数理モデル化と問題解決　40－6
（２００２．　６．　２６）

ゲノム上におけるPCR産物の特異性の高速解析

倉田 憲一 *　SAGUEZ Christian *　DINE Gerard *　中村 宏 †
Ken-ichi.Kurata@mas.ecp.fr
* エコール セントラル パリ
† 東京大学 先端科学技術研究センター

概要

ポストゲノム研究において全ゲノム情報を解析することは不可欠である。PCRはゲノム研究においてもっとも重要な実験技術の1つである。本研究で、われわれはPCRプライマーを設計するための斬新な方法を提案する。この方法はプライマー自体の特異性だけでなく、そのPCR産物のゲノム上における特異性を考慮する。PCR産物の長さの特異性を利用することにより標的特異的なプライマーペアを設計する。またハッシュおよびソートすることによりゲノム配列を高速に解析する方法も提案する。計算機実験により95％以上のORFにたいして特異的なプライマーペアを設計することに成功した。

Rapid Analysis of Specificity of PCR Product on the Whole Genome

KURATA Ken-ichi *　SAGUEZ Christian *　DINE Gerard *　and　NAKAMURA Hiroshi †
* Ecole Centrale Paris　† The University of Tokyo

Abstract

In post-genome studies, it is indispensable to analyse whole genomic information. PCR is one of the most important experimental methods in molecular biology. In this paper, we deploy a novel method to design PCR primers, which takes into account not only the specificity of the primers but also the uniqueness of the product length by using the whole genomic information. In this method, the analysis of the genome is rapidly executed in reasonable memory space by sorting and hashing the whole genomic information. We successfully designed target-specific primer-pairs for over 95% ORFs on *E. coli* genome.

1 Introduction

Thanks to the development of genetic engineering, the genomic sequences of many organisms have been unveiled. Therefore, It becomes the effective way to discover laws in the root of life things by analysing such information. In post-genome studies, it is indispensable to do research on the basis of whole genomic information. As a traditional biological method, there is Polymerase Chain Reaction (PCR), which is used broadly and usually in experiments of molecular biology, such as gene therapy, gene diagnosis, DNA sequencing and gene expression pattern observation. In order to do PCR experiments successfully, primers must be specific to the target DNA sequences. There are a couple of methods to make the genuinely target-specific sequences for probes and/or primers on the whole genome [2, 4]. However, there is no study on the method to ensure the uniqueness of production of primers on the whole genome. In hybridisation reaction, such as in northern hybridisation and in DNA chips, only the uniqueness of sequences is important. On the other hand, in PCR, we must ensure the uniqueness of the product length that is produced by a pair of primers. This point is very different from designing merely specific probes and/or primers. In order to design the primers that produce a specific product, we must take into consideration not only the specificity of primers themselves, but also the specificity of a pair of primers, namely, the uniqueness of its production in PCR.

In this paper, we propose a novel method to design a pair of target-specific primers to produce a unique length in PCR. In order to ensure the specificity of primers, the thermodynamic energy of primers is calculated. The primers that strongly hybridise with the target gene and hardly hybridise with other non-target genes are selected. We propose a sorting method like radix sort for analysing the whole genomic information in reasonable time and memory.

2 Algorithm

The goal is to find the sequences whose edge sequence is a unique concatenated sequence as short as possible on the whole genome. It is computationally infeasible to compare all the possible combinations of partial sequences of the whole genome. Hence, we take the following strategy to find such sequences at high speed. At first, the number of the combinations of sequences to be compared is reduced by hashing the genomic information. Next, an algorithm like radix sort processes the comparison among sequences.

Our method is composed of 4 steps, described as follows: (1) make a Look Up Table (LUT) from the whole genomic information; (2) make a Unique Sequence Length Table (USLT) by using the algorithm like radix sort; (3) calculate the hybridisation energy of primers; (4) select target-specific primer-pairs.

3 Result

We tried to design target-specific primers for 4405 Open Reading Frames (ORFs) of *E. coli* using the proposed method. The program was written in C++ language. It was compiled with Gnu compiler and executed on the computer having Crusoe TM5600 600 MHz CPU and 240MB memory under the Linux operating system. The length of the designed primers was 25-mer. The length of the hash-key sequence, h, was 4. The length of inner-sequence, s, was 10, namely, the length of the PCR product was 60. The limit of the minimum length of unique concatenated sequence, Th, was 7.

The total runtime of the method was about 35 minutes on 4405 ORFs of *E. coli.* The target-specific primer-pairs for 4218 ORFs were designed on *E. coli* genome. The ORF for which a target-specific primer-pair can be designed has at least one position on its sequence where $USL < Th$. The target-specific primer-pair to produce a specific band was successfully designed for over 95% ORFs on the above condition. The ORFs that had no target-specific primer-pair candidate mostly coded ribosomal RNA, transfer RNA, hypothetical protein and genes whose length was shorter than the product length to be designed.

4 Discussion

Our method directly finds genuine unique sequences on the genomic sequences. Look Up Table is made at first; that is to say, all the gene sequences are hashed. The number of the sequences to be compared among genes is lessened by using LUT and the comparison is executed in local area. Therefore, this method works fast because all the sequences having the same hash-key sequence are conceivably stored on a cache and the computation of comparison is constituted of only the local search for the sequences. Since cache speed is much faster than main memory in the current computers, this also contributes to accelerating the computation. Moreover, in our method, sequence comparison is rapidly executed by using the algorithm like radix sort. Besides that, the size of LUT and USLT is proportional to the size of the target genome. Namely, the increase of memory consumption is proportional to the size of the target genome in this method. This is very important result to design specific primers in reasonable time and memory consumption for all the gene sequences of the organism that has an enormous genome. Hereafter, the problem of memory consumption will surely become more important in bioinformatics.

As PCR primers, it is desirable to select the sequences that have as high value of Tm as possible because such sequences have high specificity to hybridise and they do not produce extra bands. In our method, the sequence that has the negative high value of hybridisation energy with its target and the negative low value of cross-hybridisation energy is selected. That is to say, the sequence that has the high value of Tm is selected.

In the near future, we will make an integrated software to support PCR experiments in molecular biology.

5 Acknowledgment

We thank Yoshida Scholarship Foundation for its support during our work.

References

[1] Allawi,H.T. Thermodynamics and NMR of Internal G.T Mismatches in DNA. *Biochemistry*, 36:10581–10594, 1997.

[2] Fugen,L. Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics*, 17:1067–1076, 2001.

[3] Griffais,R. K-tuple frequency in the human genome and polymerase chain reaction. *Nucleic Acids Research*, 19:3887–3889, 1991.

[4] Mitsuhashi,M. Oligonucleotide probe design - a new approach. *Nature*, 367:759–761, 1994.