D102–D106 *Nucleic Acids Research, 2008, Vol. 36, Database issue* *Published online 15 November 2007*
doi:10.1093/nar/gkm955

JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update

Jan Christian Bryne[1], Eivind Valen[2], Man-Hung Eric Tang[2], Troels Marstrand[2], Ole Winther[2,3], Isabelle da Piedade[4], Anders Krogh[2], Boris Lenhard[1,5,*] and Albin Sandelin[2,*]

[1]Computational Biology Unit, Bergen Center for Computational Science, University of Bergen, Thormøhlensgate 55, N-5008 Bergen, Norway, [2]The Bioinformatics Centre, Department of Molecular Biology & Biotech Research and Innovation Centre, University of Copenhagen, Ole Maaløes Vej 5, DK-2100 København Ø, [3]Informatics and Mathematical Modeling, Building 321, Technical University of Denmark, DK-2800 Kgs. Lyngby, [4]Center for Comparative Genomics, Institute of Biology, University of Copenhagen, Universitetsparken 15, DK-2100 Copenhagen Ø, Denmark and [5]Sars Centre for Marine Molecular Biology, University of Bergen, Thormøhlensgate 55, N-5008 Bergen, Norway

Received September 14, 2007; Revised October 15, 2007; Accepted October 16, 2007

ABSTRACT

JASPAR is a popular open-access database for matrix models describing DNA-binding preferences for transcription factors and other DNA patterns. With its third major release, JASPAR has been expanded and equipped with additional functions aimed at both casual and power users. The heart of the JASPAR database—the JASPAR CORE sub-database—has increased by 12% in size, and three new specialized sub-databases have been added. New functions include clustering of matrix models by similarity, generation of random matrices by sampling from selected sets of existing models and a language-independent Web Service applications programming interface for matrix retrieval. JASPAR is available at http://jaspar.genereg.net.

INTRODUCTION

Computational analysis of regulatory properties of DNA is most often based on the use of matrix models describing binding preferences of transcription factors, or other DNA patterns. Such matrices are based on sets of known or inferred sites for a DNA-binding protein, and can be scanned over genomic sequences to predict novel binding sites (1,2). JASPAR is the most comprehensive open-access database holding such models. The heart of JASPAR is the JASPAR CORE sub database, holding curated, non-redundant matrix models from multi-cellular eukaryotes. The methodology for JASPAR CORE curation has been described previously (3). JASPAR CORE is now a standard resource in gene regulation bioinformatics and is used as a matrix set in a wide variety of other services [for instance (4–9)], and large-scale projects (10,11). Besides JASPAR CORE, the database contains several sub-databases (JASPAR Collections) holding matrix models produced by different methods and for different purposes (Table 1).

Here we present the recent JASPAR expansion, which includes a significant increase of the JASPAR CORE content and an addition of three new sub-databases focusing on core promoter patterns, splice sites and motifs detected in vertebrate highly conserved non-coding elements, respectively. In addition, we present several unique functional features in the web interface aimed at both casual and power users, including statistics on expected number of predictions each matrix will yield at several different thresholds in random sequences generated by three commonly encountered sequence background models, dynamic clustering of matrices by similarity and generation of random matrices using a selected set of matrices as background model.

*To whom correspondence should be addressed. Boris Lenhard: Tel: +47 55584362; Fax: +47 55584295; Email: boris.lenhard@bccs.uib.no
Albin Sandelin: Tel: +45 52321285; Fax: +45 35325669; Email: albin@binf.ku.dk

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

Table 1. JASPAR databases

Database	Number of models	Scope	Species coverage	When to use
JASPAR CORE	138	Curated, non-redundant matrix models	Multi-cellular eukaryotes	'Standard' promoter analysis
JASPAR FAM	11	Familial 'consensus' patterns for major structural families of transcription factors	Multi-cellular eukaryotes	Matrix-to-matrix comparison and classification, or as prior knowledge for pattern finders
JASPAR PHYLOFACTS	174	Evolutionary conserved patterns in 5′ promoter regions	Multi-cellular eukaryotes	As a complement to JASPAR CORE for large-scale studies
JASPAR POLII	13	Core promoter element models	Multi-cellular eukaryotes	Core promoter analysis
JASPAR CNE	233	Motifs overrepresented in vertebrate highly conserved non-coding elements	Human	Analysis of regulatory content of long-range enhancers
JASPAR SPLICE	6[a]	Splice sites	Human[a]	Splice site analysis

[a]Expansion under way.

RESULTS

Here we briefly describe the new data and functional features; more detailed descriptions are available at the documentation at the web site.

Expansion of JASPAR CORE

The JASPAR CORE database holds a curated set of transcription factor-binding profiles from multi-cellular eukaryotes: this is a unique feature with respect to databases of similar scope. We have extended JASPAR CORE with 15 new, high-quality profiles from recent experimental literature, increasing the total number of JASPAR CORE models to 138 (Table 1). In addition, annotation for all models in the database has been updated [e.g. to standard gene symbols from Entrez Gene (12)] and expanded. Prompted by user feedback, several existing matrices have been updated or corrected.

New sub-databases

Existing and new sub-databases within JASPAR and their specific features are described in Table 1. Since the last update, we have added three new sub-databases, which are briefly described below (see the web documentation for details):

JASPAR POLII. The large body of novel data pertaining transcription start sites (13,14) has triggered a new interest in computational studies of core promoters. The JASPAR POLII sub-database holds 13 known DNA patterns linked to RNA polymerase II core promoters, such as the Inr and BRE elements, each based on experimental evidence: each model must be constructed using five or more experimentally verified sites. An important difference to the transcription factor profiles in JASPAR CORE is that patterns here do not necessarily have a specified protein that binds them [See Ref. (15) for a review on core promoter patterns]. When possible, profiles were extended by 2 nt more than the core motif. We consistently report positions relative to the TSS as the position of 5′ and 3′ edge of the matrix.

JASPAR CNE. Highly conserved non-coding elements (CNEs) are a distinctive feature of metazoan genomes. Many of them can be shown to act as long-range enhancers that drive expression of genes that are themselves regulators of core aspects of metazoan development and differentiation. Since they act as regulatory inputs, attempts at deciphering the regulatory content of these elements have started (16–18). JASPAR CNE is a collection of 233 matrix profiles derived by Xie *et al.* (19) by clustering of overrepresented motifs from human conserved non-coding elements. While the biochemical and biological role of most of these patterns is still unknown, Xie *et al.* have shown that the most abundant ones correspond to known DNA-binding proteins, among them is the insulator-binding protein CTCF. These matrix profiles will be useful for further characterization of regulatory inputs in long-range developmental gene regulation in vertebrates.

JASPAR SPLICE. This small collection contains matrix profiles of human canonical and non-canonical splice sites, as matching donor:acceptor pairs. It currently contains only six highly reliable profiles (two canonical and four non-canonical) obtained from human genome (20). In the future, we shall include additional eukaryotic species, as well as new models for exonic splicing enhancers (ESE) and inhibitors (ESI).

Extended functionality

In addition to data extension, we have implemented a number of functional improvements in the web interface of the JASPAR database. These range from static statistics, such as expected number of hits on typical DNA sequence for any factor, to dynamic tools for similarity-based profile clustering and for generating random profiles based on a subset of known profiles.

Web service interface. The JASPAR database can now be reached remotely through a new Web Service interface. Current functionality includes retrieval of profiles by name, by identifier and by searching profile annotations. The purpose of providing an external application programming interface (API) is to simplify the utilization of JASPAR in distributed applications and in scientific workflows created in workflow editors like Triana (21), BPEL (http://www.bpelsource.com/) or Taverna (22). Other benefits include platform- and language-independent access, as well as constant up-to-date access to

the database over time. The API is implemented as a WS-I compliant Web Service, identical to the technology used for the services made available through the EMBRACE Network of Excellence (www.embracegrid.info), and the Web Service technology chosen by the European Bioinformatics Institute (EBI) (23). Its basic usage is described in tutorials at the JASPAR web site. The WSDL describing this service can be found at: http://api.bioinfo.no/wsdl/JasparDB.wsdl. Further information about the Web Service, including example clients in Java and Python, is available on the Jaspar web site and in the WSDL file.

Expected predictions/base-pair statistics for all models. An important problem with genome-wide scanning with matrix models is the limited information content in a typical matrix, resulting in numerous spurious hits just due to sequence background (1,2). The number of false positives varies considerably between factors and also depends on what type of sequences that models are applied to, user-defined cutoffs and to a more limited extent on the type of scoring scheme used. For a first-glance assessment of the rate of spurious predictions of a given model, we apply the model to three distinct sequence sets: known promoters from the EPD database (24), CpG islands and randomly selected genomic DNA, respectively. For different score thresholds, we plot the mean number of hits per 1000 nt for each sequence set. The resulting bar plots are available for each JASPAR matrix (Figure 1).

Dynamic clustering by similarity and creation of familial binding profiles from a given profile subset. Many transcription factors bind similar targets and it is often helpful to cluster similar binding profiles to generate familial binding profiles—models describing a set of matrices (25). Part of this problem is matrix profile comparison and alignments, explored by several researchers (25–30). Recently, Mahony *et al*. (27,28) made a comprehensive study on alignments of matrices and construction of familial binding profiles, resulting in the STAMP tool, which is now used within JASPAR to cluster matrix models. Hierarchical clustering is performed on a selected set of matrices using the UPGMA algorithm with a Pearson Correlation Coefficient distance metric. Then the optimal number of clusters is selected using a log variant of the Calinski and Harabasz statistic [See Ref. (27) for details]. Finally, the clusters are partitioned and a familial binding profile is created for each cluster using iterative refinement (a multiple alignment method). An example is shown in Figure 1.

Dynamic random profile generation. In many computational studies, it is helpful to have a set of 'random' matrices. This is particularly true for assessment of distances between putative sites and reference points as transcription start sites, and also for matrix-to-matrix comparisons. In these cases, it is desired that the randomized matrices should share properties with the true matrix set—for instance having the same nucleotide content and/or the same general information content.

Within any JASPAR sub-database, users can select a subset of matrices, which will then be used to generate random matrices using one of two methods: (i) Permutations: Columns of the selected matrices are shuffled: either constrained to shuffling of columns within each matrix or between all selected matrices. (ii) Probabilistic sampling: This enables the users to generate random Position Frequency Matrices from selected profiles. In our model, each random column is sampled from a posterior distribution—a 4D Dirichlet mixture distribution. The posterior distribution has two contributions: a multinomial with counts of columns selected as in (i), and a Dirichlet mixture prior trained from all observed nucleotides in the JASPAR database. We assume that column positions are independent.

DISCUSSION

We have presented a significant update to the JASPAR database, including an expansion of the core database, three new sub-databases and many new utilities. The new web service interface enables easy interaction with scientific workflows and an increasing number of programming languages that support this technology. We project that the new features, together with the open-access policy, will further consolidate the JASPAR database as a standard resource in the field of gene regulation bioinformatics.

Towards a comprehensive set of models for most known transcription factors

The lack of models for the binding specificity of most transcription factors is a significant bottleneck for comprehensive computational analysis of genomes. Only a fraction of transcription factors have been characterized in enough detail to allow the construction of adequate models of their binding specificity. This problem is being solved in two principally different ways. First, tiling array approaches for measuring binding preferences *en masse* are being developed (31); these technologies show great promise and are expected to make their mark on the field in the near future. Second, a wealth of *cis*-regulatory elements, characterized in painstaking detail, is hidden in experimental literature; many of these sites are not included in any database. There is a growing awareness of this problem in the field, resulting in online open-access databases such as ORegAnno (32) and PAZAR (33), where one of the goals is to house expert-curated binding sites. We are currently developing services to enable cross-talk with these databases to enable matrix models built on curated sites that exceed a certain quality threshold. JASPAR, ORegAnno and PAZAR face the same challenge: to build models or sites, it is necessary to mine the literature, which inevitably means that the curators will miss many important studies. The only long-term solution would be a requirement by scientific journals for researchers to deposit protein–DNA interactions in public databases prior to publication, much in the same way as mRNAs must be submitted to Genbank (34). Part of such a system will be to establish a minimal standard for

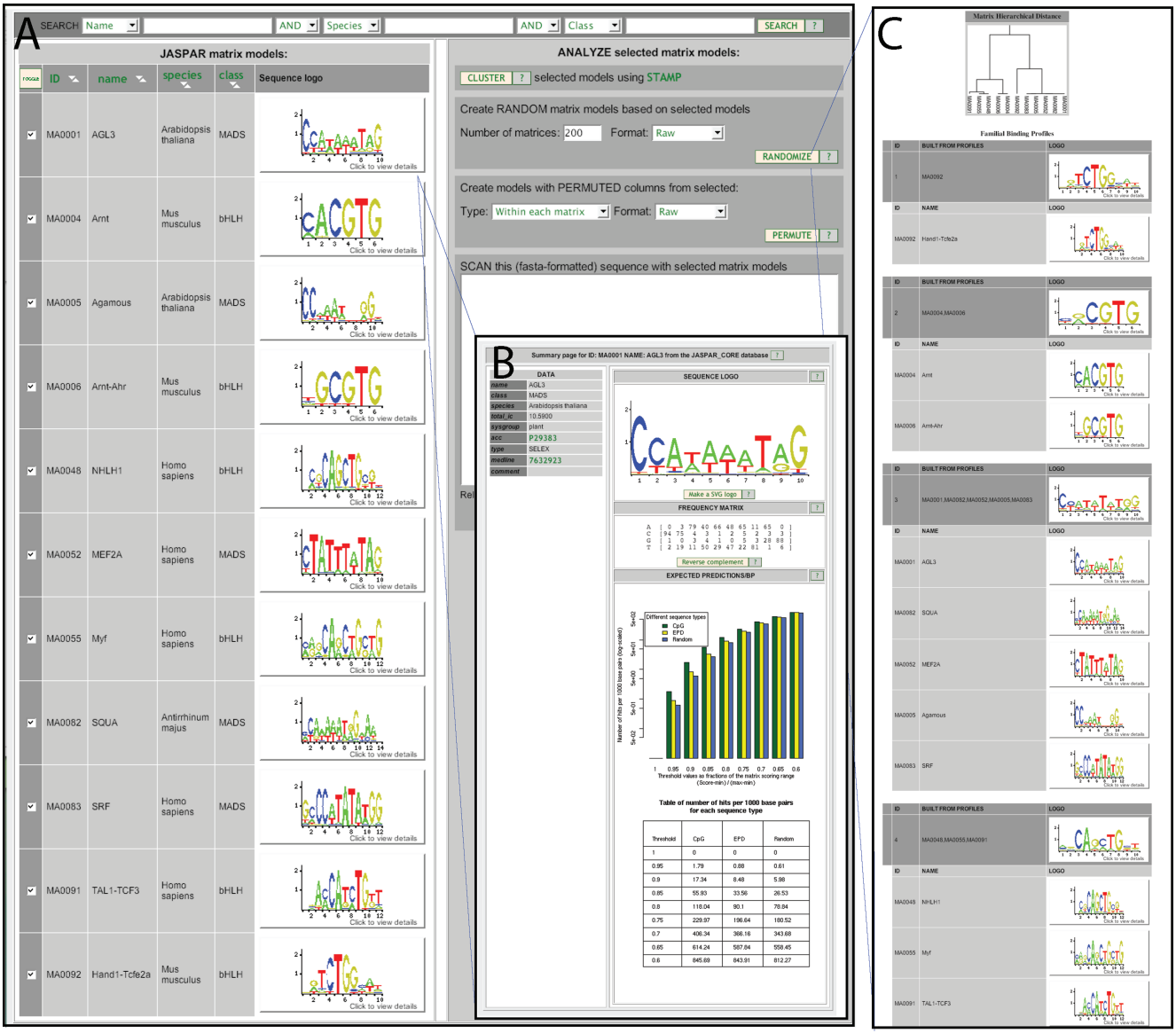

Figure 1. New features in the JASPAR database web interface. **(A)** A listing of matrices in the JASPAR-CORE database resulting from selection of MADS and bHLH-type factors. These models are used in the clustering analysis in panel C. **(B)** A pop-up window showing detail information on the MA0001 model, with expected predictions/bp statistics. **(C)** Dynamic clustering of selected profiles. At the top, a dendrogram describing the similarities of the input profiles is shown. Clusters of similar modes are merged into familial binding profiles, shown below. In this case, two larger clusters are produced, corresponding to bHLH and MADS type matrices. Two smaller clusters correspond to outliers in both groups.

reporting these interactions, much like the MIAME standard (35) for microarray data. As before, JASPAR team is always prepared to incorporate new matrices and matrix sets provided by external contributors.

Data availability

All the data in JASPAR are available without any restrictions, either from the web interface, as flat files or through the Web service interface.

ACKNOWLEDGEMENTS

Thanks to Katsuya Shigesada for pointing out errors in matrix MA0002, Shaun Mahony and Panayiotis V. Benos for generously sharing the STAMP code and general helpfulness and Vladimir B. Bajic for kindly providing the frequency matrices for JASPAR SPLICE. E.V., M.-H.E.T., T.M., O.W., A.K. and A.S. were supported by a grant from the Novo Nordisk foundation to the Bioinformatics Center. I.P. was supported by a grant from Carlsberg Foundation (21-00-0680). J.C.B. was supported by EMBRACE—an EU Sixth Framework Network of Excellence. B.L. was supported by the Functional Genomics Programme (FUGE) of the Research Council of Norway, and a core grant from the Sars Centre. Funding to pay the Open Access publication charges for this article was provided by a grant from the Novo Nordisk Foundation and the Functional Genomics Programme of the Research Council of Norway.

Conflict of interest statement. None declared.

REFERENCES

1. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
2. Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276–287.
3. 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.
4. Lenhard,B., Sandelin,A., Mendoza,L., Engstrom,P., Jareborg,N. and Wasserman,W.W. (2003) Identification of conserved regulatory elements by comparative genome analysis. *J. Biol.*, **2**, 13.
5. Montgomery,S.B., Astakhova,T., Bilenky,M., Birney,E., Fu,T., Hassel,M., Melsopp,C., Rak,M., Robertson,A.G. *et al.* (2004) Sockeye: a 3D environment for comparative genomics. *Genome Res.*, **14**, 956–962.
6. Frith,M.C., Fu,Y., Yu,L., Chen,J.F., Hansen,U. and Weng,Z. (2004) Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res.*, **32**, 1372–1381.
7. 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.
8. 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.
9. 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.
10. The ENCODE Consortium, (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
11. Carninci,P., Kasukawa,T., Katayama,S., Gough,J., Frith,M.C., Maeda,N., Oyama,R., Ravasi,T., Lenhard,B. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
12. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **35**, D26–D31.
13. Muller,F., Demeny,M.A. and Tora,L. (2007) New problems in RNA polymerase II transcription initiation: matching the diversity of core promoters with a variety of promoter recognition factors. *J. Biol. Chem.*, **282**, 14685–14689.
14. Sandelin,A., Carninci,P., Lenhard,B., Ponjavic,J., Hayashizaki,Y. and Hume,D.A. (2007) Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat. Rev. Genet.*, **8**, 424–436.
15. Smale,S.T. and Kadonaga,J.T. (2003) The RNA polymerase II core promoter. *Annu. Rev. Biochem.*, **72**, 449–479.
16. Bailey,P.J., Klos,J.M., Andersson,E., Karlen,M., Kallstrom,M., Ponjavic,J., Muhr,J., Lenhard,B., Sandelin,A. *et al.* (2006) A global genomic transcriptional code associated with CNS-expressed genes. *Exp. Cell Res.*, **312**, 3108–3119.
17. Pennacchio,L.A., Ahituv,N., Moses,A.M., Prabhakar,S., Nobrega,M.A., Shoukry,M., Minovitsky,S., Dubchak,I., Holt,A. *et al.* (2006) In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, **444**, 499–502.
18. Pennacchio,L.A., Loots,G.G., Nobrega,M.A. and Ovcharenko,I. (2007) Predicting tissue-specific enhancers in the human genome. *Genome Res.*, **17**, 201–211.
19. Xie,X., Mikkelsen,T.S., Gnirke,A., Lindblad-Toh,K., Kellis,M. and Lander,E.S. (2007) Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc. Natl Acad. Sci. USA*, **104**, 7145–7150.
20. Chong,A., Zhang,G. and Bajic,V.B. (2004) Information for the coordinates of exons (ICE): a human splice sites database. *Genomics*, **84**, 762–766.
21. Majithia,S., Shields,M., Taylor,I. and Wang,I. (2004) Triana: a graphical web service composition and execution toolkit. In *Proceedings of the IEEE International Conference on Web Services*, pp. 514–524, http://www.trianacode.org.
22. Hull,D., Wolstencroft,K., Stevens,R., Goble,C., Pocock,M.R., Li,P. and Oinn,T. (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Res.*, **34**, W729–W732.
23. Labarga,A., Valentin,F., Anderson,M. and Lopez,R. (2007) Web services at the European bioinformatics institute. *Nucleic Acids Res.*, **35**, W6–W11.
24. Schmid,C.D., Perier,R., Praz,V. and Bucher,P. (2006) EPD in its twentieth year: towards complete promoter coverage of selected model organisms. *Nucleic Acids Res.*, **34**, D82–D85.
25. 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.
26. Hughes,J.D., Estep,P.W., Tavazoie,S. and Church,G.M. (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J. Mol. Biol.*, **296**, 1205–1214.
27. Mahony,S., Auron,P.E. and Benos,P.V. (2007) DNA familial binding profiles made easy: comparison of various motif alignment and clustering strategies. *PLoS Comput. Biol.*, **3**, e61.
28. Mahony,S. and Benos,P.V. (2007) STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.*, **35**, W253–W258.
29. Pietrokovski,S. (1996) Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res.*, **24**, 3836–3845.
30. Schones,D.E., Sumazin,P. and Zhang,M.Q. (2005) Similarity of position frequency matrices for transcription factor binding sites. *Bioinformatics*, **21**, 307–313.
31. Berger,M.F., Philippakis,A.A., Qureshi,A.M., He,F.S., Estep,P.W.III. and Bulyk,M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, **24**, 1429–1435.
32. Montgomery,S.B., Griffith,O.L., Sleumer,M.C., Bergman,C.M., Bilenky,M., Pleasance,E.D., Prychyna,Y., Zhang,X. and Jones,S.J. (2006) ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. *Bioinformatics*, **22**, 637–640.
33. Portales-Casamar,E., Kirov,S., Lim,J., Lithwick,S., Swanson,M.I., Ticoll,A., Snoddy,J. and Wasserman,W.W. (2007) PAZAR: a framework for collection and dissemination of cis-regulatory sequence annotation. *Genome Biol.*, **8**, R10.
34. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2007) GenBank. *Nucleic Acids Res.*, **35**, D21–D25.
35. Brazma,A., Hingamp,P., Quackenbush,J., Sherlock,G., Spellman,P., Stoeckert,C., Aach,J., Ansorge,W., Ball,C.A. *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.*, **29**, 365–371.