BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no. 2 2003 Pages 285–286

An efficient algorithm for minimal primer set selection

Ming-Hua Hsieh, Wei-Che Hsu, Sung-Kay Chiu and Chi-Meng Tzeng*

Department of Research and Development, U-Vision Biotech Inc., 3F 132, Ln 235, Pao-Chiao Rd, Hsin-Tien City 231 Taipei, Taiwan

Received on June 12, 2002; revised on July 29, 2002; accepted on August 6, 2002

ABSTRACT

Summary: We have developed U-PRIMER, a primer design program, to compute a minimal primer set (MPS) for any given set of DNA sequences. The U-PRIMER algorithm, which uses automatic variable fixing and automatic redundant constraint elimination to tackle the binary integer programming problem associated with the MPS selection problem. The program has been tested successfully with 32 adipocyte development-related genes and 9 TB-specific genes to obtain their respective MPSs.

Availability: A free copy of U-PRIMER implemented in C++ programming language is available from http://www.u-vision-biotech.com

Contact: cmtzeng@u-vision-biotech.com

Extensive diagnostic and molecular analyses are often restricted by limited availability of biological materials. Multiplex RT-PCR is a time and reagent saving amplification technique in which multiple primer sets are used to amplify several specific targets simultaneously from a single sample. In microarray experiments, mRNA molecules from cells are reverse transcribed into cDNA by random primers before being hybridized onto the microarray. However, the cDNA generated by random primers gives a significant level of cross hybridization because of the presence of homologous sequences or genes with similar sequences in the genome. One way to overcome this problem is using RT-PCR amplification of the transcripts of interest. In the case of many genes to be monitored in a microarray experiment, many primer pairs are needed for the RT-PCR reactions. If the primer for one gene can be used to prime another gene, called a universal primer, a smaller primer set is thus needed. With this idea, multiple PCR kit and microarray could be more cost effective, convenient, and accurate. Let a complete primer set (CPS) be a primer set that contains primers that can amplify all the DNA sequences in a given set S of n DNA sequences and a minimal primer set (MPS) of S is a CPS of S that has a minimal number of primers. Thus, the goal of this paper is to design an efficient algorithm for finding an MPS for any given set of DNA sequences. We called the algorithm and program U-PRIMER. The program contains two parts. The first part takes any given number of DNA sequences as input, and screens primer candidates for both ends of each sequence of any primer length, allowed number of base-pair mismatches, GC%, and T_m. It applies standard criteria for filtering out inappropriate primer candidates that contain poly N, poly AG, and potential hairpin patterns. The remaining candidates become the input of the second part which is the implementation of U-PRIMER algorithm.

Given a set S of n DNA sequences with length $l_1, l_2, \ldots, l_n$, the MPS problem is to find a CPS that comprises a forward primer set P_f and a reverse primer set P_r that satisfy certain desired constraints and have minimal number of primers. Since P_f and P_r selection are the same in the MPS problem, it is sufficient to consider only the P_f selection problem. Let m be the length of primers and $w_l(i)$ be the length of the forward primer window of sequence i. We say a primer candidate p is eligible for a DNA fragment if its complementary sequence is a substring of the DNA fragment. However, to increase the chance of finding the universal primers, we allow a maximum of 2 base-pair mismatches between the primer candidate and the DNA fragment at user's choice. We call a forward primer candidate p_f eligible for a target DNA sequence if p_f is eligible for the 5′ end fragment (forward primer window) of the DNA sequence. Let P_f be a forward primer set. Then P_f selection can be informally stated as follows.

$$\min_{P_f} \ |P_f|$$

$$\text{subject to} \ \ \text{for all } s \in S, \text{ there exists a } p_f \in P_f \text{ such that } p_f \text{ is eligible for 5' end of } s$$

where $|P|$ denotes the number of elements of a set P. To formally define this minimization problem, we define primer candidates first. For each forward primer window,

*To whom correspondence should be addressed.

we move the primer candidate one base at a time from the beginning of the window with fragment length m until we run through the whole window. For example, if only a perfect match is allowed, there will be $M(0) = \sum_{i=1}^{n}(w_l(i) - m + 1)$ such fragments, and if up to two mismatches are allowed, there will be $M(2) = \sum_{i=1}^{n}(1 + 3m + 9m(m-1)/2)(w_l(i) - m + 1)$ such fragments. Let P_f be the set of the forward primer candidates defined above and C_f be the cover matrix of P_f. The entries of C_f are defined as

$$C_f(i, j) = \begin{cases} 1, & P_f(j) \text{ is eligble for 5' end of } S(i); \\ 0, & \text{otherwise.} \end{cases}$$

Where $P_f(j)$ denotes the jth element of P_f $(1 \le j \le k_f)$ and $S(i)$ denotes the ith element of S $(1 \le i \le n)$. Set $k_f = |P_f|$ and x_j's $(1 \le j \le k_f)$ be the decision variables for P_f. In particular, $x_j = 1$ if $P_f(j)$ is selected; otherwise $x_j = 0$. Then, the above minimization problem can be formulated as an binary integer programming (BIP) problem (see Hillier and Lieberman (2001) for more details)

$$\min_{x_1,\ldots,x_{k_f}} \sum_{j=1}^{k_f} x_j, \qquad \text{(P1)}$$
$$\text{subject to} \quad C_f x \ge 1_n, \; x_j = 0 \text{ or } 1, \; 1 \le j \le k_f,$$

where $x = (x_1, x_2, \ldots, x_{k_f})^{\mathrm{T}}$ and 1_n denotes the n-vector with all elements equal to 1.

The main difficulty of computing the exact solution of (P1) is that the number of decision variables is large even when n and m are moderately small; e.g. for $n = 32$, $m = 15$, and $w_l(i) = 100$ $(i = 1, \ldots, n)$, then $M(2) = 2\,727\,232$. It is well known that a general BIP problem with more than 1000 decision variables is computationally intractable unless a special algorithm is designed to exploit the structure of the BIP problem. To solve this problem, we adopt automatic variable fixing and automatic redundant constraint elimination techniques (see, Crowder *et al.* (1983)). We summarize the basic idea of the U-PRIMER algorithm as follows.

(1) for $k = 1$ to k_f do
for $j \ne k$ do
if $c_j \ge c_k$ (c_j and c_k denote the jth and kth column of C_f) set x_k to 0

(2) for $k = 1$ to k_f do
if $\sum_{j=1}^{k_f} C_f(i, j) = 1$ and $C_f(k, j) = 1$
set x_k to 1

(3) for all the x_k that is fixed to 1 at Step 2, remove constraint i if $C_{ik} = 1$. If there is any variable that can be fixed in Step 1 or 2 go to 1

(4) if all variables have been fixed and then the algorithm will be terminated; otherwise, use a regular BIP solver (such as a branch- and bound-solver) for the remaining variables.

Step 1 and 2 are automatic variable fixing. x_k can be fixed to zero if we can find a $j \ne k$, such that $c_j \ge c_k$. Since if $\{\cdots, k, \cdots\}$ is an MPS, then $\{\cdots, j, \cdots\}$ will be an MPS, too. On the other hand, $x_k = 1$ if $\sum_{j=1}^{k_f} C_f(i, j) = 1$ and $C_f(k, j) = 1$, since any set without k can not be a CPS and thus not an MPS. Step 3 is an automatic redundant constraint elimination.

In our Type II diabetes (NIDDM) research, we used U-PRIMER program to design an MPS for 32 genes involved in adipocyte differentiation. Also, in another application, a minimal primer set is needed for the PCR amplification of nine novel markers from the genome of Mycobacterium tuberculosis (or TB), generated by U-GET (U-Vision Biotech Inc.), a computer program that can find unique sequences in the genome of any species. With a choice of primer-size 15 and one allowable base-pair mismatch, 6 pairs of primers from the 5′ and 3′ ends for the 9 U-GET verified sequences were found and evaluated. The combination of U-PRIMER program to generate gene-specific primers and the selection of unique sequences in a genome by U-GET program makes the multiplex PCR reaction specific and accurate in clinical diagnosis.

There are several previous studies for the MPS problem: Pearson *et al.* (1996) and Talaat *et al.* (2000). They considered primer lengths from 5 or 12. By probability, the chance of matching of a random 8mer is 61 times in a 4 Mb genome. Thus the primer set of short primers is not specific enough in real biological applications. In real world applications, short primers cannot generate unique PCR products because they can cross-prime to many other sites (homologues). We increase the size up to 17 base-pairs, which can enhance the specificity for the priming reaction. We also allow one to two mismatches between the primers and the DNA sequence, for increasing the chance to find a second priming site without sacrificing the specificity.

ACKNOWLEDGEMENTS

This work was partly supported by a fund from SBIR Grant, MOEA, Taiwan (No. 12900033).

REFERENCES

Crowder,H., Johnson,E. and Padberg,M. (1983) Solving large-scale zero-one linear programming problems. *Oper. Res.*, **31**, 803–834.

Hillier,F.S. and Lieberman,G.J. (2001) *Introduction to Operations Research*, Seven edition, McGraw-Hill, New York.

Pearson,W., Robins,G., Wrege,D. and Zhang,T. (1996) On the primer selection problem for polymerase chain reaction experiments. *Disc. Appl. Math.*, **71**, 231–246.

Talaat,A., Hunter,P. and Johnston,S. (2000) Genome-directed primers for selective labeling of bacterial transcripts for DNA microarray analysis. *Nat. Biotechnol.*, **18**, 679–682.