Published online 24 April 2013 *Nucleic Acids Research, 2013, Vol. 41, Web Server issue* ***W531–W534***
doi:10.1093/nar/gkt288

PhysBinder: improving the prediction of transcription factor binding sites by flexible inclusion of biophysical properties

Stefan Broos[1,2], Arne Soete[1,2], Bart Hooghe[1,2], Raymond Moran[1,2], Frans van Roy[1,2] and Pieter De Bleser[1,2,*]

[1]Department for Molecular Biomedical Research, VIB and [2]Department of Biomedical Molecular Biology, Ghent University, B-9052 Ghent, Belgium

Received January 24, 2013; Revised March 24, 2013; Accepted March 31, 2013

ABSTRACT

The most important mechanism in the regulation of transcription is the binding of a transcription factor (TF) to a DNA sequence called the TF binding site (TFBS). Most binding sites are short and degenerate, which makes predictions based on their primary sequence alone somewhat unreliable. We present a new web tool that implements a flexible and extensible algorithm for predicting TFBS. The algorithm makes use of both direct (the sequence) and several indirect readout features of protein–DNA complexes (biophysical properties such as bendability or the solvent-excluded surface of the DNA). This algorithm significantly outperforms state-of-the-art approaches for *in silico* identification of TFBS. Users can submit FASTA sequences for analysis in the PhysBinder integrative algorithm and choose from >60 different TF-binding models. The results of this analysis can be used to plan and steer wet-lab experiments. The PhysBinder web tool is freely available at http://bioit.dmbr.ugent.be/physbinder/index.php.

INTRODUCTION

Proteins called transcription factors (TFs) are crucial for proper regulation of gene expression. They function by binding to regions of DNA called transcription factor binding sites (TFBS). Two different mechanisms contribute to the TF–DNA binding specificity needed for correct regulation of gene expression: a direct readout component caused by direct contact between the amino acids of the protein and the bases of the DNA and an indirect readout component caused by the global shape of the DNA and by conformational changes in both interaction partners (1,2). Traditional methods for predicting TFBS tend to look at the direct readout component alone and almost exclusively at the primary sequence. However, many of these widely used methods, such as positional weight matrices, are afflicted by many false positive predictions, indicating the need for incorporating other discriminative features (3). Recent evidence shows that sequence-dependent structural variations in the DNA account for a significant portion of the protein–DNA specificity (4–6). Thus, it is expected to be beneficial to include structural features and nucleotide dependencies in the prediction models. In a recent publication, we examined the effect of incorporating nucleotide position dependencies, which are related to the 3D structure of the DNA (7), on the prediction of TFBS (8). We also calculated structural features of the DNA and verified to which extent these features improve the prediction of TFBS. We found that incorporation of both types of data can substantially enhance the prediction of TFBS. Here, we present PhysBinder, a web tool based on the flexible Random Forest algorithm published in (8). We compiled >60 vertebrate TF models from various sources, but many more models will be offered in the future, as new data become available. Binding sites for these models can be visualized together with the ENCODE TFBS data track of UCSC genome (9) to get a useful insight in the genomic context of the inspected region.

INPUT AND OUTPUT

Input

The PhysBinder web tool is easy to use: for most parameters, we offer default configurations to ensure a quick and easy workflow. Users just provide their sequences of interest and select the appropriate TF model information.

*To whom correspondence should be addressed. Tel: +32 9 3313 601; Fax: +32 9 3313 609; Email: Pieterdb@DMBR.VIB-UGent.be

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

Sequences can be uploaded by one of the following means: (i) pasting a set of FASTA-formatted sequences in the input field; (ii) uploading a file with FASTA-formatted sequences; (iii) indicating genomic regions in the 'Fetch genomic regions' text field. Subsequently, a model and a threshold are to be selected. We provide three pre-calculated thresholds: 'Max. Precision', 'Max. F-Measure' and an average of these two measures. A custom threshold can also be selected.

More than 60 different TF models are now available on the PhysBinder website, but we expect to provide more models, as additional data become available. Most of the PhysBinder models are compiled from recent ENCODE data (10), but other sources were also used (see Materials and Methods for more information). TF models constructed from sequences that, according to the literature, clearly contain a sequence element associated with the TF are called 'direct evidence' models. When an alternative consensus sequence is found or when no consensus sequence is known for a particular TF, we call the models 'putative associated factors' (PAFs). Such a PAF might be a TF binding to multiple sequence elements, or it might be a common cofactor (hence 'putative associated factor'). By default, PhysBinder is configured to run in filter mode to speed up the calculations. In this mode, sequences are pre-filtered with a short positional weight matrice with low thresholds, minimizing the number of false-negative hits and effectively guaranteeing maximum recall.

Output

A summary table is given at the top of the results web page. This table can be sorted by model type or by input sequence, and, for each model or sequence, the number of hits is indicated. On this page, users can still alter the thresholds to increase or decrease the stringency of the binding site predictions. In the results section, binding sites are shown as sequences with a colored background (exemplified in Figure 1a). Clicking on the first nucleotide of such a colored sequence provides more details on the binding site. When clicked, a details window with the sequence logo of the binding site is shown (this logo was calculated on the model data), and the Random Forest score with a *P*-value is given as well. The relative position of the TFBS is shown, and if the genomic location of the sequence is known (because the user indicated this on the input page or performed a BLAT analysis of the sequence against a human or mouse reference genome), then the absolute coordinates of the binding sites are shown in the details window. Two additional options become available when the absolute position is known. For human sequences (hg18 and hg19), it is possible to integrate the most recent ENCODE data to get an overview of the transcription factors and RNA polymerase components that might bind within this genomic region. Predicted binding sites can also be visualized in the UCSC genome browser (11) (exemplified in Figure 1b). Using the aforementioned checkboxes, the sequences or those on the right side of the screen, models can be dynamically shown or hidden to aid the interpretation of the results.

Example

As an example (see Figure 1), we examined the analysis performed by Kyo *et al.* (12) of the promoter of the human TERT gene, encoding the catalytic subunit of telomerase. These researchers identified a core promoter of 181 bp responsible for the transcriptional activity of the TERT gene. This 181-bp region, consisting of the 5'-UTR and the upstream promoter region, contains two E-boxes bound by MYC *in vivo*. Between these E-boxes, Kyo *et al.* discovered and validated five GC-boxes that are bound by SP1. For illustrative purposes, we used the PhysBinder tool to look for SP1, MYC and TBP binding sites with default threshold settings in the same sequence they used (12), and we were readily able to confirm their findings. We unmistakably found the five SP1 binding sites flanked by two MYC binding sites, as reported in the initial publication. No TATA-box was found, and this promoter was reported to lack such box (13).

TECHNICAL DETAILS

Web tool

The web tool is hosted on a Linux CentOS 5 server with 32 GB of RAM, an Apache 2.2.3 web server, and PHP version 5.1.6. Web pages are written in the PHP and Javascript scripting languages. To map input sequences to mouse (mm10) or human (hg19) reference genomes, we use gfServer and Client binaries from UCSC, which makes it possible to BLAT sequences (11). ENCODE tracks are obtained from UCSC Genome (9). Sequences can be fetched from 16 different species, obtained from UCSC Genome. Extensive help documentation is available on the PhysBinder website, including guidelines and tutorials to facilitate the interpretation of the PhysBinder results.

Backend and models

The backend of PhysBinder is programmed in a combination of Perl and R-script. The Random Forest classifier used in the backend is the 'FastRandomForest' implementation. This is a multithreaded implementation of the Random Forest classifier in the Weka statistical package (14). In our models, we use a Random Forest with 100 trees. Most models are built from available ENCODE data of tier 1 cell lines, except for Esrrb (15), ETS1 (16), KLF4 (15), NANOG (15), Nmyc (15), STAT3 (15), TBP (17), Tfcp2l1 (15), TP53 (18) and Zfx (15). All sequences were first aligned using the multiple EM (expectation maximization) for motif elicitation (MEME) motif aligner (19) on the STEVIN supercomputing infrastructure of Ghent University. To ensure the quality of input data, the resulting aligned sequence motifs were then manually searched for in the literature. If a motif is not yet reported in literature, the resulting model is called a PAF. Otherwise, the model is termed a direct evidence model. When available, 100 sequences were used to build the model. The other sequences were used for validation. More information on the different steps of the algorithm and on its validation has been reported by us previously (8). Details about all models are available on

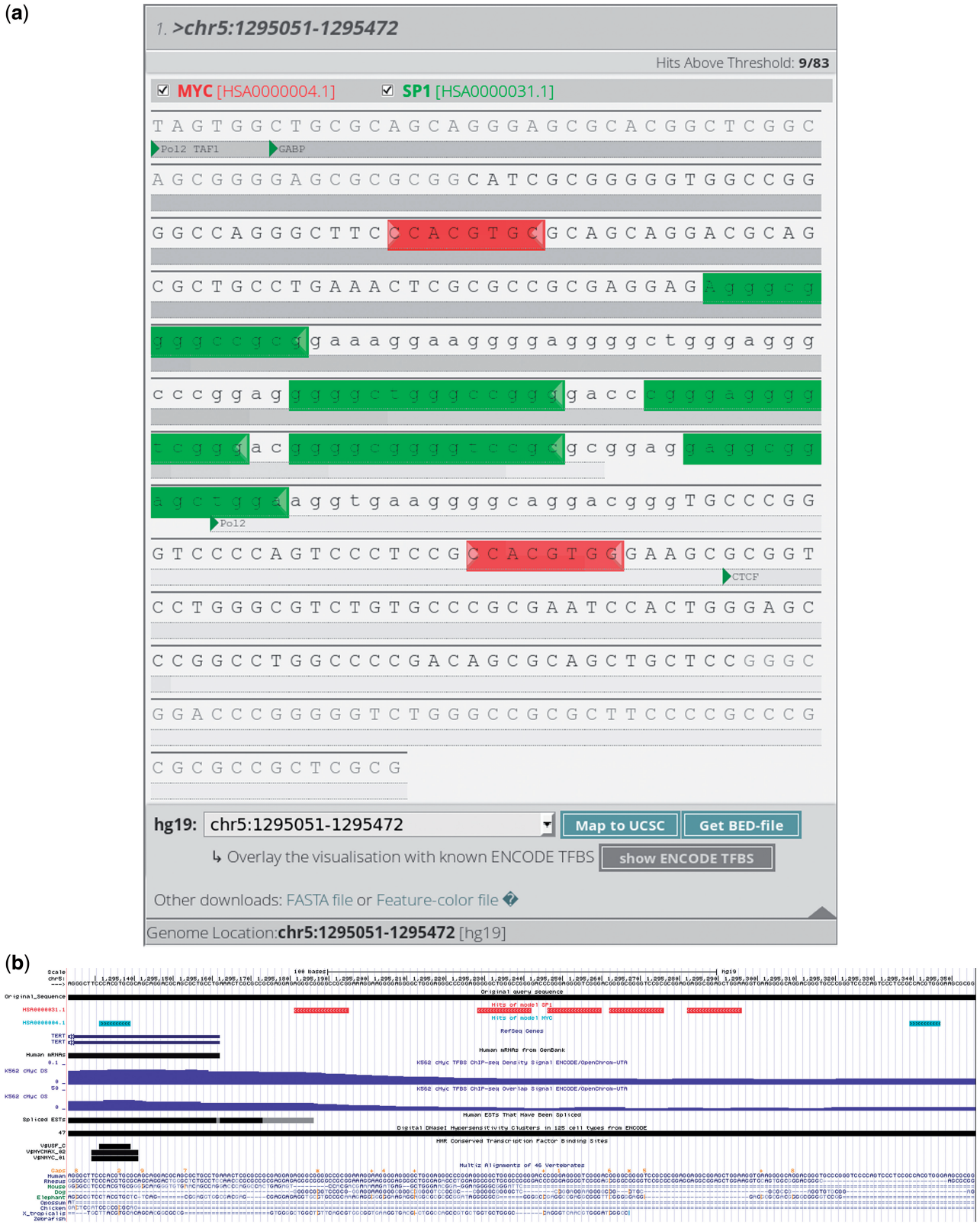

Figure 1. Example output of the PhysBinder tool. All predicted TFBS match the experimentally determined locations reported by Kyo *et al.* (12). (**a**) Detail of the results window: MYC binding sites (E-box) [HSA0000004.1] are shown in red. SP1 binding sites (GC-box) [HSA0000031.1] are shown in green. The default threshold ('Average') was used for both models. Gray shaded bars indicate overlapping ENCODE tracks (9). The checkboxes below the sequence indicate the different ENCODE tracks visualized in this sequence. (**b**) Both models were visualized in the UCSC Genome Browser (11). MYC binding sites are indicated in blue, whereas SP1 binding sites are in red.

the 'models' page, where an overview can be found of all the features contained in the models, together with performance measures that were calculated on external test sets.

ACKNOWLEDGEMENTS

The authors thank Dr. Amin Bredan for critical reading and editing of the article. The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were partially provided by Ghent University, the Hercules Foundation and the Flemish Government—Department EWI.

FUNDING

Agency for Innovation through Science and Technology in Flanders: [091213 to S.B.]; Research Foundation - Flanders (FWO): [G.0235.10 to F.vR. and P.D.B.]. Funding for open access charge: Department for Molecular Biomedical Research, VIB, Ghent, Belgium.

Conflict of interest statement. None declared.

REFERENCES

1. Michael Gromiha,M., Siebers,J.G., Selvaraj,S., Kono,H. and Sarai,A. (2004) Intermolecular and intramolecular readout mechanisms in protein-DNA recognition. *J. Mol. Biol.*, **337**, 285–294.
2. Steffen,N.R., Murphy,S.D., Tolleri,L., Hatfield,G.W. and Lathrop,R.H. (2002) DNA sequence and structure: direct and indirect recognition in protein-DNA binding. *Bioinformatics*, **18(Suppl. 1)**, S22–S30.
3. Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276–287.
4. Honig,B. and Rohs,R. (2011) Biophysics: flipping Watson and Crick. *Nature*, **470**, 472–473.
5. Rohs,R., West,S.M., Sosinsky,A., Liu,P., Mann,R.S. and Honig,B. (2009) The role of DNA shape in protein-DNA recognition. *Nature*, **461**, 1248–1253.
6. Parker,S.C., Hansen,L., Abaan,H.O., Tullius,T.D. and Margulies,E.H. (2009) Local DNA topography correlates with functional noncoding regions of the human genome. *Science*, **324**, 389–392.
7. Tomovic,A. and Oakeley,E.J. (2007) Position dependencies in transcription factor binding sites. *Bioinformatics*, **23**, 933–941.
8. Hooghe,B., Broos,S., van Roy,F. and De Bleser,P. (2012) A flexible integrative approach based on random forest improves prediction of transcription factor binding sites. *Nucleic Acids Res.*, **40**, e106.
9. Rosenbloom,K.R., Sloan,C.A., Malladi,V.S., Dreszer,T.R., Learned,K., Kirkup,V.M., Wong,M.C., Maddren,M., Fang,R., Heitner,S.G. *et al.* (2013) ENCODE Data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res.*, **41**, D56–D63.
10. Dunham,I., Kundaje,A., Aldred,S.F., Collins,P.J., Davis,C.A., Doyle,F., Epstein,C.B., Frietze,S., Harrow,J., Kaul,R. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
11. Kuhn,R.M., Haussler,D. and Kent,W.J. (2013) The UCSC genome browser and associated tools. *Brief Bioinform.*, **14**, 144–161.
12. Kyo,S., Takakura,M., Taira,T., Kanaya,T., Itoh,H., Yutsudo,M., Ariga,H. and Inoue,M. (2000) Sp1 cooperates with c-Myc to activate transcription of the human telomerase reverse transcriptase gene (hTERT). *Nucleic Acids Res.*, **28**, 669–677.
13. Horikawa,I., Cable,P.L., Afshari,C. and Barrett,J.C. (1999) Cloning and characterization of the promoter region of human telomerase reverse transcriptase gene. *Cancer Res.*, **59**, 826–830.
14. Hall,M., Frank,E., Holmes,G., Pfahringer,B., Reutemann,P. and Witten,I.H. (2009) The WEKA data mining software. *ACM SIGKDD Explorations Newslett.*, **11**, 10.
15. Chen,X., Xu,H., Yuan,P., Fang,F., Huss,M., Vega,V.B., Wong,E., Orlov,Y.L., Zhang,W., Jiang,J. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–1117.
16. Hollenhorst,P.C., Chandler,K.J., Poulsen,R.L., Johnson,W.E., Speck,N.A. and Graves,B.J. (2009) DNA specificity determinants associate with distinct transcription factor functions. *Plos Genet*, **5**, e1000778.
17. 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**, R207.
18. Gowrisankar,S. and Jegga,A.G. (2009) Regression based predictor for p53 transactivation. *BMC Bioinformatics*, **10**, 215.
19. Bailey,T.L., Williams,N., Misleh,C. and Li,W.W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, **34**, W369–W373.