Hong SUN
Phd Defense
BIOI@KULeuven
7 July 2011

Computational Discovery of *Cis*-regulatory Modules Based on Itemset Mining

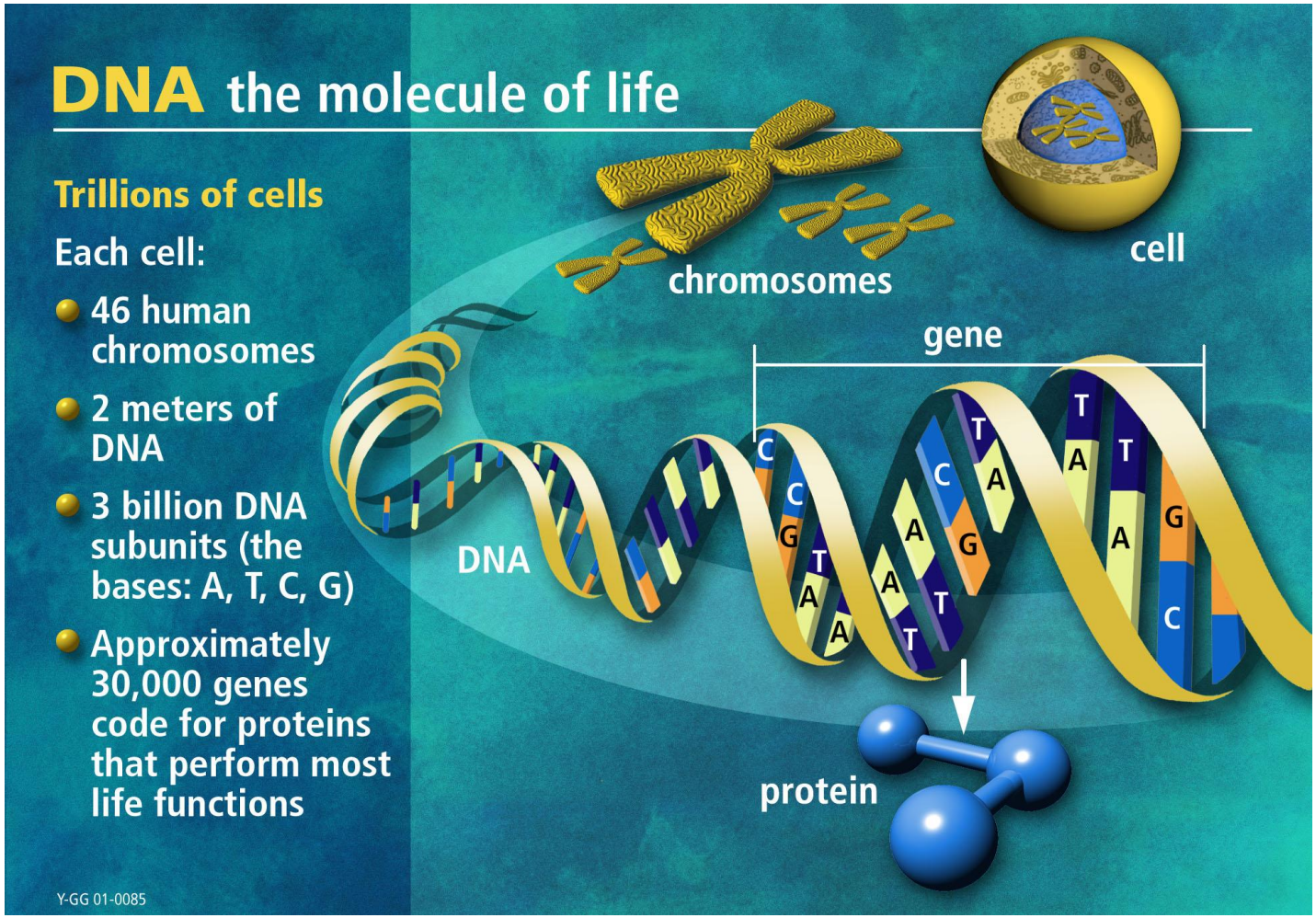
Promoters: Prof. Bart De Moor, Prof. Kathleen Marchal
Chairman: Prof. Anny Haegemans
Jury members: Prof. Bart De Moor, Prof. Kathleen Marchal
Prof. Yves Moreau , Prof. Annemieke Verstuyf
Prof. Jozef Vanderleyden, Dr. Tim Van Den Bulcke

Overview

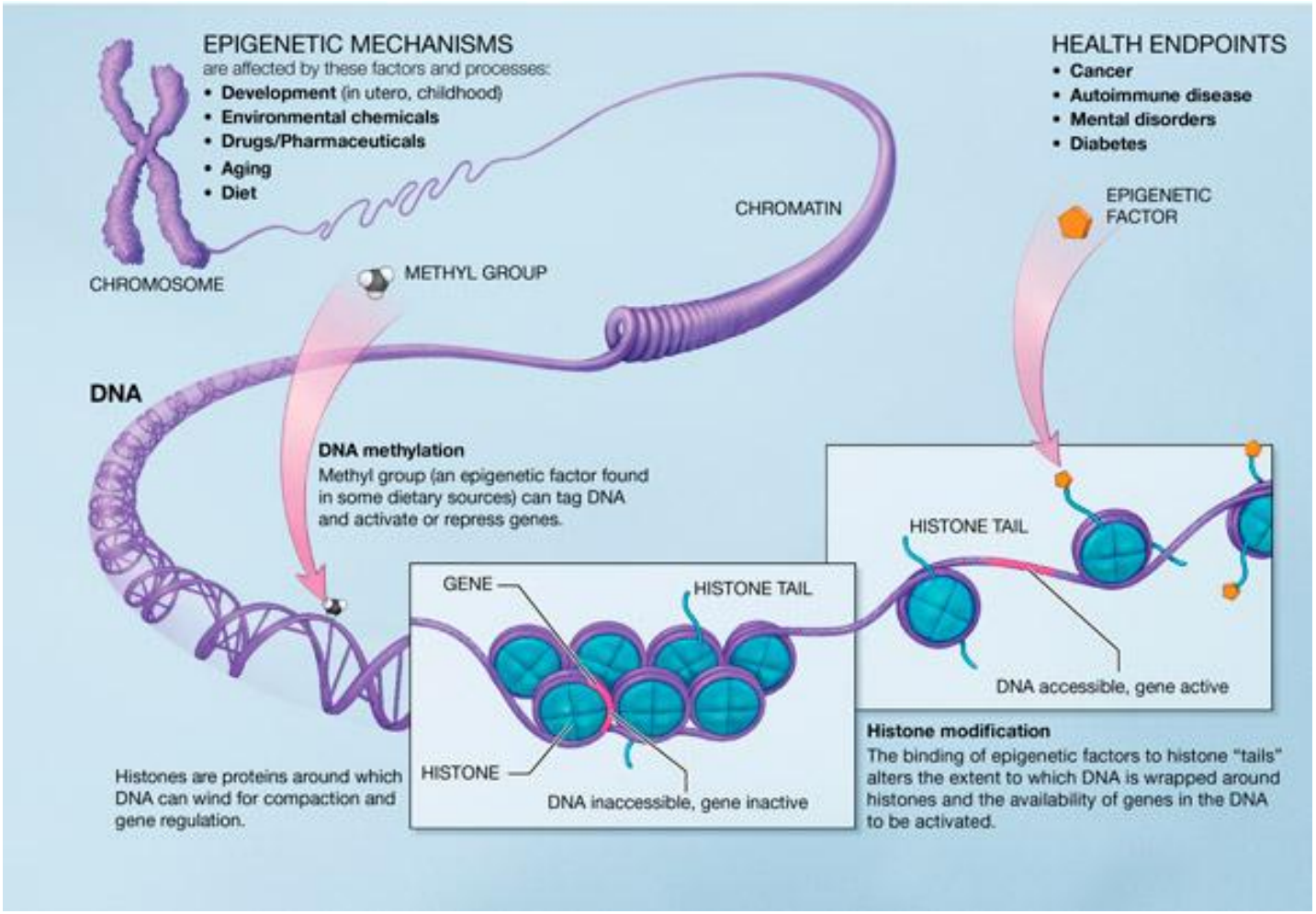
- Introduction
- ModuleDigger
- CPModule
- Conclusions
- Acknowledgements

Introduction

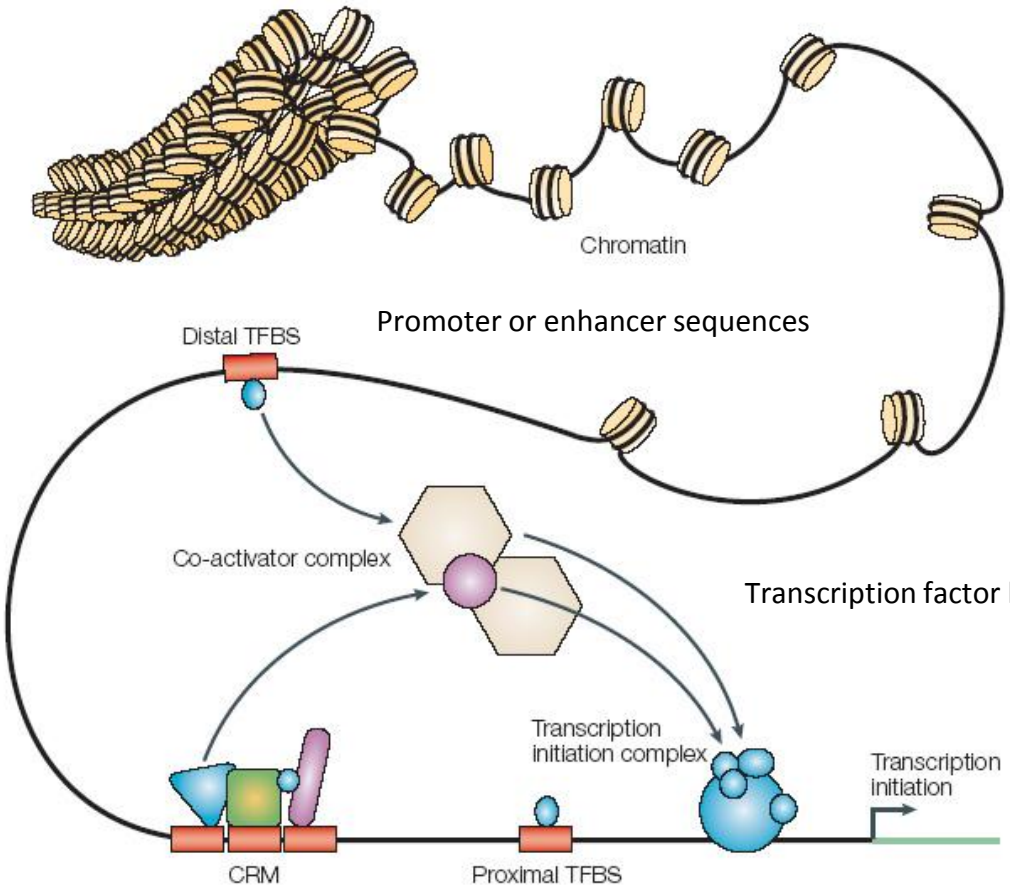
Human Genome



Human Epigenome



Cis-regulatory module (CRM)

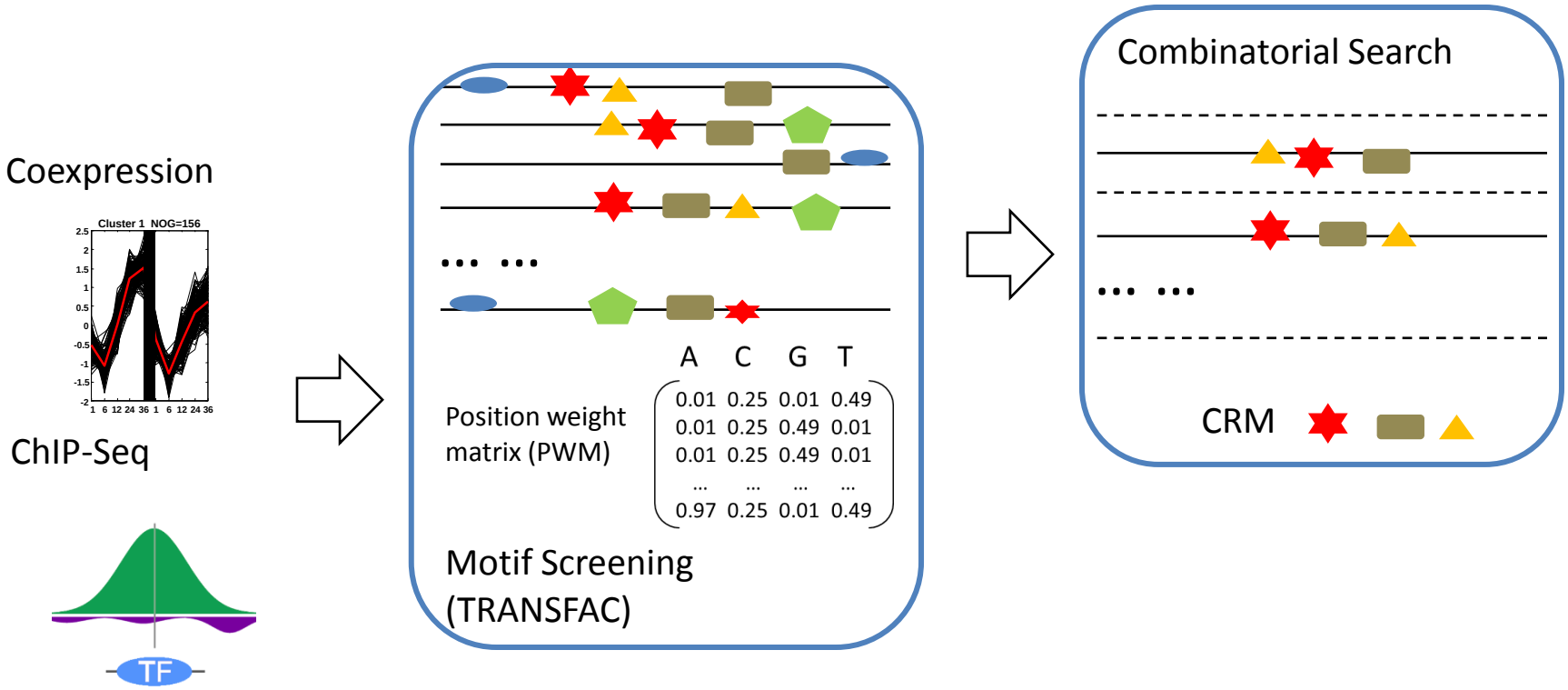


Problem statement

- CRM detection
  - Very hard combinatorial search problem
    - Given 50 motifs, there are 2^{50} combinations
  - Previous studies were restricted by computational limit
  - Problem of selecting correct coregulated sequences

We want to solve this problem using itemset mining

How is CRM detection performed?



To find the combination of motifs that occur more frequently in an input set than can be expected by chance

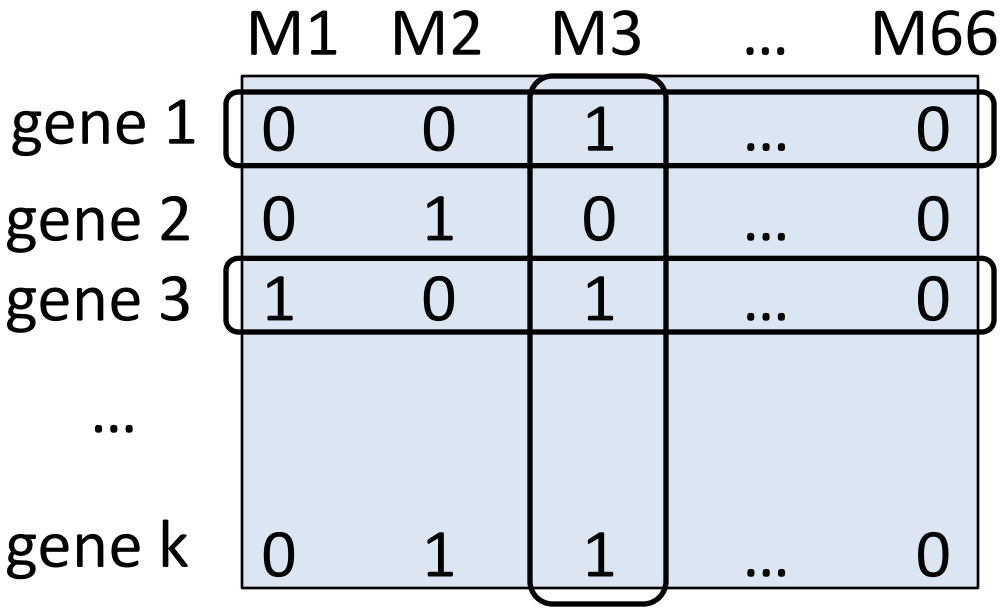
ModuleDigger

An itemset mining framework for the detection of *cis*-regulatory modules

Sun et al. 2019 *BMC Bioinformatics*

Input for ModuleDigger

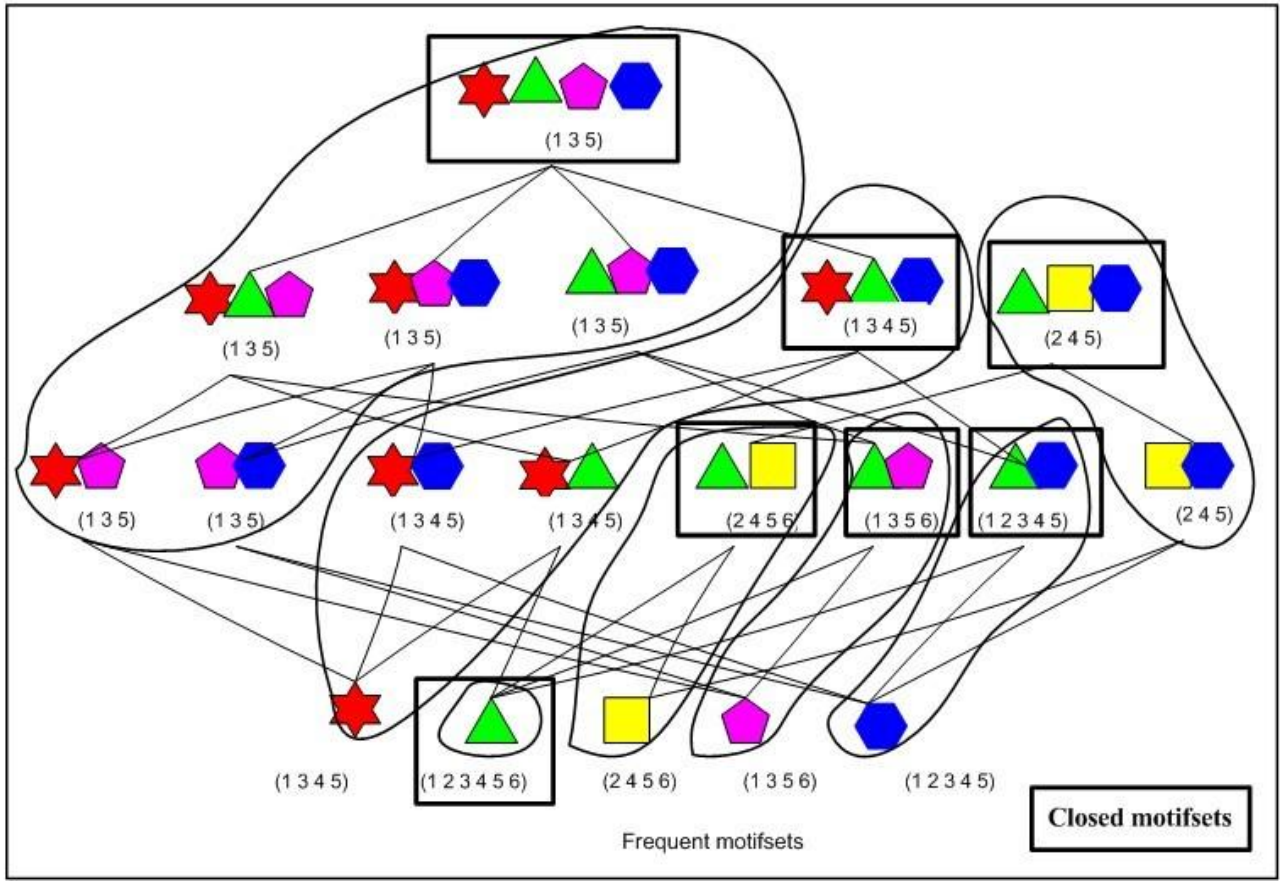
We only select the best scoring motif instance for each TF



Algorithm of ModuleDigger

Step2

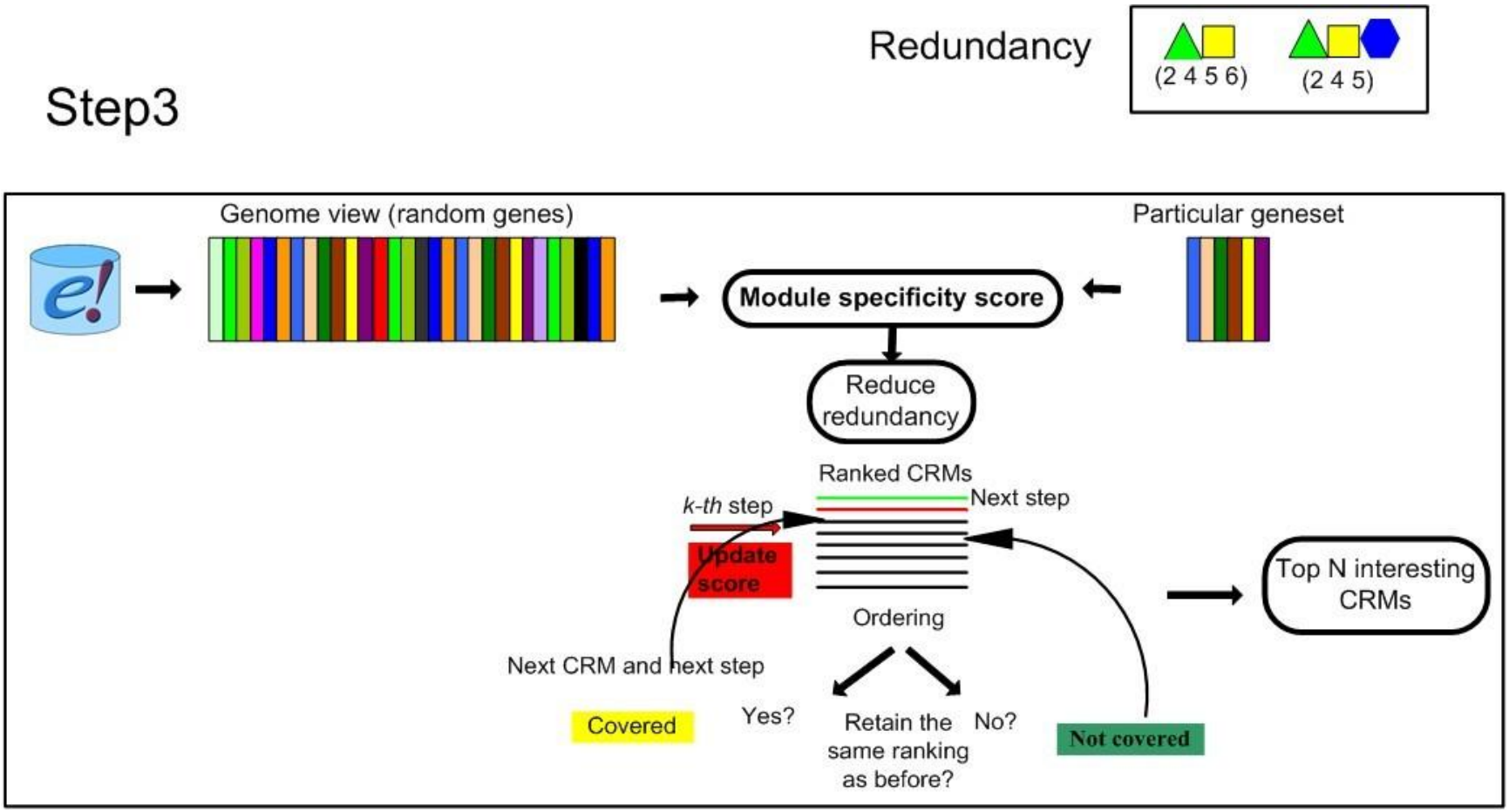
Frequency constraint
Length of motifset



Zaki & Hsiao et al. 2002 *SDM*

Algorithm of ModuleDigger

Step3



To select the biologically most interesting CRMs, we introduce a filter strategy

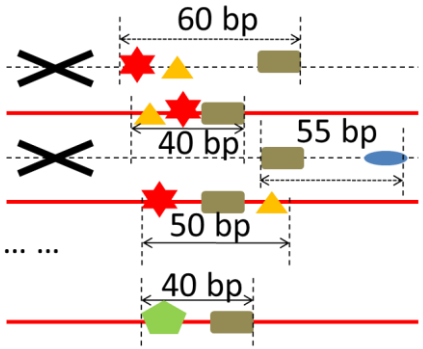
Gallo et al. 2007 *PKDD*

Limitations of ModuleDigger

- Difficulty on longer eukaryotes sequences
  - 1 motif instance per gene for a TF
- Although faster, but still computational limit
  - Because redundancy reducing wasn't addressed during the mining procedure

Solutions

- Add proximity constraint to allow for multiple instances
  - It's difficult to implement in classic itemset mining



- Incorporate redundancy reducing during the mining

CPModule

Cis-regulatory module detection based on constraint programming for itemset mining

Guns et al. 2010 *BIBM*
Sun et al. 2011 *In Revision Nucleic Acids Res*

Constraint programming (CP)

One of the most efficient general problem solving techniques

Model (by user)

Variables *V*
Domains *D(v)*
Constraints *C:* *defined on a set of variables*

Search (by solver)

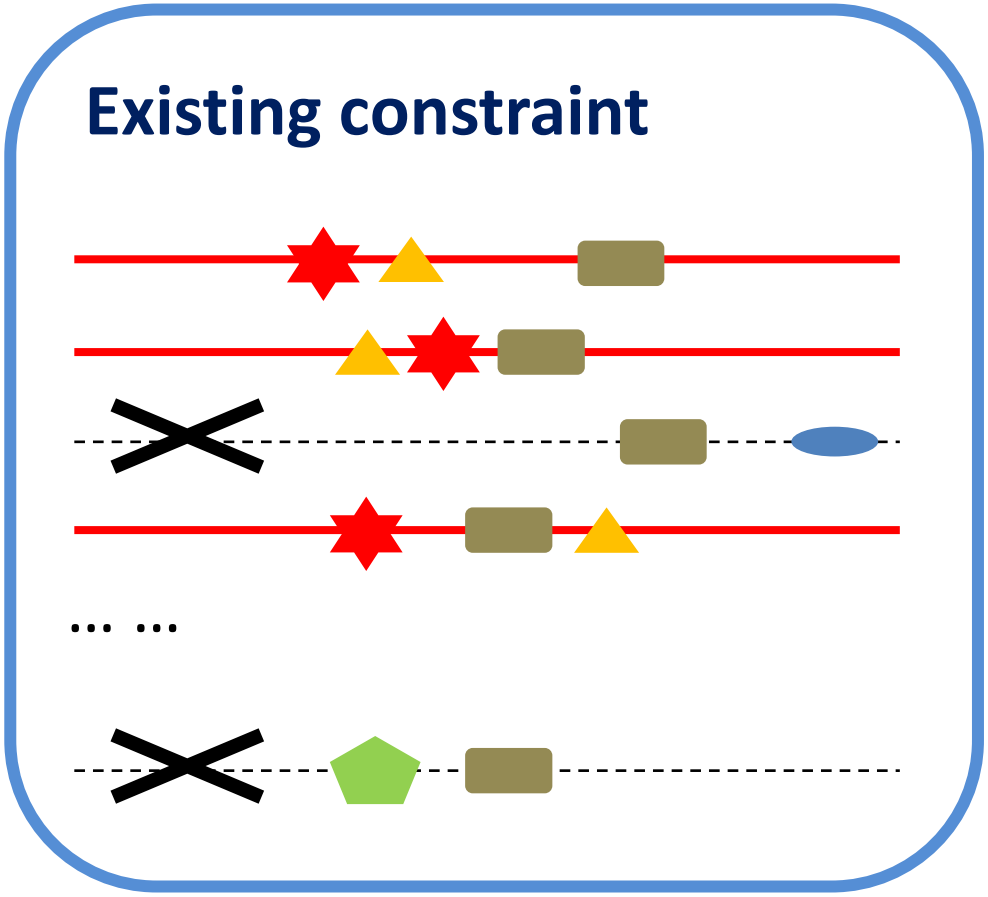
Propagation: *in which a constraint is used to remove values from the domain of variables that would violate it*
Branching: *in which a variable is assigned a value from its domain D(v)*

De Raedt et al. 2008 *KDD*

Model

- Variables
  - A Boolean variable m_i for every motif
    - Indicating whether this motif is part of the motifset
  - A Boolean variable s_j for every sequence
    - Indicating whether the motifset is a potential CRM in a sequence
  - A Boolean variable seqm_ij for every motif i and every sequence j
    - Indicating whether motif m_i is in the proximity of the motifs in motifset on sequence j
- Domains: {0, 1}
- Constraints: defined on the variables (explain latter)

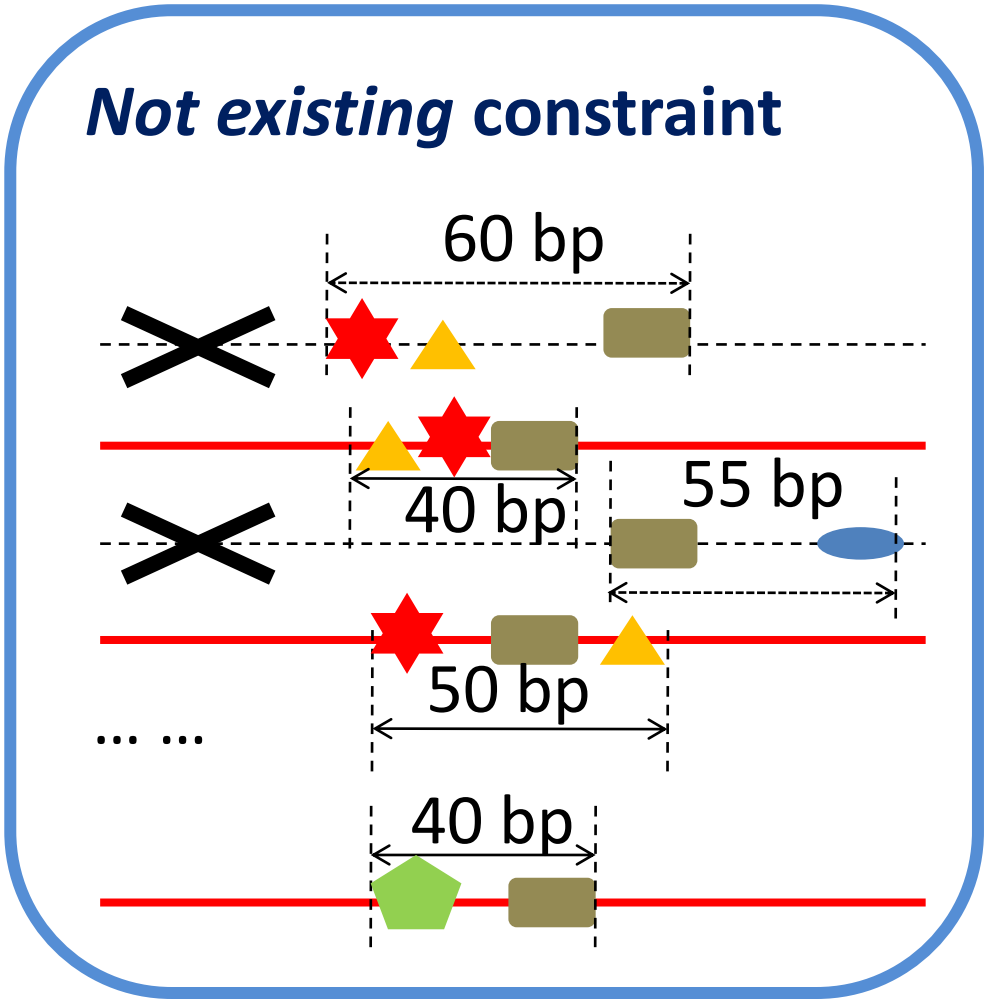
Frequency constraint



The CRM should occur in at least 2 sequences to be considered valid

Valid CRM

Proximity constraint

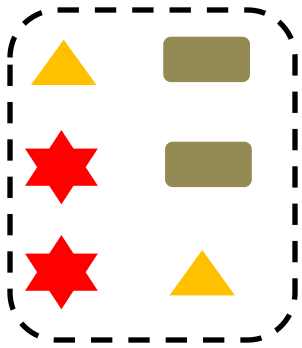


Only motif instances that occur in each others proximity can contribute to a valid CRM

Redundancy constraint

Existing constraint

Selected one



We found most of the solutions are redundant, and most occur in exact the same sequences.

The CRM with more motifs will be most specific for the sequence set and most statistically significant

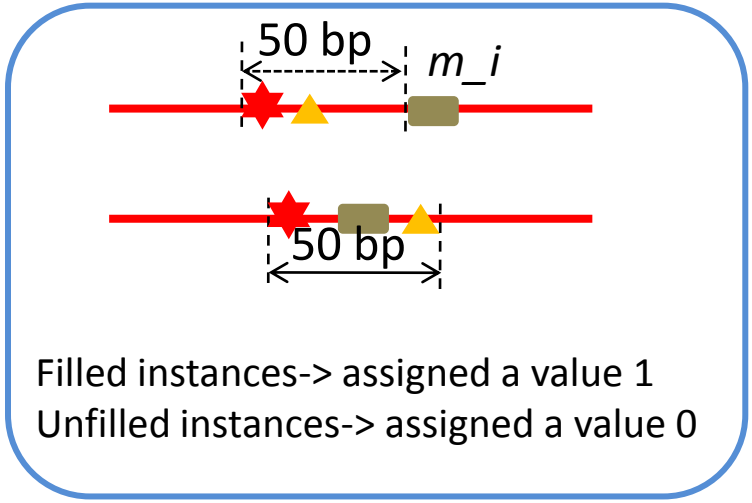
Removing redundant CRMs drastically decreases the computation time (closeness in itemset mining)

Propagation of the constraints

We focus on the propagation of the proximity constraint, we propagate changes of the motif variables to the seqm_ij variables.

Whether motif m_i is in the proximity of the motifs in motifset on sequence j?

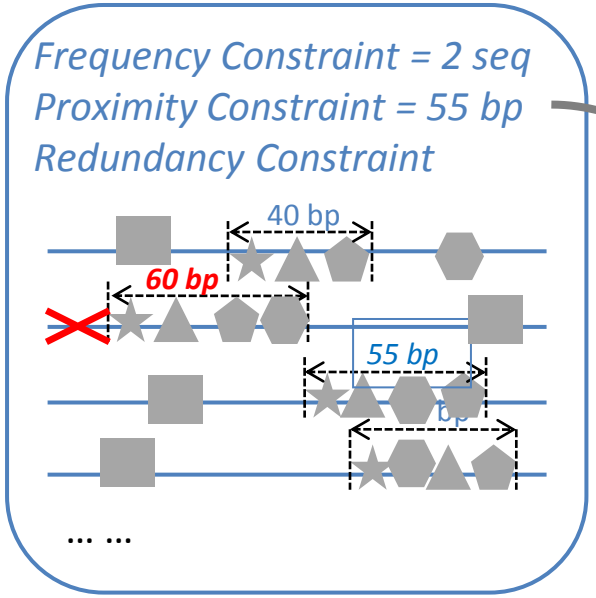
propagation

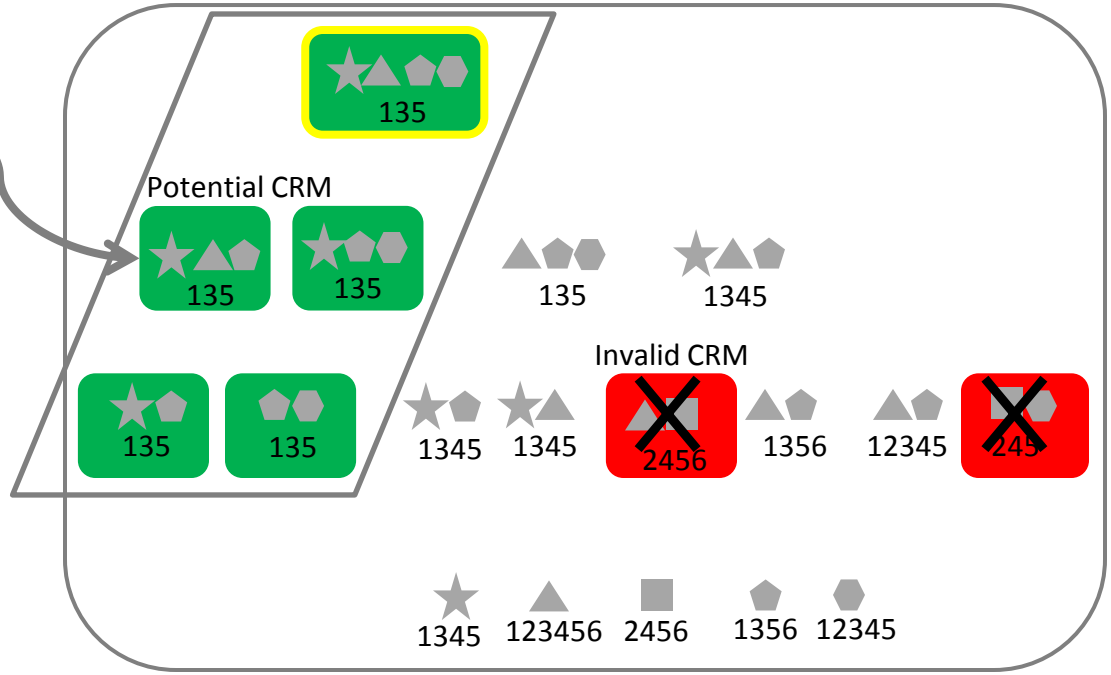


	m1	m2	m3	...	m66
seq 1	0	0	1	...	0
seq 2	0	1	0	...	0
seq 3	1	0	1	...	0
...					
seq k	0	1	1	...	0

Binary matrix dynamically updated

CPModule





Generic framework GECODE:
Generic Constraint Development Environment
http://www.gecode.org/

Guns et al. 2010 *BIBM*

Enrichment score calculation

- Similar to ModuleDigger
  - Because now the solutions we have are non-redundant, thus we can directly calculate the score for each solution and rank them accordingly

Benchmark on synthetic data

- To simply test the combinatorial search part
  - Xie et al. 2008 *Genome Res*
    - 22 sequences inserted with 3 motif instances
    - Maximal distance between the three TFBSs is 164 bp
    - Each sequence is 1000 bp in length
  - 516 TRANSFAC vertebrate PWMs
    - TFs from the same protein family are very similar
    - MotifComparison

Coessens et al. 2003 *Nucleic Acids Res*

Correlation coefficient

- Motif level and nucleotide level CC
- Based on the data
  - TP: motif/nucleotide was predicted and it is part of the CRM
  - FP: motif/nucleotide was predicted but it isn't part of the CRM
  - FN: motif/nucleotide wasn't predicted and it's part of the CRM
  - TN: motif/nucleotide wasn't predicted and it isn't part of the CRM

CC lies in the range of -1 to +1. +1 indicates the prediction corresponds to the correct answer

$$CC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}}$$

Klepper et al. 2008 *BMC Bioinformatics*

Effect of proximity on performance

The binding sites were at most 164 bp from each other

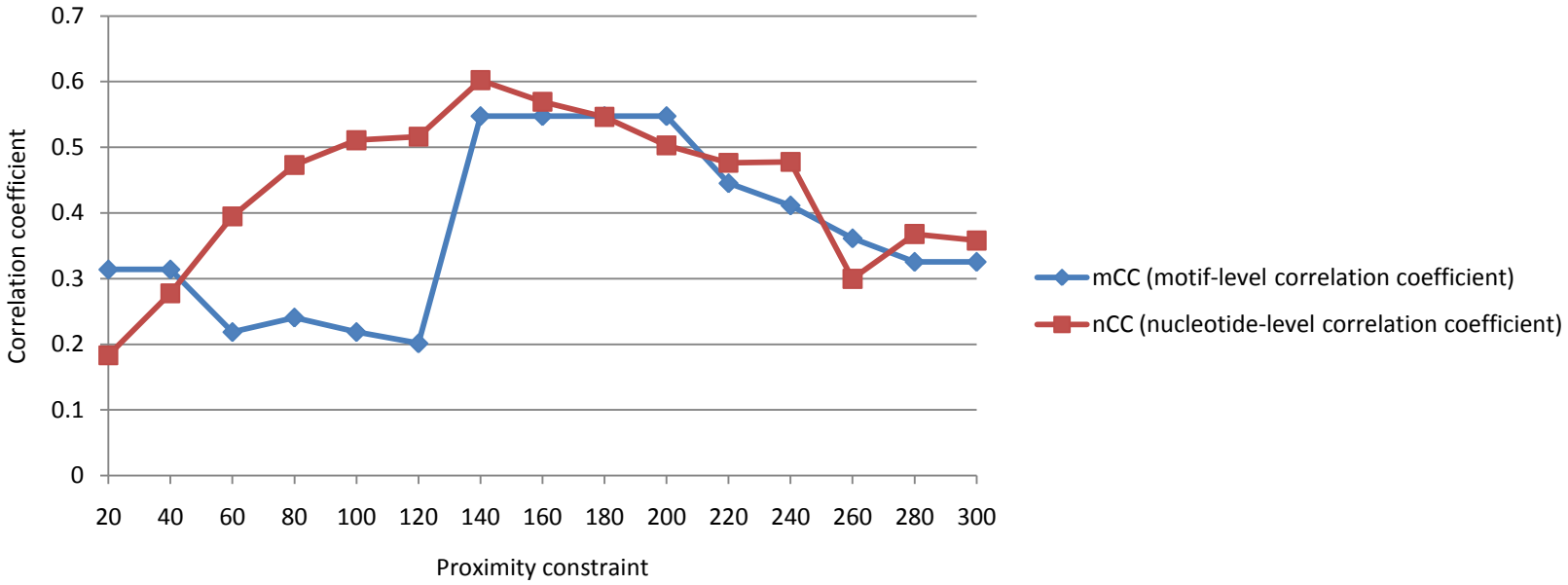


Figure 1: Effect of the proximity constraint on the quality of the results.

Benchmark with well performing tools

- Benchmark with Cister, Cluster-Buster, ModuleSearcher and Compo
  - The performance was assessed by comparing the best scoring solution of each method with the known true solutions

Table 1: Comparison of CRM prediction algorithms

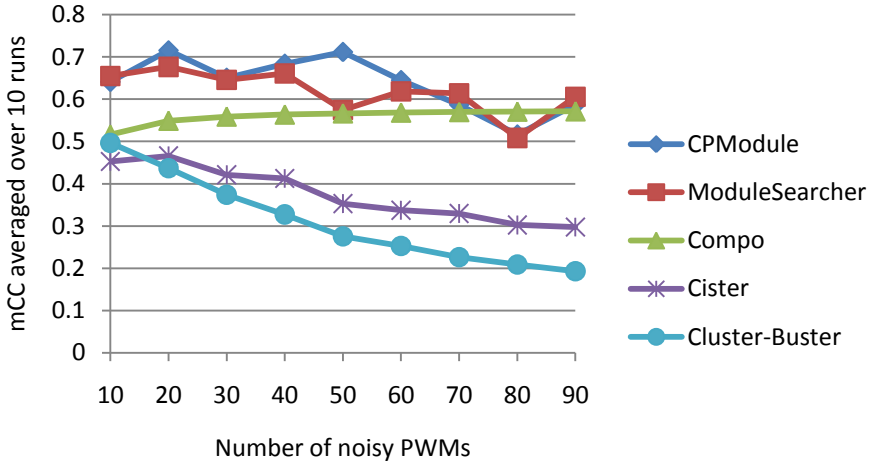
	Cister	Cluster-Buster	ModuleSearcher	Compo	CPModule
mCC	0.16	0.05	/	-	0.57
nCC	0.23	0.23	/	-	0.55

"/" indicates termination by lack of memory, "-" indicates the algorithm was still running after 2 days.

Run with increasing number of PWMs

A

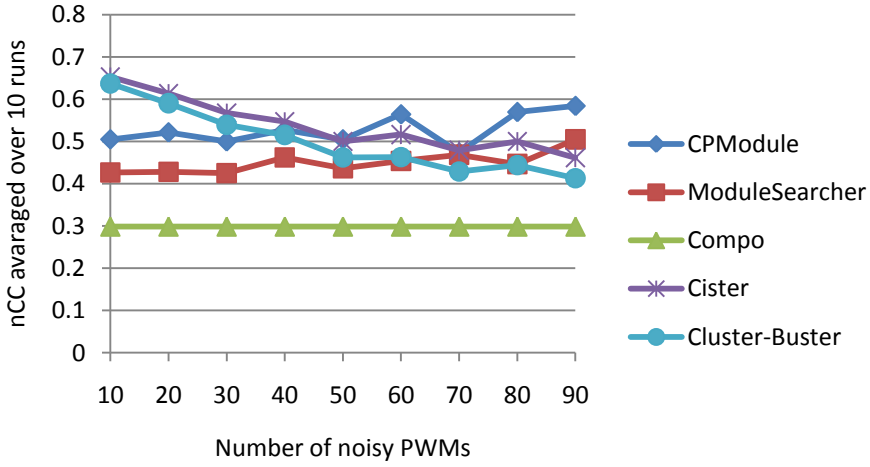
Figure 2: Effect of noise on the motif and nucleotide correlation



PWMs set = 3 inserted PWMs + several noisy PWMs randomly sampled from the remaining PWMs

Each set sampled 10 times

B



CPModule has performances similar to state of the art algorithms on a synthetic dataset BUT is able to deal with much larger sequence sets

Application

Unveiling combinatorial regulation of mouse embryonic stem cells through the combination of ChIP information and *in silico* CRM detection

Sun et al. 2011 *In Revision*

Motivations

- Combine CRM detection with ChIP-Seq
  - Largely reduce the regions of the binding sites of the assayed TF
  - Search for CRMs that at least contain one TFBS for the assayed TF (Query-based way)

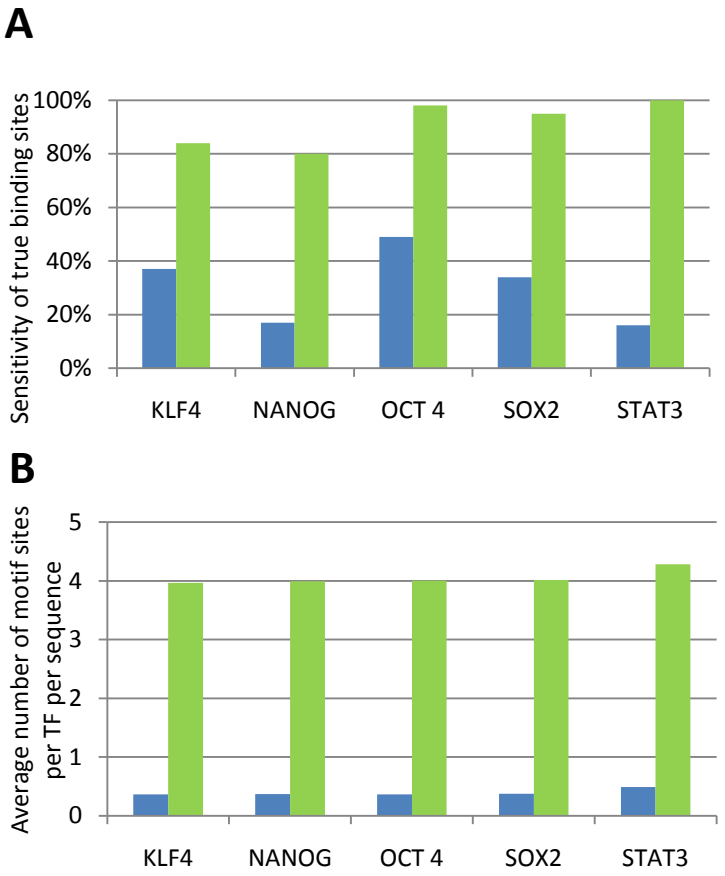
Five ChIP-Seq dataset and input

- Dataset
  - ChIP-Seq binding peaks for 5 key transcription factors: KLF4, NANOG, OCT4, SOX2 and STAT3 involved in self-renewal of mouse embryonic stem cells (ESCs)

- Input
  - Sequences
    - Top 100 binding peaks for each of the assayed TF
    - 500 bp sequence centered around each of the top 100 binding peaks
  - 517 PWMs
    - 516 TRANSFAC PWMs
    - 1 KLF4 PWM from literature

Chen et al. 2008 *Cell*
Won et al. 2010 *Genome Biology*
Wilbanks et al. 2010 *PLoS ONE*
Whitington et al. 2009 *Nucleic Acids Res*
Jiang et al. 2008 *Nature Biotechnology*

Motif screening



Assessing the effect of screening on

- Sensitivity: number of true sites recovered
- Precision: number of false positives in the screening

Use ChIP-Seq data as benchmark

Assess sensitivity: each of the sequences located around the top 100 binding peaks should contain a binding site for the assayed TF

Assess precision: the average number of binding sites per TF other than the assayed one should be restricted (on average <1 as not all TF should have a site)

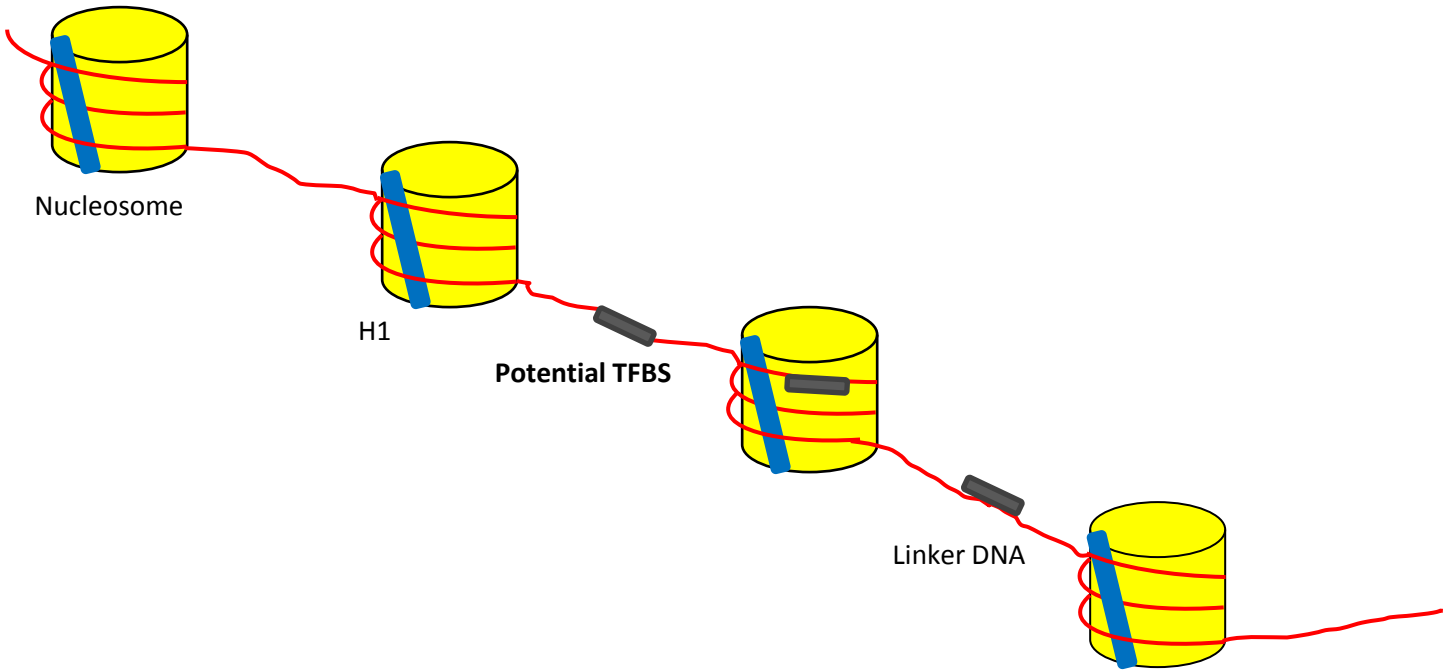
Figure 3: A stringent screening and a non-stringent screening.

Motivation for epigenetic filtering

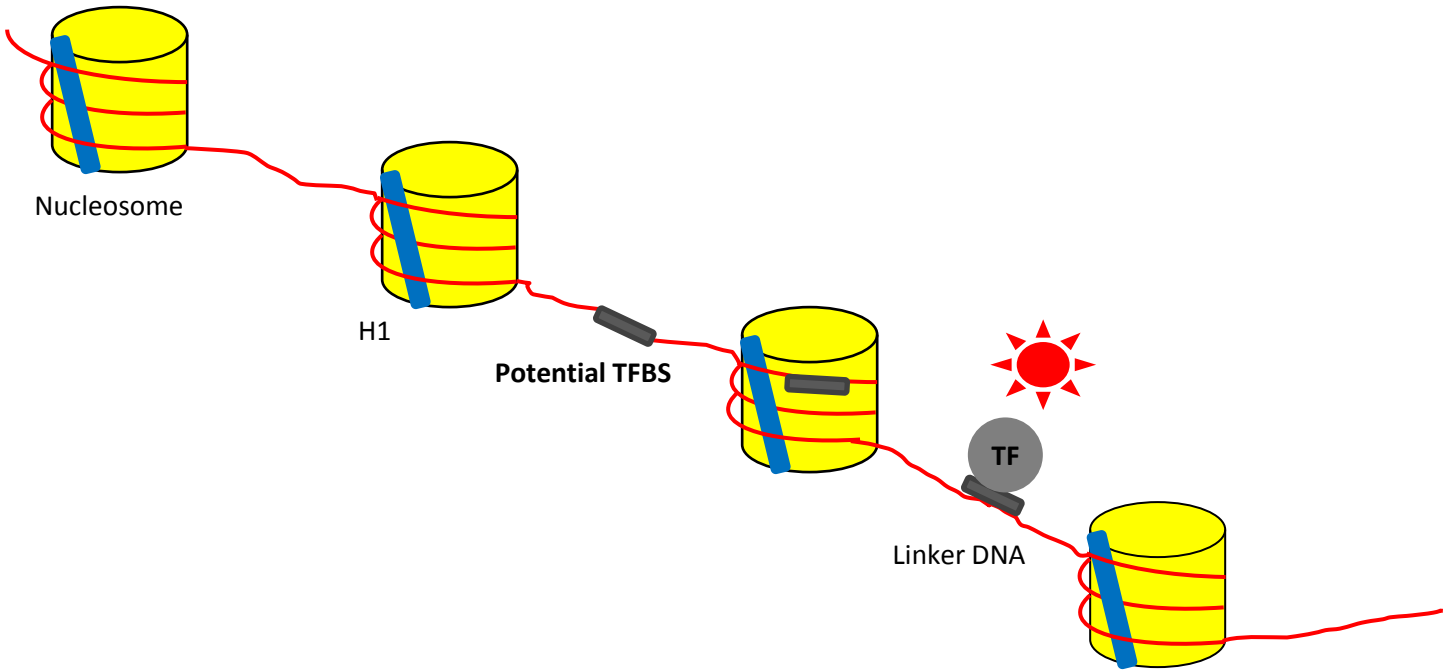
- Recent study shows
  - A ***lower binding specificity*** but a ***stable chromatin stability*** can also lead to the binding of TF

Whitington et al. 2009 *Nucleic Acid Research*
Li et al. 2011 *Genome Biology*

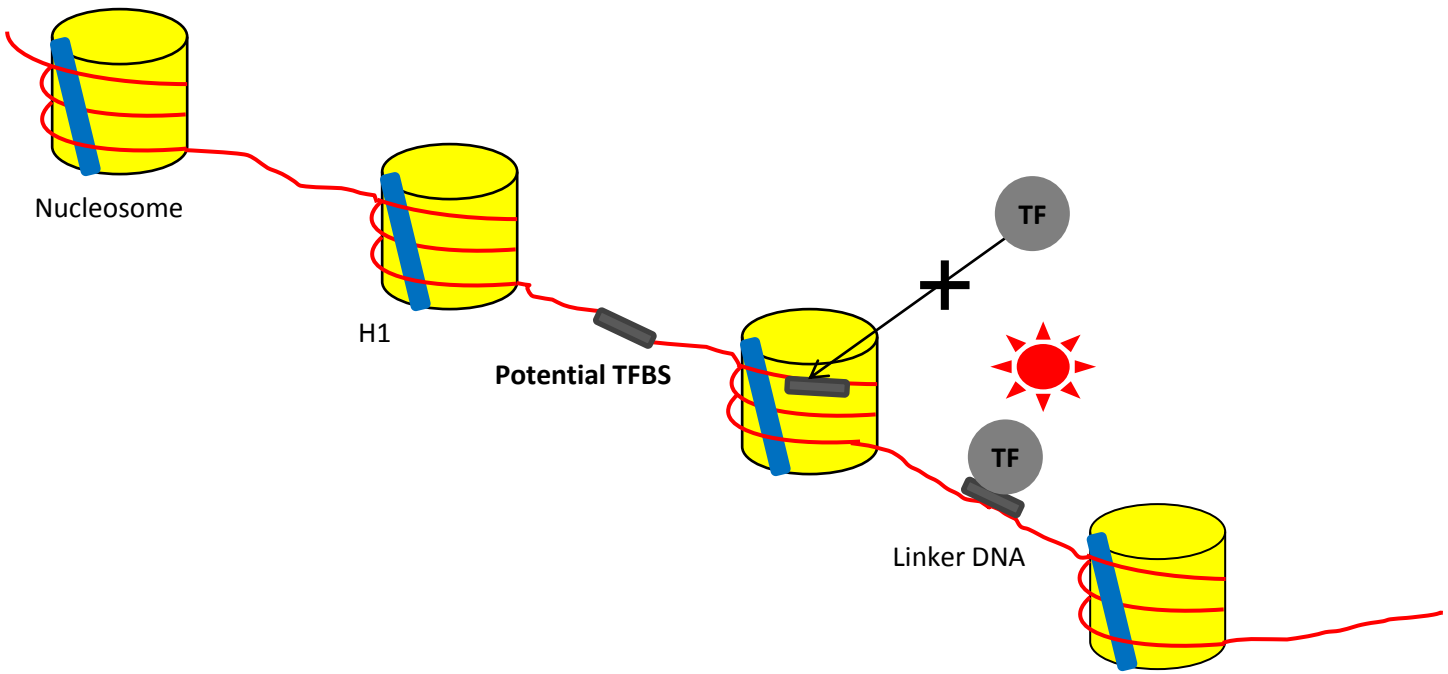
Motivation for epigenetic filtering for embryonic stem cells



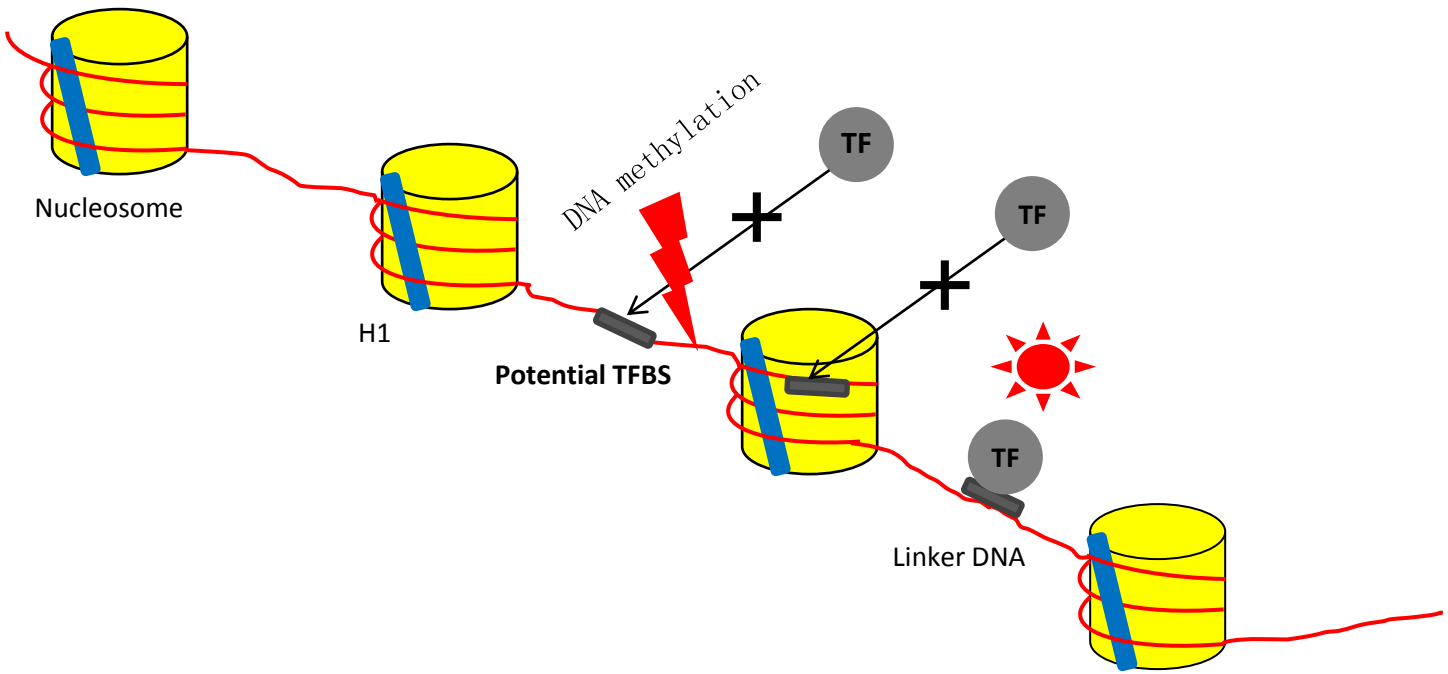
Motivation for epigenetic filtering for embryonic stem cells



Motivation for epigenetic filtering for embryonic stem cells



Motivation for epigenetic filtering for embryonic stem cells



Epigenetic scores

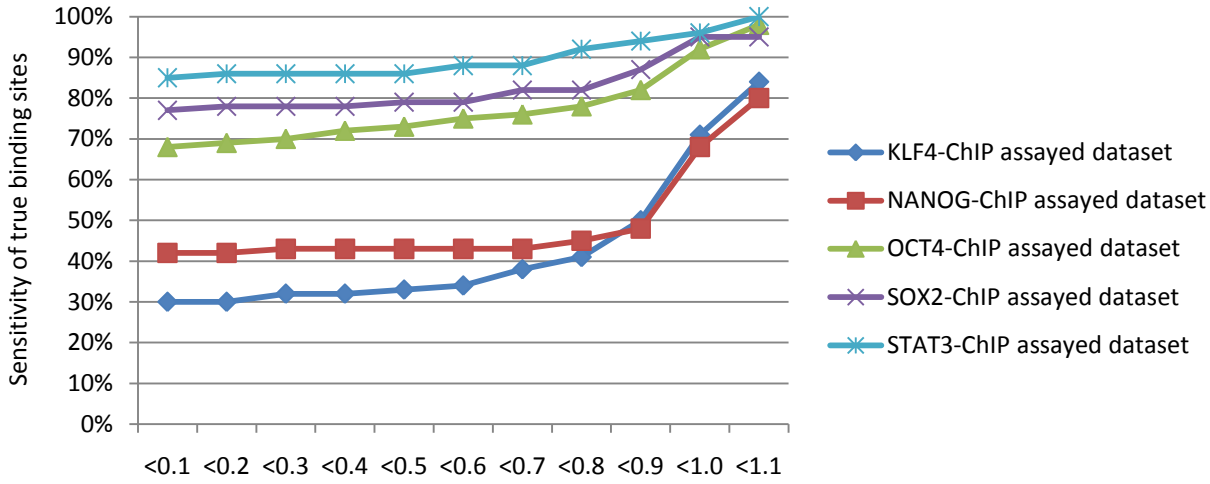
For each potential TFBS (except for the assayed TF), we calculate

- Nucleosome occupancy
  - Nucleosome occupancy score
    - First assign a probability to each base pair position of the TFBS using a prediction model
    - The nucleosome occupancy score for TFBS was calculated as the geometric mean of the probabilities at all positions of the potential TFBS
- DNA methylation level
  - GC dinucleotide score
    - The fraction of GC dinucleotides within a window of 50 bp centered around the TFBS
  - GC content score
    - The fraction of G or Cs within a window of 50 bp centered around the TFBS

Ernst et al. 2010 *Genome Research*
Ramsey et al. 2010 *Bioinformatics*
Xi et al. 2010 *BMC Bioinformatics*

Choosing thresholds for epigenetic filtering

A



(1.7 – 1.2) x 100sequences x 517pwms = 258500TFBSs

B

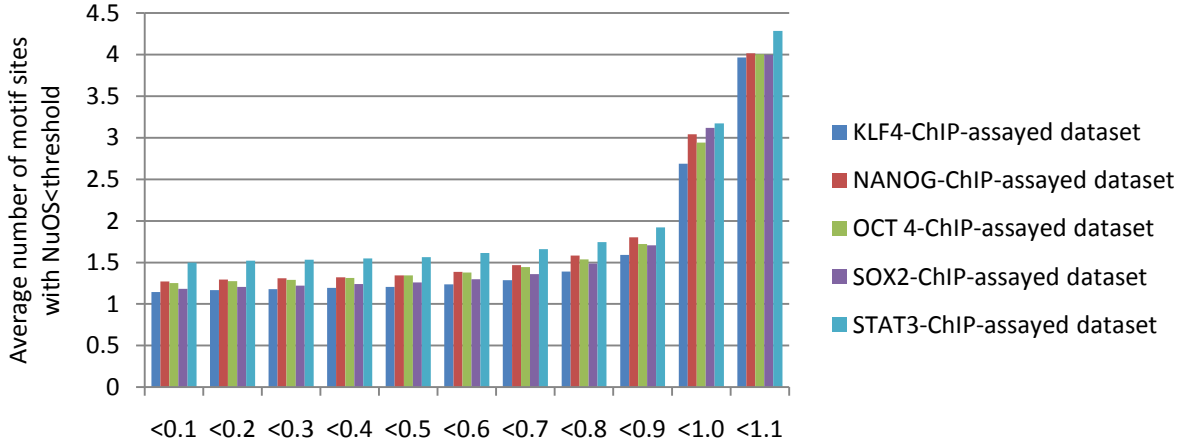
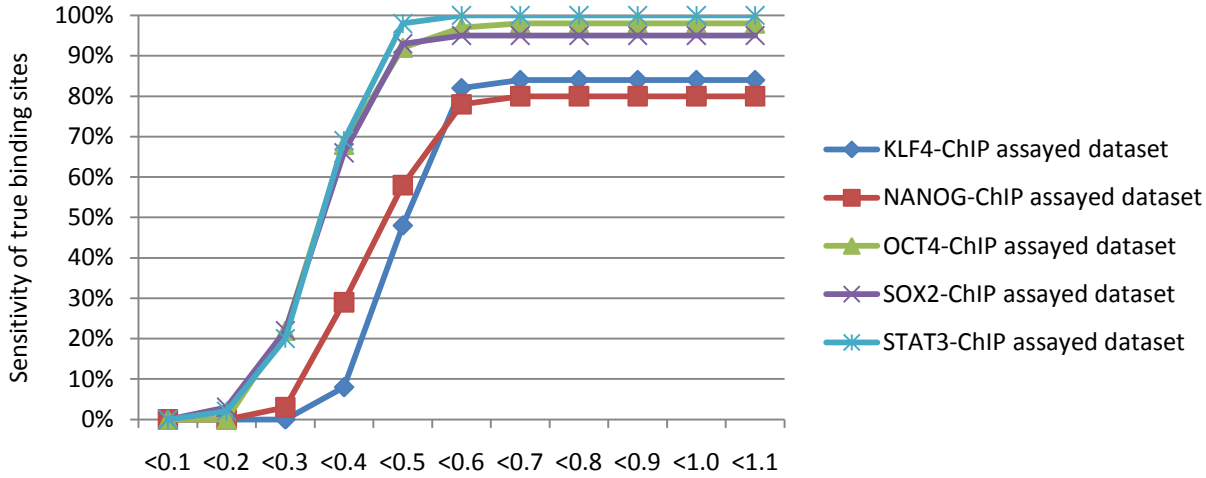


Figure 4: Effect of the nucleosome occupancy score filtering thresholds on motif screening results.

Choosing thresholds for epigenetic filtering

A



B

Figure 5: Effect of the GC content filtering thresholds on motif screening results.

Parameter settings for CPModule

- Parameter settings
  - Start our analysis by setting 150 bp as the proximity constraint, and gradually step wisely increasing 50 bp till 400 bp
  - Frequency constraint was set at 60%

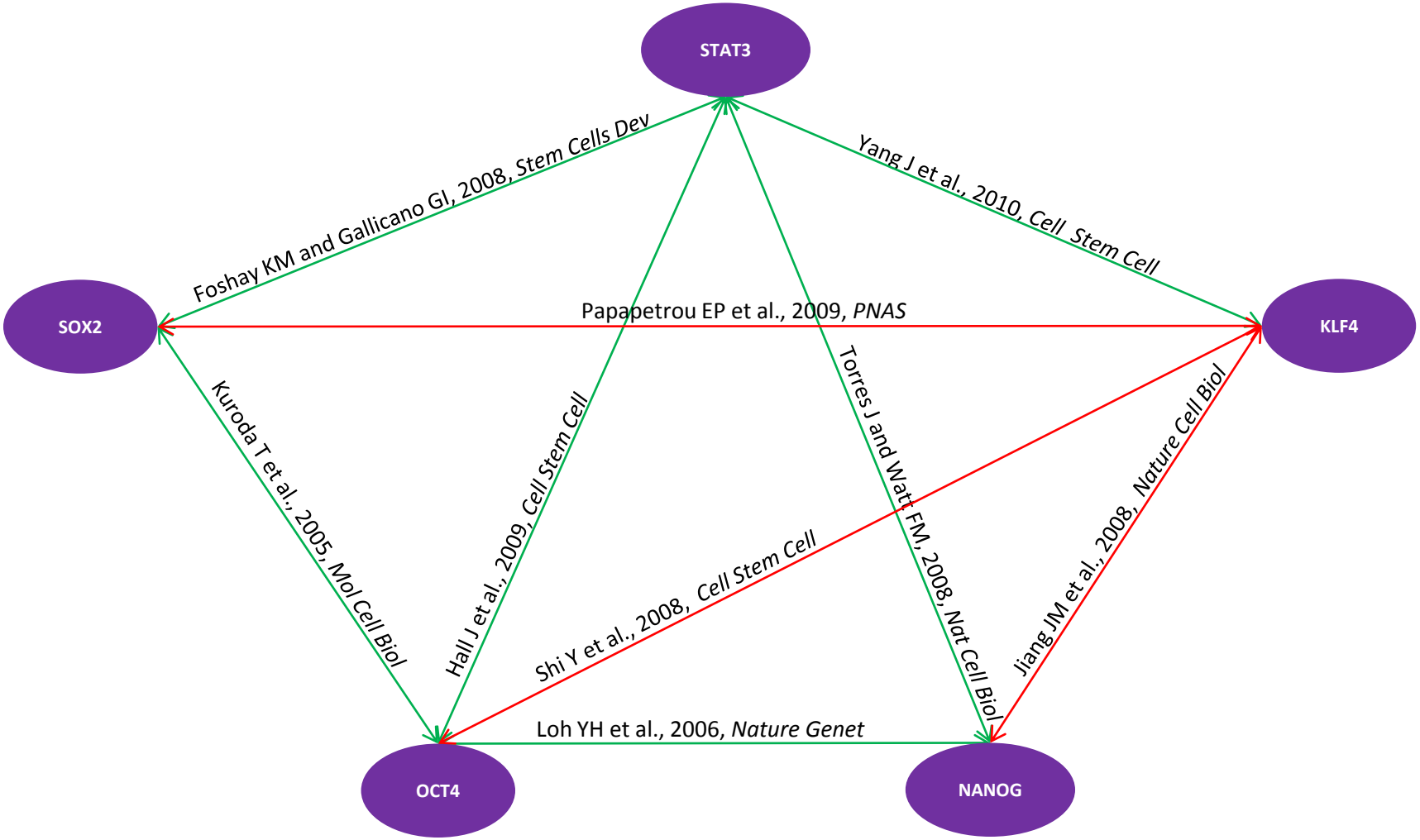
- CRM selection
  - Only the first detected CRM that consists of the assayed TF will be selected

Detected CRMs in mouse embryonic stem cell

Table 2: CRMs obtained with CPModule with non-stringent screening and filtering for non-assayed TFs

ChIP-Seq-assayed TF	CRM	Rank		Frequency constraint	Cross validation	Proximity constraint (bp)	Number of CRMs	
		All	Query-based				All	Query-based
KLF4	**KLF4**, STAT4	143	2	60%	40.00%	300	147	3
NANOG	**NANOG**, OCT1	6846	4	61%	70.49%	300	6868	17
	NANOG, STAT3	14017	10	60%	25.00%	350	14033	26
OCT4	**OCT4**, STAT1, [XFD2, STAT4, STAT6]	5	5	63%	11.10%	150	5068	613
SOX2	**SOX2**, OCT4	430	1	63%	79.40%	150	14180	22
	SOX2, STAT3, [CDXA, PAX2, STAT5A]	61807	24	60%	23.33%	250	117006	189
STAT3	**STAT3**, OCT4, [STAT1, STAT5A, STAT6]	1	1	61%	24.59%	150	1366	20

Detected interactions between the 5 TFs in mouse embryonic stem cell



Top 3 ranked CRMs

Table 3: Top three ranked CRMs for each of the five assayed TFs

Assayed TF	Proximity constraint (bp)	Rank		Frequency constraint	CRM	Reference
		All	Query-based			
KLF4	300	84	1	60%	KLF4, TBP	More details see below
		143	2	60%	KLF4, STAT4	(Bourillot&Savatier 2010)
		147	3	61%	KLF4, CAP	Not available or no comment
NANOG	300	6465	1	62%	NANOG, TTF1	More details see below
		6827	2	61%	NANOG, BRCA	More details see below
		6828	3	63%	NANOG, FAC1	More details see below
	350	13615	1	61%	NANOG, HOXA3	More details see below
		13863	2	63%	NANOG, TTF1	More details see below
		14002	3	60%	NANOG, HELIOSA	More details see below
OCT4	150	1	1	60%	OCT4, HMGIY, ELF1, XFD2	(John et al.,1995;Leger et al.,1995)
		2	2	63%	OCT4, HMGIY, ELF1, CDXA	(John et al.,1995;Leger et al.,1995)
		3	3	60%	OCT4, PAX2, HMGIY, ELF1, CDXA	(John et al.,1995;Sun et al.,2008;Gupta et al., 2006)
SOX2	150	430	1	63%	SOX2, OCT4	(Kuroda et al.,2005)
		2821	2	60%	SOX2, CDXA, AR	Not available or no comment
		4662	3	62%	SOX2, CDXA, CAP1	Not available or no comment
	250	2514	1	60%	SOX2, OCT4, CDXA, LEF1	(Kurado et al.,2005)
		4026	2	60%	SOX2, OCT4, PAX2, LEF1	(Kurado et al.,2005)
		5561	3	60%	SOX2, OCT4, PAX2, SRY	(Kurado et al.,2005)
STAT3	150	1	1	61%	STAT3, OCT4, STAT1, STAT5A, STAT6	(Hall et al., 2005)
		2	2	61%	STAT3, OCT4, STAT6, STAT5A	(Hall et al.,2005)
		3	3	60%	STAT3, OCT4, STAT5A, STAT6	(Hall et al.,2005)

Literature support for the involvement of the retrieved TFs in functions related to ESC biology

- Human **TBP** protein increases anchorage-independent growth of cells (Johnson et al., 2003).
- **STAT4** activation is involved in differentiation of type 1 helper T cells (Farrar et al., 2000).
- **CAP1** has a role in apoptosis (Wang et al., 2008).
- **TTF1** is involved in lung morphogenesis (Hosgor et al., 2002).
- **HELIOS** is expressed in the earliest hematopoietic sites of the embryo (Kelley et al., 1998).
- **HMGA1** affects embryonic stem cell lymphohematopoietic differentiation (Battista et al., 2003).
- **FOXI1** genetic and biochemical data suggest a central role in embryonic development for genes encoding forkhead proteins (Pierrou et al., 1994).
- **ELF1** plays an important and non-redundant role in the development and function of NKT cells (Choi et al., 2010). Homozygous knockout of **ELF** in mice affects development of heart, brain, liver and gastrointestinal tract (Tang et al., 2003).
- **CDX1** is involved in axial patterning and intestinal cell differentiation (Beck et al., 2010 ,Park et al., 2009).
- **AR** is required for male embryonic sexual differentiation (Holdcraft et al., 2004).
- **LEF1** regulates lineage differentiation of multipotent stem cells in skin (Merrill et al., 2001). Mouse **LEF1** is involved in differentiation of paraxial mesoderm and morphogenesis of embryonic limb (Calceran et al., 1999).
- **PAX2** is involved in nephric lineage specification (Bourchard et al., 2002) and urogenital development (Torres et al., 1995).
- **SRY** is the master switch in mammalian sex determination (Kashimada&Koopman 2010).
- The JAK1-**STAT1**-**STAT3** pathway promotes proliferation and prevents premature differentiation of myoblasts (Sun et al., 2007).
- **STAT5** is required for embryonic thymocyte production, TCRgamma gene transcription, and Peyer's patch development (Kang et al., 2004). **STAT5** promotes multilineage hematolymphoid development in vivo through effects on early hematopoietic progenitor cells (Snow et al., 2002).
- **STAT6** protein is necessary for development of T-helper cell (Wurster et al., 2000).

Literature support for the CRMs

-- We assume some of the interactions might exist based on the properties of individual TFs

- **KLF4-TBP:** We could not find direct literature support for the interaction between KLF4 and TBP, but TBP is a general TATA box-binding protein (Bertolino&Singh 2002), making the interaction is plausible.
- **NANOG-TTF1:** Recent studies in mouse models have demonstrated that SOX2 regulates airway epithelium differentiation and that SOX2 and thyroid transcription factor TTF1 are modulated in concert during the course of tracheal and esophageal development (Que et al., 2007). As NANOG at least during embryonic stem cell development belongs to the same regulatory network as SOX2, the interaction of NANOG with TFs that are also interaction partners of SOX2 is possible.
- **NANOG-BRCA:** Roles of BRCA in both homologous recombination and nonhomologous end joining DNA repair have been shown (Shafee et al., 2008; Farmer et al., 2005). Such function of BRCA might also play a role during the self-renewal process to repair DNA damage.
- **NANOG-FAC1:** The putative transcriptional regulator FAC1 is expressed in embryonic and extraembryonic tissues of the early mouse conceptus. Study showed FAC1 is essential for trophoblast differentiation during early mouse development (Goller et al., 2008). Thus there might be an interaction between NANOG and FAC1.
- **NANOG-HOXA3:** As we known, HOXA3 is involved in wound repair (Mace et al., 2009), so it might interact with NANOG in the self-renewal process.
- **SOX2-CDXA:** Binding of homeobox domain from **CDX1** protein and **SOX2** protein was shown to occur in a system of purified components (Beland et al., 2004). Although we identified a module with cdxA, cdxA and cdx1 belong to the same family and have very similar motif models.
- **STAT3, STAT6, STAT1:** Binding of human **STAT3** protein and human **STAT6** protein occurs (2-hybrid assay) (Ravasi et al., 2010). **STAT1** and **STAT3** can form heterodimers (John et al.,1995; Levy&Darneel 2002). Note however that with the STAT motif models it is difficult to make the distinction between the different STAT members.

Ingenuity pathway analysis of the retrieved TFs

For the 21 transcription factors in the list of predicted CRMs that could be mapped to Ingenuity Pathways (AR, BPTF, BRCA1, CAP1, CDX1, ELF1, FOXI1, HMGA1, HOXA3, IKZF2, LEF1, PAX2, SPTBN1, SRY, STAT1, STAT4, STAT6, STAT5A, TBP, TTF1, LEF1), we searched for known functions involving at least half of the TFs in the set

Cellular Growth and Proliferation (AR, BRCA1, CDX1, ELF1, HMGA1, HOXA3, IKZF2, LEF1, PAX2, SPTBN1, SRY, STAT1, STAT4, STAT6, STAT5A)
Cell Death (AR, BPTF, BRCA1, CDX1, FOXI1, HMGA1, HOXA3, LEF1, PAX2, STAT1, STAT4, STAT6, STAT5A, TBP, TTF1)
Cancer (AR, BRCA1, CDX1, HMGA1, HOXA3, IKZF2, LEF1, PAX2, SPTBN1, STAT1, STAT6, STAT5A)
Cellular Development (AR, BRCA1, CDX1, HMGA1, LEF1, PAX2, SRY, STAT1, STAT4, STAT6, STAT5A)
Tissue Morphology (AR, BRCA1, CDX1, FOXI1, HOXA3, LEF1, STAT1, STAT4, STAT6, STAT5A)

http://www.ingenuity.com/

Conclusions

- Powerful combinatorial search strategy based on constraint programming for itemset mining
  - Handle much larger dataset comparing with previous tools
  - Open up new application of combining CRM detection with ChIP data
    - Unveil the combinatorial regulation
    - Help biologists with the design of consequent ChIP experiment

- Potential applications
  - Combination of chromatin marks
    - Chromatin marks may act cooperatively to prepare chromatin for transcriptional activation (Zhang et al. 2008, *Nature Genetics*, Ernst&Kellis 2010, *Nature Biotechnology*)
  - Combination of single nucleotide polymorphisms (SNPs)
    - Coulet et al. 2007, *BMC Bioinformatics*
  - … …

Acknowledgements

- Promoters:
  - Prof. Dr. Bart De Moor (ESAT)
  - Prof. Dr. Kathleen Marchal (CMPG)
- CMPG-BIOI
  - Dr. Pieter Monsieurs*
  - Dr. Karen Lemmens*
  - Dr. Inge Thijs*
  - Dr. Riet De Smet*
  - Dr. Valerie Storms*
  - CMPG-BIOIers
- ESAT/SCD-BIOI
  - Prof. Dr. Yves Moreau
  - ESAT-BIOIers
- Engineering Mathematics, University of Bristol, UK
  - Prof. Dr. Tijl De Bie
- Computer Science
  - Prof. Dr. Luc De Raedt
  - Tias Guns
  - Dr. Siegfried Nijssen
- CMPG
  - Prof. Jozef Vanderleyden
- UZ Hospital-Experimental Medicine
  - Prof. Dr. Annemieke Verstuyf
- UZ Hospital-Interdepartmental Stem Cell Institute
  - Dr. Lieven Thorrez
- CMPG&VIB
  - Prof. Dr. Kevin Verstrepen
  - Elham Aslankoohi
- Universitair Ziekenhuis Antwerpen & Universiteit Antwerpen
  - Dr. Tim Van den Bulcke*

How genetic and epigenetic work …

Thank you for your attention~