# Weeder Web: a Web-Based Tool for the Discovery of Transcription Factor Binding Sites

Giulio Pavesi [(1)], Giancarlo Mauri [(2)], and Graziano Pesole[(3)]

[(1)] D.I.Co., University of Milan, Via Comelico 39, 20135 Milan - Italy
pavesi@dico.unimi.it
[(2)] Dept. of Computer Science, University of Milano-Bicocca, Via Bicocca A. 8, 20126 Milan – Italy
mauri@disco.unimib.it
[(3)] Dept. of Biomolecular Science and Biotechnology, Un. of Milan, Via Celoria 26, 20133 Milan - Italy
graziano.pesole@unimi.it

**Keywords.** Motif discovery, transcription factor binding sites, transcription initiation, gene regulation

## Introduction

Understanding the complex mechanisms governing basic biological processes requires the characterization of regulatory motifs modulating gene expression at transcriptional and post-transcriptional level. In particular, the extent, chronology and cell-specificity of transcription are modulated by the interaction of transcription factors (TFs) with their corresponding binding sites (TFBS), located in the promoter regions of the genes. The ever growing amount of genomic data, complemented by other sources of information concerning gene expression opens new opportunities to researchers.

Transcription factor binding sites are generally short (less than 12-14 bp long) and degenerate oligonucleotides, and this fact makes significantly harder their computational discovery and large-scale annotation. Hence, the need for efficient and reliable methods for detecting novel *motifs*, significantly over-represented in the regulatory regions of sets of genes sharing common properties (e.g. similar expression profile, biological function, product cellular localization, etc.), that in turn could represent binding sites for the some common TF regulating the genes.

We present here a Web server that provides access to a previously developed enumerative pattern discovery method [1] that is able to carry out an exhaustive search of significantly conserved degenerate oligonucleotide patterns with remarkable computational efficiency. Also, the interface has been designed in order avoid the explicit definition of a large number of parameters that were included in the original general-case implementation of the algorithm, as well as to produce a simpler "user-friendly" output. The parameters have been set to default values suitable for capturing TFBSs. The interface Web address is:

`http://www.pesolelab.it:8080/weederWeb`

## The Interface

The Weeder Web interface requires users to input their e-mail address, one or more sequences in FASTA format either by cutting and pasting the sequences or by uploading a file, and to provide two more quite intuitive parameters. The first one specifies only whether the motif has to appear in all the input sequences, or in some of them. The user can also select a mode in which the input is processed as a single sequence, looking for repeated oligonucleotides regardless of their distribution throughout the sequences (suitable, for example, for genome-wide analyses). The second parameter needed, describing the type of analysis desired (quick, normal or thorough), just influences the time required to obtain the results: clearly, the shorter is the time, the less accurate the analysis performed is. Finally, users must specify which organism their sequences were taken from, by selecting it from a list. This parameter is fundamental for the computation of the significance of the results, since different choices can yield completely different results on the same set of sequences. If the organism is not included in the list provided, users can contact the page administrators. The list of available organisms will be constantly updated and enlarged. Since the computation time in some cases might exceed an hour, the results of the analysis are sent by e-mail, also including a link to a Web page where the same results are presented with a nicer graphic layout.

### Program Runs

Once the "Submit" button is clicked, if all the fields are filled in correctly the Web interface automatically starts a series of runs of the Weeder algorithm, looking for motifs of length 6 and 8 (if launched in quick mode), or from length 6 to 12 (in normal mode and thorough mode). The number of sequences a motif has to appear in (the *quorum*) is determined according to the user's choice (all or some). The number of mutations allowed in the occurrences of a motif is one for motifs of length 6, two for length 8, three for length 10, and four for length 12. The "thorough" mode performs an additional scan for motifs of length 8 with three mutations and length 10 with four mutations, and considers also the reverse complement of the sequences. Experiments performed by us and by other groups (see for example [2, 3]) have shown that these values are suitable for capturing a large class of TFBSs, even in case of corrupted datasets including several sequences not containing instances of the motif.

### The Output

For each run, all the motifs satisfying the input constraints (length, error, and quorum) are scored according to a statistical measure of significance that considers simultaneously the number of sequences a motif appears in as well as the overall number of its occurrences in the input set. The measure of significance is based on the genome-wide oligo-frequency analysis of promoter regions of different species. Then, the five most significant motifs of each run are reported to the user. The results are also post processed by the program in order to select the most "interesting" motifs. This selection is based on some heuristics derived from the analysis of real TFBS. The interesting motifs are again listed at the bottom of the output file, under the heading "My Advice". For each of these, the interface also reports the frequency matrix (built by aligning all the instances of the motif found) as well as a list of its best occurrences, collected from the input sequences by using the frequency matrix.

The current implementation of the Web interface has been designed for the discovery of transcription factor binding sites. However, many extensions and improvements are planned. First, it will be extended to the detection of functional motifs in other types of sequences, like mRNA untranslated regions, by defining suitable values for the statistical analysis. Then, we are planning to include directly in the input gene expression values, like those derived from microarray experiments, so to spare the user the need of defining and selecting a priori a set of (supposedly) related genes.

## References

[1] G. Pavesi, G. Mauri, and G. Pesole, An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics*, 17 Suppl 1: S207-14, 2001.

[2] C. Narasimhan, P. LoCascio, and E. Uberbacher, Background rareness-based iterative multiple sequence alignment algorithm for regulatory element detection. *Bioinformatics*, 19(15): 1952-63, 2003.

[3] S. Sinha and M. Tompa, Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res*, 30(24): 5549-60, 2002.