Gene xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Gene

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

Review

Eukaryotic and prokaryotic promoter databases as valuable tools in exploring the regulation of gene transcription: a comprehensive overview

Małgorzata Majewska[a,*], Halina Wysokińska[a], Łukasz Kuźma[a], Piotr Szymczyk[b]

[a] Department of Biology and Pharmaceutical Botany, Medical University of Lodz, 90-151 Lodz, Poland
[b] Department of Pharmaceutical Biotechnology, Medical University of Lodz, 90-151 Lodz, Poland

ARTICLE INFO

Keywords:
Database
Promoter
Transcription factor
miRNA
Binding site
Gene regulation

ABSTRACT

The complete exploration of the regulation of gene expression remains one of the top-priority goals for researchers. As the regulation is mainly controlled at the level of transcription by promoters, study on promoters and findings are of great importance. This review summarizes forty selected databases that centralize experimental and theoretical knowledge regarding the organization of promoters, interacting transcription factors (TFs) and microRNAs (miRNAs) in many eukaryotic and prokaryotic species. The presented databases offer researchers valuable support in elucidating the regulation of gene transcription.

1. Introduction

Nowadays, one of the most important challenges in Biology is to gain a complete understanding of the regulation of gene expression. At the level of transcription, this regulation is largely dependent on promoters. Promoters are able to bind proteins called transcription factors (TFs), recruited RNA polymerase and thereby regulate gene transcription (Zawel and Reinberg, 1995; Latchman, 1997; Browning and Busby, 2004). Additionally, accumulating evidence suggests that also microRNAs (miRNAs) are involved in transcriptional gene regulation by targeting promoter elements (Huang et al., 2012; Younger and Corey, 2011; Bartel, 2004). However, the mechanism is not clear yet. The acquired knowledge about investigated promoters is often integrated into databases. By providing comprehensive information, the databases facilitate and even guide future research. Moreover, they can also act as a basis for the development of tools that enable in silico analysis of new promoter sequences with regard to gene transcription regulation (Ohler and Niemann, 2001; Qiu, 2003; Wasserman and Sandelin, 2004). Therefore, the databases have become a cornerstone of modern Biology.

The aim of the present work is to summarize forty selected, functioning publicly-available web-based databases (excluding tools used only for in silico analysis) that provide experimental and theoretical knowledge concerning promoters as regulators of transcription of genes in eukaryotes and prokaryotes. Despite our efforts, we were not able to incorporate all existing resources into the review. The data describe the organization of promoters and their interaction with TFs and miRNAs. The selected repositories were classified according to organism and the type of data they present. To provide the fullest possible analysis, several similar databases were chosen from each given area, if available, and examined together. This is a comprehensive and up-to-date review on this important topic. We believe that our manuscript can successfully assist researchers exploring the regulation of gene

Abbreviations: AGRIS, *Arabidopsis* Gene Regulatory Information Server; AtcisDB, *Arabidopsis thaliana cis*-regulatory element database; AtPAN, *Arabidopsis thaliana* Promoter Analysis Net; AtRegNet, *Arabidopsis thaliana* regulatory network; AtTFDB, *Arabidopsis thaliana* transcription factor database; BBLS, Bayesian Branch Length Score; BLAST, Basic Local Alignment Search Tool; B1H, bacterial one-hybrid system; ChIP-chip, chromatin immunoprecipitation-on-chip; ChIP-seq, chromatin immunoprecipitation-sequencing; DBTBS, The Database of Transcriptional Regulation in *Bacillus subtilis*; DCPD, *Drosophila* Core Promoter Database; DNase-seq, DNase sequencing; DoOP, Database of Orthologous Promoters; DPE, downstream promoter element; EPD, Eukaryotic Promoter Database; ESTs, expressed sequence tags; FDR, false discovery rate; FFLs, feed-forward regulatory loops; GRN, gene regulatory network; GTRD, Gene Transcription Regulation Database; hg18, human genome 18; hg19, human genome 19; HOCOMOCO, *Homo sapiens* Comprehensive Model Collection; HT-SELEX, high-throughput systematic evolution of ligands by exponential enrichment; Inr, initiator element; LDSS, local distribution of short sequences; miRNA, microRNA; MFE, minimum free energy; NGS, next-generation sequencing; NLOD, normalized log-odds; PBM, protein binding microarray; PFMs, position frequency matrices; PlantPAN, Plant Promoter Analysis Navigator; PlantProm DB, Plant Promoter Database; ppdb, plant promoter database; PPIs, protein-protein interactions; PRODORIC, Prokaryotic Database of Gene Regulation; Pro54DB, σ^{54} promoter database; PWMs, position weight matrices; RACE, 5′ rapid amplification of cDNA ends; REGs, regulatory element groups; ReIN, Regulatory Networks Interaction Module; RNA-seq, RNA sequencing; SCPD, *Saccharomyces cerevisiae* Promoter Database; SELEX, systematic evolution of ligands by exponential enrichment; SNP, single nucleotide polymorphism; sRNA, small RNA; STIFDB, Stress-Responsive Transcription Factor Database; TF, transcription factor; TFBSs, transcription factor binding sites; TFFMs, transcription factor flexible models; TRED, Transcriptional Regulatory Element Database; TSSs, transcription start sites; UniPROBE, Universal PBM Resource for Oligonucleotide Binding Evaluation; UTR, untranslated region; YCRD, Yeast Combinatorial Regulation Database; YEASTRACT, Yeast Search for Transcriptional Regulators and Consensus Tracking; YPA, Yeast Promoter Atlas

* Corresponding author at: Department of Biology and Pharmaceutical Botany, Medical University of Lodz, Muszyńskiego 1, 90-151 Lodz, Poland.
E-mail address: malgorzata.majewska1@stud.umed.lodz.pl (M. Majewska).

http://dx.doi.org/10.1016/j.gene.2017.10.079
Received 12 May 2017; Received in revised form 26 July 2017; Accepted 27 October 2017
0378-1119/ © 2017 Elsevier B.V. All rights reserved.

expression to find a suitable database by markedly reducing the time needed to identify one.

2. Eukaryotic databases

2.1. Multispecies eukaryotic databases

One of the oldest promoter databases is the Eukaryotic Promoter Database (EPD) (Cavin Perier et al., 1998). It is a collection of 4806 RNA polymerase II promoters with transcription start sites (TSSs). The data presented in the EPD are based on TSS mapping experiments published in journal articles or on expressed sequence tags (ESTs) of full-length cDNA clones used for in silico primer extension. The user is able to download selected species-specific promoter sequences (−499: +100 bp relative to TSS). In 2011, a new section called the EPDnew was added (Dreos et al., 2013). In contrast to the EPD, the EPDnew contains promoters automatically assembled from next-generation sequencing (NGS) and from high-throughput promoter mapping experiments. The EPDnew currently supports 10 species: six animal species (*Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Apis mellifera*, *Caenorhabditis elegans*, *Danio rerio*), two plant species (*Arabidopsis thaliana*, *Zea mays*), and two fungus species (*Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*). The EPDnew provides over 105,000 promoter sequences and information on their motifs, i.e., TATA box, GC box, CCAAT box, initiator element (Inr), CpG island. The user is free to use various analysis tools to examine promoters (frequency and occurrence of motifs) and selection tools to select subsets of promoters. The EPD and EPDnew are regularly updated, with the most recent update on May 2017.

ECRbase provides a collection of vertebrate promoter sequences with transcription factor binding sites (TFBSs); it contains *H. sapiens*, *Macaca mulatta*, *Pan troglodytes*, *Bos taurus*, *Canis familiaris*, *Monodelphis domestica*, *Rattus norvegicus*, *M. musculus*, *Gallus gallus*, *Xenopus tropicalis*, *D. rerio* and *Takifugu rubripes* annotations (Loots and Ovcharenko, 2007). TFBSs are recognized based on position weight matrices (PWMs) extracted from the TRANSFAC database (Matys et al., 2006) and the tfSearch algorithm (Loots and Ovcharenko, 2004). To mitigate the problem of false positive results, the authors independently optimized thresholds for different TFBSs. The database gives the possibility to download the data. It was recently updated in 2009.

The Transcriptional Regulatory Element Database (TRED) is a repository of human, rat, and mouse *cis*- and *trans*-regulatory elements (Zhao et al., 2005; Jiang et al., 2007). The user is able to retrieve 139,379 promoter sequences with TSSs. The promoters are extracted from experimental literature and are derived by computational predictions. It is also possible to search for TFBSs and obtain information on their sequence, localization, and corresponding TFs. Furthermore, the database provides access to orthologous genes and gene regulatory networks (GRNs) constructed by 34 TF families implicated in cancer pathways and their target genes. The TRED is simple in construction and comprehensive database. The last update of the database could not be determined.

The mechanisms of regulation of genes encoding miRNAs are depicted by DIANA miRGen (Georgakilas et al., 2016). It is a unique repository of 276 precise TSSs of miRNA genes, and over 19,000 binding sites for 202 TFs obtained from nine cell lines and six tissues of *H. sapiens* and *M. musculus*. The user is able to receive information on miRNAs, the genomic position of TSSs and binding sites, binding motif logos and expression of TFs. The promoters incorporated into the database were obtained from experimental and computational studies. The TSSs were predicted by using the microTSS algorithm (Georgakilas et al., 2014) on chromatin immunoprecipitation-sequencing (ChIP-seq), RNA sequencing (RNA-seq), and DNase sequencing (DNase-seq) data. The algorithm is able to identify miRNA TSSs to a resolution of one nucleotide, with 93,6% sensitivity and 100% precision. Binding sites for TFs were estimated by scanning promoter sequences based on position frequency matrices (PFMs). The last update of the database could not be determined.

The CircuitsDB connects transcriptional and post-transcriptional interactions of TFs, miRNAs and genes to establish feed-forward regulatory loops (FFLs) (Friard et al., 2010). A master TF regulates miRNA in the loop, and both regulate target protein-coding genes. The circuits were created here based on ab-initio sequence analysis of *M. musculus* and *H. sapiens* genomes. The CircuitsDB provides detailed information on target genes and miRNAs (including disease involvement), TFs, binding sites for miRNAs in the 3´untranslated region (3′UTR), putative TFBSs in promoters, and tissue expression of loop components. Conserved overrepresented oligos (putative TFBSs) were identified by promoter scanning based on consensus sequences from the TRANSFAC database (Matys et al., 2006). In addition, the CircuitsDB includes 21,446 human (21,316 protein-coding and 130 for pre-miRNA) and 21,944 mouse (21,814 protein-coding and 130 pre-miRNA) promoter regions. In total, the transcriptional network consists of 43,453 genes and 4062 binding sites for 230 TFs, while the post-transcriptional network involves 33,021 genes and 360 binding sites for 283 mature miRNAs. The user can explore transcriptional and post-transcriptional networks separately. The database is flexible, it is possible to query by TF, by miRNA or by target gene and download the data. The last update of the database could not be determined.

A comparable database is RegNetwork (Liu et al., 2015). It is a comprehensive collection of experimentally confirmed and predicted transcriptional and post-transcriptional regulatory networks extracted from 25 databases, including JASPAR (Bryne et al., 2008) and the TRED (Jiang et al., 2007) described here. TF-gene, TF-miRNA, TF-TF, miRNA-TF, miRNA-gene interactions are considered. Each of the interactions is labeled with a degree of confidence (high, medium or low). The human part of RegNetwork consists of 1456 TFs, 1904 miRNAs and 19,719 target genes. The mouse unit comprises 1328 TFs, 1290 miRNAs and 18,120 target genes. The obtained results can be downloaded. The database was recently updated on 1st July 2017.

The microPIR2 database provides a number of possible miRNA binding sites in promoters: about 80 million human and 40 million mouse (Piriyapongsa et al., 2014). The sites were predicted using the RNAhybrid program (Kruger and Rehmsmeier, 2006). The user can also search for conserved miRNA binding sites that can be powerful candidates for subsequent experimental analysis. As a comprehensive database, microPIR2 supplies detailed knowledge about miRNAs, target genes and binding sites (length, position, overlapping with other important sequences, minimum free energy (MFE), *p*-value, conserved score), together with links to other databases. The user can query by gene, miRNA, disease and binding site characteristics. The last update of the database could not be determined.

A comparable repository is miRWalk2.0 (Dweep et al., 2011). It comprises miRNA binding sites in human, mouse, and rat promoters. The sites were identified with the miRWalk algorithm and an additional nine programs (DianamT, miRanda, miRDB, Pictar4, Pictar5, PITA, RNA22, RNAhybrid, Targetscan), to improve prediction. Apart from miRNA binding sites, the database provides information on miRNAs and target genes. The user can search by target gene, miRNA, pathway, class of gene and protein, gene ontology, disorders, diseases and human phenotype ontology. miRWalk2.0 is regularly updated. The last update was on June 2017.

The Database of Orthologous Promoters (DoOP) (Barta et al., 2005) and its tool DoOPSearch (Sebestyen et al., 2009) were created to identify evolutionarily-conserved promoter sequences (possible TFBSs) of a given gene. This method is called phylogenetic footprinting. The conserved sequences were found by multiple alignments using the DIALIGN2 program (Morgenstern, 1999). The DoOP comprises over one million such orthologous sequences (6–50 bp long) from chordate and plant promoters. The database offers the possibility to search for specific promoter clusters (500, 1000 or 3000 bp long) or sequence motifs using the DoOPSearch tool, giving a list of characterized genes and a

Download English Version:

https://daneshyari.com/en/article/8645719

Download Persian Version:

https://daneshyari.com/article/8645719

Daneshyari.com