Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W665–W669***
doi:10.1093/nar/gkl297

PrimerStation: a highly specific multiplex genomic PCR primer design server for the human genome

Tomoyuki Yamada[1,*], Haruhiko Soma[1,2] and Shinichi Morishita[1]

[1]Department of Computational Biology, Graduate School of Frontier Sciences, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8562, Japan and [2]Life Science Laboratory, Materials Laboratories, Sony Corporation, Shinagawa, Tokyo 144-0001, Japan

Received February 14, 2006; Revised March 1, 2006; Accepted April 7, 2006

ABSTRACT

PrimerStation (http://ps.cb.k.u-tokyo.ac.jp) is a web service that calculates primer sets guaranteeing high specificity against the entire human genome. To achieve high accuracy, we used the hybridization ratio of primers in liquid solution. Calculating the status of sequence hybridization in terms of the stringent hybridization ratio is computationally costly, and no web service checks the entire human genome and returns a highly specific primer set calculated using a precise physicochemical model. To shorten the response time, we precomputed candidates for specific primers using a massively parallel computer with 100 CPUs (SunFire 15 K) about 3 months in advance. This enables PrimerStation to search and output qualified primers interactively. PrimerStation can select highly specific primers suitable for multiplex PCR by seeking a wider temperature range that minimizes the possibility of cross-reaction. It also allows users to add heuristic rules to the primer design, e.g. the exclusion of single nucleotide polymorphisms (SNPs) in primers, the avoidance of poly(A) and CA-repeats in the PCR products, and the elimination of defective primers using the secondary structure prediction. We performed several tests to verify the PCR amplification of randomly selected primers for ChrX, and we confirmed that the primers amplify specific PCR products perfectly.

INTRODUCTION

There are pressing needs to detect genomic polymorphisms and alterations with high accuracy. For example, recent studies have revealed that large-scale copy number polymorphisms in the human genome contribute to human genetic variation (1), and genome alterations could be a latent cause of cancer (2,3). Despite the long history of research, the selection of highly specific genomic PCR primers that do not hybridize anywhere else in the genome, except for the target, remains a challenging task. Although some software programs consider the specificity against target sequences, some of these programs (4–8) simply check the specificity in terms of sequence similarity using text-matching algorithms for the prompt selection of primers but fails to take into account real primer hybridization in a liquid solution.

Since the precise quantification of real primer hybridization is difficult to achieve, many software programs (9–18) evaluate the specificity of a primer in terms of a rough indicator, the melting temperature at which 50% of the copies of the primer hybridize the target sequence in a liquid solution. With this alternative standard, a better primer has a larger margin that separates the melting temperature for the target sequence and those for off-targets. Therefore, we should select the target-specific primer that maximizes the margin. Currently this physicochemical model is accepted as the best for designing primers.

Stringent requirements for genomic PCR primers

Care has to be taken to select an annealing temperature because a lower annealing temperature may yield unrelated PCR products of off-target genes while a higher value is likely to fail to amplify the target. Annealing temperature must be in between the two extremes in order to guarantee effective amplification of the target genomic sequence while minimizing the risk of cross-reaction with off-targets. To amplify the target genomic sequence effectively, the hybridization ratio (19) of the primer to the target should be close to 1, e.g. >0.99, while to avoid generating false-positive PCR products, the hybridization ratio to any off-target should be minimized, e.g. <0.05. Let us call the temperature for the primer 'executable' if this requirement is satisfied. We

*To whom correspondence should be addressed. Tel: +81 47 136 3985; Fax: +81 47 136 3977; Email: yamada@cb.k.u-tokyo.ac.jp

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

require that the annealing temperature should be executable. This demand may appear to be too stringent. Nevertheless, we hold that it is a prerequisite for designing genomic PCR primers with executable temperatures because the human genome is always 'expressed' and is much longer than the total of its coding regions. Hence, cross-reactions with off-targets must be eliminated effectively. Unfortunately, the design scheme based on melting temperature may not meet this prerequisite because it only ensures the weaker condition that the hybridization ratio of the primer to the target is >0.5, while that to any off-target is <0.5.

Considering the stringency of executable temperatures, one might be concerned with the feasibility of selecting primers with executable temperatures. However, we obtained a positive answer to this question. Figure 1 illustrates how the hybridization ratios (*y*-axis) of a particular primer to the target itself and four representative off-targets drop as the temperature (*x*-axis) increases. Note that when the temperature is below 60°C, the hybridization ratios approach 1.0, indicating that the primer is likely to hybridize to off-targets as well as to the target. The vertical band in the middle that ranges from 71 to 73°C displays executable temperatures because the hybridization ratio of the primer to the target is over 0.99 and the hybridization ratios to off-targets are no more than 0.05. In this respect, a better primer has a wider range of executable temperatures. Therefore, we developed a program for selecting the best primer with the widest range of executable temperatures, which allowed us to design qualified primers for human genes.

This positive result motivated us to develop a system called PrimerStation for designing multiplex genomic PCR primers because there is no such comprehensive web service for the human genome. In addition, these primers are applicable to a variety of other studies, such as designing sequence tagged site (STS) markers. Although there exist some software programs (20,21) that calculate the hybridization ratio, PrimerStation focused on the possible annealing temperature considering the hybridization ratio.

Multiplex genomic PCR primers

To design primers for multiplex PCR, we need to adjust the amplification conditions for each primer so that the PCR product sizes of individual target sequences are separated sufficiently. Most existing multiplex primer design software programs use melting temperature for this adjustment. However, the large discrepancy between the melting temperatures of designed primers makes it difficult to perform PCR amplification under the same conditions. To overcome this problem, we need to look for a set of primers for individual target sequences in the human genome that share common executable temperatures. Once such a set of primers is found, any common executable temperature suffices for PCR amplification using all the primers in the set under the same conditions. Further improvement can be achieved by selecting the optimal set of primers that maximize the range of common executable temperatures. PrimerStation attempts to output the optimal set of primers for the given target sequences.

Experimental result

Before we describe PrimerStation in detail, we present an experimental result that uses primers designed using PrimerStation. Figure 2 shows the result of 7% acrylamide gel electrophoresis for randomly selected multiplex PCR primers on human ChrX. We confirmed that the primers magnified specific PCR products perfectly. Lanes b–q are the results of the amplification. Lanes l–q are the result of multiplex genomic PCR; each band in the lane was amplified by the primers corresponding to the lane {d,e,h}, {g,j,k}, {b,c,i}, {d,e,g,h,k}, {b,c,f,i,j}, {b,c,d,e,f,g,h,i,j,k}.

Pre-calculation of qualified primer candidates

PrimerStation calculates multiplex genomic PCR primers for given query target sequences in the human genome. Since the calculation demands extremely costly computations, it cannot be executed online. To guarantee a prompt response on the web server, we preprocessed the human genome to enumerate

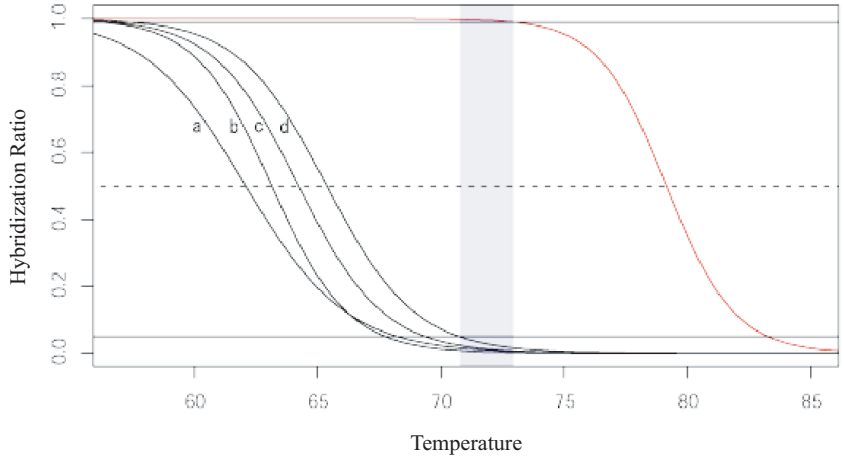

Figure 1. The relationship between the hybridization ratio and temperature. The horizontal axis denotes the temperature and the vertical axis the hybridization ratio. The red line depicts the hybridization ratio of the primer to the target sequence. The black lines show the hybridization ratio of the primer to the off-target sequences. The primer and target sequences are TCGCCAGGAAGTAACTGGGAGCAG and the off-target sequences are (**a**) CAGCTCCCAGTTACTCCCAGGCTG, (**b**) CTGCCCCCAGTTCCCTCCTGGAGG, (**c**) CTGCTCCCAGTTATTTCCTGGTGG, and (**d**) TGGGCAGGGAGGTACTGGGAGCAG. The blue region indicates the executable temperatures at which the hybridization ratio exceeds 0.99 to the target sequence and the hybridization ratios to the off-targets are no more than 0.05.

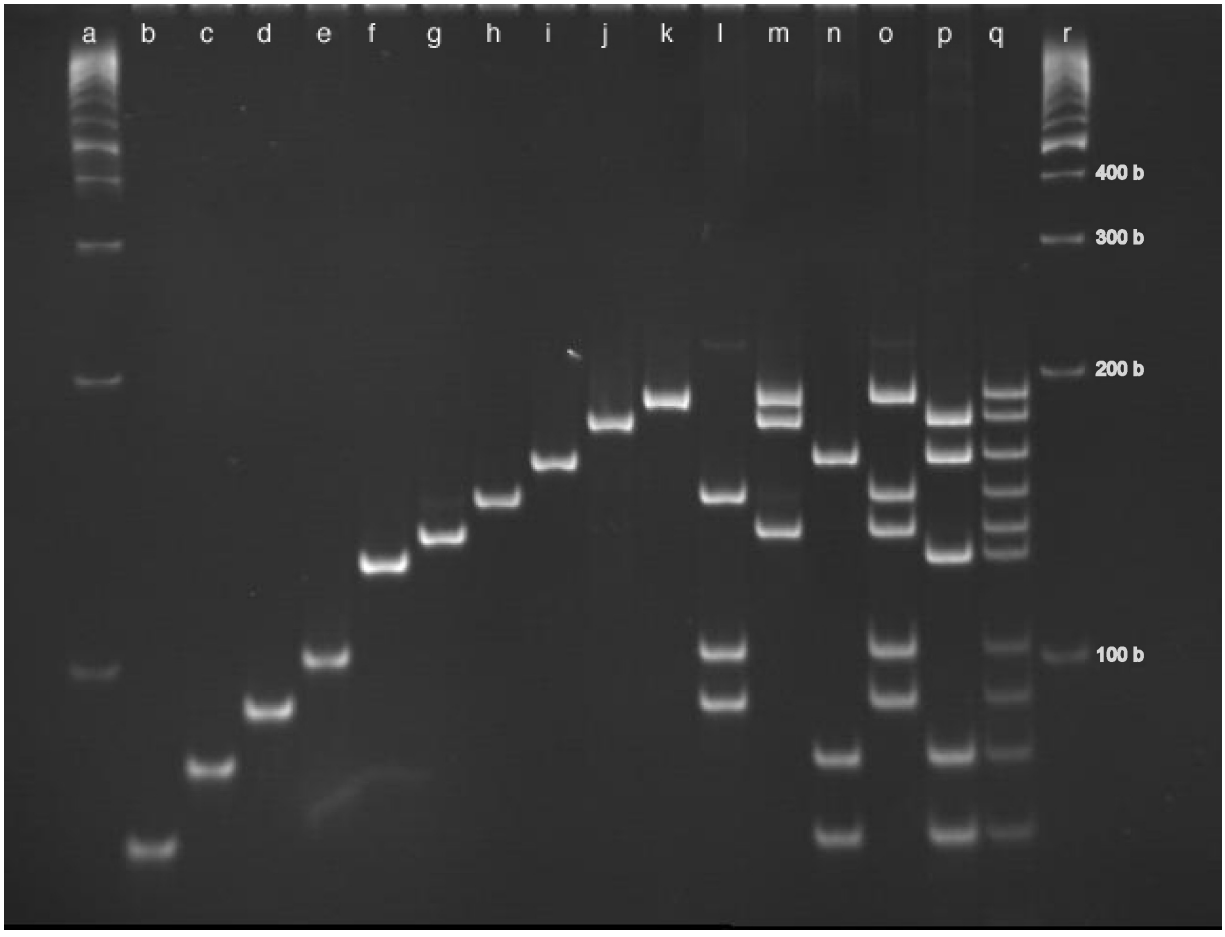

Figure 2. Multiplex genomic PCR. Lanes b–q are the electrophoresis results of multiplex genomic PCR on the human ChrX. Ten primers were used, and the multiplicity of primers for each result is (b–k) 1, (l–n) 3, (o and p) 5 and (q) 10. Primers were mixed before PCR amplification. The bands in lanes a and r are ladder markers. Observe that the designed primers amplified a single target sequence from the human genome, and that primers b–k amplified highly specific bands.

highly specific 25 base substrings in gene coding regions, such that their hybridization ratios to any off-target sequences are less than 0.05. The pre-calculation took 80 days using a massively parallel computer with 100 CPUs (SunFire 15 K) because off-target candidates were searched for as carefully as possible in the pre-calculation step. In particular, we examined off-target sequence candidates with at most four mismatches before the thermodynamic calculation using an efficient text-matching algorithm (22).

Seven percent of 25 base substrings of the gene coding regions were qualified. Although the figure 7% may appear to be small, in reality, these qualified 25 base substrings suffice to cover most of base pairs in the gene coding regions because 74% of base pairs in the gene coding regions are covered by at least one 60–600 base-long PCR product of qualified primers. The remaining 26% of the gene coding regions are difficult to amplify due to repetitive elements.

PrimerStation web service

Figure 3 shows a snapshot of PrimerStation. To specify specific positions in the human genome, PrimerStation accepts a list of RefSeq accession numbers of genes to design primers for multiplex genomic PCR. The users can also input configure options for the PCR product involving the product size range, the minimum product size differences among the set of designed primer pairs, the exclusion of primers with known single nucleotide polymorphisms (SNPs), the avoidance of $(A)_n$ and $(CA)_n$ repeats and the elimination of defective primers using the secondary structure prediction by Mfold (23). It also allows us to set PCR condition options, such as the cation concentration and primer concentration. We are able to fit these parameters to the real experiment by changing the values.

After inputting all of the parameters, the design of multiplex PCR primers is requested by pressing the submit button located below the gene ID textbox. PrimerStation attempts to output an optimal primer set for multiplex genomic PCR for the given gene set, or it reports the failure to find a set. The result includes primer sequences for individual genes, their product size, the second maximum hybridization ratio against off-targets, the melting temperatures of the forward and reverse primers, the minimum executable temperature and information about amplifying the target chromosome. The primer sequences can be downloaded as a fasta or csv formatted file.

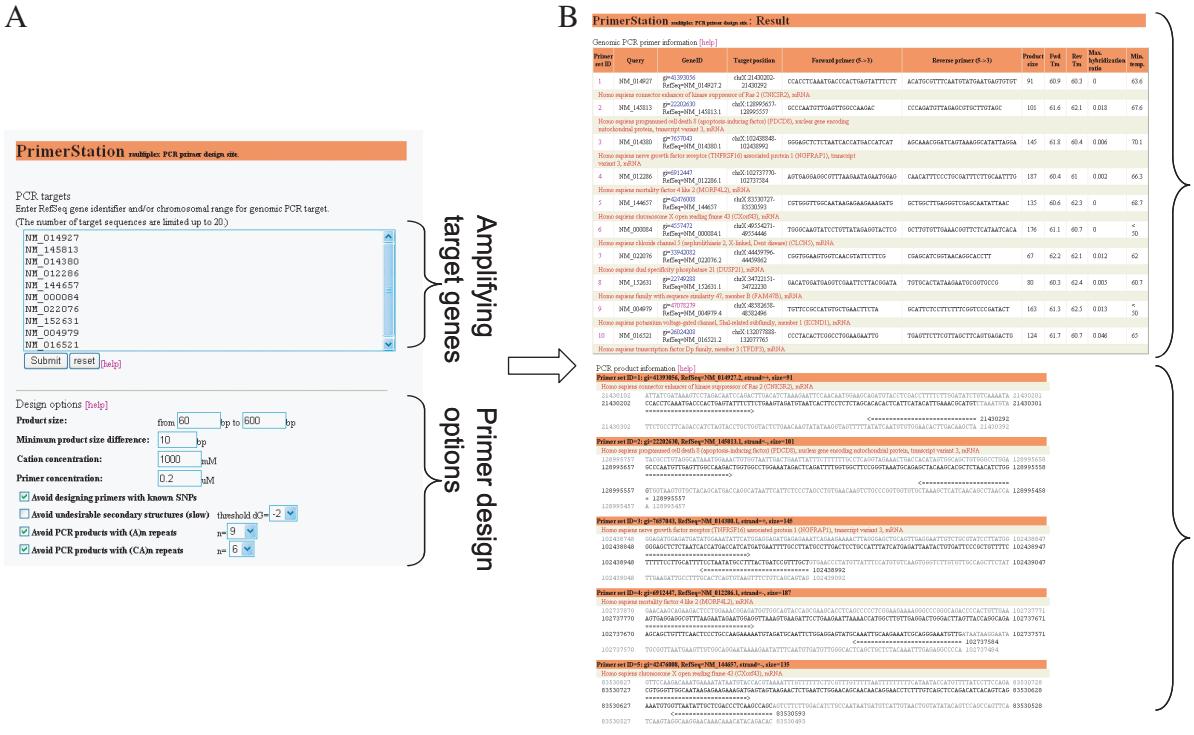

Figure 3. Flowchart of multiplex primer design using PrimerStation. (**A**) Input arbitrary gene IDs that represent target genomic sequence positions. Subsequently, provide primer conditions, such as the product size range, the minimum product size differences, and the avoidance of known SNPs, and $(A)_n$ or $(CA)_n$ repeats. The cation concentration and primer concentration are also adjustable. (**B**) The result of multiplex PCR primer design. The upper area shows primer information including GeneID, the target chromosome and position, forward and reverse primer sequences, the product size, melting temperatures of forward and reverse primers, second maximum hybridization ratio, minimum executable temperature and free energy changes for the most stable secondary structures of the primers. The lower area shows product information. The product that would be amplified by the designed primer is highlighted in black, and the surrounding sequences are colored gray.

MATERIALS AND METHODS

Hybridization ratio

Here, we describe the idea of the hybridization ratio in detail. The primer–template dissociation reaction is given by the following equation:

$$[\text{Primer}] + [\text{Template}] \leftrightarrow [\text{Primer} + \text{Template}].$$

The left side of the equation means that the primer and template sequences are dissociated. The right side of the equation indicates that the primer and template sequences are hybridized to each other. In a real solution of primer and template sequences, equilibrium of the two states is established.

To calculate the hybridization ratio, it is essential to know the dissociation constant, a value expressing the extent to which a primer–template complex dissociates in solution:

$$K = e^{-\frac{\Delta G}{RT}} = \frac{C_{\text{p\&t}}}{C_{\text{p}} \cdot C_{\text{t}}} = \frac{f}{(C_{\text{p}} - f \cdot C_{\text{t}}) \cdot (1 - f)},$$

where ΔG is the free energy change, which can be calculated using the nearest neighbor method (24); R is the molar gas constant; T is the temperature; $C_{\text{p\&t}}$, C_{p} and C_{t} are the molar concentrations of the duplex of primer and template, the single strand of primer and the single strand of template, respectively; and f denotes the hybridization ratio, which is the ratio of hybridized primer to the total amount of primer.

If C_{p} is much larger than C_{t}, the hybridization ratio f is approximated by

$$f \approx \frac{C_{\text{p}} \cdot K}{1 + C_{\text{p}} \cdot K} = \frac{C_{\text{p}} \cdot e^{-\frac{\Delta G}{RT}}}{1 + C_{\text{p}} \cdot e^{-\frac{\Delta G}{RT}}}.$$

The hybridization ratio f is determined by the primer concentration C_{p}, the free energy change ΔG and the temperature T. The temperature at which the hybridization ratio becomes the specified value can also be calculated from C_{p}, ΔG and the hybridization ratio f. The temperature at which the hybridization ratio is 50% ($f = 0.5$) is the melting temperature.

The conditions for the multiplex PCR in Figure 2

The reaction consisted of 200 mM Tris–HCl (pH 8.4), 500 mM KCl, 100 mM $MgCl_2$, 2.5 mM dNTPs, 6.0 µl ddH_2O, 10 µM concentration of each primer, 36 ng of human genomic DNA and 2 U of Bio *Taq* HS DNA polymerase (Takara Bio, Tokyo, Japan). The final reaction volume was 10 µl. The cycling conditions were 1 min at 94°C, 40 cycles of 30 s melting at 94°C, and 2 min of polymerization at 72°C, followed by a final 7 min extension at 72°C.

DISCUSSION

To evaluate the completeness of the design using PrimerStation, we attempted to process 28 516 human RefSeq genes

(November 2004, RefSeq gene set). We found that our primer sets were able to amplify 74% of base pairs in the gene coding regions. From the candidates of the primers, PrimerStation selects the best primer set from the candidates. The maximum hybridization ratio against off-targets was set to 0.05, although relaxing the threshold did not improve the coverage of genes for which primers were designed successfully because these genes have highly homologous sequences in their family genes.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors are grateful to Fumihito Miura and Takashi Ito for valuable input and helping H.S. perform biological experiments. The authors also thank anonymous referees for their valuable comments. This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas (C) to S.M., the Leading Project for Biosimulation from the Ministry of Education, Culture, Sports, Science and Technology of Japan to S.M., and a joint research project with SONY Corporation. Computational time was provided by the Super Computer System, Human Genome Center, Institute of Medical Science and the University of Tokyo. Funding to pay the Open Access publication charges for this article was provided by the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Conflict of interest statement. None declared.

REFERENCES

1. Sebat,J., Lakshmi,B., Troge,J., Alexander,J., Young,J., Lundin,P., Maner,S., Massa,H., Walker,M., Chi,M. *et al*. (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525–528.
2. Futreal,P.A., Coin,L., Marshall,M., Down,T., Hubbard,T., Wooster,R., Rahman,N. and Stratton,M.R. (2004) A census of human cancer genes. *Nature Rev. Cancer*, **4**, 177–183.
3. Weir,B., Zhao,X. and Meyerson,M. (2004) Somatic alterations in the human cancer genome. *Cancer Cell*, **6**, 433–438.
4. Li,F. and Stormo,G.D. (2001) Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics*, **17**, 1067–1076.
5. Rahmann,S. (2003) Fast large scale oligonucleotide selection using the longest common factor approach. *J. Bioinform. Comput. Biol.*, **1**, 343–361.
6. Pozhitkov,A.E. and Tautz,D. (2002) An algorithm and program for finding sequence specific oligonucleotide probes for species identification. *BMC Bioinformatics*, **3**, 9.
7. Yamada,T. and Morishita,S. (2004) Computing highly specific and noise tolerant oligomers efficiently. *J. Bioinform. Comput. Biol.*, **2**, 21–46.
8. Heikki,H., Martti,J. and Mauno,V. (2005) Genome-wide selection of unique and valid oligonucleotides. *Nucleic Acids Res.*, **33**, e115–e115.
9. Kaderali,L. and Schliep,A. (2002) Selecting signature oligonucleotides to identify organisms using DNA arrays. *Bioinformatics*, **18**, 1340–1349.
10. Xu,D., Li,G., Wu,L., Zhou,J. and Xu,Y. (2002) PRIMEGENS: robust and efficient design of gene-specific probes for microarray analysis. *Bioinformatics*, **18**, 1432–1437.
11. Haas,S.A., Hild,M., Wright,A.P.H., Hain,T., Talibi,D. and Vingron,M. (2003) Genome-scale design of PCR primers and long oligomers for DNA microarrays. *Nucleic Acids Res.*, **31**, 5576–5581.
12. Nielsen,H.B., Wernersson,R. and Knudsen,S. (2003) Design of oligonucleotides for microarrays and perspectives for design of multi-transcriptome arrays. *Nucleic Acids Res.*, **31**, 3491–3496.
13. Rouillard,J.M., Zuker,M. and Gulari,E. (2003) OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucleic Acids Res.*, **31**, 3057–3062.
14. Wang,X. and Seed,B. (2003) Selection of oligonucleotide probes for protein coding sequences. *Bioinformatics*, **19**, 796–802.
15. Reymond,N., Charles,H., Duret,L., Calevro,F., Beslon,G. and Fayard,J.M. (2004) ROSO: optimizing oligonucleotide probes for microarrays. *Bioinformatics*, **20**, 271–273.
16. Chou,H.H., Hsia,A.P., Mooney,D.L. and Schnable,P.S. (2004) PICKY: oligo microarray design for large genomes. *Bioinformatics*, **20**, 2893–2902.
17. Rouchka,E.C., Khalyfa,A. and Cooper,N.G.F. (2005) MPrime: efficient large scale multiple primer and oligonucleotide design for customized gene microarrays. *BMC Bioinformatics*, **6**, 175.
18. Nordberg,E.K. (2005) YODA: selecting signature oligonucleotides. *Bioinformatics*, **21**, 1365–1370.
19. Marky,L.A. and Breslauer,K.J. (1987) Calculating thermodynamic data for transitions of any molecularity from equilibrium melting curves. *Biopolymers*, **26**, 1601–1620.
20. Miura,F., Uematsu,C., Sakaki,Y. and Ito,T. (2005) A novel strategy to design highly specific PCR primers based on the stability and uniqueness of 3′ end subsequences. *Bioinformatics*, **21**, 4363–4370.
21. SantaLucia,J., Jr and Hicks,D. (2004) The thermodynamics of DNA Structural Motifs. *Annu. Rev. Biophys. Biomol. Struct.*, **33**, 415–440.
22. Yamada,T. and Morishita,S. (2005) Accelerated off-target search algorithm for siRNA. *Bioinformatics*, **21**, 1316–1324.
23. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
24. SantaLucia,J.,Jr (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl Acad. Sci. USA*, **95**, 1460–1465.