ELSEVIER

Expert Systems with Applications 30 (2006) 129–136

Expert Systems with Applications

www.elsevier.com/locate/eswa

An intelligent primer design system for multiplex reverse transcription polymerase chain reaction and complementary DNA microarray

Ming-Hua Hsieh *, Ray Tsaih [1], Chiung-Yu Huang [2]

Department of Management Information Systems, National Chengchi University, Taiwan, Republic of China

Abstract

Various biological and clinical applications require the quantification of the messenger RNA (mRNA) abundance of the biological materials under investigation. Reverse transcription-polymerase chain reaction (RT-PCR) is a popular technique for mRNA quantification. However, in these applications, multiple quantifications of mRNA are required from the same sample. Multiplex RT-PCR uses multiple primer pairs to amplify the specific targets simultaneously from a single sample. Therefore, multiplex RT-PCR is suitable for such applications. Complementary DNA (cDNA) microarray is a robust and versatile tool that allows researchers to examine the expression of hundreds or thousands of genes for the analysis of gene function, disease diagnosis, and drug discovery. Multiplex RT-PCR is usually performed for the cDNA preparation during the cDNA microarray experiments. Primer design is the most crucial step in multiplex RT-PCR. Therefore, the goal of this study was to implement an intelligent primer design system. The core of the system was based on mathematical techniques. The system can guide users to design cost effective primer collection. The system had also been successfully applied to two real world examples with biological and clinical significance.

Keywords: Intelligent system; Mathematical model; Primer design; Polymerase chain reaction; Microarray

1. Introduction

Traditional assays for measuring the messenger RNA (mRNA) abundance, such as Northern blot analysis (Lodish, Berk, Zipursky, Matsudaira, Baltimore and Darnell, 2000) and nuclease protection assay (Lodish et al., 2000) demand a large amount of mRNA. These assays are not good choices for low-abundance transcripts or limited tissues. Reverse transcription-polymerase chain reaction (RT-PCR) is a more sensitive technique (Alcamo, 1999) for mRNA quantification. RT-PCR consists of two steps: synthesis of complementary DNA (cDNA) from mRNA by reverse transcription and amplification of the specific cDNA by polymerase chain reaction (PCR). The PCR step amplifies the amount of cDNA. Therefore, RT-PCR avoids the need for large amounts of mRNA and provides sensitive detection in small tissue samples (Siebert & Larrick, 1992). The cDNA serves as a template in the PCR. The key components of the PCR include the cDNA and a pair of oligonucleotide primers specific for the cDNA of interest. The cDNA is amplified exponentially via PCR cycles of denaturation, annealing, and extension. Since the amplification is exponential, loading differences are amplified as well. Therefore, primer selection and careful optimization of the PCR conditions are essential for achieving reliable amplification.

Various biological and clinical applications require the quantification of the mRNA abundance of the biological materials under investigation. However, the amount of biological materials is usually limited. Further, quantifications of multiple mRNA are required from the same sample. Multiplex RT-PCR (Downing, Khandekar, Shurtleff, Head, Parham and Webber, 1995) uses multiple primer pairs to amplify several specific targets simultaneously from a single sample. Therefore, multiplex RT-PCR is suitable for such applications.

cDNA microarray is a robust and versatile tool that allows researchers to examine the expression of hundreds or thousands of genes for the analysis of gene function, disease diagnosis, and drug discovery (Chen, Dougherty, & Bittner, 1997; Eisen, & Brown, 1999; Eisen, Spellman, Brown, & Botstein, 1998; Schena, Shalon, Davis, & Brown, 1995; Shalon, Smith, & Brown, 1996). In cDNA microarray experiments, DNA molecules representing individual genes are spotted onto a

* Corresponding author. Tel.: +886 2 29393091 81236; fax: +886 229393754.

E-mail addresses: mhsieh@mis.nccu.edu.tw (M.-H. Hsieh), tsaih@mis.nccu.edu.tw (R. Tsaih), hcy@mis.nccu.edu.tw (C.-Y. Huang).

[1] Tel.: +886 2 29393091 81036; fax: +886 2 29393754.

[2] Tel.: +886 2 29393091 81242; fax: +886 2 29393754.

0957-4174/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2005.09.055

glass surface. Usually, mRNA molecules from cells are reversely transcribed into cDNA by random primers before hybridized onto the microarray. However, the cDNA generated by random primers gives a significant cross hybridization because of the presence of homologous sequences in the genome (Talaat, Hunter, & Johnston, 2000). One way to overcome this problem is using multiplex RT-PCR to amplify transcripts of interest.

From the discussions above, it is clear that primer selection plays a crucial role for both multiplex RT-PCR and cDNA microarray experiments. It is well known (Antao, Lai, & Tinoco, 1991; Antao & Tinoco, 1992) that the selected oligonucleotide primers have to satisfy essential criteria such as T_m (temperature of melting), GC content, and secondary structures in order to optimize RT-PCR. For multiplex RT-PCR and cDNA microarray experiments, there are two other important factors to be considered: the cost of the primers and the probability of cross hybridization among primers. The cost of a commercial set of oligonucleotide primers is determined by the number of distinct primers, k, and the units of each type of primers, q_i with $1 \leq i \leq k$. Let P be the (fixed) cost for ordering a primer type and C the cost for a unit of primer. Then, the cost equals $Pk + C\sum_{i=1}^{k} q_i$, in which the value of Pk dominates the cost. Thus, reducing the value of k can lower the cost. In addition, the probability of cross hybridization among primers is positively correlated to k. Therefore, reducing the value of k can lower both the cost and the probability of cross hybridization among primers. In case of n genes (or DNA sequences) to be monitored in the multiplex RT-PCR or in the cDNA microarray experiment, then n primer pairs are needed. However, if we can find the so-called multi-gene primers that can be used for priming multiple genes, then the required types of primers can be reduced. The adoption of multi-gene primer can reduce the value of k.

The goal of this study was to implement an intelligent primer design system that utilizes the idea of multi-gene primer and thus reducing the value of k. The intelligent primer design system contains the following two major parts. The first part is similar to other traditional primer design systems, such as PRIMER3 (Rozen & Skaletsky, 2000). Such primer design systems considered the traditional primer design criteria (Antao, Lai, & Tinoco, 1991; Antao & Tinoco, 1992) and are useful for RT-PCR. The first part provides interfaces for inputting parameters, which set up the relevant criteria, the primer length, the length of priming windows, the number of allowable base-pair mismatches, and DNA sequences. Certainly, the number of allowable base-pair mismatch might have impact on the primer's specificity. To mitigate this concern, the system left the choice to the user. For long primers (more than 15 nucleotides), one or two base-pair mismatches have little effect on the experimental result. It was shown in Wallace (1979) that the base-pair mismatch problem can be resolved by tuning the PCR conditions. Thus, the system can set the number of allowable base-pair mismatches up to 2.

The first part also provides a mechanism which, according to user's choices of T_m, GC content, the primer length, and the number of allowable base-pair mismatches, screens primer candidates for both ends of each sequence. This mechanism applies standard criteria mentioned in (Antao, Lai, & Tinoco, 1991; Antao & Tinoco, 1992) for filtering out inappropriate primer candidates that contain poly N, poly AG, and potential hairpin patterns. After filtering out inappropriate primer candidates, the other primers become the input of the second part of the system.

The base of the second part of the system was a mathematical model, which utilized the concept of a multi-gene primer. The second part was also the core of the intelligent systems. The techniques involved included techniques in operations research and computer programming; see Section 3 for the detail. There were also many other intelligent systems for biological and clinical applications. These systems were based on diversified techniques. The following are a few random examples: data mining (Alonso, Caraça-Valente, González, & Montes, 2002; Wang et al., 2005), total management (Chae, Park, Park, & Young, 1998), machine-learning (Wang, Hong, & Tseng, 1996), ontology (Delgado, Fajardo, Gibaja, & Pérez-Pérez, 2005), fuzzy logic (Daniels, Cayton, Chappell, & Tjahjadi, 1997), and soft computing (Li, Dong, I, & Un, 2005).

The rest of the paper is organized as follows. In Section 2, we discuss the existing primer design systems for multiplex RT-PCR. In Section 3, we present the multi-gene primer selection problem. We give its mathematical formulation and propose an intelligent algorithm for solving it. In Section 4, we show the application to two real cases of clinical and biological significance. In Section 5, we give the managerial implication to bioinformatics. At the end, we summarize our research contributions and discuss future directions in bioinformatics.

2. Existing primer design systems for multiplex RT-PCR

A number of primer design systems for multiplex RT-PCR have been developed. (Kampke, Kieninger, & Mecklenburg, 2001) described an efficient algorithm for selecting primer collection for multiplex RT-PCR experiments. Their criteria for selecting the primer collection was similar to the traditional criteria for selecting a single pair of primers; i.e. they considered the trade-off between the primer's GC content, primer's T_m, and cross hybridization effect between primers. They did not put k into consideration. Therefore, if there were n genes (or DNA sequences) to be monitored in the multiplex RT-PCR, the system would create n primer pairs.

Pearson, Robins, Wrege, and Zhang (1995) showed that multi-gene primer selection problem is hard to solve computationally because it is NP-complete. They proposed a heuristic to find a sub-optimal solution for multi-gene primer selection problem. In 1996, Pearson, Robins, Wrege, and Zhang proposed two more heuristics for multi-gene primer selection problem (pearson et al., 1996). They ran the heuristics on 30 examples of DNA sequences using a primer length of just 5. They showed empirically that the two heuristics found only 19 optimal solutions out of the 30 examples. These heuristics could be classified as greedy methods.