EuPA Open Proteomics 13 (2016) 14–23

Contents lists available at ScienceDirect

EuPA Open Proteomics

journal homepage: www.elsevier.com/locate/euprot

Prediction of protein-DNA interactions of transcription factors linking proteomics and transcriptomics data

Yu. Kondrakhin[a,b], T. Valeev[a,c], R. Sharipov[a], I. Yevshin[a], F. Kolpakov[a,c], A. Kel[a,d,e,*]

[a] Institute of Systems Biology, Ltd, Novosibirsk, Russia
[b] Design Technological Institute of Digital Techniques, SB RAS, Novosibirsk, Russia
[c] Institute of Informatics Systems, SB RAS, Novosibirsk, Russia
[d] geneXplain GmbH, Wolfenbuettel, Germany
[e] Institute of Chemical Biology and Fundamental Medicine, SBRAN, Novosibirsk, Russia

ARTICLE INFO

Article history:
Received 2 December 2015
Received in revised form 2 August 2016
Accepted 6 September 2016
Available online 15 September 2016

Keywords:
Protein-DNA interactions
Proteomics versus transcriptomics
Transcription factor binding site
ChIP-Seq
Position weight matrix approach
The ROC curve
Area under curve

ABSTRACT

We compared positional weight matrix-based prediction methods for transcription factor (TF) binding sites using selected fraction of ChIP-seq data with the help of partial AUC measure (limited to false positive rate 0.1, that is the most relevant for the application of the TF search in the genome scale). Comparison of three prediction methods—additive, multiplicative and information-vector based (MATCH) showed an advantage of the MATCH method for majority of transcription factors tested. We demonstrated that application of TF site identifying methods can help to connect the proteomics and phosphoproteomics world of signaling networks to gene regulation and transcriptomics world.

1. Introduction

Transcription factors (TFs) are proteins of crucial importance for regulation of all processes in human and other organisms. A rigorous classification of human transcription factors was published recently [1], summarizing many years of proteomics research attempting to understand the molecular mechanisms of functioning of transcription factors through their binding to DNA target sites and consecutive regulation of transcription of all genes in the human genome.

The poor correlation between proteomics and transcriptomics data is extensively discussed in proteomics literature [2]. Lack of such correlation making it extremely difficult to use high throughput and easy to generate transcriptomics data in understanding many cellular mechanisms acting mostly on protein level. Dynamic changes of abundance of proteins as well as changes of the status of their posttranslational modifications (such as phosphorylation of many regulatory proteins, including transcription factors) govern many biological processes. Direct measurements of such proteins and their modifications (often related to their activity) with the help of proteomics methods is very tedious, expensive and not always possible at all, often due to the lack of enough biological material necessary for proteomics and phosphoproteomics experiments.

Activity of such important proteins as transcription factors (TFs) can be estimated by their ability to bind DNA at their specific binding sites in genomes. TFs are often triggered in the cells by specific posttranslational modifications (phosphorylation), that enable TFs to bind to their specific sites at DNA. So, by measuring such interactions of TFs with DNA we can deduce activity status of these proteins. Such DNA-binding assay experiments can be combined sometimes with proteomics experiments measuring specific phosphorylation events that can give a lot of information to the researchers about exact mechanisms of acting of this class of proteins. Multiple cascades of phosphorylation and de-phosphorilation events happening in the cell signal transduction system leading to the activation of considered transcription factors. Therefore phosphoproteome data can be also combined with prediction of signal transduction pathways upstream of transcription factors to discover causative mechanism of acting of such transcription factors under particular signaling triggering cells to differentiation or to other cellular fate.

* Corresponding author at: Institute of Chemical Biology and Fundamental Medicine, SBRAN, Novosibirsk, Russia.
E-mail address: alexander.kel@biosoft.ru (A. Kel).

http://dx.doi.org/10.1016/j.euprot.2016.09.001

Since its introduction in 2007 [3], ChIP-Seq has become the most powerful experimental technique for genome-wide study of interactions between TFs and DNA. As a rule, a single ChIP-Seq experiment generates millions of short DNA 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 resulting 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) [4] and SISSRs (Site Identification from Short Sequence Reads) [5]. The reproducibility of nine peak detection algorithms including MACS and SISSRs was studied in [6] 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 correspondence profiles fitted by a copula model.

A comparative analysis of nine peak detection algorithms including MACS and SISSRs was performed in [7]. This comparison demonstrated that biological conclusions could change dramatically when the same raw ChIP-Seq dataset was processed using different algorithms. The results also indicated that the optimal choice of algorithm depends heavily on the selected dataset. Eleven different peak detection algorithms including MACS and SISSRs were also compared on common data sets [8]. This study offered a variety of ways to assess the performance of each algorithm and addressed the question 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.

The ChIP-Seq approach was designed as an experimental tool for identifying TF-binding regions in genome. 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 (5–15 bp). Second, some TF-binding regions are spurious due to the false positive rates of methods for read mapping and peak detection. Third, an unknown fraction of TF-binding regions should not contain the TF-binding sites because of tethered binding [9]. In this case, transcription factor bound to a 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 the PWM approach was introduced [10], it has become the most common and widely used for the computational analysis of TF-binding sites, see [11] for a review. A number of methods for the prediction of TF-binding sites have been developed within this approach. In particular, PWM algorithms were implemented in the computational tools such as MATCH [12], MatInspector [13], MATRIX SEARCH [14], ANN-Spec [15] and MEME [16]. There are several repositories that accumulate many matrices for the representation of TF-binding sites, in particular, TRANSFAC [17], JASPAR [18], Factorbook [19], UniPROBE [20] and HOCOMOCO [21]. Usually these matrices were derived from 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. The 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 [22,23]. It is a good way of visualizing the correspondence between sensitivity and false positive rate of a detection method. The area under the ROC curve, known as the AUC, is currently considered the standard measure to assess the accuracy of prediction methods, including those for the prediction of TF-binding sites. Currently it is common practice to reduce a comparison of different prediction methods to a comparison of the corresponding AUCs [24–26]. 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 the reasons mentioned above (including tethered binding).

We have developed an approach for reliable comparison of TFBS prediction methods under the condition that an unknown fraction of the ChIP-Seq data does not contain genuine TF-binding sites. In this article we have performed a comparative analysis of three existing PWM based methods, namely the common additive, common multiplicative methods, and the method that uses an information vector. We also vary two peak detection algorithms, MACS and SISSR. 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 a collection of non-redundant matrices from TRANSFAC (rel.2012.4). The analysis has revealed that all three methods perform rather similarly on the same sets of data. For the majority of PWMs the additive method gave slightly higher AUC values compared to the other two methods. Still both multiplicative and information vector based methods showed higher AUC values for some of the PWMs of the library. A comparison of the methods using partial AUC measure, which compare methods inside of their applicability domain, revealed that the information vector based method often outperforms other site search methods in the area of low false positive rate, whereas methods that don't use information vector are better for the area of parameter giving a low false negative rate. It is interesting to see that the general results obtained are invariant with respect to choice of peak detection algorithm despite dissimilarities between MACS and SISSRs that were revealed in this work.

Finally, to demonstrate the utility of the TF site prediction methods for proteomics research we combined the TF site analysis with phosphoproteomics and transcriptomics (RNA-seq) data (from PRIDE database) from the recently published experiment of treatment of MCF7 cell line with retinoic acid (RA) [27]. Promoters of differentially expressed genes (from RNA-seq analysis) were analyzed for TF-site frequency using the MATCH method following the approach published earlier [28]. Revealed overrepresented TF-sites indicate to us those transcription factors that are potentially activated (usually through phosphorylation of specific positions in the proteins) in the given cells under stimulation of the cells by RA. Next, we demonstrated that the revealed by this analysis transcription factors are connected to the network of signal transduction cascades identified by phosphoproteomics analysis of the cytoplasmic and nuclear fractions of those cells.

Therefor we can conclude that the methods of computational prediction of protein-DNA interactions of transcription factors that are described in this paper help researchers to find the missing link between the transcriptomics and proteomics (phosphoproteomics) data.

2. Materials and methods

2.1. Data

Human TF-binding region sets that were used in this study are stored in the GTRD database. GTRD collected raw ChIP-Seq data (sequenced reads) from literature, Gene Expression Omnibus (GEO), [29], Sequence Read Archive (SRA) [30], and the 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 the reference genome (build 37) using Bowtie release 1.1.1 [31], and the sets of the TF-binding regions were generated independently with the help of two peak detection algorithms, MACS release 1.4.2 and SISSRs version 1.4.

The transcriptomic and phosphoproteomic data of the experiment of treatment of MCF7 cell line with retinoic acid (RA) [27] were extracted from following data repositories: e RNA-seq data are available from the GEO institutional Data Access: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81814. The mass spectrometry proteomics data are available in PRIDE database with the dataset identifier PXD004357.

2.2. The ROC curves and AUCs as basis of comparison

According to common practice, the areas under the ROC curves are used in order to compare the site models. In turn, each ROC curve represents the correspondence between sensitivity of the model and false positive rate. In general, it is necessary to have a representative sample of genuine TF-binding sites in order to calculate the sensitivity. ChIP-seq derived TF-binding regions can be used for this purpose. It is assumed that TF-binding regions revealed by ChIP-seq experiments contain genuine TF-binding sites. Therefore the sensitivity was computed as a relative number of the TF-binding regions containing one or more TF-binding sites predicted. The false positive rate was computed on the basis of artificially generated sequences with the help of 10-fold permutations of nucleotides in each TF-binding region. The false positive rate was determined then as the relative number of such artificially generated sequences containing one or more TF-binding sites predicted. For AUC calculation we used the sets of the TF-binding regions that are stored in GTRD.

2.3. Scheme of site model comparison

According to common practice, the comparison of site models is reduced to a comparison of AUCs. In turn, AUCs are calculated on the sets of the TF-binding regions. However, the direct use of the full sets of TF-binding region for the AUCs calculation does not seem advisable because some TF-binding regions can be “empty”, i.e. they do not contain genuine TF-binding sites. To model such a situation we introduced a parameter τ, which defines a percentage of TF-binding regions that are not “empty” and contain at least one genuine TF-binding site. The following scheme of site model comparison takes into account the assumption about the existence of empty TF-binding regions.

First, we prepare the sets of TF-regions in such a way that all regions had the same length. If the TF-binding regions are longer then 200 bp, we redefine them as regions of the lengths 200 bp with the centers in summits of distributions of the number of matched reads. If the TF-binding regions are shorter then 200 bp, we extend them to the total length 200 bp adding respective flanks.

In 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 the fragment of the TF-binding region where the site model obtained the maximal score among all scores calculated for every possible fragments of the TF-binding region. Then, for each site model, a top list of the τ percent (τ is given) of the ‘best sites’ with the highest scores is constructed and the so-called τ-union of the ‘best sites’ is composed as a union of all such top lists for all three site models considered in the study. Then, the so-called the τ-union of the TF-binding regions is defined as the merged union of such TF-binding regions that contained at least one ‘best site’ from τ-union of the ‘best sites’. Finally, the ROC curves are generated on the τ-union of the TF-binding regions and the corresponding AUC values are calculated.

2.4. Implementation

The proposed approach for comparing the TF site prediction methods was implemented with the help of the open source BioUML platform (http://biouml.org/). We have created the following Java modules:

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

The ‘ROC curves for best sites union’ module generates the ROC curves and calculates the corresponding AUCs for the user-selected set of site models when the value of parameter τ ($1 \leq \tau \leq 100$) and the set of the TF-binding regions are specified. The user interface allows for selecting the site model (additive model, multiplicative model or information vector based model, see Site model section above for details). The resulting ROC curves and corresponding AUCs are computed by the java modules and are stored within a user-specified folder in the platform.

The ‘Summary on AUCs’ tool performs a comparative analysis of site models when the value of the parameter τ is pre-specified. Initially all appropriate AUC values calculated by the ‘ROC curves for best sites union’ tool are read in all available tables. Then a comparison of AUC values is performed with the help of the non-parametrical Friedman and Wilcoxon signed rank tests [32]. In the case of the Friedman test, a chi-squared distribution with (k-1) degrees of freedom is used for assessing the statistical significance of the difference between AUCs, where k denotes the number of site models. In the case of the 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 the kernel density estimator [33] with Epanechnikov kernel and are plotted for the user.

The ‘Peak finders comparison’ tool performs a comparative analysis of two peak detection algorithms. To compare two peak detection algorithms, this tool carries out a 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 the 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 the 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 a summit is determined by MACS as the precise binding location within a given TF-binding region. The probability density is estimated by the kernel density estimator with Epanechnikov kernel.

2.5. Three site models available for comparative analysis

Currently, three site models that represent PWM approach are available for comparative analysis. For a given TF they share the same position frequency matrix MAT $= (m_{ij})$, $i = \{A,C,G,T\}$, $j = 1, \ldots, l$ but produce diverse scores for a fixed DNA fragment $S = (s_1, \ldots,$

s_l). In other words, the models represent different scoring algorithms.

2.5.1. Additive model

This model calculates the common additive score x defined by the formula

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

where the values score(j), $j = 1, \ldots, l$, 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 $s_j = T\}$.

2.5.2. Multiplicative model

For a fragment S this model calculates the common multiplicative score x_m

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

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

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

where the values score*(j), $j = 1, \ldots, l$, 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 a logarithm of zero we preliminarily found minimal a 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.

2.5.3. Information vector-based model (MATCH model)

This model is determined by the popular PWM method MATCH for TF-binding site prediction. This model calculates the so-called matrix similarity score *mSS* defined in [12]. Actually, this model is a common additive model, which uses a transformed matrix instead of an initial matrix, where each column of the transformed matrix was determined with the help of weighting the corresponding initial column by information content. More specifically, the j-th column of the weight matrix is equal (up to the constant (*–Min*/(*Max-Min*))) to the product of the j-th column of the frequency matrix and the value *I*(*j*)/(*Max-Min*), $j = 1, \ldots, l$, where I(*j*), *Min*, and *Max* were defined in [12].

2.6. Software availability

The site search algorithms described in this paper are available for free in BioUML/geneXplain platform. The anonymous access to the platform is available here:

http://gtrd.biouml.org/bioumlweb/#anonymous=true

Individualized access to the platform with secure space for your data is available for free upon registration at the URL:

http://www.genexplain.com/genexplain-platform-registration

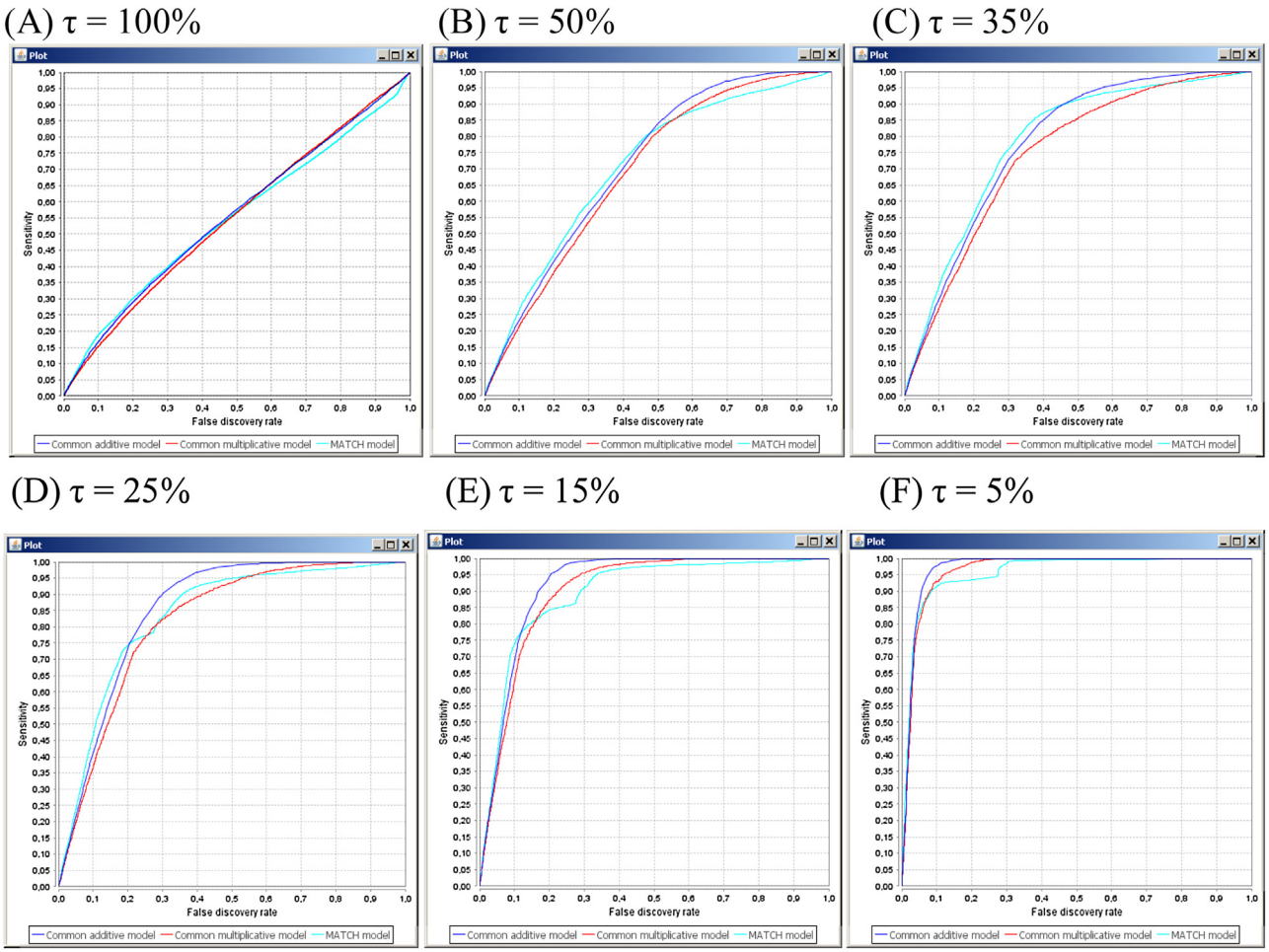

Fig. 1. The ROC curves obtained for different values of τ on the YY1-binding regions that were generated by MACS peak detection algorithm. Dark blue lines correspond to the additive model, red lines to the multiplicative model, and light blue lines to the information vector based model (MATCH model).

3. Results

3.1. Selection of τ-union parameter

The key step of the proposed scheme of AUCs calculation is the construction of the τ-union of the TF-binding regions, where the percentage τ is a free parameter. In general, the following relationship exists between τ values and the shapes of the ROC curves: the smaller the percentage τ, the more convex the ROC curve is, and the higher the AUC values are. 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%. An example is shown in Fig. 1, where the ROC curves were generated on the YY1-binding regions (processed by MACS). In turn the corresponding values of AUCs are close to 0.5 when τ tends to 100%, while these values are close to 1.0 when τ tends to 5%, see Table 1.

It is important to note that the shown relationship between τ and the shape of the ROC curve can be interpreted as follows. According to the definition of the τ-union of TF-binding regions, it consists of those TF-binding regions that contain the 'best sites' with the highest scores. In other words, the TF-binding regions that contain TF sites with the smallest scores only are removed from further analysis (so-called "empty" regions). Obviously, the higher the percentage τ, the smaller the number of regions that are classified as empty, see also the first and the last columns of Table 1.

3.2. Comparative analysis of three site models

We performed a comparative analysis of the following three site models that represent the PWM approach: additive model, multiplicative model and information vector based model (MATCH model). For this analysis we have selected 266 TFs that have got 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 the USF1-binding region set in GTRD, we did not include it in the analysis, because there is no appropriate matrix for the USF1-binding sites in TRANSFAC that contains the matrices V USF_01, VUSF_02, VUSF_C, VUSF_Q6 and V$USF_Q6_01 derived for the sites of the USF family.

A comparative analysis was performed independently on 265 sets of TF-binding regions generated by MACS, and on 263 sets generated by SISSRs. In the case of SISSRs we excluded 2 sets from our analysis because of their small sizes (<200). TF-binding regions produced by MACS and SISSRs were trimmed or enlarged to 200 bp according to the procedure described in the Method Section.

For the first comparative analysis we have considered the following five values of τ: 100%, 35%, 25%, 15% and 5%. We computed AUC values for each of three site models applied to each of the TF-binding region set with five values of τ. Next, we compared results generated by three site models with the help of the Friedman test which compares distributions of generated AUC values by applying the site models to all analyzed PWMs. We used a Chi-squared distribution with two degrees of freedom for assessing the significance of differences between three site models (see Table 2). As one can see from the result of this comparison, the three site models produce statistically significantly different results. This difference increases with the increase of τ. Therefore it is important to understand which site model is the method of choice in the further analysis of biological data.

Table 2
Comparison of three site models with the help of Friedman test using two peak detection algorithms. P-value show the statistical significance of the value of the Friedman test statistic showing global difference of the distributions of AUCs for 265 (for MACS) and 263 (for SISSRs) TF-binding ChIP-seq data.

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	$<10^{-12}$

As the next step, we performed a more detailed comparative analysis of the generated ROC curves and AUC values in order to understand which site models are preferable for each of the PWMs and under which conditions. Here, we choose a value of τ equal to 25%, since most of the site models give reasonably high values of AUC for all of the PWMs (0.7–0.9).

A more detailed consideration of computed ROC curves for each PWM shows that, in fact, the actual difference of AUC values for different matrices is relatively small. This conclusion is invariant with respect to the choice of peak detection algorithm. So, with a few exceptions, we can say that although the general difference of the performance of all three site models for all PWMs altogether is statistically significant, the absolute values of the differences of AUC for each individual PWM are quite small. In Table A1 in the Appendix (see Supplementary data), we provide all values of AUC. We also indicate which site model gives better AUC for each PWM. Also, we annotate this table with the name of the TF antibody which was used in each ChIP-seq experiment, the cell line, the classification of TF according to their DNA binding domain using the classification of human transcription factors [1]. We also computed and presented in the table the total and average entropy and the length of each PWM.

Next, we computed several partial-AUC values for each of the PWMs. This means that we summed up the areas under the ROC curve for particular ranges of FP and FN values only. The reason of computing a partial AUC is well described previously [34,35]. Such a partial-AUC attempts to estimate the performance of the recognition method in the area of true positive and false positive rates that are actually applied in the data analysis in the majority of cases. We considered the two most frequent use cases of the

Table 1
AUCs calculated for different values of τ on the YY1-binding regions that were generated by MACS peak detection algorithm.

	Site model			Percentage of regions that are classified as "empty"
Percentage, τ	Information vector based model	Multiplicative model	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

Table 3
Results of comparison of three site model methods applied to the TRANSFAC PWMs on respective ChIP-seq data sets. Three measures of site recognition methods were applied—full AUC and two partial-AUCs. We computed the number of PWMs that gives maximal value of the measure (full AUC or partial-AUC) for the given site model. The last row gives the number of PWMs when all three methods produced equal values for the respective measure. In bold we indicate a method that gives the highest number of PWMs with maximal AUC_FP0.1 criteria.

A) MACS

Site model method.	Number of PWMs with maximal AUC	Number of PWMs with maximal partial AUC_TP0.8	Number of PWMs with maximal partial AUC_FP0.1
Additive	152	154	40
Multiplicative	61	58	92
MATCH	42	43	**113**
All three methods give the same AUC value	10	10	20

B) SISSRs

Method name	Number of PWMs with maximal AUC	Number of PWMs with maximal partial AUC_TP0.8	Number of PWMs with maximal partial AUC_FP0.1
Additive	138	134	45
Multiplicative	62	65	85
MATCH	52	53	**107**
All three methods give the same AUC value	11	11	26

application of site recognition models. The first one corresponds to values of false positive rates equal or lower then 0.1. It applies to all potential searches of TF binding sites in full genomes, or at least in relatively long genomic regions. It is reasonable to assume that in such an application of the site search method it makes no sense to allow for false positive rate higher then 0.1 (which means on average one site prediction in every tenth position). Normally much lower false positive rates are used in such genome scanning methods to minimize potential huge noise. The second use case corresponds to the values of true positive rate higher or equal to 0.8. It applies to those rather rare use cases when one should not miss practically any of the true sites irrespectively of how many false positives it also finds. Such site searches are applied in cases of analysis of relatively short genome regions (e.g. one individual promoter or enhancer), with consideration of further validation of all found sites by independent experimental or computational methods (for instance, by cross-species comparison [36] or by analysis of site combinations [37]).

We compared the performance of three site models using three measures—AUC, partial-AUC_TP0.8 (which corresponds to the area under the ROC curve of true positive rates higher or equal to 0.8) and partial-AUC_FP0.1 (which corresponds to the area under the ROC curve of false positive rates lower or equal to 0.1) (see Table 3). It is interesting to see that depending on the measure we get rather different results. In case of the application of full AUC, the highest value is provided by the additive site model method for most of the PWMs. The partial-AUC_FP0.1 however gives a completely different picture. For most of the PWMs, the highest values are provided by the information-vector based site model method (MATCH method). Application of partial-AUC_TP0.8 gives a very similar result to the full AUC.

3.3. Application of TF site prediction models to link transcriptomics and phosphoproteomics data

In order to demonstrate the usefulness of the described TF site prediction methods for proteomics research we jointly analyzed phosphoproteomics (from PRIDE database) and transcriptomics (RNA-seq) data from recently published experiment of treatment of MCF7 cell line with retinoic acid (RA) [27]. Since the change of expression of the genes measured by transcriptomics upon treatment by RA must be clearly dependent on the changes of activity of transcription factors we, first of all, analyzed promoters of differentially expressed genes for TF-site frequency using the MATCH method following the approach published earlier [28]. Here we used MATCH models described in the current paper as most specific for the given type of analysis of multiple promoter sequences. Revealed overrepresented TF-sites in promoters of differentially expressed genes in comparison to the promoters of genes with no change of expression indicated to us those transcription factors that are potentially activated or inhibited (usually through phosphorylation of specific positions in their protein sequence) in the given cells under stimulation of the cells by RA. (see Table 3).

In the next step we applied graph algorithms described earlier [28] in order to identify potential common regulators of the activity of predicted set of transcription factors in the signaling network of the cells under study. Statistical significance of such common regulators is confirmed by random shuffling of the input TF lists. Among such common regulators we expect to find protein kinases and other components of signal transduction cascades that can phosphorylate multiple transcription factors or other intermediate signaling molecules and therefore play a role as such common regulators of the activity of the set of TFs under study. In turn, an indicator of activity of such protein kinases often could be their phosphorylation status which is measured in the phosphoproteomics experiments. So, we were interested to find links between the signal transduction proteins detected by phosphoproteomics measurements in the cytoplasm or in the nucleus of the cells and the TFs predicted by our promoter analysis. Indeed, we confirmed such links between identified common regulators and phosphoproteomics measurements. (see Table 4). One can see that almost all found common regulators (9 out of 11) have been identified by the phosphoroteomics experiment (Table 5).

On Fig. 2. we show the diagram that connects two most significant common regulators (light red nodes at the top of the diagram) and TFs (light blue nodes in the middle and at the bottom of the diagram) whose sites found overrepresented in the promoters of differentially expressed genes. With red, blue and gray decoration of several nodes in the diagram we annotate the phosphorylation of the respective proteins detected in the phosphoproteomics experiment. The left part of the decoration circle corresponds to the protein phosporialytion observed in the cytoplasm of the cells and the right side corresponds to the protein phosphorylation observed in the nucleus. The red color corresponds to the increased level of phosphorialtion after treatment of

Table 4
Transcription factors found by the combined analysis of transcriptomics and phosphoprotyomics data. With the help of MATCH algorithm we identified overrepresented TF binding sites in promoters of differentially expressed genes (DEG) (from transcriptomics data). TRANSPAC PWM—name of the position weight matrix from TRANSFAC database which was used by MATCH; Yes-No ratio—the ratio of TF site frequency in promoters of DEG compared to the promoters of non-changed genes; p-value—statistical significance of the Yes-No ratio; Phospho Cytoplasm/Nucleus—detection of the phosphorypation of the TF in cytoplasm or in nucleus of the cells (p- phosphorylation was detected, p-up—phosphorylation was found increased upon treatment by RA, p-dn—decreased by RA).

Gene symbol	TF name	TRANSFAC PWM	Yes-No ratio	P-value	UniProt ID	Phospho Cytoplasm	Phospho Nucleus	Gene description
RELA	RelA-p65	V$RELA_Q6	1.22	2.78E-04	Q04206	p	p	v-rel reticuloendotheliosis viral oncogene homolog A (avian)
RXRA	RXR-alpha	V$DR4_Q2	1.34	8.36E-15	P19793	p	p	retinoid X receptor, alpha
SP1	Sp1	V$SP1_Q6_01	2.37	1.36E-85	P08047	p	p-dn	Sp1 transcription factor
CTCF	ctcf	V$CTCF_01	1.71	1.75E-16	P49711	p	p	CCCTC-binding factor (zinc finger protein)
RXRB	RXR-beta	V$DR4_Q2	1.34	8.36E-15	P28702	p	p	retinoid X receptor, beta
TRIM28	RNF96	V$RNF96_01	2.54	6.71E-43	Q13263	p-up	p-dn	tripartite motif containing 28
NFYC	NF-YC	V$NFY_Q3	1.67	1.16E-04	Q13952	p	p	nuclear transcription factor Y, gamma
SP3	Sp3	V$SP1_Q6_01	2.37	1.36E-85	Q02447	p	p	Sp3 transcription factor
RREB1	RREB-1	V$RREB1_01	1.33	1.28E-12	Q92766	p	p-dn	ras responsive element binding protein 1
NR2F2	COUP-TF2	V$DR4_Q2	1.34	8.36E-15	P24468	p	p	nuclear receptor subfamily 2, group F, member 2
KLF4	GKLF	V$GKLF_Q4	1.63	4.06E-135	O43474	p	p-dn	Kruppel-like factor 4 (gut)
PATZ1	PATZ	V$MAZR_01	2.14	1.90E-11	Q9HBE1	p	p	POZ (BTB) and AT hook containing zinc finger 1

Table 5
Statistically significant common regulators found by the graph algorithm of the geneXplain platform (www.genexplain.com) by searching upstream of TFs listed in Table 5 in the signal transduction network of TRANSPATH database [38]. TF-reached—number of TFs (out of 12 from Table 5) that are reached in the network downstream from the respective common regulator; Score—score of the common regulator calculated on the basis of the number of reached TFs and topology of the network [28]; FDR and Z-score are calculated by multiple randomization of input set of TFs [28].(FDR <0.05 AND Z-Score > 1.0 AND TF-reached > 7).

TRANSPATH ID	Name of common regulator	TF reached	Score	FDR	Z-Score	Phospho Cytoplasm	Phospho Nucleus
MO0000056714	HDAC1	8	0.623	0.036	1.031	p-up	p-up
MO0000257368	SUSP1	8	0.555	0.031	1.354	p	p-dn
MO0000103308	CKI-gamma1	8	0.530	0.035	1.093		
MO0000019363	RelA-p65	7	0.484	0.030	1.679	p	P
MO0000132731	PP4C	7	0.445	0.047	1.068	p	P
MO0000140900	ing4	8	0.434	0.050	1.613		
MO0000272358	ctcf{sumo}	7	0.390	0.035	1.455	p	P
MO0000284804	RNF96{p}	7	0.341	0.047	1.590	p-up	p-dn
MO0000107711	RXR-alpha{sumo}	8	0.337	0.033	1.549	p	P
MO0000272357	ctcf{sumo}	7	0.275	0.049	1.564	p	P
MO0000284833	RNF96{pS473}{pS824}	7	0.250	0.040	1.833	p-up	p-dn

cells by RA, blue color corresponds to decreased level and gray—the same level of phosphorylation of these proteins after the RA treatment.

We can show here that such important signaling proteins as “histone deacetylase 1 (HDAC1)”, whose level of phosphorylation is rapidly increasing after treatment of the cells by RA, and “SUMO-1-specific protease 1 (SUSP1)”, whose level of phosphorylation is high and stable in the cytoplasm and decreasing in the nucleus, are involved in this cellular system in triggering signal transduction pathways towards activity of particular transcription factors. Among them there are the number of important transcription factors such as RelA, Sp-1, RXR, CTCF, GKLF, RNF96 that are characterized by the high and often changing level of phosphorylation in cytoplasm and especially important, in nucleus and evidently as a result of such signal transduction cascade changing their activity during RA treatment and consequently up-regulating expression of their target genes. It was also interesting to see that HDAC1 was actually one of the top proteins whose phosphorylation status most significantly increased after RA treatment (11 additional phosphopeptides detected in nucleus after the treatment by RA). And it was also independently identified as the top common regulator in our analysis.

Therefor we can conclude that the methods of computational prediction of protein-DNA interactions of transcription factors that are described in this paper help researchers to find the missing link between the transcriptomics and proteomics (phosphoproteomics) data.

4. Discussion

Currently the AUC values are considered the standard measures to assess the predictive abilities of site models. Certainly, for an accurate calculation of precise AUCs it is necessary to have representative samples of genuine TF-binding sites. Available TF-binding regions from ChIP-seq experiments processed by peak calling algorithms provide a good resource for such computations. But the direct use of the raw initial sets of the TF-binding regions for the AUC calculations is not reasonable because many of the TF-binding regions can be “empty” (not actually containing genuine TF-binding sites) mainly due to various experimental and data pre-

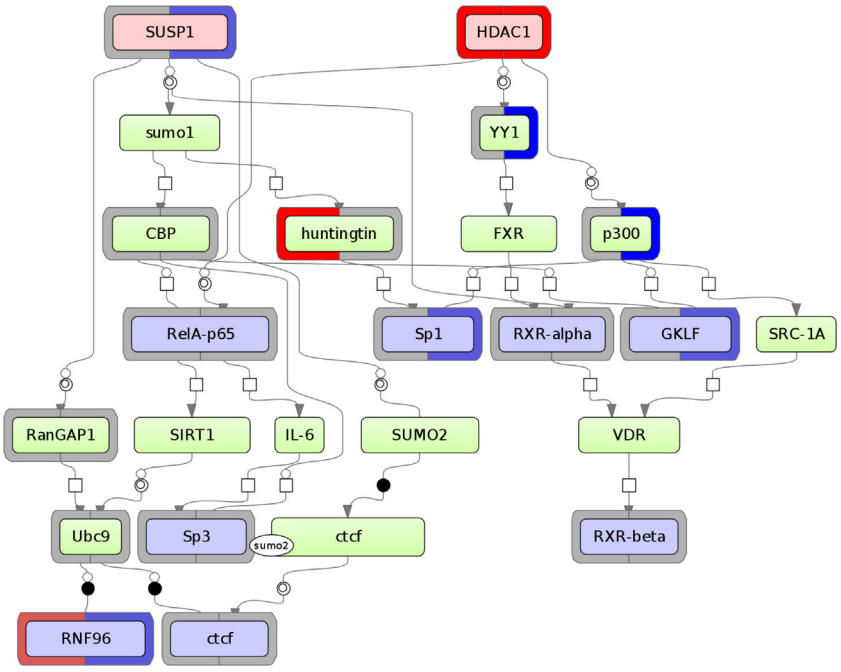

Fig. 2. Signal transduction diagram that connects two most significant common regulators (light red nodes at the top of the diagram) and TFs (light blue nodes in the middle and at the bottom of the diagram) whose sites found overrepresented in the promoters of differentially expressed genes. Red, blue and gray decoration of the odes in the diagram annotates the phosphorylation of the respective proteins detected in the phosphoproteomics experiment. The left part of the decoration circle corresponds to the protein phosporialytion observed in the cytoplasm of the cells and the right side corresponds to the protein phosphorylation observed in the nucleus. The red color of the decoration corresponds to the increased level of phosphorialtion after treatment of cells by RA, blue color corresponds to decreased level and gray – the same level of phosphorylation of these proteins after the RA treatment.

processing uncertainties discussed above. Indeed, it turned out that when taking full sets of ChIP-seq TF-regions for the majority of the selected TFs, the values of the computed AUCs of all applied PWM-based methods were close to 0.5 (see Table A2 Supplementary data), and the shapes of the ROC curves were approximately linear (see, for instance, Figure A3 Supplementary data).

It becomes clear that such sets of sequences are not directly suitable as an ideal set for the comparison of different TF-site recognition algorithms. In this paper we have suggested the τ-union approach for selecting subsets of TF-binding regions suitable for the sheer purpose of comparing the performance of different site models to each other. Of course this does not guarantee the selection of all true TF-binding sites out of the initial sets of TF-binding regions. This method just provides a platform for a relatively unbiased comparison of different methods for TF-site recognition.

Certainly, the construction of the τ-union of the TF-binding regions is just one of several possible ways to compose refined sets of TF-binding regions that can be used for site model comparison. One of the alternative ways to compose refined sets is to select the most "reliable" TF-binding regions according to external characteristics obtained in the ChIP-seq data preprocessing. We demonstrated (see Appendix 4.4 Supplementary data) that the use of such external characteristics coming from the peak detection algorithm, as 'FDR', 'Fold enrichment', 'Tag number', 'Score' and 'p-value', does not actually provide suitable platform for comparing TF site prediction methods.

As has been described in detail in the Method Section, the τ-union approach allows for preparing subsets of TF-binding regions that contain an unbiased mixture of DNA motifs for TF-binding sites as they are recognized by different PWM site models. This way we create a good platform for comparing different site models to each other using the same set of sequences, which makes such a comparison most objective and unbiased. At the same time, such a comparison is done on the basis of natural genomic sequences, experimentally shown to be bound by the given transcription factors (directly or indirectly), rather then on the basis of some artificially prepared sequences as has been done elsewhere. This provides a higher reliability of such a comparison of methods and a better basis for choosing the method for a real analysis of genomic sequences.

The final comparison of PWM site models was done on the τ-union sets of TF-binding regions with a relatively low value of τ equal of 25%. This means that only about 25% of TF binding regions obtained from the ChIP-seq experiments were used for such a comparison. Our choice of this value was based on the average values obtained of the AUCs for most of the PWM site models (see Table A1 in the Appendix Supplementary data), which were mainly above 0.7 (with some small exceptions); this is considered to be a borderline for relatively good quality for a diagnostic test [39].

The use of the AUC value for comparing the precision of different recognition methods and diagnostics tests is well accepted in the machine learning community [39], and is widely used for comparison of various bioinformatics methods including TF site recognition methods [40]. However, this practice has recently been questioned [41,42]. Certain important parameters should be carefully taken into account when applying AUC for comparison of different recognition methods. When comparing two methods by their ROC curves problems arise when the interest does not lie in the entire range of false-positive rates. Often in bioinformatics and other applications it is more useful to look at a specific region of the ROC curve rather than at the whole curve. To overcome these difficulties the approach of computing partial AUC has been proposed earlier [34,35]. In this approach one focuses for instance on the low false positive rates only, which is often of prime interest for population or genome screening tests, and calculates the value of "partial AUC" by calculating the area under the ROC curve only in the respective part of the curve [34,35].

In our work we applied two partial AUCs that correspond to two of the most frequently used cases of applying TF-site recognition methods. In the first case, we compute the area under the ROC curve only in the region of false positive rates from 0.0 to 0.1. In this way we focus our attention on the cases of TF binding sites searches in full genomes or at least in relatively long genomic regions. We assume that in such applications of full genome screening it makes no sense to allow false positive rate higher then 0.1. Otherwise the results will be flooded with millions of false positive hits and will become useless in practical applications. In the second use case we focus our attention on the alternative part of the scale when the values of true positive rate should be higher of equal to 0.8. Such use cases correspond to the TF-site analysis in relatively short genome regions (e.g. in an individual promoter or enhancer) when one should minimize the loss of real sites. We implemented two measures of partial AUC—"partial AUC_FP0.1" and "partial AUC_TP0.8", respectively.

Using these partial AUC measures as well as traditional full AUC we compared the efficiency of three different PWM-based site models for recognition of binding sites for more then 260 different human transcription factors. Such a full-scale comparison has not been done so far. Our results provide a basis for the choice of the TF site identification methods for various future applications.

In order to find a rationale for the higher performance of a certain PWM-based site model for recognition of sites for different transcription factors we compared the results of AUC calculations with various characteristics of transcription factors and their respective PWMs. In Table A1 in the Appendix (see Supplementary data) we summarized several characteristics, including: TF classification index [1], name of the TF antibody and cell line used in the respective ChIP-seq experiments, the length of PWM, mean and sum entropy of the PWM. Our attempts to find any correlation between those characteristics and the performance of one of the tested TF-site model failed. For instance, no significant difference was found while comparing the average entropy of those PWMs that showed superior results for "additive site model" with the average entropy of PWMs showing superior results for the "site model based on information vector". Also, it was interesting to observe that even for very similar transcription factors belonging to one family, different family members can display absolutely different preferences to one or another TF site model. For instance, the FOX family of transcription factors is characterized by very similar PWMs. Although for most of the family members the highest values of full AUC correspond to the additive site model, for the factor FoxM1 the highest value was achieved by the multiplicative site model, and for the factor FoxO4 it was taken by the site model based on information vector.

Generally the application of the full AUC measure gives the highest values for the "additive site model method" for most of the tested PWMs. Still our results show that for the actual most frequent applications of the PWM method, e.g. in the use cases of searches of TF sites in long genomic sequences, the supreme method is the site model which is based on information vector (which is implemented in the popular MATCH algorithm [12]), since it gives the higher values of the respective partial AUC.

Therefore, in this paper we successfully applied a novel unified method for comparing different approaches of computing TF site models based on PWM.

Finally, to demonstrate the utility of the TF site prediction methods for proteomics research we combined the TF site analysis with phosphoproteomics and transcriptomics data. We analysed promoters of the differentially expressed genes (from RNA-seq) using the MATCH site prediction method and predicted those transcription factors that are potentially activated in these cells. Next, using graph analysis algorithm we connected these transcription factors to the network of signal transduction cascades identified by phosphoproteomics analysis of the cytoplasmic and nuclear fractions of those cells. This example of analysis of two "-omics" datasets allowed us to conclude that the methods of computational prediction of protein-DNA interactions of transcription factors that are described in this paper can indeed help researchers to find the missing link between the transcriptomics and proteomics (phosphoproteomics) data.

We hope that our results will contribute to an improvement of efficiency in the application of computational methods for understanding the molecular mechanisms of functioning of such an important group of proteins as transcription factors and will contribute to the growing field of proteomics research.

Conflict of interest

None.

Acknowledgements

This work was supported by a grant of the Federal Targeted Program "Research and development on priority directions of science and technology in Russia, 2014-2010", grant number: 14.604.21.0101 to the Institute of Chemical Biology and Fundamental Medicine, SBRAS.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.euprot.2016.09.001.

References

[1] E. Wingender, T. Schoeps, M. Haubrock, J. Dönitz, TFClass: a classification of human transcription factors and their rodent orthologs, Nucleic Acids Res. 43 (2015) D97–102, doi:http://dx.doi.org/10.1093/nar/gku1064.
[2] G. Chen, T.G. Gharib, C.C. Huang, J.M. Taylor, D.E. Misek, S.L. Kardia, T.J. Giordano, M.D. Iannettoni, M.B. Orringer, S.M. Hanash, D.G. Beer, Discordant protein and mrna expression in lung adenocarcinomas, Mol. Cell. Proteomics 1 (4) (2002) 304–313.
[3] D.S. Johnson, A. Mortazavi, R.M. Myers, B. Wold, Genome-wide mapping of in vivo protein-DNA interactions, Science 316 (2007) 1497–1502.
[4] Y. Zhang, T. Liu, C.A. Meyer, J. Eeckhoute, D.S. Johnson, B.E. Bernstein, C. Nussbaum, R.M. Myers, M. Brown, W. Li, X.S. Liu, Model-based analysis of ChIP-seq (MACS), Genome Biol. 9 (1) (2008) R137.1–R137.9.
[5] R. Jothi, S. Cuddapah, A. Barski, K. Cui, K. Zhao, Genome-wide identification of in vivo protein–DNA binding sites from ChIP-seq data, Nucleic Acids Res. 36 (2008) 5221–5231.
[6] Q. Li, J.B. Brown, H. Huang, P.J. Bickel, Measuring reproducibility of high-throughput experiments, Ann. Appl. Statist. 5 (2011) 1752–1779.
[7] T.D. Laajala, S. Raghav, S. Tuomela, R. Lahesmaa, T. Aittokallio, L.L. Elo, A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments, BMC Genomics 18 (December) (2009), doi: http://dx.doi.org/10.1186/1471-2164-10-618 (10:618).
[8] E.G. Wilbanks, M.T. Facciotti, Evaluation of algorithm performance in ChIPseq peak detection, PLoS One 5 (7) (2010) e11471.
[9] J. Wang, J. Zhuang, S. Iyer, X. Lin, T. Whitfield, 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. Weng, Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors, Genome Res. 22 (2012) 1798–1812.
[10] G.D. Stormo, T.D. Schneider, L. Gold, A. Ehrenfeucht, Use of the 'perceptron' algorithm to distinguish translational initiation sites in E. coli, Nucleic Acids Res. 10 (1982) 2997–3011.
[11] G.D. Stormo, Modeling the specificity of protein-dna interactions, Quant. Biol. 1 (2013) 115–130.
[12] A.E. Kel, E. Gossling, I. Reuter, E. Cheremushkin, O.V. Kel-Margoulis, E. Wingender, MATCH™: a tool for searching transcription factor binding sites in DNA sequences, Nucleic Acids Res. 31 (2003) 3576–3579.
[13] K. Quandt, K. Frech, H. Karas, E. Wingender, T. Werner, MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data, Nucleic Acids Res. 11–12 (23) (1995) 4878–4884 (PMID:8532532).
[14] K. Chen Q, Z. Hertz G, D. Stormo G, s. author, MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices, Comput. Appl. Biosci. 5 (11) (1995) 563–566 (PMID:8590181).

[15] T. Workman C, D. Stormo G, ANN-Spec: a method for discovering transcription factor binding sites with improved specificity, Pac Symp. Biocomput. (2000) 467–478 (PMID:10902194).
[16] T.L. Bailey, N. Williams, C. Misleh, W.W. Li, MEME: discovering and analyzing DNA and protein sequence motifs, Nucleic Acids Res. 34 (Web Server issue) (2006) 1–7, doi:http://dx.doi.org/10.1093/nar/gkl198 (PMID:16845028).
[17] V. Matys, O.V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A.E. Kel, E. Wingender, TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes, Nucleic Acids Res. 1–1 (Database issue) (2006) 34, doi:http://dx.doi.org/10.1093/nar/gkj143 (PMID:16381825).
[18] Elodie Portales-Casamar, Supat Thongjuea, Andrew T. Kwon, David Arenillas, Xiaobie Zhao, Eivind Valen, Dimas Yusuf, Boris Lenhard, Wyeth W. Wasserman, Albin Sandelin, JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles, Nucleic Acids Res. 11–11 (Database issue) (2009) 38, doi:http://dx.doi.org/10.1093/nar/gkp950 (PMID:19906716).
[19] Jie Wang, Jiali Zhuang, Sowmya Iyer, Xin Ying Lin, Troy W. Whitfield, Melissa C. Greven, Brian G. Pierce, Xianjun Dong, Anshul Kundaje, Yong Cheng, Oliver J. Rando, Ewan Birney, Richard M. Myers, Williams S. Noble, Michael Snyder, Zhiping Weng, Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors, Genome Res. 2 (9) (2012) 1798–1812, doi:http://dx.doi.org/10.1101/gr.139105.112 (PMID:22955990).
[20] Kimberly Robasky, Martha L. Bulyk, UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions, Nucleic Acids Res. 30–10 (Database issue) (2010) 39, doi:http://dx.doi.org/10.1093/nar/gkq992 (PMID:21037262).
[21] Ivan V. Kulakovskiy, Yulia A. Medvedeva, Ulf Schaefer, Artem S. Kasianov, Ilya E. Vorontsov, Vladimir B. Bajic, Vsevolod J. Makeev, HOCOMOCO: a comprehensive collection of human transcription factor binding sites models, Nucleic Acids Res. 21–11 (2012) 41, doi:http://dx.doi.org/10.1093/nar/gks1089 (PMID:23175603).
[22] K. Fukunaga, Introduction to Statistical Pattern Recognition, 2nd edition, Academic Press, San Diego, 1990.
[23] C.W. Therrien, Decision Estimation and Classification: An Introduction to Pattern Recognition and Related Topics, John Wiley and Sons, 1989.
[24] A. Mathelier, W.W. Wasserman, The next generation of transcription factor binding site prediction, PLoS Comput. Biol. 5–9 (9) (2013) 9, doi:http://dx.doi.org/10.1371/journal.pcbi.1003214 (PMID:24039567).
[25] L. Smeenk, S.J. van Heeringen, M. Koeppel, M.A. Driel, S.J.J. van Bartels, R.C. Akkers, S. Denissov, H.G. Stunnenberg, M. Lohrum, Characterization of genome-wide p53-binding sites upon stress response, Nucleic Acids Res. 28–5 (11) (2008) 3639–3654, doi:http://dx.doi.org/10.1093/nar/gkn232 (ISSN: 0305-1048).
[26] D. Alamanova, P. Stegmaier, A. Kel, Creating PWMs of transcription factors using 3D structure-based computation of protein-DNA free binding energies, BMC Bioinf. 11 (1) (2010) 225, doi:http://dx.doi.org/10.1186/1471-2105-11-225 (ISSN: 1471-2105).
[27] M. Carrier, M. Joint, R. Lutzing, A. Page, C. Rochette-Egly, Phosphoproteome and transcriptome of RA-responsive and RA-resistant breast cancer cell lines, PLoS One 11 (6) (2016) e0157290, doi:http://dx.doi.org/10.1371/journal.pone.0157290.
[28] A. Kel, N. Voss, R. Jauregui, O. Kel-Margoulis, E. Wingender, Beyond microarrays: find key transcription factors controlling signal transduction pathways, BMC Bioinf. 6–7 (September (Suppl. 2)) (2006) S13.
[29] T. Barrett, S.E. Wilhite, P. Ledoux, C. Evangelista, I.F. Kim, M. Tomashevsky, K.A. Marshall, K.H. Phillippy, P.M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C.L. Robertson, N. Serova, S. Davis, A. Soboleva, NCBI GEO: archive for functional genomics data sets–update, Nucleic Acids Res. 27–11 (Database issue) (2012) 41, doi:http://dx.doi.org/10.1093/nar/gks1193 (PMID:23193258).
[30] L. Wheeler D, T. Barrett, A. Benson D, H. Bryant S, K. Canese, V. Chetvernin, M. Church D, M. Dicuccio, R. Edgar, S. Federhen, M. Feolo, Y. Geer L, W. Helmberg, Y. Kapustin, O. Khovayko, D. Landsman, J. Lipman D, L. Madden T, R. Maglott D, V. Miller, J. Ostell, D. Pruitt K, D. Schuler G, M. Shumway, E. Sequeira, T. Sherry S, K. Sirotkin, A. Souvorov, G. Starchenko, L. Tatusov R, A. Tatusova T, L. Wagner, E. Yaschenko, s. author, Database resources of the National Center for Biotechnology Information, Nucleic Acids Res. 27–11 (Database issue) (2012) 41, doi:http://dx.doi.org/10.1093/nar/gks1189 (PMID:23193264).
[31] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biol. 4–3 (3) (2009) 10, doi:http://dx.doi.org/10.1186/gb-2009-10-3-r25 (PMID:19261174).
[32] M. Hollander, D.A. Wolfe, Nonparametric statistical methods, Nonparametric Statistics, 8)17, John Wiley & Sons, 1973, pp. 526, doi:http://dx.doi.org/10.1002/bimj.19750170808 (ISSN: 00063452).
[33] L. Wasserman, All of Statistics: A Concise Course in Statistical Inference, Springer, New York, 2004, doi:http://dx.doi.org/10.1007/978-0-387-21736-9 (ISBN: 0-387-40272-1).
[34] Donna Katzman McClish, Analyzing a portion of the ROC curve, Med. Decision Making 9 (3) (1989) 190–195, doi:http://dx.doi.org/10.1177/0272989x8900900307 (PMID 2668680).
[35] E. Dodd Lori, S. Pepe Margaret, Partial AUC estimation and regression, Biometrics 59 (3) (2003) 614–623, doi:http://dx.doi.org/10.1111/1541-0420.00071 (PMID 14601762. Retrieved 2007-12-18).
[36] E. Cheremushkin, A. Kel, Whole genome human/mouse phylogenetic footprinting of potential transcription regulatory signals, Pac. Symp. Biocomput. 29 (2003) 1–302.
[37] T. Waleev, D. Shtokalo, T. Konovalova, N. Voss, E. Cheremushkin, P. Stegmaier, O. Kel-Margoulis, E. Wingender, A. Kel, Composite module analyst: identification of transcription factor binding site combinations using genetic algorithm, Nucleic Acids Res. 34 (July (1)) (2006) W541–W545 (Web Server issue):W541-5. PMID: 16845066.
[38] C. Choi, M. Krull, A. Kel, O. Kel-Margoulis, S. Pistor, A. Potapov, N. Voss, E. Wingender, TRANSPATH–a high quality database focused on signal transduction, Comp. Funct. Genomics 5 (2) (2004) 163–168, doi:http://dx.doi.org/10.1002/cfg.386.
[39] J.A. Hanley, B.J. McNeil, A method of comparing the areas under receiver operating characteristic curves derived from the same cases, Radiology 148 (3) (1983) 839–843, doi:http://dx.doi.org/10.1148/radiology.148.3.6878708 (PMID 6878708).
[40] I.V. Kulakovskiy, I.E. Vorontsov, I.S. Yevshin, A.V. Soboleva, A.S. Kasianov, H. Ashoor, W. Ba-Alawi, V.B. Bajic, Y.A. Medvedeva, F.A. Kolpakov, V.J. Makeev, HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models, Nucleic Acids Res. 19 (November) (2015) (pii: gkv1249. [Epub ahead of print] PMID: 26586801).
[41] J.M. Lobo, A. Jiménez-Valverde, R. Real, AUC: a misleading measure of the performance of predictive distribution models, Global Ecol. Biogeogr. 17 (2008) 145–151.
[42] D. Berrar, P. Flach, Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them), Brief. Bioinform. 13 (1) (2012) 83–97, doi:http://dx.doi.org/10.1093/bib/bbr008.