Web Watch

TREP: a database for Triticeae repetitive elements

Thomas Wicker, David E. Matthews and Beat Keller

Genomes of the important crop plants wheat, barley and rye are large (5×10^9 to 17×10^9 bp) and contain 80% repetitive sequences. Although research on the molecular genetics of Triticeae concentrates on gene-rich regions, most genomic and some EST sequences isolated from these species are repetitive. Rapid identification of repetitive elements could significantly speed up the process of gene discovery and chromosome walking. Unfortunately, wading through the quagmire of repetitive sequences can be a difficult and uninspiring task. Classification and naming of such elements has been somewhat arbitrary and occasionally the same elements have even been classified and named differently by different researchers (e.g. '*Sabrina*' [1] and '*XA*' [2]). Often, repetitive elements are not present as complete copies but are fragmented by the insertion of other elements or by deletions, adding to the complexity of the analysis. In addition, frequently, only PCR fragments of conserved domains in repetitive elements (e.g. sequences encoding reverse transcriptase) are available.

'…wading through the quagmire of repetitive sequences can be a difficult and uninspiring task.'

GrainGenes (http://wheat.pw.usda.gov) initiated the TREP (Triticeae Repeat Sequence) database to help in the rapid identification of repetitive elements in genomic or cDNA sequences of Triticeae (http://wheat.pw.usda.gov/ITMI/Repeats). All elements entered in the database were either characterized from large genomic sequences or were found in studies specifically devoted to repetitive elements. TREP provides a high-quality standard because individual entries are annotated and contain only the element that is described in the annotation header and do not contain any flanking sequences, making it easier to determine the borders of repetitive elements in the sequence being searched against the database.

Individual entries are deposited under a unique identifier, the TREP accession number. It consists of the prefix 'TREP' followed by a number. In an effort to establish a consistent and relatively simple classification system, all elements are classified into five main groups, which themselves are further divided into subgroups (Table 1). The discovery of new classes of elements in the future might necessitate the addition of new groups and subgroups. All entries have an annotation header, which contains the classification, optional remarks and references as to where the sequence was originally published and deposited (GenBank accession number). The most important features of the sequences, such as long terminal repeats or coding regions, are annotated. In some cases, the researcher who contributed the sequence provides a detailed annotation.

TREP is divided into two databases, a complete and a non-redundant database. The non-redundant database can be searched by BLAST and contains only one or two copies of each type of element to make the search more efficient. Abundant element types, which can be divided into subgroups, such as *Stowaway* [3] or *BARE-1* elements [4], are present in two copies for each defined subgroup. This should give a better reference for each subgroup because repetitive elements often contain many mutations. Therefore, this division of the database could also be described as 'quasi non-redundant'. If available, only complete copies of elements have been used for this database.

The complete database contains all entries and is intended to allow more in-depth studies of the different element classes (Table 1). The TREP homepage provides a table of contents of all entries, with information about classification, source and the name of the researcher or institution that submitted the sequence. All sequences with their annotations can be downloaded, either individually or as one large file.

'…the number of described repetitive elements is expected to be in the thousands in the near future.'

The amount of available sequences from Triticeae is growing rapidly, and the number of described repetitive elements is expected to be in the thousands in the near future. To avoid confusion over terminology, common guidelines for classifying repetitive elements need to be established. A database for repetitive elements is the first step in this direction and serves as a

Table 1. Overview on the classification and number of repetitive elements in the TREP database[a]

Main group	Subgroup I	Subgroup II	Entries
Retrotransposon	LTR	*copia*	58
	LTR	*gypsy*	42
	LTR	Others	32
	LTR	TRIM	2
	Non-LTR	LINE	19
Foldback element	MITE	*Stowaway*	334
	MITE	*Tourist*	3
	MITE	Others	8
	LITE		7
	Others		1
Tandem repeat	*Afa*		88
	SCEN		69
	*Tai*I		45
	Others		28
Transposon			53
Unclassified			19
Total entries as of 4 October 2002			808

[a]Abbreviations: LINE, long interspersed nuclear element; LITE, large inverted-repeat transposable element; LTR, long terminal repeat; MITE, miniature inverted-repeat transposable element; SCEN, Saccharum centromeric sequence; TRIM, terminal-repeat retrotransposon in miniature.

 1360-1385/02/$ – see front matter PII: S1360-1385(02)02372-5

valuable tool in several different ways: (1) EST databases contain a considerable amount of repetitive sequences [5]. The information from TREP can be used to identify such sequences and mask them out. (2) The presence of multiple copies of repetitive elements often causes misalignment of shotgun sequences when large genomic regions from Triticeae are being sequenced. This database can be helpful and often essential in closing gaps and arranging the subcontigs into a contiguous sequence. (3) TREP allows systematic comparison of large sets of sequences contributed from varying sources and should increase our understanding of the evolution and possible function of these elements.

Thomas Wicker
Institute of Plant Biology, Zollikerstrasse 107, University Zurich, 8008 Zurich, Switzerland.

David E. Matthews
USDA-ARS, Dept of Plant Breeding, Cornell University, Ithaca, NY 14853, USA.

Beat Keller
Institute of Plant Biology, University Zurich, Zollikerstrasse 107, 8008 Zurich, Switzerland.
e-mail: bkeller@botinst.unizh.ch

References

1 Shirasu, K. *et al.* (2000) A contiguous 66 kb barley DNA sequence provides evidence for reversible genome expansion. *Genome Res.* 10, 908–915
2 Wicker, T. *et al.* (2001) Analysis of a contiguous 211 kb sequence in diploid wheat (*Triticum monococcum* L.) reveals multiple mechanisms of genome evolution. *Plant J.* 26, 307–316
3 Bureau, T. and Wessler, S.R. (1994) *Stowaway*: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. *Proc. Natl. Acad. Sci. U. S. A.* 9, 1411–1415
4 Manninen, I. and Schulman, A.H. (1993) *BARE*-1, a *copia*-like retroelement in barley (*Hordeum vulgare* L.). *Plant Mol. Biol.* 22, 829–846
5 Echenique, V. *et al.* (2002) Frequencies of Ty1-*copia* and Ty3-*gypsy* retroelements within the Triticeae EST databases. *Theor. Appl. Genet.* 104, 840–844

RIKEN *Arabidopsis* full-length cDNA database

Motoaki Seki, Masakazu Satou, Tetsuya Sakurai and Kazuo Shinozaki

Full-length cDNAs are essential, not only for correctly annotating encoded genes in a genomic sequence, but also for functional analysis of genes and their products. We isolated 155 144 RIKEN *Arabidopsis* full-length (RAFL) cDNA clones. The 3′-end ESTs of 155 144 RAFL cDNAs were clustered into 14 668 non-redundant cDNA groups, ~60% of predicted genes. We also obtained 5′-ESTs from 14 034 non-redundant cDNA groups and constructed a promoter database. Using the full-length cDNA microarray, we are studying expression profiles of *Arabidopsis* genes in response to various environmental stress conditions and hormone treatments. The database for sequence information of RAFL cDNA clones, promoter sequences and expression profiles is available at http://pfgweb.gsc.riken.go.jp/index.html. Here, we briefly explain the RAFL cDNA database.

RAFL cDNA database

Arabidopsis thaliana has been adopted as a model organism because of its small size, short generation time and high efficiency of transformation [1]. Furthermore, its complete genome sequence was determined in 2000 in the *Arabidopsis* genome sequencing project [2].

We constructed full-length cDNA libraries from *Arabidopsis* using the biotinylated CAP-trapper method, using trehalose-thermoactivated reverse transcriptase [3–5]. Nineteen full-length cDNA libraries were constructed from *Arabidopsis* plants grown under various stress, hormone and light conditions, from plants at various developmental stages, and from various plant tissues [5]. Single-pass sequencing of the cDNA clones was performed from the 3′-end. A total of 155 144 3′-ESTs were clustered and then mapped onto the *Arabidopsis* genome [5]. Finally, 14 668 non-redundant RAFL cDNA clones were identified and mapped on the *Arabidopsis* genome [5]. Assuming that the total number of *Arabidopsis* genes is ~25 000, the RAFL clones should account for ~60% of all *Arabidopsis* genes. The sequence information of the 14 668 RAFL cDNA clones is available at http://pfgweb.gsc.riken.go.jp/index.html. The sequence database of the RAFL cDNAs is useful, not only for correct annotation of predicted transcription units and gene products, but also for promoter analysis of *cis*-regulatory elements in transcriptional regulation.

Promoter database

Promoter sequences of transcripts can be obtained by comparing the 5′-end sequences of the RAFL cDNAs with the *Arabidopsis* genomic sequences. We also constructed a promoter database [5] on 14 034 RAFL cDNA clones using the PLACE (plant *cis*-acting regulatory DNA elements) database [6]. The 5′-ESTs of the 14 034 RAFL cDNAs were mapped onto *the Arabidopsis* genome using the BlastN program. The criterion for mapping the 5′-ESTs was sequence identity >98% within >100 bp-overlap. The genomic sequences hit with the 5′-ESTs with the highest score were used for the construction of the promoter database. The genomic sequences that were observed in the 1000-bp upstream regions of the 5′-termini of the RAFL cDNA clones were regarded as the promoter sequence of each RAFL cDNA clone. We then searched for ~300 known plant *cis*-acting elements in the 1000-bp promoter sequences of corresponding genes for the RAFL cDNA clones using the PLACE database [6]. The promoter database is useful for analyses of *cis*-acting elements based on expression profiles obtained with the microarray [7–10].

Microarray database

We are studying the expression profiles of *Arabidopsis* genes in response to various stress conditions and hormone treatments using the full-length cDNA microarray [7–10], as well as that of various mutants and transgenics. Recently, to publish microarray analysis data, preparation of supplemental tables on the expression profiling data is often requested. After the publication of our papers on expression profiling, supplemental tables on this work will be available at http://pfgweb.gsc.riken.go.jp/index.html. Our microarray database should be useful for analysing the expression profiles of

http://plants.trends.com 1360-1385/02/$ – see front matter PII: S1360-1385(02)02378-6