TargetFinder: searching annotated sequence databases for target genes of transcription factors

Giovanni Lavorgna\textsuperscript{2}, Alessandro Guffanti\textsuperscript{3}, Giuseppe Borsani\textsuperscript{3}, Andrea Ballabio\textsuperscript{3} and Edoardo Boncinelli\textsuperscript{1,2}

\textsuperscript{1}DIBIT, HSR Via Olgettina 60, 20132 Milano, \textsuperscript{2}Centro per lo studio della Farmacologia Cellulare e Molecolare, CNR, Via Vanvitelli 32, 20129 Milano and \textsuperscript{3}TIGEM, San Raffaele Biomedical Science Park, Via Olgettina 58, 20132 Milano, Italy

Received on June 3, 1998; revised on August 18, 1998; accepted on August 25, 1998

Abstract

Summary: TargetFinder is a new software tool to search a database of annotated sequences for transcription factor binding sites located in context with other important transcription regulatory signals and regions, like the TATA element, the promoter, and so on, thereby greatly reducing the background usually associated with this kind of search.

Availability: The TargetFinder Web service is available at http://hercules.tigem.it/TargetFinder.html.

Contact: giovanni.lavorgna@hsr.it

The exponential accumulation of biological information in sequence databases has created the need for periodic screening with query sequences of known or unknown function, looking for statistically significant (Karlin and Altschul, 1990) new hits. When looking for short sequences, such as transcription factor (TF) binding sites, the situation is complicated since such sites have a rather high chance of randomly occurring not only within gene regulatory regions, but also almost everywhere else in the sequence, giving rise to a rather high noise level during the search. This problem can be reduced by considering some factors inherent to the search, such as:

1. It can be assumed that the majority of the regulatory elements are located somewhere near the eukaryotic and prokaryotic genes regulated by them.

2. It is often not necessary to search the entire database of DNA sequences. For example, when looking for putative target genes of a vertebrate transcription factor, it makes little sense to analyze sequences belonging to bacteria, phages or plants.

3. A new generation of computer software (reviewed in Frech et al., 1997) allows the detection of protein-binding motifs in DNA sequences, no longer exclusively relying on IUPAC string searches, but using the more sophisticated weight matrix searches.

Thus we have developed a TargetFinder, to facilitate the search for TF target genes. The task performed by TargetFinder can be divided into two parts, sequence retrieval and sequence analysis, which are accomplished by two separate programs.

The TargetFinder program fetches all GenBank sequences in which the localization of at least one of the following 13 features has been annotated in the corresponding database entry and records the position of the feature. This step is pre-calculated, in order to save CPU cycles at run-time. The rationale for the choice of keywords is to use frequently annotated features in GenBank, to locate important transcription regulatory signals and regions, like enhancers, promoters, transcription start sites, etc., that can help restrict the search for binding sites to specific gene regions. Features used for sequence retrieval (and their number of occurrences in Release N. 107 of GenBank, June 1998) are: ‘TATA’ (6064), ‘promoter’ (5082), ‘CAAT’ (1968), ‘enhancer’ (525), ‘exon’ (37 894), ‘intron’ (26 341), ‘mRNA’ (20 650), ‘prim_transcript’ (4598), ‘precursor_RNA’ (1113), ‘5UTR’ (8351), ‘3UTR’ (7388), ‘gene’ (69 786) and ‘CDS’ (128 779). As an option, TargetFinder can also search sequences belonging to the specialized Eucaryotic Promoter Database (EPD) (Cavin Perier et al., 1998).

TargetFinder runs the MatInspector program (Quandt et al., 1995) on the sequence sets fetched in the previous step that have been selected by the user from the HTML form. The algorithm built into MatInspector (Version 2.1) allows not only the searching of DNA sequences for simple IUPAC-coded strings, but also for whole nucleotide distribution matrices which are known to be a more precise representation of consensus patterns (Frech et al., 1997). From the MatInspector result file, TargetFinder determines which sequences (if any) have these binding sites inside and/or within a certain distance from the selected feature(s). All the sequences matching these criteria will be candidate target genes of the transcription factor being analyzed and will be reported to the user by e-mail, together with the detected match. The user can also perform a search for
clustered binding sites, which is an effective way to impose less restriction on the binding sites to be found, without sacrificing search selectivity (Wagner, 1997).

TargetFinder was used to screen the vertebrate sequences of Release N. 107 of GenBank for target genes of the family of vertebrate E2F transcription factors. These proteins have been implicated in the control of cell proliferation. In fact, functional E2F binding sites have been found within nuclear oncogenes and within genes that control cell cycle progression and engagement of DNA synthesis (Helin, 1998). E2F sites are often located in proximity to the transcription start site of genes carrying TATA-less promoters, according to the proposed E2F role in the formation of pre-initiation complex as an initiator binding protein (Helin, 1998). The preferential localization of E2F sites was exploited by limiting TargetFinder search to regions surrounding the transcription start site of GenBank sequences.

The matrix similarity was set to the lowest value (0.896) capable of re-identifying all the sequences that were used for deriving the nucleotide distribution matrix. The TRANSFAC V$E2F_02 matrix was used as query to the program. It was also interesting to compare TargetFinder performance with that of existing programs. Therefore, we also used MatInspector and ProfileSearch and FindPatterns (Wisconsin Package Version 9.1, Genetics Computer Group, Madison, WI), to perform the same screening of GenBank sequences for potential E2F target sequences. In all cases, the database screened was GenBank, Release N. 107. We also used all the above programs to screen Release N. 55 of the EPD (July 1998), using the same search conditions employed for the screening of GenBank sequences. Results are detailed in Table 1. The number of real E2F target genes detected by TargetFinder in GenBank (14) is comparable to that found by the other programs, despite the considerable diminution of output size. This an important issue, since a program capable of finding all possible target sequences of a TF, that also yields a great search background (i.e. thousands of spurious matches), would be useless for formulating experimental testable hypotheses. While the reduction of the amount of output by including additional search criteria is due to the TargetFinder algorithm, the increase in selectivity is due to the underlying MatInspector algorithm. In fact, this program alone scores better than ProfileSearch (15 versus 13 real E2F target genes found) using the same nucleotide distribution matrix. A good performance is also achieved by TargetFinder when searching the EPD database. In fact, while only seven known target genes are found, the total number of reported matches, 72, is rather small, yielding a good signal-to-noise ratio. However, the result also reflects the fact that only a limited number of E2F target genes have been entered into the EPD database at the present time.


**Acknowledgements**

We are indebted to Thomas Werner and Kerstin Quandt for permission to use MatInd and MatInspector programs. We are grateful to Gyorgy Simon for the excellent management of the TIGR net computer resource. We thank Ms Monica Sottocorno for skilful secretarial assistance. This work was supported by grants from the EC BIOTECH and BIOMED Programs, the Telethon-Italia Program and the AIRC to E.B.

**References**


