Motif databases and comparison

Morgane Thomas-Chollier

Computational systems biology - IBENS

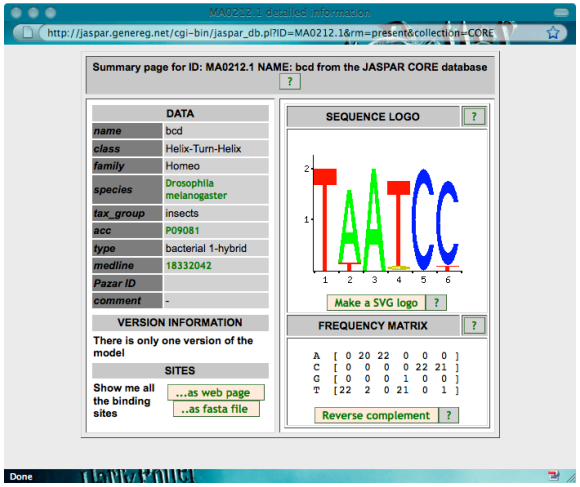
mthomas@biologie.ens.fr

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

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

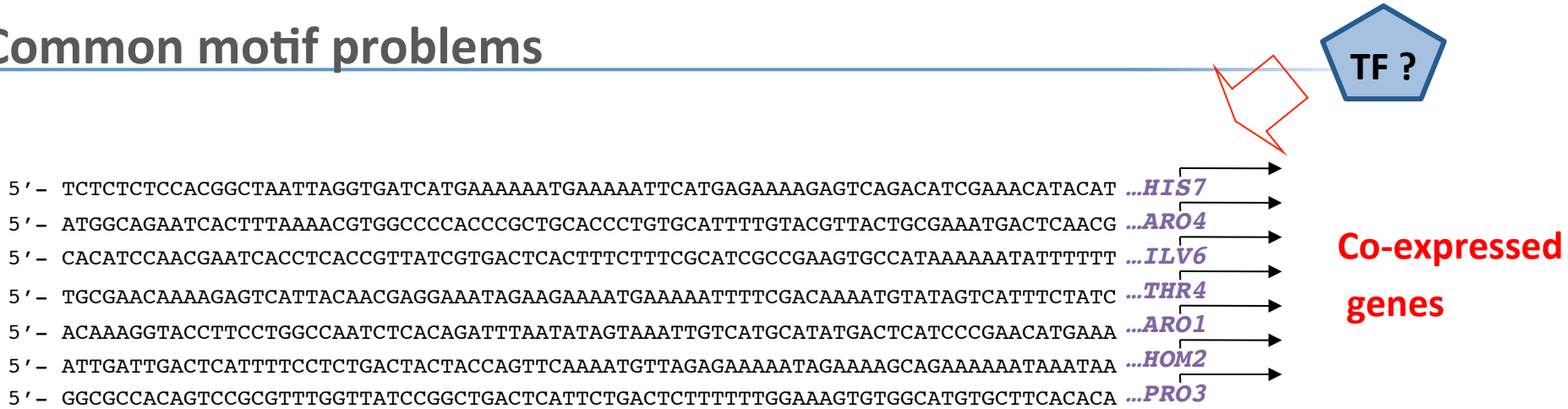
Aim of the course

1 - Collections of motifs



2 - General principle of motif comparison

Common motif problems



Questions

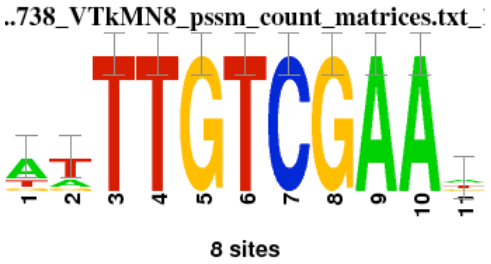
- Could we **discover** some signals (motifs) on the basis of these sequences ?
  - » This is a problem of **motif discovery** ("ab initio" motif detection)

- 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**

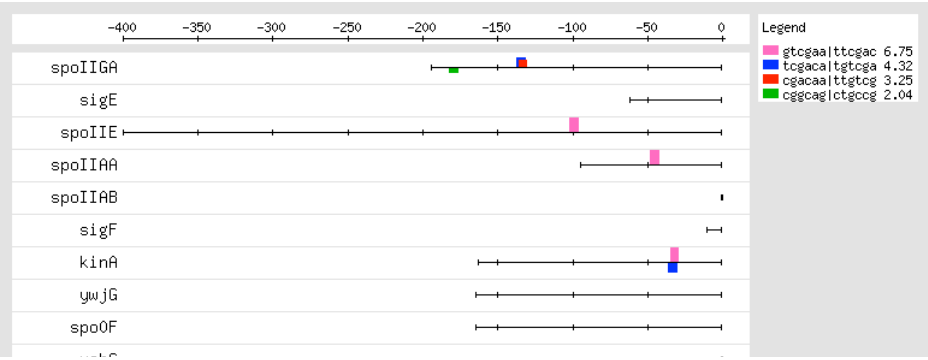
I have predicted a motif, what's next ?

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



Discovered motif

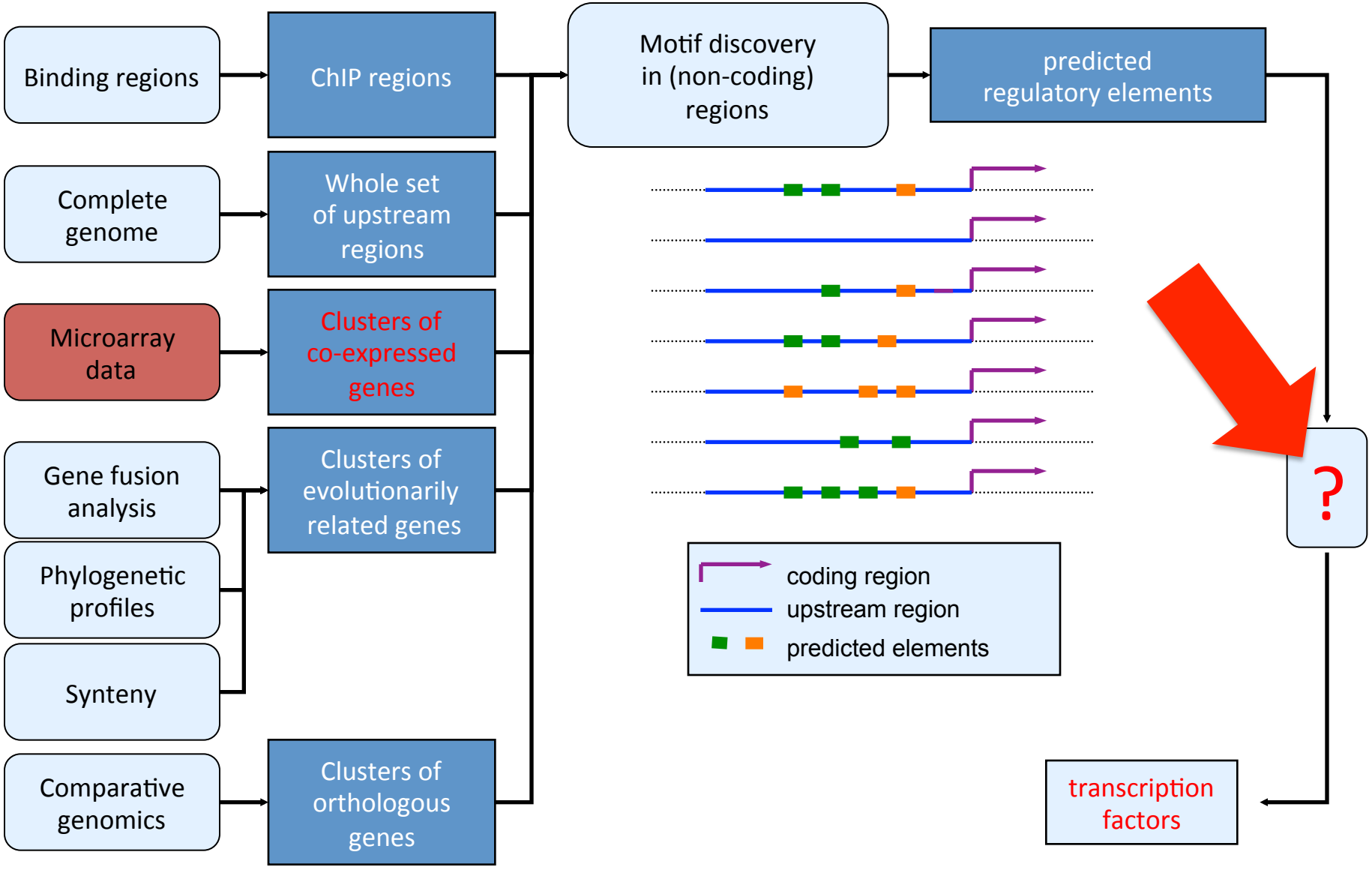
locate the instances of these discovered motifs in the input sequences



predict the transcription factor

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

Predict the transcription factor



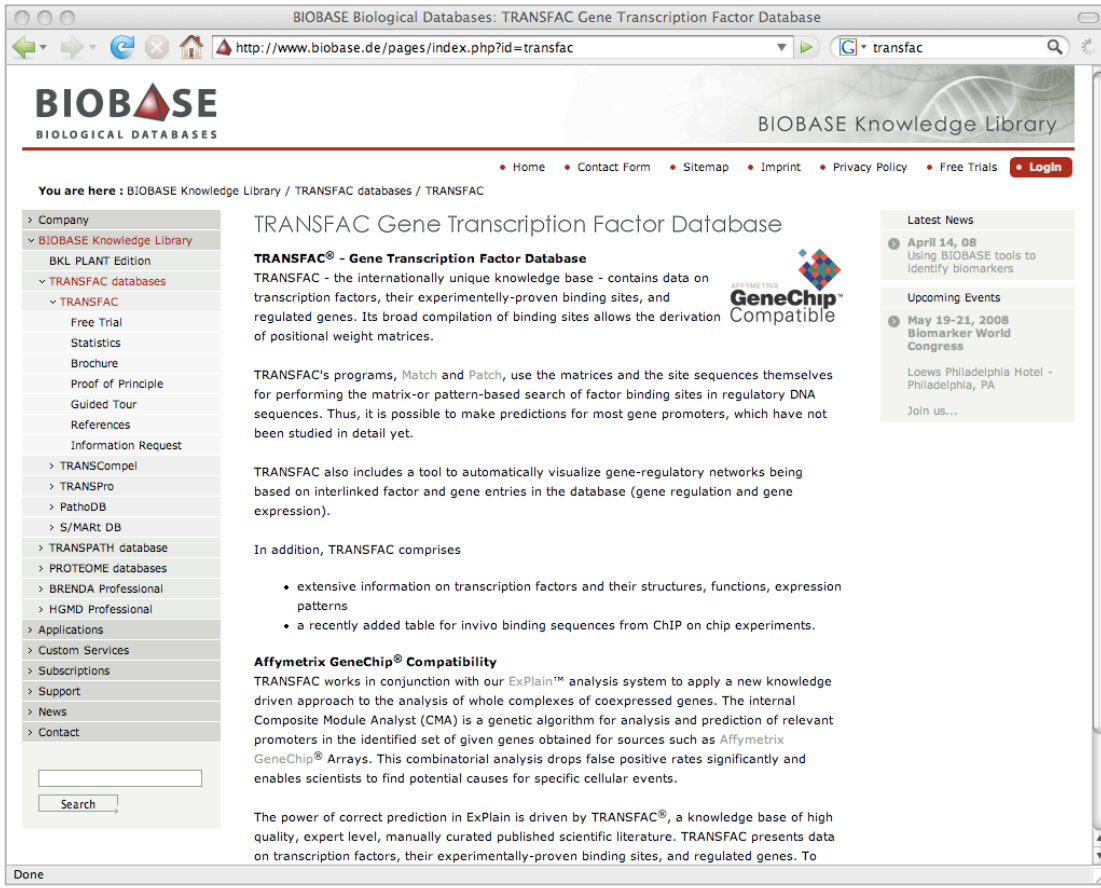
Collections of known motifs

- 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 BioBaseTM
    - – http://www.biobase.de/
- Data content
  - » Transcription factors
  - » Binding sites
    - – **Evidences !**
    - – **Publications !**
  - » Position-specific scoring matrices
- Pattern matching tools (patch, match)



TRANSFAC – matrix example – V$SOX2_Q6

Field descriptions

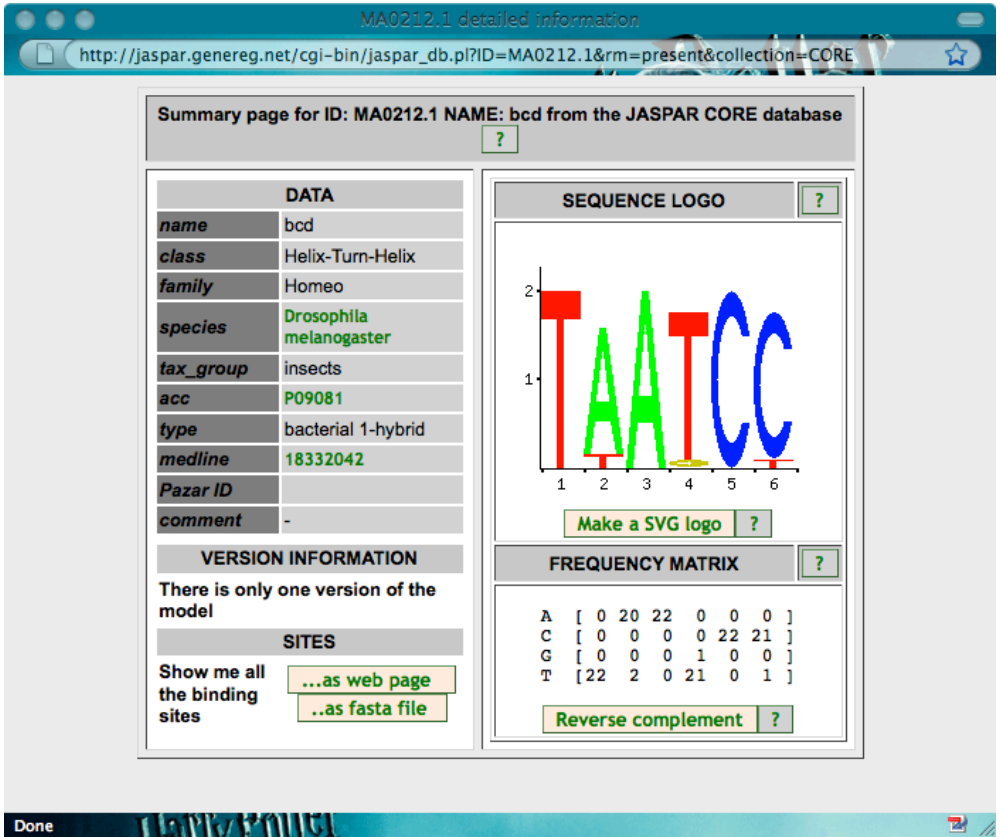
AC Accession no.
XX (field separator)
ID Identifier
DT Date; author
NA Name of the binding factor
DE Short factor description
BF List of linked factor entries

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  aaactctttgtttgga; R19367; 20; 16;; n.
BS  atggacattgtaatgc; 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



Sequences for model MA0212.1

Site	Occurences
tgt**TAATCC**c	1
tg**GGATTA**ta	1
tta**cTAATCC**	1
gct**TAATCC**g	1
ggt**TAATCC**g	1
agc**TTATCC**	1
gaga**TAATCC**	1
gtcc**TAATCC**	1
cgt**TAATCT**c	1
at**GGATTA**ga	2
cgct**TAATCC**	1
cggg**TAATCC**	1
GGCTTAagcc	1
tgt**TAATCC**g	1
tgt**TAATCC**	1
tct**TAATCC**c	1
gg**TTATCC**g	1
gcg**TAATCC**a	1
gggt**TAATCC**	1
tcta**TAATCC**	1
ggtt**TAATCC**	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

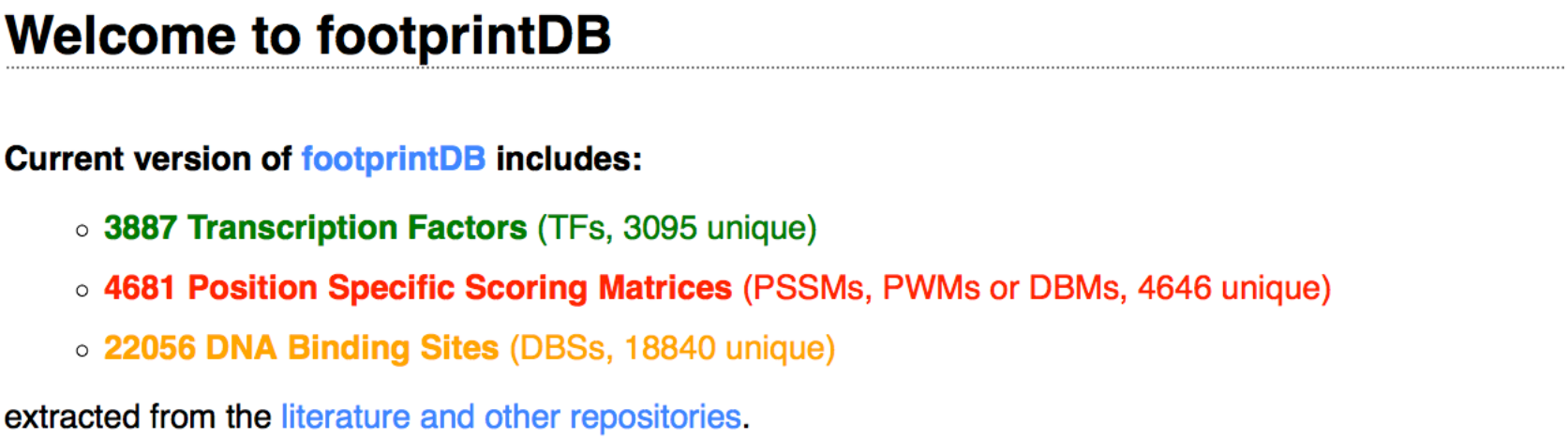
Other databases

- PAZAR http://www.pazar.info/
  - » Unification of independent collection of transcription factor binding sites and motifs.
- **YeasTract http://www.yeastract.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/)



Redundancy in databases

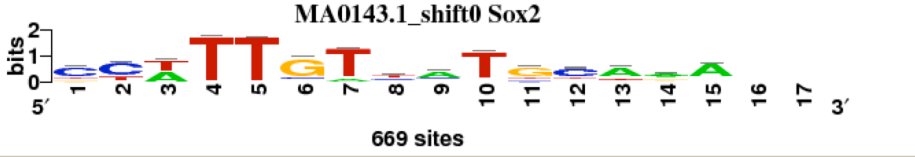
- High level of redundancy
  - » Within a database (several entries for a given TF)
  - » Across databases (very similar or identical motif listed)

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

Matrix name	Aligned logos
oligos_7nt_mkv5_m2_shift6 (oligos_7nt_mkv5_m2)	
MA0142.1_shift1 (Pou5f1)	
MA0143.1_shift0 (Sox2)	

- 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