Discriminative de novo motif discovery from high-throughput data

Jan Grau[1], Stefan Posch[1], Ivo Grosse[1], and Jens Keilwagen[2]

[1]Institute of Computer Science, Martin Luther University Halle–Wittenberg, Germany
[2]Julius Kühn-Institut (JKI) - Federal Research Centre for Cultivated Plants, Quedlinburg, Germany

CSC ChIP- and DNase-seq data analysis workshop

Transcriptional regulation by transcription factors

Biological question

- reasons for phenotypic observations
- regulation of gene expression
- first step: transcriptional regulation

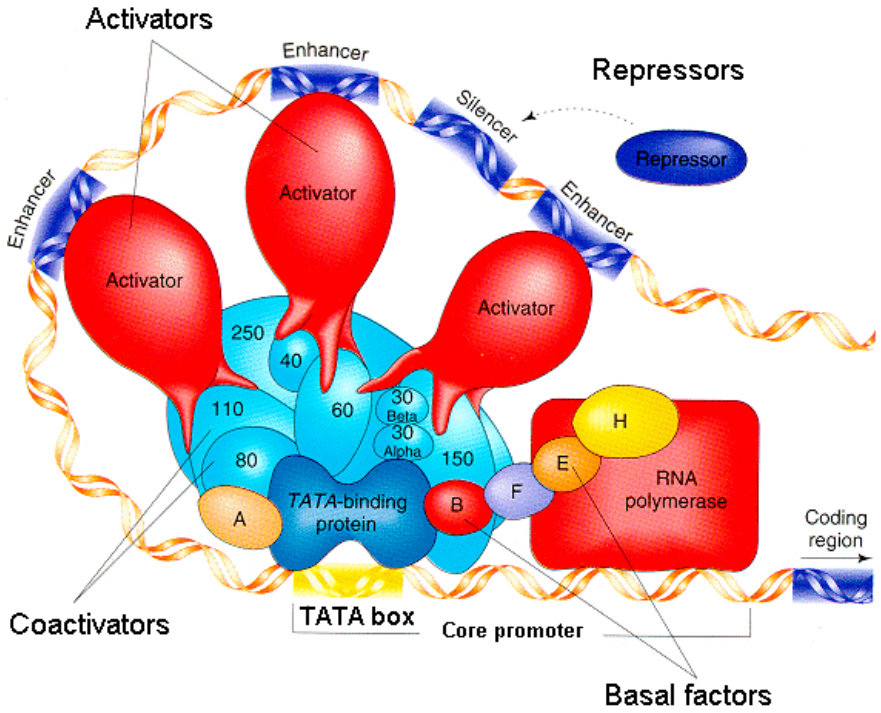
$\Rightarrow$ transcription factor binding sites

De-novo motif discovery

without knowledge of

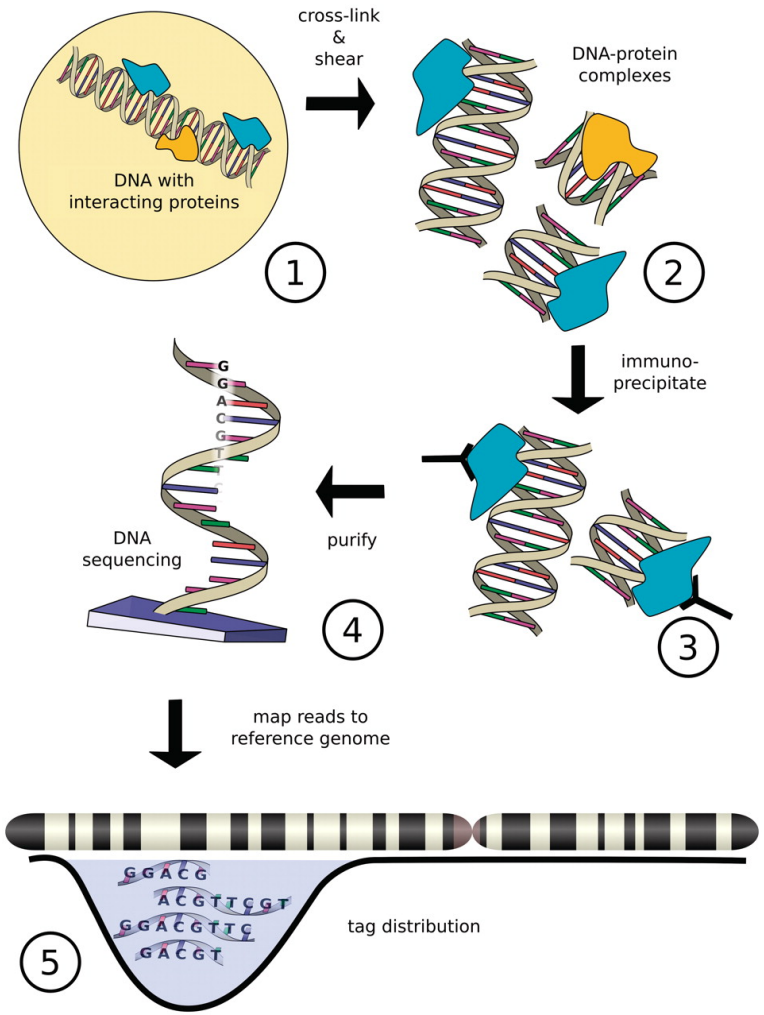
- motif
- extact location of sites

from set of input sequences



[Based on Robert Tjian, "Molecular Machines that Control Genes"]

Experimental techniques - ChIP-seq



[Szalkowski & Schmid, Brief Bioinform, 2010]

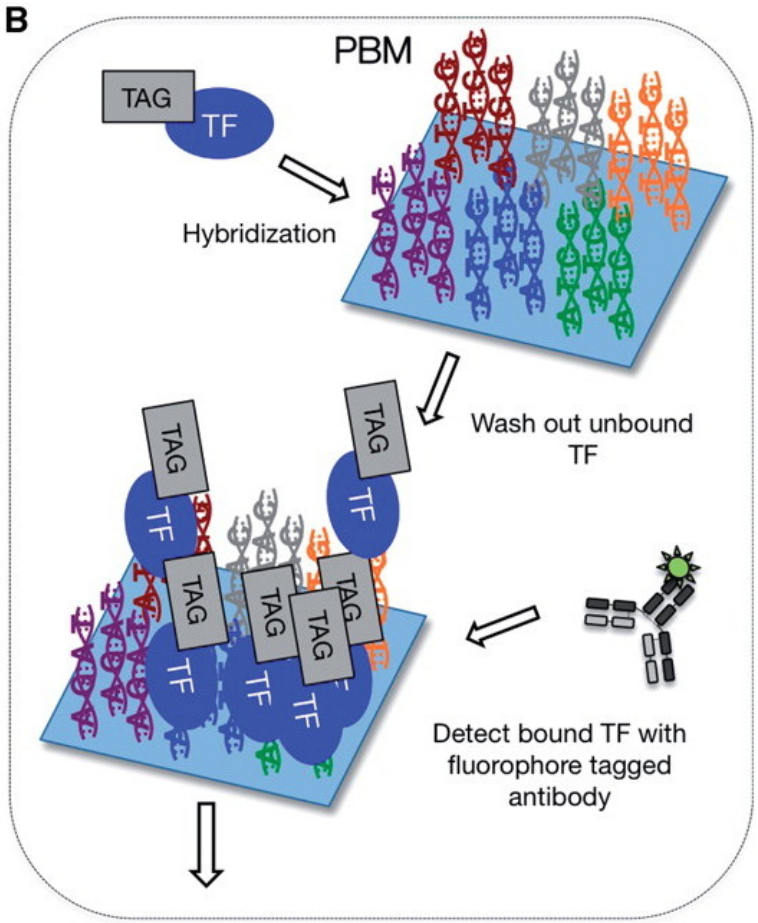
Data

- ChIP-seq peaks:
  approximate binding regions

$\Rightarrow$ extract sequences under peaks

- ChIP-seq peak statistics:
  information about TF
  abundance at binding region

Experimental techniques - PBM



[Geertz & Maerkl, Brief Func Genom, 2010]

Data

- PBM probes:
  contain all possible DNA 10-mers

$\Rightarrow$ probe sequences
(length 35 bp + linker)

- Probe intensities:
  information about TF binding frequency

Requirements for a novel approach

Use all sequences

thresholding to extract top peaks/probes arbitrary
$\Rightarrow$ use all peaks and probe sequences, respectively

Use all information present in the data

ChIP-seq	sequence under peak
	peak statistics
	binding more likely around peak center
PBM	probe sequence (including part of linker)
	probe intensities

Use discriminative learning principle

which often yield better results than generative principles

Allow for flexible choice of motif models

e.g., position weight matrices, weight array matrices,...

Retain acceptable runtime

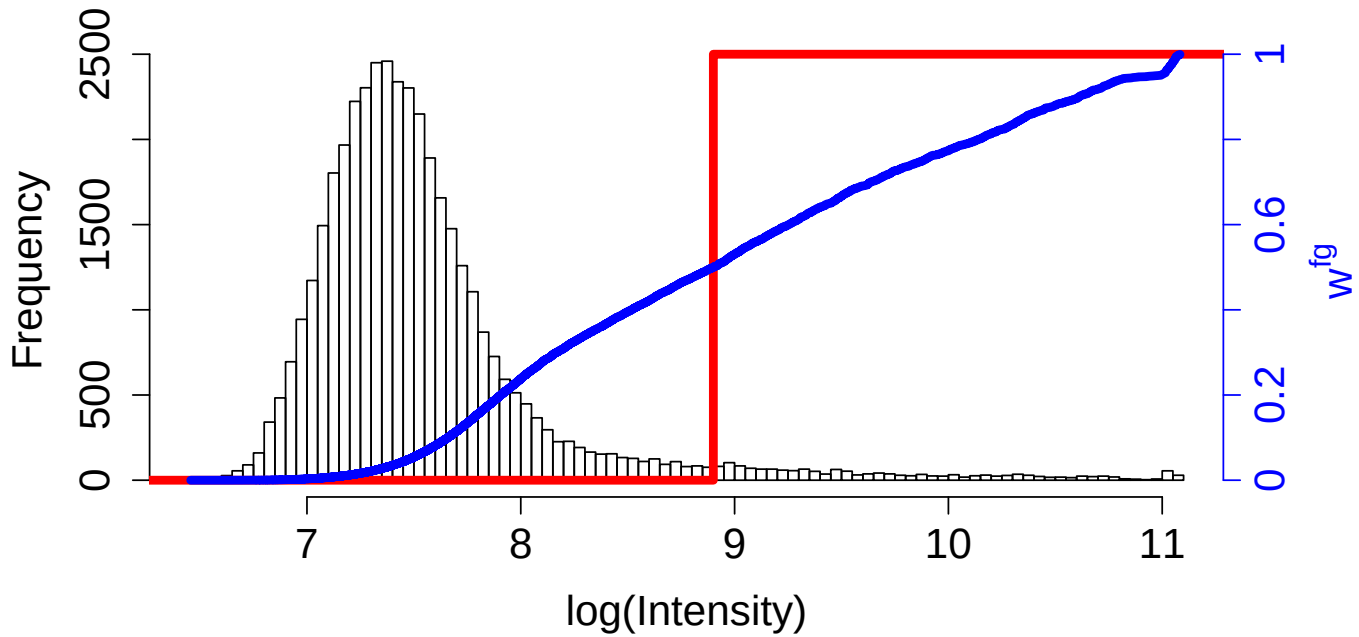
below 1h for majority of data sets

Weighting schema for integrating ChIP and PBM data

allows for using ChIP peak statistics and PBM probe intensities in a common approach

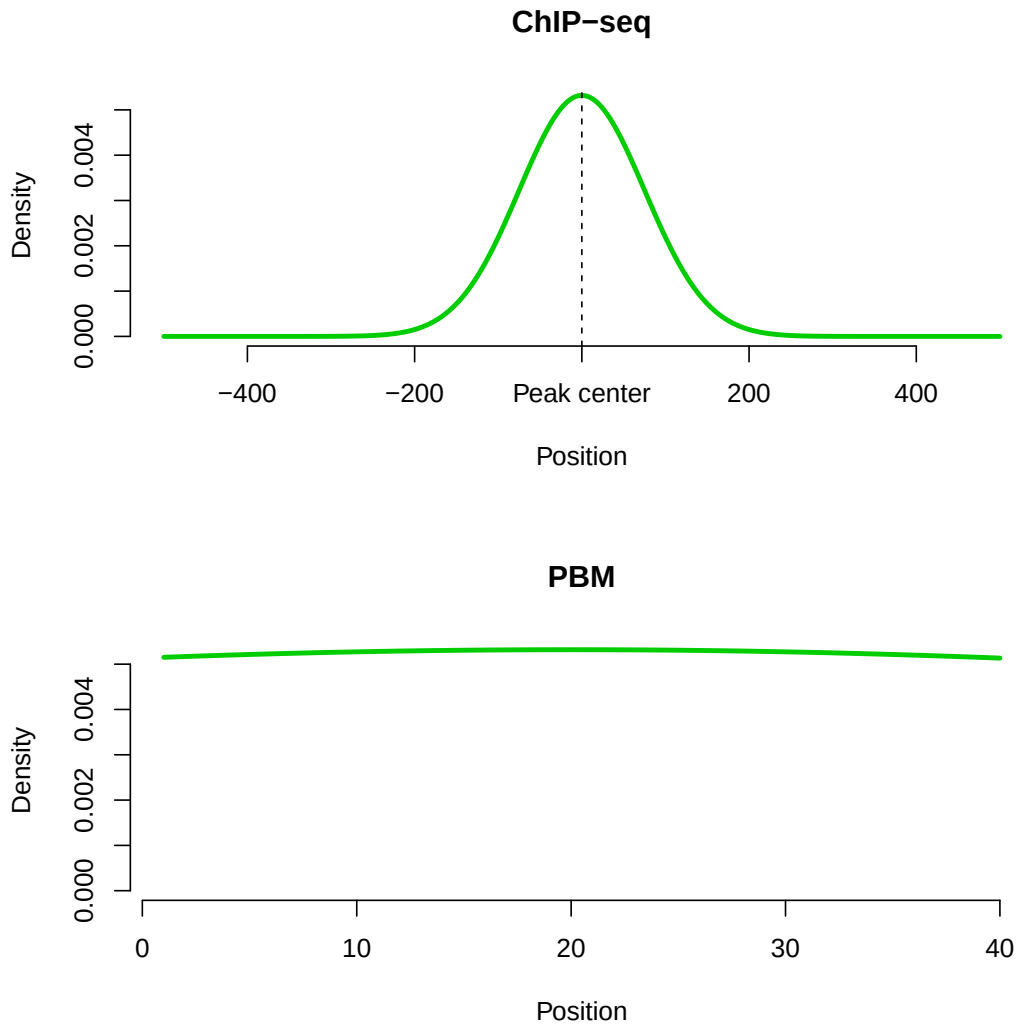
$$w_n^{fg} := \frac{1}{1 + \frac{h_n}{1-h_n} \cdot \frac{1-q}{q}}, \qquad w_n^{bg} := 1 - w_n^{fg}$$

h_n: relative rank of sequence $\mathbf{x}_n$ based on peak statistic or probe intensity,
q: weighting factor, i.e., a-priori fraction of foreground sequences



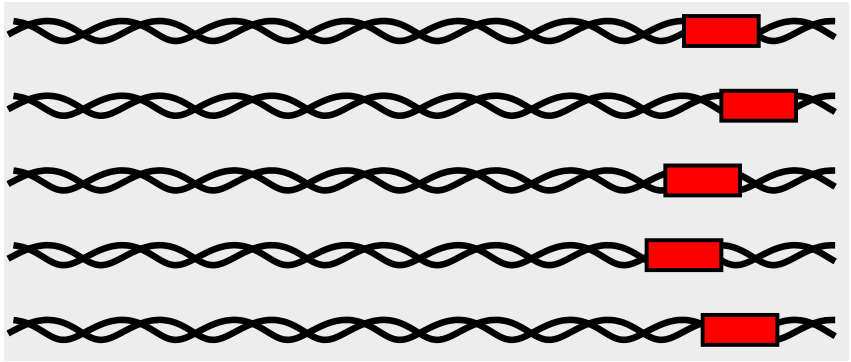
A-priori position distribution

represents that binding occurs close to peak center



Discriminative learning - Motivation

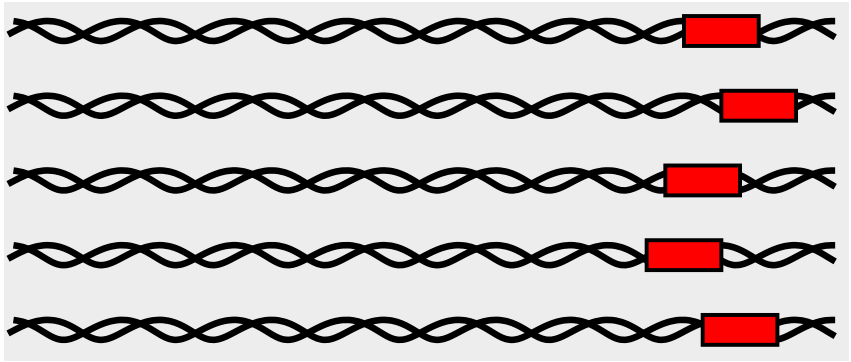
ChIP-seq positives



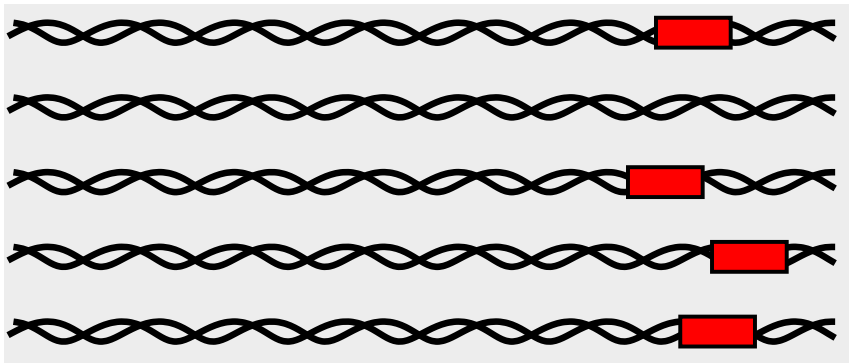
over-represented

Discriminative learning - Motivation

ChIP-seq positives



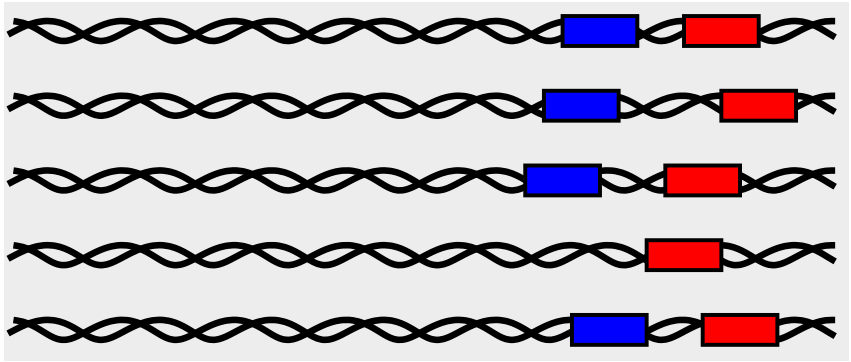
ChIP-seq negatives



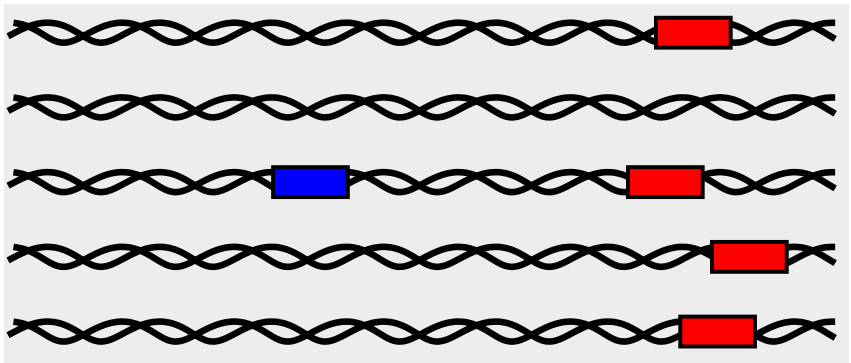
over-represented

Discriminative learning - Motivation

ChIP-seq positives



ChIP-seq negatives



over-represented
differentially abundant

$\Rightarrow$ discriminative learning

Discriminative learning - Objective function

Discriminative weighted maximum supervised posterior principle

$$\hat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda}}{\operatorname{argmax}} \underbrace{\sum_{n=1}^{N} \sum_{c \in \mathcal{C}} w_n^c \log \left(\frac{P(c|\boldsymbol{\lambda}) P_c(\mathbf{x}_n|\boldsymbol{\lambda})}{\sum_{\tilde{c} \in \mathcal{C}} P(\tilde{c}|\boldsymbol{\lambda}) P_{\tilde{c}}(\mathbf{x}_n|\boldsymbol{\lambda})} \right)}_{\textit{Weighted conditional likelihood}} + \underbrace{Q(\boldsymbol{\lambda}|\boldsymbol{\alpha})}_{\textit{Prior}},$$

where $\mathcal{C} = \{fg, bg\}$: set of classes,

$Q(\boldsymbol{\lambda}|\boldsymbol{\alpha})$: prior on the parameters $\boldsymbol{\lambda}$ given hyper-parameters $\boldsymbol{\alpha}$,

$P(c|\boldsymbol{\lambda})$: a-priori class probability, and

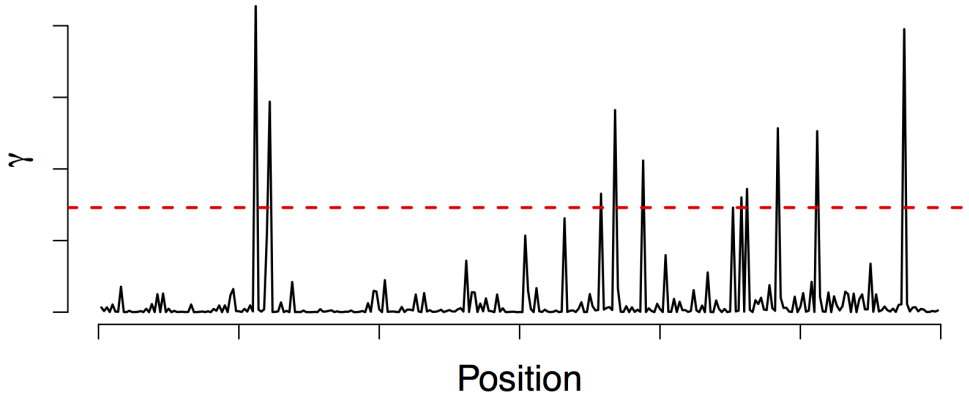
$P_c(\mathbf{x}_n|\boldsymbol{\lambda})$: class-conditional likelihood, "model"

Dimont models

$$P_{fg}(\mathbf{x}|\boldsymbol{\lambda}) = P(\text{motif}|\boldsymbol{\lambda}) \cdot \frac{1}{|\Sigma|^{L-w}} \sum_{\ell \in \mathcal{L}} P(\ell) P_{\text{motif}}(x_\ell, \ldots, x_{\ell+w-1}|\boldsymbol{\lambda}) + (1 - P(\text{motif}|\boldsymbol{\lambda})) \cdot \frac{1}{|\Sigma|^L}$$

- Dimont uses standard ZOOPS model ($P_{fg}(\mathbf{x}|\boldsymbol{\lambda})$)
- sequence flanking the motif: uniform, i.e., all nucleotides with equal probability
- motif model: strand model enclosing
  - position weight matrix (PWM): assumes nucleotide independence or
  - weight array matrix (WAM): allows dependencies between neighboring nucleotides or
  - higher-order Markov models
- background model: uniform or Markov model ($P_{bg}(\mathbf{x}|\boldsymbol{\lambda})$)

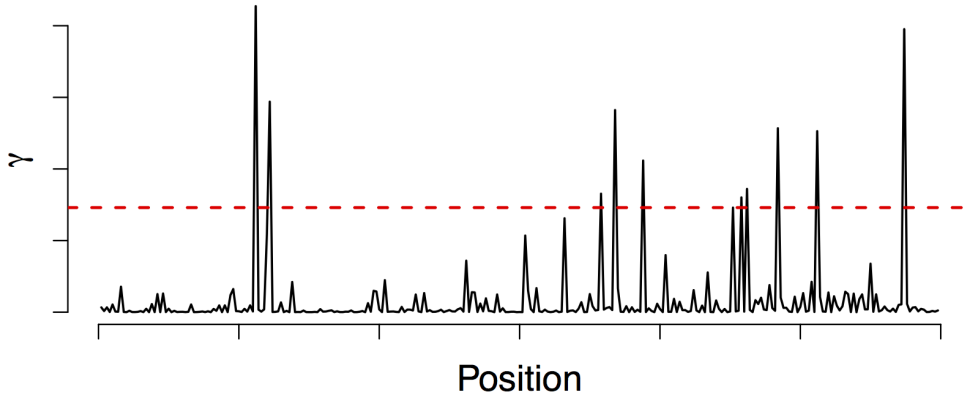
Speed-up strategies



Idea:

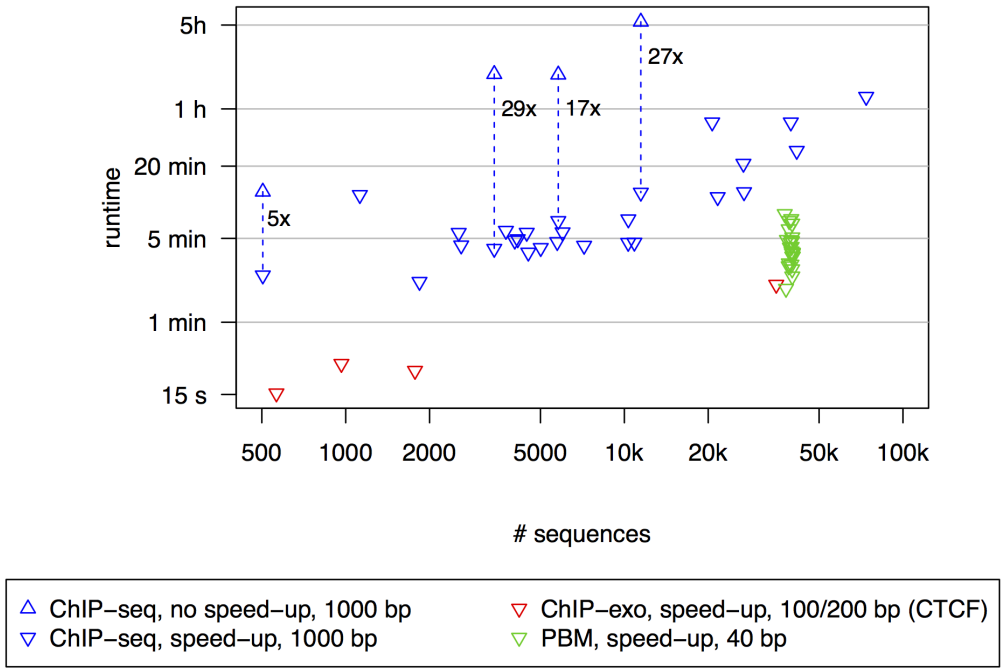
- pre-optimization on reduced data set
- evaluation of only highest-scoring motif occurrences

Speed-up strategies



Idea:

- pre-optimization on reduced data set
- evaluation of only highest-scoring motif occurrences



Benchmark on PBM data

66 PBM data sets (Weirauch *et al.*)

- protein binding microarray data for 66 TFs
- two different array designs (HK/ME) with different probes

Task:

Learn motif on one design, predict binding intensities for other design

Algorithm	Pearson corr.	AUC-ROC	Final
Dimont	0.695	**0.951**	**1.002**
FeatureREDUCE	0.693	0.949	0.997
Team_D	0.691	0.938	0.984
Team_E	**0.696**	0.906	0.952

Benchmark on ChIP-seq data

26 ChIP-seq data sets (Ma *et al.*)

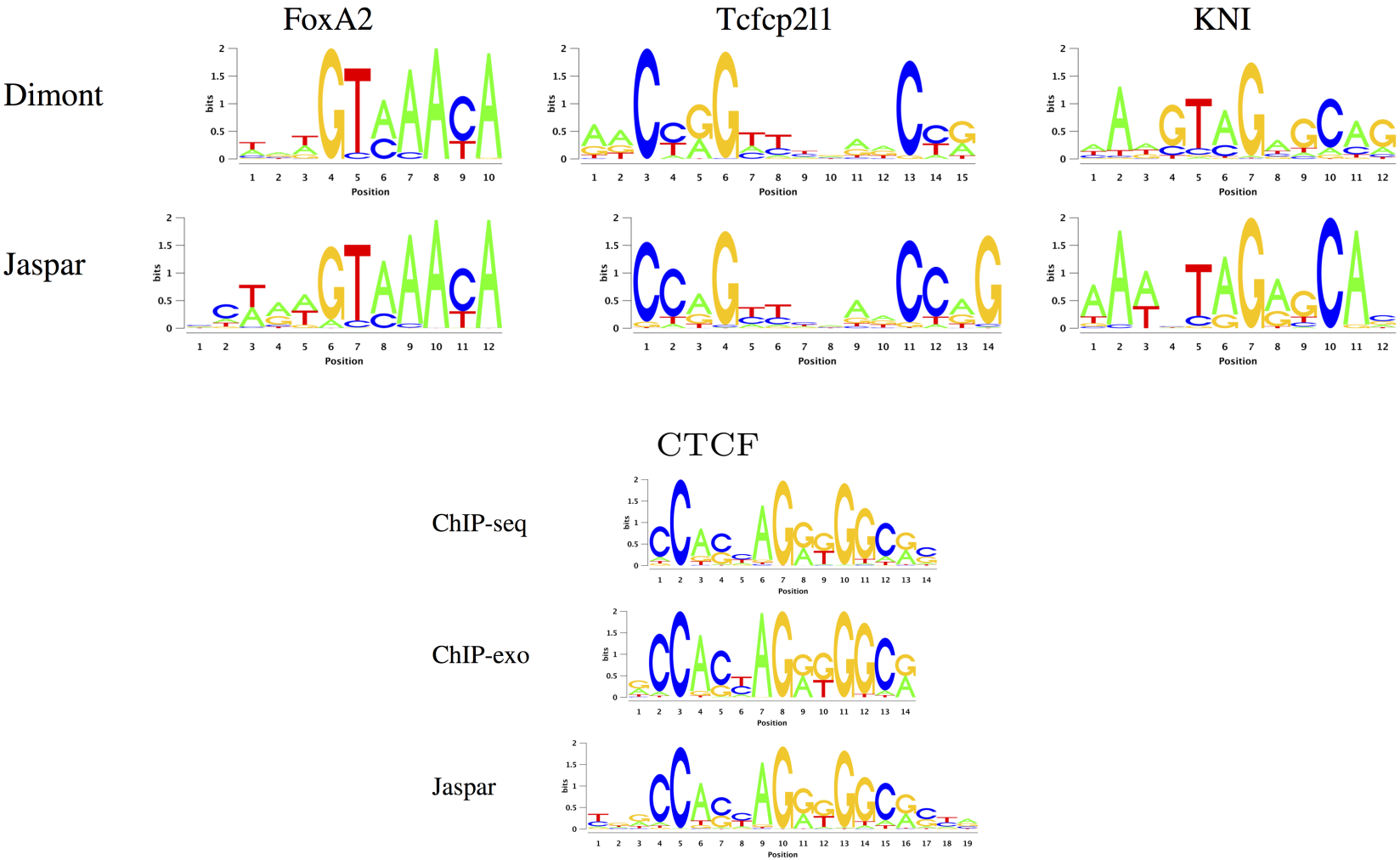
- 26 ChIP-seq data sets for TFs with known motifs
- human, mouse, fly

Task:

Discover motif consistent with literature

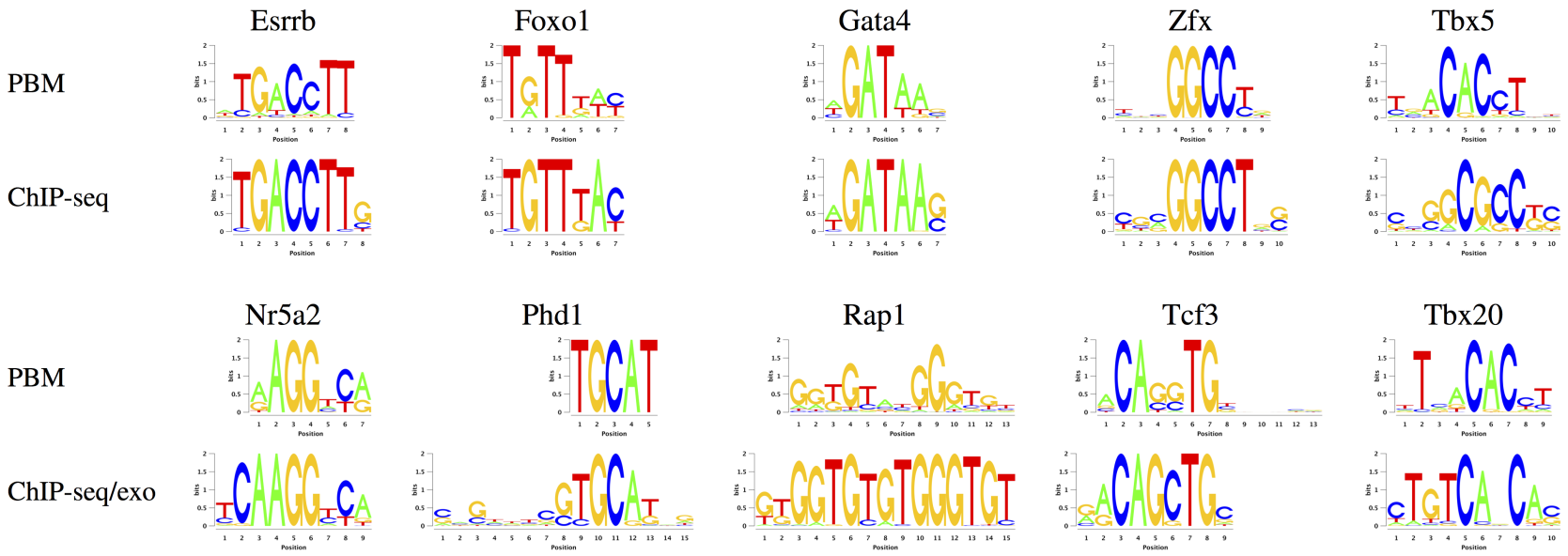
Algorithm	Total successes	Average rank
Dimont	**26**	1.23
POSMO	23	1.00
ChIPMunk	23	1.00
MEME	22	1.32
DME	22	1.45
DREME	22	1.45
HMS	12	1.00

Example motifs



$\Rightarrow$ most motifs fit the literature well

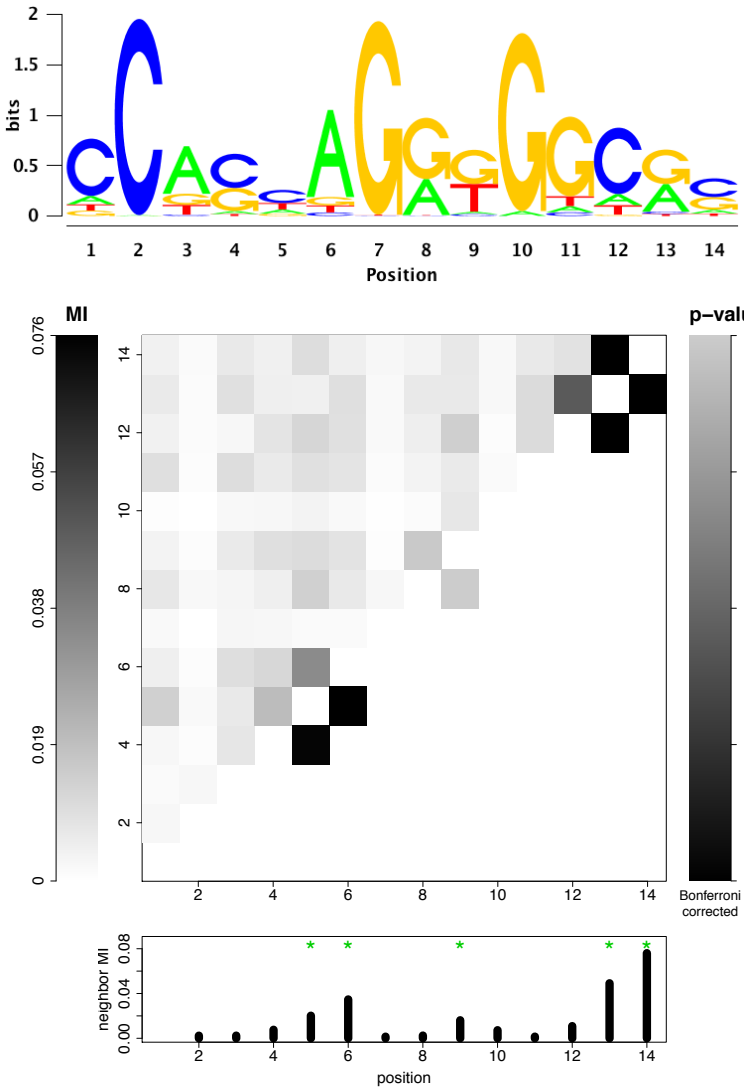
In-vivo vs in-vitro binding



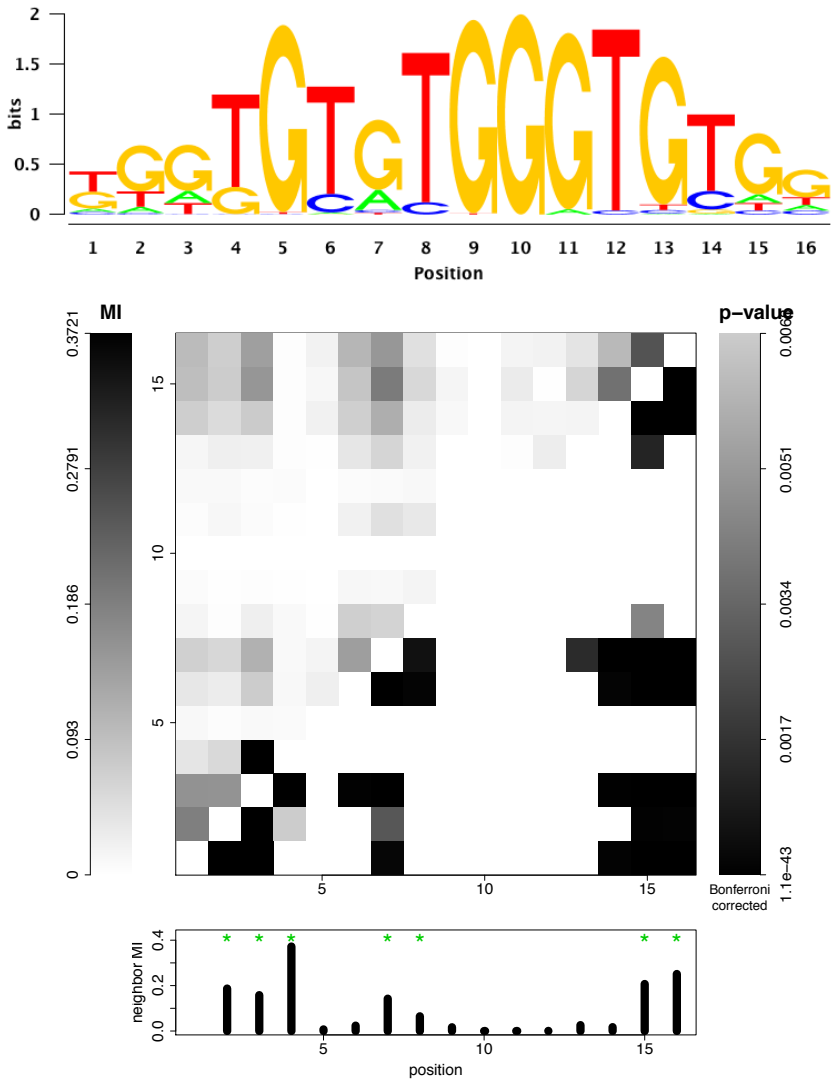
$\Rightarrow$ good accordance between *in-vivo* and *in-vitro* binding, but notable exceptions

Dependencies between neighboring positions

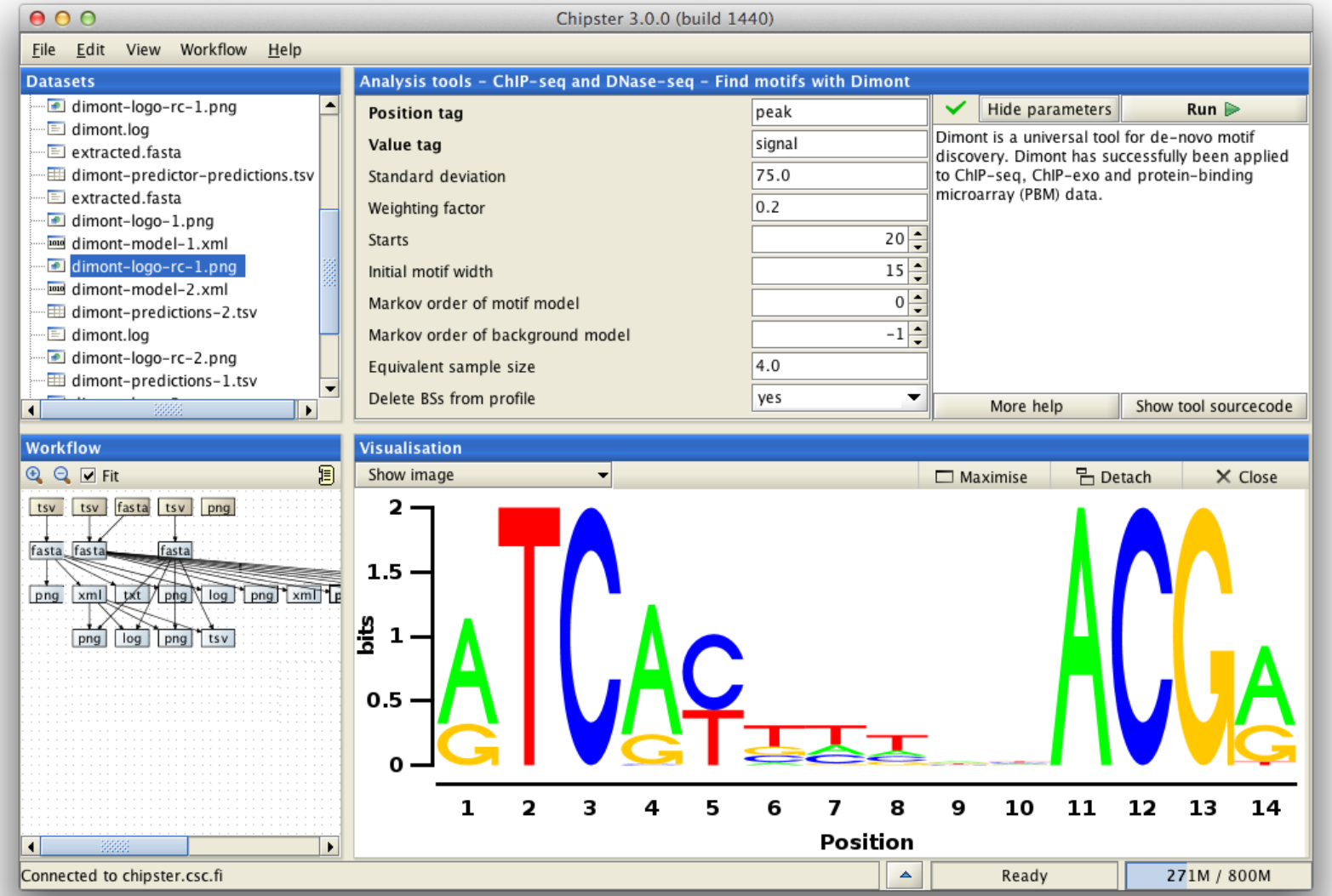
CTCF



Rap1

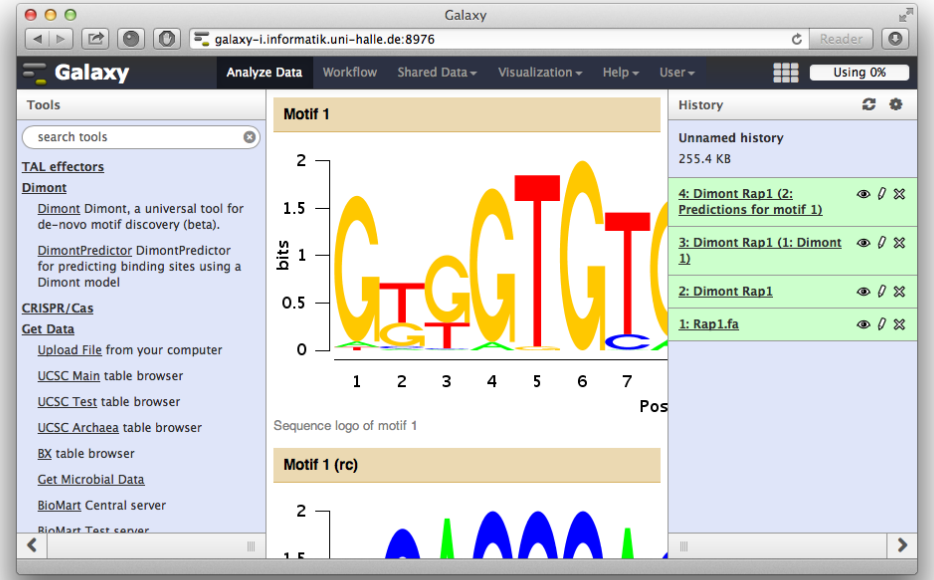


Dimont@Chipster



Galaxy application

- public server
- convenient user interface
- also available in Galaxy Tool-Shed



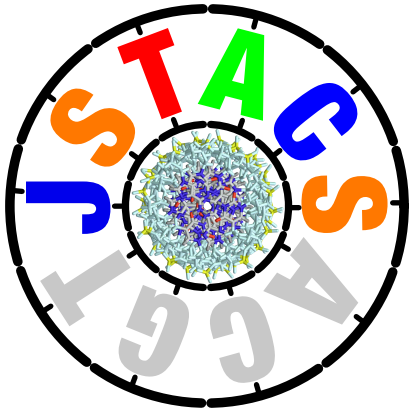
`galaxy.informatik.uni-halle.de`

- Galaxy-Server: 45 registered users, 500 runs (est.)
- Galaxy Tool-Shed: 60 clones

Command line

Command line application

- <key>=<value> interface
- easily scriptable
- multi-threaded




```
java -jar Dimont.jar data=myseqs.fa infix=myresult position=peak
value=signal threads=8
```

- available from www.jstacs.de/index.php/Dimont
- 290 downloads of command line program

Conclusions

Dimont, a general approach for motif discovery

- reliably discovers motifs from ChIP-seq and PBM data
- achieves an acceptable runtime

In-vitro and in-vivo binding

- often in good accordance
- but notable exceptions

Availability

- Chipster since version 2.11
- public Galaxy at `galaxy.informatik.uni-halle.de` and Galaxy Tool-Shed
- command line application: `www.jstacs.de/index.php/Dimont`

Conclusions

Dimont, a general approach for motif discovery

- reliably discovers motifs from ChIP-seq and PBM data
- achieves an acceptable runtime

In-vitro and in-vivo binding

- often in good accordance
- but notable exceptions

Availability

- Chipster since version 2.11
- public Galaxy at `galaxy.informatik.uni-halle.de` and Galaxy Tool-Shed
- command line application: `www.jstacs.de/index.php/Dimont`

Thank you for your attention!