BIOINFORMATICS

Vol. 19 no. 17 2003, pages 2191–2198
DOI: 10.1093/bioinformatics/btg286

Automatic design of gene-specific sequence tags for genome-wide functional studies

Vincent Thareau[1,3], Patrice Déhais[2,§], Carine Serizet[1,3], Pierre Hilson[2,3], Pierre Rouzé[1,] and Sébastien Aubourg[2,3]*

[1]Laboratoire Associé de l'Institut National de Recherche Agronomique (France) and [2]Department of Plant Systems Biology, Flanders Interuniversity Institute of Biotechnology, Ghent University, Technologie Park 927, B-9052 Gent, Belgium and [3]Unité de Recherche en Génomique Végétale, UMR INRA-CNRS, 2, rue Gaston Crémieux, CP 5708, F-91057 Evry, France

Received on December 9, 2002; revised on March 6, 2003; accepted on May 8, 2003

ABSTRACT

Motivation: The availability of complete genome sequences allows the identification of short DNA segments that are specific to each annotated gene. Such unique gene sequence tags (GSTs) replace advantageously cDNAs in microarray transcript profiling experiments. In particular, probes corresponding to individual members of multigene families can be chosen carefully to avoid cross-hybridization events.

Results: The Specific Primer and Amplicon Design Software (SPADS) was constructed to delineate the more divergent regions in each gene by comparing them with a completely annotated genome sequence and to select optimal primer pairs for the polymerase chain reaction amplification of one divergent region per gene. SPADS is a unique integrated tool to design specific GSTs from any public or private genome sequences and allows the user to fine-tune GST size and specificity. SPADS has been used to obtain probes for whole genome and family-wide transcript profiling, as well as inserts for gene-specific knock-out experiments.

Availability: The GENOPLANTE™ SPADS source code and web interface are available upon request. The online version is accessible via http://genoplante-info.infobiogen.fr/spads and via http://oberon.fvms.ugent.be:8080/SPADS/

Contact: pierre.rouze@psb.ugent.be

1 INTRODUCTION

Duplicated genes and gene families are common in all sequenced model genomes. The fraction of genes that belong to paralogous groups in *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *Arabidopsis thaliana* is 15, 40 and 60%, respectively (Wolfe and Shields, 1997; Semple and Wolfe, 1999; Arabidopsis Genome Initiative, 2000). Duplicated genes share between 24 and 100% identity at the nucleotide level (Wolfe and Shields, 1997; Semple and Wolfe, 1999). Studies based on the hybridization between spotted DNA probes and labeled nucleic acid samples are hampered by high identity levels between gene family members. For example, cDNA probes from close paralogous genes may yield results that are difficult to interpret because the signal from a member expressed at a low level can be masked by cross-hybridization with a highly expressed transcript corresponding to a member of a multigene family. Furthermore, microarrays based on cDNA clones [mainly identified by expressed sequence tags (ESTs)] only represent a fraction of the genes of a given species and rely on information and clone tracking that are often not flawless (see, for instance, Halgren *et al*., 2001). To avoid having to rely on cDNA collections, microarrays can be produced with oligonucleotides (Lockhart *et al*., 1996; Hughes *et al*., 2001) or polymerase chain reaction (PCR) products amplified from genomic DNA templates. Such probe arrays are exhaustive when designed based on structural annotation of a whole genome and allow the distinction between similar sequences when chosen in divergent regions of closely related paralogous genes.

Programs that take into account gene structure to design PCR primers have been described (Podowski and Sonnhammer, 2001; Varotto *et al*., 2001), but none of them intend to select a primer pair on the criterion of amplicon specificity. On the other hand, some computer programs are able to select specific regions within genes (Mitsuhashi *et al*., 1994; Li and Stormo, 2001), albeit only for the design of long oligonucleotides. More recently, two recent programs, which take specificity into account during gene sequence tag (GST) selection, have been published [PRIMEGENS (Xu *et al*., 2002) and PROBEWIZ (Nielsen and Knudsen, 2002)], but the concept of exon–intron gene structure is missing from both programs. PROBEWIZ aims at designing probes from

*To whom correspondence should be addressed.
§ Present address: INRA–AGENA, BP 27, F-31326 Castanet-Tolosan Cedex, France.

cDNAs, and PRIMEGENS at searching for probes in open reading frames (ORFs), which typically focuses and restricts its use to the genomes of prokaryotes. The basic novelty of the software is that specific PCR probes for eukaryotic genes can be designed without prior need of the cognate transcripts to be cloned and to be available, by using genome sequences as input and providing GSTs and PCR primers for their amplification as output.

We have developed the Specific Primer and Amplicon Design Software (SPADS) to automatically select PCR amplicons in the least conserved regions within a group of genes. Each unique GST is defined by a pair of primers chosen to maximize the efficacy and the reliability of its PCR amplification. The user enters the reference sequence sets to select divergent regions and design primers that may range from a few paralogous genes to a complete genome. Because SPADS works on the genomic sequence, taking into account the intron–exon gene structure it can process eukaryotic gene models. Here, the SPADS construction is presented and how it performs in designing unique GSTs from the annotated genomes of two model organisms, *Arabidopsis* and yeast. GSTs have already been used to obtain genome-wide gene-specific probes for expression microarrays and inserts for knock-out experiments as well. Additional capabilities and applications are discussed.

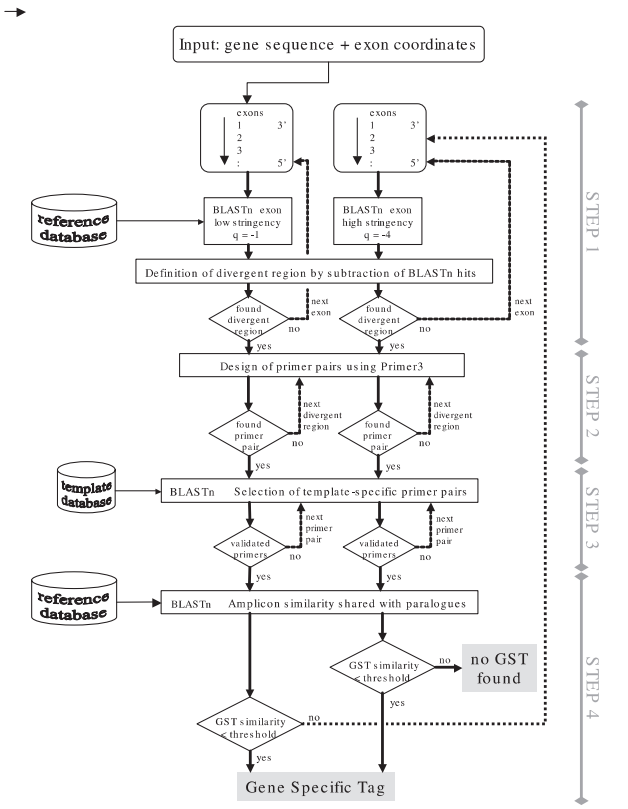

Fig. 1. SPADS flow chart. Thick and thin arrows represent the successive steps in the process and the inputs from the user or from databases, respectively. At each step, the procedure loops back (dashed arrows) and the next 5′ element (exon, primer pair, and amplicon) is processed when a given test (in a diamond) is not successful. The graph shows two parallel GST selection pipelines, the right one less stringent, with the BLAST mismatch parameter set at $q = -4$ and operative only when the left one, set at $q = -1$, fails to design a unique GST.

2 ALGORITHM

2.1 Overview of SPADS

SPADS selects gene-specific GSTs automatically based on gene models in which exons and introns are properly located on the sequence. It has been implemented in PERL and integrates the BLAST (Altschul *et al*., 1997) and Primer3 (Rozen and Skaletsky, 2000) algorithms. The procedure, schematically presented as a flow chart in Figure 1, can be summarized as four successive steps.

2.1.1 Search for divergent regions The exons of each gene are sequentially matched with BLASTn against a 'reference database', containing the full genome sequence (in the implementation described here, but, potentially, any other gene set). Segments with homology hits are removed so that primer pairs can be selected in the remaining regions.

2.1.2 Primer design The Primer3 software selects primer pairs in the divergent regions.

2.1.3 Selection of template-specific primer pairs Oligonucleotides designed by Primer3 are tested for specificity with BLASTn against a 'template database' that contains the DNA sequence corresponding to the template on which the PCR is to be performed. Primer pairs are discarded when matches indicate the potential risk of unwanted PCR amplification. The 'template database' can preferably be a subset of the genome, such as a bacterial artificial chromosome (BAC) clone, depending on the resources available for the experiment. When the PCR reaction cannot be restricted to a subset of the genomic DNA, the 'template database' used is identical to the 'reference database' described above (default option).

2.1.4 Analysis of GST similarity towards foreign sequences Amplicons defined by template-specific primer pairs are matched with BLASTn to the 'reference database' to determine their identity (%) with all possible paralogous sequences. An amplicon is discarded when its identity is above a user-defined threshold. At each step, SPADS scans every gene model from 3′ to 5′ until a validated GST that meets the identity constraint is identified.

2.2 Input

To design GSTs, the user must provide (i) the source sequence in FASTA format containing the target genes together with (ii) the corresponding exon position(s) (standard GenBank

format). A single input may include multiple genes in the source sequence without any limits in sequence size nor in gene numbers. Thus, all the genes on a very large sequence (for instance, a whole chromosome) can be processed as a single input. The 'reference' and 'template' databases have also to be chosen among those available or have to be provided by the user.

2.3 Step 1: search for divergent regions

To find the regions homologous with sequences from other genes, the nucleic acid sequence of an exon extracted from a gene model is compared with the 'reference database' by using BLASTn (Fig. 1). The mismatch parameter is fixed at $q = -1$, which is lower than the BLASTn default parameter; this low stringency results in long matching regions (hits). All hits larger than 40 base pairs (bp) or with more than 95% identity are taken into account. The coordinates of the divergent region(s) are deduced by subtraction of these hits, keeping the exonic regions on each side when large enough. Step 1 is completed for each gene, one exon at the time, starting with the most 3′ sequence to privilege the choice of unique GSTs in the 3′ end of the transcription unit. Indeed, 3′ tags are more likely to hybridize to their cognate transcripts, because most transcript labeling protocols based on polyadenylated mRNA samples include oligo-dT priming. A divergent region is used as input for Step 2, when its size is above the minimal user-determined GST length.

When no validated unique GST can be selected after this first scan across a given gene, the mismatch parameter of BLASTn is decreased to $q = -4$ to only delineate segments with stringent homology. This second run of BLASTn results in fewer and shorter hits and, by subtraction, identifies novel or larger divergent regions available for primer design.

In certain genes, the exons are too short to encompass a divergent region in which unique GSTs can be designed. In such cases, an option is left to the user to select a GST from the entire gene sequence, including introns, but still positioning both primers in exons. The maximum fraction of intronic sequence within the GST can be chosen to guarantee hybridization to transcribed sequences.

2.4 Step 2: primer design

The input of the Primer3 software (Rozen and Skaletsky, 2000) consists of substrings of gene sequences that are delimited by the coordinates of the divergent region (Step 1). Primer3 selects optimal PCR primerpairs and excludes oligonucleotide sequences favoring primer-dimers and hairpins. In SPADS, the Primer3 default values are set for the parameters that calculate maximum self-complementarity, maximum 3′ self-complementarity, 3′ end duplex stability, and Tm calculation. All 'general primer picking conditions' of Primer3 (such as Tm difference, primer length, G + C % in primer, and product length) can be modified by the user except for the parameters involved in the amplicon position because they are managed by SPADS itself. When no primer pair can be designed in a given divergent region, the next 5′ region is processed until an acceptable primer pair is identified.

2.5 Step 3: selection of template-specific primer pairs

This step is intended to remove primers that would potentially yield a contaminant amplicon in addition to the proper GST in PCR amplification. Each primer pair selected in Step 2 is matched by BLASTn (with mismatch penalty $q = -1$ and E-value cut-off $E = 200$) to the 'template database', which contains the template sequence used for PCR. Any primer pair with a hit located at a distance below 10 kb on the opposite strand and larger than 60% of the primer length is rejected, except when the hit contains a mismatch in the five 3′ bases. As mentioned above, this step can be tuned according to the PCR template source, genomic DNA, or individual (BAC) clones. BAC clones may be preferred to reduce the risk of spurious amplifications and to improve PCR yield.

2.6 Step 4: analysis of GST similarity towards foreign sequences

An amplicon defined by a primer pair in Step 3 is a candidate gene-specific GST. As last validation, each candidate GST is compared to the 'reference database' sequences with BLASTn to determine how similar it is to the closest gene with which it could potentially cross-hybridize. Identity is computed by summing up the matching bases in every local alignment of the amplicon with its closest paralogue. Because all matching bases have to be scored, the BLASTn sensitivity must be high. Therefore, to identify alignments in otherwise poorly conserved regions, mismatch penalty is low ($q = -1$) to enlarge local alignments, word size is short ($W = 7$) and E-value cut off is very high ($E = 500$). When the identity level is above the user-defined threshold, the amplicon is discarded and the next 5′ divergent region is processed.

As an alternative to the use of genomic sequences from the 'reference database', the GST identity check in Step 4 can be compared with all annotated transcripts for that given species, which significantly reduces computing time. Another open possibility is to match candidate GSTs only to all paralogous genes in related families, speeding up the process considerably by comparing only sequences for which the search of specific probes is most problematic. Yet, we favour genome-based analyses because they do not depend on gene prediction and annotation that are often erroneous or incomplete (Aubourg and Rouzé, 2001; Mathé *et al*., 2002).

2.7 Output

Figure 2 shows the standard format of the SPADS output file. It provides the sequence of the GST and of the corresponding PCR primer pair selected in a particular gene model, as well as additional information about them.

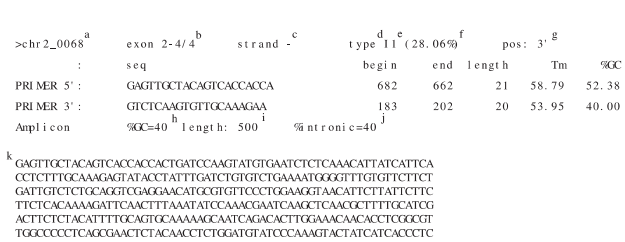

Fig. 2. SPADS output. The GST features are: (**a**) GST identification number; (**b**) exon(s) covered by the GST (in the example, the GST overlaps with exons 2, 3, and 4 out of a total of four exons in the gene model); (**c**) gene strand (+ or −); (**d**) GST type, with 'E' and 'I' for GST entirely in one exon or containing at least one intron, respectively; (**e**) GST specificity class, referring to the level of identity with the closest paralogous sequence (1, for 0–40%; 2, for 40–70%; and 3, for more than 70%); (**f**) actual % identity between the GST and the closest paralogue; (**g**) relative position in the transcript (5′, central, or 3′; for details, see Table 1); (**h**) G + C %; (**i**) GST length in nucleotides; (**j**) percentage of intronic sequence in the GST; (**k**) sequence of the GST. Lines 2, 3, and 4 describe the features of the 5′ and 3′ primers: sequence, coordinates in the input sequence, length, Tm based on nearest-neighbor calculation (Breslauer *et al*., 1986), and G + C %.

3 METHODS

3.1 System

The main body of SPADS is written in Perl (v5.004_04) for Sun4 Solaris. It integrates NCBI-BLAST (v2.2.1 or 2.2.2) and Primer3 (v0.9). Evaluation was computed on a Sun SPARC station running the Solaris operating system. SPADS may be operated with the command line of a UNIX system or a web interface built for limited requests.

3.2 SPADS parameters

The programmed parameters are either fixed or user defined. Default values are chosen to ensure successful amplification of sequence tags that yield gene-specific nucleic acid molecular hybridization. The threshold of maximum similarity of a unique GST with any other sequences has been set at 70% identity as default, which is an acceptable limit to avoid cross-hybridization (Girke *et al*., 2000; Miller *et al*., 2002). For more details, see Section 2.

3.3 SPADS evaluation

3.3.1 Parameters All GSTs were designed with the option that allow them to contain intron(s) up to 50% of the total sequence. The parameters for primers were: length, 18–25 bp, G + C %, 40–80%, and Tm, 50–65°C, with a maximum Tm difference of 4°C between paired primers. For SPADS evaluation, no identity threshold was applied, whereas for experimental validation, a threshold of 70% identity was used for all unique GSTs to similar sequences.

3.3.2 Computing time The selection of GSTs for the 1830 yeast genes within the test set, including the genome-wide search of template-specific primers, required 13 h (CPU time) on a Sun Enterprise 5500 with 6 GB of RAM and one 450 MHz SPARC CPU.

3.4 Sequences

3.4.1 Gene sets. For *Arabidopsis*, we used a set of 1814 genes for which the intron–exon structures had been experimentally defined by cognate full-length cDNA (http://genoplante-info.infobiogen.fr/Databases/PlantGene/).

The yeast set of 1830 genes, extracted from the genome annotation files available at the Saccharomyces Genome Database (SGD; ftp://genome-ftp.stanford.edu/pub/yeast/data_download/chromosomal_feature/archive) is defined by the gene coordinates on chromosomes I, II, III, IV, and V. The annotation files were downloaded on November 14, 2001.

3.4.2 BLAST reference databases The complete nuclear genome used as *Arabidopsis* reference database contains the five pseudo-molecule sequences (accession numbers 68 170, 51 595, 68 173, 68 164, and 68 172) downloaded from the web site (ftp://ftp.tigr.org/pub/data/a_thaliana/ath1) of The Institute of Genome Research on January 28, 2001, whereas that used as yeast reference database contains the 16 chromosome sequences (accession numbers NC_001133, NC_001134, NC_001135, NC_001136, NC_001137, NC_001138, NC_001139, NC_001140, NC_001141, NC_001142, NC_001143, NC_001144, NC_001145, NC_001146, NC_001147 and NC_001148) downloaded from the SGD web site (ftp://genomeftp.stanford.edu/pub/yeast/data_download/sequence/genomic_sequence/chromosomes/fasta/archive) on August 27, 2001. Because NCBI-BLAST (v2.2.1 and 2.2.2) does not work properly on sequences larger than 16 Mb (http://hg.wustl.edu/info/README.html), the *Arabidopsis* genome sequences were fragmented in two Megabase regions with 200 kb overlaps, prior to formatting the BLAST database with the Formatdb software.

4 RESULTS

4.1 Evaluation of SPADS

4.1.1 Gene sets Two random sets of genes were automatically processed with SPADS in series of gene-specific GST design tests: the 1814 experimentally characterized *Arabidopsis* genes and the 1830 genes predicted on chromosomes I, II, III, IV and V of yeast. The gene sets were large enough to provide relevant information at the genome level, but small enough for reiterative parameter evaluation. In all tests, both the reference and template databases contained the complete nuclear genome sequence of the organism under study. The comparative assessment of the SPADS results for the two eukaryotic genomes shows that the software is applicable to a wide range of organisms.

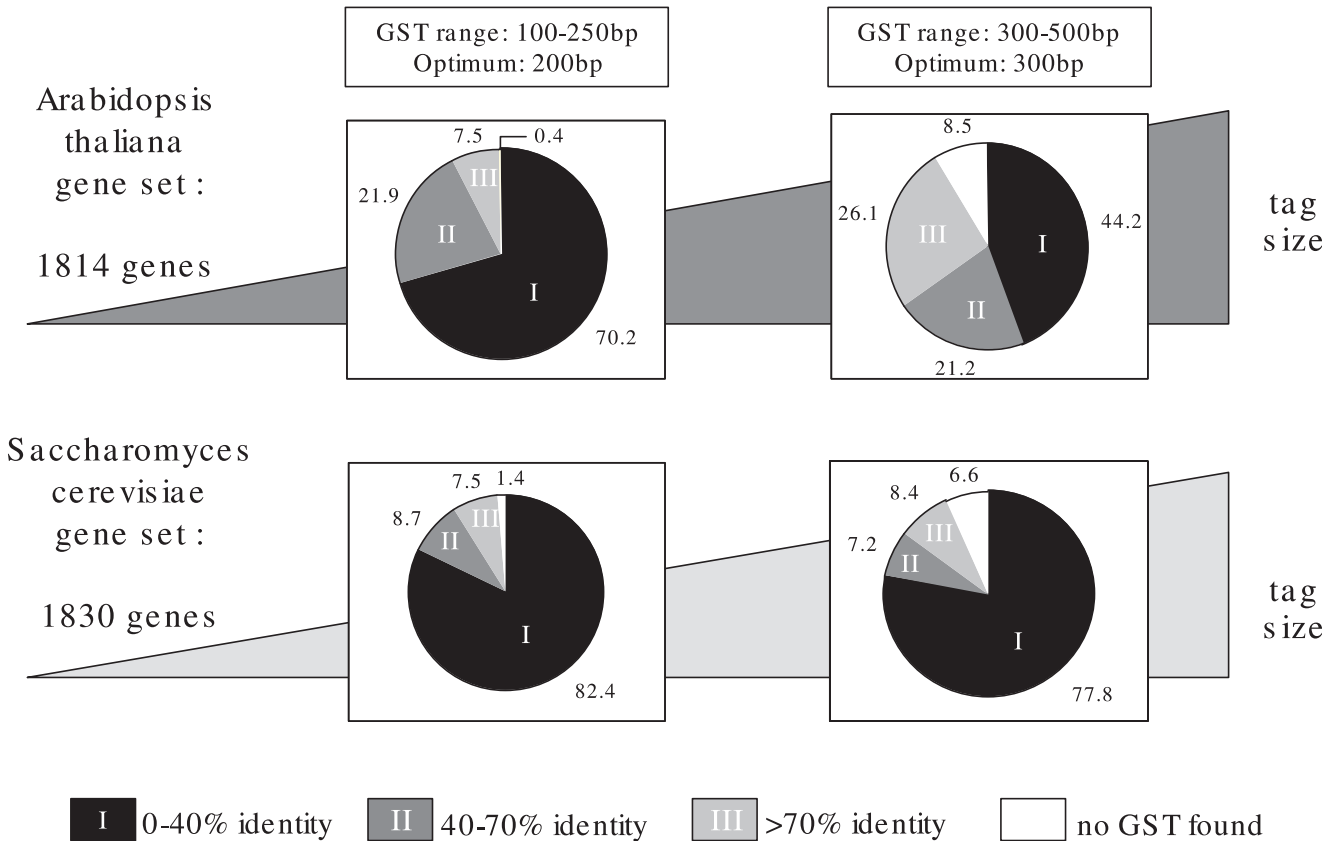

Fig. 3. Effect of GST size on the number and specificity of GSTs obtained for the *Arabidopsis* and yeast genomes. The design of GSTs was performed on two gene sets from *Arabidopsis* and yeast, with two parameters defining GST size range (Primer3 input parameters). Each GST was classified according to its specificity class, referring to the level of identity with the closest paralogous sequence [see Fig. 2(e)]. The remaining genes are those for which no GST was found.

4.1.2 GST size and specificity Smaller probes are more likely to be gene specific, but are expected to yield less stable hybridization. Therefore, we studied the effect of size on unique GST design. Figure 3 illustrates the SPADS success rate in finding a unique GST for each gene in the two test sets according to different GST size ranges. Each GST was ranked in one of the three classes according to the percentage of identity with its closest paralogous sequence: below 40% (class I), from 40 to 70% (class II), and over 70% (class III). GSTs in class III may potentially hybridize with paralogous sequences. An additional class gathered genes for which no GST could be found. Two GST size ranges were tested: 100–250 and 300–500 bp.

SPADS successfully designed a GST in most genes from both genomes (Fig. 3). The yeast gene set yielded a comparable class distribution for GSTs ranging from 300 to 500 bp, although the number of genes for which no GST was found increased up to 6.6% compared to 1.4% for the smaller size range. In contrast, the distribution of 300–500 bp GSTs extracted from the *Arabidopsis* gene set was markedly altered as the proportion of class III GSTs increased to 26.1% versus 7.5% for the smaller size range. Finally, in all tests, class II was larger in *Arabidopsis* than in yeast. The differences between the two organisms result from the fact that in *Arabidopsis* more genes belong to paralogous groups and genes are more fragmented by introns.

4.1.3 Relative position of unique GSTs In protocols based on oligo-dT/polyA annealing, the 3′ part of transcripts is preferentially labeled. Therefore, SPADS scans each gene from the 3′ to the 5′ end until it designs a GST fulfilling all selection criteria. To assess the results of this procedure, the distribution of unique GST positions for the GST size range 150–300 bp was computed in two manners. First, the distance between the end of each gene and its GST was measured and 85.4 and 96.4% fall in the last kb for the *Arabidopsis* and yeast test sets, respectively. Second, unique GSTs were ranked in three classes depending on their location in the first (5′), middle

Table 1. Distribution of GSTs according to their location on the cognate transcript

Species	5′	Center	3′	Total
A.thaliana	354	160	1274	1788
	19.9%	8.9%	71.2%	100%
S.cerevisiae	356	304	1133	1793
	19.9%	16.9%	63.2%	100%

For each gene, the virtual transcript is deduced from genome annotation and divided in three equal parts (5′, center, and 3′). The relative position of the GST in either of these parts is then computed with the the GST size range of 150–300 bp.

(centre), or last (3′) third of the corresponding transcript. The distribution of both the *Arabidopsis* and yeast test sets indicates that at least two-thirds of the GSTs were located in the 3′ part of the transcript (Table 1).

4.2 Experimental validation of unique GST design

The quality of primer pair design was confirmed experimentally. GSTs were designed with SPADS on a set of 29 787 annotated genes in the *Arabidopsis* genome. GSTs in the 150–500 bp range, with a specificity threshold of 70% identity, and BAC-specific primer pairs had been found for 21 120 of the predicted genes. The PCR amplification was performed with purified BAC clone DNA as template or with genomic DNA, when PCR with BAC clone had failed. Based on DNA agarose gel size analysis of all PCR products, 20 338 (96.5%) unique GSTs were successfully amplified. As a further control, 1039 randomly chosen amplicons were sequenced and all had the expected sequence. Out of the 21 120 primer pairs, only 26 (0.12%) led to unspecific multiple bands or an unexpected size of the PCR product (Crowe *et al*., 2003). To evaluate the primer specificity towards the full genome, a random subset of 380 primer pairs was used for PCR amplification with full genomic DNA as template. In total, 339 (89.2%) yielded a single PCR product of the expected size, whereas 31 (8.1%) yielded multiple PCR products and 10 (2.6%) no amplification (B.Chalhoub, personal communication).

5 DISCUSSION

To study the expression of every single gene in an organism, gene-specific probes are required. SPADS allows such probes to be obtained by designing primers to amplify unique GSTs with genomic DNA as template, the main selection criterion being the amplicon specificity. With SPADS, this specificity is tested at several levels. First, a candidate GST is searched for inside a window of genome-wide gene specificity, the BLASTn-divergent region. The primer pairs themselves are tested for specificity towards the PCR template to further reduce the risk of amplifying unwanted PCR products. This important step also allows SPADS to design specific primer pairs for quantitative reverse transcription (RT)-PCR experiments. Finally a cross-hybridization boundary is set, by checking the specificity towards the whole genome of the GST proper. To reach these separate goals, SPADS allows the user to build two sets of sequences, each in a proper database, the ‘reference’ set that defines and controls the GST hybridization specificity and the ‘template’ set that is needed to produce these GSTs through PCR amplification in the most specific way, depending on practical template availability.

In the genomes of *Arabidopsis* (Arabidopsis Genome Initiative, 2000) and yeast (Llorente *et al*., 1999), in which many multigenic families with a large number of members are found, SPADS succeeded in finding a GST with genome-wide specificity (i.e. a probe with a similarity with the closest paralog in the genome lower than 70%) in 92% of the test sets for both organisms, when the minimal length for the GST is set at 100 bp. As expected, increasing the amplicon size decreases the success rate. Nevertheless, for the *Arabidopsis* genome, SPADS ends up with a unique GST for 65% of the genes when the minimal GST length is set at 300 bp. This result is significantly higher for the yeast genome with 85% for probes of at least 300 bp. SPADS fails to design a unique GST for 7.9–34.6% of the genes (of classes I and II in Fig. 3), depending on the GST length range and the organism. Such a failure can be expected when genes in a given family or subfamily are too highly conserved for SPADS to find a divergent region and design a gene-specific GST, such as the ribosomal protein-encoding genes. The approach is also expected to fail when genes are split into small exons, especially with the largest GSTs.

The SPADS process ends up with most unique GSTs falling into the 3′ extremity of the genes, followed by the 5′, and finally the central part of the genes (Table 1). This result is foreseen because the central region is usually the more conserved among homologous genes. On the contrary, the terminal regions contain the more divergent sequences and the privileged search of gene-specific GSTs in the 3′ extremity appears to be very efficient.

The suitability of SPADS for designing unique GSTs has been demonstrated for the *Arabidopsis* genome. Of the 1394 primer pairs tested, 96% produced the expected unique amplicon. The occurrence of multiple or incorrectly sized bands was below 1%. Kurth *et al*. (2002) have obtained quite similar results with a set of 1898 GSTs designed with GST-PRIME, but without constraint on amplicon specificity (Varotto *et al*., 2001). SPADS can be useful for other experiments involving the selection of specific primers or probes, such as the design of specific probes for only a few target genes, namely a gene family or genes suspected to be involved in a common physiological function. To study a gene family, the reference database used in Step 1 can be reduced to the set of sequences of the relevant paralogous genes, either genomic DNA or cDNAs and ESTs. Because the whole genomic sequence is

not always required, the reduction of the reference database to the similar sequences saves computing time. SPADS can also be exploited for its capacity to design specific primer pairs to perform specific PCR reactions from genomic DNA or from mRNA samples (in the case of RT-PCR, for instance). For such purposes, a 'phase' option is offered to constrain SPADS to take into consideration the coding frame. With this option, the 3′-most nucleotide of the primers corresponds to the first base of a codon. Because the first base is always more conserved in evolution than the two others, PCR reactions can be designed using DNA from a neighbor species as template (B.Chalhoub, personal communication).

SPADS aims at producing gene-specific amplicons, not only to build genome-wide microarrays, but also for smaller gene sets and smaller scale experiments, a frequent task for many biologists for which not easy tool is available. The fact that SPADS is concerned by specificity all along the process of probe design can also be exploited to select unique and long oligonucleotides (such as 50–80 bp) for DNA arrays. In this case, risks of hairpin conformation should be tested with single-stranded DNA secondary structure prediction tools, such as Mfold (Zucker *et al*., 1999).

The power of the systems biology approach in genomics, aiming at understanding not only the function of individual genes but also their interactions, requires reliable expression data to be collected for as many genes as possible, ideally all of them. To achieve this goal we developed SPADS. This tool has been used for the generation of a large GST collection (Crowe *et al*., 2003) and is currently used for the European Complete Arabidopsis Transcriptome Micro-Array consortium (CATMA; http://www.catma.org) that aims at designing and producing unique GSTs covering most *Arabidopsis* genes for the comprehensive analysis of their expression, as anticipated in the CAGE project (http://www.psb.ugent.be/CAGE/). These gene-specific GSTs are also used for a project of systematic RNAi knockout in *Arabidopsis* (http://www.agrikola.org/).

ACKNOWLEDGEMENTS

The authors thank Sébastien Reboux for his help in implementing the CGI web interface, Boulos Chalhoub for the conceptual idea of the 'phase' parameter of SPADS and for performing the 380 PCRs using *Arabidopsis* genomic DNA as template, and Martine De Cock, Ian Small, and the referees for helpful comments and critical reading of the manuscript. The French GENOPLANTE™ program supported this work. GENOPLANTE™ SPADS code has been deposited at the French Agency for Program Protection under number IDDN.FR.001.350016.000.D.P.2002.000.10000.

REFERENCES

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.

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

Aubourg,S. and Rouzé,P. (2001) Genome annotation. *Plant Physiol. Biochem.*, **39**, 181–193.

Breslauer,K.J., Frank,R., Blocker,H. and Markey,L.A. (1986) Predicting DNA duplex stability from the base sequence. *Proc. Natl Acad. Sci. USA*, **83**, 3746–3750.

Crowe,M.L., Serizet,C., Thareau,V., Aubourg,S., Rouzé,P., Hilson,P., Beynon,J., Weisbeek,P., van Hummelen,P., Reymond,P. *et al*. (2003) CATMA—A complete *Arabidopsis* GST database. *Nucleic Acids Res*., **31**, 156–158.

Girke,T., Todd,J., Ruuska,S., White,J., Benning,C. and Ohlrogge,J. (2000) Microarray analysis of developing *Arabidopsis* seeds. *Plant Physiol*., **124**, 1570–1581.

Halgren,R.G., Fielden,M.R., Fong,C.J. and Zacharewski,T.R. (2001) Assessment of clone identity and sequence fidelity for 1189 IMAGE cDNA clones. *Nucleic Acids Res*., **29**, 582–588.

Hughes,T.R., Mao,M., Jones,A.R., Burchard,J., Marton,M.J., Shannon,K.W., Lefkowitz,S.M., Ziman,M., Schelter,J.M., Meyer,M.R. *et al*. (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol*., **19**, 342–347.

Kurth,J., Varotto,C., Pesaresi,P., Biehl,A., Richly,E., Salamini,F. and Leister,D. (2002) Gene-sequence-tag expression analyses of 1800 genes related to chloroplast functions. *Planta*, **215**, 101–109.

Li,F. and Stormo,G.D. (2001) Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics*, **17**, 1067–1076.

Llorente,B., Fairhead,C. and Dujon,B. (1999) Genetic redundancy and gene fusion in the genome of the baker's yeast *Saccharomyces cerevisiae*: functional characterization of a three-member gene family involved in the thiamine biosynthetic pathway. *Mol. Microbiol*., **32**, 1140–1152.

Lockhart,D.J., Dong,H., Byrne,M.C., Follettie,M.T., Gallo,M.V., Chee M.S., Mittmann,M., Wang,C., Kobayashi,M., Horton,H. and Brown,E.L. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol*., **14**, 1675–1680.

Mathé, C., Sagot,M.F., Schiex,T. and Rouzé, P. (2002) Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res*., **30**, 4103–4117.

Miller,N.A., Gong,Q., Bryan,R., Ruvolo,M., Turner,L.A. and LaBrie,S.T. (2002) Cross-hybridization of closely related genes on high-density macroarrays. *Biotechniques*, **32**, 620–625.

Mitsuhashi,M., Cooper,A., Ogura,M., Shinagawa,T., Yano,K. and Hosokawa,T. (1994) Oligonucleotide probe design—a new approach. *Nature*, **24**, 759–761.

Nielsen,H.B. and Knudsen,S. (2002) Avoiding cross hybridization by choosing nonredundant targets on cDNA arrays. *Bioinformatics*, **18**, 321–322.

Podowski,R.M. and Sonnhammer,E.L. (2001) MEDUSA: large scale automatic selection and visual assessment of PCR primer pairs. *Bioinformatics*, **17**, 656–657.

Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol*., **132**, 365–386.

Varotto,C., Richly,E., Salamini,F. and Leister,D. (2001) GST-PRIME: a genome-wide primer design software for the generation of gene sequence tags. *Nucleic Acids Res*., **29**, 4373–4377.

Semple,C. and Wolfe,K.H. (1999) Gene duplication and gene conversion in the *Caenorhabditis elegans* genome. *J. Mol. Evol*., **48**, 555–564.

Wolfe,K.H. and Shields,D.C. (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, **387**, 708–713.

Xu,D., Li,G., Wu,L., Zhou,J. and Xu,Y. (2002) PRIMERGENS: robust and efficient design of gene-specific probes for microarray analysis. *Bioinformatics*, **18**, 1432–1437.

Zucker,M., Mathews,D.H. and Turner,D.H. (1999) Algorithms and thermodynamics for RNA secondary structure prediction. In Barciszewski,J., Clark,B.F.C. and Clark (eds), *A Practical Guide in RNA Biochemistry and Biotechnology*. Kluwer Academic NATO ASI Series edition, Dordrecht.