Nucleic Acids Research, 2003, Vol. 31, No. 13 ***3605–3607***
DOI: 10.1093/nar/gkg517

Dragon ERE Finder version 2: a tool for accurate detection and analysis of estrogen response elements in vertebrate genomes

Vladimir B. Bajic*, Sin Lam Tan, Allen Chong, Suisheng Tang, Anders Ström[1], Jan-Åke Gustafsson[2], Chin-Yo Lin[3] and Edison T. Liu[3]

Knowledge Extraction Laboratory, Institute for Infocomm Research, Singapore 119613, [1]Center for Biotechnology and [2]Department of Medical Nutrition, Karolinska Institute, Novum, S-141 57 Huddinge, Sweden and [3]Genome Institute of Singapore, Singapore 117528

Received February 2, 2003; Revised and Accepted March 18, 2003

ABSTRACT

We present a unique program for identification of estrogen response elements (EREs) in genomic DNA and related analyses. The detection algorithm was tested on several large datasets and makes one prediction in 13 300 nt while achieving a sensitivity of 83%. Users can further investigate selected regions around the identified ERE patterns for transcription factor binding sites based on the TRANSFAC database. It is also possible to search for candidate human genes with a match for the identified EREs and their flanking regions within EPD annotated promoters. Additionally, users can search among the extended promoter regions of ~11 000 human genes for those that have a high degree of similarity to the identified ERE patterns. Dragon ERE Finder version 2 is freely available for academic and non-profit users (http://sdmc.lit.org.sg/ERE-V2/index).

INTRODUCTION

Beyond its well recognized role as a female sex hormone, estrogen is involved in the regulation of cell proliferation and differentiation in a variety of tissues in both males and females (1–3). Estrogen receptors (ERs) mediate hormone functions by binding to target gene promoters at sites named the estrogen response elements (EREs) (4). The signal pathways of estrogen are currently under debate (5,6). At least four mechanisms are possible in the estrogen reaction:

1. Classical pathway where estrogen-activated ERs bind directly to EREs and induce changes in gene expression.
2. Hormone-independent pathways where ERs are activated by growth factors but still bind directly to EREs.
3. Activated ERs bind indirectly to non-ERE sites (AP-1 and Sp-1 sites) through interaction with other transcription factors such as Jun and Fos.
4. Cell membrane signaling pathways which result in fast tissue responses without involving gene expression.

Dragon ERE Finder focuses on the EREs utilized for the first and second mechanisms and specific subsets of ER target genes. The consensus sequence of ERE is composed of palindromic half-sites intervened by three nucleotides (7). However, all other EREs identified to date have one or more base deviations from the consensus (8,9). In addition, the length of interval between the two half-sites, as well as the immediate flanking dinucleotide sequences are also important in determining ER binding and hormone induction (8,10–12). Identification and characterization of transcriptional mechanisms of genes regulated by estrogen are of fundamental importance in understanding the diverse functions of the hormone. Thus, it is crucial to have a computational tool that can pinpoint candidate ERE patterns across anonymous DNA and help to identify genes regulated by direct ER binding to their promoters. Tools for this purpose should possess a high sensitivity for the detection of putative functional ERE patterns, but, at the same time, they should not make so many predictions that the interpretation of results is made impossible. We have developed the Dragon ERE Finder (DEREF) program which fits the above criteria and is now freely available on the web (http://sdmc.lit.org.sg/ERE-V2/index) for academic and non-profit users. Details of the implemented recognition algorithm are provided on the web site.

PROGRAM DESCRIPTION

The program uses a probabilistic model to detect ERE patterns. The model is described on the program's web page. Its portal allows users to input the sequence in FASTA, GenBank, EMBL, IG, GCG or plain format by either pasting it into an input box or by reading it from a file. The current version 2.0 of the program uses the same algorithm for detection of the ERE patterns, as version 1.0. However, the report files and

*To whom correspondence should be addressed. Tel: +65 68748800; Fax: +65 67748056; Email: bajicv@i2r.a-star.edu.sg

aims of these program versions are different. Version 1.0 (http://sdmc.lit.org.sg/ERE/) is suitable for the analysis of long DNA sequences, up to 100 000 nt (nucleotides), in one session, but analyses only the forward strand of the query sequence. Its sample report file is shown in the Supplementary Material (Fig. S1). Version 2 of the program (http://sdmc.lit.org.sg/ERE-V2/index) is primarily aimed at giving more detailed interactive analyses of the identified ERE patterns, although it is possible to run it in a non-interactive mode (see sample report file in Fig. S2) from the same window. We describe here the use of version 2.0. When a pattern is detected along a DNA sequence, a page with the following information is displayed:

1. the position of the pattern on the DNA;
2. the actual nucleotide composition of the identified ERE pattern;
3. the nature of the identified pattern (i.e. does it belong to a known pattern used for training or is it a new ERE pattern not used for training).

A sample report page is shown in Figure S3.

Users can carry out further analyses of a selected region around the identified ERE pattern. On the next page, Figure S4, the user faces several options. Firstly, using an integrated BLAST program (13), it is possible for the user to search for a match of an identified ERE pattern (17 nt in length) with its flanking 20 nt on either side against the promoter sequences of EPD (14). The sample report page, with the identified ERE pattern highlighted for easy inspection of the quality of the BLAST matches, is in Figure S5. Sequences from the EPD database cover proximal promoter region [−500, +100] relative to the transcription start site (TSS). Similarly, the same 57 nt sequence used to match the EPD data can be matched, using BLAST, against about ∼11 000 human promoters covering the region of [−3000, +1000] relative to the start of the gene, as defined by the FIE2 program (http://sdmc.lit.org.sg/FIE2.0) (15). The results of the BLAST analysis are presented with the identified ERE pattern highlighted (sample report page is shown in Fig. S6). Finally, the user can select a region around the identified ERE pattern and subject that sequence to analysis by the MATCH program (16) which uses TRANSFAC database version 6.1 (16). The transcription factor binding sites (TFBSs) found with MATCH in the selected region are presented to the user through a graphical display with all pertinent information appended. This can help the user in identifying potentially important TFBSs in the immediate neighborhood of the ERE site (a sample report page is shown in Fig. S7).

Users have to pay attention to the following: (i) if the query sequence is shorter than 17 nt, the program will not make predictions; and (ii) if you are in the interactive mode and request BLAST of the found pattern against either the EPD or the internal Human Promoter database, while the sequence submitted to BLAST is shorter than 57 nt, no result will be produced. This is because the internal BLAST operation is based on the basic ERE motif (17 nt) flanked by 20 nt from each side.

SUGGESTIONS FOR EFFICIENT PROGRAM USE

If the interest is only in pinpointing the location of the putative ERE patterns in long segments of DNA, then the use of the non-interactive mode of DEREF v.2.0 is suggested due to its simple output format. Sequences which are longer than 100 000 nt would need to be segmented. The recommended sensitivity is 0.83, but the user can change it to meet desired requirements. The data correlating sensitivity to expected frequency of predictions is presented on the program's web page at http://sdmc.lit.org.sg/ERE-V2/DERE2_test.htm.

Moreover, if users are interested in finding genes controlled by direct ER binding to the promoters, they can use the Dragon Promoter Finder programs (17–19, http://sdmc.lit.org.sg/promoter/), to pinpoint potential TSSs, and then localize the search for EREs within these promoters. Since the ERE patterns could be in the range of [−3000, 1000] relative to the TSS, it is recommended that a region covering [−5000, +3000] relative to the found TSS be used for the ERE search.

However, if the interest is in exploring the specific genes, then the interactive mode of DEREF version 2.0 is far more suitable. This will give the user option for finding the candidate transcription factors that potentially interact with ER, by the analysis of the immediate surrounding region of the putative ERE pattern. If the examined sequence is of human origin, an internal database of ∼11 000 human gene promoters of high quality allows for the search of candidate paralogous genes among them, or for those genes that have the found ERE pattern and its close neighborhood highly preserved. Thus, it is possible to find other genes that have a high chance of being part of the gene network controlled by direct ER binding to promoter region.

If in the query sequence, ERE patterns are not reported by the system, the user may increase the sensitivity (the allowed range is 0.7–1.0) and repeat the analysis. Conversely, the user can reduce the sensitivity level if the detected ERE patterns seem to be false positive predictions. Reduction in sensitivity should decrease the number of potential false positives.

ACCURACY AND PREDICTIONS OF FUNCTIONAL SITES

The program performance is presented on http://sdmc.lit.org.sg/ERE-V2/DERE2_test.htm and it is obtained from the analysis over several diverse data-sets. The page also explains how test sequences and results were obtained and the main findings. In summary, at the sensitivity of 83%, the program makes one prediction per 13 300 nt, based on the analysis of the whole human chromosome 21. Chromosome 21 has a G + C content of 41% which is less than the average G+C content for the human genome. Prediction of ERE patterns is 2.8-fold higher in the promoter region of estrogen responsive versus non-responsive genes obtained through a gene array experiment. While one cannot assume that every new ERE pattern detected by the system is functional, we have confirmed that Dragon ERE Finder is capable of detecting new functional ERE patterns. The system correctly detected recently characterized functional EREs (20) in human and mouse lipocalin 2 genes at the default sensitivity setting. Also, the system detects ERE motifs (p17d1, p17daTA, p17d2) for which binding of ERα and ERβ is experimentally shown (9,21). The first two motifs are detected at the default sensitivity, while the

third one is detected at a sensitivity of 0.87. The p17d2 ERE motif is shown to be functional in COS-1 and HepG2 transfected cells (21).

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

REFERENCES

1. Grumbach,M.M. (2000) Estrogen, bone, growth and sex: a sea change in conventional wisdom. *J. Pediatr. Endocrinol. Metab.*, **13** (Suppl. 6), 1439–1455.
2. Rochira,V., Balestrieri,A., Faustini-Fustini,M. and Carani,C. (2001) Role of estrogen on bone in the human male: insights from the natural models of congenital estrogen deficiency. *Mol. Cell. Endocrinol.*, **178**, 215–220.
3. Nilsson,S., Mäkelä,S., Treuter,E., Tujague,M., Thomsen,J., Andersson,G., Enmark,E., Pettersson,K., Warner,M. and Gustafsson,J.-Å. (2001) Mechanisms of estrogen action. *Phys. Rev.*, **81**, 1535–1565.
4. Nilsson,S. and Gustafsson,J.-Å. (2000) Estrogen receptor transcription and transactivation: basic aspects of estrogen action. *Breast Cancer Res.*, **2**, 360–366.
5. Sanchez,R., Nguyen,D., Rocha,W., White,J.H. and Mader,S. (2002) Diversity in the mechanisms of gene regulation by estrogen receptors. *Bioessays*, **24**, 244–254.
6. Diel,P. (2002) Tissue-specific estrogenic response and molecular mechanisms. *Toxicol. Lett.*, **127**, 217–224.
7. Klein-Hitpass,L., Schorpp,M., Wagner,U. and Ryffel,G.U. (1986) An estrogen-responsive element derived from the 5′ flanking region of the Xenopus vitellogenin A2 gene functions in transfected human cells. *Cell*, **46**, 1053–1061.
8. Klinge,C.M. (2001) Estrogen receptor interaction with estrogen response elements. *Nucleic Acids Res.*, **29**, 2905–2919.
9. Driscoll,M.D., Sathya,G., Muyan,M., Klinge,C.M., Hilf,R. and Bambara,R.A. (1998) Sequence requirements for estrogen receptor binding to estrogen response elements. *J. Biol. Chem.*, **273**, 29321–29330.
10. Kraus,R.J., Ariazi,E.A., Farrell,M.L. and Mertz,J.E. (2002) Estrogen-related receptor α1 actively antagonizes estrogen receptor-regulated transcription in MCF-7 mammary cells. *J. Biol. Chem.*, **277**, 24826–24834.
11. Klinge,C.M., Jernigan,S.C., Smith,S.L., Tyulmenkov,V.V. and Kulakosky,P.C. (2001) Estrogen response element sequence impacts the conformation and transcriptional activity of estrogen receptor alpha. *Mol. Cell. Endocrinol.*, **174**, 151–166.
12. Wood,J.R., Likhite,V.S., Loven,M.A. and Nardulli,A.M. (2001) Allosteric modulation of estrogen receptor conformation by different estrogen response elements. *Mol. Endocrinol.*, **15**, 1114–1126.
13. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
14. Praz,V., Perier,R., Bonnard,C. and Bucher,P. (2002) The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. *Nucleic Acids Res.*, **30**, 322–324.
15. Chong,A., Zhang,G. and Bajic,V.B. (2002) Information and sequence extraction around the 5′-end and translation initiation site of human genes. *In Silico Biology*, **2**, 461–465.
16. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V., *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
17. Bajic,V.B., Seah,S.H., Chong,A., Zhang,G., Koh,J.L.Y. and Brusic,V. (2002) Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoters. *Bioinformatics*, **18**, 198–199.
18. Bajic,V.B., Chong,A., Seah,S.H. and Brusic,V. (2002) Intelligent system for vertebrate promoter recognition. *IEEE Intelligent Systems Magazine*, **17**, 64–70.
19. Bajic,V.B., Seah,S.H., Chong,A., Krishnan,S.P.T., Koh,J.L.Y. and Brusic,V. (2003) Computer model for recognition of functional transcription start sites in polymerase II promoters of vertebrates. *J. Mol. Graph. Model.*, **21**, 323–332.
20. Seth,P., Porter,D., Lahti-Domenici,J., Geng.Y., Richardson,A. and Polyak,K. (2002) Cellular and molecular targets of estrogen in normal human breast tissue. *Cancer Res.*, **62**, 4540–4544.
21. Yi,P., Driscoll,M.D., Huang,J., Bhagat,S., Hilf,R., Bambara,R.A. and Muyan,M. (2002) The effects of estrogen-responsive element- and ligand-induced structural changes on the recruitment of cofactors and transcriptional responses by ER alpha and ER beta. *Mol. Endocrinol.*, **16**, 674–693.