Toolkit for ChIP-Seq based comparative analysis of the PWM methods for prediction of transcription factor binding sites

Yu. Kondrakhin[1,2], T. Valeev[1,3], R. Sharipov[1,*], I. Yevshin[1] and F. Kolpakov[1,2]

[1] Institute of Systems Biology, Ltd, Novosibirsk, Russia
[2] Design Technological Institute of Digital Techniques, SB RAS, Novosibirsk, Russia
[3] Institute of Informatics Systems, SB RAS, Novosibirsk, Russia
[*]To whom correspondence should be addressed: yvkondrat@mail.ru.

Key words: Transcription factor binding site, ChIP-Seq, position weight matrix approach, the ROC curve, area under curve.

Abstract

Despite wide application of the powerful ChIP-Seq technology for experimental identification of transcription factor (TF) binding sites, the computational prediction of the TF-binding sites is also relevant. Many methods for the prediction of the TF-binding sites have been developed over the last decades. Some of them represent position weight matrix (PWM) approach that is the most common and widely used. However, there exists little guidance in the choice among these methods because of a comprehensive comparison of existing methods is still challenging in practice. Thus, the direct use of the ChIP-Seq data for assessing predictive ability of the methods does not seem advisable because of such reasons as the tethered binding or false positive rates of peak detection algorithms. We have developed computational toolkit for reliable comparison of prediction methods under condition that unknown fraction of the ChIP-Seq data do not contain genuine TF-binding sites. On the base of developed toolkit, we have performed comparative analysis of three existing methods that represent PWM approach. The analysis has revealed that MATCH performed significantly worse than two other methods while common additive method outperformed others.

Introduction

Since its introduction in 2007 (Johnson *et al.*, 2007), ChIP-Seq has become the most powerful experimental technique for the genome-wide study of interactions between TFs and DNA. As a rule, a single ChIP-Seq experiment generates millions of short reads. Then the sequenced reads are aligned (mapped) to a reference genome and the TF-binding regions are identified by applying a peak detection algorithm (or peak finder) to the resulted set of tags (aligned reads). Until now a number of peak detection algorithms have been proposed, in particular, MACS (Model-based Analysis of ChIP-Seq) (Zhang *et al.*, 2008) and SISSRs (Site Identification from Short Sequence Reads) (Jothi *et al.*, 2008). The reproducibility of nine peak detection algorithms including MACS and SISSRs was studied in (Li *et al.*, 2011) on two repeated ChIP-seq experiments for CTCF.

It was inferred that MACS is one of the highest reproducible algorithm while SISSRs is the least reproducible. This conclusion was made with the help of the correspondence profiles fitted by copula model.

A comparative analysis of nine peak detection algorithms including MACS and SISSRs was performed in (Laajala *et al.*, 2009). This comparison demonstrated that biological conclusions could change dramatically when the same raw ChIP-Seq dataset was processed using different algorithms. It was also indicated that the optimal choice of algorithm depends heavily on selected dataset. Eleven different peak detection algorithms including MACS and SISSRs were also compared on common data sets (Wilbanks and Facciotti, 2010). This study offered a variety of ways to assess the performance of each algorithm and addressed the questions as to how to select the most suitable among several available methods. In general, one can conclude that currently it is impossible to choose the most reliable and well-validated algorithm for peak detection.

Despite the emergence of ChIP-Seq technology, application of the theoretical methods for prediction of TF-binding sites is also relevant. Initially ChIP-Seq approach was designed as experimental tool for identification of TF-binding sites. Unfortunately, some TF-binding regions do not represent genuine TF-binding sites because of, at least, the following three reasons. First, peak detection algorithms can produce much wider TF-binding regions (500 – 2000 bp or longer) than actual TF-binding sites (5-15bp). Second, some TF-binding regions are spurious due to false positive rates of methods for read mapping and for peak detection. Third, unknown fraction of the TF-binding regions should not contain the TF-binding sites because of tethered binding (Wang *et al.*, 2012). In this case, transcription factor bound to DNA fragment not because it recognized its site, but because it bound (due to protein-protein interaction) to another transcription factor that, in turn, bound to DNA.

In the 30 years since PWM approach was introduced (Stormo *et al.*, 1982), it has become the most common and widely used for computational analysis of the TF-binding sites, see (Stormo, 2013) for a review. A number of methods for prediction of the TF-binding sites have been developed within this approach. In particular, PWM algorithms were implemented in the computational tools such as MATCH (Kel *et al.*, 2003) MatInspector (Quandt *et al.*, 1995), MATRIX SEARCH (Chen *et al.*, 1995), ANN-Spec (Workman and Stormo, 2000) and MEME (Bailey *et al*., 2006). There are several repositories that accumulate many matrices for representation of TF-binding sites, in particular, TRANSFAC (Matys *et al*., 2006), JASPAR (Portales-Casamar *et al*., 2010), Factorbook (Wang *et al.*, 2012), UniPROBE (Robasky and Bulyk, 2012) and HOCOMOCO (Kulakovskiy *et al*., 2013) are among them. Usually these matrices were derived from the experimentally identified TF-binding sites (or regions) obtained by gel-shift analysis, SELEX, plasmid construction assays, ChIP-Seq, universal protein binding microarray technology (PBM) and other experimental techniques. Majority of those PWMs are represented as position frequency matrices.

In general, the Receiver Operating Characteristic (ROC) curve has long been used in signal detection theory (Fukunaga, 1990; Therrien, 1989). It is a good way of visualizing the correspondence between sensitivity and false positive rate (or False Discovery Rate, FDR) of a detection method. The area under the ROC curve, known as the AUC, is currently considered as the standard measure to assess the accuracy of prediction methods including those for prediction of the TF-binding sites. Currently it is common practice to reduce comparison of different prediction methods to comparison of the corresponding AUCs (Mathelier and Wasserman, 2013; Smeenk *et al*., 2008; Alamanova *et al*., 2010). It is important to note that it is necessary to have a representative sample of genuine TF-binding sites in order to evaluate the sensitivities of the comparable methods. Unfortunately, the direct use of the TF-binding region sets for sensitivity estimation does not seem advisable because of, at least, three reasons (including tethered binding) mentioned above. The main goal of our article is to work out a toolkit for reliable comparison of methods for prediction of the TF-binding sites under condition that unknown fraction of the TF-binding regions do not contain genuine TF-binding sites. On the base of developed toolkit, we have performed comparative analysis of the following three site models that represent PWM approach: common additive model, common multiplicative model and MATCH model. This analysis was carried out on 266 sets of human TF-binding regions from GTRD (Gene Transcription Regulation Database; http://wiki.biouml.org/index.php/GTRD) and matrices from TRANSFAC. The analysis has revealed that MATCH performed significantly worse than two other methods while common additive method outperformed others. It is important to note that inference of our comparative analysis is invariant with respect to choice of peak detection algorithm despite dissimilarities between MACS and SISSRs that were revealed by our toolkit.

Materials and methods

Data

Our toolkit intensively uses the human TF-binding region sets as input data. These sets, in turn, are stored in GTRD database. The GTRD collected raw ChIP-Seq data (sequenced reads) from literature, Gene Expression Omnibus (GEO), (Barrett *et al*., 2013), Sequence Read Archive (SRA), (Wheeler *et al*., 2008) and ENCODE project (http://www.nature.com/nature/journal/v489/n7414/full/nature11247.html). Currently GTRD contains 1450 human raw ChIP-Seq data sets and the ChIP-Seq controls (such as input DNA or IgG) are available for 1291 (89%) sets. The sequenced reads were aligned to reference genome (build 37) using Bowtie (Langmead *et al*., 2009) and the sets of the TF-binding regions were generated independently with the help of MACS and SISSRs.

The ROC curves and AUCs as basis of comparison

According to common practice, the areas under the ROC curves are widely used in order to compare the site models. In turn, each ROC curve represents the correspondence between sensitivity of model and FDR (False Discovery Rate). In general, it is necessary to have a representative sample of genuine TF-binding sites in order to calculate the sensitivity. However, only sets of the TF-binding regions are available instead of the required samples. It is assumed that each TF-binding contains genuine TF-binding site.

Therefore the sensitivity was computed as a relative number of the TF-binding regions that contain one or more TF-binding sites predicted. The FDR was determined as the relative number of the TF-binding regions containing false positives among all TF-binding regions containing site predictions. It was calculated with the help of 10-fold permutations of nucleotides in each TF-binding region. For UACs calculation we have used the sets of the TF-binding regions that are stored in GTRD.

Scheme of site model comparison

According to common practice, the comparison of site models is reduced to comparison of AUCs. In turn, AUCs are calculated on the sets of the TF-binding regions. However, the direct use of the full TF-binding region sets for the AUCs calculation does not seem advisable because some TF-binding regions can be empty, i.e. do not contain genuine TF-binding sites. The following scheme of site model comparison takes into account the assumption about existence of empty TF-binding regions.

We have developed the computational toolkit for ChIP-Seq based comparison of the PWM methods therefore the given position frequency matrix and the set of the TF-binding regions are the input for the AUCs calculation; see Figure 1. Thus, the site models share the same matrix but represent distinct algorithms for site scoring. Then the given set of the TF-binding regions can be modified, if necessary. Namely, all the TF-binding regions can be shortened or lengthened depending on a priori information about them.

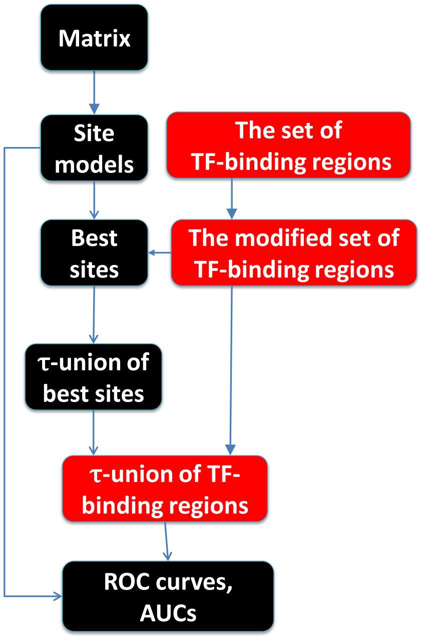

Matrix
Site models
The set of TF-binding regions
Best sites
The modified set of TF-binding regions
τ-union of best sites
τ-union of TF-binding regions
ROC curves, AUCs

Figure 1. Flowchart of the AUCs calculation.

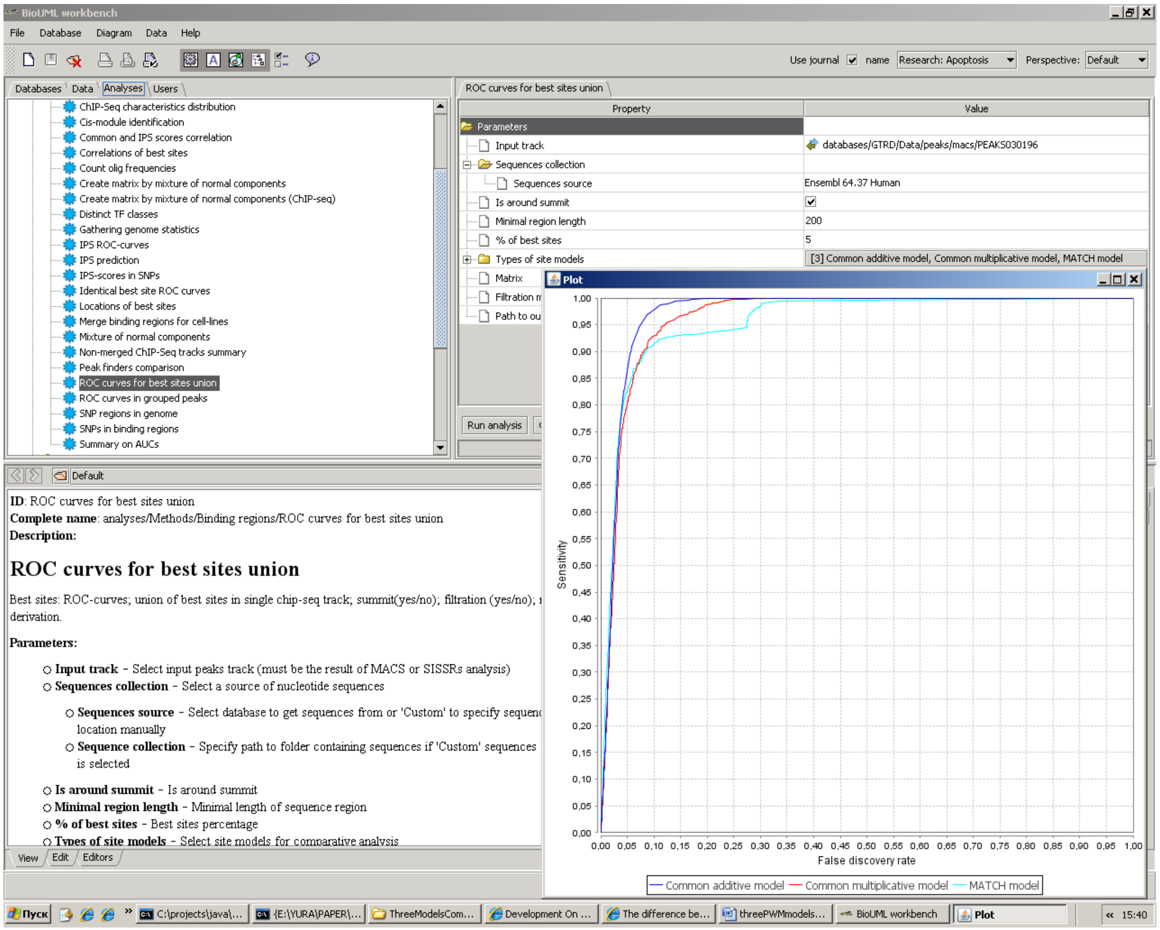
At the next step, each site model predicts its so-called ‘best site’ in every modified TF-binding region. The ‘best site’ of the given site model is defined as fragment of the TF-binding region where site model evaluated maximal score among all scores calculated for every possible fragments of the TF-binding region. Then top list of the τ percent (τ is given) of ‘best sites’ with the highest scores is selected for each site model and the so-called τ-union of the ‘best sites’ is composed as a union of all top lists selected. Finally, the so-called the τ-union of the TF-binding regions is defined as merged union of such TF-binding regions that contain at least one ‘best site’ from τ-union of the ‘best sites’. At last, the ROC curves are generated on the τ-union of the TF-binding regions and the corresponding AUC values are calculated.

Implementation

The proposed toolkit has been designed not only to perform the site model comparative analysis but also to reveal some fruitful features of the site models and the TF-binding regions. The toolkit consists of the following five independent computational modules (tools) implemented with the help of the open source BioUML / geneXplain plug-in framework (http://biouml.org/; http://genexplain.com/):

1. ‘ROC curves for best sites union’
2. ‘Summary on AUCs’
3. ‘Peak finders comparison’
4. ‘Locations of best sites’
5. ‘ROC curves in grouped peaks’.

The ‘ROC curves for best sites union’ module is a key tool in the toolkit. According to the flowchart in Figure 1, it generates the ROC curves (see, for example, Figure 2) and calculates the corresponding AUCs for the user-selected set of site models when value of parameter τ ($1 \leq \tau \leq 100$) and the set of the TF-binding regions are pre-specified. To form the set of site models, the toolkit provides user with the following basic list of the five available site models that share the same input matrix and represent PWM approach: Common additive model, Common multiplicative model, MATCH model, IPS model and Multiplicative IPS model, see Appendix for details. In order to modify (if necessary) the initial set of the TF-binding regions, toolkit provides user with appropriate input parameters, see Table A1 in Appendix for details. The resulted ROC curves and corresponding AUCs will be stored within user-specified folder.



Property	Value
Parameters	
Input track	databases/GTRD/Data/peaks/macs/PEAKS030196
Sequences collection	
Sequences source	Ensembl 64.37 Human
Is around summit	☑
Minimal region length	200
% of best sites	5
Types of site models	[3] Common additive model, Common multiplicative model, MATCH model
Matrix	
Filtration m	
Path to ou	

ID: ROC curves for best sites union
Complete name: analyses/Methods/Binding regions/ROC curves for best sites union
Description:

ROC curves for best sites union

Best sites: ROC-curves; union of best sites in single chip-seq track; summit(yes/no); filtration (yes/no); r derivation.

Parameters:

- **Input track** – Select input peaks track (must be the result of MACS or SISSRs analysis)
- **Sequences collection** – Select a source of nucleotide sequences
  - **Sequences source** – Select database to get sequences from or 'Custom' to specify sequenc location manually
  - **Sequence collection** – Specify path to folder containing sequences if 'Custom' sequences is selected
- **Is around summit** – Is around summit
- **Minimal region length** – Minimal length of sequence region
- **% of best sites** – Best sites percentage
- **Types of site models** – Select site models for comparative analysis

Figure 2. Screenshot of ‘ROC curves for best sites union’ tool.

The ‘Summary on AUCs’ tool performs comparative analysis of site models when value of parameter τ is pre-specified. Initially all appropriate AUC values calculated by ‘ROC curves for best sites union’ tool are read in all available tables. Then comparison of AUC values is performed with the help of non-parametrical Friedman test and Wilcoxon signed rank test (Hollander and Wolfe, 2003). In the case of Friedman test, chi-squared distribution with (k-1) degrees of freedom is used for assessing the statistical significance of difference between AUCs, where k denotes number of site models. In the case of Wilcoxon test, the significances of the differences are assessed with the help of normal approximations of the test statistics. Probability densities of differences between paired AUCs are estimated by kernel density estimator (Wasserman, 2004) with Epanechnikov kernel and are plotted for user.

The ‘Peak finders comparison’ tool performs comparative analysis of two peak detection algorithms. To compare two peak detection algorithms, this tool carries out comparative analysis of the matched sets of the TF-binding regions where the numbers and mean lengths of the TF-binding regions are analyzed independently with the help of Wilcoxon signed rank test. The statistical significances are assessed on the base of normal approximations of the test statistics. Additionally, the impact of the ChIP-Seq controls (such as input DNA or IgG) on the performance of peak detection algorithms is analyzed. Probability densities of the numbers and mean lengths of the TF-binding regions are estimated by kernel density estimator with Epanechnikov kernel and are plotted for user.

The ‘Locations of best sites’ tool estimates and plots the probability density of the ‘best site’ locations along the TF-binding regions around the so-called summits where summit is determined by MACS as precise binding location within given TF-binding region. Probability density is estimated by kernel density estimator with Epanechnikov kernel.

The ‘ROC curves in grouped peaks’ tool was developed to analyze the relationships between the ROC curves and reliability characteristics that were assigned by peak detection algorithm to each TF-binding region. The tool rearranges the given TF-binding regions in increasing order of the reliability characteristic and divides the ordered set into several groups of the same size. Then the ROC curves are generated and the corresponding AUCs are calculated on each group.

Application

Comparison of MACS and SISSRs

On the one hand, comparative analysis of peak detection algorithms has an independent (substantive) interest. On the other hand, this analysis can reveal some features of the TF-binding region sets and the revealed features, in turn, can be appropriately taken into account in site model comparison in order to increase the reliability of conclusions.

For comparison of MACS and SISSRs, the ‘Peak finders comparison’ tool carried out comparative analysis of 1450 pairs of the human TF-binding regions sets stored in GTRD. Two characteristics, namely the numbers and mean lengths of the TF-binding regions were analyzed independently with the help of Wilcoxon signed rank test. Statistical significances of the differences were assessed with the help of normal approximations of the test statistics.

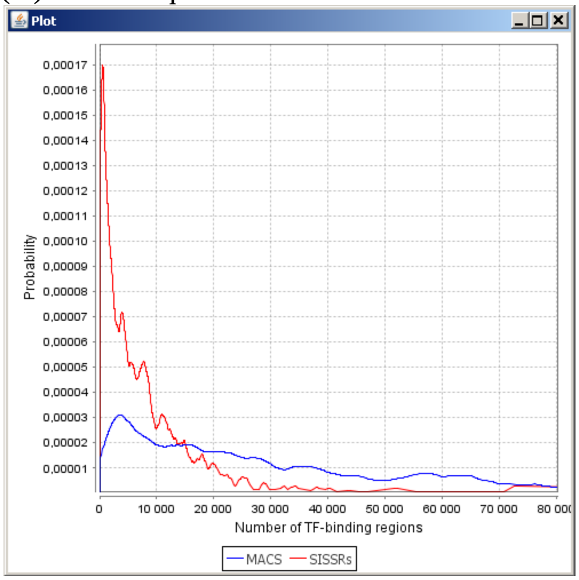
The performed analysis has revealed the following two dissimilarities between MACS and SISSRs. First, MACS generated significantly more the TF-binding regions than SISSRs when the ChIP-Seq controls were available, see Table 1. However, if ChIP-Seq controls were not available then SISSRs generated significantly more the TF-binding regions than MACS, see Table 1. Figure 3(A, B) demonstrates the probability densities of numbers of the TF-binding regions.

Second, comparative analysis has revealed that SISSRs generated significantly shorter TF-binding regions than MACS and this second dissimilarity is invariant with respect to presence/absence of the ChIP-Seq controls, see Table 1 and Figure 3(C, D). According to revealed dissimilarities we made conclusion that MACS and SISSRs have processed differently the same raw ChIP-Seq data.

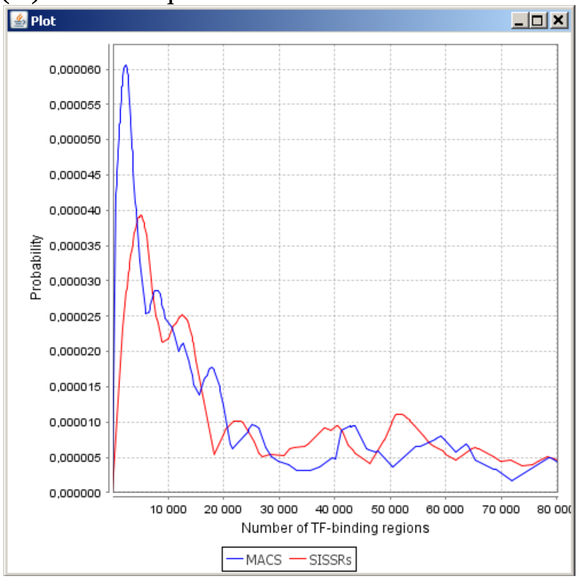
Table 1. Comparative analysis of MACS and SISSRs with the help of Wilcoxon signed rank test.

Comparable characteristic	ChIP-Seq control availability	Average characteristic for MACS	Average characteristic for SISSRs	Wilcoxon statistic (normal approximation)	p-value
Number of the TF-binding regions	Available	34013	7887	30.483	$<10^{-10}$
	Not available	28359	40839	6.069	1.3×10^{-9}
Mean length of the TF-binding regions	Available	811	105	31.123	$<10^{-10}$
	Not available	714	137	10.937	$<10^{-10}$

(A) ChIP-Seq control is available



(B) ChIP-Seq control is not available



(C) ChIP-Seq control is available

(D) ChIP-Seq control is not available

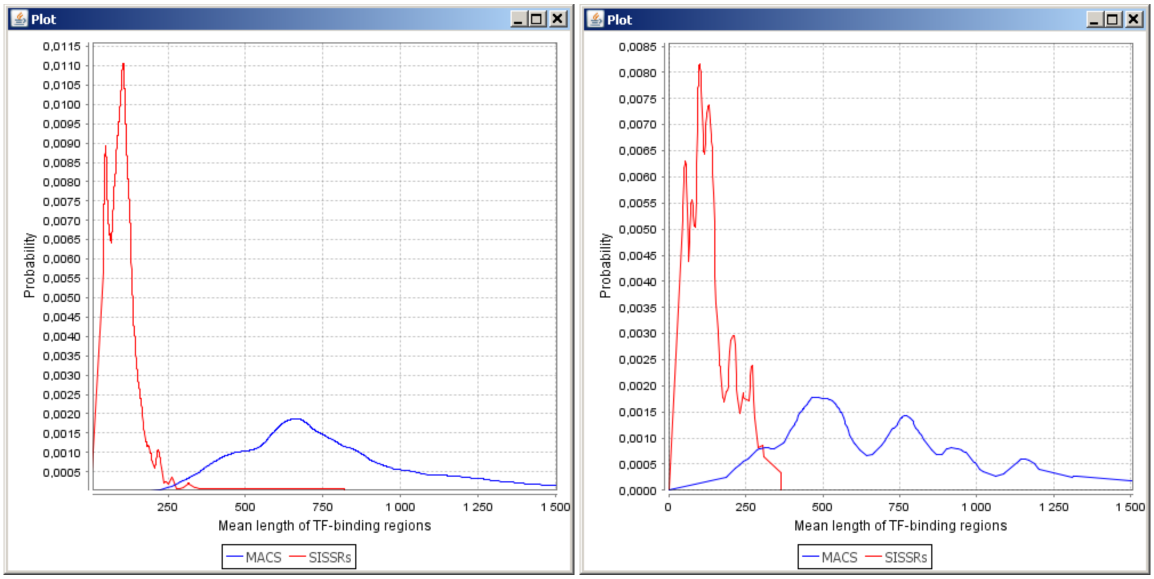


Figure 3. Probability densities of (A, B) number of the TF-binding regions and (C, D) mean length of the TF-binding regions.

Comparative analysis of three site models

On the base of developed toolkit, we have performed comparative analysis of the following three site models that represent PWM approach: common additive model, common multiplicative model and MATCH model, see their description in Appendix. For this analysis we have selected 266 TFs for whom we found simultaneously matrices in TRANSFAC (release 2012.4) and human TF-binding region sets in GTRD. It is important to note that we did not consider matrices derived for TF families. For example, despite the availability of USF1-binding region set in GTRD, we did not involve it into analysis because there is no appropriate matrix for the USF1-binding sites in TRANSFAC that contains matrices VUSF_01, VUSF_02, VUSF_C, VUSF_Q6 and V$USF_Q6_01 derived for the USF family.

Comparative analysis was performed independently on 266 sets of the TF-binding regions generated by MACS and on 214 sets generated by SISSRs. In the case of SISSRs we excluded 52 sets from our analysis because of their small sizes (<500). According to the flowchart in Figure 1, the ‘ROC curves for best sites union’ tool has calculated three AUCs on the given set of the TF-binding regions when value of parameter τ was specified. We have considered independently the following five values of τ: 100%, 35%, 25%, 15% and 5%. According to Table 1 and Figure 3(C, D), MACS produced much wider TF-binding regions than actual TF-binding sites. Therefore the initial set of the TF-binding regions was modified as follows. If the TF-binding regions were processed by MACS then we redefined them as regions of the lengths 200bp with the centers in summits. If the TF-binding regions were processed by SISSRs then all short (<200bp) regions are extended to 200bp.

After the AUC calculations the ‘Summary on AUCs’ tool has carried comparative analysis of site models with the help of Friedman and Wilcoxon tests. Chi-squared distribution with two degrees of freedom was used for assessing the significance of differences between three site models, see Table 2. On the base of this test, we made the conclusion that there exists significant difference between site models. This conclusion is invariant with respect to the choice of peak detection algorithm. However, this test is not intended to identify outperformance (superiority) of particular site model.

To get idea about site model outperformance, we analyzed all three possible pairs of site models with the help of Wilcoxon signed rank test, see Table 3. This analysis has revealed that MATCH performed significantly worse than two other models while common additive model outperformed others. For instance, when $\tau=25$ in the case of MACS the common additive model outperformed MATCH for 78.6% TFs and common multiplicative model outperformed MATCH for 66.5% TFs, see last column of Table 3. Probability densities of differences between AUCs also demonstrate that MATCH performed worse. It is important to note that, as in the case of Friedman test, the conclusions again do not depend on the choice of peak detection algorithm.

Table 2. Comparison of three site models with the help of Friedman test.

Peak detection algorithm	Percentage τ	Friedman test statistic	p-value
MACS	100	17.556	1.541×10^{-4}
	35	108.076	$<10^{-12}$
	25	139.908	$<10^{-12}$
	15	163.188	$<10^{-12}$
	5	218.362	$<10^{-12}$
SISSRs	100	15.165	5.093×10^{-4}
	35	51.732	5.843×10^{-12}
	25	91.103	$<10^{-12}$
	15	92.104	$<10^{-12}$
	5	106.150	$<10^{-12}$

Table 3. Comparative analysis of three site models with the help of Wilcoxon test.

1-st site model	2-nd site model	Peak detection algorithm	Percentage τ	Wilcoxon statistic (normal approximation)	p-value	Portion (in %) of TFs for which 1-st site model outperforms 2-nd site model
		MACS	100	3.875	1.067×10^{-4}	61.3
			35	11.238	$<10^{-15}$	75.9
			25	10.652	$<10^{-15}$	78.6
			15	11.593	$<10^{-15}$	80.5

Common additive model	MATCH		5	12.056	$<10^{-15}$	78.6
		SISSRs	100	3.434	5.941×10^{-4}	59.3
			35	7.414	1.226×10^{-13}	69.6
			25	8.653	$<10^{-15}$	75.7
			15	8.971	$<10^{-15}$	72.4
			5	9.112	$<10^{-15}$	71.0
Common multiplicative model	MATCH	MACS	100	3.250	0.001	59.4
			35	4.080	4.512×10^{-5}	61.7
			25	5.145	2.676×10^{-7}	66.5
			15	5.951	2.667×10^{-9}	67.7
			5	6.405	1.507×10^{-10}	72.2
		SISSRs	100	3.626	2.877×10^{-4}	61.7
			35	3.622	2.926×10^{-4}	62.6
			25	4.627	3.702×10^{-6}	65.4
			15	4.546	5.466×10^{-6}	66.8
			5	4.539	5.649×10^{-6}	68.2
Common additive model	Common multipli-cative model	MACS	100	0.074	0.941	50.8
			35	7.472	7.927×10^{-4}	71.4
			25	8.831	$<10^{-15}$	71.1
			15	9.740	$<10^{-15}$	71.4
			5	11.580	$<10^{-15}$	77.1
		SISSRs	100	1.825	0.068	47.2
			35	5.183	2.181×10^{-7}	61.7
			25	6.359	2.034×10^{-10}	66.4
			15	7.692	1.443×10^{-14}	68.2
			5	7.849	4.219×10^{-15}	67.3

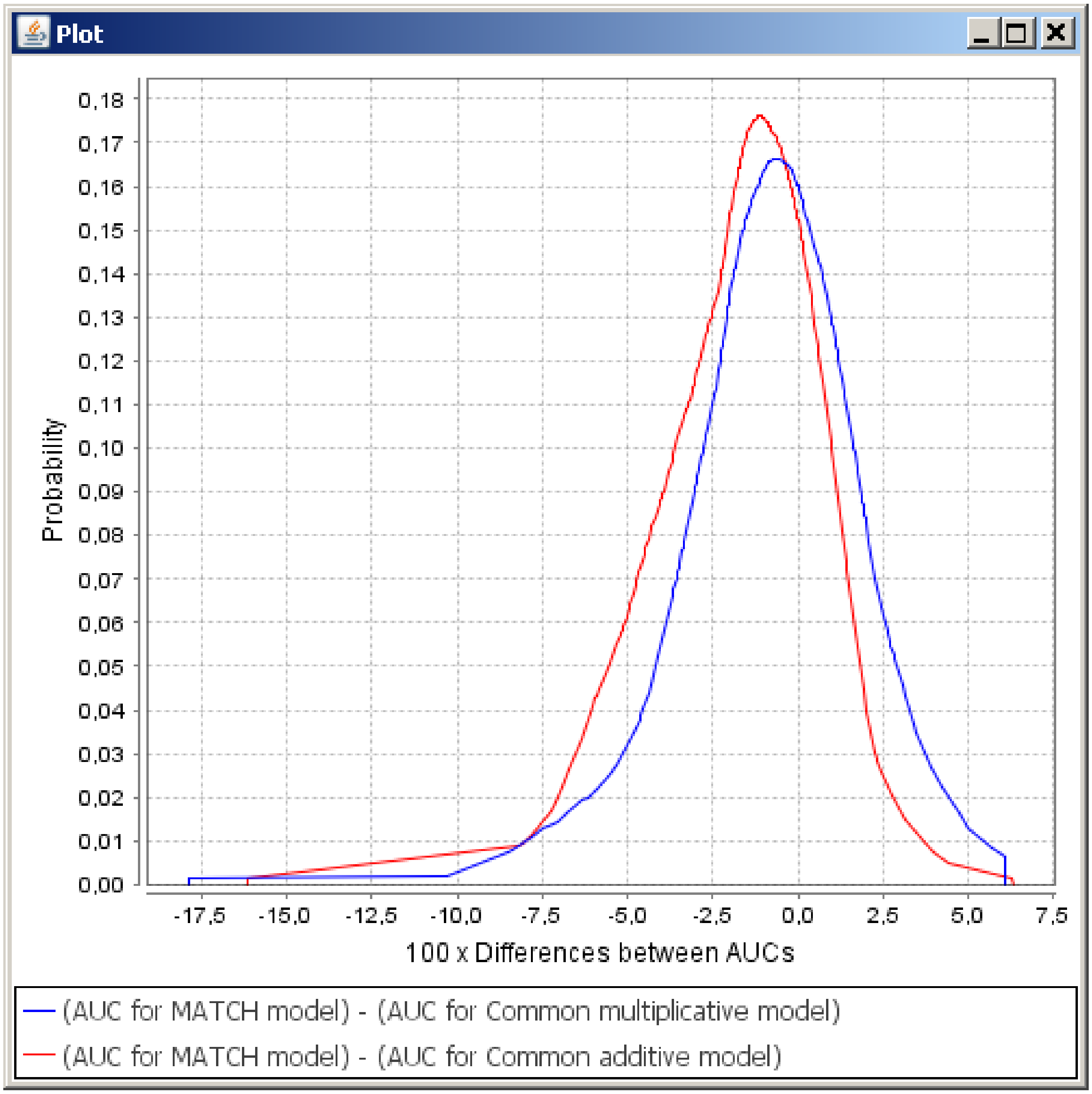


Figure 4. Probability densities of differences between AUCs when τ=25.

Discussion

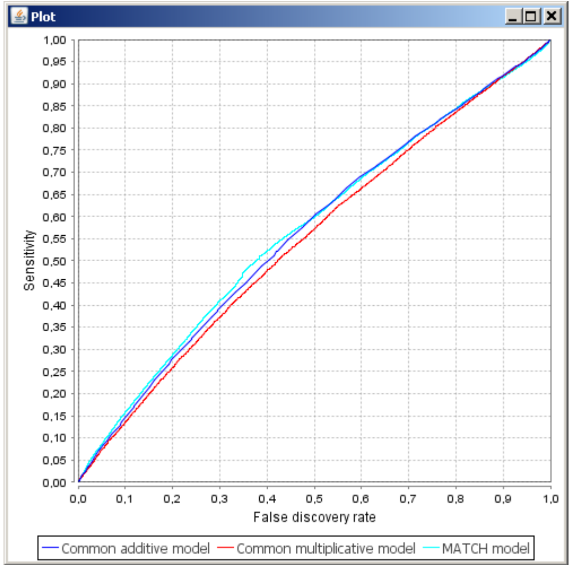
Currently the AUCs values are considered as the standard measures to assess the predictive abilities of site models. Certainly, for accurate calculation of precise AUCs it is necessary to have the representative samples of genuine TF-binding sites. Unfortunately, only sets of the TF-binding regions are available instead of the required samples. One can expect that direct use of initial sets of the TF-binding regions for the AUC calculations is not reasonable because some of the TF-binding regions can be empty. Indeed, it turned out that for majority of the selected TFs the values of AUCs were closed to 0.5 (see, for instance, Table 4) while the shapes of the ROC curves were approximately linear (see, for instance, Figure 5) when we directly used initial sets of the TF-binding regions. The low

AUC values have actually indicate a need for development of the special toolkit for comparison of site models on ChIP-Seq data.

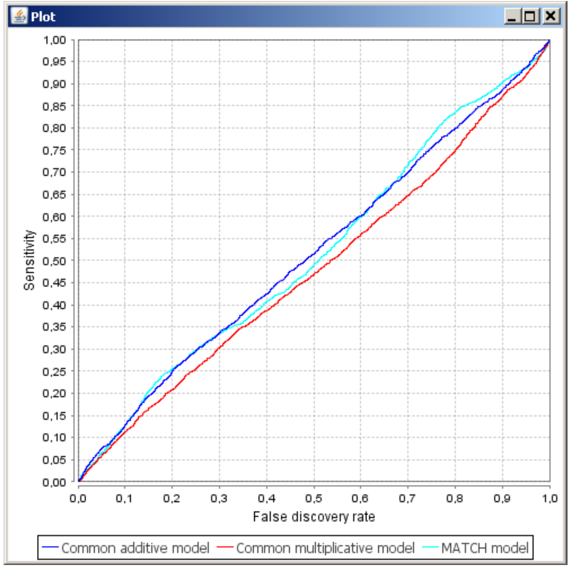
Table 4. AUCs calculated on YY1- and STAT1-binding regions. Matrices V$YY1_01 and V$STAT1_01 as well as the corresponding sets of the TF-binding regions with GTRD' IDs PEAKS030196 and PEAKS010470 were used for calculation of AUCs.

TF	Peak detection algorithm	AUCs for site models		
		MATCH	Common additive model	Common multiplicative model
YY1	MACS	0.569	0.564	0.549
	SISSRs	0.569	0.574	0.570
STAT1	MACS	0.515	0.515	0.480
	SISSRs	0.475	0.494	0.468

(A) YY1, MACS



(B) STAT1, MACS



(C) YY1, SISSRs

(D) STAT1, SISSRs

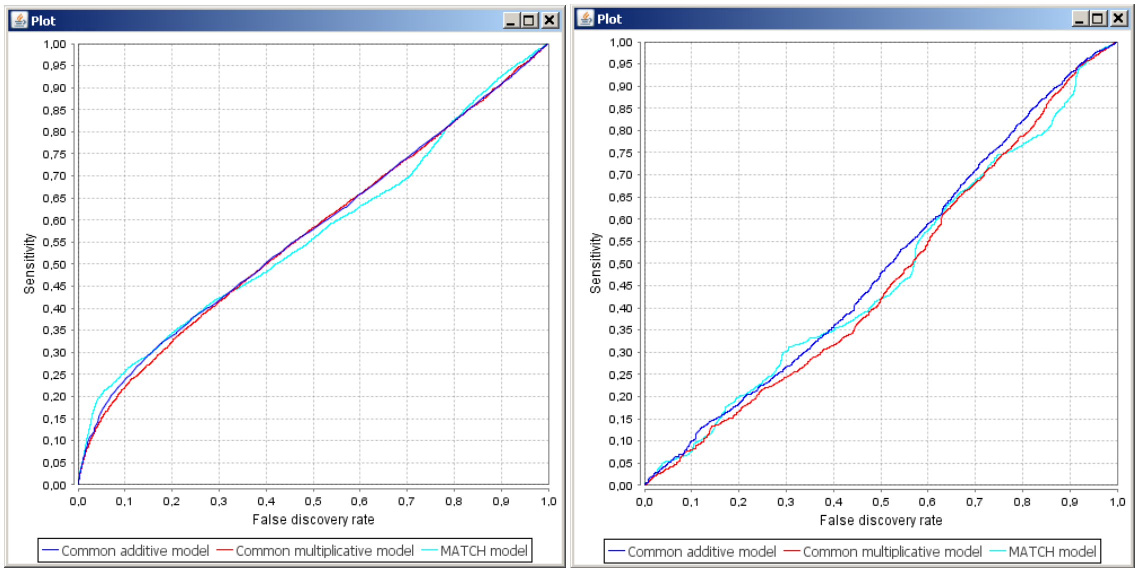


Figure 5. The ROC curves obtained on YY1- and STAT1-binding regions that were generated by MACS and SISSRs.

A shape of the ROC curve and the AUC value can be affected not only by empty TF-binding regions but also by lengths of the TF-binding regions. One can expect that the wider TF-binding regions, the higher FDR and the less convex the ROC curve. According to Table 1 and Figure 3(C, D), MACS produced much wider TF-binding regions than genuine TF-binding sites. In order to find an appropriate way to shorten reasonably the TF-binding regions generated by MACS, 'Locations of best sites' tool has estimated the probability densities of 'best sites' locations around the summits with the help of kernel density estimator. For majority of the selected 266 TFs it appeared that 'best sites' of each site model preferred to locate near summits and the maximal values of densities were observed approximately in the range [-100bp, 100bp] with respect to summits. Figure 6 demonstrates, for instance, the probability densities of 'best sites' locations around the summits within YY1- and STAT1-binding regions.

(A) YY1 (B) STAT1

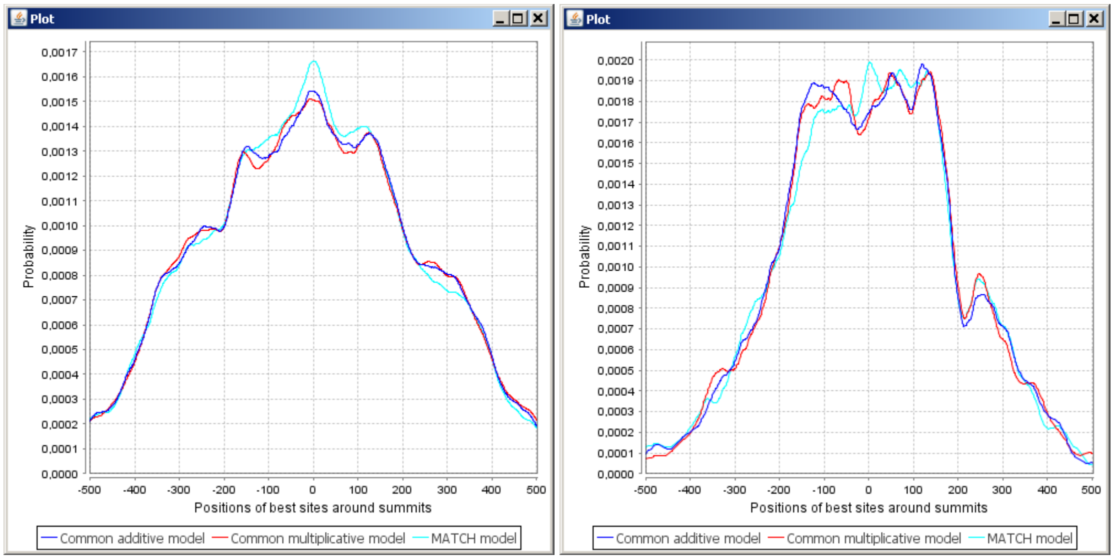
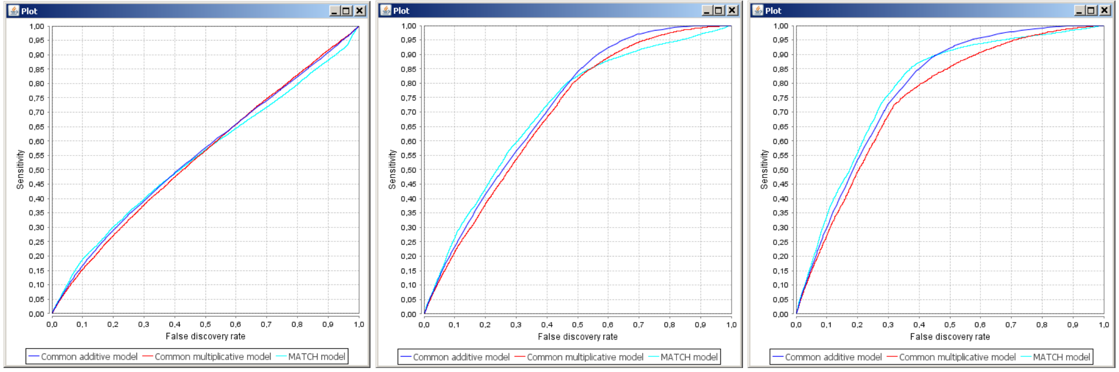


Figure 6. Probability densities of 'best sites' locations around summits for (A) YY1 and (B) STAT1.

The key step of the proposed scheme of the AUCs calculation (see Figure 1) is the construction of the τ-union of the TF-binding regions, where the percentage τ is free parameter. In general, there exists the following relationship between τ values and the shapes of the ROC curves: the smaller percentage τ, the more convexity of the ROC curve and the higher AUC values. Thus, for small values of τ (5% - 15%) the ROC curves, as a rule, are strongly convex while the shapes of the ROC curves became approximately linear when τ tends to 100%, see, for example, Figure 7 where the ROC curves were generated on the YY1-binding regions (processed by MACS). In turn, the corresponding values of AUCs are closed to 0.5 when τ tends to 100% while these values are closed to 1.0 when τ tends to 5%, see Table 5.

(A) τ = 100% (B) τ = 50% (C) τ = 35%



(D) τ = 25% (E) τ = 15% (F) τ = 5%

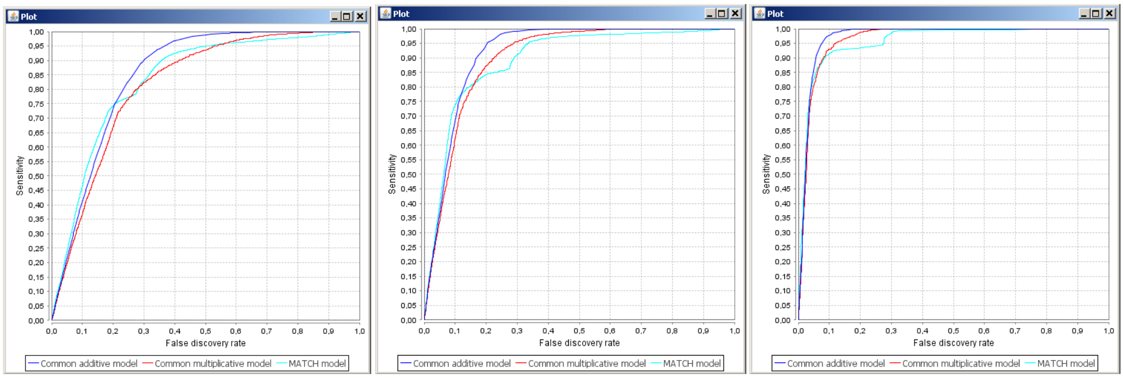


Figure 7. The ROC curves obtained for different values of τ on the YY1-binding regions that were generated by MACS.

Table 5. AUCs calculated for different values of τ on the YY1-binding regions that were generated by MACS.

Percentage, τ	Site model			Percentage of regions that are classified as empty
	MATCH	Common multiplicative model	Common additive model	
100	0.548	0.550	0.555	0
50	0.707	0.694	0.716	37.5
35	0.782	0.744	0.778	51.5
25	0.835	0.817	0.852	65.4
15	0.892	0.899	0.918	78.8
5	0.956	0.963	0.972	92.9

It is important to note that the shown relationship between τ and shape of the ROC curve can be interpreted as follows. According to definition of the τ-union of the TF-binding regions, it consists of such TF-binding regions that contain ‘best sites’ with the highest scores. In other words, the TF-binding regions containing ‘best sites’ with the smallest scores are removed. The removed TF-binding regions, in turn, represent empty regions from the point of view of all site models considered. Obviously, The higher percentage τ, the smaller number of regions that are classified as empty, see also first and last columns of Table 5. In this connection, it is interesting to note the following tendency presented in Table 3: the higher percentage τ, the lower statistical significance of differences between site models. In other words, the higher percentage τ, the more noisy τ-union of the TF-binding regions. Moreover, as a single exception, Wilcoxon test was not able to identify significant difference between common additive and multiplicative models on the full sets of the TF-binding regions (i.e. when τ =100%). However, this exception just confirms the assumption that full sets of the TF-binding regions can be noisy due to empty regions.

Certainly, the construction of the τ-union of the modified TF-binding regions is just one of the possible ways to compose the refined sets of the TF-binding regions that can be used for site model comparison. One of the alternative ways to compose the refined sets is to select the most reliable TF-binding regions and this way has been implemented in ‘ROC curves in grouped peaks’ tool.

As a rule, a peak detection algorithm assigns several characteristics (such as ‘FDR’, ‘Fold enrichment’, ‘Tag number’, ‘Score’ and ‘p-value’) of reliability to each TF-binding regions identified. ‘ROC curves in grouped peaks’ tool rearranged all TF-binding regions in the individual set in increasing order of the reliability characteristic and divided the ordered set into six groups of the same size. One can expect that shapes of the ROC curves have to change visibly in transition from first group to sixth group. However, serious changes were not observed for majority of TFs; see, for instance, Figure 8 that demonstrates the ROC curves created on the STAT-binding regions.

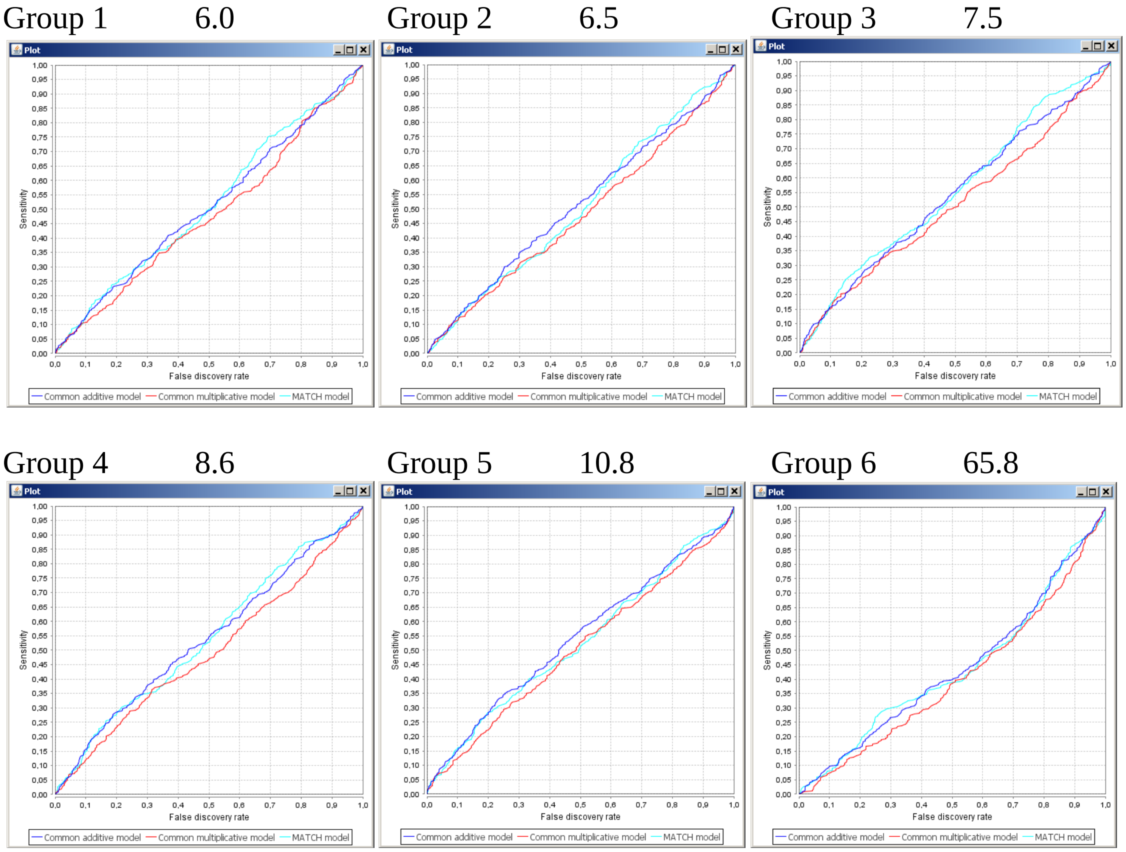


Figure 8. The ROC curves created on six groups of the STAT1-binding regions that were generated by SISSRs. ‘Tag number’ characteristic was used for division into groups. Average ‘Tag number’ is also shown for each group.

Appendix

Five site models available for comparative analysis

Currently, five site models that represent PWM approach are available for comparative analysis. For given TF they share the same position frequency matrix MAT = (m_{ij}), i={A,C,G,T}, j=1,...,*l* but produce diverse scores for fixed DNA fragment S = $(s_1,...,s_l)$. In other words, the models represent different scoring algorithms.

1. Common additive model. This model calculates the common additive score x defined by formula

$$x = x(MAT) = \Sigma_{j=1,...,l} \ \ score(j),$$

where the values score(j), j=1,…,*l*, are determined as follows:

score(j) = {m_{Aj}, if s_j=A; m_{Cj}, if s_j=C; m_{Gj}, if s_j=G; m_{Tj}, if s_j=T}.

2. Common multiplicative model. For fragment S this model calculates the common multiplicative score x_m,

$$x_m = \prod_{j=1,...,l} \ \ score(j).$$

This model can be converted to equivalent additive model by taking logarithms of matrix elements, i.e.

$$x_{ln} = \Sigma_{j=1,...,l} \ \ score^*(j),$$

where the values score*(j), j=1,…,*l*, are determined as follows:

score*(j) = {ln(m_{Aj}), if s_j=A; ln(m_{Cj}), if s_j=C; ln(m_{Gj}), if s_j=G; ln(m_{Tj}) if s_j=T}.

In order to avoid taking logarithm of zero we preliminarily found minimal non-zero element of matrix MAT. Then we replaced all zero values of MAT by this value and re-normed all changed columns of MAT in such a way that the sum of frequencies in each changed column was equal to unit.

3. MATCH model. This model is determined by popular PWM method MATCH for TF-binding site prediction. This model calculates the so-called matrix similarity score *mSS* defined in (Kel *et al*., 2003). Actually, this model is common additive model, which uses transformed matrix instead of initial matrix, where each column of transformed matrix was determined with the help of weighting the corresponding initial column by information content. More specifically, the j-th column of weight matrix is equal (up to the constant (–*Min* / (*Max-Min*))) to product of the j-th column of frequency matrix and the value *I(j)* / (*Max-Min*), j=1,...,*l*, where *I(j)*, *Min*, and *Max* were defined in (Kel *et al* 2003).

4-5. IPS model and Multiplicative IPS model. Briefly, in addition to the common additive/multiplicative scores, these models take into account the nucleotide content of the both flanks of the site cores. The detailed description of these models will be published in next volume of Virtual Biology.

Description of the inputs in tools

Table A1. Description of the inputs in tools.

Tool	Name of the input	Description
'ROC curves for best sites union'	Input track	Path to GTRD track that contains the initial set of the TF-binding regions.
	Sequences source	The genome build (has to be selected).
	Is around summit (IAS)	If IAS = true and summit exists, then each TF-binding region will be redefined as region of the length l_{min} with the center in summit.
	Minimal region length (l_{min})	If IAS = false, then all short ($<l_{min}$) regions will be extended to l_{min}. If IAS = false and l_{min} = 1 then the TF-binding regions will not be modified.
	% of best sites	Percentage τ ($1\leq \tau \leq 100$)
	Types of site models	Basic list of the site models available for comparison. User can select the subset of site models.
	Matrix	Path to position frequency matrix.
	Filtration matrix	Path to matrix for identification site motifs associated with the Alu repeats. If this path is not empty then the TF-binding regions containing such motifs will be removed from the AUC calculation.
	Path to output folder	The resulted ROC curves and AUCs will be stored within this folder. In particular, AUCs will be written into table with name 'AUCs'.
'Summary on AUCs'	Path to collection of folders	Each folder has to contain table with the name 'AUCs' that stores the AUC values. It is assumed that these tables were created by 'ROC curves for best sites union' tool.
	% of best sites	Percentage τ ($1\leq \tau \leq 100$)
	Path to output folder	The results of statistical analysis will be stored within this folder.
	Are revised	This input was developed in advance for future applications. Currently it is not functional.
	Minimal size (n_{min})	The AUC value will be omitted if it was calculated on the TF-binding region set with size less than n_{min}.

‘Peak finders comparison’	Species	In general, GTRD contains human, mouse and rat ChIP-Seq data. User can select a species.
	First folder containing ChIP-seq tracks	This folder has to contain the sets of the TF-binding regions processed by first peak detection algorithm.
	Second folder containing ChIP-seq tracks	This folder has to contain the sets of the TF-binding regions processed by second peak detection algorithm.
	Path to output folder	The results of comparison of peak detection algorithms will be stored within this folder
‘Locations of best sites’	Input track	The same as in ‘ROC curves for best sites union’ tool
	Sequences source	
	Is around summit	
	Minimal region length	
	% of best sites	
	Types of site models	
	Matrix	
	Path to output folder	The chart of probability density and table with positions of ‘best sites’ will be stored within this folder
‘ROC curves in grouped peaks’	Input track	The same as in ‘ROC curves for best sites union’ tool
	Sequences source	
	Is around summit	
	Minimal region length	
	% of best sites	
	Types of site models	
	Matrix	
	Number of groups	How many groups of the TF-binding regions will be created.
	Path to output folder	The resulted ROC curves and AUCs will be stored within this folder.

References

Alamanova, D., Stegmaier, P. and Kel, A. (2010) Creating PWMs of transcription factors using 3D structure-based computation of protein-DNA free binding energies, BMC Bioinformatics, 11:225.

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H/, Sherman, P.M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C.L., Serova, N., Davis, S. and Soboleva, A. (2013) NCBI GEO: archive for functional genomics data sets—update, Nucleic Acids Res., 41: D991-5.

Fukunaga, K. (1990) Introduction to statistical pattern recognition, Academic Press, San Diego, CA, Second Edition.

Johnson, D.S., Mortazavi, A., Myers, R.M. and Wold, B. (2007) Genome-wide mapping of in vivo protein-DNA interactions, Science, 316, p.1497-1502.

Jothi, R., Cuddapah, S., Barski, A., Cui, K. and Zhao, K. (2008) Genome-wide identification of in vivo protein–DNA binding sites from ChIP-seq data, Nucleic Acids Res., 36, p.5221-5231.

Hollander, M. and Wolfe, D.A. (1973) Nonparametric Statistics, New York: J. Wiley.

Kel, A.E., Gobling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O.V. and Wingender, E. (2003) MATCH^{TM}: a tool for searching transcription factor binding sites in DNA sequences, Nucleic Acids Res., 31, p.3576-3579.

Kulakovskiy, I.V., Medvedeva, Y.A., Schaefer, U., Kasianov, A.S., Vorontsov, I.E., Bajic, V.B. and Makeev, V.J. (2013) HOCOMOCO: a comprehensive collection of human transcription factor binding sites models, Nucleic Acids Res., 41: D195-202

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biol., 10: R25.

Li, Q., Brown, J.B., Huang, H. and Bickel P.J. (2011) Measuring reproducibility of high-throughput experiments, Ann. Appl. Statist., 5, p. 1752–1779.

Mathelier, A. and Wasserman, W.W. (2013) The next generation of transcription factor binding site prediction, PLoS Comput Biol 9(9).

Portales-Casamar, E., Thongjuea, S., Kwon, A.T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W.W. and Sandelin, A. (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles, Nucleic Acids Res., 38: D105–D110.

Robasky, K. and Bulyk, M.L. (2011) UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions, Nucleic Acids Res., 39, D124-8.

Smeenk, L., van Heeringen, S.J., Koeppel, M., van Driel, M.A., Bartels, S.J.J., Akkers, R.C., Denissov, S., Stunnenberg, H.G. and Lohrum, M. (2008) Characterization of genome-wide p53-binding sites upon stress response, Nucleic Acids Res., 36, p.3639-3654.

Stormo, G.D., Schneider, T.D., Gold, L. and Ehrenfeucht, A. (1982) Use of the ‘Perceptron’ algorithm to distinguish translational initiation sites in E. coli, Nucleic Acids Res., 10, p.2997–3011.

Stormo, G.D (2013) Modeling the specificity of protein-dna interactions, Quantitative Biology 1, p.115–130.

Therrien, C.W. (1989) Decision estimation and classification: an introduction to pattern recognition and related topics, John Wiley and Sons.

Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T,W., Greven, M.C., Pierce, B.G., Dong, X., Kundaje, A., Cheng, Y., Rando, O.J., Birney, E., Myers, R.M., Noble, W.S., Snyder, M. and Weng, Z.. (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors, Genome Res., 22, p.1798-1812.

Wasserman L. (2004) All of Statistics: A Concise Course in Statistical Inference, New York, Springer, ISBN: 0-387-40272-1.

Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., Dicuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L.Y., Helmberg, W., Kapustin. Y., Khovayko, O., Landsman, D., Lipman, D.J., Madden, T.L, Maglott, D.R., Miller, V., Ostell, J., Pruitt, K.D., Schuler, G.D., Shumway, M., Sequeira, E., Sherry, S.T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R.L., Tatusova, T.A., Wagner, L. and Yaschenko, E. (2008) Database resources of the National Center for Biotechnology Information, Nucleic Acids Res., 36: D13-21.

Wilbanks, E.G and Facciotti, M.T. (2010) Evaluation of algorithm performance in ChIPseq peak detection, PLoS ONE, 5(7):e11471.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W. and Liu, X.S. (2008) Model-based analysis of ChIP-seq (MACS), Genome Biol., 9: R137.1-R137.9.