BIOINFORMATICS

Vol. 18 no. 11 2002
Pages 1432–1437

PRIMEGENS: robust and efficient design of gene-specific probes for microarray analysis

Dong Xu[1,], Guangshan Li[2], Liyou Wu[2], Jizhong Zhou[2] and Ying Xu[1]*

[1]Protein Informatics Group, Life Sciences Division and [2]Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

Received on March 6, 2002; revised on April 29, 2002; accepted on May 2, 2002

ABSTRACT

Motivation: DNA microarray is a powerful high-throughput tool for studying gene function and regulatory networks. Due to the problem of potential cross hybridization, using full-length genes for microarray construction is not appropriate in some situations. A bioinformatic tool, PRIMEGENS, has recently been developed for the automatic design of PCR primers using DNA fragments that are specific to individual open reading frames (ORFs).

Result: PRIMEGENS first carries out a BLAST search for each target ORF against all other ORFs of the genome to quickly identify possible homologous sequences. Then it performs optimal sequence alignment between the target ORF and each of its homologous ORFs using dynamic programming. PRIMEGENS uses the sequence alignments to select gene- specific fragments, and then feeds the fragments to the Primer3 program to design primer pairs for PCR amplification. PRIMEGENS can be run from the command line on Unix/Linux platforms as a stand-alone package or it can be used from a Web interface. The program runs efficiently, and it takes a few seconds per sequence on a typical workstation. PCR primers specific to individual ORFs from *Shewanella oneidensis* MR-1 and *Deinococcus radiodurans* R1 have been designed. The PCR amplification results indicate that this method is very efficient and reliable for designing specific probes for microarray analysis.

Availability: The software is available at http://compbio.ornl.gov/structure/primegens/.

Contact: xud@ornl.gov

INTRODUCTION

Various genome-scale sequencing projects have generated vast amounts of sequence data. The next important step is to determine the function of each gene and gene regulatory networks. One of the most powerful tools for investigating gene functions and regulatory networks is DNA microarrays (DeRisi *et al.*, 1997; Duggan *et al.*, 1999; Moch *et al.*, 2001), which measure gene expression levels through DNA–DNA hybridization. Due to homologous sequences in a genome, the same probe on microarrays may cross-hybridize with different homologous genes if there is a certain degree of sequence identity between them. Such cross-hybridization is problematic for microarray data interpretation. To avoid this problem, researchers typically do not use full-length genes as targets but rather use gene-specific fragments on a microarray, i.e. a fragment of DNA sequence of a gene that does not have high sequence identity to any other sequence in the genome sequence pool. For this purpose, it is necessary to carry out two computational tasks: (1) to identify a fragment specific to an open reading frame (ORF), and (2) to design forward and reverse primers based on the selected gene-specific fragment to allow ORF-specific amplification by polymerase chain reaction (PCR). The amplified DNA fragments can then be used as probes specific to individual genes on microarrays.

Several studies (Kane *et al.*, 2000), as well as several computer programs such as Primer Master (Proutski and Holmes, 1996), *PrimeArray* (Raddatz *et al.*, 2001), and *GST-PRIME* (Varotto *et al.*, 2001), have been published that are related to these computational tasks. Of the related studies and programs available, none are capable of automatically performing the above two computational tasks at the genome scale. Therefore, researchers presently must use multiple tools (sequence alignment tools and primer design tools) in a semi-manual manner to identify gene-specific fragments. Outputs of sequence alignments must be manually checked, and inputs for primer design of each gene must be entered manually as well. This process is very time-consuming and the results may not be reliable.

This study focuses on developing programs for designing ORF-specific probes for microarray analysis in an automatic and high throughput fashion. To accomplish the task of target sequence selection, a computer

*To whom correspondence should be addressed.

program called PRIMEGENS (PRIMEr design using GENe Specific fragments) has been developed. PRIMEGENS automatically designs PCR primers to amplify target DNA fragments specific to individual ORFs. The program is available to academic users free of charge (see http://compbio.ornl.gov/structure/primegens/). In this paper, the algorithm used in PRIMEGENS and an application of the program using ORFs from *Shewanella onidensis* MR-1 and *Deinococcus radiodurans* R1 are presented.

MATERIALS AND METHODS

Overview of PRIMEGENS

Figure 1 summarizes how PRIMEGENS works for the two computational tasks discussed above, i.e. to identify a gene-specific fragment and to design the primers for the target fragment. PRIMEGENS first carries out the heuristic BLAST (Altschul *et al.*, 1997) search[†] for each ORF (query) against all other ORFs to quickly identify possible homologous sequences. Then it performs optimal alignment between the query and each of its homologous ORF sequences using the dynamic programming technique. Based on the alignment, PRIMEGENS selects the gene-specific fragments. For those ORFS whose sequences are specific themselves, the entire ORF sequence will be used as the gene-specific fragment. For the second task, PRIMEGENS uses a third-party software, Primer3 (Rozen and Skaletsky, 2000), which takes a DNA fragment selected in the first task and designs PCR primer pairs for PCR amplification based on user- specified parameters for the primers (e.g. primer size, melting temperature, GC content, and self complementarity). To further ensure that the primer will not amplify multiple sequences, gapless sequence alignments are carried out between the two primers and all the ORFs. If two primers amplify multiple sequences, an alternative gene-specific fragment will be used to design new primers.

Mathematical Formulation

The problem of gene-specific fragment identification can be formulated as follows. Assume we have a collection of N sequences in the genome, denoted as S, which consists of $s^1, s^2, \ldots, s^N$. For sequence $s^m (1 \leq m \leq N)$ with n^m nucleotide bases, the goal is to find the longest fragment of nucleotide bases between positions a and b, i.e. $s^m_{a,b} (1 \leq a < b \leq N)$, such that the following three conditions are satisfied: (1) the maximum sequence identity between $s^m_{a,b}$ and $s^k (k \neq m)$ is lower than a user-specified value $\text{sim}_{\max}$; (2) the fragment has at least $l_{\min}$ bases, i.e. $b - a > l_{\min}$; and (3) if s^m and $s^k (k \neq m)$ have a local sequence alignment with a minimum length of f_{mix} and an expectation value of $E_{\max}$ or less (by BLAST), the aligned portion of s^m should not overlap with the $s^m_{a,b}$ fragment. The purpose of the third condition is to avoid the potential problem that a continuous stretch of sequences (such as more than 25 nt) may cause cross-hybridization even if the entire region is less than the defined value, e.g. 75%. After the fragment is identified, two primers are designed ($p5^m$ for the 5′ end and $p3^m$ for the 3′ end) according to user-specified parameters. The two primers should not amplify any $s^k (k \neq m)$, i.e. if the complement of $p5^m$ has a high sequence identity with s^k at position t, then the reverse complement of $p3^m$ cannot have high sequence identity with s^k at any position after t.

[†] We plan to test another heuristic search algorithm PatternHunter (Ma *et al.*, 2002), which is faster and more sensitive than BLAST, for this purpose.

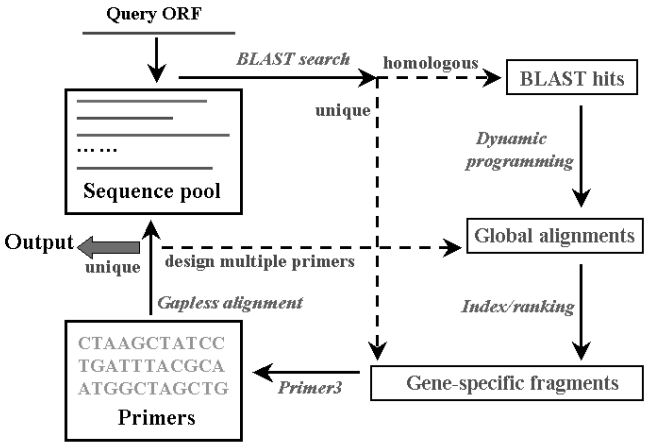


Fig. 1. A flowchart to show how PRIMEGENS works. A solid arrow indicates that the process is required, and a dashed arrow indicates that the process is performed under a particular condition.

Algorithm

The algorithm first uses the fast local alignment by BLAST to search each ORF against the database of all the ORFs in the genome. This procedure identifies a list of ORFs that are potentially similar to the target ORF. Then an alignment is performed using dynamic programming (Smith and Waterman, 1981) between the target ORF and each ORF in the selected list. Dynamic programming is slower than BLAST but it guarantees to find an optimal solution for the alignment. The strategy of combining both alignment methods allows a quick comparison of the target ORF to all the ORFs in the genome, and the accurate identification of gene-specific fragments based on optimal alignments between the target ORF and a limited number of selected ORFs. The following three steps are carried out to solve the problem formulated above:

- Carry out a local alignment for each $s^m (1 \leq m \leq N)$ against the whole sequence database S using BLAST to find all the local alignments with an expectation

value of $E_{\max}$ or lower. If there is no region aligned with other genes, the entire sequence (s^m) of the ORF will be used for designing ORF-specific primers and included in the sequence set of the gene-specific fragments. Otherwise the sequence is included in the sequence set Q. If the local alignment has a length longer than f_{mix}, the aligned portion of s^m is recorded to ensure that the gene-specific fragment of s^m will not overlap with this region.

- Let q^m be a sequence with n^m nucleotide bases in Q that has local alignments (other than the self alignment) with w^m sequences in Q with an expectation value of $E_{\max}$ or lower. We denote these aligned sequences $p_\nu^m (1 \leq \nu \leq w^m)$. An alignment is carried out between q^m and each p_ν^m using dynamic programming. For a fragment $q_{i,j}^m (1 \leq i \leq n^m - l_{\min} - 1, i + l_{\min} - l \leq j \leq n^m)$ in q^m, if the three conditions described in *Mathematical Formulation* are satisfied for the optimal alignment between q^m and every p_ν^m, q_{ij}^m is recorded as a possible gene-specific fragment. Then all the possible gene-specific fragments are ranked according to length. If no q_{ij}^m is found to satisfy the three conditions, the user may adjust the specified parameters (e.g. to increase $\text{sim}_{\max}$) to redo the calculation. PRIMEGENS indicates the minimum $\text{sim}_{\max}$ needed for each ORF whose q_{ij}^m was not found. If the two sequences are too similar so that it is impossible to find a gene-specific fragment, the user may remove one of them and use the other one as the representative to redesign primers based the gene fragment specific to these two sequences. In this way, even though the fragment identified is not specific to the two ORFs, no cross-hybridization is expected with the rest of the ORFs in the genome.

- If the identified gene fragments are shorter than the maximum fragment size of the user-specific value such as 1 kb, the longest fragment among all the gene-specific fragments will be used to design $p5^m$ and $p3^m$ according to user-specified parameters. Given that the fragment used in the primer design is gene specific, it is not likely that the designed primers can amplify another ORF sequence. On the other hand, since the primers are much shorter (typically around 20 nt) than the fragment, such a possibility still exists. To be certain that the two primers only amplify s^m, $p5^m$ is searched against all $s^k (k \neq m)$ using gapless alignments. If the complement $p5^m$ aligns with ORF s^k at position t with a high sequence identity, then the reverse complement of $p3^m$ is used to check whether it has high sequence identify with s^k at any position after t. If it has, the second longest fragment among all the gene-specific fragments is chosen to repeat the process until $p5^m$ and $p3^m$ do not amplify another sequence in the database.

PRIMEGENS Implementation

The computer program is written in C programming language. The program also links to the executables of the *primer3* program for automated primer design. It has been tested on various Unix/Linux platforms, including Sun, DEC, and Linux PC. PRIMEGENS can run as a stand-alone package or it can be used from a Web server (http://compbio.ornl.gov/structure/primegens/). For the Web server, the user needs to specify an email address. When the calculation is finished, the user will receive an email showing an URL where the result is saved. On average, it takes a few seconds for each sequence to find a pair of primers. The program reads a file of all sequences in the FASTA format. For the stand-alone package, a user has more options than with the Web server to specify parameters either at the command line or in a parameter file. The default values for the parameters are

$$\text{sim}_{\max} = 75\%;\ l_{\min} = 100 \text{ nt};\ E_{\max} = 10^{-15};\ f_{mix} = 50 \text{ nt}.$$

The output results can be viewed from a plain text editor or from a Microsoft Excel spread sheet.

RESULTS

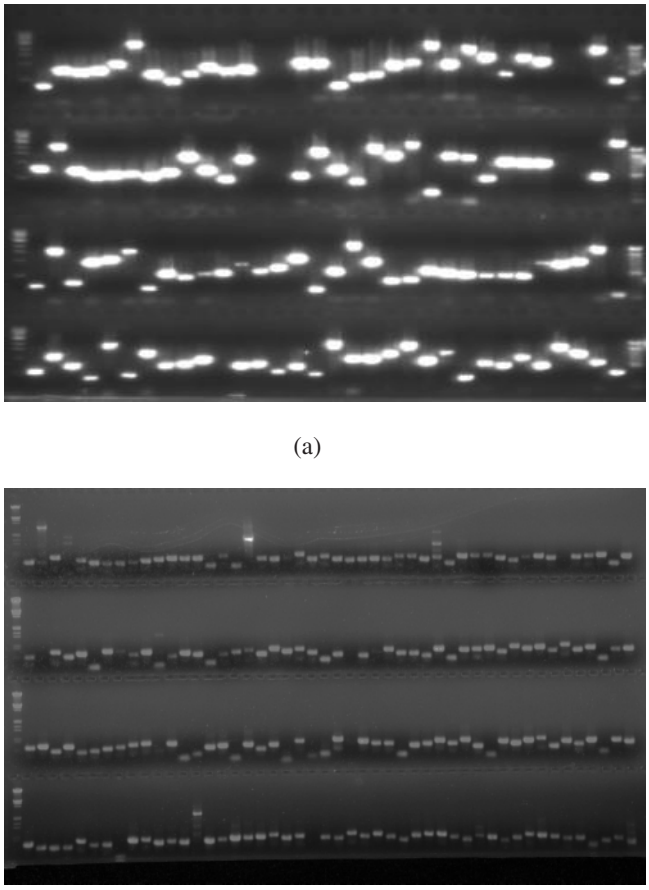
To experimentally evaluate the performance of the PRIMEGENS program, PCR amplification primers are designed for constructing whole genome microarrays for several different bacteria. Comparing the target gene with all other genes in a genome first identifies a DNA fragment specific to each ORF. The initial cutoff as an ORF-specific fragment is set to a sequence identity of less than 75% as demonstrated by several other studies (DeRisi *et al.*, 1997; Wu *et al.*, 2001). The PCR primers are designed based on the unique fragments using the PCR primer design program 'Primer 3'. The parameters for designing optimal forward and reverse primers to generate PCR products specific to each of the selected ORFs are generally set as follows: (1) each primer contains 18–25 oligonucleotides. (2) to simplify the PCR amplifications, the primers melting temperatures are set within the range of 62–68°C. (3) The maximum amplified PCR product size for microarray fabrication is generally set at 1000 nt and the minimum is set at 200 nt. The selected primers should also meet the optimal criteria defined in the Primer 3 program (GC content: 40–75%, max 3′ self end: 4, max poly-X: 5, and salt concentration: 50%). The primer designing process is repeated for the homologous genes with the sequence identity of >75% by setting the identity cutoff at 85%. If no optimal primers can be obtained for individual homologous genes, a pair of PCR primers is designed based on a DNA fragment specific to the group

of homologous genes. Finally, as described above, the designed primers are compared against the whole genome sequence and any primer pairs with multiple priming sites are eliminated and redesigned to minimize potential non-specific PCR amplification. All of the above process is performed in an automatic, high-throughput fashion. On a typical Unix/Linux workstation, less than five hours are needed to process 5000-gene genomes.

This program was first used to design PCR amplification primers for the metal-reducing bacterium *Shewanella oneidensis* MR-1 (Thompson *et al.*, 2002; Beliaev *et al.*, 2002). *S. oneidensis* is a ubiquitous, gram-negative bacterium that has been isolated from a variety of environments. It is capable of using a variety of compounds (oxygen, iron, manganese, uranium, nitrate, nitrite, fumarate, thiosulfate, dimethyl sulfoxide, trimethylamine N-oxide and elemental sulfur) as electron acceptors. Due to the great environmental importance of this strain, the MR-1 genome was sequenced. The current annotation of the 4.9 Mb *S. oneidensis* MR-1 genome predicted approximately 4700 protein-encoding genes. The average size of the open reading frame was about 770 bp, and the GC content of this organism was about 46%. In total, specific primers for 95.6% of the genes were obtained. Primers were designed for 4704 ORFs, which covers 95.6% of the whole genome of 4921 ORFs (annotated as of August, 2001). A portion of these PCR primers were synthesized at the PAN Facility of Stanford University. The whole genome primers were synthesized by MWG Biotech, Inc (High Point, NC).

The experimental conditions for PCR amplification were as follows. Each reaction contained 100 μL of: 1 $\times$ PCR buffer [Promega], 20 mM $MgCl_2$, 0.2 mM dNTPs, 20 ng genomic DNA, 5 units of Taq DNA polymerase [Promega], and 0.5 μM of each primer. Thermal-cycling conditions included: initial denaturation at 95°C for 2 min; then 30 cycles of 94°C for 30 s, 56°C for 30 s, and 72°C for 1 min; and a final incubation at 72°C for 5 min.

Typically, 90% of the primers yielded specific PCR products (Figure 2a), with 85–95% amplification success. Altogether, 93.9% of the designed primers yielded good PCR products. In total, single, specific and strong amplification products were obtained for 4020 ORFs (85.5%), while 21 ORFs (0.4%) gave more than two bands of PCR products, 39 ORFs (0.8%) yielded single faint bands, 18 ORFs (0.4%) gave multiple bands, and 394 ORFs (8.4%) gave one strong band as well as one faint band. No amplifications were obtained for 200 ORFs (4.3%). Approximately 99% of the 4020 ORFs that were single band and high yield PCR product showed the expected ORF size. The primers which did not generate good specific PCR products were resynthesized and/or or redesigned. The majority of PCR amplification failures appeared to be due to the primer synthesis process because



(a)

(b)

Fig. 2. Representative gel images of amplified PCR products created with the primers designed by PRIMEGENS. The PCR amplification conditions were described in the text. (a) PCR amplification with the primers from *S. oneidensis* MR-1. DNA marker at the left side: 1 Kb DNA marker from Invitrogen (Carlsbad, CA); DNA maker on the right side: 100 bp DNA marker from Invitrogen (Carlsbad, CA). (b) PCR amplification with the primers from *D. radiodurans* R-1. The DNA maker on the left side: Lambda DNA digested with HindIII obtained from Promega (Madison, WI).

95% of the resynthesized primers yielded good results. Partial genome microarrays containing genes involved in energy metabolism have been constructed and used to analyze gene expression profiles under different growth conditions for both wild type and mutant cells (*etrA*$^-$ and *fur*$^-$). The results indicated that the changes of expression profiles of many genes in wild type and mutant cells are consistent with those from other studies (Thompson *et al.*, 2002; Beliaev *et al.*, 2002). This implies that the program PRIMEGENS is able to select specific DNA probes for microarrays.

Genome GC content could affect the performance of the

primers designed by PRIMEGENS. To evaluate whether PRIMEGENS works well for genomes with high GC content, genome-wide PCR amplification primers have been designed for the radiation-resistant bacterium, *D. radiodurans* R1. Its genome size is about 3.2 Mb with 3186 predicted ORFs. The average size of the open reading frame is about 940 bp, and the GC content is about 68%. Primers were obtained for 3045 ORFs but no primers could be designed for the rest of the ORFs because either no appropriate primer location was found (72 ORFs), or the sequence identity of the ORFs was similar to other ORFs by more than 75% (69 ORFs). The primers were synthesized by MWG Biotech, Inc (High Point, NC). A touchdown PCR program from 60°C to 55°C with 95°C denaturing temperature, 2 min extension and 35 cycles total was performed to amplify DNA fragments in a reaction buffer (50 mM KCl, 0.1% [v/v] Triton X-100, 2.5 mM $MgCl_2$, 10 mM Tris-HCl, pH 9.0) containing 0.2 mM dNTPs, 1% (w/v) bovine serum albumin (BSA), 1% (v/v) dimethyl sulfoxide (DMSO), and 1.32 M Betaine. For every ORF, first a 50 μl PCR reaction was carried out with 10 ng of the genomic DNA, followed by four 100 μl of re-amplification PCRs using 0.2 μl of the initial PCR reaction as templates.

Single, specific and strong amplification products were obtained for 2682 ORFs (89.3%). While 25 ORFs (0.8%) gave more than two bands of PCR products, 217 (7.2%) ORFs yielded single faint bands. Seventy-nine ORFs (2.6%) gave one strong band as well as one faint band. No amplifications were obtained for 42 ORFs (1.4%). Figure 2b showed a representative gel image. Microarray analysis indicated that the gene expression profiles of many known genes after high dose radiation were consistent with previous results (Liu et al. manuscript in preparation). These results suggested that the probes designed by PRIMEGENS work well for monitoring gene expression differences under different conditions.

In addition, since the web server has been up and running, a few dozen requests have been received. Some users of the server have commented that they were satisfied with the output results of this program. They reported that PRIMEGENS saved them much time and the primers designed worked out very well. Some biologists with little computer skill used the Web server without any problem.

SUMMARY

PRIMEGENS provides an easy-to-use software for biologists to select gene-specific fragments and to design PCR primers. Biologists with little computer skills can easily use the web server to design the primers. The PCR amplification results of *S. oneidiensis* MR-1 and *D. radiodurans* R1 indicate that the PRIMEGENS program works successfully and efficiently. The microarray hybridization results from *S. oneidiensis* MR-1 and *D. radiodurans* R1 also support the fact that PRIMEGENS is suitable for high throughput primer design. It is expected that it will work equally well for other bacterial genomes, provided that the sequence for each open reading frame is available. Such applications will benefit other high-throughput primer designs, such as oligo arrays and comparative genomic hybridization arrays. Given the massive applications of microarray technologies available for microbial genome research in the post-genome era, PRIMEGENS will play an important role as a powerful tool.

ACKNOWLEDGMENTS

We would like to thank our colleagues Manesh Shah, Serguei Passovets, Yongqing Liu, Joel A. Klappenbach, Debbie Arnett, and Julia Stair for their help and insightful comments. We also thank Christal Yost for a critical reading of this manuscript. This study was supported by the Microbial Genome Program, the Microbial Cell Project, and the Natural and Accelerated Bioremediation Research Program (NABIR), Biological and Environmental Research (BER) and the U.S. Department of Energy. Oak Ridge National Laboratory is managed by UT-Battelle, LLC for DOE under contract # DE-AC05-96OR22464.

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.

Beliaev,A.S., Thompson,D.K., Giometti,C.S., Li,G.S., Yates III,J., Nealson,K.H., Tiedje,J.M., Heidelberg,J.F. and Zhou,J.Z. (2002) Gene and protein expression profiles of *Shewanella oneidensis* during anaerobic growth with different electron acceptors. *Omics: a Journal of Integrative Biology*, **6**, 39–60.

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.

Duggan,D.J., Bittner,M., Chen,Y., Meltzer,P. and Trent,J.M. (1999) Expression profiling using cDNA microarrays. *Nature Genet.*, **21**, 10–14.

Kane,M.D., Jatkoe,T.A., Stumpf,C.R., Lu,J., Thomas,J.D. and Madore,S.J. (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res.*, **28**, 4552–4557.

Ma,B., Li,M. and Tromp,J. (2002) PatternHunter: faster and more sensitive homology search. *Bioinformatics*, **18**, 440–445.

Moch,H., Kononen,T., Kalllioniemi,O.P. and Sauter,G. (2001) Tissue microarrays: what will they bring to molecular and anatomic pathology? *Adv. Anat. Pathol.*, **8**, 14–20.

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.

Raddatz,G., Dehio,M., Meyer,F.T. and Dehio,C. (2001) PrimeArray: genome-scale primer design for DNA-microarray construction. *Bioinformatics*, **17**, 98–99.

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

Smith,T.F. and Waterman,M.S. (1981) Comparison of biosequences. *Adv. Appl. Math.*, **2**, 482–489.

Thompson,D.K., Beliaev,A.S., Giometti,C.S., Tollaksen,S.L., Khare,T., Lies,D.P., Nealson,K.H., Lim,H., Yates,III,J., Brandt,C.C., Tiedje,J.M. and Zhou,J.Z. (2002) Transcriptional and Proteomic Analysis of a Ferric Uptake Regulator (Fur) Mutant of *Shewanella oneidensis*: Possible Involvement of Fur in Energy Metabolism, Transcriptional Regulation, and Oxidative Stress. *Appl. Environ. Microbiol.*, **68**, 881–892.

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.

Wu,L.Y., Thompson,D.K., Li,G.S., Hurt,R.A., Tiedje,J.M. and Zhou,J.Z. (2001) Development and evaluation of functional gene arrays for detection of selected genes in the environment. *Appl. Environ. Microbiol.*, **67**, 5780–5790.