ABSTRACT
Summary: TFBS is a set of integrated, object-oriented Perl modules for transcription factor binding site detection and analysis. It implements objects representing specificity profile matrices, binding sites and sets thereof, pattern generators, and pattern database interfaces. The modules are interoperable with the BioPerl open source system.
Availability and Supplementary Information: The module package with documentation and example scripts are available at http://forkhead.cgb.ki.se/TFBS/.
Contact: Boris.Lenhard@cgb.ki.se

INTRODUCTION
Eukaryotic regulatory regions are characterized by the presence of multiple transcription factor binding sites, which can be described as sequence patterns with varying degrees of degeneracy. For computational analysis of regulatory regions, most approaches can be described by three sequential phases. First, a pattern is described for the set of target sequences known to be bound by a specific transcription factor. Second, a set of DNA sequences are analysed to determine the location(s) of sequences consistent with the described binding pattern (Staden, 1990). Finally, in advanced cases, predictive statistical models of regulatory regions are constructed based on multiple occurrences of the detected patterns (Wasserman and Fickett, 1998).

Frequency matrices have proven to be the most successful quantitative method for representing the binding specificity of a given transcription factor (Berg and von Hippel, 1988). From this foundational unit, a number of quantitative variations have been successfully applied, including weight matrices for easier statistical analysis and information content matrices well-suited for graphical representations (Schneider and Stephens, 1990) of the binding pattern. An annotated scheme for recording specificity of a sequence in a matrix and its transformation to other types of matrices is depicted in a figure at http://forkhead.cgb.ki.se/TFBS/matrixfigure.html.

OVERVIEW AND EXAMPLES
The classes constituting TFBS encompass the storage, manipulation and analysis objects for nucleotide sequence patterns and the corresponding nucleotide sequence features. Currently available classes are listed in Table 1. Among these are pattern classes, pattern set classes and pattern generators. Pattern classes hold the actual matrix profiles with the associated information. Pattern set classes are aggregate classes for storage and manipulation of multiple matrices (e.g. for scanning nucleotide sequences with multiple matrices, when one wants to analyse the resulting set as a whole). Pattern generators are factory classes for constructing new matrices from a set of

*To whom correspondence should be addressed.
sequences using different algorithms, some of them implemented by existing external programs. For pattern classes, the basic functionality is shown in annotations to the figure at http://forkhead.cgb.ki.se/TFBS/matrixfigure.html.

The following two code snippets demonstrate the ease of use of TFBS objects:

- a script that retrieves a sequence from GenBank using BioPerl, a C/EBP position weight profile from TRANSFAC, scans the sequence with the matrix and outputs the detected sites in GFF format:

```perl
#!/usr/bin/env perl -w
use Bio::DB::GenBank;
use TFBS::DB::TRANSFAC;
my $seq = Bio::DB::GenBank->new(-file=>'sequences.fa', -threshold=>'80%');
my $db = TFBS::DB::TRANSFAC->connect();
my $pwm = $db->get_Matrix_by_ID('V$CEBPA_01','PWM');
my $siteset = $pwm->search_seq(-seqobj=>$seq, -threshold=>">80%" hunger=>1);
print $siteset->GFF();
```

- a script that identifies new patterns from a set of DNA sequences stored in the file ‘sequences.fa’ and stores them in a simple flat-file database:

```perl
#!/usr/bin/env perl -w
use TFBS::DB::FlatFileDir;
use TFBS::PatternGen::Gibbs;
my $gibbs = TFBS::PatternGen::Gibbs->new(-file=>'sequences.fa', motif_length=>10);
my $db = TFBS::DB::FlatFileDir->create('NewPatterns');
$db->store_Matrix($gibbs->all_patterns());
```

The annotated versions of the above and more complex example scripts are available at http://forkhead.cgb.ki.se/TFBS/.

### FUTURE DEVELOPMENT

The TFBS functionality can be extended and expanded in a multitude of directions. Two additions currently under construction are:

- TFBS::DB::SimpleXML: database interface for a flexible single-file exchange format for matrix patterns
- TFBS::Analysis::LRA:—a set of classes for model building and statistical evaluation using logistic regression

Although originally designed for transcription factor patterns and binding sites, the overall design readily lends itself to the utilization of diverse patterns classes.

### REFERENCES


