BIOINFORMATICS APPLICATIONS NOTE
Vol. 16 no. 8 2000
Pages 739–740

TRES: comparative promoter sequence analysis

Mukund V. Katti[1], Meena K. Sakharkar[2], Prabhakar K. Ranjekar[1] and Vidya S. Gupta[1,3]

[1]Plant Molecular Biology Unit, Division of Biochemical Sciences, National Chemical Laboratory, Pune 411 008, INDIA and [2]Bioinformatics Centre, National University of Singapore, Singapore

Received on February 9, 2000; revised on March 30, 2000; accepted on April 12, 2000

Abstract

Summary: *Comparative promoter analysis is a promising strategy for elucidation of common regulatory modules conserved in evolutionarily related sequences or in genes showing common expression profiles. To facilitate such analysis, we have developed a software tool that detects conserved transcription factor binding sites, cis-elements, palindromes and k-tuples simultaneously in a set of sequences, and thus helps to identify putative motifs for designing further experiments.*

Availability: *The program TRES can be accessed at URL: http://www.bic.nus.edu.sg:8888/tres*

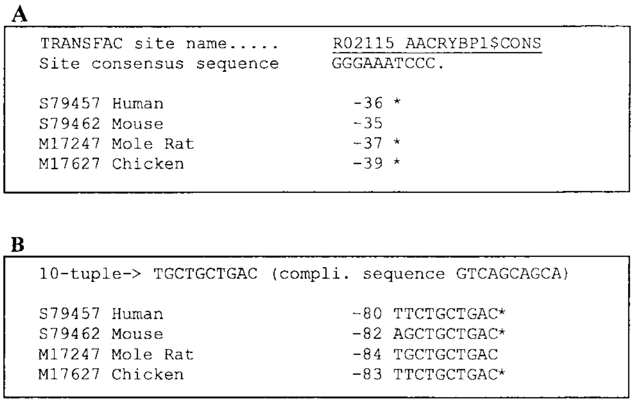
Contact: *vidya@ems.ncl.res.in*

Introduction

Computational search of promoter DNA sequences helps to identify putative sequence motifs possibly involved in transcription regulation. Several computational tools are available to search a given promoter sequence for potential transcription factor (TF) binding sites or cis-acting elements (e.g. Chen *et al.*, 1995; Quandt *et al.*, 1995; Prestridge, 1996; Frech *et al.*, 1997a,b, and references therein). However, when a single promoter sequence is searched using these methods, one often finds several motifs conserved all over the sequence making it difficult to analyse each of them experimentally.

Rather than searching a single sequence, simultaneous analysis of several related sequences can be more informative and useful to identify common regulatory modules. Considering effectiveness of this approach, we developed a web based software tool useful for comparative analysis of evolutionarily or functionally related promoter sequences.

Program organization



```
TRANSFAC site name.....       R02115 AACRYBP1$CONS
Site consensus sequence       GGGAAATCCC.

S79457 Human                    -36 *
S79462 Mouse                    -35
M17247 Mole Rat                 -37 *
M17627 Chicken                  -39 *
```

```
10-tuple-> TGCTGCTGAC (compli. sequence GTCAGCAGCA)

S79457 Human                    -80 TTCTGCTGAC*
S79462 Mouse                    -82 AGCTGCTGAC*
M17247 Mole Rat                 -84 TGCTGCTGAC
M17627 Chicken                  -83 TTCTGCTGAC*
```

Fig. 1. Part of TRES output file showing (a) AACRYBP1 site and (b) a 10-tuple corresponding to DE-1 element, conserved in human, mouse, mole rat and chicken eye lens specific alpha-A crystallin promoter sequences. Site locations are relative to TATA box and * indicates one mismatch.

The program TRES (Transcription Regulatory Element Search) is written in ‘C’ and implemented on a Unix server. Using TRES, as many as 20 promoter sequences, each of maximum 1000 bp, can be simultaneously searched for putative regulatory elements. TRES has the following four analysis tools:

1. **Matrix search:** This program scans the input sequences for conserved TF binding sites using matrices described in TRANSFAC database (Heinemeyer *et al.*, 1999). From the nucleotide distribution matrices, the position weights and matrix similarity scores are calculated essentially according to Quandt *et al.* (1995).
2. **IUPAC-string search:** Using this program input sequences can be searched for TF binding sites or cis-acting elements based on IUPAC consensus sequences described for the sites. Currently, a total of 3980 TF binding sites from TRANSFAC database (Heinemeyer *et al.*, 1999), 5919 sites from ooTFD database (Ghosh, 2000) and 240 plant cis-acting elements from PLACE database (Higo *et al.*, 1999) can be searched.

[3]To whom correspondance should be addressed Vidya S. Gupta, Division of Biochemical Sciences, National Chemical Laboratory, Pune-411 008, INDIA

3. **Palindrome search:** This program detects palindromic motifs conserved in the sequences. If $b_i \in \{A, T, G, C\}$, N is any base and c_i is the complementary base to b_i , then the program searches palindromes of the form $b_1b_2(0\text{-}6N)c_2c_1$, $b_1b_2b_3(0\text{-}6N)c_3c_2c_1$, $b_1b_2b_3b_4(0\text{-}6N)c_4c_3c_2c_1$ and $b_1b_2b_3b_4b_5(0\text{-}6N)c_5c_4c_3c_2c_1$.
4. **k-tuple search:** This program searches all possible sub-strings (k-tuples) of length 6 to 50 bases conserved in a given set of sequences. Essentially each k-tuple is a window of size 'k' sliding over all the sequences and searched on both the strands at a given mismatch level.

The interactive web interface enables the user to select program module, choose search parameters and submit sequences for online search. For all the conserved TRANSFAC, ooTFD and PLACE sites reported in the output file, hyperlinks are provided to the corresponding entry in the respective database and thereby further information can be retrieved.

Discussion

TRES makes use of known information on transcription factor binding sites/cis-elements and at the same time can detect new putative motifs (palindromes, k-tuples or phylogenetic footprints). The main advantage of TRES over other available programs is that it can analyse many related sequences at a time and report only the sites that are conserved in all or in the majority of the sequences. Thus motifs that occur only in one or few sequences, possibly due to chance, can be filtered. On the other hand, TRES identifies motifs conserved across the diverse sequences and such motifs can be expected to be functionally important.

Comparative analysis of orthologous sequences, phylogenetic footprinting, has been shown to be effective to identify evolutionarily conserved functional motifs (Gumucio *et al.*, 1996). However, for such study it is necessary that selected sequences be from moderately diverse species so that there has been sufficient evolutionary time for mutations to accumulate in non-specific regions (Duret and Bucher, 1997).

TRES can be also useful to identify common regulatory modules in genes that show similar patterns of expression. Recent developments in micro-array based mRNA quantification techniques make it possible to identify genes with common regulatory programs (Bucher, 1999). Thus, with ever-increasing sequence information available from diverse species, comparative promoter analysis appears to be a promising strategy to identify regulatory modules in genes of interest.

Acknowledgements

MVK acknowledges the Council of Scientific and Industrial Research, New Delhi, India for an award of a Senior Research Fellowship. We wish to thank Dr Edger Wingender, Dr David Ghosh and Dr Kenichi Higo for making available the site libraries from TRANSFAC, ooTFD and PLACE databases, respectively.

References

Bucher,P. (1999) Regulatory elements and expression profiles. *Curr. Opinion Struct. Biol.*, **9**, 400–407.

Chen,Q.K., Hertz,G.Z. and Stormo,G.D. (1995) MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Comput. Appl. Biosci.*, **11**, 563–566.

Duret,L. and Bucher,P. (1997) Searching for regulatory elements in human noncoding sequences. *Curr. Opinion Struct. Biol.*, **7**, 399–406.

Frech,K., Dietze,P. and Werner,T. (1997a) ConsInspector 3.0: new library and enhanced functionality. *Comput. Appl. Biosci.*, **13**, 109–110.

Frech,K., Quandt,K. and Werner,T. (1997b) Software for the analysis of DNA sequence elements of transcription. *Comput. Appl. Biosci.*, **13**, 89–97.

Ghosh,D. (2000) Object oriented Transcription Factor Database (ooTFD). *Nucleic Acids Res.*, **28**, 308–310.

Gumucio,D.L., Shelton,D.A., Zhu,W., Millinoff,D., Gray,T., Bock,J.H., Slightom,J.L. and Goodman,M. (1996) Evolutionary strategies for the elucidation of *cis* and *trans* factors that regulate the developmental switching programs of the β-like globin genes.. *Mol. Phylogenet. Evol.*, **5**, 18–32.

Heinemeyer,T., Chen,X., Karas,H., Kel,A.E., Kel,O.V., Liebich,I., Meinhardt,T., Reuter,I., Schacherer,F. and Wingender,E. (1999) Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms. *Nucleic Acids Res.*, **27**, 318–322.

Higo,K., Ugawa,Y., Iwamoto,M. and Korenaga,T. (1999) Plant cis-acting regulatory DNA elements (PLACE) database:1999. *Nucleic Acids Res.*, **27**, 297–300.

Prestridge,D.S. (1996) SIGNAL SCAN 4.0: additional databases and sequence formats. *Comput. Appl. Biosci.*, **12**, 157–160.

Quandt,K., Frech,K., Karas,H., Wingender,E. and Werner,T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.*, **23**, 4878–4884.