JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework

To cite this version:

HAL Id: hal-01646126
https://hal-amu.archives-ouvertes.fr/hal-01646126
Submitted on 10 Jan 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License
JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework

Aziz Khan1,†, Oriol Fornes2,†, Arnaud Stigliani3,†, Marius Gheorghe1, Jaime A. Castro-Mondragon1, Robin van der Lee2, Adrien Bessy3, Jeanne Chèneby4,5, Shubhada R. Kulkarni6,7,8, Ge Tan9,10, Damir Baranasic9,10, David J. Arenillas2, Albin Sandelin11,*, Klaas Vandepoele6,7,8, Boris Lenhard9,10,12,*, Benoît Ballester3, Wyeth W. Wasserman2,*, François Parcy3 and Anthony Mathelier1,13,*

1Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo, 0318 Oslo, Norway, 2Centre for Molecular Medicine and Therapeutics, Department of Medical Genetics, BC Children's Hospital Research Institute, University of British Columbia, 950 28th Ave W, Vancouver, BC V5Z 4H4, Canada, 3University of Grenoble Alpes, CNRS, CEA, INRA, BIG-LPCV, 38000 Grenoble, France, 4INSERM, UMR1090 TAGC, Marseille, F-13288, France, 5Aix-Marseille Université, UMR1090 TAGC, Marseille, F-13288, France, 6Ghent University, Department of Plant Biotechnology and Bioinformatics, Technologiepark 927, 9052 Ghent, Belgium, 7VIB Center for Plant Systems Biology, Technologiepark 927, 9052 Ghent, Belgium, 8Bioinformatics Institute Ghent, Ghent University, Technologiepark 927, 9052 Ghent, Belgium, 9Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, London W12 0NN, UK, 10Computational Regulatory Genomics, MRC London Institute of Medical Sciences, London W12 0NN, UK, 11The Bioinformatics Centre, Department of Biology and Biotech Research & Innovation Centre, University of Copenhagen, DK2200 Copenhagen N, Denmark, 12Sars International Centre for Marine Molecular Biology, University of Bergen, N-5008 Bergen, Norway and 13Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, 0310 Oslo, Norway

Received September 25, 2017; Revised October 17, 2017; Editorial Decision October 18, 2017; Accepted October 27, 2017

JASPAR (http://jaspar.genereg.net) is an open-access database of curated, non-redundant transcription factor (TF)-binding profiles stored as position frequency matrices (PFMs) and TF flexible models (TFFMs) for TFs across multiple species in six taxonomic groups. In the 2018 release of JASPAR, the CORE collection has been expanded with 322 new PFMs (60 for vertebrates and 262 for plants) and 33 PFMs were updated (24 for vertebrates, 8 for plants and 1 for insects). These new profiles represent a 30% expansion compared to the 2016 release. In addition, we have introduced 316 TFFMs (95 for vertebrates, 218 for plants and 3 for insects). This release incorporates clusters of similar PFMs in each taxon and each TF class per taxon. The JASPAR 2018 CORE vertebrate collection of PFMs was used to predict TF-binding sites in the human genome. The predictions are made available to the scientific community through a UCSC Genome Browser track data hub. Finally, this update comes with a new web framework with an interactive and responsive user-interface, along with new features. All the underlying data can be retrieved programmatically using a RESTful API and through the JASPAR 2018 R/Bioconductor package.

INTRODUCTION

Transcription factors (TFs) are sequence-specific DNA-binding proteins involved in the transcriptional regulation of gene expression (1). TFs bind to DNA through their DNA-binding domain(s) (DBDs), which are used for TF classification (2). DNA regions at which TFs bind are defined as TF-binding sites (TFBSs) and can be identified
in vivo by methods such as chromatin immunoprecipitation (ChIP) or in vitro by methods based on binding of large pools of DNA fragments (e.g. Systematic evolution of ligands by exponential enrichment (SELEX) or protein-binding microarrays (PBM)) (reviewed in (3)). Analysis of TFBSs for a given TF provides models for its specific DNA-binding preferences, which in turn can be used to predict TFBSs in DNA sequences (4). This is important as experiments can only identify TFBSs that are bound in the cell and state analyzed.

The computational representation of TF binding preferences has evolved over the years, from simple consensus sequences to position frequency matrices (PFMs). A PFM summarizes experimentally determined DNA sequences bound by an individual TF by counting the number of occurrences of each nucleotide at each position within aligned TFBSs. Such matrices can be converted into position weight matrices (PWMs), also known as position-specific scoring matrices, which are probabilistic models that can be used to predict TFBSs in DNA sequences (reviewed in (5)).

PFMs/PWMs have been the standard models for describing binding preferences of TFs for many years. The JASPAR database is among the most popular and longest maintained databases for PFMs and a standard resource in the field. In particular, the JASPAR CORE collection of the database, which is the most used, stores non-redundant TF binding profiles, providing a single representative DNA binding model per TF decided by expert curators. Exceptionally, multiple TF-binding profiles are associated to a TF when it is known to interact with DNA with multiple distinct sequence preferences, due to differential splicing for example (6,7). JASPAR was created and persists under three guiding principles: (i) unrestricted open-access; (ii) manual curation and non-redundancy of profiles; and (iii) ease-of-use. The 2016 release of the JASPAR CORE collection stored 1082 non-redundant and manually curated TF-binding profiles as PFMs for TFs from six different taxonomic groups (vertebrates, plants, insects, nematodes, fungi and urochordata) (8).

An intrinsic limitation to PFMs/PWMs is that they ignore inter-nucleotide dependencies within TFBSs (9–13). TF–DNA interaction data derived from next-generation sequencing assays has improved the computational modeling of TF binding (14–19). For example, the TF flexible models (TFFMs) (14), based on first-order hidden Markov models, capture dinucleotide dependencies within TFBSs and were introduced in the 2016 release of the JASPAR database.

In this report, we describe the seventh release of JASPAR (8,20–24), which comes with a major expansion and update of the CORE collection of TF-binding profiles as PFMs and TFFMs. These models have been manually assessed by expert curators who reconciled recent high-throughput data with available literature and linked the models to the classification of their TF DBDs from TFClass (2). The CORE collection expansion is supported by a range of new functionalities and resources, including PFM clustering, genome-wide UCSC tracks of predicted TFBSs and fully redesigned user and programming interfaces.

### EXPANSION AND UPDATE OF THE JASPAR CORE COLLECTION

In this 2018 release of the JASPAR database, we added 355 new PFMs for TFs from plants (270), vertebrates (84) and insects (1) to the JASPAR CORE collection (Table 1). Specifically, we added 322 PFMs (262 for plants, a 118% increase and 60 for vertebrates, an 11% increase) for TF monomers and dimers that were not previously present in JASPAR and updated 33 (8 in plants, 3% of JASPAR 2016, 24 in vertebrates, 5% of JASPAR 2016 and 1 in insects). The PFMs were manually curated using independent external literature supporting the candidate TF-binding preferences, as previously described in (23). The curated PFMs were derived from ChIP-seq (from ReMap (25) and (26–30)), DAP-seq (31), SMiLE-seq (32), PBM (33) and HT-SELEX (34) experiments. The JASPAR CORE collection now includes 1404 non-redundant PFMs (579 for vertebrates, 489 for plants, 176 for fungi, 133 for insects, 26 for nematodes and 1 for urochordata) (Table 1).

We continued with the incorporation of TFFM models, initiated in JASPAR 2016. In this release of JASPAR, we introduced 316 new TFFMs for vertebrates (95), plants (218) and Drosophila (3), which represents a 243% increase in the number of non-redundant TFFMs stored in the JASPAR CORE collection.

### HIERARCHICAL CLUSTERING OF TF-BINDING PROFILES

While the non-redundancy of binding profiles is one of the guiding principles of JASPAR, TFs with similar DBDs often have similar binding preferences (35,36). To facilitate the exploration of similar profiles in the JASPAR CORE collection, we performed hierarchical clustering of PFMs using the RSAT matrix-clustering tool (37). Specifically, the tool was applied to PFMs in each taxon independently as
Figure 1. JASPAR PFM clustering. (A) Radial tree representing the clusterization of the JASPAR CORE vertebrate PFMs. (B) Zoom in view of the radial tree where the predicted clusters are highlighted at the branches and the TF classes are indicated with different colors at the leaves. (C) Clicking on a leaf in the radial tree will open a link to the corresponding motif description page on the JASPAR website (the MA0148.3 profile associated to FOXA1 is provided here as an example).
Figure 2. Overview of the JASPAR 2018 new web interface with interactive searching activity. (A) A quick and detailed search feature on the homepage. (B) A responsive table lists the searched profile(s), which can be further selected and added to the cart listed on the right panel for users to perform their own analyses. (C) A detailed page for the GATA3 matrix profile, which is divided into sub-panels including the profile summary, sequence logo, PFM, TF-binding information, external links, version information, ChIP-seq centrality, TFFM and other details. (D) The PFM for the GATA3 profile (MA0037.2) is downloaded in MEME format using the RESTful API.

well as in each TF class per taxon. The clustering results are provided as radial trees (Figure 1), which can further be explored through dedicated web pages (http://jaspar.genereg.net/matrix-clusters).

JASPAR UCSC TRACKS FOR GENOME-WIDE ANALYSES OF TFBSs

A typical application of JASPAR TF-binding profiles in gene regulation studies is the identification of TFBSs in DNA sequences for further analyses. Although, we recognize that genome-wide PWM-based predictions contain a high number false positives, we believe that they are a powerful resource for the research community in the context of a variety of genomic information, including transcription start site activity, DNA accessibility, histone marks, evolutionary conservation or in vivo TF binding (38–46).

To facilitate such integrative analyses, we have performed TFBS predictions on the human genome using the JASPAR CORE vertebrate PFM (see Supplementary Data for details on the computation). The predicted TFBSs are publicly available through a UCSC Genome Browser data hub (47) containing tracks for the human genome assemblies hg19 and hg38 (http://jaspar.genereg.net/genome-tracks/).
A NEW, POWERFUL AND USER-FRIENDLY WEB INTERFACE

A new web interface

The JASPAR 2018 release comes with a completely redesigned web interface that meets modern web standards. This interactive web framework is implemented using Django, a model-view-controller based web-framework for Python. We used MySQL as a backend database to store profile metadata and Bootstrap as a frontend template engine. We have greatly improved the visibility and usability of existing functionality, created easier navigation with semantic URLs, and enhanced browsing and searching. On the homepage, we provide a dynamic tour of JASPAR 2018, walking users through the main features of the new website. A video of the tour is available at http://jaspar.genereg.net/tour. The database can be browsed for individual collections by using the navigation links on the left sidebar. Moreover, it can be searched for each of the six different taxonomic groups included in the JASPAR CORE collection using the tabs available on the homepage (Figure 2). TF-binding profiles can be further filtered through the case insensitive search option available on the homepage. In addition, through the ‘Advanced Options’, the search criteria can be further restricted (Figure 2A). Search results are presented in a responsive and paginated table along with sequence logos of the PFMs, which can be selected for download or to perform a variety of analyses available on the right panel (Figure 2B). All information in the tables can be downloaded as comma-separated value files. Profile IDs and sequence logos can be clicked to view the detailed profile pages (Figure 2C). PFMs can be downloaded in several formats including JASPAR, TRANSFAC and MEME (Figure 2D). Furthermore, we have incorporated new features to the web interface, such as ‘Add to Cart’, where users can add TF profiles of interest for download or further analyses (Figure 2B). Finally, we have introduced semantic URLs to facilitate external linking to the detailed pages of individual profiles (e.g. http://jaspar.genereg.net/matrix/MA0059.1). We have implemented a URL redirection mechanism to correctly direct the links pointing to previous JASPAR URL patterns from external resources.

RESTful API

In previous releases, the underlying data could be retrieved as flat files or by using programming language-specific modules. Associated with this release, we introduced a RESTful API to access the JASPAR database programmatically (see https://www.biorxiv.org/content/early/2017/07/06/160184 for details). The RESTful API enables programmatic access to JASPAR by most programming languages and returns data in seven widely used formats: JSON, JSONP, JASPAR, MEME, PFM, TRANSFAC and YAML. Further, it provides a browsable interface and access to the JASPAR motif inference tool for bioinformatics tool developers. The RESTful API is implemented in Python using the Django REST Framework and is freely accessible at http://jaspar.genereg.net/api. The source code for the website and RESTful API are freely available at https://bitbucket.org/CBGR/jaspar under GPL v3 license.

CONCLUSION AND PERSPECTIVES

In this seventh release of the JASPAR database, we continue our commitment to provide the research community with high-quality, non-redundant TF-binding profiles for TFs in six taxa. As in previous releases, we have greatly expanded the number of available profiles in the database, both for PFMs and TFFMs. We also greatly improved user experience through a new easy-to-use website and a RESTful API that grants universal programmatic access to the database. Moreover, for the PFMs in the JASPAR CORE collection, we provide a hierarchical clustering and genome-wide TFBs predictions for the hg19 and hg38 human genome assemblies as UCSC tracks.

During the curation process, hundreds of PFMs were discarded because our curators failed to find any support from existing literature. As new experiments and data become available, binding preferences for these TFs will be considered for JASPAR incorporation. For instance, we re-examined data from (34) to incorporate seven previously excluded PFMs into JASPAR 2018. In the future, we would like to engage the scientific community in the curation process to increase our capacity to introduce new TF-binding profiles in JASPAR. We plan to dedicate a specific section of the website to hosting the profiles that were not introduced into JASPAR, to encourage researchers to perform experiments and/or point us to literature that our curators missed in order to support these profiles. We believe that the engagement of the scientific community to support JASPAR will further improve our capacity to expand the collection of high quality TF-binding profiles.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the scientific community for performing experimental assays of TF–DNA interactions and for publicly releasing the data. We thank Georgios Magklaras and his team for IT support. We thank José Manuel Franco for sharing the plant PBM data, Jens De Ceukeleire for help with plant ChIP-seq data processing and José Luis Villanueva-Cañas for sharing the Drosophila TFFMs prior to publication. We thank Rachelle Farkas for proofreading the manuscript.

FUNDING

Norwegian Research Council, Helse Sør-Øst, and University of Oslo through the Centre for Molecular Medicine Norway (NCMM) (to A.M., M.G., A.K.); Genome Canada and Canadian Institutes of Health Research (On-Target Grants) [255ONT and BOP-149430 to W.W., O.F., R.v.d.L., D.J.A.]; Natural Sciences and Engineering Research Council of Canada (Discovery Grant) [RGPIN-2017–06824 to W.W., O.F., A.S.]; Weston Brain Institute [20R74681 to O.F.]; Agence Nationale de la Recherche [ANR-10-LABX-49–01 to F.P., A.S.]; IDEX graduate schoool (to A.S.); CNRS (to A.B., F.P.); Research Foundation–Flanders Grant [G001015N to S.R.K.]; French Ministry


