A new string matching method for genomic sequencingdata

XXX1, Zhang Zhang1,\*

1CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, China

2School of Computer Science and Technology, XXX

\***Corresponding authors**:

Zhang Zhang (zhangzhang@big.ac.cn): CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics,Chinese Academy of Sciences, No.7Beitucheng West Road, Building G, Chaoyang District, Beijing100029, China,Tel & Fax: +86-10-82995427

Running head: Parallel construction of multiple protein-coding DNA alignments

### Abstract

String matching algorithm plays the vital role in the Bioinformatics. In this research, we propose a fast pattern matching algorithm for biological sequences:The SSE Hash Shift algorithm.The algorithm use specialized word-size packed string matching instructions, based on the Intel streaming SIMD extensions (SSE) technology to compute the shift distance.The experimental results show the algorithm is superior to other algorithms especially when the length of the pattern is short then 64 bytes.

Key words: Exact string matching ,SSE,biological sequences

### Introduction

Pattern matching is one of the most important research subjects that have been studied in computer science.With the development of sequencing techniques, it has become easy to obtain the sequence, i.e. the linear arrangement of residues (nucleotides or amino-acids), of DNA, RNA, or protein molecules.

There are two types of pattern matching: single pattern matching and multiple patterns matching,here we focus on the single pattern matching.

Classical single pattern matching algorithm includes KMP algorithm [2] and BM algorithm [2]. Other algorithm [2] is mostly improved on these classical algorithms In BM algorithm, pattern string moves from left to right, while character is compared from right to left. When it loses matching, the predefined excursion function adopts the max value to decide right shift value. The efficiency of an algorithm depends on two phases: the pre-processingphase and the searching phase.The aim of a good algorithm is to minimize the work done during each attempt and to maximize the length of the shifts.

In the recent years,a new solutions called packed string matching technique has been used on the problem. In the packed string matching technique, multiple characters are packed into one larger word, so that the characters can be compared in bulk rather than individually.And in the modern CPU, SIMD instructions exist in many recent microprocessors supporting parallel execution of some operations on multiple data simultaneously via a set of special instructions working on limited number of special registers.

In this case ,we use Intel streaming SIMD extensions (SSE) technology to achive the goal.

### Material and methods

##### Algorithm

In this section we present the SSE Hash Shift algorithm. The algorithm is designed to be faster for media biological sequences. The algorithm is shown in Fig. 1.

, (1)

where  and .

##### Estimating XXX

XXX

, , ,  (2)

XXX

##### Data collection

X我们选取5种串匹配算法做对比试验，epsm，tvsbs，ufndmq，hashq，fsbndmq，简单介绍，测试集选取，

XXX

##### Section XXX

XXXX

### Results

##### Section 1

XXX

##### Section 2

XXX.

##### Section 3

XXX.

### Discussion

##### Section 1

XXX

##### Section 2

XXX.

##### Section 3

XXX.

### ~~Acknowledgments~~

~~We thank XXX for YYY. This work was supported by the “100-Talent Program” of Chinese Academy of Sciences (Y1SLXb1365; ZZ).~~

### References

[1] W.-H. Li, Molecular Evolution, Sinauer Associates, Sunderland, Massachusetts, 1997.

[2] B. Rannala, Z. Yang, Phylogenetic inference using whole genomes, Annu Rev Genomics Hum Genet 9 (2008) 217-231.

[3] Z. Yang, Inference of selection from multiple species alignments, Curr Opin Genet Dev 12 (2002) 688-694.

[4] J.P. Townsend, F. Lopez-Giraldez, R. Friedman, The phylogenetic informativeness of nucleotide and amino acid sequences for reconstructing the vertebrate tree, J Mol Evol 67 (2008) 437-447.

### Tables

Table 1.

Experimental results for searching different length of patterns on a ecoli genome sequence using six algorithm. Running times are expressed in millisecond, best results have been boldfaced and padded.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| len(m) | 8 | 12 | 16 | 20 | 24 | 28 | 32 | 36 | 40 | 44 | 48 | 64 | 128 | 256 | 512 | 1024 | 2048 |
| new | 146.55 | 95.9 | 72.35 | 58.2 | 49.25 | 43.1 | 38.9 | 35.9 | 33.7 | 32.25 | 30.95 | 29.2 | 26.4 | 41.25 | 32.1 | 17.65 | 11.65 |
| epsm | 158.05 | 157.8 | 123.55 | 124.75 | 65.05 | 66.1 | 46.35 | 47.15 | 37.2 | 37.5 | 31.3 | 25.9 | 21.45 | 17.95 | 13.4 | 10.95 | 9.15 |
| tvsbs | 531.29 | 409.43 | 372.43 | 325.86 | 292 | 276.29 | 300.43 | 278.86 | 245 | 256.71 | 244.29 | 268.71 | 277.14 | 293.86 | 270.57 | 301 | - |
| ufndmq | 527.86 | 407.29 | 370 | 324.86 | 291 | 275.57 | - | 262.57 | 225 | 274.86 | 286.71 | 243 | 254.57 | 274.86 | 280.43 | 254.71 | - |
| hashq | 5022.86 | 1015.71 | 574.43 | 400.86 | 307.71 | 253.14 | 216.14 | 187 | 167.14 | 149.57 | 138 | 103.86 | 63.71 | 45.29 | 43.71 | 40.29 | 40.29 |
| fsbndmq | 376.57 | 296.14 | 226.14 | 184 | 152.71 | 142.14 | 126.29 | 120.57 | 120.57 | 113.86 | 113.14 | 94.14 | 94.29 | 95.86 | 94 | 97.14 | 96.71 |

Table 2.

Experimental results for searching different length of patterns on a rice genome sequence using six algorithm. Running times are expressed in millisecond, best results have been boldfaced and padded.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| len(m) | 8 | 12 | 16 | 20 | 24 | 28 | 32 | 36 | 40 | 44 | 48 | 64 | 128 | 256 | 512 | 1024 | 2048 |
| new | 148.05 | 96.3 | 71.95 | 58.5 | 49.8 | 44.05 | 39.6 | 36.4 | 34.4 | 33.35 | 31.7 | 30 | 26.95 | 42.9 | 32.45 | 18.1 | 11.45 |
| epsm | 161.5 | 164.45 | 122.6 | 124.25 | 65.15 | 65.9 | 46.75 | 47.4 | 37.4 | 38.1 | 31.75 | 26.25 | 21.8 | 18.7 | 13.7 | 11.75 | 9.65 |
| tvsbs | 502 | 385.86 | 371.86 | 369.71 | 306.14 | 283 | 281.86 | 260.71 | 273.43 | 224 | 216 | 211.57 | 198.43 | 263.43 | 235.71 | 144.86 | - |
| ufndmq | 500.29 | 385.57 | 370.29 | 368.43 | 304.14 | 283.29 | - | 293 | 228 | 239.57 | 219.29 | 247.14 | 213 | 202 | 229.29 | 219.29 | - |
| hashq | 5033 | 1016.14 | 570.86 | 399.71 | 308.71 | 252.71 | 213.71 | 186.43 | 165.29 | 149.86 | 136.71 | 102 | 62.14 | 44.86 | 42.86 | 40 | 39.57 |
| fsbndmq | 317 | 280.29 | 224.43 | 180 | 148.14 | 139.71 | 124.57 | 116.57 | 121.43 | 117.86 | 107.43 | 93.43 | 92.14 | 94.43 | 91.71 | 91.43 | 93.71 |

Table 3.

Experimental results for searching different length of patterns on a human genome sequence using six algorithm. Running times are expressed in millisecond, best results have been boldfaced and padded.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| len(m) | 8 | 12 | 16 | 20 | 24 | 28 | 32 | 36 | 40 | 44 | 48 | 64 | 128 | 256 | 512 | 1024 | 2048 |
| new | 146.5 | 95.9 | 71.75 | 57.85 | 49.25 | 42.9 | 39.05 | 35.9 | 34 | 32.2 | 31.2 | 29.7 | 27.2 | 44.1 | 33.1 | 18.3 | 12.15 |
| epsm | 163.05 | 159.2 | 123.45 | 125.6 | 65.5 | 66.8 | 47.1 | 47.35 | 37.9 | 38.2 | 32.1 | 26.65 | 22.25 | 18.85 | 14.05 | 11.75 | 9.9 |
| tvsbs | 509.71 | 426.86 | 370.86 | 302.86 | 299.71 | 281.86 | 284.29 | 293.71 | 293.43 | 241.43 | 229.43 | 237.14 | 231.57 | 169.86 | 211 | 239.29 | - |
| ufndmq | 505.43 | 424.29 | 369 | 301.29 | 297.29 | 280.43 | - | 281.43 | 265.57 | 284.43 | 263.43 | 249.71 | 234.14 | 233.29 | 240.14 | 216.14 | - |
| hashq | 5028.29 | 1015.71 | 571.71 | 400.57 | 308.86 | 252.29 | 214.57 | 186.57 | 167 | 149 | 137.86 | 102.57 | 62.57 | 44.29 | 43 | 40.29 | 39.57 |
| fsbndmq | 360.29 | 281.29 | 215.14 | 182.86 | 166.14 | 137.14 | 126.86 | 118.57 | 120.14 | 117.14 | 113.43 | 91.14 | 96.29 | 96.14 | 95.43 | 96.29 | 93.43 |

Table 4.

Experimental results for searching different length of patterns on a protein genome sequence using six algorithm. Running times are expressed in millisecond, best results have been boldfaced and padded.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| len(m) | 8 | 12 | 16 | 20 | 24 | 28 | 32 | 36 | 40 | 44 | 48 | 64 | 128 | 256 | 512 | 1024 | 2048 |
| new | 128.15 | 84.6 | 64.05 | 52.2 | 44.4 | 39.95 | 36.25 | 33.6 | 32 | 30.65 | 29.9 | 28.1 | 29 | 40.4 | 29.6 | 17.05 | 10.8 |
| epsm | 123 | 122.8 | 122.55 | 125.6 | 64.85 | 65.85 | 46.25 | 46.85 | 37.25 | 37.6 | 31.5 | 26 | 21.8 | 18.6 | 13.4 | 10.7 | 8.9 |
| tvsbs | 222.29 | 152 | 120.86 | 103.57 | 90.86 | 81.71 | 70.57 | 64.43 | 59.14 | 54.71 | 50.86 | 42.14 | 31.57 | 29.29 | 31.57 | 28 | - |
| ufndmq | 222.57 | 152.43 | 120.29 | 103.71 | 90.43 | 81 | - | 66.14 | 60 | 56 | 50 | 43 | 31.86 | 28.71 | 31.43 | 29.14 | - |
| hashq | 5018 | 1010.14 | 566.14 | 431.43 | 317.29 | 249.71 | 211.43 | 184.71 | 163 | 147 | 133.71 | 100.14 | 59.71 | 42.71 | 41.43 | 39.43 | 39.43 |
| fsbndmq | 126.86 | 86.86 | 68.57 | 57.86 | 52.14 | 46.57 | 43.57 | 40.29 | 39.86 | 40.43 | 41.29 | 41.43 | 40.14 | 39.86 | 39.29 | 38 | 39 |

Table 5.

### Figure legends

Fig. 1 Parallelization scheme of ParaAT.

Fig. 2 Speedup (dotted lines) and running time (solid lines) for constructing protein-coding DNA alignments using 1–16 CPUs.

### Figures

(Font: Arial Narrow; Size: >=12pt; Format: EPS;)

Fig. 1

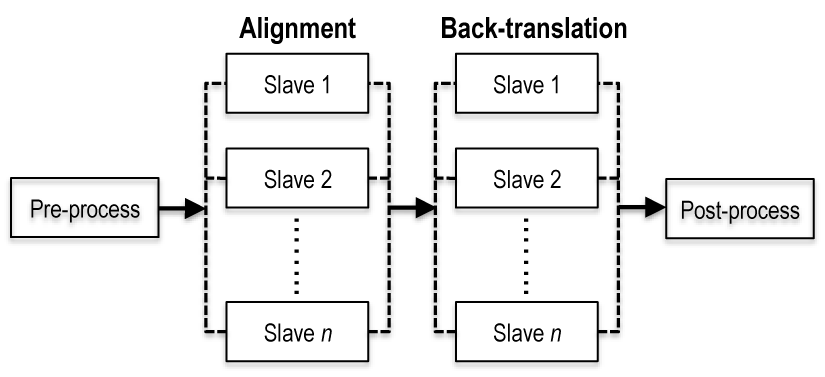


Fig. 2

