In Motif Databases:

- TFBMs come from different sources
  (e.g. Footprint Assays, Microarrays, Selex, ChIP-seq, …)
- TFBMs can come from data of several species.
- TFs from the same Family use to have conserved structure of DBD.
- Errors in annotation.

MA0142.1	Pou5f1::Sox2	10090	Helix-Turn-Helix	Homeo	 Click to view details
MA0143.1	Sox2	10090	Other Alpha-Helix	High Mobility Group box (HMG)	 Click to view details
MA0143.2	Sox2	10090	Other Alpha-Helix	High Mobility Group box (HMG-box)	 Click to view details
MA0143.3	Sox2	10090	Other Alpha-Helix	High Mobility Group box (HMG)	 Click to view details
MA0442.1	SOX10	10090,10116,9606	Other Alpha-Helix	High Mobility Group box (HMG)	 Click to view details

SOX2_DBD_2 (HumanTF 1.0)	SOX2	a**TGAATG**ky**ATTCA**k	Homo sapiens	Site type: dimeric; SELEX cycle: 4; Possible binding to single
SOX2_DBD_3 (HumanTF 1.0)	SOX2	r**ATCAATG**ky**ATTGA**tr	Homo sapiens	Site type: dimeric; SELEX cycle: 4; Possible binding to single
SOX2_full_1 (HumanTF 1.0)	SOX2	s**AACAATA**a**CATTGTT**c	Homo sapiens	Site type: dimeric; SELEX cycle: 4; Possible binding to single
SOX2_full_2 (HumanTF 1.0)	SOX2	h**ATCAATA**m**CATTGAT**m	Homo sapiens	Site type: dimeric; SELEX cycle: 4; Possible binding to single
SOX2_full_3 (HumanTF 1.0)	SOX2	a**TGAAT**am**CATTCA**t	Homo sapiens	Site type: dimeric; SELEX cycle: 4; Possible binding to single
SOX2_f1 (HOCOMOCO v9)	SOX2	**T**tgcAtr**ACAA**w**R**g	Homo sapiens Mus musculus	
MA0142 (JASPAR 2014)	Pou5f1::Sox2	yw**TT**swy**ATGCA**a**A**t	Mus musculus	Chip-seq
MA0143 (JASPAR 2014)	Sox2	**CC**w**TTGT**y	Mus musculus	ChiP-seq; ChiP-seq; ChIP-seq

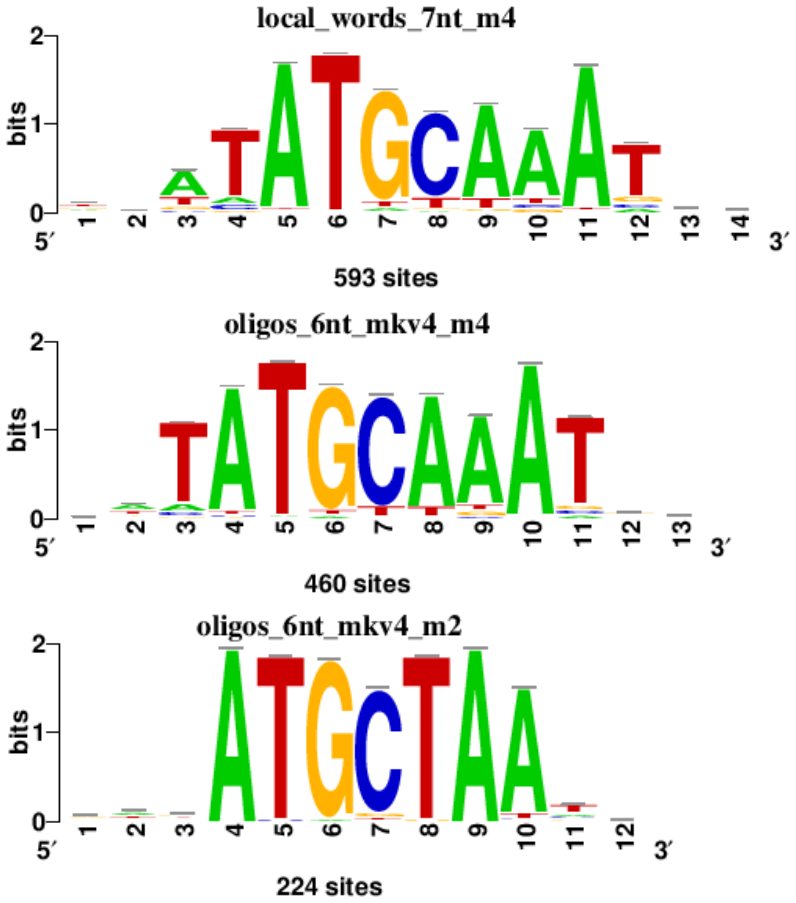
Screenshot from Jaspar and FootprintDB: Query **Sox2**

Introduction – Motif comparison

Similarity Metric	Formula
Pearson correlation coefficient (PCC)	$PCC(X,Y) = \frac{\sum_{b=A}^{T}(f_X(b)-\bar{f}_X)\cdot(f_Y(b)-\bar{f}_Y)}{\sqrt{\sum_{b=A}^{T}(f_X(b)-\bar{f}_X)^2\cdot\sum_{b=A}^{T}(f_Y(b)-\bar{f}_Y)^2}}$
Chi-square (pCS) (1−*p-value of*)	$\chi^2_3(X,Y) = \sum_{K=\{X,Y\}}\sum_{b=A}^{T}\frac{(n_K(b)-n^e_K(b))^2}{n^e_K(b)}$
Average Kullback–Leibler (AKL)	$AKL(X,Y) = 10 - \frac{\sum_{b=A}^{T} f_X(b)\cdot log\frac{f_X(b)}{f_Y(b)} + \sum_{b=A}^{T} f_Y(b)\cdot log\frac{f_Y(b)}{f_X(b)}}{2}$
Sum of squared distances (SSD)	$SSD(X,Y) = 2 - \sum_{b=A}^{T}(f_X(b)-f_Y(b))^2$
Average log-likelihood ratio (ALLR)	$ALLR(X,Y) = \frac{\sum_{b=A}^{T} n_X(b)\cdot log\frac{f_Y(b)}{p_{ref}(b)} + \sum_{b=A}^{T} n_Y(b)\cdot log\frac{f_X(b)}{p_{ref}(b)}}{\sum_{b=A}^{T}(n_X(b)+n_Y(b))}$
ALLR with lower limit (ALLR_LL)	Same as above, but a lower limit of −2 is imposed on the score (see text)

How similar are these motifs?

- Different width
- Different orientation
- Different number of sites
- Differences in nucleotide frequencies.
- Different information content



Mahony S, et al (2007) **DNA Familial Binding Profiles Made Easy: Comparison of Various Motif Alignment and Clustering Strategies**. PloS Comp Biol

Introduction – Motif comparison : a battlefield

- 25 papers published (since 2004) with new methods to measure motif similarity.
- Some metrics works well with monomers, dimer, spaced motifs, high information content columns, etc.
- Each publication says that its method outperforms the others!

- **Conclusion**: No a standard metric to measure motif similarity. A combination of them could help to measure correctly the similarity.

FISim: A new similarity measure between transcription factor binding sites based on the fuzzy integral
Fernando Garcia*, Francisco J Lopez, Carlos Cano and Armando Blanco

A Novel Alignment-Free Method for Comparing Transcription Factor Binding Site Motifs
Minli Xu¤, Zhengchang Su*

Jaccard index based similarity measure to compare transcription factor binding site models
Ilya E Vorontsov[2,3†], Ivan V Kulakovskiy[1,2*†] and Vsevolod J Makeev[1,2,4]

Quantifying similarity between motifs
Shobhit Gupta*, John A Stamatoyannopoulos*, Timothy L Bailey† and William Stafford Noble*‡

A Discriminative Approach for Unsupervised Clustering of DNA Sequence Motifs
Philip Stegmaier[1,2¤], Alexander Kel[2], Edgar Wingender[2,3], Jürgen Borlak[4]

Natural similarity measures between position frequency matrices with an application to clustering
Utz J. Pape[1,2,*], Sven Rahmann[3,4] and Martin Vingron[1]

SPIC: A novel similarity metric for comparing transcription factor binding site motifs based on information contents
Shaoqiang Zhang[1*], Xiguo Zhou[1], Chuanbin Du[2], Zhengchang Su[2*]

Alignment-free clustering of transcription factor binding motifs using a genetic-k-medoids approach
Pilib Ó Broin[1,2], Terry J Smith[1] and Aaron AJ Golden[1,3*]

Motivation – From comparison to clustering

- The (redundant) discovered motifs are compared against several motif databases (which also contain redundant motifs).

- The motif comparison results can be used to cluster the similar motifs.

- The clustering of motifs would allow to:
  a) Filter out the redundant motifs.
  b) Classify the motifs (TF families)

- A few programs are available to cluster the TFBMS (stamp[1] , m2match[2] , matlign[3], GMACS[4])

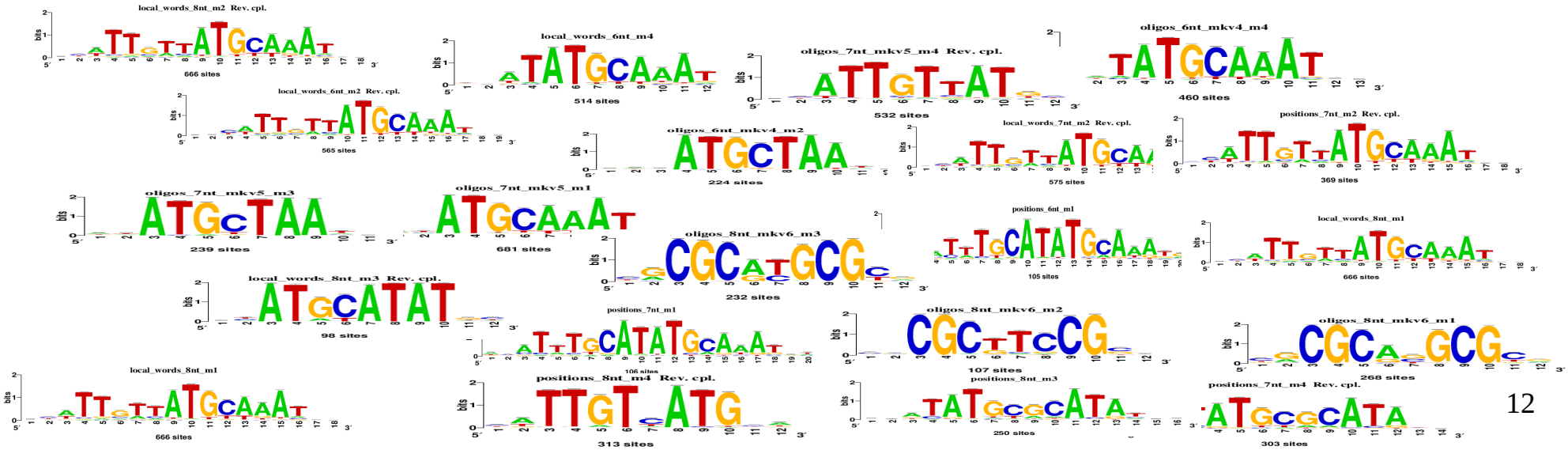
  Limitations:
  - Low number of input motifs.
  - One (or few) metrics can be used.
  - No visualization of the clustered TFBMs.
  - Only one dataset of TFBMs can be grouped.

1.- Mahony S and Benos P. (2007) **STAMP: a web tool for exploring DNA-binding motif similarities.** NAR.
2.- Kankainen M and Löytynoja A. (2007) **MATLIGN: a motif clustering, comparison and matching tool.** BMC Bioinformatics.
3.- Stegmaier P et al. (2013) **A Discriminative Approach for Unsupervised Clustering of DNA Sequence Motifs.** PloS Computational Biology.
4.- Broin P, Smith TJ and Golden AJ. (2015) **Alignment-free clustering of transcription factor binding motifs using a genetic-k-medoids approach.** BMC Bioinformatics.

Motivation – Clustering of Motifs

Questions:

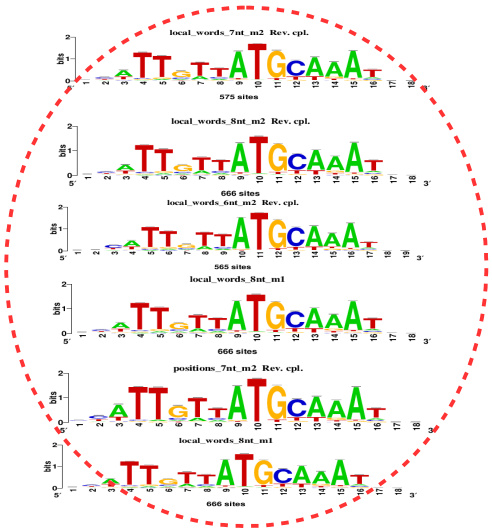
- Which metric use to compare/cluster the motifs ?
- Can we group the similar motifs ?
- Does the clustering reflect binding properties of TFs ?
- Does the grouped TFBMs correspond to the same Family ?
- Is it possible to create a non-redundant collection of motifs ?
- Does the public collections of motifs encompass the private ones ?

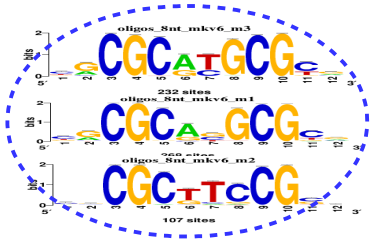


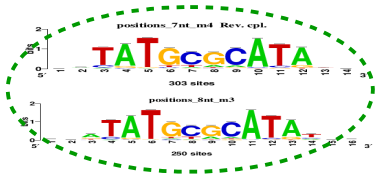
Motivation – Clustering of Motifs

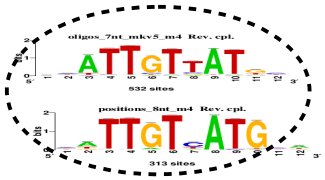
Questions:

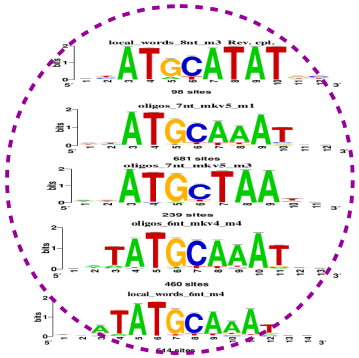
- Which metric use to compare/cluster the motifs ?
- Can we group the similar motifs ?
- Does the clustering reflect binding properties of TFs ?
- Does the grouped TFBMs correspond to the same Family ?
- Is it possible to create a non-redundant collection of motifs ?
- Does the public collections of motifs encompass the private ones ?











Objectives

- Create a tool to identify and visualize groups of similarities among a set of input TFBMs.

Specific goals:

(1) Partitioning the input set of TFBMs into distinct clusters.

(2) Different TFBMs format should be used as input.

(3) Cluster more than one collection of motifs.

(4) Align the TFBMs to highlight common/different positions.

(5) Provide intuitive and dynamic visualization of motifs and their relationships.

Applications:

(1) Simplify the interpretation of motif discovery results.

(2) Identify groups of similar motifs (TF families).

(3) Create non-redundant collections of TFBMs.

Choosing a metric

The metrics to measure motif similarity can be categorized in two groups:

- **Correlations**

  Range: (-1 … 0 … +1)

  - Pearson Correlation (**cor**)
  - Information Content Correlation (**Icor**)
  - ...

- **Distances**

  Range: (0 … no obvious bound value for the maximum distance)

  - Euclidean Distance (**dEucl**)
  - Sum of Squared distances (**SSD**)
  - Sandelin-Wasserman Similarity (**SW**)
  - Kullback-Leibler distance (**dKL**)
  - ...

Choosing a metric - Drawbacks

- Spurious alignments (alignment of non-informative columns)

cor >= 0.8

Cluster 1

Cluster 2

Choosing a metric - Drawbacks

- Spurious alignments (alignment of non-informative columns)

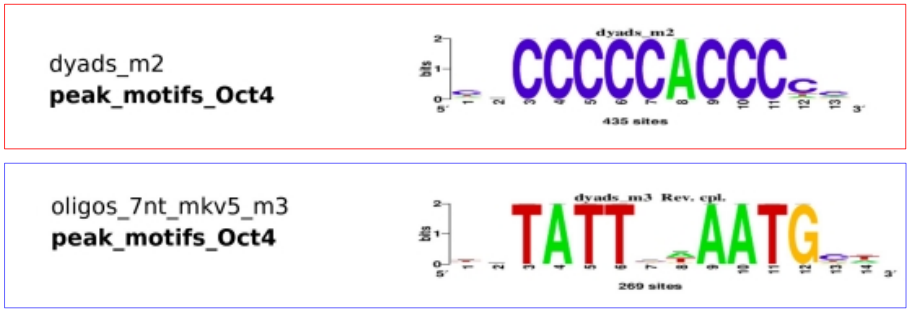
cor >= 0.8

Cluster 1

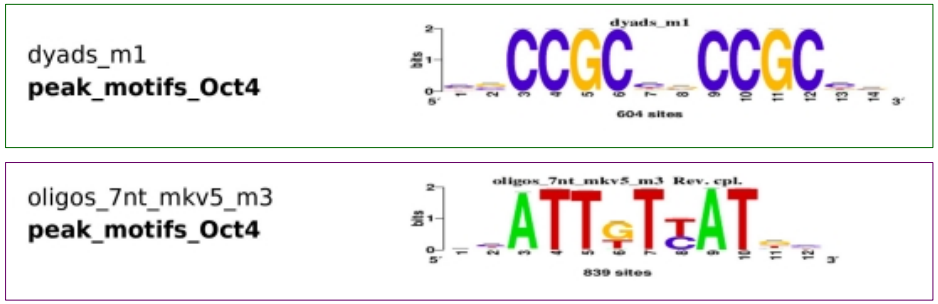
Cluster 2

- Including the width of the aligned segment avoid the spurious alignments.
  But the motifs are not aligned.

cor >= 0.8 + w >= 5



Cluster 1 and 2

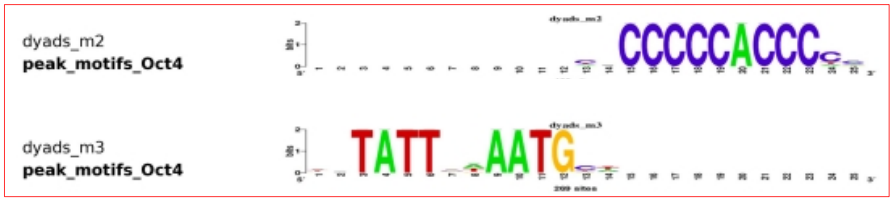


Cluster 3 and 4

Choosing a metric - Drawbacks

- Spurious alignments (alignment of non-informative columns)

cor >= 0.8

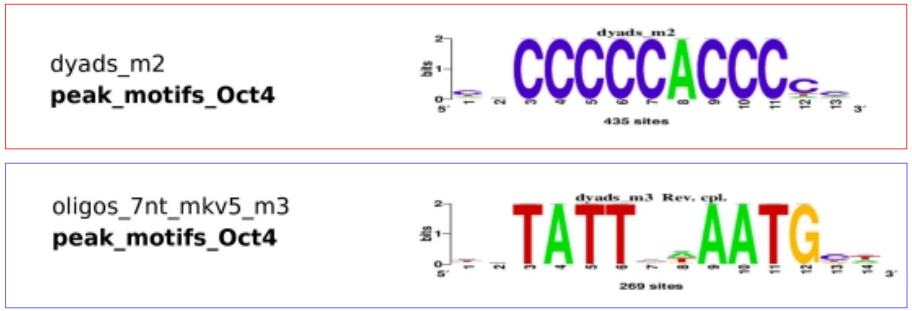


Cluster 1

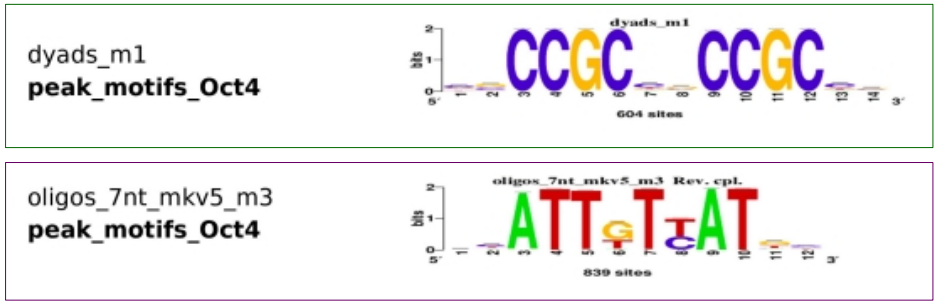
Cluster 2

- Including the width of the aligned segment avoid the spurious alignments.
  But the motifs are not aligned.

cor >= 0.8 + w >= 5



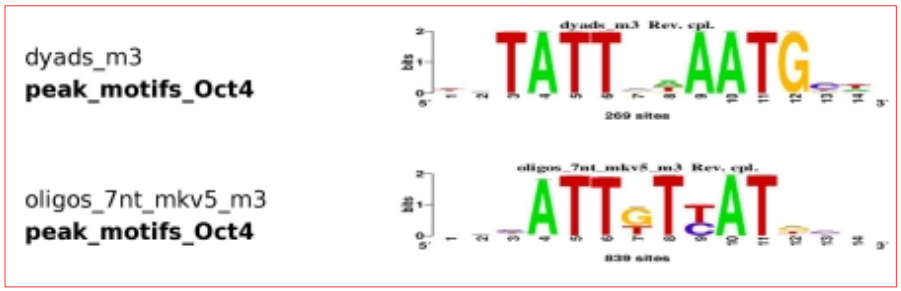
Cluster 1 and 2



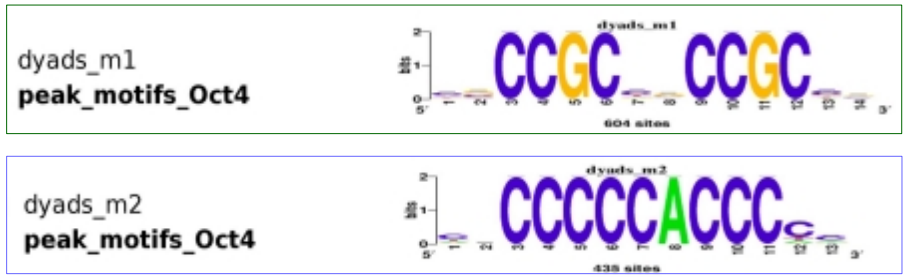
Cluster 3 and 4

- A correction in these metrics (e.g. normalization with the aligned segment) improves the results.

Ncor >= 0.6



Cluster 1



Cluster 2 and 3

matrix-clustering

- *matrix-clustering* : dynamic visualization of clusters of TFBMs in RSAT[1].

  1) Segmentation of the input set of TFBMs into separated trees (forest) rather than a single tree.

  2) Displays the cluster of motifs as logo alignments.

  3) Generate Branch-wise motifs (also known as Familial Binding Profiles)

  4) Supports for a large series of metrics to measure TFBM (dis)similarities.

  5) Possibility to set a custom combination of these metrics to compute an integrative threshold to separate the TFBMs.

  6) Multiple and dynamic representation of clusters (logo alignment, heatmaps).

  7) Possibility to cluster two ore more collection of TFBMs.

- Available in RSAT website (http://www.rsat.eu/)

1.- Medina-Rivera A et al. (2015) **RSAT 2015: Regulatory Sequence Analysis Tools**. NAR

Matrix-clustering - algorithm

Collection of Motifs

Motifs from several motif discovery tools

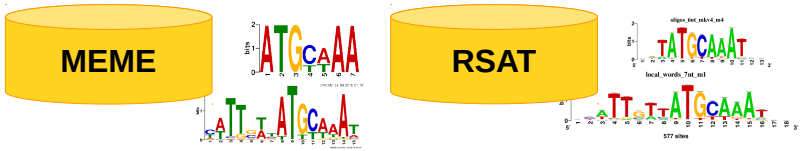
Motifs from different experiments/conditions

Motif Databases
(Jaspar, Transfac, CisBP, Hocomoco, FootprintDB)

Matrix-clustering - algorithm

Collection of Motifs

Motifs from several motif discovery tools

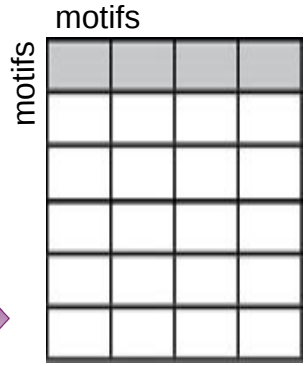


Motifs from different experiments/conditions

Motif Databases
(Jaspar, Transfac, CisBP, Hocomoco, FootprintDB)

Comparison of all motifs

Various metrics to calculate motif similarity

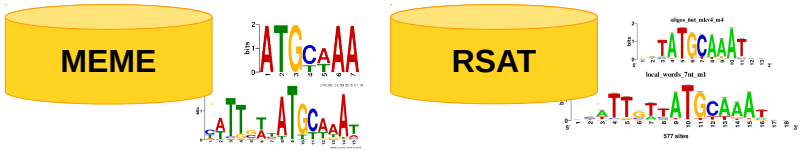


Distance matrix

Matrix-clustering - algorithm

Collection of Motifs

Motifs from several motif discovery tools

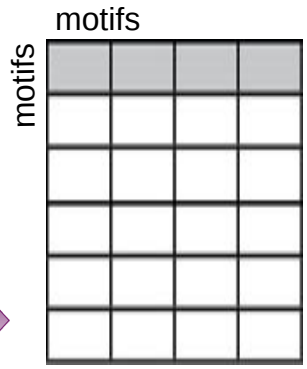


Motifs from different experiments/conditions

Motif Databases
(Jaspar, Transfac, CisBP, Hocomoco, FootprintDB)

Comparison of all motifs

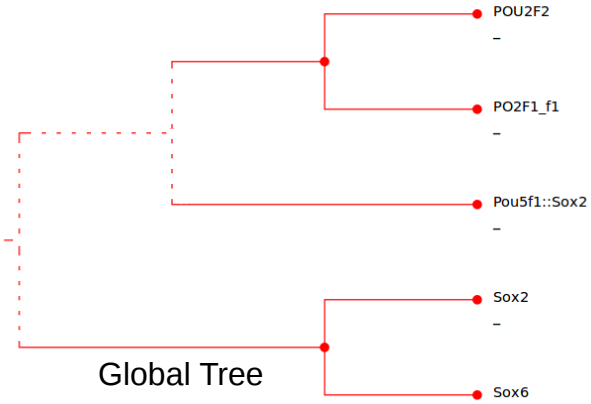
Various metrics to calculate motif similarity



Distance matrix

Hierarchical Clustering

One similarity metric + linkage rule



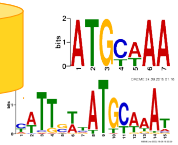
Global Tree

Matrix-clustering - algorithm

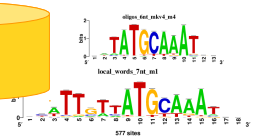
Collection of Motifs

Motifs from several motif discovery tools

MEME



RSAT



Motifs from different experiments/conditions

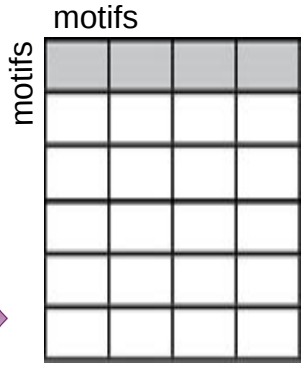
Oct4 ChIP-seq

Sox2 ChIP-seq

Motif Databases
(Jaspar, Transfac, CisBP, Hocomoco, FootprintDB)

Comparison of all motifs

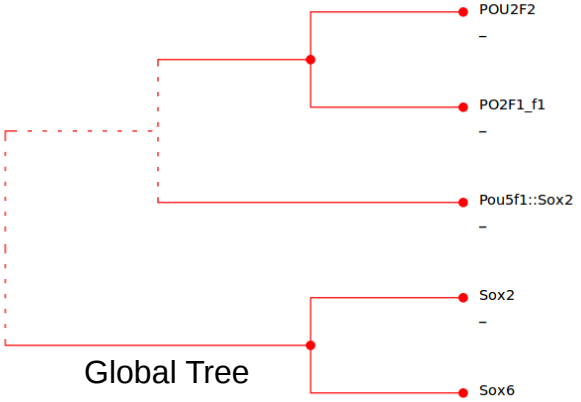
Various metrics to calculate motif similarity



Distance matrix

Hierarchical Clustering

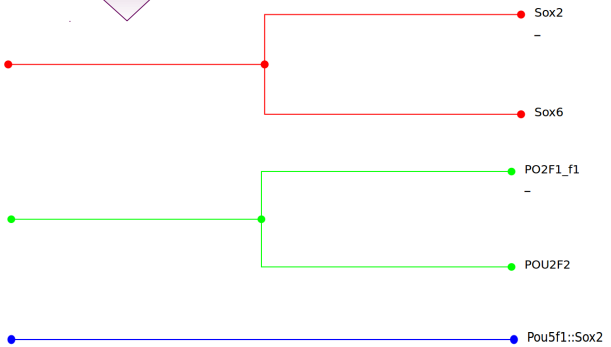
One similarity metric + linkage rule



Global Tree

Partitioning

Integrative threshold
with multiple metrics



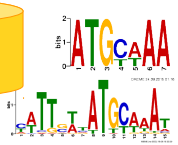
Multiple clusters

Matrix-clustering - algorithm

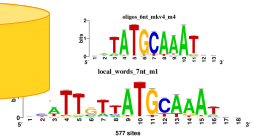
Collection of Motifs

Motifs from several motif discovery tools

MEME



RSAT



Motifs from different experiments/conditions

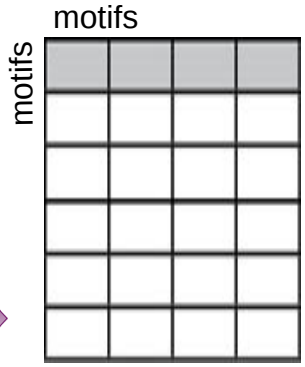
Oct4 ChIP-seq

Sox2 ChIP-seq

Motif Databases
(Jaspar, Transfac, CisBP, Hocomoco, FootprintDB)

Comparison of all motifs

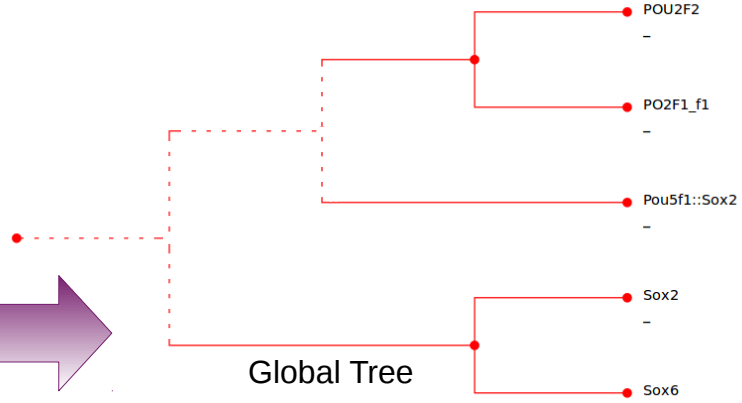
Various metrics to calculate motif similarity



Distance matrix

Hierarchical Clustering

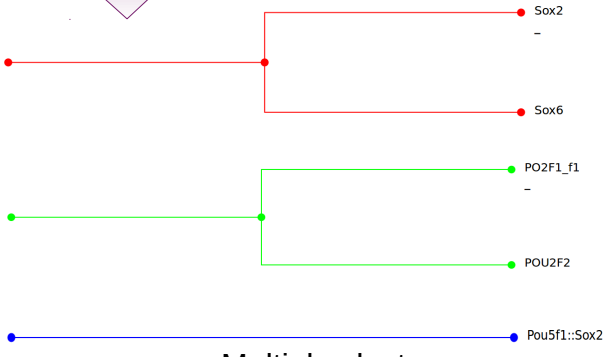
One similarity metric + linkage rule



Global Tree

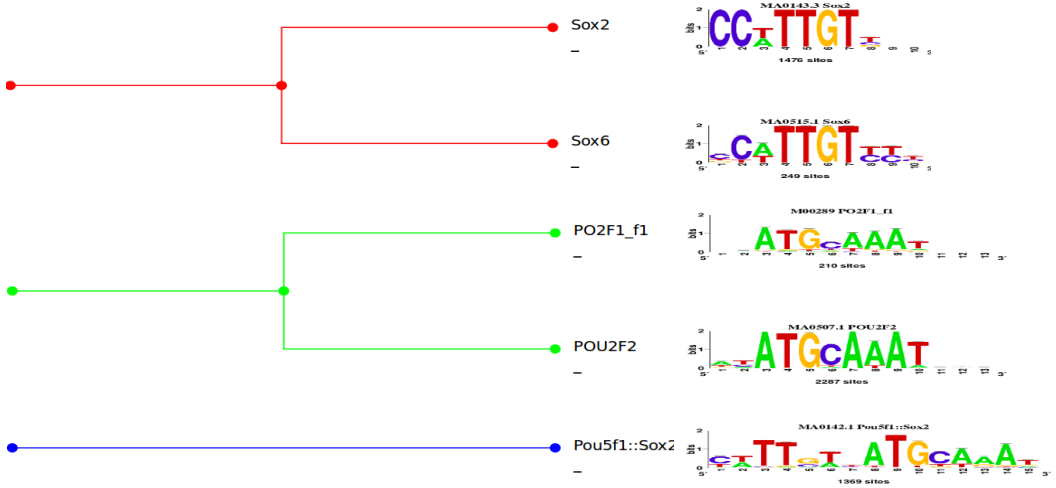
Partitioning

Integrative threshold with multiple metrics



Multiple clusters

Alignment



Matrix-clustering - algorithm

Collection of Motifs

Motifs from several motif discovery tools

MEME

RSAT

Motifs from different experiments/conditions

Oct4 ChIP-seq

Sox2 ChIP-seq

Motif Databases
(Jaspar, Transfac, CisBP, Hocomoco, FootprintDB)

Comparison of all motifs

Various metrics to calculate motif similarity

motifs

motifs

Distance matrix

Hierarchical Clustering

One similarity metric + linkage rule

POU2F2
PO2F1_f1
Pou5f1::Sox2
Sox2
Sox6

Global Tree

Partitioning

Integrative threshold with multiple metrics

Sox2
Sox6
PO2F1_f1
POU2F2
Pou5f1::Sox2

Multiple clusters

Alignment

Sox2
Sox6
PO2F1_f1
POU2F2
Pou5f1::Sox2

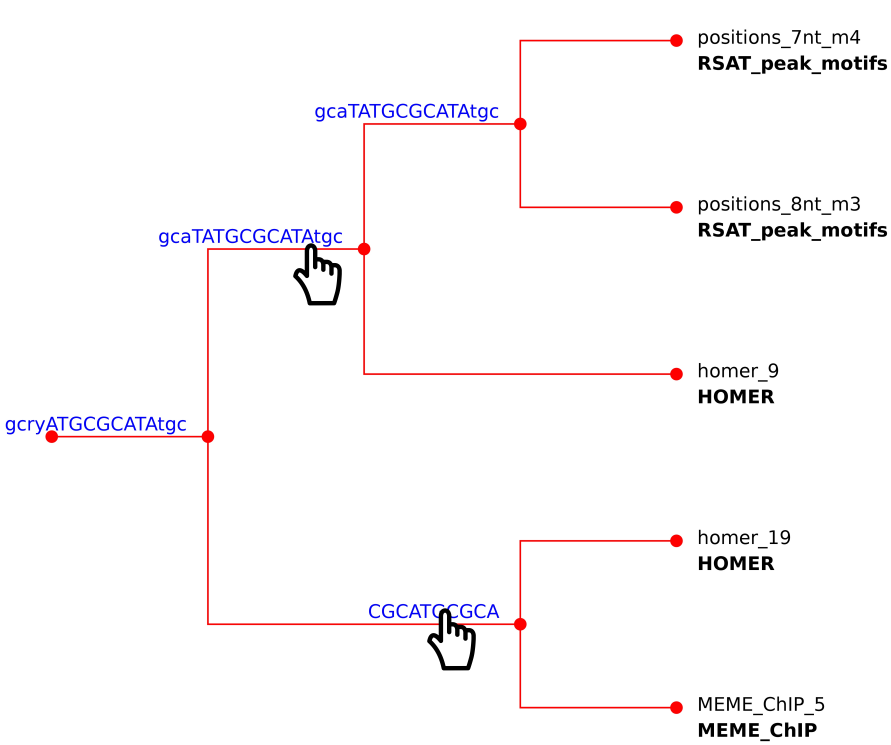
Branch-wise Motifs and Collapsing

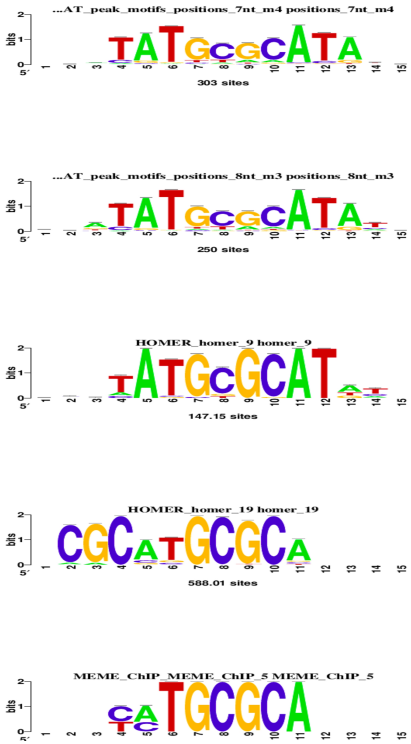
cluster1_node_1_2motifs
cluster2_node_1_2motifs
Pou5f1::Sox2

Non-redundant Motifs

Matrix-clustering : output example

Expanded Tree





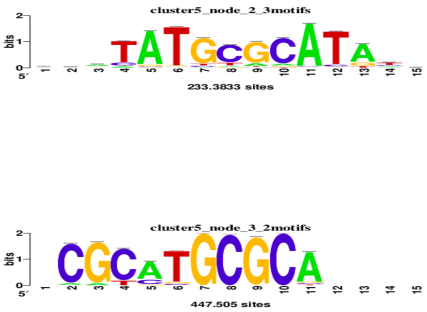
The Familial Binding Profiles are calculated at each branch.

The '**collection**' tag indicates the provenance of the motif.

Click on each branch to collapse it

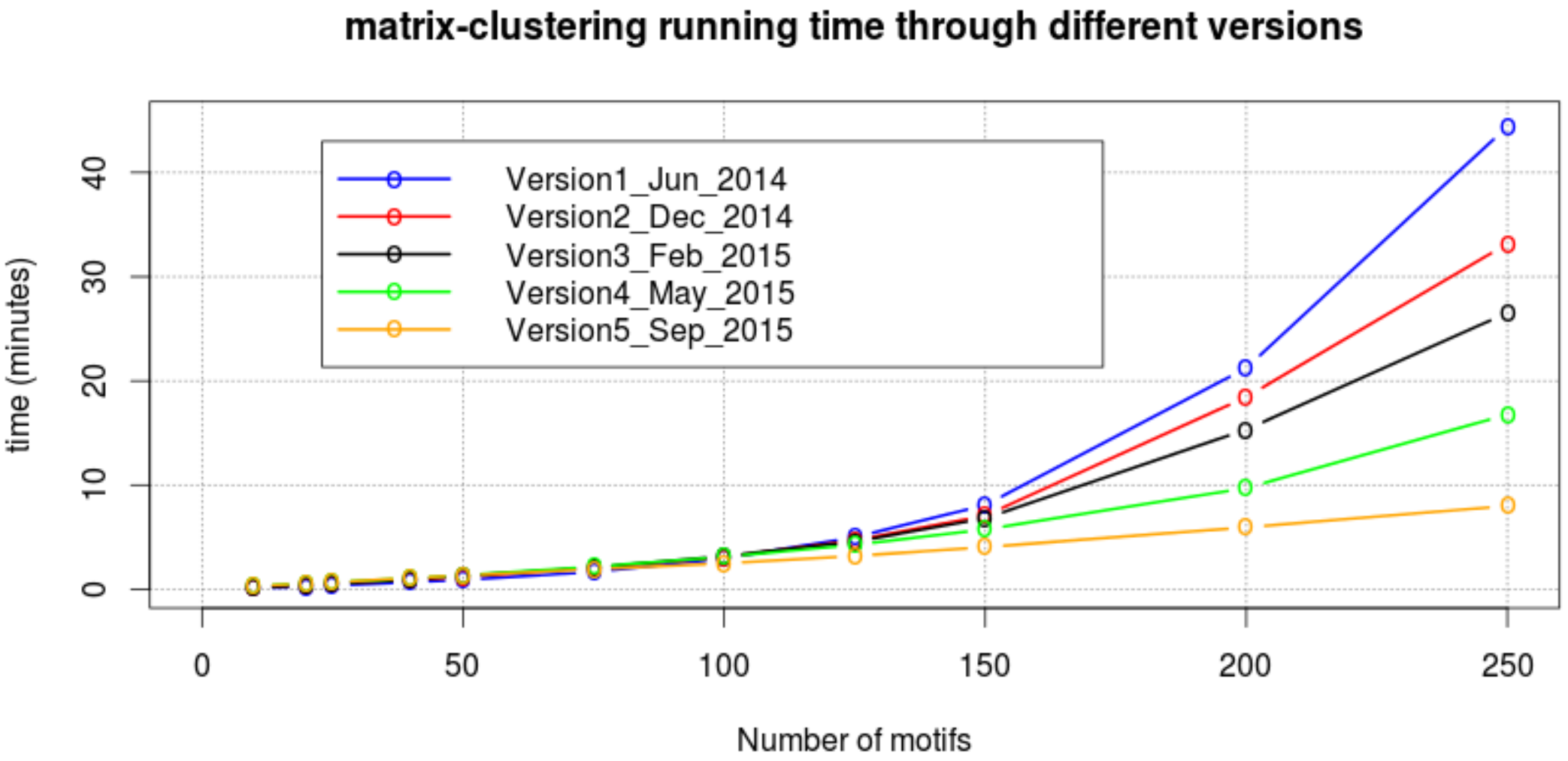
The hierarchical tree is collapsed/expanded dynamically.

Collapsed Tree



Non-Redundant Motifs: they can be selected manually by the user.

matrix-clustering



Example 1 : clustering several motif sets

- **Main Objective**:

  Discover and Cluster motifs in Oct4 ChIP-seq peaks with different algorithms (RSAT + MEME + HOMER).

- **Specific Objectives**:

  - Cluster several collections (files) of motifs coming from different sources (Databases, Algorithms, etc).
  - Reduce the redundancy in multiple collections of motifs.
  - Obtain a representative (non-redundant) set of motifs.
  - Identify the Oct variants (Monomer, homodimer, heterodimer, etc).
  - Use STAMP[1] as alternative tool to cluster the motifs.

1.- Mahony S and Benos P. (2007) **STAMP: a web tool for exploring DNA-binding motif similarities.** NAR.

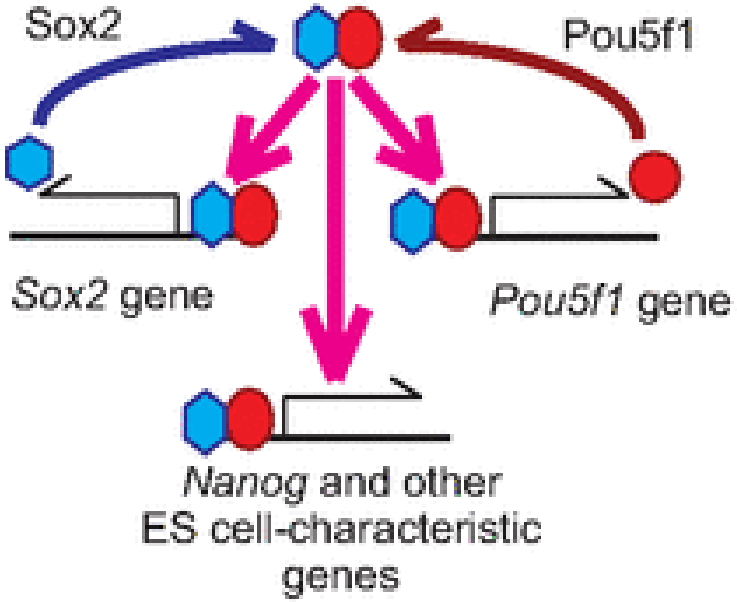
Oct4 and Sox2

- Forced expression of four TFs expressed in ES Cells (c-Myc, Klf4, Oct4, Sox2) can reprogram mouse embryonic fibroblast to pluripotent cells.

- OCT-TFs bind to target sequences either by homodimerisation or by heterodimerisation with other TFs.

- Oct4 (Pou5f1) can heterodimerise with Sox2. (SOx + OCt = SOCT).

```
Sox2  = CATTGTT
SOCT  = CATTGTTATGCAAAT
OCT4  =        ATGCAAAT
```



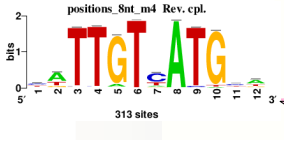
- SOCT is the top motif identified with *de novo* motif discovery algorithms independently in Sox2 and Oct4 ChIP-seq peaks from mouse ES cells.

1.- Kamachi Y and Kondoh H. (2013) **Sox proteins: regulators of cell fate specification and differentiation**. Development.

Motifs found in Oct4 ChIP-seq peaks

```
>mm9_chr1_3473041_3473370_+
ctgtctctctatcttgcttaataaaggat
ctctttgtattggaaattggttgtttggg
tatatcctgtgcctaatttgcatatgga
```

Oct4 ChIP-seq peaks[2]

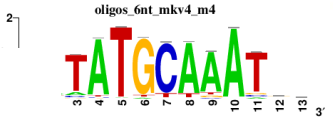


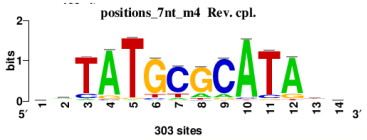
Motif Discovery

MEME — Multiple Em for Motif Elicitation

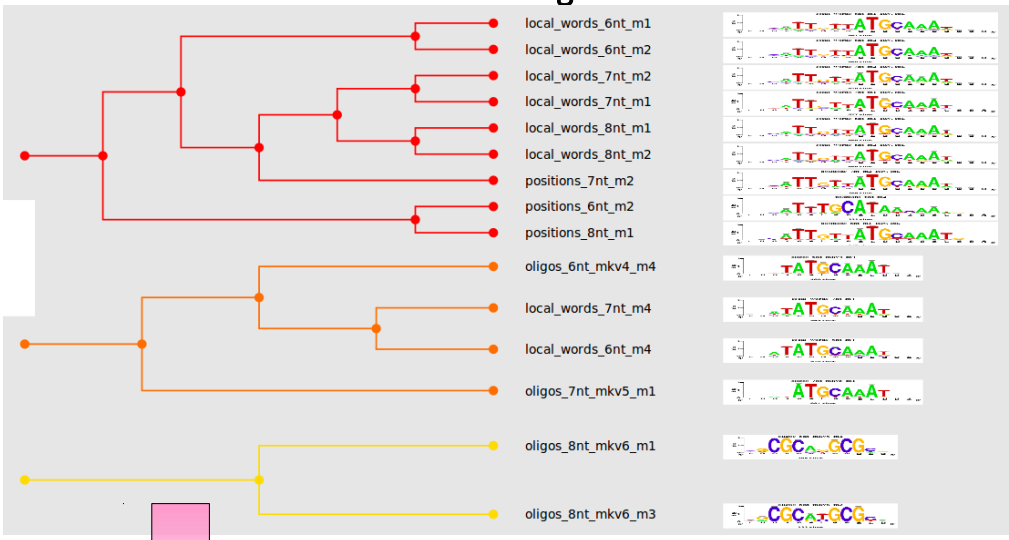
RSAT

Various motif discovery algorithms

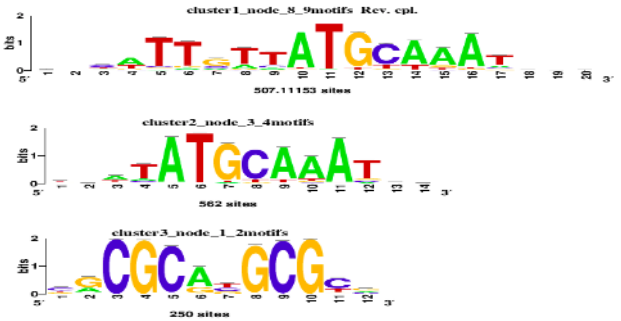


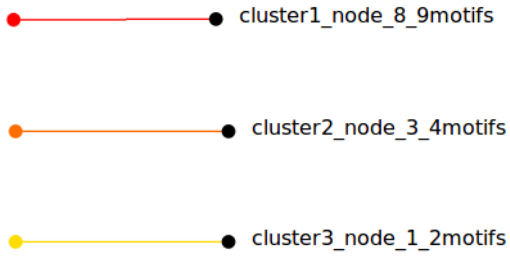


Clustering of motifs

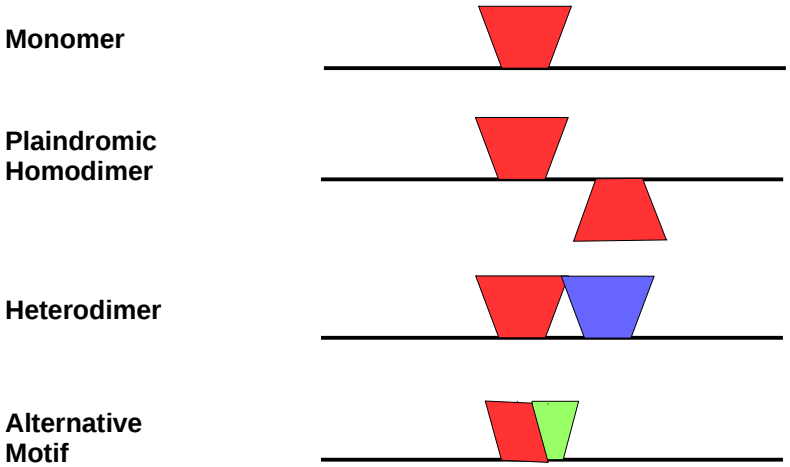


Reduce Motif Redundancy





Identification of Oct variants



Non-redundant motifs compared with DataBases

1.- Adapted from Thomas-Chollier et al. (2011) **RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets**. Nucleic Acids Research.

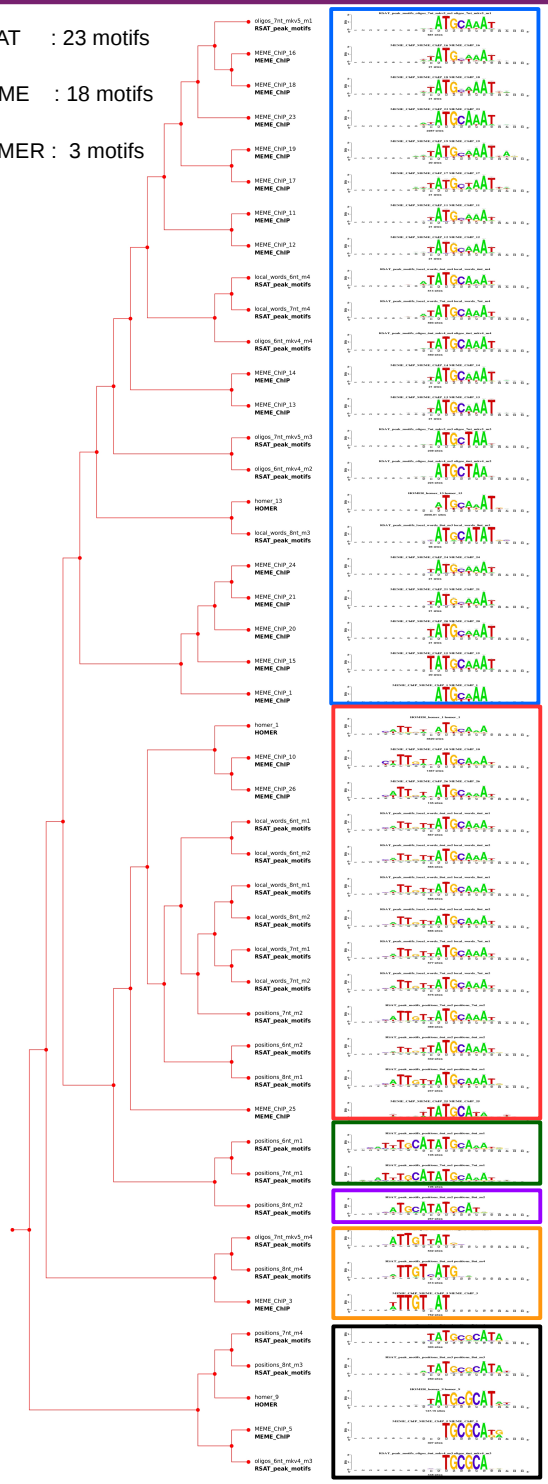
2.- Chen X et al. (2008). **Integration of External Signaling Pathwayswith the Core Transcriptional Network in Embryonic Stem Cells**. Cell

One cluster with Oct4 motifs

RSAT : 23 motifs

MEME : 18 motifs

HOMER : 3 motifs

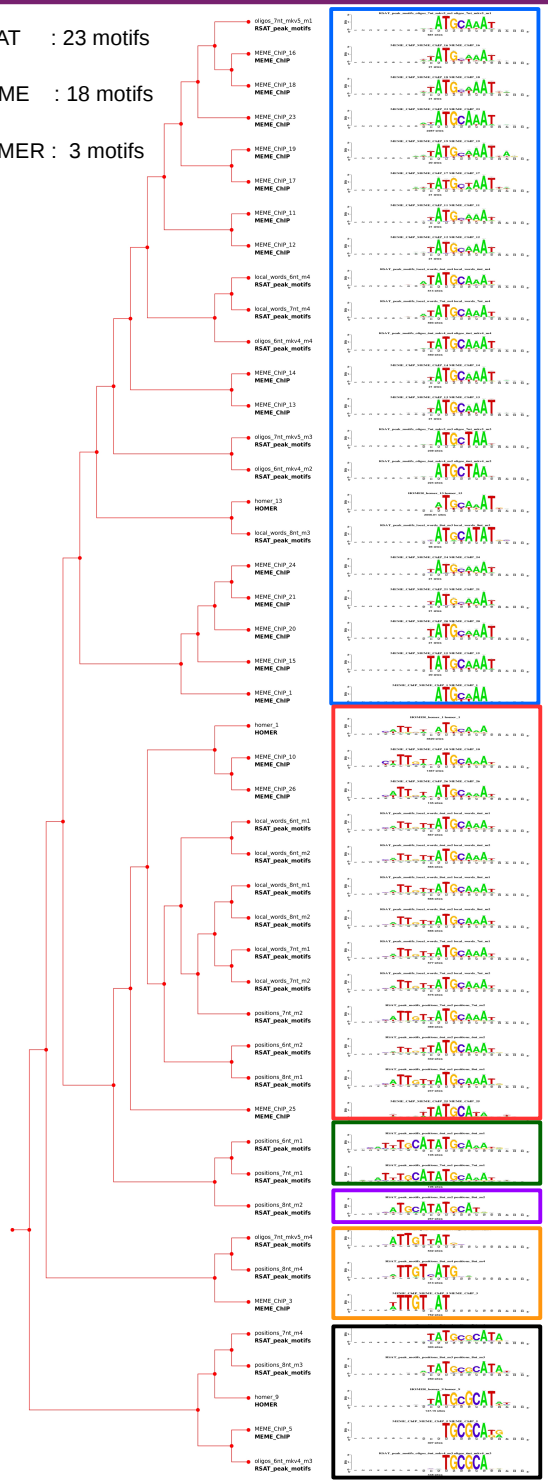


One cluster with Oct4 motifs

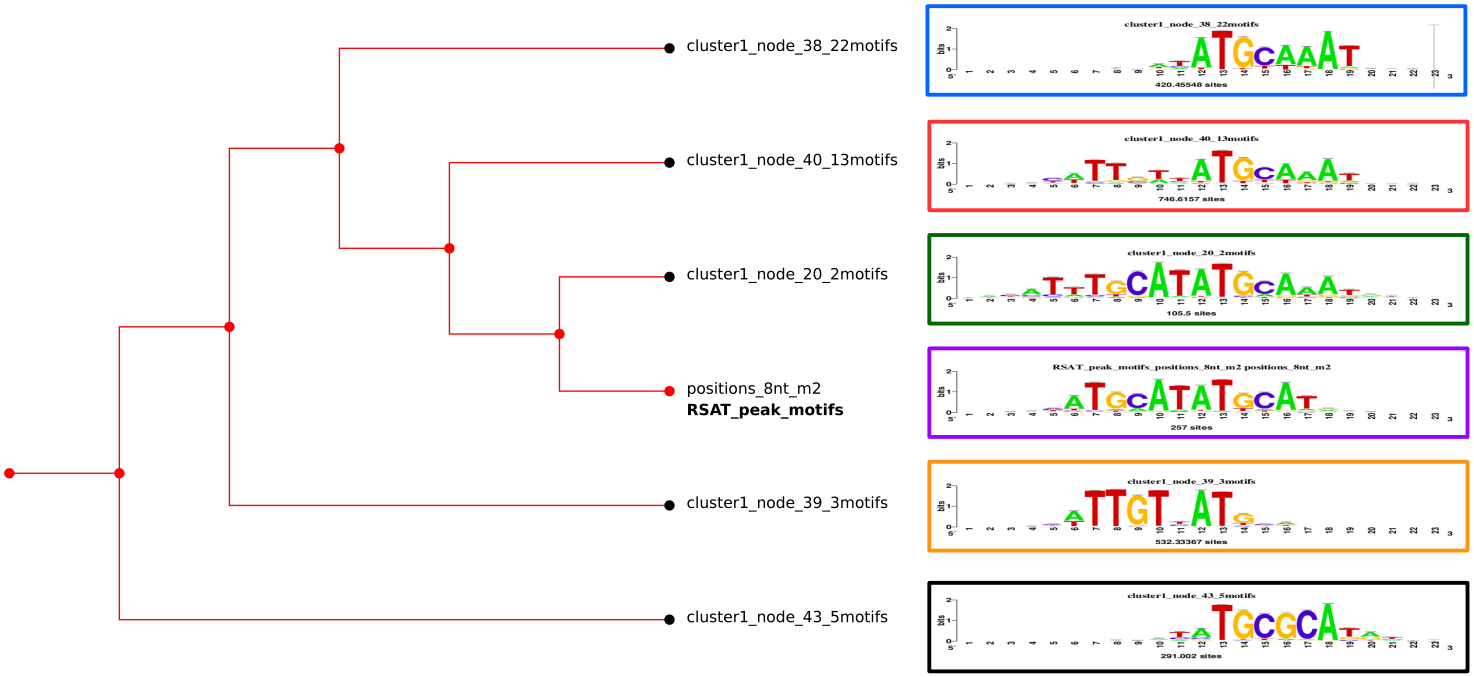
RSAT : 23 motifs

MEME : 18 motifs

HOMER : 3 motifs



Collapsed tree with non-redundant motifs

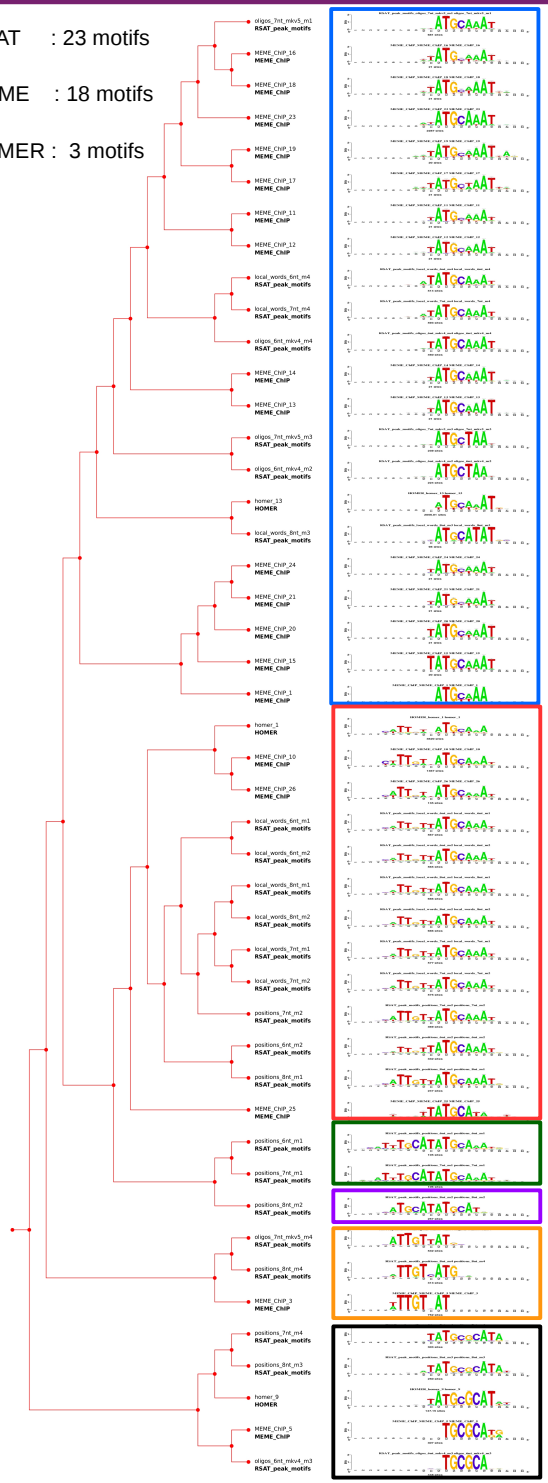


One cluster with Oct4 motifs

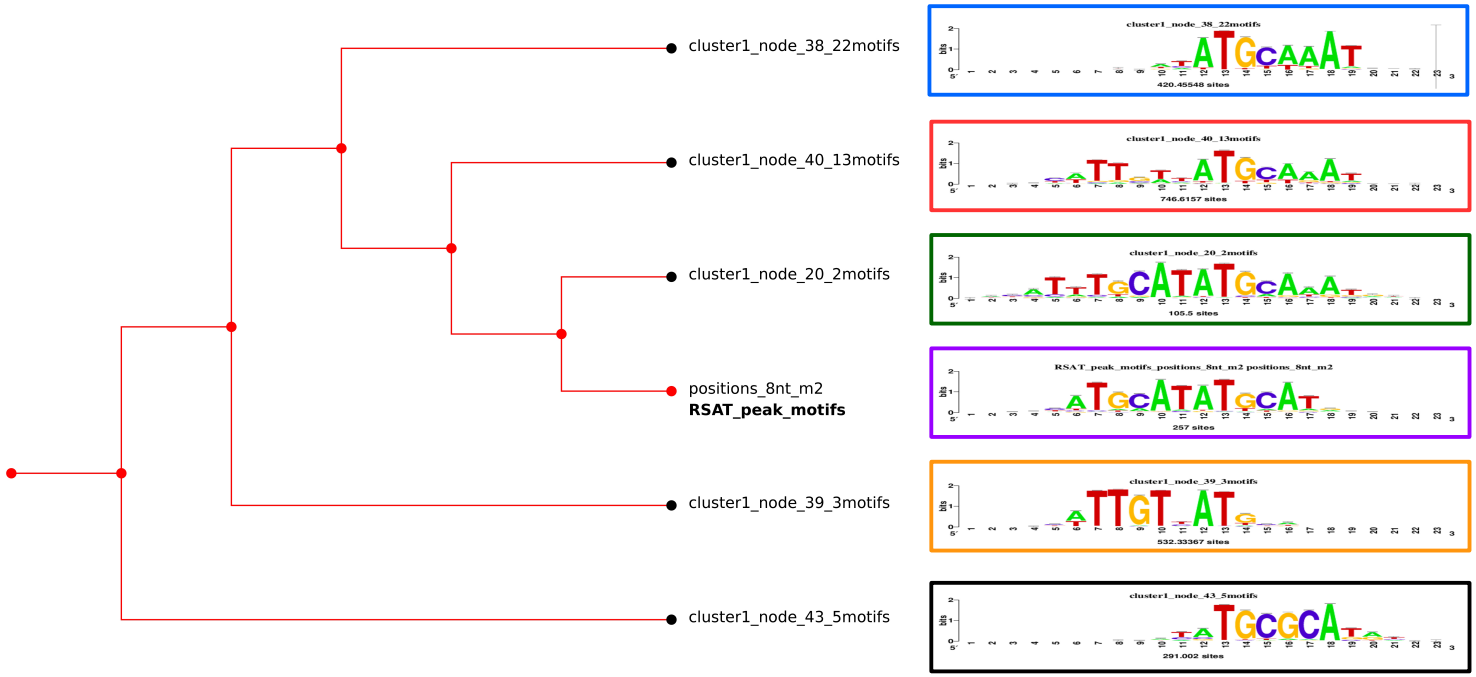
RSAT : 23 motifs

MEME : 18 motifs

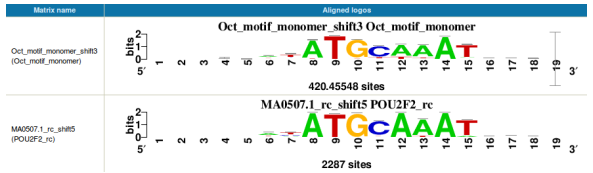
HOMER : 3 motifs



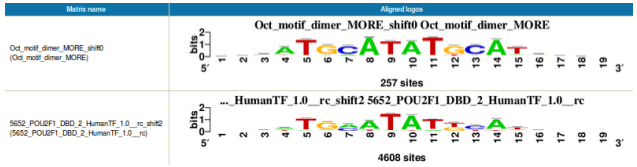
Collapsed tree with non-redundant motifs



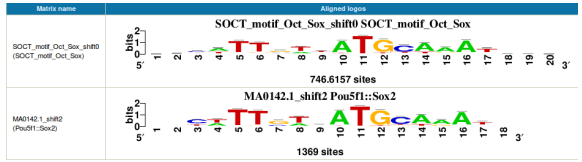
Canonical Oct



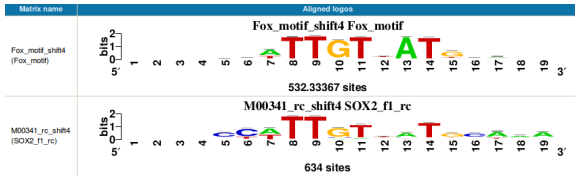
MORE (More palindromic Oct factor Recognition Element)



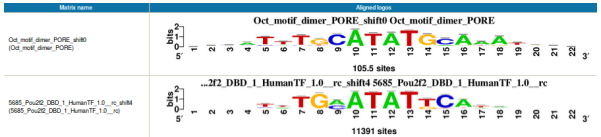
SOCT (Sox + Oct)



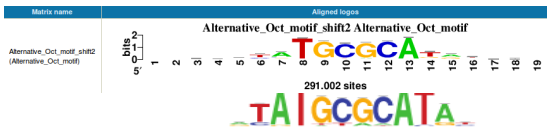
Sox



PORE (Palindromic Oct factor Recognition Element)



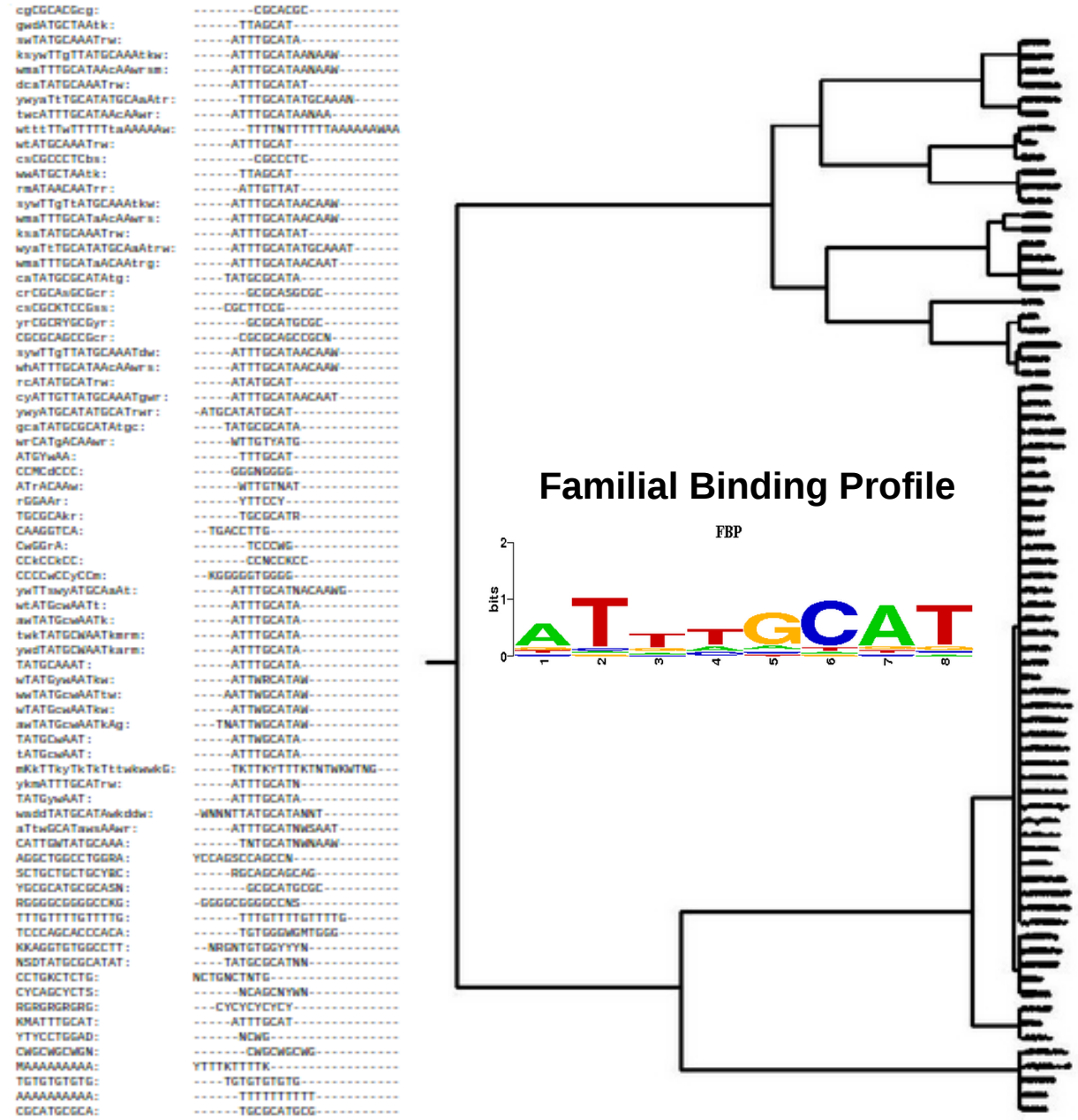
Alternative Oct motif



(The same) Motifs clustered with STAMP[1]

Alignment

Tree



Familial Binding Profile

- Pros:
  a) The fastest
  b) Several metrics can be selected
  c) Several alignment types allowed.
- Cons:
  a) The tree is not partitioned in a forest.
  b) The motifs are forced to be aligned in a single alignment.
  c) Only one FBP can be obtained.

1.- Mahony S and Benos P. (2007) **STAMP: a web tool for exploring DNA-binding motif similarities.** NAR.

Example 2 : extended analysis with 12 TFs

- **Main Objective**:

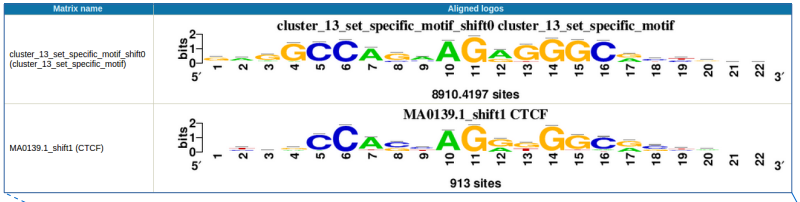
  Discover and Cluster motifs in 12 TF ChIP-seq peaks taken from Chen et al[1].
  (Oct4, Sox2, Nanog, c-Myc, m-Myc, CTCF, E2f1, Esrrb, Klf4, Stat3, Tcfcp21, Zfx)

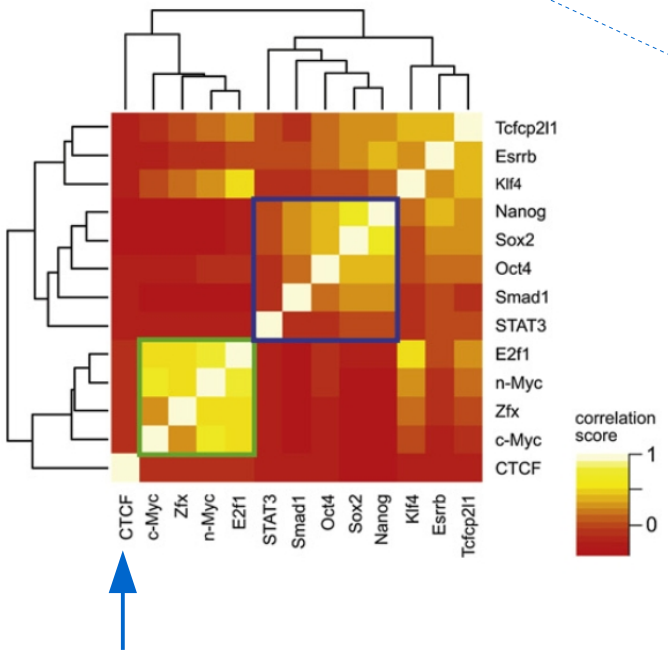
- **Specific Objectives**:

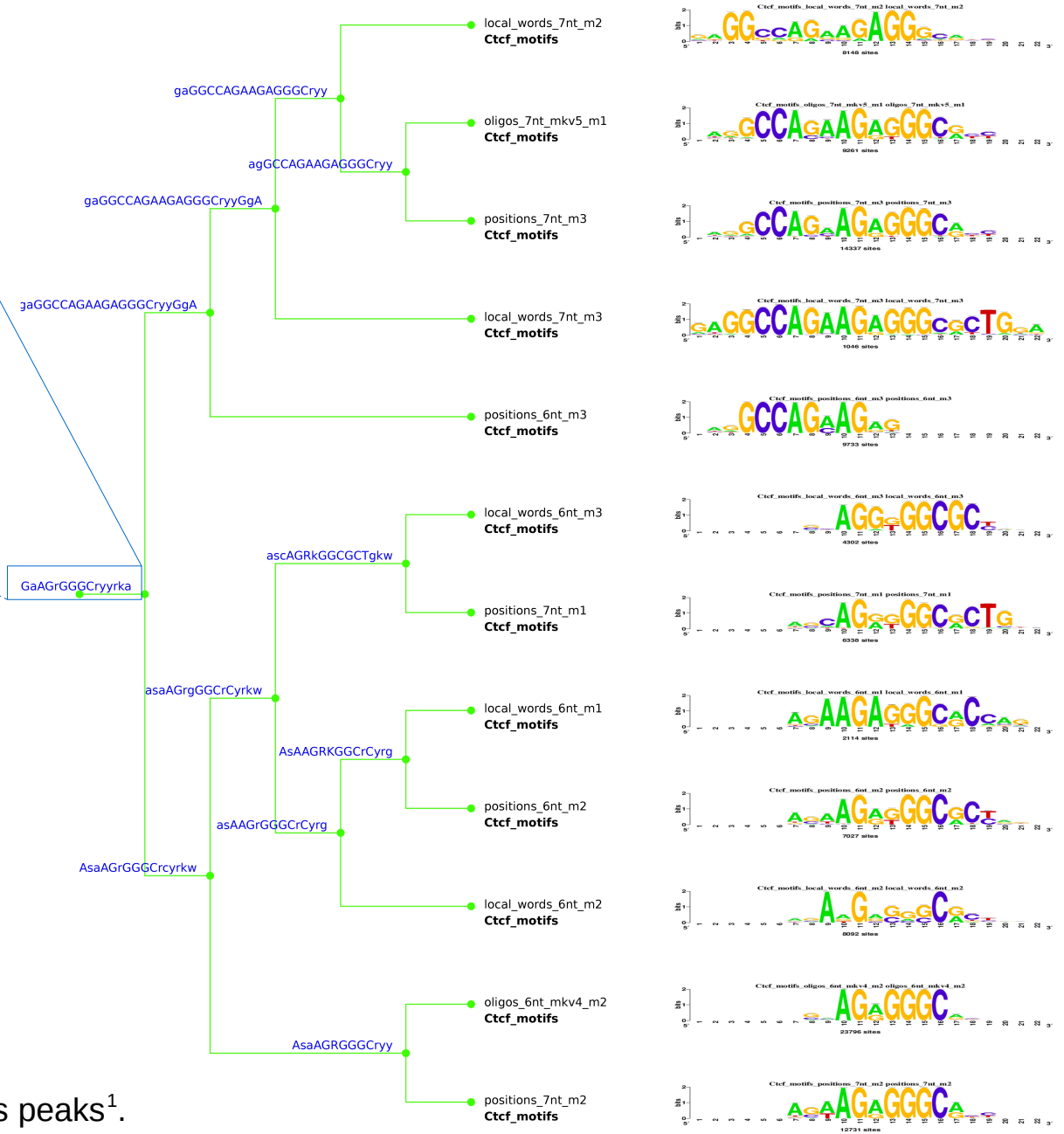
  - Identify the motifs found exclusively in one peak collection (set-specific).
  - Identify the motifs found several peak collections (recurrent).

1.- Chen X et al. (2008). **Integration of External Signaling Pathwayswith the Core Transcriptional Network in Embryonic Stem Cells**. Cell

Example 2 : ChIP-seq peaks specific motifs







No overlap between CTCF peaks with other TFs peaks[1].

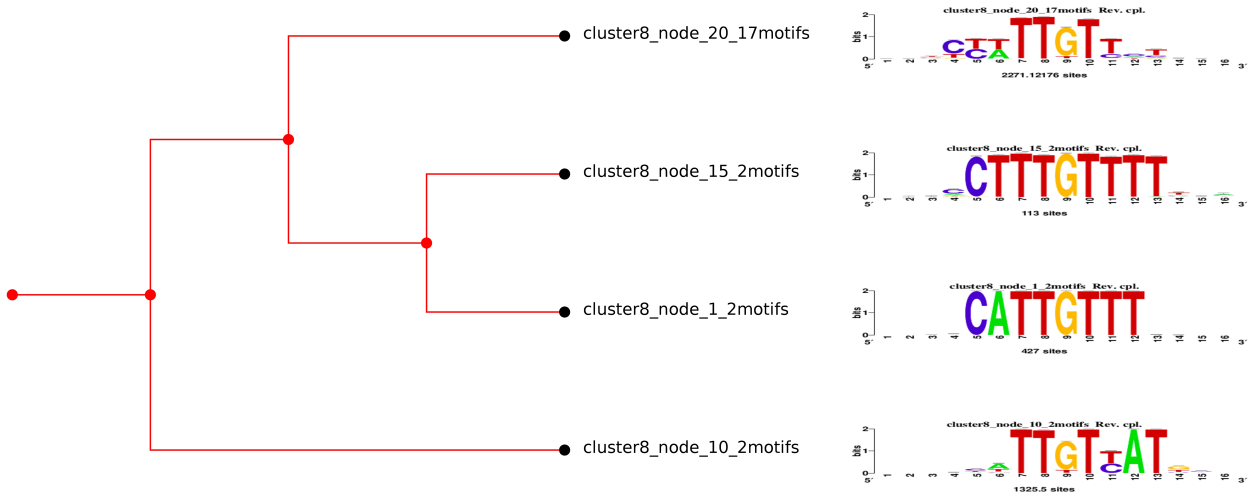
Figure taken from Chen *et al*.

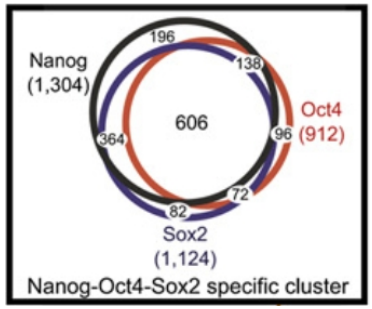
1.- Chen X et al. (2008). **Integration of External Signaling Pathwayswith the Core Transcriptional Network in Embryonic Stem Cells**. Cell

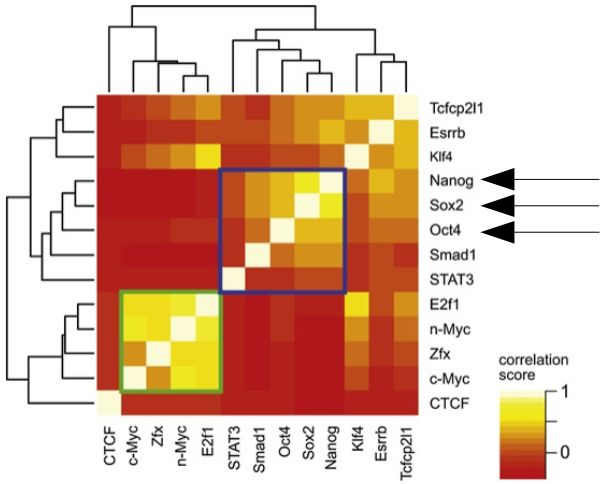
Example 2 : ChIP-seq peaks recurrent motifs

Sox2_motifs::13_motifs
Nanog_motifs::9_motifs
Oct4_motifs::1_motif

Sox Motifs







Oct4_motifs::15_motifs
Nanog_motifs::1_motif
Sox2_motifs::2_motifs

Oct Motifs

- High overlap between Nanog-Sox2-Oct4 peaks[1].
  Figures taken from Chen *et al*.
- The clusters corresponding to Oct4, Sox2 and Nanog are encompassed by motifs found only in these TF peaks.

1.- Chen X et al. (2008). **Integration of External Signaling Pathwayswith the Core Transcriptional Network in Embryonic Stem Cells**. Cell

Example 1 and 2: conclusions

- One motif can be found several times by different algorithms (redundancy).
- Some motifs were found only by one algorithm (complementarity).
- The logo alignment allows to identify the different Oct variants, highlighting their differences/similarities.
- The dynamic tree view allows to select a branch encompassing a coherent alignment.
- The collapsible tree allow to identify the non-redundant motifs.
- A large collection of redundant motifs can be reduced to a small non-redundant collection.
- Matrix-clustering can be used to find condition-specific or recurrent motifs.

TakeHome Messages

- There is no a 'best' metric to measure motif similarity. (But combining several metrics can help to properly compare or classify the motifs)
- The *motif redundancy issue* after using motif discovery tools can be faced with clustering of motifs.
- A large collection of redundant motifs can be reduced to a small non-redundant collection, which in practice reduce the
- Clustering several motif files (condition A vs condition B; RSAT vs MEME, Database1 vs Database2) allow to identify shared motifs or those corresponding to one collection.

Acknowledgments

- Jacques van Helden
- Morgane Thomas-Chollier
- Denis Thieffry
- Henry Buttler
- Nishant Thakur