Infection, Genetics and Evolution 11 (2011) 769–777

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

Identification of key mechanisms controlling gene expression in *Leishmania* infected macrophages using genome-wide promoter analysis

Kais Ghedira [a,b], Klaus Hornischer [b], Tatiana Konovalova [b], Ahmed-Zaki Jenhani [a], Alia Benkahla [a,*], Alexander Kel [c,d]

[a] *Laboratory of Immunology, Vaccinology, and Molecular Genetics, Institut Pasteur de Tunis, 13, place Pasteur BP 74, 1002 Tunis, Tunisia*
[b] *BIOBASE GmbH, Halchtersche Strasse 33, Wolfenbüttel 38304, Germany*
[c] *geneXplain GmbH, Am Exer 10b, Wolfenbüttel 38302, Germany*
[d] *Institute of Chemical Biology and Fundamental Medicine, Lavrentiev Ave. 8, Novosibirsk 630090, Russia*

ARTICLE INFO

Article history:
Received 25 May 2010
Received in revised form 18 October 2010
Accepted 19 October 2010
Available online 18 November 2010

Keywords:
TFs
TFBSs
Promoters
Phylogenetic foot printing
Orthology
TSSs
Gene regulation
ChIP-Seq
Microarray analysis

ABSTRACT

The present study describes the *in silico* prediction of the regulatory network of *Leishmania* infected human macrophages. The construction of the gene regulatory network requires the identification of Transcription Factor Binding Sites (TFBSs) in the regulatory regions (promoters, enhancers) of genes that are regulated upon *Leishmania* infection. The promoters of human, mouse, rat, dog and chimpanzee genes were first identified in the whole genomes using available experimental data on full length cDNA sequences or deep CAGE tag data (DBTSS, FANTOM3, FANTOM4), mRNA models (ENSEMBL), or using hand annotated data (EPD, TRANSFAC). A phylogenetic footprinting analysis and a MATCH analysis of the promoter sequences were then performed to predict TFBS. Then, an SQL database that integrates all results of promoter analysis as well as other genome annotation information obtained from ENSEMBL, TRANSFAC, TRED (Transcription Regulatory Element Database), ORegAnno and the ENCODE project, was established. Finally publicly available expression data from human *Leishmania* infected macrophages were analyzed using the genome-wide information on predicted TFBS with the computer system ExPlain™. The gene regulatory network was constructed and activated signal transduction pathways were revealed. The Irak1 pathway was identified as a key pathway regulating gene expression changes in *Leishmania* infected macrophages.

1. Introduction

Promoters are genomic DNA sequences located in the upstream region of genes and containing specific DNA sequence elements, which are targets of proteins called Transcription Factors (TFs). The interaction of the TFs with TFBSs leads to the transcriptional regulation of the nearby gene. An accurate identification and a detailed knowledge of TFs and their target sites in promoters (cis-regulatory elements) contributes to the understanding of functioning of the Gene Regulatory Network (GRN) in a cell under various normal and disease conditions. The construction of the GRN in eukaryotes is based on the use of different sources of information such as the databases about positions of promoter regions in genomes, databases on experimentally known functional TFBSs, use of comparative genomics to search for phylogenetically conserved regions, study of the cis-regulatory modules, use of information of co-regulated genes from micro-array data, and integration of results from ChIP-chip or ChIP-Seq data.

During the last years many researchers have focused on the analysis of gene expression and on the prediction of the GRN. In a given cell, the GRN represents how genes "interact" with each others; it describes how the products of one set of genes (regulators) affects the expression of another set of genes (their targets) (Steele et al., 2009). The GRN is often represented as a graph in which nodes are genes/gene products (TF and their target genes) and edges correspond to the relation between the nodes. Over the last years, a number of different models for reverse engineering of GRNs from gene expression data have been proposed (Hecker et al., 2009). Qu et al. (2007) used the Nonlinear Dynamical System to analyze a breast cancer microarray experiment and to construct the corresponding GRN. Chen et al. (2009)

Abbreviations: DNA, desoxyribo nucleic acid; TF, transcription factor; TFBS, transcription factor binding site; GRN, gene regulatory network; TSS, transcription start site; PWM, position weight matrices; cDNA, complementary desoxyribo nucleic acid; EST, expressed sequence tags; EPD, eukaryotic promoter database; DBTSS, database of transcription start sites; CAGE, cap analysis gene expression; ChIP, chromatin immuno precipitation; ChIP-Seq, chromatin immuno precipitation followed by massively parallel DNA sequencing.

* Corresponding author. Tel.: +216 22938439.
E-mail address: alia.benkahla@pasteur.rns.tn (A. Benkahla).

1567-1348/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2010.10.015

used applied Network Component Analysis (Liao et al., 2003) to infer TF activities from microarray databases and partial TF-gene connectivity information for cytokines and related genes. Rather limited success of such computational techniques of reverse engineering is due to disregarding prior knowledge on promoter architecture and information on known protein–DNA and protein–protein interactions. Studies taking into account such prior information were done by Debily et al. (2009) and Smith et al. (2008), who combined pathway analysis with the analysis of promoter motifs to construct the GRN in breast cancer cells.

The first objective of the present work is to provide an accurate knowledge concerning promoter regions of genes in human, mouse, rat, dog and chimpanzee genomes, TFs and their target genes, and experimentally verified protein–DNA interaction data available in different databases. The next step is then to analyze the acquired knowledge together with tools for pathway analysis, such as ExPlain™ (Kel et al., 2008), to define the GRN of processes affected through *Leishmania* infection. ExPlain™ is a tool which relies on a knowledge database (TRANSFAC, TRANSPATH, TRANSCOMPEL) of curated functional and regulatory interactions extracted from the literature and which provides an integrated topological representation of the biological relationships between genes and their products.

The identification of promoter regions is considered as an important step in the analysis of the GRN. Early bioinformatic tools tried to predict the exact locations of Transcription Start Sites (TSSs) and promoter regions by applying rules which are based on some statistical characteristics of promoter sequences, such as the presence of particular consensus sequences (e.g. TATA-box, CG-element). These tools proved to be highly inaccurate, because the rules they are based upon are not universally applicable (TATA-box like sequences, for instance, are statistically seen located every 250 bp, not all transcription initiation sites are proximal to CpG islands, etc.) and can help to identify promoters, but cannot be exclusively used for this purpose. A technique which proved to be a much more reliable indicator for promoter regions is the mapping of 5′ EST's and/or full length cDNAs and/or Deep Cage tags on genome sequences. Several databases provide access to the TSS positions for human, mouse and other organisms, e.g. DBTSS (Wakaguri et al., 2008), EPD (Schmid et al., 2004), PromoSer (Halees and Weng, 2004), FANTOM (Kawaji et al., 2009; Severin et al., 2009), TRANSPRO (Matys et al., 2006) and TiProD (Chen et al., 2006).

The next step in the construction of a GRN is the detection of TFBSs located in the promoter regions of regulated genes. Computational prediction of TFBSs in DNA sequences is usually done using Position Weight Matrices (PWMs) also known as Position Specific Scoring Matrices or Position Specific Weight Matrices, which are probabilistic models that characterize the DNA binding preferences of TFs. The PWMs are usually constructed from a collection of experimentally identified TFBSs that are expected to be bound by the same TF. A large collection of PWMs is present in the TRANSFAC database. Associated tools, such as MATCH (Kel et al., 2003) and PMATCH (Chekmenev et al., 2005) and a number of similar methods are using PWMs from the TRANSFAC database for the prediction of TFBSs in DNA sequences. A major problem of PWM-only methods for prediction of TFBSs is their high rate of false positive predictions due to the short typical length of PWMs and the high level of (potential) site redundancy.

Phylogenetic footprinting, can also be used to identify genomic regions of potential regulatory impact. It is based on the assumption that functionally important regulatory elements are subject to a higher evolutionary pressure than the general genomic sequence "background". This approach has proven to be highly effective in decreasing the rate of false positives prediction and improving the accuracy of prediction (Doerwald et al., 2004; Polavarapu et al., 2008) by focusing on sequences under selective pressure. In 2007, Davies et al. (2007) used PWM to identify known/novel cis-regulatory motifs, and used phylogenetic footprinting to increase the sensitivity of their predictions. A number of software tools have been developed for the prediction of conserved TFBSs (Blanchette and Tompa, 2003; Dubchak and Ryaboy, 2006; Sandelin and Wasserman, 2004). While extremely valuable, this approach has to be used with caution, since the very dynamic evolution of promoters of eukaryotic genes may lead to a high variability of both position and structure of TFBSs. In this work we used the program FootPrinter (Blanchette and Tompa, 2003) to perform phylogenetic footprinting. This method relies on a motif discovery approach between sets of unaligned homologous promoter sequences, and is therefore less prone to yield false negatives due to variability of site positions.

Experimental identification of functional TFBS in promoter regions is time consuming, very expensive and impractical for genome wide transcription regulation studies. Advances in high-throughput approaches such as Chromatin Immuno-Precipitation (ChIP) experiments followed by array hybridization (Choi et al., 2009; Whitington et al., 2009) or ChIP-Seq experiments (Barski et al., 2007; Boyle et al., 2008; Robertson et al., 2007), give valuable information for the characterization of genome-wide binding of TFs to their genomic sites and are very potent techniques to improve the identification of functional TFBSs and the decrease of false positive rates. Integration of the ChIP-Seq data generated in the frame of the ENCODE project, to other large-scale genome-wide data should help to increase accuracy of computational prediction of TFBSs.

In spite of the use of the above data and techniques, our knowledge about TFs, and especially their binding sites and target genes, is still limited and much remains unknown about the regulation of transcription in mammals. Known information on experimentally verified binding sites is available currently for many organisms in databases like TRANSFAC (Wingender et al., 1997; Matys et al., 2006), TRED (Jiang et al., 2007) and ORegAnno (Griffith et al., 2008).

The combination of TFBSs prediction tools, phylogenetic footprinting, ChIP-Seq data and microarray analysis can help to accurately identify TFBSs and decrease the rate of false positive predictions. Our present study uses data from different resources to identify promoter sequences in the genomes of 5 mammals (*Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Canis familiaris* and *Pan troglodytes*). Once the promoters were identified, we combined different tools (MATCH, FootPrinter) and the ENCODE project ChIP-Seq data to predict regulatory elements being located in promoters. We then used TRANSFAC, TRED and ORegAnno databases to evaluate the predictions based on more than two data types. Finally, we used the ExPlain system (Kel et al., 2008) and microarray data from pathogen infected immune cells (Chaussabel et al., 2003) to analyze gene expression and to predict key networks in *Leishmania major* (*L. major*) infected macrophages.

2. Materials and methods

2.1. TSS identification and promoter delimitation

TSS data (5′-end sequences of full-length cDNAs, deep CAGE tags) for human and mouse were collected from different available resources by extracting full length cDNA sequences or deep CAGE tag data (DBTSS, FANTOM3, FANTOM4), mRNA models (ENSEMBL), or using hand annotated data (EPD, TRANSFAC). These sequences were mapped to human and mouse genomes sequences (hg18 and mm9, respectively) using BLASTN, in order to delimit TSS positions. Sequences sharing $\geq$95% homology with human and mouse