

# Motif databases and comparison

**Morgane Thomas-Chollier**

*Computational systems biology - IBENS*

mthomas@biologie.ens.fr

The logo for IBENS (Institut de Biologie de l'École Normale Supérieure) features the acronym "IBENS" in a bold, black, sans-serif font. The text is centered within a circular area composed of a dense field of small, light blue and grey dots, creating a textured, particle-like effect. A thin horizontal line is positioned directly below the circular graphic.

IBENS

**M2 – Computational analysis of cis-regulatory sequences 2015/2016**

Denis Thieffry, Jacques van Helden and Carl Herrmann kindly shared some of their slides.

## 1 - Collections of motifs



## 2 - General principle of motif comparison

# Common motif problems

TF ?

5' - TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATCGAAACATACAT ...*HIS7*  
5' - ATGGCAGAATCACTTTAAACGTGGCCCCACCCGCTGCACCCTGTGCATTTTGTACGTTACTGCGAAATGACTCAACG ...*ARO4*  
5' - CACATCCAACGAATCACCTCACCGTTATCGTGACTCACTTTCTTTTCGCATCGCCGAAGTGCCATAAAAAATATTTTTT ...*ILV6*  
5' - TGCGAACAAAAGAGTCATTACAACGAGGAAATAGAAGAAAATGAAAAATTTTCGACAAAATGTATAGTCATTTCTATC ...*THR4*  
5' - ACAAAGGTACCTTCCTGGCCAATCTCACAGATTTAATATAGTAAATTGTCATGCATATGACTCATCCCGAACATGAAA ...*ARO1*  
5' - ATTGATTGACTCACTTTTCTCTGACTACTACCAGTTCAAATGTTAGAGAAAAATAGAAAAGCAGAAAAAATAAATAA ...*HOM2*  
5' - GGCGCCACAGTCCGCGTTTGGTTATCCGGCTGACTCATTCTGACTCTTTTTTGGAAAGTGTGGCATGTGCTTCACACA ...*PRO3*

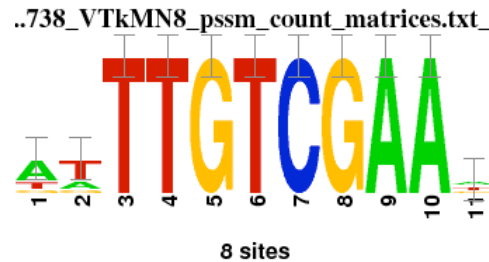
Co-expressed  
genes

## Questions

- Could we **discover** some signals (motifs) on the basis of these sequences ?
  - » This is a problem of **motif discovery** (“ab initio” motif detection)
- Can we afterwards **locate** the instances of these discovered motifs in the input sequences ?
  - » This is a problem of **pattern matching**.
- Can we **predict the transcription factor** that would bind the discovered motifs ?
  - » By comparison with a collection of known factors => **motif comparison problem**

# I have predicted a motif, what's next ?

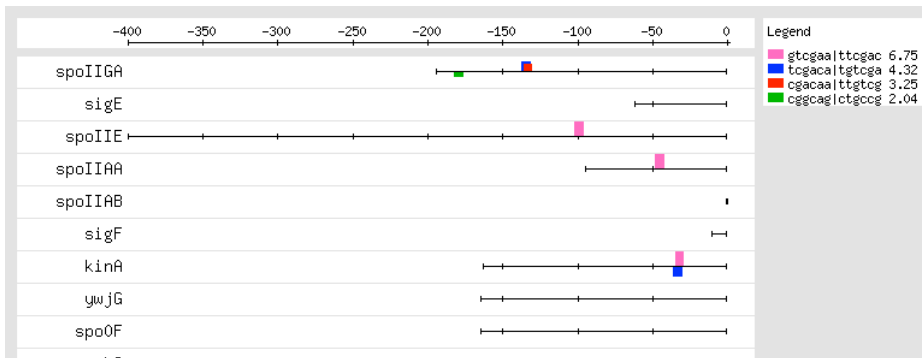
seq	identifier
gtcgaa	gtcgaa   ttcgac
tcgaca	tcgaca   tgtcga
cgacaa	cgacaa   ttgtcg
cggcag	cggcag   ctgccg



**Discovered motif**



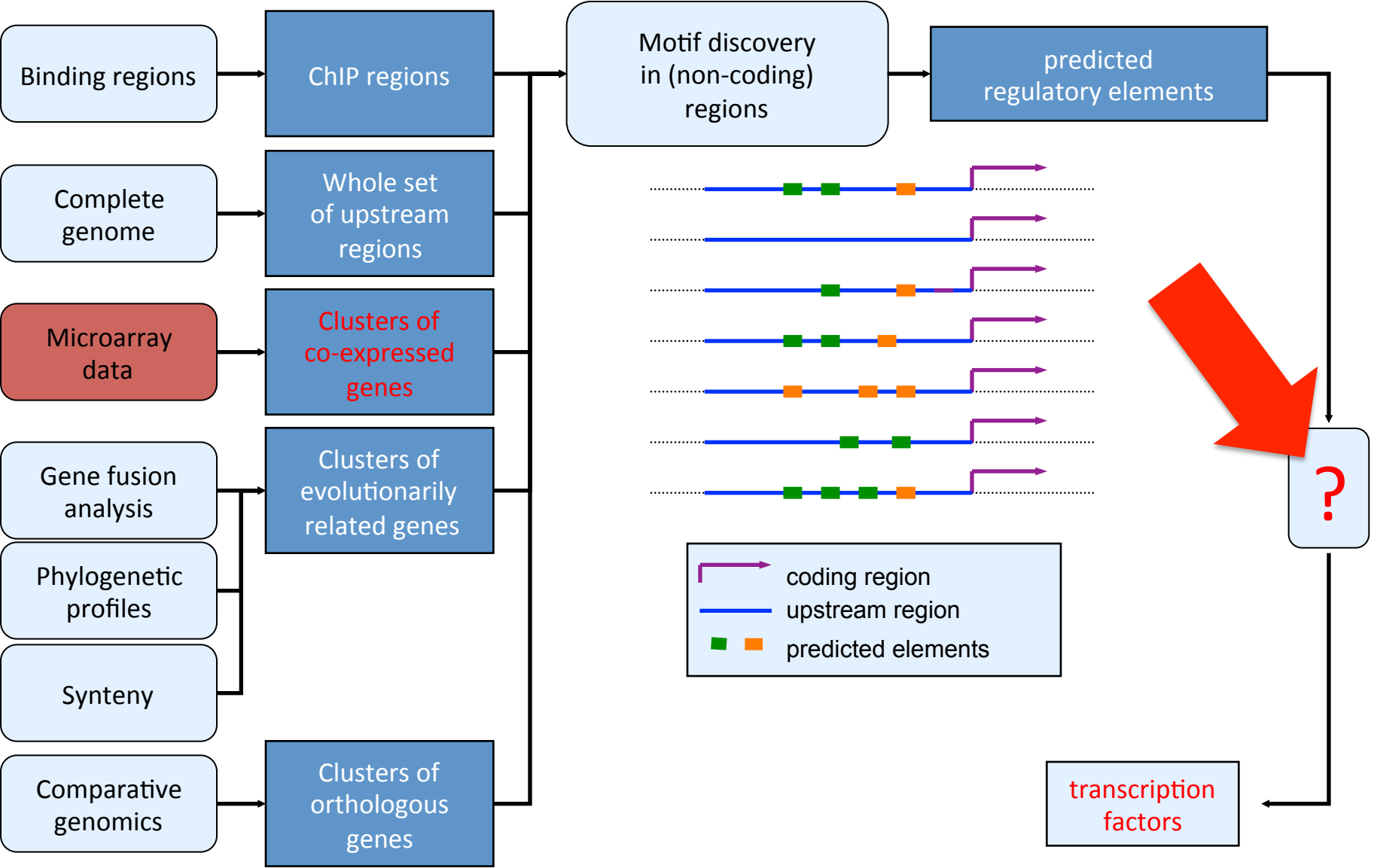
**locate** the instances of these discovered motifs in the input sequences



**predict the transcription factor**

- Compare motif against collections of known motifs
- Get more information from external sources

# Predict the transcription factor



## Collections of known motifs

- Public (free) vs commercial databases
- General vs organism or experiment specific databases
- Many different databases (not united, information spreaded over many different sources)
- Often linked to analyses tools... warning, they can be very basic tools !
  
- **Characteristics of all databases**
  - » Incomplete !!
  - » Redundant !!
  - » Heterogenity in terms of quality (mixture of matrices resulting from small- and large-scale experiments)
  - » Different file formats
  - » Original sequences used to make the matrix is often not accessible

# TRANSFAC - Gene transcription factor database

- Organisms
  - » Eukaryotes
  - » Particular emphasis on mammals (specially human, mouse, rat)
- Distribution
  - » The public version is not updated anymore (since 2007 !)
  - » **Commercial** version (TRANSFAC PRO)
  - » Distributed by BioBase™
    - <http://www.biobase.de/>
- Data content
  - » Transcription factors
  - » Binding sites
    - **Evidences !**
    - **Publications !**
  - » Position-specific scoring matrices
- Pattern matching tools (patch, match)

BIOBASE Biological Databases: TRANSFAC Gene Transcription Factor Database

http://www.biobase.de/pages/index.php?id=transfac

**BIOBASE**  
BIOLOGICAL DATABASES

BIOBASE Knowledge Library

Home • Contact Form • Sitemap • Imprint • Privacy Policy • Free Trials • **Login**

You are here: BIOBASE Knowledge Library / TRANSFAC databases / TRANSFAC

TRANSFAC Gene Transcription Factor Database

**TRANSFAC® - Gene Transcription Factor Database**

TRANSFAC - the internationally unique knowledge base - contains data on transcription factors, their experimentally-proven binding sites, and regulated genes. Its broad compilation of binding sites allows the derivation of positional weight matrices.

TRANSFAC's programs, Match and Patch, use the matrices and the site sequences themselves for performing the matrix-or pattern-based search of factor binding sites in regulatory DNA sequences. Thus, it is possible to make predictions for most gene promoters, which have not been studied in detail yet.

TRANSFAC also includes a tool to automatically visualize gene-regulatory networks being based on interlinked factor and gene entries in the database (gene regulation and gene expression).

In addition, TRANSFAC comprises

- extensive information on transcription factors and their structures, functions, expression patterns
- a recently added table for in vivo binding sequences from ChIP on chip experiments.

**Affymetrix GeneChip® Compatibility**

TRANSFAC works in conjunction with our ExPlain™ analysis system to apply a new knowledge driven approach to the analysis of whole complexes of coexpressed genes. The internal Composite Module Analyst (CMA) is a genetic algorithm for analysis and prediction of relevant promoters in the identified set of given genes obtained for sources such as Affymetrix GeneChip® Arrays. This combinatorial analysis drops false positive rates significantly and enables scientists to find potential causes for specific cellular events.

The power of correct prediction in ExPlain is driven by TRANSFAC®, a knowledge base of high quality, expert level, manually curated published scientific literature. TRANSFAC presents data on transcription factors, their experimentally-proven binding sites, and regulated genes. To

Done

# TRANSFAC – matrix example – V\$SOX2\_Q6

## Field descriptions

AC Accession no.  
 XX (field separator)  
 ID Identifier  
 DT Date; author  
 NA Name of the binding factor  
 DE Short factor description  
 BF List of linked factor entries

PO A C G T Position within the aligned sequences,  
 01 frequency of A, C, G, T residues, resp.;  
 02 last column: deduced consensus in  
 03 IUPAC 15-letter code

BA Statistical basis  
 BS Factor binding sites underlying the matrix  
 BS (SITE accession no.; Start position for matrix sequence;  
 length of sequence used;  
 BS number of gaps inserted; strand orientation)  
 CC Comments  
 RX MEDLINE ID  
 RN Reference no.  
 RA Reference authors  
 RT Reference title  
 RL Reference data  
 //

```
AC M01272
XX
ID V$SOX2_Q6
XX
DT 08.07.2009 (created); dtc.
CO Copyright (C), Biobase GmbH.
XX
NA SOX2
XX
BF T09507; Sox-xbb1; Species: mouse, Mus musculus.
BF T01836; Sox2; Species: mouse, Mus musculus.
BF T04915; Sox2; Species: human, Homo sapiens.
BF T01837; Sox2; Species: chick, Gallus gallus.
BF T10231; Sox2; Species: Mammalia.
BF T09970; Sox2; Species: human, Homo sapiens.
BF T10885; Sox2; Species: monkey, Cercopithecus aethiops.
XX
P0      A      C      G      T
01      6      2      4      4      N
02      7      2      3      4      N
03      4      6      2      4      N
04      4      5      4      3      N
05      2      9      1      4      C
06      0      12     0      4      C
07      8      0      0      8      W
08      0      0      0      16     T
09      0      0      0      16     T
10      0      0      16     0      G
11      0      0      0      16     T
12      2      2      2      9      T
13      7      2      0      6      W
14      0      2      2      11     T
15      1      0      9      5      K
16      4      6      3      2      N
XX
BA 16 compiled sequences
XX
BS gccctcattgttatgc; R15133; 13; 16;; n.
BS AAActCTTTGTTTGGa; R15201; -1; 16;; p.
BS ttcaccattgttctag; R15231; 11; 16;; n.
BS GACTCTATTGTCTCTG; R15267; 11; 16;; p.
BS GATATCTTTGTTTCTT; R16367; -4; 16;; p.
BS tgcacctttgttatgc; R17099; 5; 16;; n.
BS aattccattgttatga; R19276; 15; 16;; n.
BS aaactctttgtttggga; R19367; 20; 16;; n.
BS atggacattqtaatqc; R19510; 15; 16;; n.
```



# JASPAR

- <http://jaspar.genereg.net/>
- Public database
- Data content
  - » PSSM
  - » “sites” (i.e. sequences having served to build the matrix, but no genomic position)
  - » Core: transcription factor-specific matrices
  - » Collection: matrices for families of transcription factors
- Tools
  - » Pattern matching, matrix randomization

MA0212.1 detailed information

http://jaspar.genereg.net/cgi-bin/jaspar\_db.pl?ID=MA0212.1&rm=present&collection=CORE

Summary page for ID: MA0212.1 NAME: bcd from the JASPAR CORE database

DATA	
name	bcd
class	Helix-Turn-Helix
family	Homeo
species	<i>Drosophila melanogaster</i>
tax_group	insects
acc	P09081
type	bacterial 1-hybrid
medline	18332042
Pazar ID	
comment	-

VERSION INFORMATION

There is only one version of the model

SITES	
Show me all the binding sites	<a href="#">...as web page</a>
	<a href="#">...as fasta file</a>

SEQUENCE LOGO

Make a SVG logo

FREQUENCY MATRIX

A	[ 0 20 22 0 0 0 ]
C	[ 0 0 0 0 22 21 ]
G	[ 0 0 0 1 0 0 ]
T	[ 22 2 0 21 0 1 ]

Reverse complement

Sequences for model MA0212.1

Site	Occurences
tgt <b>TAATCC</b> c	1
tg <b>GGATTA</b> ta	1
ttac <b>TAATCC</b>	1
gct <b>TAATCC</b> g	1
gg <b>TAATCC</b> g	1
agc <b>TTATCC</b>	1
gaga <b>TAATCC</b>	1
gtcc <b>TAATCC</b>	1
cgt <b>TAATCT</b> c	1
at <b>GGATTA</b> ga	2
cgct <b>TAATCC</b>	1
cg <b>ggTAATCC</b>	1
<b>GGCTTA</b> agcc	1
tgt <b>TAATCC</b> g	1
tgt <b>TAATCC</b>	1
tct <b>TAATCC</b> c	1
gg <b>TTATCC</b> g	1
g <b>cgTAATCC</b> a	1
gggt <b>TAATCC</b>	1
tcta <b>TAATCC</b>	1
gg <b>ttTAATCC</b>	1

# RegulonDB : Transcriptional regulation in *Escherichia coli*

- RegulonDB Web site
  - » <http://regulondb.ccg.unam.mx/>
- Model organism: *Escherichia coli*
- Data content
  - » Transcription factors
  - » Transcription factor binding sites (TFBS)
  - » Position-specific scoring matrices (PSSM)
  - » Promoters
  - » Operons
- Collaboration with EcoCyc
  - » EcoCyc is the reference database about metabolism in *Escherichia coli*
  - » RegulonDB is integrated in the EcoCyc database

The screenshot shows the RegulonDB 6.1 website in a browser window. The address bar displays <http://regulondb.ccg.unam.mx/>. The page features a navigation menu with links for 'Main Page', 'Using RegulonDB', 'Tools', 'Downloads', and 'About RegulonDB'. A search bar is located in the top right corner, with 'Gene' selected in the dropdown and a 'Go' button. The main content area is dominated by a large blue graphic of a hand touching a DNA double helix. Text on the graphic reads: 'Currently the **major** electronically-encoded **regulatory network** of any free-living **organism**'. Below this, three bullet points describe the database's features: 1. Simplified navigation streams for genes, operons, and regulons. 2. Continuously updated curated knowledge of original scientific literature complemented with comprehensive computational predictions. 3. A graphic and text-integrated environment with friendly navigation where regulatory information is always at hand. To the right of the main graphic, there are several sidebar sections: 'News' (RegulonDB 6.1 update), 'Get started RegulonDB' (learn what you can do), 'Tools' (Nebulon Tool, Genome Browser), 'Contact information' (Contact us, Suggestions), and 'Funding' (NIH grant R01-GM71962). At the bottom left, there are logos for the UNAM and CCG (Centro de Ciencias Genómicas). At the bottom right, there are four small thumbnail images representing different biological or computational aspects.

## Other databases

---

- PAZAR <http://www.pazar.info/>
  - » Unification of independent collection of transcription factor binding sites and motifs.
- YeasTract <http://www.yeasttract.com/>
  - » **Yeast-specific database. Factors, binding sites and motifs + tools.**
- FlyReg <http://www.flyreg.org/>
  - » Drosophila DNase I Footprint Database
- PlantCARE  
<http://bioinformatics.psb.ugent.be/webtools/plantcare/html/>
  - » Plant Cis-Acting Regulatory Elements

## Some new collections

---

- CHIP-seq => generation of high-quality motifs for many TFs
  - » Plenty of new databases / collections (supp data of articles)
- Meta-databases: to regroup all these little collections
  - » FootprintDB (<http://floresta.eead.csic.es/footprintdb/>)

## Welcome to footprintDB

---

Current version of **footprintDB** includes:

- **3887 Transcription Factors** (TFs, 3095 unique)
- **4681 Position Specific Scoring Matrices** (PSSMs, PWMs or DBMs, 4646 unique)
- **22056 DNA Binding Sites** (DBSs, 18840 unique)

extracted from the [literature and other repositories](#).

# Redundancy in databases

---

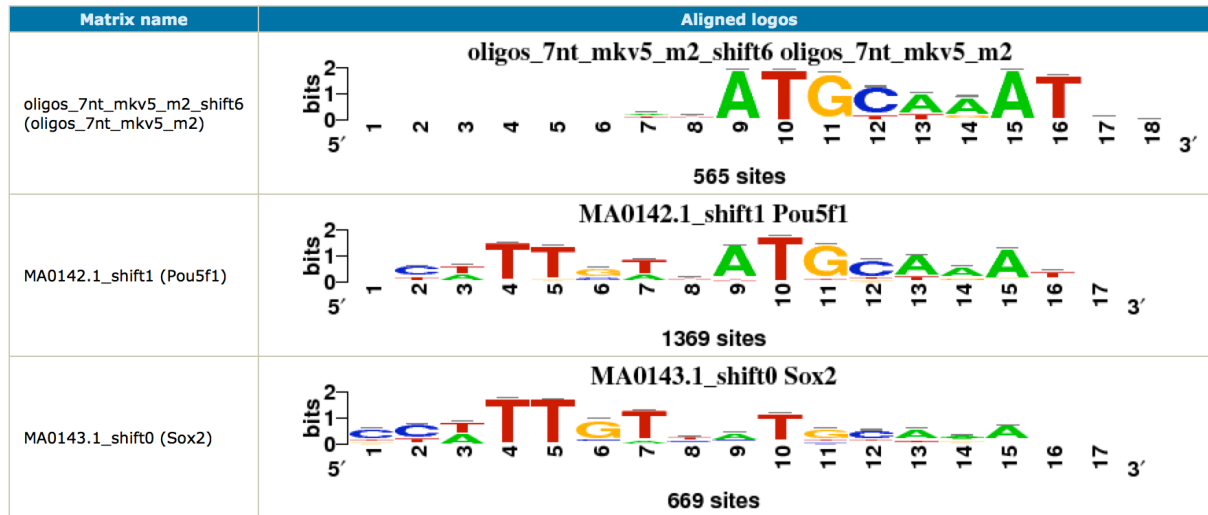
- High level of redundancy
  - » Within a database (several entries for a given TF)
  - » Across databases (very similar or identical motif listed)

*How to obtain a non-redundant dataset ?*

- Compare the motif + cluster them
  - » Often a manual and not reproducible work
  - » Very few motif clustering program

# Matrix-matrix comparison

- Basic **algorithm**:
  - » Shifting a matrix against another one to align it best
  - » Score the matrix alignment => distance
  - » Return the minimal distance and the shift position

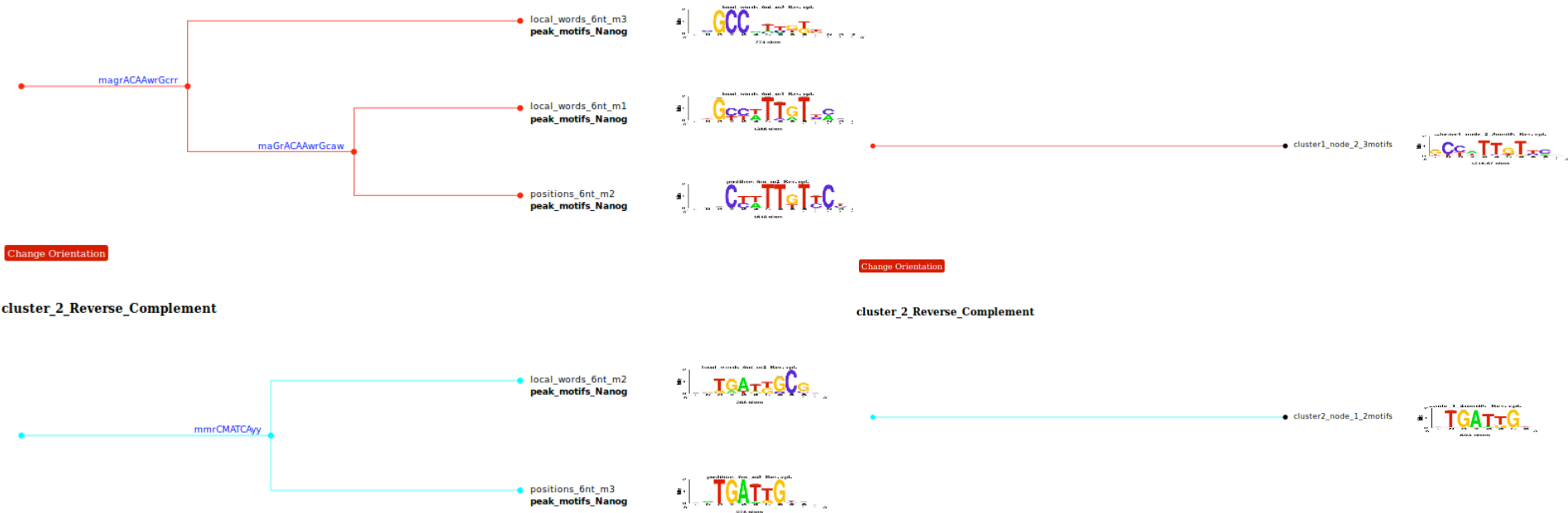


- In RSAT: **compare-matrices**
- Other tools :
  - » STAMP (<http://www.benoslab.pitt.edu/stamp/>)
  - » Tomtom (<http://meme.nbcr.net/meme/cgi-bin/tomtom.cgi>)

# Matrix clustering

- New program in RSAT:
  - » Matrix-clustering (still active development, unpublished)

Dynamic visualisation of the clusters, allows collapsing



# Applications of motif comparison and problems

- Interpretation of discovered patterns (e.g. from microarray clusters)
  - » Compare discovered patterns with annotated cis-acting elements in order to predict potential trans-acting factors.
- Compare patterns discovered in different datasets (e.g. co-expressed clusters)
- Compare patterns discovered in different organisms
- **Issues**
  - » Types of comparisons
    - String-based versus string-based
    - Matrix-based versus matrix-based
    - Comparison between string-based and matrix-based patterns
  - » Scoring the matching
    - Boolean matching (TRUE or FALSE)
    - Count of matching residues
    - P-value to estimate the significance of the matching