Nucleic Acids Research, 2004, Vol. 32, Web Server issue **W257–W260**
DOI: 10.1093/nar/gkh396

BEARR: Batch Extraction and Analysis of *cis*-Regulatory Regions

Vinsensius B. Vega*, Dhinoth Kumar Bangarusamy, Lance D. Miller, Edison T. Liu and Chin-Yo Lin

Genome Institute of Singapore, 60 Biopolis Street, #02-01, Singapore 138672

Received February 9, 2004; Revised and Accepted March 24, 2004

ABSTRACT

Transcription factors play important roles in regulating biological and disease processes. Microarray technology has enabled researchers to simultaneously monitor changes in the expression of thousands of transcripts. By identifying specific transcription factor binding sites in the *cis*-regulatory regions of differentially expressed genes, it is then possible to identify direct targets of transcription factors, model transcriptional regulatory networks and mine the dataset for relevant targets for experimental and clinical manipulation. We have developed web-based software to assist biologists in efficiently carrying out the analysis of microarray data from studies of specific transcription factors. Batch Extraction and Analysis of *cis*-Regulatory Regions, or BEARR, accepts gene identifier lists from microarray data analysis tools and facilitates identification, extraction and analysis of regulatory regions from the large amount of data that is typically generated in these types of studies. The software is publicly available at http://giscompute.gis.a-star.edu.sg/~vega/BEARR1.0/.

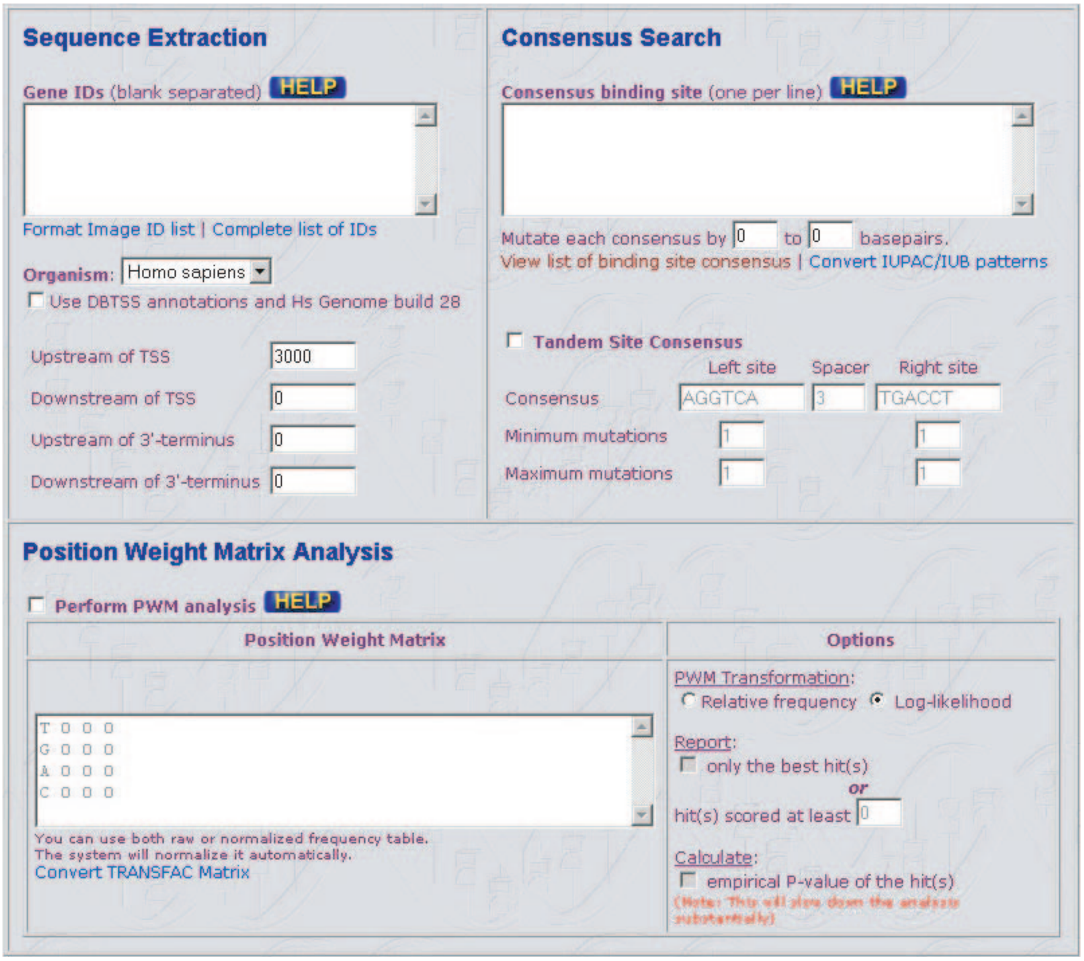
INTRODUCTION

Microarray technology has been applied extensively to interrogate changes in gene expression patterns that underlie biological and disease processes. The observed alterations in transcript levels are mediated by intracellular signalling pathways and downstream transcription factors that interact with specific regulatory regions of the genome and the transcription machinery. Therefore, DNA microarrays are especially well suited for experiments where specific transcription factors are known to play important roles. Transcription factors that are involved in cancers, such as p53 and Myc, have been studied using this new technology. By combining expression data and computational analysis of regulatory regions, it is then possible to begin identifying the molecular mechanisms of transcriptional regulation, to map the network of transcription factors and their target genes which affect the phenotypes being studied, and to generate additional hypothesis for testing in the laboratory [see comments in (1)].

Some online computational tools for mammalian gene regulatory region extraction and analysis are available to researchers, although with varying degrees of accessibility, capability and focus. The most comprehensive tools, including TOUCAN (2), EZRetrieve (3), the Genomatix suite (www.genomtix.de) and TRANSPLORER (www.biobase.de) from BIOBASE (the former two are from academic sources and the latter two are only available commercially for batch analysis), usually employ annotation-based sequence retrieval and position weight matrices, typically from the TRANSFAC database (4) or specialized databases, for sequence extraction (e.g. the upstream region extraction tool developed by the Harvard-Lipper Center, http://arep.med.harvard.edu/labgc/adnan/hsmmupstream) and binding site identification, respectively. There are also standalone [e.g. MEME (5), BioProspector (6), MDScan and REDUCE] and integrated [e.g. MotifSampler and phylogenetic footprinting modules in TOUCAN] motif discovery tools designed to uncover conserved sequences in extracted regulatory regions (7,8). While the sophisticated approaches in these programs represent innovations in comprehensive binding site prediction and discovery and are useful for detailed analysis and mining of refined or idealized datasets—for example members of the same gene family or known targets of specific transcription factors—nevertheless, based on our experience, they may be of limited use against the large amount of 'noisy' data generated in microarray studies that often confound the interpretation of the analytical results or fail to yield conserved regulatory sequences. Whereas these tools favour comprehensive approaches for identifying all known binding sites or conserved sequence discovery in the dataset, we propose

*To whom correspondence should be addressed. Tel: +65 6478 8041; Fax: +65 6478 9058; Email: vegav@gis.a-star.edu.sg

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors



Sequence Extraction

Gene IDs (blank separated) HELP

Format Image ID list | Complete list of IDs

Organism: Homo sapiens

□ Use DBTSS annotations and Hs Genome build 28

Upstream of TSS	3000
Downstream of TSS	0
Upstream of 3'-terminus	0
Downstream of 3'-terminus	0

Consensus Search

Consensus binding site (one per line) HELP

Mutate each consensus by 0 to 0 basepairs.

View list of binding site consensus | Convert IUPAC/IUB patterns

□ Tandem Site Consensus

	Left site	Spacer	Right site
Consensus	AGGTCA	3	TGACCT
Minimum mutations	1		1
Maximum mutations	1		1

Position Weight Matrix Analysis

□ Perform PWM analysis HELP

Position Weight Matrix

```
T 0 0 0
G 0 0 0
A 0 0 0
C 0 0 0
```

You can use both raw or normalized frequency table.
The system will normalize it automatically.

Convert TRANSFAC Matrix

Options

PWM Transformation:
○ Relative frequency ● Log-likelihood

Report:
□ only the best hit(s)
or
hit(s) scored at least 0

Calculate:
□ empirical P-value of the hit(s)
(Note: This will slow down the analysis substantially)

Figure 1. BEARR web interface.

that a critical and complementary step prior to a more comprehensive analysis and experimental validation is a hypothesis-driven and focused computational strategy based on the experimental design. Furthermore, although batch options are available for some of the tools noted, they have not been optimized for batch operations and are not suitable for the large amount of data that can be generated from genome-scale studies. Therefore, we created Batch Extraction and Analysis of *cis*-Regulatory Regions, or BEARR, to assist biologists in performing batch extraction and analysis of *cis*-Regulatory regions of hundreds or even thousands of differentially expressed genes identified in microarray studies. Here, we describe the system design and functionalities.

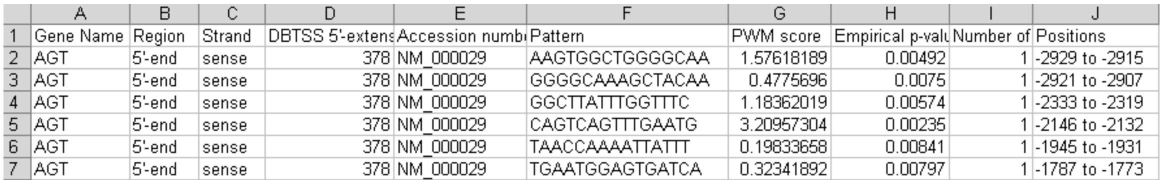
SYSTEM DESIGN

The system is divided into two parts: (i) the regulatory region sequence extraction module and (ii) the sequence analysis module. Essentially, the sequence extraction module generates the nucleotide sequences, using annotation and genomic sequence databases, to be queried by the analysis module, based on transcription factor binding sites or any conserved regions defined by the user. The modular system design allows for maximum component reusability and extensibility. To ensure platform independence and ease of installation, no specialized library or programs were employed. An interactive web-based user interface was developed for ease of use, interoperability and accessibility. Figure 1 shows a screenshot of the web interface. The system schema of BEARR is also available in the FAQ section of website.

REGULATORY REGION EXTRACTION

RefSeq accession numbers, recognizable by the NM_ and XM_prefixes, are the primary gene identifiers employed in the modules, although the system is capable of translating other identifiers from microarray analysis tools, including the UniGene cluster ID and the IMAGE clone ID. The list of gene identifiers can be easily copied from the output files of most expression analysis tools and pasted into the input window of the extraction module. This input feature allows users the flexibility of utilizing the optimal microarray data analysis tools to identify differentially expressed genes and then efficiently moving into the extraction and analysis of their respective regulatory regions. Currently, BEARR accepts identifiers for human and mouse genes.

The sequence extraction module utilizes NCBI's LocusLink and RefSeq (9) databases to identify the genes and



	A	B	C	D	E	F	G	H	I	J
1	Gene Name	Region	Strand	DBTSS 5'-extens	Accession numb	Pattern	PWM score	Empirical p-valu	Number of	Positions
2	AGT	5'-end	sense	378	NM_000029	AAGTGGCTGGGGCAA	1.57618189	0.00492	1	-2929 to -2915
3	AGT	5'-end	sense	378	NM_000029	GGGGCAAAGCTACAA	0.4775696	0.0075	1	-2921 to -2907
4	AGT	5'-end	sense	378	NM_000029	GGCTTATTTGGTTTC	1.18362019	0.00574	1	-2333 to -2319
5	AGT	5'-end	sense	378	NM_000029	CAGTCAGTTTGAATG	3.20957304	0.00235	1	-2146 to -2132
6	AGT	5'-end	sense	378	NM_000029	TAACCAAAATTATTT	0.19833658	0.00841	1	-1945 to -1931
7	AGT	5'-end	sense	378	NM_000029	TGAATGGAGTGATCA	0.32341892	0.00797	1	-1787 to -1773

Figure 2. Sample output from BEARR when used to identify potential functional Estrogen Response Element, based on TRANSFAC ERE position weight matrix. Output of consensus pattern searches is similar, with the exception of PWM score and empirical *P*-value columns.

pinpoint their loci within the genome. These annotations were chosen for their comprehensiveness, in terms of number of annotated genes, and their consistency with the current state of the NCBI contig databases, the underlying genomic sequence database used in BEARR. Using the transcription start site (TSS), defined as the 5′-most nucleotide in the reference transcript, and the 3′ terminus of the transcript as reference points, the module can extract user-defined regions both upstream and downstream of the references. Extraction speed has been enhanced by utilizing downloaded databases on the BEARR server rather than accessing the information on the NCBI servers via the internet. The resulting sequences are saved in the FASTA format and are downloadable by users. It is important to note that the TSS locations annotated in LocusLink and RefSeq may only approximate true start sites due to incomplete information at the 5′ ends of some reference sequences. The large number of annotated genes in LocusLink and RefSeq, however, provides the advantage of maximizing the number of regulatory regions we are able to extract for further analysis. To better annotate human start sites, we have incorporated the information from DBTSS (10), a database of experimentally extended 5′ sequences in human transcripts, to assist users in orienting the locations of predicted sites with reference to the more accurate TSS. It is also expected that as the sequence and annotation information is updated regularly in LocusLink and RefSeq, the accuracy of the current extraction method will improve.

SEQUENCE ANALYSIS

Once the desired sequences are obtained by the extraction module, the sequence analysis subsystem queries the input for putative transcription factor binding sites. Although it may be possible to map all known binding sites by incorporating information from TFD (11), TRANSFAC (12) and TRRD (13), BEARR is specifically designed to efficiently detect a focused group, as determined by experimental design or the microarray data, of transcription factor binding sites. We have, however, provided links to the information in TFD and TRANSFAC to enable users to extract consensus sites or matrices of their choice for use in BEARR. Users also have the option of inputting their own binding sites derived from the latest experimental data rather than relying on the potentially dated information in the databases. Extracted sequences can be queried using single or multiple consensus binding sites, including tandem sites, or position weight matrices of binding sites [PWMs; (14)]. Stringency is adjusted by defining the number of mismatches or spacer sequences allowed in the consensus searches or the similarity score threshold in the PWM output. The PWM score signifies the goodness of the match, where higher scores correspond to better matches. An optional empirical *P*-value calculation is also available for users who wish to assess the statistical significance of the motifs being identified against the null hypothesis of finding the motifs in randomly generated sequences. The analysis results are displayed in a tab-delimited text file and include the detected putative binding sites, number of hits and their positions relative to either 5′ end or 3′ terminus. Additional columns of PWM score, empirical *P*-value, and DBTSS 5′-extension are also included under the appropriate settings (see Figure 2). Further information of the analysis module can be found in the Help sections of the website.

In some cases, the user may find it necessary to redefine the published consensus sequences or matrices for binding sites because the reported derivations may be based on outdated or incomplete information. It is also highly recommended that the users optimize the stringency thresholds, using a test set of known target genes of the transcription factors of interest, prior to the analysis of the extracted sequences. These parameters will directly impact the sensitivity and the specificity of the analysis. For further detailed and comprehensive analysis of the extracted sequences, users can upload the FASTA sequence files into other available regulatory region analysis tools [e.g. MEME (5)] described above from the links on the website.

ANALYSIS EXAMPLES

BEARR was originally created to assist the discovery of potential estrogen receptor binding sites in the regulatory regions of hormone-responsive genes identified in our microarray experiments. We have subsequently used BEARR to analyse expression data from studies of other nuclear receptors and transcription factor families in order to identify target genes and downstream pathways. In general, this software is well suited for the analysis of data generated from study designs where transcription factor activity is manipulated experimentally. The software has also been applied in the analysis of the role of transcription factors identified in microarray studies of clinical samples in mediating the observed alterations in gene expression profiles of diseased tissues. In addition to the analysis of regulatory regions of genes identified in array studies, genome-wide surveys (all genes annotated in LocusLink and RefSeq) of transcription factor binding sites of interest, within defined regulatory regions, have been carried out to determine the frequency and distribution of putative binding sites. In a typical analysis flow, a user starts by formulating the underlying hypothesis, followed by designing and carrying out the appropriate experiments while at the same time studying the literature for related transcription factor binding sites to construct an initial binding site model. The experiments enable the user to group interesting genes, from which BEARR tries to identify potential functional binding sites based on the input transcription factor binding site model. Refinement of the model might also be necessary. Further biological investigations and experimentations could be performed on the binding sites found. Examples, tutorials and analysis workflows can be found in the online supplementary material on the website.

FUTURE DEVELOPMENTS

As more genomes are sequenced and annotated in LocusLink and RefSeq, additional databases will be integrated into BEARR to enable users to extract and analyse regulatory regions of genes from other model organisms. These upgrades will enhance the comparative analysis of conserved transcription factor binding sites. Other areas of development include motif search algorithms that will allow users to identify conserved sequences within the regulatory regions of co-regulated genes without prior knowledge of binding sites, i.e. consensus sequences or defined matrices. These additional functions will necessitate the development of a graphical results output for optimal data visualisation and analysis.

ACKNOWLEDGEMENTS

We especially thank Philip M. Long and Joshy George for the valuable inputs. The research and computational resource described here are supported by funding from the Biomedical Research Council (BMRC) of the Agency for Science, Technology and Research (A*STAR) in Singapore.

REFERENCES

1. Werner,T. (2001) The Promoter Connection. *Nature Genet.*, **29**, 105–106.
2. Aerts,S., Thijs,G., Coessens,B., Staes,M., Moreau,Y. and De Moor,B. (2003) TOUCAN: deciphering the *cis*-regulatory logic of coregulated genes. *Nucleic Acids Res.*, **31**, 1753–1764.
3. Zhang,H., Ramanathan,Y., Soteropoulos,P., Recce,M.L. and Toias,P.P. (2002) EZRetrieve: a web-server for batch retrieval of coordinate-specific human DNA sequences and underscoring putative transcription factor-binding sites. *Nucleic Acids Res.*, **30**, e121.
4. Wingender,E., Dietze,P., Karas,H., and Knuppel,R. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **28**, 316–319.
5. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, Menlo Park, CA, pp. 28–36.
6. Liu,X., Brutlag,D.L. and Liu,J.S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Proceedings of the Pacific Symposium on Biocomputing*, World Scientific Publishing, Hawaii, pp. 127–138.
7. Liu,X.S., Brutlag,D.L., and Liu,J.S. (2002) An algorithm for finding protein–DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature Biotech.*, **20**, 835–839.
8. Bussemaker,H.J., Li,H. and Siggia,E.D. (2001) Regulatory element detection using correlation with expression. *Nature Genet.*, **27**, 167–171.
9. Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acid Res.*, **29**, 137–140.
10. Suzuki,Y., Yamashita,R., Nakai,K., and Sugano,S. (2002) DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Res.*, **30**, 328–331.
11. Ghosh,D. (1993) Status of the transcription factors database (TFD). *Nucleic Acids Res.*, **21**, 3117–3118.
12. Wingender,E., Dietze,P., Karas,H. and Knuppel,R. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **28**, 316–319.
13. Kolchanov,N.A., Ignatieva,E.V., Ananko,E.A., Podkolodnaya,O.A., Stepanenko,I.L., Merkulova,T.I., Pozdnyakov,M.A., Podkolodny,N.L., Naumochkin,A.N. and Romashchenko,A.G. (2002) Transcription Regulatory Regions Database (TRRD): its status in 2002. *Nucleic Acids Res.*, **30**, 312–317.
14. Stormo,G.D. (1990) Consensus patterns in DNA. *Methods Enzymol.*, **183**, 211–221.