

An Efficient Algorithm for Unique Signature Discovery on Whole-Genome EST Databases

Hsiao Ping Lee^{*}, Tzu Fang Sheu[†], Yin Te Tsai[‡], Ching Hua Shih[§], Chuan Yi Tang[¶]

Abstract

ESTs can be used to accelerate the various research activities for the discovery of new genes. Unique oligonucleotides are signatures that distinguish an EST from all the others. An important application of those short signatures is to be used in PCR primer design and microarray experiments. In this research, we propose an efficient approach to enhance our previous work on unique signature discovery to handle the dataset of whole-genome scale. The performances of our method are evaluated by the experiments on human chromosome EST databases.

1. Introduction

ESTs were found to be an invaluable resource for the discovery of new genes, particularly those involved in human disease processes. An *unique signature* can be considered as a landscape that distinguishes an EST from all the others. Unique signatures are particularly valuable as locus-specific PCR primers for placement of ESTs at single positions on a genetic linkage map, on microarrays for studies of the expression of specific genes without signal interference from other genes, and to probe genomic libraries in search of specific genes. For example, specific oligonucleotides have already been used in a PCR method for the identification of 14 human pathogenic yeast species [1]. This raises the question of whether it is possible to find a short, specific DNA sequence for human ESTs of whole-genome scale.

In this paper, we propose an efficient approach to extend our unique signature discovery algorithm, *IMUS*, for

whole-genome EST databases. Our method is also applicable on extending the other existing discovery algorithm [2]. To evaluate the performance, we make several simulations on human chromosome EST databases with a regular PC. Based on the statistics, our approach accomplishes the discovery over the EST dataset of 156M bases in one day. Furthermore, the extraction over the largest human chromosome EST database of 230M bases requires only 55 hours.

The rest of the paper is organized as follows. In Section 2, we define the unique signature problem formally. Our algorithm is presented in Section 3. Performance statistics for mining human chromosome EST databases can be found in Section 4. In Section 5, we draw some concluding remarks.

2. Problem Definition

Let $D = \{y_1, y_2, \dots, y_k\}$ denote the input dataset, where the generic string y_i , $1 \leq i \leq k$, is an EST sequence over the alphabet set $\sigma = \{A, C, G, T\}$. l and d are two positive integers. an l -pattern $p \subseteq y_i$ is referred to as a *unique signature* in D if there is no other l -pattern q in D such that the hamming distance between p and q is less than or equal to d . The unique signature discovery problem is to extract all unique patterns from the input database.

3. Algorithm

Let l , d and w be positive integers, where $d < l < w$. A *run* is an arranged partition of the input database D , which consists of a *core*, a specially selected w -segment, and l -patterns in the w -segments that are similar to the core. Assume p is an l -pattern from D . We say that p is (l, d) -mismatched to a run r if there exists at least one other l -pattern q in r such that the hamming distance between p and q is less than or equal to d . p is referred to as a *local signature* of r if $p \in r$ and p is not (l, d) -mismatched to r . We have the following observation.

Observation. Let l and d be two positive integers. An l -pattern is a unique signature of the input dataset if and only if it is a local signature of one run and not (l, d) -mismatched

* Department of Computer Science, National Tsing-Hua University, Taiwan, ROC. E-mail: shopping@cs.nthu.edu.tw

† Institute of Communication Engineering, National Tsing-Hua University, Taiwan, ROC. E-mail: sunnie@totoro.cs.nthu.edu.tw

‡ Department of Computer Science and Information Management, Providence University, Taiwan, ROC. E-mail: yttsai@pu.edu.tw

§ Department of Life Science, National Tsing-Hua University, Taiwan, ROC. E-mail: stewardshih@yahoo.com.tw

¶ Corresponding author: Department of Computer Science, National Tsing-Hua University, Taiwan, ROC. E-mail: cychang@cs.nthu.edu.tw

EST database	Size(M)	# signature(M)	Time(hh:mm:ss)
Chromosome 1	230.385	16.887	54:42:56.04
Chromosome 11	156.274	12.147	23:56:02.37
Chromosome 3	127.466	11.735	18:31:00.34
Chromosome 4	86.157	8.268	09:12:50.51
Chromosome 18	31.561	3.778	01:20:51.77
Chromosome Y	1.496	0.212	00:00:35.06

Table 1. The statistics for the unique signature discovery on some human chromosome EST databases.

to the other runs. The pattern is *duplicated* if and only if it is (l, d) -mismatched to at least one run.

The key idea underlying our approach is that if we can reduce the amount of candidates by identifying the duplicated patterns early, the efficiency of unique signature discovery will be improved. With some similarity criteria, we partition the input dataset into the set $R = \{r_1, \dots, r_n\}$ of n runs. The order of runs is carefully arranged to help to identify the duplicated patterns as early as possible. The extraction process is partitioned into two iterations of n passes. We load $r_i \in R, 1 \leq i \leq n$ into memory in pass i of each iteration. If it is in iteration 1, we apply the IMUS method to discover the local signatures s_i from r_i . Otherwise, we add the surviving patterns in s_i into the pool of unique signatures. Then, the currently existing sets s_g s of local unique signatures which have not been examined with r_i are enumerated. We verify the patterns in each s_g with r_i , and discard the patterns that are (l, d) -mismatched to r_i from s_g . The removal can be achieved by a simplified process of IMUS.

4. Experimental Results

Our approach is implemented on a PC of Red Hat Linux release 9, equipped with one XEON 2.8GHz CPU, 1GB DDR memory, and 80GB disk space. All our programs are coded in ANSI C, and compiled by GCC. The datasets employed in our experiments are human chromosome ESTs from NCBI. As a preprocess, the universal characters, for example the 'don't care' character, and the EST sequences that are short than 36 bases are discarded from the datasets.

Our experiments are made under the condition that the pattern length $l = 24$, maximum mismatches for duplicated patterns $d = 4$, width of run core $w = 96$ and number of runs $n = \lfloor \text{size of dataset} / 8.5M \rfloor$. Due to the limitation on paper size, we present partial result statistics in Table 1. The entries in the table are sorted by the execution time.

5. Conclusions

In this paper, we propose an efficient approach to extend our IMUS algorithm for the dataset of whole-genome scale.

With a regular PC of limited resources, our method extracts unique signatures from the largest human chromosome EST database of 230M bases within 55 hours, and accomplishes the discovery on the EST dataset of 156M bases in one day. These are great improvements, either in dataset scale or discovery efficiency. In the near future, we shall apply our approach in biological applications, such as PCR primer design.

References

- [1] B. M. Kiryu and C. P. Kiryu. Rapid identification of *Candida albicans* and other human pathogenic yeasts by using oligonucleotides in a PCR. *J. Clin. Microbiol.*, 73:1634-1641, 1998.
- [2] J. Zheng, T. Close, T. Jiang and S. Lonardi. Efficient Selection of Unique and Popular Oligos for Large EST Databases. *Proc. of Symposium on Combinatorial Pattern Matching (CPM'03)*, pages 273-283, 2003.