Nucleic Acids Research, 2006, Vol. 34, Database issue ***D63–D67***
doi:10.1093/nar/gkj116

ABS: a database of Annotated regulatory Binding Sites from orthologous promoters

Enrique Blanco[1,2,*], Domènec Farré[1,2], M. Mar Albà[1], Xavier Messeguer[2] and Roderic Guigó[1]

[1]Grup de Recerca en Informàtica Biomèdica, Institut Municipal d'Investigació Mèdica/Universitat Pompeu Fabra/Centre de Regulació Genòmica, C/Doctor Aiguader 80, 08003 Barcelona, Spain and
[2]Grup d'algorísmica i genètica, Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, C/Jordi Girona 1-3, 08034 Barcelona, Spain

Received August 1, 2005; Revised September 19, 2005; Accepted October 18, 2005

ABSTRACT

Information about the genomic coordinates and the sequence of experimentally identified transcription factor binding sites is found scattered under a variety of diverse formats. The availability of standard collections of such high-quality data is important to design, evaluate and improve novel computational approaches to identify binding motifs on promoter sequences from related genes. ABS (http://genome.imim.es/datasets/abs2005/index.html) is a public database of known binding sites identified in promoters of orthologous vertebrate genes that have been manually curated from bibliography. We have annotated 650 experimental binding sites from 68 transcription factors and 100 orthologous target genes in human, mouse, rat or chicken genome sequences. Computational predictions and promoter alignment information are also provided for each entry. A simple and easy-to-use web interface facilitates data retrieval allowing different views of the information. In addition, the release 1.0 of ABS includes a customizable generator of artificial datasets based on the known sites contained in the collection and an evaluation tool to aid during the training and the assessment of motif-finding programs.

INTRODUCTION

Expression of genes is regulated at many different levels, transcription of DNA being one of the most critical stages. Specific configurations of transcription factors (TFs) that interact with gene promoter regions are recruited to activate or modulate the production of a given transcript. Many of these TFs possess the ability to recognize a small set of genomic sequence footprints called TF-binding sites (TFBSs). These motifs are typically 6–15 bp long and in some cases, they show a high degree of variability. In addition, many motifs may ambiguously be recognized by members of different TF families. Because of these flexible binding rules, computational methods for the identification of regulatory elements in a promoter sequence tend to produce an overwhelming amount of false positives. However, the identification of conserved regulatory elements present in orthologous gene promoters (also called phylogenetic footprinting) has proved to be more effective to characterize such sequences (1–3). In fact, the ever-growing availability of more genomes and the constant improvement of bioinformatics algorithms hold great promise for unveiling the overall network of gene interactions of each organism (4).

Typically, computational methods to detect regulatory elements use their own training set of experimental annotated TFBSs. These annotations are usually collected from bibliography or from general repositories of gene regulation information, such as JASPAR (5) and TRANSFAC (6). However, each program establishes different criteria and formats to retrieve and display the data that forms the final training set, which makes the comparison between different methods very difficult. The construction of a good benchmark to evaluate the accuracy of several pattern discovery methods is therefore not a trivial procedure (7).

Although important efforts are being carried out to standardize the construction of collections of promoter regions (8) or the presentation of experimental data (9), there is a clear necessity to provide stable and common datasets for future algorithmic developments. In this direction, we present here the release 1.0 of the ABS database constructed from literature annotations that have been experimentally verified in human, mouse, rat or chicken.

DATABASE CONSTRUCTION

We have gathered from the literature a collection of experimentally validated binding sites that are conserved in at least

*To whom correspondence should be addressed. Tel: +34 93 2240891; Fax: +34 93 2240875; Email: eblanco@imim.es

two orthologous vertebrate promoters. The sites and the promoter sequences have been manually curated to ensure data consistency. The compiled data are suitable for training both classical pattern discovery programs and new emerging comparative methods. Flat files accomplishing the GFF standard format were used to store and query the information.

The GenBank accession number of the sequences in each bibliographical reference was utilized to retrieve the promoters. Such sequences were mapped on to the corresponding RefSeq annotations to ensure we were retrieving the actual promoter. The DBTSS database (10) was finally used to refine the annotation of the TSSs. Since it is the region in which most experimental studies have been focused on, we considered the sequence 500 bp immediately upstream the annotated TSS, as the promoter region in this first release.

For each annotated promoter, we only included experimentally tested sites in this proximal region whose motifs were correctly identified in at least two species, i.e. orthologous sites. Every known binding site was mapped on to the corresponding promoter sequence by BLASTN (11). Those matches that exhibited <80% of identity between the sequence of the original site and the mapped motif in the promoter region were rejected.

We computed BLASTN (11), CLUSTALW (12), AVID (13) and LAGAN (14) alignments of the orthologous promoters

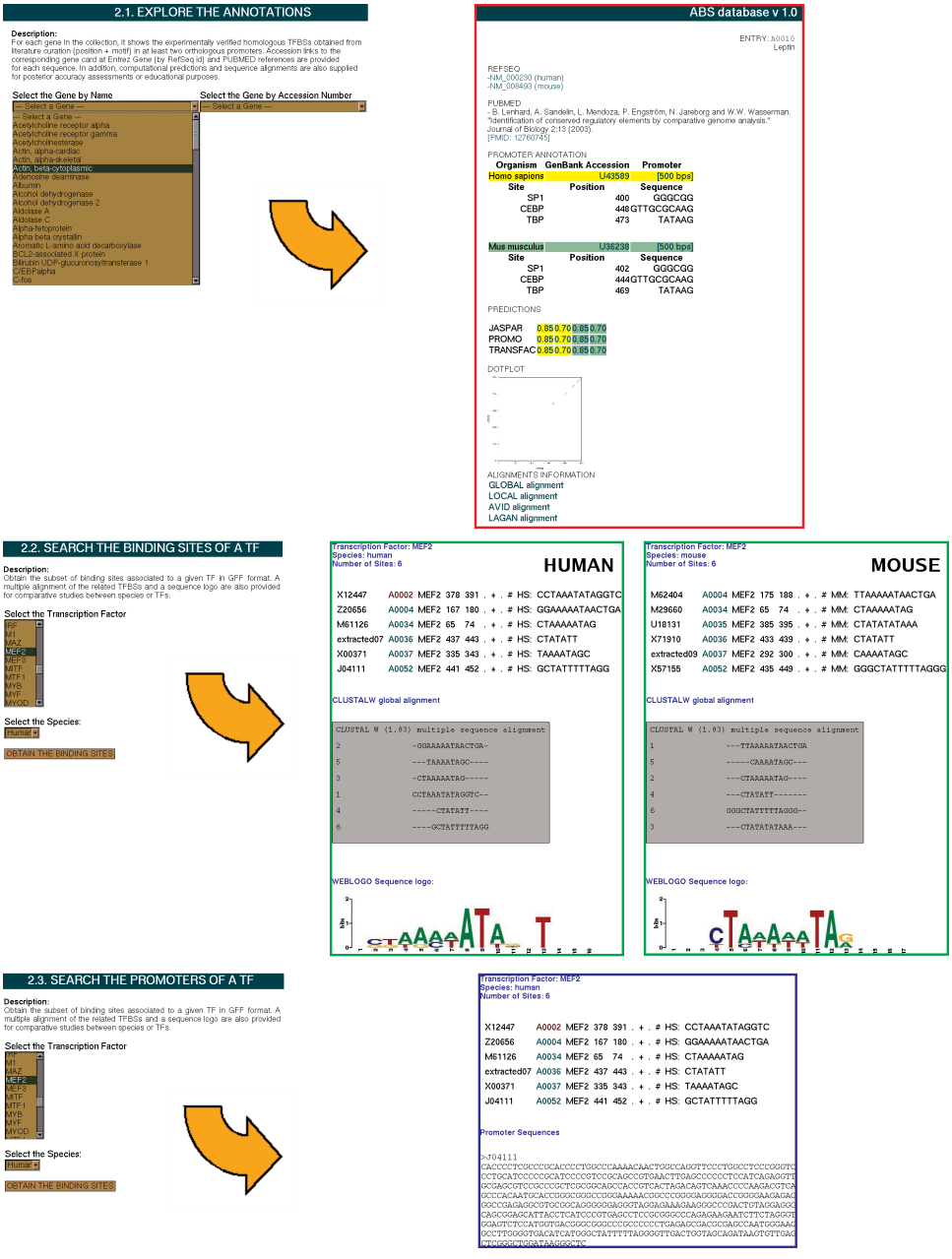

Figure 1. Examples of the ABS data retrieval system showing the annotation of a gene, the set of binding motifs from a given TF in human and mouse and the extraction of the promoter sequences containing such annotations.

from each gene. Moreover, we produced a dotplot of word matches with EMBOSS (15) to visualize unusually conserved regions. For comparative assessments, computational predictions using the JASPAR (5), TRANSFAC (6) and PROMO (16) collections of position weight matrices were calculated. A very restrictive threshold of 0.85 was used to remove those predicted TFBSs whose score was below this value, creating a first group of more reliable predictions. A second group of predictions was produced using a more flexible threshold of 0.70 (see the ABS website for further information about the scoring method).

DATABASE CONTENTS

Release 1.0 of ABS database contains 100 annotated orthologous genes, each entry corresponding to two or more species. The total number of promoter sequences is 211 (105 500 nt). There is a clear predominance of human and mouse annotations: 73 entries contain at least annotations for human–mouse orthologs. A total of 650 experimental binding sites from 68 different TFs are associated with ABS entries, covering 8624 nt. In average, three TFBSs per sequence have been mapped on to the promoters with an average length of 13.3 nt per site. The majority of the TFBSs are found near

Figure 2. Protocol to evaluate the accuracy of an external motif-finding program on a synthetic dataset generated by planting motifs from ABS in randomly generated sequences.

the TSS, as expected. The TFs that appear more frequently are TBP (14.6% of sites), SP1 (13.6%) and CEBP (5.6%). Those TFs are known to be part of the core of many eukaryotic promoters (see the ABS documentation for further details about the contents of the database).

WEB INTERFACE

Data retrieval

The ABS database can be accessed through a simple CGI/Perl-based web interface at http://genome.imim.es/datasets/abs2005/index.html. On-line documentation and tutorials are provided for each web service. The following functionalities are implemented in the current release (see Figure 1):

(i) For each gene in the collection, show the orthologous promoters and a list of experimentally verified TFBSs annotated on the corresponding sequence. Promoter sequence alignments, computational predictions, dotplots and cross-references to other well-known databases, such as GenBank, Entrez Gene and PubMed, are also provided for each annotation.
(ii) Retrieve all of the binding motifs associated with a given TF, filtering by species. Moreover, a global alignment of the motifs is provided and the corresponding sequence logo representation is displayed by using WebLogo (17). This information could be used to produce new profiles for subsequent detection of this TF in other promoters.
(iii) Retrieve all of the promoter sequences in which at least one binding site for a given TF was annotated. These sequences and the associated motifs could be used to generate datasets based on known sites to train motif-finding programs.
(iv) The gene catalogue, the promoter sequences, the collection of annotations, the sequence alignments and the computational predictions are also individually distributed in several flat files.

Benchmarking and evaluation tools

The ABS database aims to become a platform to evaluate new algorithms for the discovery of novel regulatory elements in a set of related gene promoters (e.g. orthologous promoters or co-regulated genes from microarray experiments). In addition to the data retrieval functions, two on-line applications are available to perform the benchmarking of such algorithms (see Figure 2):

(i) Constructor is a web server to produce synthetic datasets based on the ABS annotations. The design of the benchmark is highly flexible allowing to customize the number of sequences, their length, the background nucleotide distribution, the number of motifs to plant on them, the probability to plant a motif, the species and the TFs for which the associated motifs will be randomly selected from the known sites collection. The output consists of the artificial sequences with the embedded motifs, the list of motifs and a graphical representation of the occurrences in the sequences produced with the program gff2ps (18).
(ii) Evaluator is a web server to determine the accuracy of a set of predicted motifs in several sequences using a list of known binding sites as a reference set. Both sets must be provided by the user in GFF format. A complete accuracy assessment at both nucleotide and site levels is computed using the standard measures in the field (7,19).

CONCLUSIONS AND FUTURE WORK

The ABS database has been developed to fill the existing gap in the availability of consistent datasets to train and compare different pattern discovery programs. The lack of standard collections of TFBSs is specially serious in the case of phylogenetic footprinting data. The collection described here contains 650 experimental TFBSs identified in human, mouse, rat and chicken genes. Orthologous promoter sequences and their binding sites have been manually curated from bibliography. Supplementary information about the promoters is also provided for each entry. In addition, two web applications (Constructor and Evaluator) are included in this first release to facilitate the development of new motif-finding programs using the ABS annotations. In the next release, we plan to increase the number of annotations adding known sites in regulatory regions different from the proximal promoter and eventually incorporate binding motifs from other species.

ACKNOWLEDGEMENTS

E.B. is recipient of a predoctoral fellowship from Ministerio de Ciencia y Tecnología (Spain). D.F. is recipient of a predoctoral fellowship from Instituto Nacional de Bioinformática (Spain). This work has been supported by grants from Plan Nacional de I+D (BIO2000-1358-C02-02 and BIO2002-04426-C02-01), Ministerio de Ciencia y Tecnología (Spain) and from a FBBVA Bioinformatics grant. Funding to pay the Open Access publication charges for this article was provided by Ministerio de Cienciá y Tecnología (Spain).

Conflict of interest statement. None declared.

REFERENCES

1. Blanchette,M. and Tompa,M. (2002) Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.*, **12**, 739–748.
2. Elnitski,L., Hardison,R.C., Li,J., Yang,S., Kolbe,D., Eswara,P., O'Connor,M.J., Schwartz,S., Miller,W., Chiaromonte,F. *et al.* (2003) Distinguishing regulatory DNA from neutral sites. *Genome Res.*, **13**, 64–72.
3. Dermitzakis,E.T. and Clark,A.G. (2002) Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.*, **19**, 1114–1121.
4. Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nature Rev. Genet.*, **5**, 276–287.
5. 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.
6. 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.
7. Tompa,M., Li,N., Bailey,T.L., Church,G.M., De Moor,B., Eskin,E., Favorov,A.V., Frith,M.C., Fu,Y., Kent,W.J. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.
8. Barta,E., Sebestyen,E., Palfy,T.B., Toth,G., Ortutay,C.P. and Patthy,L. (2005) DoOP: Databases of Orthologous Promoters, collections of

clusters of orthologous upstream sequences from chordates and plants. *Nucleic Acids Res*., **33**, D86–D90.
9. Zhao,F., Xuan,Z., Liu,L. and Zhang,M.Q. (2005) TRED: a Transcriptional Regulatory Element Database and a platform for *in silico* gene regulation studies. *Nucleic Acids Res*., **33**, D103–D107.
10. Suzuki,Y., Yamashita,R., Sugano,S. and Nakai,K. (2004) DBTSS, DataBase of Transcriptional Start Sites: progress report 2004. *Nucleic Acids Res*., **32**, D78–D81.
11. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
12. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.
13. Bray,N., Dubchak,I. and Patcher,L. (2003) AVID: a Global Alignment Program. *Genome Res*., **13**, 97–102.
14. Brudno,M., Do,C., Cooper,G., Kim,M.F., Davydov,E., Green,E.D., Sidow,A. and Batzoglou,S. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignments of genomic DNA. *Genome Res*., **13**, 721–731.
15. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*., **16**, 276–277.
16. Farré,D., Roset,R., Huerta,M., Adsuara,J.E., Rosello,L., Albà,M.M. and Messeguer,X. (2003) Identification of patterns in biological sequences at the ALGEN server: PROMO and MALGEN. *Nucleic Acids Res*., **31**, 3651–3653.
17. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res*., **14**, 1188–1190.
18. Abril,J.F. and Guigó,R. (2000) gff2ps: visualizing genomic annotations. *Bioinformatics*, **16**, 743–744.
19. Burset,M. and Guigó,R. (1996) Evaluation of gene structure prediction programs. *Genomics*, **34**, 353–367.