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/