Nucleic Acids Research, 2006, Vol. 34, Database issue ***D111–D114***
doi:10.1093/nar/gkj052

ARED 3.0: the large and diverse AU-rich transcriptome

Tala Bakheet, Bryan R. G. Williams[2] and Khalid S. A. Khabar[1,2,*]

Department of Biostatistics, Epidemiology, and Scientific Computing (Bioinformatics Section), [1]Department of Biological and Medical Research, King Faisal Specialist Hospital and Research Center, Riyadh 11211, Saudi Arabia and [2]Department of Cancer Biology, Lerner Research Institute, Cleveland Clinic Foundation, Cleveland, OH 44195, USA

Received August 16, 2005; Revised and Accepted October 4, 2005

ABSTRACT

A comprehensive search that utilized a large set of mRNA data from human genome databases and additionally, expressed sequence tag (EST) database characterized this latest update of AU-rich elements (AREs) containing mRNA database (ARED). A large number of ARE-mRNA, as much as 4000, were recovered and include many of ARE alternative forms. This number represents as much as 5–8% of the human genes depending on the entire number of genes. The new ARED does not contain only larger and diverse number of ARE-mRNAs but additional functionality and enhanced search capabilities are given in the database website http://rc.kfshrc.edu.sa/ared/. These include class and cluster of AREs, source mRNAs, EST evidence, buildup information, retrieval of lists of genes, and integration with current and new NCBI data, such as Entrez ID and Unigene. Gene Ontology analysis shows there are significant differences in functional diversity of ARED when compared with the overall genome. Many of ARE-genes mediate regulatory processes, reactions to outside stimuli, RNA metabolism, and developmental processes particularly those of early and transient responses. The wide interest in mRNA turnover and importance of AREs in health and disease signify the compilation of ARE-genes.

INTRODUCTION

The availability of entire human genome and millions of records of expressed sequence tags (ESTs) made it possible to expand the repertoire of AU-rich mRNA information. The mRNAs that contain adenylate uridiylate (AU)-rich elements in their 3′-untranslated region (3′-UTR) comprise an important structural class of mRNAs with diverse functional repertoire. The wide interest in mRNA turnover and biological importance of AU-rich elements (AREs) in health and disease (1,2) necessitate further compiling of ARE-mRNAs.

In this latest ARED update, ARED 3.0, more than 4000 ARE-mRNAs have been computationally mapped to the human genome. This significant increase of ARE-mRNAs over previous ARED versions reflects both the breadth of human genome databases utilized and the computational extraction procedures. The new database comprised ARE-genes that were obtained from mRNA records and EST clustered data and represent broad functional classes with overrepresentation in genes involved in regulation, stress responses and other critical processes.

METHODS

Computational extraction of mRNA records from GenBank and RefSeq databases

The strategy for extracting GenBank mRNA records was previously outlined in ARED 2.0 methods (3). Briefly, GenBank release 135 (April 15, 2003) which contained 24 027 936 records and EMBL release 74 (March, 2003) which contained 23 234 788 records were utilized. The mRNA records were specifically extracted by first comprehensively extract all possible records followed by extracting records with the molecule type, mRNA. Subsequently, computational extraction of 3′-UTR from those with complete CDS using PERL program and GCG (Wisconsin Package) was performed; mRNA records with ARE search pattern were extracted and compiled. Redundancy was removed using CLEANUP program (4) and further refined by exclusion of duplicates corresponding to Unigene records. The corresponding Unigene records were extracted using PERL program that match the GenBank accession number with Unigene record.

Human RefSeq mRNA records (October, 2003) were used for further extraction of ARE-mRNA records using the same strategy above. The mRNA records from both sources (GenBank and RefSeq) mRNAs were further processed for removing redundancy by integration with Unigene data.

*To whom correspondence should be addressed. Tel: +1 966 1 442 7876, Fax: +1 966 1 442 7858; Email: khabar@kfshrc.edu.sa

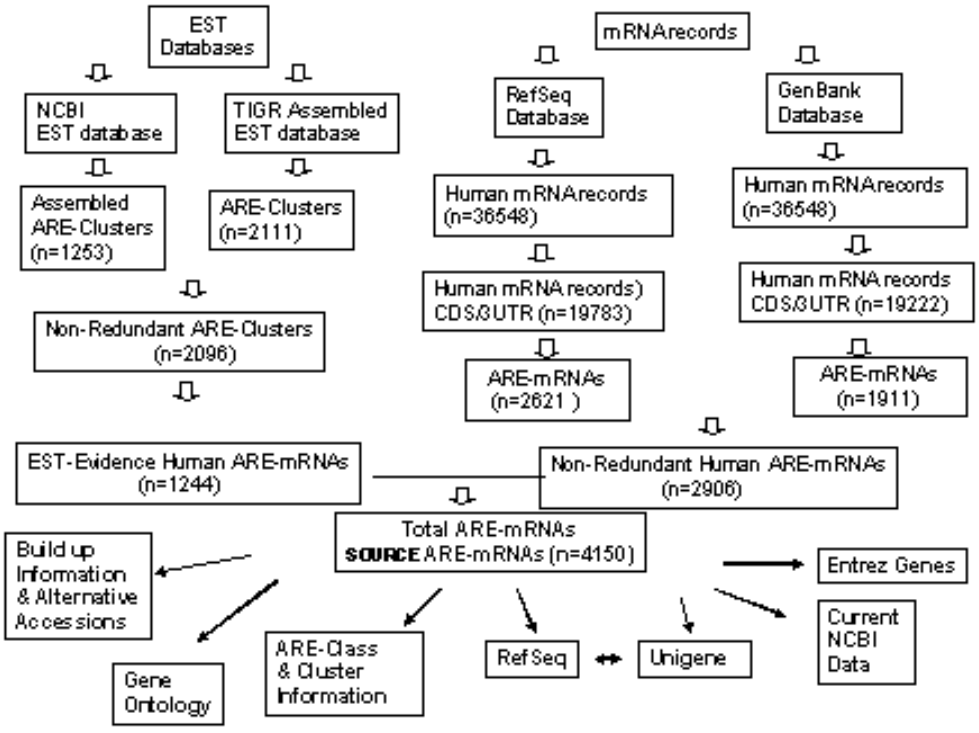

Figure 1. ARED 3.0 buildup. A schematic chart showing the stages of the ARED 3.0 buildup. Methods and computer programs used are described in Methods. *n* denotes number of sequences.

A later Unigene release (January, 2005) was used to update the database.

Computational extraction and clustering of ARE-mRNAs

Computational extraction of 3′-UTR was performed using ASSEMBLE and PERL codes. The 3′-UTRs were searched for the 13 bp pattern WWWU(AUUUA)UUUW with mismatch = −1 in the pentamer flanking regions. This pattern was computationally and functionally derived as described previously (5,6). Similarly, 3′ ESTs containing this pattern were retrieved from EST databases and mRNAs matching their cluster consensuses are described below. ARE classification was performed by two methods. One method is Chen ARE classification as Class I and Class II (7) and the other is based on number of pentamers as we previously described elsewhere (4,8). Classification of ARE-genes into Chen's ARE classification of Class I and Class II (9) was performed. If the 13 bp ARE pattern matches only one dispersed pentamer in the 3′-UTR, then the mRNA belongs to Class I whereas if the 13 bp pattern matches two pentamers or more, the mRNA belongs to Class II.

Computational extraction of ARE-mRNAs from EST databases

EST clustering buildup and the matching mRNAs were previously detailed (5). TIGR clustered EST transcripts were also used http://www.tigr.org/tigr-scripts/tgi/T_index.cgi?species=human. In both cases, ARE-mRNAs were extracted by choosing records in which ARE and poly(A) signal that are in correct order and orientation (5). Matching mRNA records with significant overlapping and similarity were retrieved from GenBank and RefSeq records using MEGABLAST at 95% identity and using the following criteria: word size of 32 bp, minimum Expect value of ($P < 0.00001$), limit by Entrez query of Homo sapiens [ORGANISM] and biomol_mRNA[PROP]. A PERL program was used to retrieve mapped Unigene records. Redundant Unigene records of the two mRNA sets (ours and TIGR) were removed. The data were integrated in ARED 3.0. Specifically, ARE-mRNAs, in which their ARE evidence are from EST evidence alone, were categorized as 'EST evidence' in the database records. In addition, EST evidence ARE-mRNAs were also classified to Class I and Class II and ARE bioinformatics clusters as explained above.

ARED 3.0 structure and data links

For each ARE-mRNA the following information and data link were associated: Buildup Unigene, Current Unigene, chromosome number, Gene Ontology (GO) information and ARE classification. The database is constructed using relational database structure. ARED 3.0 also contains different search formats including retrieval of a list of genes or accession numbers. All search results are downloadable as tab-delimited tables.

Gene Ontology analysis

ARE-genes with available GO information were used to analyze the representation of functional categories. This analysis was performed using FatiGO and GOStat programs (10,11). The *P*-values for significance of difference in GO categories

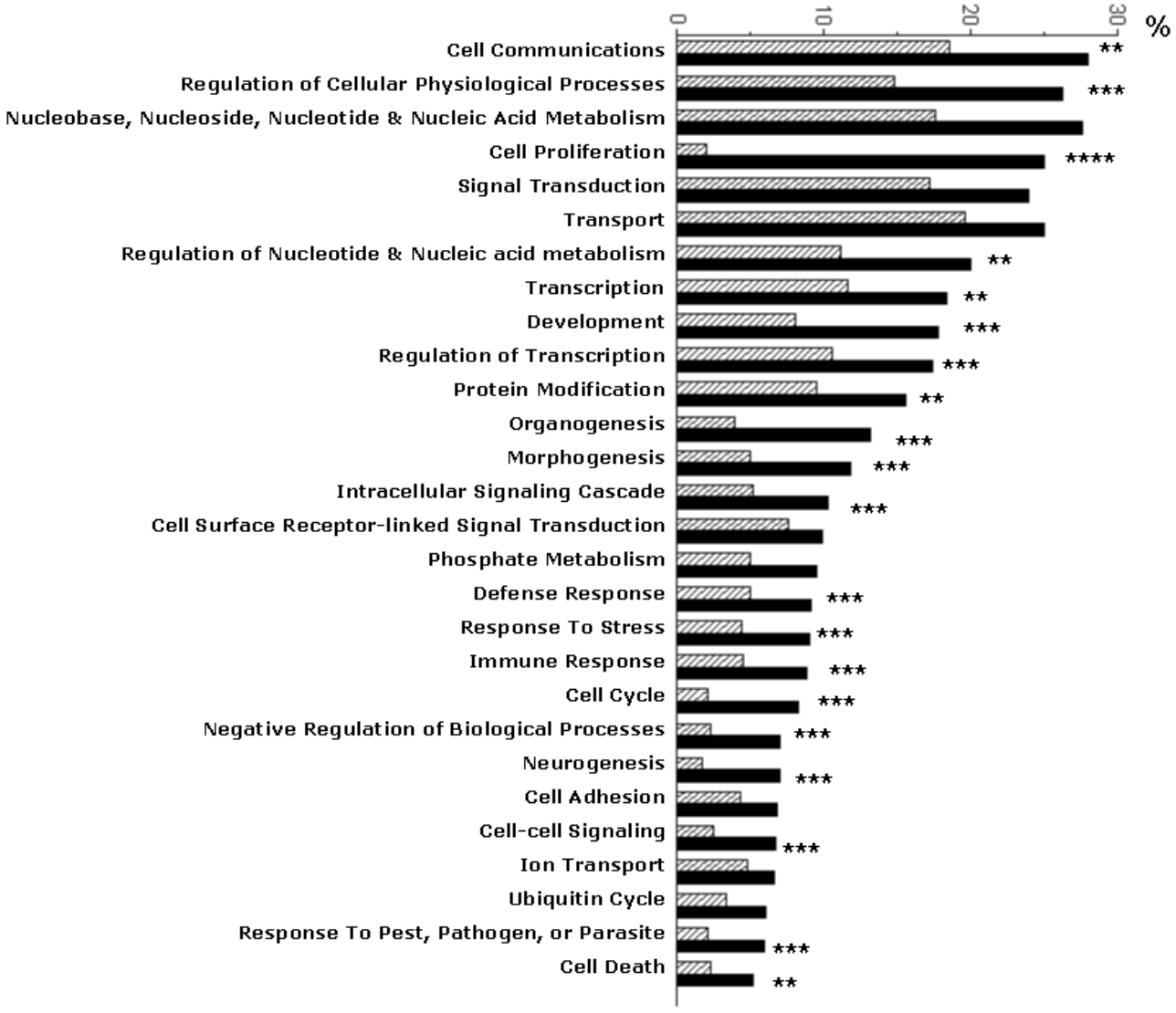

Figure 2. Gene Ontology (GO) of ARE-genes. Annotation of ARE-genes according to GO was performed using FatiGO program (http://www.fatigo.org) and compared with GO of overall genome using GOStat (http://gostat.wehi.edu.au). **, *** and **** denote *P*-values of $P < 0.01$, $P < 0.001$ and $P < 0.0001$, respectively, according to GOStat algorithm.

between ARED 3.0 and whole genome were imported from GOStat program.

RESULTS AND DISCUSSION

Extraction of ARE-mRNA records was performed using two sources: RefSeq and GenBank databases. While there were many common ARE-mRNAs in both databases, their compilation (both the numbers and content) were higher than if ARE-mRNAs were extracted from either of the database (Figure 1). We have made the search ARE pattern more stringent when compared with ARE patterns previously used in ARED 1.0 and ARED 2.0. This is because accumulation of several studies that associate ARE patterns with functional characteristics of AREs, e.g. primary and transient responses (6,12) resulted in more refined ARE pattern. This pattern WWWT(ATTTA)TTTW was unique to the 3′-UTR when compared with other regions of the mRNAs, i.e. 5′-UTR and coding region (4,5).

ARED 3.0 has additional database source, namely, EST database. Specifically, we utilized the information that was previously compiled using our 3′ end EST clustering approach (5,8). The 3′ ESTs constitute a rich source for 3′-UTR-specific AREs. This strategy allowed us to extract more than one thousand putative transcripts that include alternative forms due to 3′-UTR completeness, variant polyadenylation and splicing (5,8). Compilation with EST clusters from TIGR assembled database yielded a total of ∼2000 non-redundant ARE clusters (Figure 1). When compiled with the ARE-mRNAs from GenBank and RefSeq, there were ∼4000 mRNAs. These mRNAs clustered into more than 2500 unique ARE-genes and 400 non-HS mapped records.

Classification of ARE-genes into Chen's ARE classification of Class I and Class II, (9) in which Class I ARE-mRNA

contains one pentamer repeat in U-rich context and Class II is two pentamers or more, revealed that the majority belong to Class I (70%). This ratio is similar to the content of an ARE-cDNA library generated in the laboratory (8). The database in the website also provides information about the bioinformatics, in which Class II can be further classified into four cluster groups that are dependent of the number of pentamers (4).

It should be noted that the sequence-based ARE assignment in ARED is putative in nature and does not necessarily mean that every ARE in ARED confers regulatory control. Though it is more likely that ARE with longer stretches found in Class II AREs are functional (6,13,14), these and many of the other ARE-mRNAs need to be experimentally validated. The selection criteria used do not take into account either AREs loosely related to the 13 bp ARE search pattern, such as c-myc, nor AREs lacking the AUUUA motif, such as c-jun, which lack specific recognizable pattern derived from computational biology studies (6,11,15,16). Though the location of the AREs is not given in ARED due to heterogeneity of the length of the component sequences, a good resource to map Class II AREs can be found in UTRdb and UTRSCAN http://bighost.area.ba.cnr.it/BIG/UTRHome/ (17).

Analysis of Gene Ontology, which assigns genes into functional categories, was performed on ARED 3.0. The largest functional categories occur with those of regulatory processes, such as cell communications (28%), regulation of cellular physiological processes (26%), nucleobase, nucleoside, nucleotide and nucleic acid metabolism (27%), signal transduction (24%) and transcription (20%). Cell proliferation which comprised 25% of the gene category is significantly over-represented by 5-fold ($P < 0.0001$) when compared with overall genome. Similarly, processes involving developmental processes such as morphogenesis, organogeneses and neurogenesis, are significantly over-represented ($P < 0.001$) in ARE-genes (Figure 2). Many of significantly represented ARE-genes also include those of responses to outside stimuli including immune response and stress and of regulatory functions such as regulation of nucleotide metabolism and transcriptional regulation (Figure 2).

The ARED 3.0 website database includes a multitude of information and links, such as ARE classification, source of ARE-mRNA, Gene Ontology, Unigene and RefSeq information (Figure 1). Buildup information for each ARE-mRNA based on Unigene mapping is included with each entry whenever available. Since Unigene automated assembly and information can change with each release, a link to current Unigene is given. Information on EST evidence for those ARE-mRNAs in which their AREs were extracted using EST clustering are given in Buildup details. In addition to the enhanced search capability of ARED 3.0, the data can be downloaded as tab-delimited table.

The large and diverse repertoire of ARE-genes expands the array of the involvement of ARE-gene products in health and disease. Dysregualtion of ARE-mRNA stability is mediated predominantly by AREs. Thus, knowledge about which genes code for AREs in their mRNAs may lead to finding new pathways that operate during disease conditions such as cancer and inflammation.

ACKNOWLEDGEMENTS

The authors thank Ms Elcin Asyali for her excellent work in the databasing. The work is supported by King Abdelaziz City for Science and Technology (KACST grant AT-24-56 to K.S.A.K). Funding to pay the Open Access publication charges for this article was provided by KACST.

Conflict of interest statement. None declared.

REFERENCES

1. Khabar,K.S. (2005) The AU-rich transcriptome: more than interferons and cytokines, and its role in disease. *J. Interferon Cytokine Res.*, **25**, 1–10.
2. Espel,E. (2005) The role of the AU-rich elements of mRNAs in controlling translation. *Semin Cell Dev. Biol.*, **16**, 59–67.
3. Bakheet,T., Williams,B.R. and Khabar,K.S. (2003) ARED 2.0: A update for AU-rich element mRNA database. *Nucleic Acids Res.*, **31**, 421–423.
4. Bakheet,T., Frevel,M., Williams,B.R., Greer,W. and Khabar,K.S. (2001) ARED: human AU-rich element-containing mRNA database reveals an unexpectedly diverse functional repertoire of encoded proteins. *Nucleic Acids Res.*, **29**, 246–254.
5. Khabar,K.S., Bakheet,T. and Williams,B.R. (2005) AU-rich transient response transcripts in the human genome: expressed sequence tag clustering and gene discovery approach. *Genomics*, **85**, 165–175.
6. Raghavan,A., Dhalla,D., Bakheet,T., Ogilvi,R.L., Vlasova,I.A., Khabar,K.S., Williams,B.R. and Bohjanen,P.R. (2004) Patterns of coordinate down-regulation of ARE-containing transcripts following immune cell activation. *Genomics*, **12**, 1002–1013.
7. Chen,C.Y. and Shyu,A.B. (1994) Selective degradation of early-response-gene mRNAs: functional analyses of sequence features of the AU-rich elements. *Mol. Cell. Biol.*, **14**, 8471–8482.
8. Khabar,K.S., Dhalla,M., Al-Haj,L., Bakheet,T., Sy,C. and Naemmuddin,M. (2004) Selection of AU-rich transiently expressed sequences: reversal of cDNA abundance. *RNA*, **10**, 747–753.
9. Chen,C.Y. and Shyu,A.B. (1995) AU-rich elements: characterization and importance in mRNA degradation. *Trends Biochem. Sci.*, **20**, 465–470.
10. Al-Shahrour,F., Diaz-Uriarte,R. and Dopazo,J. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.
11. Beissbarth,T. and Speed,T.P. (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, **20**, 1464–1465.
12. Yang,E., Van Nimwegen,E., Zavolan,M., Rajewsky,N., Schroeder,M., Magnasco,M. and Darnell,J.E.,Jr (2003) Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Res.*, **13**, 1863–1872.
13. Frevel,M.A., Bakheet,T., Silva,A.M., Hissong,J.G., Khabar,K.S. and Williams,B.R. (2003) p38 Mitogen-activated protein kinase-dependent and -independent signaling of mRNA stability of AU-rich element-containing transcripts. *Mol. Cell. Biol.*, **23**, 425–436.
14. Raghavan,A., Ogilvie,R.L., Reilly,C., Abelson,M.L., Raghavan,S., Vasdewani,J., Krathwohl,M. and Bohjanen,P.R. (2002) Genome-wide analysis of mRNA decay in resting and activated primary human T lymphocytes. *Nucleic Acids Res.*, **30**, 5529–5538.
15. Conklin,D., Jonassen,I., Aasland,R. and Taylor,W.R. (2002) Association of nucleotide patterns with gene function classes: application to human 3′ untranslated sequences. *Bioinformatics*, **18**, 182–189.
16. Bakheet,T., Frevel,M., Williams,B.R.G., Greer,W. and Khabar,K.S.A. (2001) ARED: human AU-rich element-containing mRNA database reveals an unexpectedly diverse functional repertoire of encoded proteins. *Nucleic Acids Res.*, **29**, 246–254.
17. Pesole,G., Liuni,S., Grillo,G., Licciulli,F., Larizza,A., Makalowski,W. and Saccone,C. (2000) UTRdb and UTRsite: specialized databases of sequences and functional elements of 5′ and 3′ untranslated regions of eukaryotic mRNAs. *Nucleic Acids Res.*, **28**, 193–196.