Theor Chem Acc (2010) 125:583–591
DOI 10.1007/s00214-009-0643-8

REGULAR ARTICLE

PromoterSweep: a tool for identification of transcription factor binding sites

Coral del Val · Oliver Pelz · Karl-Heinz Glatting · Endre Barta · Agnes Hotz-Wagenblatt

Received: 25 March 2009 / Accepted: 12 September 2009 / Published online: 1 October 2009

Abstract There are many tools available for the prediction of potential promoter regions and the transcription factor binding sites (TFBS) harboured by them. Unfortunately, these tools cannot really avoid the prediction of vast amounts of false positives, the greatest problem in promoter analysis. The combination of different methods and algorithms has shown an improvement in prediction accuracy for similar biological problems such as gene prediction. The web-tool presented here uses this approach to perform an exhaustive integrative analysis, identification and annotation of potential promoter regions. The combination of methods employed includes searches in different experimental promoter databases to identify promoter regions and their orthologs, use of TFBS databases and search tools, and a phylogenetic footprinting strategy, combining multiple alignment of genomic sequences together with motif discovery tools that were tested previously in order to get the best method combination. The pipeline is available for academic users at the HUSAR open server http://genius.embnet.dkfz-heidelberg.de/menu/biounit/open-husar/. It integrates all of this information and identifies among the huge number of TFBS predictions those, which are more likely to be potentially functional.

Keywords Promoter · Transcription factor · Motif discovery · Annotation

Dedicated to Professor Sandor Suhai on the occasion of his 65th birthday and published as part of the Suhai Festschrift Issue.

C. del Val · O. Pelz · K.-H. Glatting · E. Barta · A. Hotz-Wagenblatt (✉)
Molecular Biophysics, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 580, 69120 Heidelberg, Germany
e-mail: hotz-wagenblatt@dkfz.de; hotz-wagenblatt@dkfz-heidelberg.de

C. del Val
Computer Science and Artificial Intelligence, Informatics Faculty, University of Granada, Daniel Saucedo Aranda, 18071 Granada, Spain

E. Barta
Agricultural Biotechnology Center, Szent-Györgyi Albert u. 4, 2100 Gödöllő, Hungary

Present Address:
E. Barta
Apoptosis and Genomics Research Group of the Hungarian Academy of Sciences, Research Center for Molecular Medicine, Medical and Health Science Center, University of Debrecen, Debrecen, Hungary

Abbreviations

TFBS	Transcription factor binding site
TSS	Transcriptional start site
ID	Identifier
TP	True positive
TN	True negative
FP	False positive
FN	False negative
SN	Sensitivity
SP	Specificity
CC	Correlation coefficient
XML	Extensible markup language

1 Introduction

The availability of numerous eukaryotic genome sequences has led to the current challenge of understanding the regulatory networks underlying gene expression. As a

consequence, a rapid increase in the number of available databases, methods and programs dedicated to promoter analysis has emerged during the past few years.

Even with the great amount of data and algorithms available, in genome annotation the identification of the core promoter and the localization of the transcription start site (TSS) remain one of the most challenging problems [1–4]. The core promoter elements of protein-encoding genes are sites of assembly for protein factors required for the transcription initiation, a process also known as TSS selection [5, 6]. In metazoan organisms, the core promoter typically contains more than one sequence motif, such as the TATA box, the initiator, a transcription factor recognition element, and a downstream core promoter element. All these DNA elements together form the transcription preinitiation complex which covers the TSS and some bases of the 5′ untranslated region of the mRNA. This complex facilitates the recruitment of the RNA polymerase II for TSS selection and the interaction with additional regulatory transcription factors bound to enhancers or silencers. The core promoter region, where the transcription-initiation complex assembles [5, 6] is located just around the TSS spanning 50–100 bp at each side. However, the proximal promoter region, which is responsible for the transcription regulation of the gene, and is linked with the regulation according to cell state or tissue, can span several kbp upstream from the TSS [7]. Both, the proximal promoter region and the core promoter are located upstream from the coding part of the gene, which is then transcribed into mRNA and translated into a protein. In this context, the use of orthologous promoter sequences may help in the recognition of conserved and, therefore, potentially functional sequence motifs. This approach is based on the assumption that gene regulatory regions and elements are often preferentially conserved during evolution with the drawback for the approach that selective pressure on orthologous genes must be similar in each respective organism.

Although gene-coding sequences are readily identified by their overall high degree of conservation, the identification of short regulatory elements such as transcription factor binding sites (TFBSs) requires a special approach, especially at the level of the initial alignment of orthologous sequences. These TFBSs are short stretches of DNA of usually 6–20 bp. Vertebrate TFBSs usually degenerate and a number of sequence variations of a TFBS can show binding affinity. This implies that the nature has evolved a system for maintaining a robust response independent of binding specificity, such that the impact of most mutations within the site is relatively small and not lethal. In some cases, differences in binding affinity play a role in subtle regulation of gene expression. Because the TFBSs are small, degenerate sequences, the selection of the correct TFBS predictions in the resulting low signal-to-noise ratio environment is a major problem. Unfortunately, the tools that are currently available cannot really avoid the greatest problem in promoter analysis: the prediction of vast amounts of false positives. In this context, integrative approaches become essential for in-depth promoter analysis. The web-tool PromoterSweep presented here performs an exhaustive integrative analysis, identification and annotation of potential promoter regions to extract potential TSSs and TFBSs with high sensitivity using a combination of different methodologies. The methodologies employed include: (1) searches in different experimental promoter databases with extraction of annotated TSS and TFBS information, (2) the use of TFBS databases and search tools, and (3) motif discovery by a phylogenetic footprinting strategy (see Fig. 1). Phylogenetic footprinting is a technique used to identify TFBSs within a non-coding region of DNA of interest by comparing it with the orthologous sequences in different species using a combination of the multiple alignment of genomic sequences together with motif discovery tools. Phylogenetic footprinting is based on the preferential conservation of functional sequences over the course of evolution by selective pressure. Mutations are more likely to be disruptive if they appear in functional sites, resulting in a measurable difference in evolution rates between functional and non-functional genomic segments.

After having applied the methodologies, the results are integrated by a handcrafted rule-based decision system to classify identified TFBSs. The intention for combining different methods and subsequent integration and classification of the results is the reduction of false positive TFBS predictions. The pipeline is available for academic users at the HUSAR open server http://genius.embnet.dkfz-heidelberg.de/menu/biounit/open-husar/.

2 Materials and methods

2.1 Design of the PromoterSweep pipeline

PromoterSweep is a pipeline that predicts regulatory sequence sections in potential promoter regions through the integration of three different methodologies: promoter database search (methodology 1), de novo motif discovery (methodology 2), and profile matrix search with known TFBSs (methodology 3) (see Fig. 1, columns leading to Results 1, 2, 3).

2.2 Input in the PromoterSweep pipeline

A sequence containing a potential promoter is used as input and one can choose between human and mouse as the organism of origin (see top of Fig. 1, input sequence). The user can additionally provide multiple known co-regulated

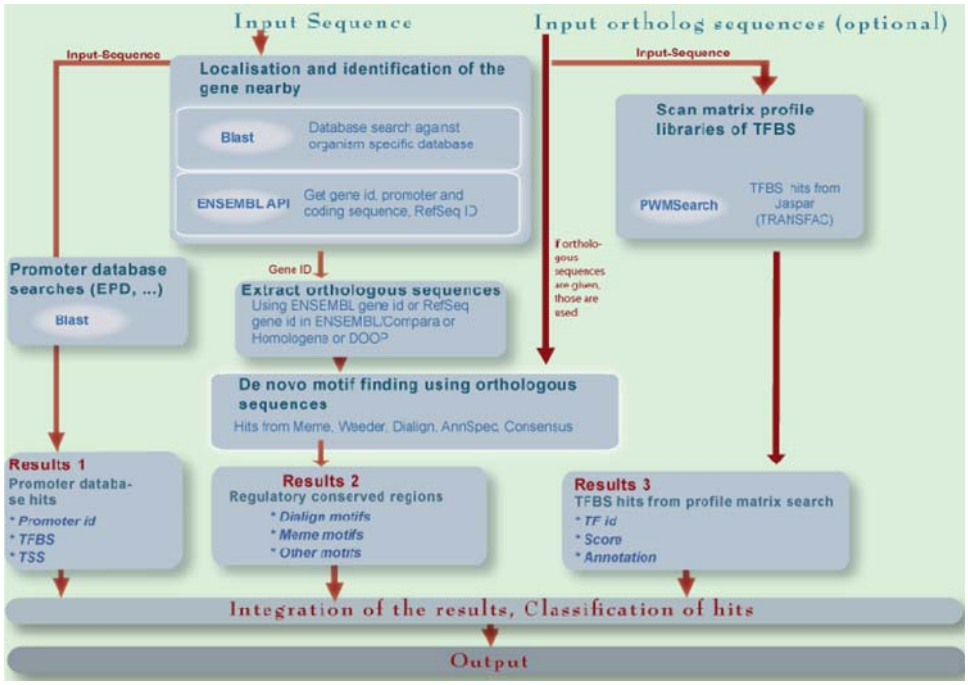

Fig. 1 Data flow of the PromoterSweep pipeline

or homologous promoter regions as a second input (see top of Fig. 1, input ortholog sequences). Both input formats are checked for correctness.

2.3 Promoter databases and homology searches

We chose from the available promoter databases those fulfilling the following criteria:

1. they contain information about experimentally found promoters and TSSs
2. their annotations are independent from other databases and they follow regular updates.

The selected databases are: EPD [8], DBTSS [9], MpromDB[10], DooP [11], and CisRed [12] (see Table 1). EPD, MpromDB and DBTSS contain information about the TSSs, but the definition of the TSS varies between databases. DBTSS defines the TSS as the farthest 5′ position in the genome, which can be aligned with the 5′ end of a cDNA from the corresponding gene. In contrast, MpromDB and EPD consider the most frequent cDNA 5′ end as the TSS, and EPD further applies a specialized algorithm to infer multiple promoters for a given gene [8].

These databases were locally installed in order to run homology searches. These searches are performed using the BLAST algorithm [13], Gapped BLAST [13] and PSI-BLAST [13]. The user-supplied potential promoter sequence which in many cases spans a stretch of 2,000 bases in front of an annotated gene is run against the promoter databases. The resulting hits are only accounted for and used for annotation transfer of the TFBSs if the alignment is at least 200-bp long and the homology is at least 99%. These default values of the parameters used to set up the requirements for the inclusion of a hit can be changed by the user.

2.4 Identification of orthologous sequences

A set of orthologous promoter regions is needed for the de novo motif discovery (see Fig. 1, middle columns leading to Result 2, methodology 2). The user has the option to provide an additional input file with orthologous or co-expressed promoter sequences. If this file is not provided, the pipeline follows up with an automatic search for orthologous sequences. In order to automatically find orthologous promoter regions the potential promoter sequence is mapped to the appropriate genome, mouse or human. The distances of the input sequence to genes on both DNA strands are taken into account for identification of the gene linked to the promoter, and the gene closest to the promoter hit is selected if certain conditions are fulfilled, such as the distance to the gene being smaller than 20 bp or the overlapping coding sequence smaller than 200 bp according to [27]. All these parameters can be changed by the user. If the gene is found on the opposite strand, then the input sequence is reversed and complemented for further use [11]. With the selected gene id, orthologous promoter regions are extracted following one of the three different databases (Ensembl, Homologene, or DooP):

1. ENSEMBL: a BLAST search against the ENSEMBL 49 Core database [14] (http://www.ensembl.org) is

Table 1 Characteristics of the different promoter and promoter based databases used in PromoterSweep

Database	Nr of entries	Data Type	Promoter region	Additional information
EPD	4,806	TSS, Promoter	(−400 up to +100)	Experimental, in silico
DBTSS	1,77,996	TSS	(−1000 up to +2009	Experimental
MPromDB	26,351	Promoter, TSS, TFBS (mammal)	(−3000 up to +1000)	Experimental
cisRED	18,779	Promoter motifs (human)	(−1500 up to +200)	In silico
Doop	22,415/4,75,266	Promoter/conserved motifs (chordate)	(length 500; 2,000; 3,000)	Experimental, in silico

performed to extract the ENSEMBL gene id to which the promoter belongs. This gene id is used to query all orthologous sequences of this gene stored in the ENSEMBL Compara database [15], and to extract their corresponding promoter regions.

2. HOMOLOGENE: the ENSEMBL gene id, obtained like in case 1 is translated into its corresponding ENTREZ gene id (http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome). This ENTREZ gene id is used to extract the orthologs from the HOMOLOGENE database (http://www.ncbi.nlm.nih.gov/homologene/).
3. DooP: a BLAST search against the DooP database [11] is performed with the potential promoter sequence. DooP is a database of eukaryotic promoter sequences, aiming to facilitate the recognition of regulatory sites conserved between species. It contains orthologous promoter sequences from *Viridiplantae* and *Chordata* species based on the *Arabidopsis thaliana* and *Homo sapiens* genome annotation. Then the orthologous promoter sequences, annotated in the obtained DooP hit, are extracted.

2.5 Motif discovery tools

We provide different methods based on different operating principles to get shared motifs of orthologous or co-regulated sequences: Meme [18], Gibbs Motifs Sampler [19], Weeder and WeederH [20], AnnSpec [21] and Consensus [22]. A short description of the different tools can be found in Table 2.

Apart from these motif discovery tools, we use the alignment program Dialign2 [23] to find common motifs and to extract them from the alignment as conserved regions.

Dialign2 and Meme are the tools selected for motif discovery by default. Dialign2 was the best one according to the positive predictive value (PPV, see below) in the evaluation results in Table 3. But the user can select any combination of six implementations of different algorithms.

2.6 Matrix profile search

The identification of known TFBSs is carried out with PWMsearch, a program out of the collection "TFBS Perl" [16] (http://tfbs.genereg.net/), and the Jaspar Database of TFBS profiles [17] (http://jaspar.genereg.net/). This database is locally installed as a mySQL database. The JASPAR CORE database contains a curated, non-redundant set of 138 profiles from published articles. All profiles are derived from published collections of experimentally located transcription factor binding sites for multi-cellular eukaryotes. The database represents a curated collection of target sequences. The binding sites were determined either in SELEX experiments, or by the collection of data from the experimentally determined binding regions of actual regulatory regions. This distinction is clearly marked in the profiles' annotation. As far as possible, the collection is non-redundant. The prime difference to similar resources like TRANSFAC (Biobase, http://biobase-international.com) consists of the open data access, non-redundancy and quality: JASPAR is a smaller, curated non-redundant dataset.

Due to licensing issues our second tool for the matrix profile search, the commercial "TRANSFAC Professional" database together with the Match program (both Biobase, http://biobase-international.com), can only be used by local users and not via the HUSAR open server. TRANSFAC® is a database on eukaryotic *cis*-acting regulatory DNA elements and *trans*-acting factors. It covers the whole range of species from yeast to human. The data have been generally extracted from the original literature, but occasionally they have been taken from other compilations. The Match™ tool is designed for searching potential TFBSs in any DNA sequence which may be of interest, and uses a library of mononucleotide weight matrices from TRANSFAC®.

2.7 Evaluation of the prediction

The performance of the selected TFBS prediction programs is assessed using a human promoter test-set. Most promoter test-sets were specifically selected for the particular programs analyzed in the publications [24, 25], however, we chose a robust test-set of promoters found in different organisms from existing literature [26] (http://bio.cs.washington.edu/assessment/). This dataset was created for the evaluation of motif discovery tools for finding TFBSs in promoter regions. The human subset, our test-set,

Table 2 Motif discovery methods used by PromoterSweep

Name	Web source	Description
Meme [18]	http://meme.sdsc.edu	Meme optimises the *E*-value of a statistic related to the information content of the motif and uses the product of *P*-values of column information contents. The motif search consists of performing expectation maximization from starting points derived from each subsequence occurring in the input sequences
Gibbs Motifs Sampler [19]	http://bayesweb.wadsworth.org/gibbs	Gibbs uses a probabilistic framework which not only estimates the expected number of motif instances in the sequence but also extracts the motifs
Weeder and WeederH [20]	http://159.149.109.9/modtools/	Weeder applies a consensus based method that enumerates exhaustively all the oligos up to a maximum length and collects their occurrences, with substitutions, from input sequences
AnnSpec [21]	http://www.cbs.dtu.dk/services/DNAarray/ann-spec.php	AnnSpec applies a neural network that models the DNA-binding specificity of a transcription factor using a weight matrix
Consensus [22]	http://bifrost.wustl.edu/consensus	Consensus models motifs using weight matrices, searching for the matrix with the maximum information content
Dialign2 [23]	http://bibiserv.techfak.uni-bielefeld.de/dialign/	Dialign2 constructs multiple alignments by comparing entire *segments* of the sequences. Extracting the common segments the conserved motifs can be found

Table 3 Results on the selected human dataset using individual different motif-prediction tools and PromoterSweep

Field	Individual methods							Promoter-sweep results				
	PwmSearch	AnnSpec	Consensus	GibbsMotif Sampler	Weeder	Meme	Dialign	All categories but weak	Most reliable	Reliable	High	Conserved
TP	6,948	375	139	148	827	119	565	**1,162**	256	217	8	681
FP	2,95,324	6,633	7,119	7,481	14,304	2,932	9,468	13,822	773	3,196	2	9,851
FN	3,473	4,496	4,732	4,723	4,044	4,753	4,307	3,957	4,863	4,902	5,111	4,438
TN	22,947	2,66,494	2,66,008	2,65,646	2,58,824	2,70,195	2,63,660	2,68,059	2,81,108	2,78,685	2,81,879	2,72,030
Sn	0.067	0.077	0.029	0.030	0.169	0.024	0.115	**0.227**	0.050	0.042	0.002	0.133
Sp	0.072	0.976	0.974	0.972	0.945	0.989	0.965	0.951	0.997	0.989	1.000	0.965
CC	−0.165	0.044	0.002	0.002	0.068	0.017	0.057	**0.106**	0.105	0.038	0.035	0.069
PPV	0.023	0.053	0.019	0.019	0.054	0.039	0.056	**0.078**	0.2488	0.063	0.800	0.065

All numbers are calculated at the nucleotide level

Bold values indicate where PromoterSweep (all categories but weak) revealed the best value compared with the other tools

contains 217 experimentally verified human promoter sequences in 26 orthologous groups. The 217 selected sequences together with their orthologous alignments were processed by the individual motif discovery tools and by PromoterSweep using Meme and Dialign. The specificity and sensitivity are calculated according to the number of common nucleotides between annotated and predicted motifs. True positive nucleotides (TP) is the number of nucleotides from an annotated motif that are predicted, false positive nucleotides (FP) is the number of nucleotides that are predicted but do not belong to the known motif, true negative nucleotides (TN) is the number of nucleotides that are not predicted and do not belong to a known motif, and false negative nucleotides (FN) is the number of nucleotides that are not predicted and do belong to the motif. Sensitivity (SN), specificity (SP) and their correlation coefficient (CC) and positive predictive value (PPV) are calculated as:

$$\mathrm{SN} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$$

$$\mathrm{SP} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP})$$

$$\mathrm{PPV} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$$

$$cc = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}$$

We compare the TFBS motifs identified by PromoterSweep in terms of SN, SP, PPV and CC with the motifs obtained with the individual programs: PWMsearch, Dialign, Weeder, Meme, Annspec, Gibbs sampler and Consensus.

3 Results

3.1 Implementation of the three methodologies for identifying TFBSs used by PromoterSweep

At the beginning, the pipeline performs homology searches with the input sequence against the promoter databases EPD, DBTSS, MpromDB, DooP and CisRed. This represents the first methodology of promoter database searches leading to Results 1 (see Fig. 1). If a hit with a known promoter is found to fulfil the requirements of having more than 99% identity over more than 200 bases, the annotation and the location of the TSS are transferred to the input sequence (leading to Results 1 in Fig. 1). In the case of hits in the DooP database, the conserved motifs constituting potential TFBSs could also be recovered.

A set of orthologous promoter regions is needed for de novo motif discovery (see Fig. 1, middle columns leading to Result 2, methodology 2). The orthologous promoter regions, that are either automatically found or provided by the user, are then used for de novo TFBS motif discovery. The user can select the programs which should be used. In our example, Meme, AnnSpec, GibbsSampler and the Dialign2 alignment tool were applied to discover common motifs of length 6–20 bp (see Fig. 1, leading to Results 2).

While the de novo motif detection is carried out, the input sequence is concurrently used for the identification of known TFBSs using profile-matrix searches in the Jaspar database or in TRANSFAC. The profile-matrix search is the third methodology and yields Results 3 (see Fig. 1).

3.2 The integration of the results of the different methodologies

Due to the fact that the promoter annotations obtained through the three different strategies are independently created, a subsequent step for the integration of the results is needed. This integration results in a classification of the hits that provides the user with information about how much an identified TFBS is supported by the three different methods. The annotations found by hits to the promoter databases (methodology 1), the de novo discovered motifs (methodology 2), and the TFBSs identified by the profile-matrix search (methodology 3) are combined and corresponding hits in the three result sets are merged and subsequently classified by a handcrafted rule-based decision system. The rules are based on the test results of de novo predictions described in Tompa et al. [26] and on the experiences of the authors. Combined TFBSs are assigned to five classes representing the quality of a hit defined by the following rules:

- "Most reliable" is the highest quality class. An extended hit in an experimentally annotated promoter of the promoter database which contains annotated TFBSs is needed (strongly supported by methodology 1). Those TFBSs are displayed. In this class, the results of the other methodologies are not accounted for.
- "Reliable" is the next quality class. A motif in this class needs to be a hit in an experimentally annotated promoter of the promoter database (supported by methodology 1), a hit in the profile-matrix search (supported by methodology 3), a Dialign hit and a hit with another motif discovery tool (supported by methodology 2).
- "High" is used to classify a motif with a hit in a non-experimental part of the promoter databases (weakly supported by methodology 1), a hit in profile-matrix search (supported by methodology 3), and a Dialign hit and a hit with another motif discovery tool (supported by methodology 2).
- "Conserved" refers to a motif with a hit in a non-experimental part of the promoter databases (weakly supported by methodology 1), no hit in profile-matrix search (not supported by methodology 3), and a Dialign hit and a hit with another motif discovery tool (supported by methodology 2).
- "Weak" describes a motif if the hit is in front of a gene and both a Dialign motif (supported by methodology 2) and a hit in profile-matrix search (supported by methodology 3) are found.

If a motif qualifies for more than one class, the highest class is taken and shown in the output.

3.3 The web output of PromoterSweep

The web output of PromoterSweep, is divided in seven sections (see Table 4 and Fig. 2):

1. General Information: displays the input sequence name and length as well as the parameter settings selected by the user.
2. Best Genomic Mapping: shows the genomic localization of the transcript identified to be linked with the promoter input sequence and the distances between them.

Table 4 PromoterSweep web output description; some sections may be missing depending on the results found

Output sections	Fields
General information	Input sequence, sequence length, species, homology database, type of sequences for motif search, type of profile search tool
Best genomic mapping	Gene ID, gene symbol, description, chromosome, distance to promoter and coding region, chromosome, start and end of promoter and gene on the chromosome
Graphical overview	TFBS classification: weak < conserved < high < reliable < most reliable.
	Promoter, TSS, combined motifs, links
Combined binding sites	Id, classification, start, end, motif discovery and TFBS search link
TSS and exon information	Type, start, database
Profile matrices overlapping combined results	Combined motif, transcription factor name/link, Database name, Start, End, Score
Generated output files	Links to all output files generated (alignments, search results, etc.)

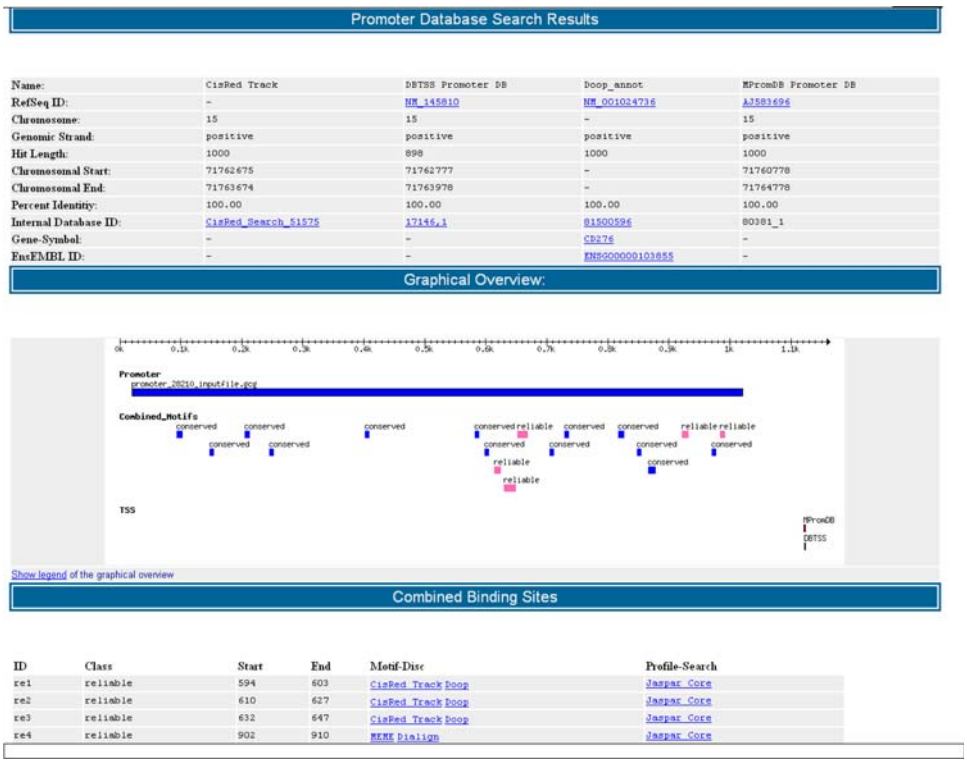

Fig. 2 Part of the output of the PromoterSweep pipeline—promoter database search results and graphical overview

3. Promoter Database Search Results: shows the database search results with links to the product and promoter id entries found.
4. Graphical Overview: graph showing the distribution of the found TFBSs in the input sequence as well as their classification. At the bottom of the Graphical Overview section there is a link to an explanatory legend.
5. Tables for combined TFBSs: this table contains the TSSs.
6. Tables for combined TFBSs: this table contains the profile-matrices found.
7. List of output files: links to all output files are listed.

In the web output, the user has immediate access to all complete application outputs and database entries via hyperlinks, at the bottom of the web output page there are also links to each of the text output files containing all the information generated in the process.

3.4 Technical implementation of PromoterSweep

PromoterSweep is implemented using the W3H task framework [28], which allows the execution of compound

jobs using work and data flow descriptions in a heterogeneous bioinformatics environment. The system allows the design of high complexity bioinformatics tasks, and stores the results of the individual applications together with the derived results obtained by integrating them. The final output of the task is an XML file which contains all relevant information generated. The XML information is transformed by means of the post-processing mechanism of the web interface W2H [29] into an HTML page (see Fig. 2). Furthermore, the XML output can also be requested and used for further analysis, e.g. direct integration in user's databases or additional pipeline analysis.

3.5 Evaluation of the pipeline PromoterSweep

For this purpose we used the Tomba et al. [26] dataset, a robust promoter test-set originally created for the evaluation of motif discovery tools applied to promoter regions. We selected the human subset for the evaluation, which contains 217 experimentally verified human promoter sequences in 26 orthologous groups. The 217 selected sequences together with their orthologous alignments were processed by the individual motif discovery tools and by PromoterSweep using Meme and Dialign. The obtained results are summarized in Table 3. The first PromoterSweep result column corresponds to the sum of all prediction classes but the weak class (see Table 3). Adjacent columns show the motif discovery performance of the PromoterSweep TFBSs classes (see Table 3). When using all classes of combined TFBSs except the weak class, PromoterSweep identified the highest number of true positive nucleotides, got the highest sensitivity, the best positive predictive value and the best correlation coefficient. The specificity was slightly reduced but still higher than 0.95. Concerning PromoterSweep TFBS classifications, it can be seen, that the class "most reliable" has the highest CC value, but a sensitivity of only 0.05. Only 25% of the TP nucleotides are found, but also only 6% of the false positive nucleotides compared to the values of the combined classes. In the class "conserved" we get about 50% of the TP nucleotides, but also about 75% of the FP nucleotides. The class "high" does not play a role due to its very low incidence, and the class "reliable" presents intermediate sensitivity and specificity values between those of "most reliable" and "conserved".

4 Discussion

PromoterSweep is a pipeline which allows the annotation of DNA sequences for TFBSs. The general problem of identifying TFBSs is the huge amount of false positives, which is a result of the weak definition of those motifs. PromoterSweep is the only tool in the field, which integrates promoter database searches with profile-matrix searches of known TFBSs and de novo motifs of orthologous or co-regulated sequences. An implemented rule-based decision system allows for the integration and classification of the hits.

The different length ranges (500–4,000 bp) of the promoter databases' sequences as well as their different TSS definitions constitute one of the main problems in comparing, integrating and transferring annotations from various promoter database hits to the input sequence. Additionally, the coverage of the different databases, meaning the number of entries for an organism, varies extremely. EPD has almost full coverage with 4,800 promoters from 213 different organisms, while DBTSS contains 1,77,800 TSS entries only from human and mouse, and DooP contains 4,75,000 conserved motifs from 22,415 promoters of the phylum chordata. A further complication is the mixture of experimentally proven promoters and those that are only predicted in some databases, a fact that produces differences in the quality of the annotation. For these reasons, not only the annotation of the hit was taken into account, but also its reliability according to the database and its annotation quality.

The identification of TFBSs by profile-matrices is accomplished in the open server using the Jaspar database, which is open source, though it has a lower coverage compared to TRANSFAC. Nevertheless, the differences between the profile-matrix libraries Jaspar and TRANSFAC are not as big as anticipated, with more than 65% of the TFBSs present in both libraries [30], rendering Jaspar a good alternative to the use of TRANSFAC.

The implemented classification system evaluates the information's quality depending on its sources, giving the highest scores to experimental data originating from promoter databases. Besides the "traditional" motif discovery algorithms we use an approach already implemented in the DooP database [11], realized as a multiple alignment step with Dialign followed by a motif extraction step. The extra treatment of the Dialign motifs in the classification system is based on the experience with DooP that this approach is a more reliable solution for finding collinear conserved motifs in orthologous promoter regions than using de novo motif-prediction tools. Concerning the de novo motif-prediction tools, we got CC values smaller than 0.1, values that are comparable with those presented in [26] and which are not convincing. Therefore, the predicted motifs are used only as supporting information in the classification system.

The pipeline is designed with the goal to annotate potential promoter sequences with a higher sensitivity than that shown by individual existing TFBS prediction methods. The combination of motif discovery tools with orthologous sequences has been previously reported to be

effective with a 9% raise of sensitivity and specificity in TFBS prediction [31]. However, it is the first time that searches against different promoter databases have been integrated in the process using the TSS and TFBS hits' annotation. The results obtained with PromoterSweep are promising and show an improvement in sensitivity and accuracy in terms of CC and PPV when compared to individual methods. However, the number of false positives is still very high (see Table 3). Future work will include the further development and improvement of the rule-based classification system using machine learning classification approaches.

References

1. Bajic VB, Brent MR, Brown RH, Frankish A, Harrow J, Ohler U, Solovyev VV, Tan SL (2006) Genome Biol 7(Suppl 1):S3.1–S313
2. Sonnenburg S, Zien A, Rätsch G (2006) Bioinformatics 22:e472–e480
3. Wang X, Bandyopadhyay S, Xuan Z, Zhao X, Zhang MQ, Zhang X (2007) Comput Syst Bioinformatics Conf 6:183–193
4. Xie X, Wu S, Lam KM, Yan H (2006) Bioinformatics 22:2722–2728
5. Pedersen AG, Baldi P, Chauvin Y, Brunak S (1999) Comput Chem 23:191–207
6. Smale ST, Kadonaga JT (2003) Annu Rev Biochem 72:449–479
7. Choi CH, Kalosakas G, Rasmussen KO, Hiromura M, Bishop AR, Usheva A (2004) Nucleic Acids Res 32:1584–1590
8. Schmid CD, Perier R, Praz V, Bucher P (2006) Nucleic Acids Res 34(database issue):D82–D85
9. Yamashita R, Suzuki Y, Wakaguri H, Tsuritani K, Nakai K, Sugano S (2006) Nucleic Acids Res 34(database issue):D86–D89
10. Sun H, Palaniswamy SK, Pohar TT, Jin VX, Huang TH, Davuluri RV (2006) Nucleic Acids Res 34(database issue):D98–D103
11. Barta E, Sebestyén E, Pálfy TB, Tóth G, Ortutay CP, Patthy L (2005) Nucleic Acids Res 33(database issue):D86–D90
12. Robertson G, Bilenky M, Lin K, He A, Yuen W, Dagpinar M, Varhol R, Teague K, Griffith OL, Zhang X, Pan Y, Hassel M, Sleumer MC, Pan W, Pleasance ED, Chuang M, Hao H, Li YY, Robertson N, Fjell C, Li B, Montgomery SB, Astakhova T, Zhou J, Sander J, Siddiqui AS, Jones SJ (2006) Nucleic Acids Res 34(database issue):D68–D73
13. Altschul SF, Madden TL, Schaeffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Nucleic Acids Res 25:3389–3402
14. Hubbard TJP, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Gräf S, Haider S, Hammond M, Holland R, Howe K, Jenkinson A, Johnson N, Kähäri A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Rios D, Schuster M, Slater G, Smedley D, Spooner W, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wilder S, Zadissa A, Birney E, Cunningham F, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Kasprzyk A, Proctor G, Smith J, Searle S, Flicek P (2009) Nucl Acids Res 37:D690–D697
15. Vilella AJ, Severin J, Ureta-Vidal A, Durbin R, Heng L, Birney E (2009) Genome Res 19:327–335
16. Lenhard B, Wasserman WW (2002) Bioinformatics 18:1135–1136
17. Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B (2004) Nucleic Acids Res 32(database issue):D91–D94
18. Bailey TL, Elkan C (1995) Proc Int Conf Intell Syst Mol Biol 3:21–29
19. Thompson W, Palumbo MJ, Wasserman WW, Liu JS, Lawrence CE (2004) Genome Res 14:1967–1974
20. Pavesi G, Zambelli F, Pesole G (2007) BMC Bioinformatics 8:46
21. Workman CT, Stormo GD (2000) Pac Symp Biocomput 2000:464–478
22. Stormo GD, Hartzell GW (1989) Proc Natl Acad Sci USA 86:1183–1187
23. Morgenstern B (1999) Bioinformatics 15:211–218
24. Sinha S, Tompa M (2003) Nucleic Acids Res 31:3586–3588
25. Favorov AV, Gelfand MS, Gerasimova AV, Ravcheev DA, Mironov AA, Makeev VJ (2005) Bioinformatics 21:2240–2245
26. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Régnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z (2005) Nat Biotechnol 23:137–144
27. Endre B (2007) Methods Mol Biol 395:319–328
28. Ernst P, Glatting KH, Suhai S (2003) Bioinformatics 19:278–282
29. Senger M, Flores T, Glatting KH, Ernst P, Hotz-Wagenblatt A, Suhai S (1998) Bioinformatics 14:452–457
30. Kielbasa SM, Gonze D, Herzel HP (2005) BMC Bioinformatics 6:237
31. Li X, Zhong S, Wong WH (2005) PNAS 102:16945–16950