Nucleic Acids Research, 2001, Vol. 29, No. 21 4373–4377

GST-PRIME: a genome-wide primer design software for the generation of gene sequence tags

Claudio Varotto, Erik Richly[1], Francesco Salamini[1] and Dario Leister[1,*]

Zentrum zur Identifikation von Genfunktionen durch Insertionsmutagenese bei *Arabidopsis thaliana* (ZIGIA) and [1]Abteilung für Pflanzenzüchtung und Ertragsphysiologie, Max-Planck-Institut für Züchtungsforschung, Carl-von-Linné Weg 10, 50829 Köln, Germany

Received July 12, 2001; Revised and Accepted September 4, 2001

ABSTRACT

The availability of sequenced genomes has generated a need for experimental approaches that allow the simultaneous analysis of large, or even complete, sets of genes. To facilitate such analyses, we have developed *GST-PRIME*, a software package for retrieving and assembling gene sequences, even from complex genomes, using the NCBI public database, and then designing sets of primer pairs for use in gene amplification. Primers were designed by the program for the direct amplification of gene sequence tags (GSTs) from either genomic DNA or cDNA. Test runs of *GST-PRIME* on 2000 randomly selected *Arabidopsis* and *Drosophila* genes demonstrate that 93 and 88% of resulting GSTs, respectively, fulfilled imposed length criteria. *GST-PRIME* primer pairs were tested on a set of 1900 *Arabidopsis* genes coding for chloroplast-targeted proteins: 95% of the primer pairs used in PCRs with genomic DNA generated the correct amplicons. *GST-PRIME* can thus be reliably used for large-scale or specific amplification of intron-containing genes of multicellular eukaryotes.

INTRODUCTION

The complete genome sequences of many prokaryotic and eukaryotic organisms (1–6) and a draft sequence of the human genome (7) have become available recently. The collection, analysis and distribution of sequence data is now a major challenge and several public databases, such as GenBank (8), the EMBL Nucleotide Sequence Database (9) and the DNA Data Bank of Japan (DDBJ) (10) are devoted to this task.

The accumulation of extensive genomic DNA sequence information, together with the collection of expressed sequence tags (ESTs), allows high-throughput experiments on large, or even complete, collections of genes (11–14). The parallel analysis of the expression of large sets of genes is a prototype of the new, genomic, approach to biology. In such an analysis, EST clones derived from cDNA libraries are frequently used as targets for expression screening. In addition, for organisms whose genome sequences are known, genes of interest can be directly PCR amplified from genomic DNA, allowing the coverage of genes not represented in cDNA libraries and minimizing the number of sequences that need to be spotted on solid supports for screening (15). Direct gene amplification by PCR generates collections of products called gene sequence tags (GSTs). To facilitate this, there is a pressing need for freely available software that allows the retrieval and assembly of large sets of gene sequences, and enables the systematic design of appropriate primers. Several programs, such as PRIDE (16), PRIMER MASTER (17) and PRIMO (18), can be used to design primers for large-scale sequencing projects, but none of them are specialized on the automated and genome-wide design. The PRIMEARRAY software (19) enables the systematic design of primer pairs for DNA microarray construction; however, its application is restricted to cDNA collections or genes without introns, as contained in bacterial genomes, and it depends on pre-assembled collections of DNA sequences as input files.

The software package *GST-PRIME*, described here, has been designed to allow, for the first time, the detection and avoidance of introns and splicing sites during the primer design subroutine. Thus, efficient and automated design of primers even for complex genomes containing genes with introns is possible, and primer pairs designed by *GST-PRIME* can be used for the amplification from both cDNA and genomic DNA. In addition, a novel sequence retrieval and editing subroutine step has been embedded in the *GST-PRIME* program package, allowing primer design for the amplification of large gene sets, starting from a list of accession numbers of protein sequences. The software package was tested on representative sets of protein accessions from *Arabidopsis thaliana* and *Drosophila melanogaster*, two organisms whose genome sequences are known. *GST-PRIME* primer pairs (~1900) were tested in PCRs to estimate the efficiency of the software.

MATERIALS AND METHODS

Databases and analyses of sequences and their redundancy

For the large-scale retrieval of nucleotide sequences and annotations, the NCBI Batch Entrez system (8; http://www.ncbi.nlm.nih.gov/entrez/batchentrez.cgi?db=Nucleotide) was employed. Protein sequences used for the design and synthesis of the 1898 primer pairs were selected for the presence of a transit peptide motif and have been described previously (20).

*To whom correspondence should be addressed. Tel: +49 221 5062 415; Fax: +49 221 5062 413; Email: leister@mpiz-koeln.mpg.de

Programming

The main body of *GST-PRIME* was written in Visual Basic (v5.0; Microsoft). The subroutine for the sequence download was implemented in Perl (v5.005) and the Windows Executable program file was generated with Perl2Exe (v4.03; IndigoSTAR).

DNA analyses

Isolation of *Arabidopsis* DNA was performed as described previously (21). Amplifications were performed using *Taq* polymerase (Eurogentec) and the following cycling conditions: initial denaturation at 94°C for 2 min, followed by 35 cycles of denaturation at 93°C for 30 s, annealing at 55°C for 1 min and elongation at 72°C for 90 s. For re-amplification of PCR products using universal 15mer tail primers, cycling conditions were modified to the following: initial denaturation at 94°C for 15 s, followed by 30 cycles of denaturation at 93°C for 30 s, annealing at 42°C for 1 min and elongation at 72°C for 90 s. Products of exponential PCR were separated by electrophoresis on 1.2% (w/v) agarose gel. Sequencing of PCR products was performed after gel purification, using an ABI prism 377 sequencer.

Evaluation of secondary structure of oligos and primer–primer interactions

Secondary structure formation of oligos was determined using the Mfold program (program parameters: temperature, 37°C; increment, 10; window, 2) of GCG (22). Prediction of annealing events between forward and reverse primers was performed by using the default scoring matrix of the Bestfit program of GCG (22). As a measure for the annealing of forward and reverse primers, the product of the quality score obtained by Bestfit [quality score = (10 × total matches) + (–9 × total mismatches) – (gap creation penalty × gap number) – (gap extension penalty × total length of gaps)] and the relative length of the complementary region was used.

Operating environments and availability of *GST-PRIME*

GST-PRIME should run under any Windows environment and has been successfully tested under Windows NT 4.0 (service pack 6), Windows 2000 (service pack 2) and Windows 98 SE. An executable version including an installation script is available by request from the authors.

RESULTS

GST-PRIME allows the design of large numbers of primer pairs starting from a list of protein accession numbers. The program provides for: (i) retrieval of DNA sequences corresponding to the selected protein sequences (in combination with the NCBI databases Batch Entrez and GenBank); (ii) extraction and assembly of DNA sequences into gene sequences with and without introns; and (iii) design of primers that are complementary to assembled gene sequences and suitable for use in PCR and RT–PCR applications.

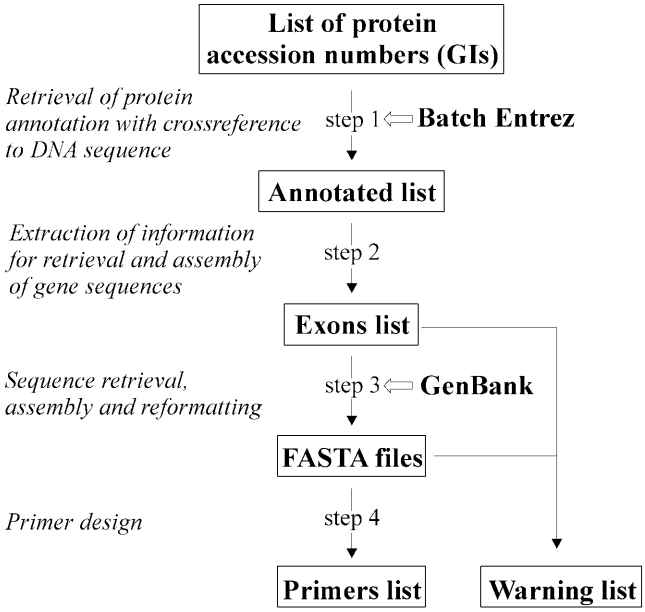
Automatic sequence retrieval and assembly

The starting point for *GST-PRIME* is a list of protein (GI) accession numbers. A text file ('annotated list') containing protein sequences and cross-references to the DNA sequences necessary for extracting the corresponding coding regions is obtained by using the Batch Entrez sequence retrieval system (Fig. 1A, step 1). From the 'annotated list' the accession numbers of the DNA sequences of interest, and the positions and boundaries of the embedded coding sequences, are extracted by *GST-PRIME* and saved as a file ('exons list'; Fig 1A, step 2). Nucleotide sequences are then retrieved from the GenBank sequence database and the 'exons list' is used for their assembly into 'gene sequences' (including introns and exons) and 'coding regions' (containing only exons). To avoid overloading of GenBank, a delay time of 2 s has been implemented between the download of two individual sequence files. The entire sequence download process can also be delayed ('delayed start option'), allowing it to run during the night or on weekends. Both 'gene sequences' and 'coding regions' are reformatted into FASTA format ('FASTA files'; Fig. 1A, step 3). The orientation of all DNA sequences can be standardized, allowing the conversion of all sequences into either forward (5′→3′) or reverse orientation (3′→5′).

Automatic primer design

GST-PRIME designs primers suitable for the amplification of both DNA and cDNA. The following default constraints were incorporated into the program: (i) annealing sites are located exclusively in exons to allow reverse transcription experiments (such as cDNA first-strand synthesis and RT–PCR); and (ii) standard primer length is 20 nt with a G+C content of 50%.

GST-PRIME employs the 'exons list' and 'FASTA files' to design the primer pairs (Fig. 1A, step 4). DNA sequences not suitable for primer design (<120 bp), or without start or stop codons, are identified and listed in a 'warning file'. To obtain amplification products with a preferential size of ~500 bp, by default forward primers are designed to anneal within the first 180 bp of the gene sequence and reverse primers between position 480 and 720. For the forward primer, search commences at position 1 and stops after the first suitable primer has been identified. For the reverse primer, all suitable primers in the window from position 480 to 720 are identified and, if more than one reverse primer is found, the primer allowing amplification of a GST sized closest to 500 bp is selected. For sequences <720 bp, forward (reverse) primers are designed within the first (last) 50% of the sequence. Forward and reverse primers localized on putative exon–intron borders are rejected. If no primer can be identified based on the criteria listed above, the search is repeated for 21mer primers with nine G+C residues or 22mer primers with eight G+C. The final criterion for reverse primer design is an annealing site that will allow, in combination with the forward primer, the amplification of a product of preferentially 500 bp at cDNA level. For both primers the annealing sites in genomic DNA are determined by using the 'gene sequence' files to predict the expected genomic amplicon length. *GST-PRIME* primers are designated according to the protein GI accession number, with a suffix indicating the annealing site relative to the gene sequence. The final output file ('primer list') contains the primer sequences and the calculated lengths of amplicons at both DNA and cDNA level (Fig. 1B). The primer design subroutine of *GST-PRIME* can be applied independently from the large-scale sequence retrieval and editing step.



B

Input:
GI: 4262227
GI: 6598621

Output:

G+C	Sense	Sequence	Anti-sense	Sequence	DNA	cDNA	DNA-Accession	Oligo for RT
10	4262227-1s	ATGAAACTTACCTCCCACCG	4262227-452as	TCCTCCAAATGCCTTCTCTG	548	452	AC006200	Anti-sense
10	6598621-24s	TATCCAACTGTTCTTCGGCC	6598621-544as	CCTCAAGTTGCAGATGGAAG	643	521	AC006837	Sense

Figure 1. Flow diagram for *GST-PRIME* subroutines. (**A**) The left column indicates the tasks performed during the design of a large set of primer pairs starting from a list of protein accession numbers. The right column explains the files and databases employed. (**B**) An example for input and output files of *GST-PRIME*. The input file is a list of protein accession numbers (GIs). In the output file, the following information about the primer pairs designed is given (from left to right): G+C, content of G+C; Sense/Anti-sense, designation of the s (sense)/as (anti-sense) primers indicating the site of annealing relative to the template (e.g. 4262227-1s refers to sense strand position 1 as start for the forward primer, while 4262227-452as indicates that position 452 of the anti-sense strand is the start of the reverse primer); Sequence, sequences of forward and reverse primers; DNA/cDNA, predicted lengths of amplicons at the DNA/cDNA level; DNA-Accession, the accession number of the DNA from which the gene sequence was retrieved; Oligo for RT, the primer to be used for reverse transcription experiments: forward (sense) or reverse (anti-sense).

Primer design for *Arabidopsis* and *Drosophila* genes

Two lists of 2000 randomly selected protein accession numbers, corresponding to *Arabidopsis* and *Drosophila* genes, were generated and used as input files for *GST-PRIME*. DNA sequence retrieval was performed using the 'delayed start option'. For the 2000 *Arabidopsis* and for 1997 *Drosophila* proteins, corresponding DNA sequences were retrieved (Table 1) and fed into the primer design subroutine of *GST-PRIME*. For *Arabidopsis*, 1868 suitable primer pairs (94.4%) were designed, and 1756 (87.8%) were designed for *Drosophila*. Failure of primer design was mainly due to inability to meet the length criteria for predicted amplification products. In the case of *Arabidopsis*, 3.4% of predicted products were too short (<150 bp at cDNA level); in *Drosophila*, 5.7% were too long (>2050 bp at genomic DNA level) (Table 1).

Frequency distributions of the predicted cDNA amplicons were similar for the two organisms: more than 70% of amplifications were predicted to amplify cDNA fragments between 451 and 550 bp long, with the average predicted length for amplified cDNAs being equal in both species (473 bp; Fig. 2).

Table 1. Comparison of the efficiency of primer design for *Arabidopsis* and *Drosophila*

	Arabidopsis	*Drosophila*
Number of protein accessions	2000 (100%)	2000 (100%)
Number of DNA sequences downloaded	2000 (100%)	1997 (99.9%)
Failure of primer pair design	27 (1.4%)	81 (4.1%)
Predicted amplicon <150 bp cDNA	67 (3.4%)	47 (2.4%)
Predicted amplicon >2050 bp DNA	38 (1.9%)	113 (5.7%)
Primer pairs suitable for GST generation	1868 (93.4%)	1756 (87.8%)

Distributions of predicted genomic amplicon lengths differed significantly in the two species. The most prominent amplicon class for genomic DNA again had a predicted length between 451 and 550 bp. However, the average predicted genomic amplicon length in *Drosophila* was significantly smaller than in *Arabidopsis* (632 versus 739 bp).

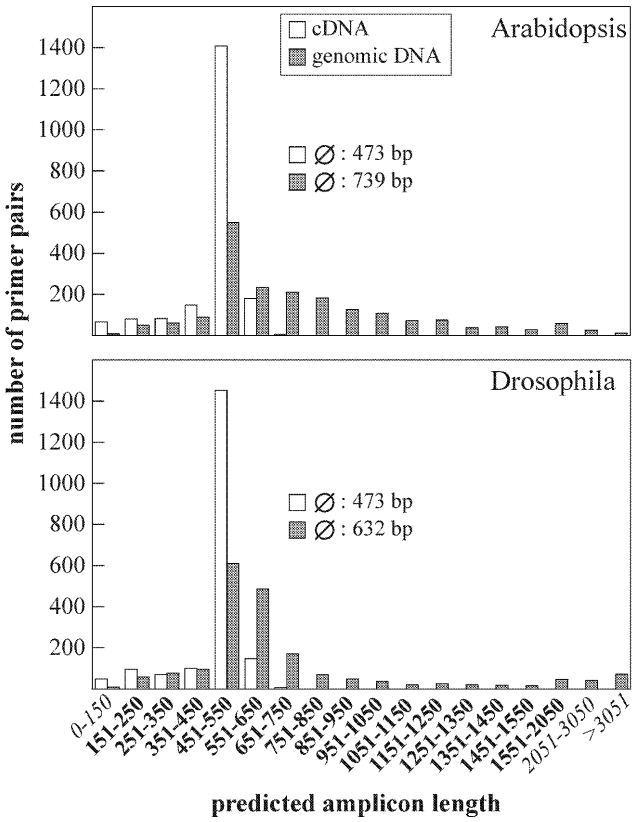


Figure 2. Distribution of predicted lengths of amplicons for *Arabidopsis* and *Drosophila*. Lengths suitable for PCR-based amplification of GSTs are indicated in bold, whereas lengths not considered for GST generation are indicated in italics.

GST-PRIME primer testing by PCR amplification of 1900 GSTs

The identification of 2047 GenBank entries for *Arabidopsis* proteins related to chloroplast functions has been described previously (20). Sequence retrieval or primer design by *GST-PRIME* failed for 149 of the 2047 sequences (7.5%). The remaining 1898 primer pairs were synthesized as 35mers, consisting of a gene-specific 20mer sequence provided by *GST-PRIME* and a universal 5′-located 15mer tail sequence suitable for re-amplification with universal tail primers. PCRs with 1804 (95%) of the primer pairs from genomic DNA resulted in the amplification of products that could be identified by agarose gel electrophoresis. The approximate size of 200 PCR products was determined in gels and correlated with the size predicted by *GST-PRIME*: for 190 PCR products the predicted and actual sizes were identical, while 10 PCR products (5%) were larger than predicted. Re-evaluation of the corresponding protein and DNA sequences revealed that the latter were exclusively derived from cDNA sequencing projects and therefore lacked any cross-reference to the corresponding genomic sequence. Sequence data for these cDNAs compared to those for the PCR products demonstrated that size differences were due to the presence of introns. Sequencing of 50 randomly chosen PCR products confirmed that target gene and PCR product were identical.

Next, we tested whether the failure of 94 of the 1898 PCR reactions in *Arabidopsis* was correlated to primer–primer interactions or to self-complementarity of oligos (such as stem–loops). When these 94 primer pairs and a control group of randomly selected primers, which were successfully employed in amplifications, were compared for their tendency to develop secondary structures using Mfold of the GCG package (22), no significant difference between the two groups was predicted. However, the group of 'failed' primers showed a slight increase in its tendency for primer–primer interactions as predicted by the program Bestfit of GCG. Taken together, for the vast majority of the 94 primer pairs, failure to result in PCR products was not predicted to be due to self-complementarity or primer–primer interactions.

DISCUSSION

A pre-requisite for the large-scale generation of GSTs from DNA or cDNA is the establishment of a standard procedure based on: (i) selection of a set of genes/proteins of interest; (ii) retrieval of the corresponding gene sequences, including genomic sequence and (predicted) transcribed regions; and (iii) design of primers that can be used for the amplification of genes from DNA or cDNA.

Several programs are currently available to design primers for large-scale sequencing projects (16–18) or primer pairs for DNA microarray construction (19), but they cannot be applied to the automated and genome-wide primer design starting from protein sequence collections. Furthermore, none of these programs allows the design of primer pairs suitable for the amplification of GSTs from genomic DNA containing intron sequences. The *GST-PRIME* package includes, for the first time, an automated sequence retrieval and assembly subroutine and a subroutine detecting and avoiding introns and splicing sites, thus allowing the systematic design of primer pairs for the amplification of GSTs, even from intron-containing gene collections, either from DNA or cDNA.

The systematic classification of genes according to the biological function of their products is based on the analysis of protein sequences rather than DNA sequences. For this reason we have established a routine that permits the automatic retrieval of DNA sequences that encode the specified protein sequences. The download and sequence assembly subroutine of the *GST-PRIME* program has been tested on two eukaryotic model organisms. Both for *Arabidopsis* and *Drosophila*, sequence retrieval and assembly were performed successfully. The sequence retrieval step of *GST-PRIME* can be carried out overnight using the 'delay start' function of the program, which relieves the pressure on public databases and improves the performance of the download process.

The primer design subroutine was successful in 87.8 and 93.4% of cases for *Drosophila* and *Arabidopsis*, respectively, providing predicted GSTs with coding regions >150 bp and a maximal genomic length of 2050 bp. Surprisingly, predicted genomic amplicons were significantly smaller in *Drosophila* compared to *Arabidopsis*, although fruit fly genes are in general larger than *Arabidopsis* genes (5,6). This discrepancy can be attributed to the high frequency of introns of 59–63 bp in *Drosophila* genes (5); the finding of a prominent class of genomic amplicons with a size between 551 and 650 bp (Fig. 2) may reflect the presence of one or two of such relatively short introns.

The suitability of *GST-PRIME* primers for generating amplicons from genomic DNA was demonstrated in *Arabidopsis*. A total of 1900 primer pairs were tested, and 95% of reactions resulted in PCR products. That these were derived from the appropriate genomic sequences was verified by size and sequence analysis. The sizes of a few PCR products were not correctly predicted by *GST-PRIME*. However, this could be attributed to sequence annotations that lacked cross-references to the genomic DNA including intron sequences. Of the 5% of reactions that failed to result in PCR products, only a small fraction was predicted to be due to the primer–primer interactions or self-complementarity of oligomers. This fraction might be decreased by implementing additional selection steps in the primer design subroutine of *GST-PRIME*. However, we decided against this option, since more elaborate design parameters would inevitably lead to a significant decrease in the number of GSTs fulfilling our length criteria. Furthermore, variation of PCR conditions, with respect to annealing temperature or buffer composition, was sufficient to obtain PCR products for ~50% of the 94 reactions that failed initially (data not shown), supporting the strategy chosen for primer design. Future versions of *GST-PRIME* may therefore include a modified output file that highlights GSTs with high G+C contents or stable secondary structures interfering with efficient PCR-based amplification.

Future modifications of *GST-PRIME* will depend on the particular scope for its application and the choice of species for GST generation. Sequencing and gene annotation revealed the presence of many gene families in the genome of *A.thalania* (6). Future upgrades of *GST-PRIME* may therefore include a subroutine to identify gene regions specific for individual members of gene families, allowing the design of gene-specific primers for the amplification of GSTs with minimal cross-hybridization. Extension of the *GST-PRIME* program to other species depends on the state of genome research in each organism. For organisms with sequence databases based in part on cDNA sequences, the output of *GST-PRIME* may be modified to flag cDNA-derived primer pairs, allowing to interpret the presence of GSTs with a larger size than predicted. For yeast, generation of amplicons of all 6000 genes has been accomplished successfully (23). This was made easier by the low frequency of intron-containing genes in this species, which allowed the use of a relatively unsophisticated primer design software. *Homo sapiens* represent a more challenging task for systematic primer design. The average human gene has a coding sequence of 1.3 kb, and includes, on average, 3.4 kb of intron sequences (7). In the case of generating human GSTs from genomic DNA, predicted coding region lengths significantly smaller than the 500 bp used for *Arabidopsis* and *Drosophila* in this study will have to be used to maximize coverage. *GST-PRIME* should nevertheless be suitable for primer design even for the human genome, but the preferential cDNA amplicon size constraint of the program will have to be adapted to the architecture of the human genome.

ACKNOWLEDGEMENTS

We thank Paul Hardy for his critical comments on the manuscript. Grateful acknowledgements are extended to Eurogentec, Koen Dekker and Heinz Saedler. This work was supported by the award of a Habilitation Stipend of the Deutsche Forschungsgemeinschaft to D.L.

REFERENCES

1. Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M. *et al.* (1996) Life with 6000 genes. *Science*, **274**, 563–567.
2. Kaneko,T., Sato,S., Kotani,H., Tanaka,A., Asamizu,E., Nakamura,Y., Miyajima,N., Hirosawa,M., Sugiura,M., Sasamoto,S. *et al.* (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.*, **3**, 109–136.
3. Blattner,F.R., Plunkett,G.,III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.
4. The *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, **282**, 2012–2018.
5. Adams,M.D., Celniker,S.E., Holt,R.A., Evans,C.A., Gocayne,J.D., Amanatides,P.G., Scherer,S.E., Li,P.W., Hoskins,R.A., Galle,R.F. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.
6. The *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
7. The International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
8. Wheeler,D.L., Church,D.M., Lash,A.E., Leipe,D.D., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Tatusova,T.A., Wagner,L. and Rapp,B.A. (2001) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **29**, 11–16.
9. Stoesser,G., Baker,W., van Den Broek,A., Camon,E., Garcia-Pastor,M., Kanz,C., Kulikova,T., Lombard,V., Lopez,R., Parkinson,H. *et al.* (2001) The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **29**, 17–21.
10. Tateno,Y., Miyazaki,S., Ota,M., Sugawara,H. and Gojobori,T. (2000) DNA data bank of Japan (DDBJ) in collaboration with mass sequencing teams. *Nucleic Acids Res.*, **28**, 24–26.
11. Brown,P.O. and Botstein,D. (1999) Exploring the new world of the genome with DNA microarrays. *Nature Genet.*, **21**, 33–37.
12. Duggan,D.J., Bittner,M., Chen,Y., Meltzer,P. and Trent,J.M. (1999) Expression profiling using cDNA microarrays. *Nature Genet.*, **21**, 10–14.
13. Eisen,M.B. and Brown,P.O. (1999) DNA arrays for analysis of gene expression. *Methods Enzymol.*, **303**, 179–205.
14. Schaffer,R., Landgraf,J., Perez-Amador,M. and Wisman,E. (2000) Monitoring genome-wide expression in plants. *Curr. Opin. Biotechnol.*, **11**, 162–167.
15. Penn,S.G., Rank,D.R., Hanzel,D.K. and Barker,D.L. (2000) Mining the human genome using microarrays of open reading frames. *Nature Genet.*, **26**, 315–318.
16. Haas,S., Vingron,M., Poustka,A. and Wiemann,S. (1998) Primer design for large scale sequencing. *Nucleic Acids Res.*, **26**, 3006–3012.
17. Proutski,V. and Holmes,E.C. (1996) Primer Master: a new program for the design and analysis of PCR primers. *Comput. Appl. Biosci.*, **12**, 253–255.
18. Li,P., Kupfer,K.C., Davies,C.J., Burbee,D., Evans,G.A. and Garner,H.R. (1997) PRIMO: a primer design program that applies base quality statistics for automated large-scale DNA sequencing. *Genomics*, **40**, 476–485.
19. Raddatz,G., Dehio,M., Meyer,F.T. and Dehio,C. (2001) PrimeArray: genome-scale primer design for DNA-microarray construction. *Bioinformatics*, **17**, 98–99.
20. Abdallah,F., Salamini,F. and Leister,D. (2000) A prediction of the size and evolutionary origin of the proteome of chloroplasts of *Arabidopsis*. *Trends Plant Sci.*, **5**, 141–142.
21. Liu,Y.G., Mitsukawa,N., Oosumi,T. and Whittier,R.F. (1995) Efficient isolation and mapping of *Arabidopsis thaliana* T-DNA insert junctions by thermal asymmetric interlaced PCR. *Plant J.*, **8**, 457–463.
22. Devereux,J., Haeberli,P. and Smithies,O. (1984) A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.*, **12**, 387–395.
23. DeRisi,J.L., Iyer,V.R. and Brown,P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.