Nucleic Acids Research, 2006, Vol. 34, Database issue ***D95–D97***
doi:10.1093/nar/gkj115

A new generation of JASPAR, the open-access repository for transcription factor binding site profiles

Dominique Vlieghe, Albin Sandelin[1], Pieter J. De Bleser, Kris Vleminckx, Wyeth W. Wasserman[2], Frans van Roy and Boris Lenhard[3,*]

Department for Molecular Biomedical Research (DMBR), VIB—Ghent University, Technologiepark 927 B-9052 Ghent (Zwijnaarde), Belgium, [1]Genome Exploration Research Group, RIKEN Genomic Sciences Center (GSC), RIKEN Yokohama Institute,1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan, [2]Centre for Molecular Medicine and Therapeutics, Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada and [3]Computational Biology Unit, Bergen Center for Computational Science, University of Bergen, Thormøhlensgate 55, N-5008 Bergen, Norway

Received September 13, 2005; Revised and Accepted October 18, 2005

ABSTRACT

JASPAR is the most complete open-access collection of transcription factor binding site (TFBS) matrices. In this new release, JASPAR grows into a meta-database of collections of TFBS models derived by diverse approaches. We present JASPAR CORE—an expanded version of the original, non-redundant collection of annotated, high-quality matrix-based transcription factor binding profiles, JASPAR FAM—a collection of familial TFBS models and JASPAR phyloFACTS—a set of matrices computationally derived from statistically overrepresented, evolutionarily conserved regulatory region motifs from mammalian genomes. JASPAR phyloFACTS serves as a non-redundant extension to JASPAR CORE, enhancing the overall breadth of JASPAR for promoter sequence analysis. The new release of JASPAR is available at http://jaspar.genereg.net.

INTRODUCTION

Methods for computational discovery and analysis of regulatory sequences are becoming increasingly important for the interpretation of genome and transcriptome data. Reliable prediction of *cis*-regulatory elements is critically dependent upon access to high-quality models for the binding specificity of transcription factors (TFs) (1). These models are predominantly defined by ungapped alignments of bona fide TF binding sites (TFBSs), summarized as count matrices (also referred to as matrix *profiles*) (2).

The JASPAR database, the largest open-access collection of TFBS matrix profiles, is used as a fundamental component within a growing number of bioinformatic tools (3–8). The initial release of the JASPAR database (9) contained a collection of extensively curated, non-redundant profiles collected from published collections of TFBS from multicellular eukaryotes. Those high-quality profiles remain the collection of choice for the detection of putative binding sites resembling target sequences of known TFs.

Since laboratory-based elucidation of bona fide TFBSs is time-consuming and labor-intensive, only a fraction of all TFs have a defined binding profile. Based on the slow influx of conclusively validated data, expansion of the curated JASPAR binding profiles is lethargic. In the meantime, binding model collections based on other approaches, such as *in silico* pattern discovery, have emerged. While researchers may prefer to use highly curated profiles from bona fide TFBSs, the new collections offer great utility for genome-scale analysis. Compared with the original JASPAR collection, such datasets will differ both in terms of methods used for generation as well as in the level of biological evidence. They should therefore not be indiscriminately added to the collection, but are nevertheless valuable for the exploration of the content of regulatory regions. To address the need and desire for access to a broader range of binding profiles, we present an expansion of JASPAR into a meta-repository for TF binding profiles, within which profiles are divided into distinct subsets that differ in data generation methodology.

*To whom correspondence should be addressed. Tel: +47 55 84362; Fax: +47 555 84295; Email: Boris.Lenhard@bccs.uib.no

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

APPLYING THE JASPAR PARTITIONS FOR GENOME ANALYSIS

In this release, JASPAR contains three distinct subsets (Table 1):

JASPAR CORE. This collection corresponds to an expansion of the original set of JASPAR profiles. The 123 binding models in JASPAR CORE are non-redundant and based on experimentally defined TFBSs from published reports, subject to scrutinous curation. Methods used for collection and alignments are described previously (9). In this release, names of some of the profiles changed to match the official HGED or MGED symbols of the corresponding TFs where applicable; their JASPAR IDs were not changed.

JASPAR FAM. The JASPAR FAM partition houses familial binding profiles (also referred to as 'consensus profiles') for 11 major structural classes of factors. The collection facilitates prediction of TF binding domain structures based on profile information alone (10). These models are especially suitable for gene- and genome-wide exploratory searches in cases where there is no prior knowledge of cognate factors. As only a fraction of TFs are well characterized, factor-specific profiles are lacking for most TFs. The familial models in JASPAR FAM can be used as proxy profiles for uncharacterized TFs within TF structural families known to bind similar target sequences. The construction and application of familial binding profiles is described in Ref. (10).

JASPAR phyloFACTS. This new subset of the database contains a set of matrices that are derived from evolutionarily conserved sequences in the regulatory regions of mammalian genes. The profiles were based on a recent comprehensive systematic survey of regulatory motifs (11), which used the phylogenetic relationship between human, mouse, rat and dog to discover conserved and overrepresented sequence motifs in the region 2 kb upstream and downstream from the RefSeq-based transcription start site of human genes. To construct the 174 matrix models, we scanned multiple sequence alignments from Ref. (11) for the conserved motifs reported in the paper and used the detected sites to derive matrices that can be regarded as common mammalian matrix profiles. The resulting matrices represent numerous putative binding sites, providing potentially high-matrix granularity. We compared the 174 phyloFACTS matrices to the existing matrices in the JASPAR CORE using Pearson correlation coefficient (PCC) as a measure for matrix similarity. In total, similarity higher than 0.8 (an empirical value PCC often used to indicate strong correlation) was observed for 27% of the JASPAR CORE mammalian matrices. Inspection of this significant, but limited overlap indicates that potential binding sites for many TFs computationally detected in Ref. (11) have not been experimentally characterized to date; on the other hand, their computational method is unable to detect many of the experimentally verified binding sites from JASPAR CORE that either have low information content or are predominantly found in long-range enhancers. This and the validation procedure (see below) strongly indicate that the JASPAR phyloFACTS and CORE sets complement each other.

COMBINATION OF JASPAR CORE AND phyloFACTS DATABASES ENHANCES BINDING SPACE COVERAGE

We wanted to assess the predictive power of the combination of phyloFACTS and JASPAR CORE, and to compare the coverage of the union against the coverage of vertebrate subset of the JASPAR CORE itself, as well as the TRANSFAC database (version 8.4) (12). Co-regulated gene expression is a consequence of the co-occurrence of similar features, such as common TFBSs, in the promoter regions of a gene set. If a collection of matrix profiles contains relevant data for these features, we should be able to distinguish a set of co-regulated genes from a random set. We compiled two sets of co-regulated genes: one set of 16 genes known to be important in the Wnt signalling pathway and a set of 20 histone genes [the results on the Wnt dataset are explained in more detail in P.J. de Bleser *et al*. (submitted for publication)]. By applying a feature selection and classification procedure described in detail in Supplementary Data, we were able to show that the use of JASPAR CORE/phyloFACTS increases both specificity and sensitivity of predictions compared with either JASPAR CORE alone, or with TRANSFAC. The complementarity of JASPAR CORE and phyloFACTS is further shown in the selection of matrix profiles used as classification attributes. Of the four chosen attributes for the Wnt dataset, two are from JASPAR CORE and two are from phyloFACTS. One of the selected matrices of phyloFACTS (JASPAR ID: PF0073) comprises a motif that is associated with the TF Lef1, known to be essential in the Wnt signalling pathway. This motif is absent from the attributes selected by either JASPAR CORE or TRANSFAC. For the histones, the 44 selected attributes are evenly distributed between the two matrix sets (20 from JASPAR CORE and 24 from phyloFACTS),

Table 1. Summary of the database components

Data collection	JASPAR CORE	JASPAR FAM	JASPAR phyloFACTS
Keywords	Non-redundant, literature curated models (9)	Meta-models for structural classes of TFs (10)	Data-mined profiles using phylogenetic pattern finding (11)
Number of models	123	11	174
Mean information content (bits)	12.1	8.1	15.6
Mean profile sequence depth	33.9	100[a]	1598.5
Number of structural TF classes	26	11	NA[b]
Anonymous MySQL access[c]	JASPAR_CORE_2005	JASPAR_FAM_2005	JASPAR_PHYLOFACTS_2005

[a]The sequence depths for the meta-models in JASPAR FAM is normalized to 100.
[b]As the patterns in JASPAR phyloFACTS are not experimentally linked to cognate factors, the structural classes are unknown.
[c]The MySQL server is at jasper.genereg.net (user: anonymous, password: jaspar).

indicating that phyloFACTS serves as a complementary profile collection to JASPAR CORE (see Supplementary Data for details of the validation procedure).

AVAILABILITY, API AND DISTRIBUTION

The JASPAR web portal address provides a graphical interface for casual users, enabling browsing and database search functions, as well as basic sequence search functionality for selected profiles. In addition, novel profiles entered by users can be compared to profiles in the three datasets using matrix alignment algorithms (10).

The TFBS module for the Perl programming language (13) has extensive support for the JASPAR database and can be considered an application programming interface to the database. This approach is recommended for power users.

The JASPAR database and underlying datasets are available for download with no restrictions from the JASPAR portal. In addition, users can access the underlying MySQL database anonymously (Table 1).

FUTURE DIRECTIONS: EXPANSION, USER SUBMISSION AND UNIVERSAL DATA MODEL

In the future, we shall see an increasing number of TFBS models produced by diverse approaches. Those methods will vary in scope, reliability and the depth of biological validation. For instance, the recently launched ENCODE project (14) is anticipated to produce a large number of genome-wide chip–CHIP (15) experiments for a wide selection of TFs, potentially leading to a new cadre of profiles. Therefore, the new meta-database model provides the required flexibility for the growth of the JASPAR collections. JASPAR CORE will remain faithful to the original purpose as the central open-access repository of high-quality, experimentally verified profiles that will continue to expand by expert curation.

As a community resource, JASPAR is introducing a user data submission mechanism which enables users to submit (i) individual models with sufficient experimental evidence to JASPAR CORE, or (ii) whole collections of annotated matrix profiles that share the scope, origin and mechanism of generation. The submission form is available at http://jaspar.genereg.net.

As part of a larger effort, the development of a universal data model and the associated input and curation tools for annotated TF binding data is under way. The ultimate goal is not only to have an open-access binding profile repository, but also the most comprehensive collection of information about TFs and their binding sites. We strongly encourage researchers supporting the open-access database model to use JASPAR as a means for sharing models and datasets with the research community.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The research of D.V., P.J.dB., K.V. and F.vR. was supported by VIB, the Fund for Scientific Research (FWO)-Flanders, the Geconcerteerde Onderzoeksacties of Ghent University, and Interuniversity Attraction Pools Programme—Belgian Science Policy. B.L. was supported by Pharmacia Corporation (now Pfizer), Swedish Research Council (Vetenskapsrådet), and Functional Genomics Programme (FUGE) in the Research Council of Norway. Funding to pay for the Open Access publication charges for this article was also provided by FUGE.

Conflict of interest statement. None declared.

REFERENCES

1. Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nature Rev. Genet*., **5**, 276–287.
2. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
3. Aerts,S., Van Loo,P., Thijs,G., Mayer,H., de Martin,R., Moreau,Y. and De Moor,B. (2005) TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis. *Nucleic Acids Res*., **33**, W393–W396.
4. Berezikov,E., Guryev,V. and Cuppen,E. (2005) CONREAL web server: identification and visualization of conserved transcription factor binding sites. *Nucleic Acids Res*., **33**, W447–W450.
5. Ho Sui,S.J., Mortimer,J.R., Arenillas,D.J., Brumm,J., Walsh,C.J., Kennedy,B.P. and Wasserman,W.W. (2005) oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. *Nucleic Acids Res*., **33**, 3154–3164.
6. Marinescu,V.D., Kohane,I.S. and Riva,A. (2005) MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes. *BMC Bioinformatics*, **6**, 79.
7. Roepcke,S., Grossmann,S., Rahmann,S. and Vingron,M. (2005) T-Reg Comparator: an analysis tool for the comparison of position weight matrices. *Nucleic Acids Res*., **33**, W438–W441.
8. Sandelin,A., Wasserman,W.W. and Lenhard,B. (2004) ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res*., **32**, W249–W252.
9. Sandelin,A., Alkema,W., Engstrom,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*., **32**, D91–D94.
10. Sandelin,A. and Wasserman,W.W. (2004) Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J. Mol. Biol*., **338**, 207–215.
11. Xie,X., Lu,J., Kulbokas,E.J., Golub,T.R., Mootha,V., Lindblad-Toh,K., Lander,E.S. and Kellis,M. (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
12. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al*. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*., **31**, 374–378.
13. Lenhard,B. and Wasserman,W.W. (2002) TFBS: computational framework for transcription factor binding site analysis. *Bioinformatics*, **18**, 1135–1136.
14. ENCODE Project Consortium (2004), The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**, 636–640.
15. Mukherjee,S., Berger,M.F., Jona,G., Wang,X.S., Muzzey,D., Snyder,M., Young,R.A. and Bulyk,M.L. (2004) Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nature Genet*., **36**, 1331–1339.