A new string matching method for genomic sequencing data

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.7 Beitucheng West Road, Building G, Chaoyang District, Beijing 100029, China, Tel & Fax: +86-10-82995427

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

### Abstract

Constructing multiple homologous alignments for protein-coding DNA sequences is crucial for a variety of bioinformatic analyses but remains computationally challenging. XXX

Key words: parallel, alignment, back-translation, homolog, protein-coding DNA alignment

### Introduction

Alignments of homologous sequences within and among species are of utmost importance for comparative genomics, molecular evolution and phylogenetic reconstruction [[1](#_ENREF_1),[2](#_ENREF_2),[3](#_ENREF_3),[4](#_ENREF_4)].

XXX.

Here we present XXX.

### Material and methods

##### Algorithm

XXX

, (1)

where  and .

##### Estimating XXX

XXX

, , ,  (2)

XXX

##### Data collection

XXXX

##### 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

Table 2

### 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

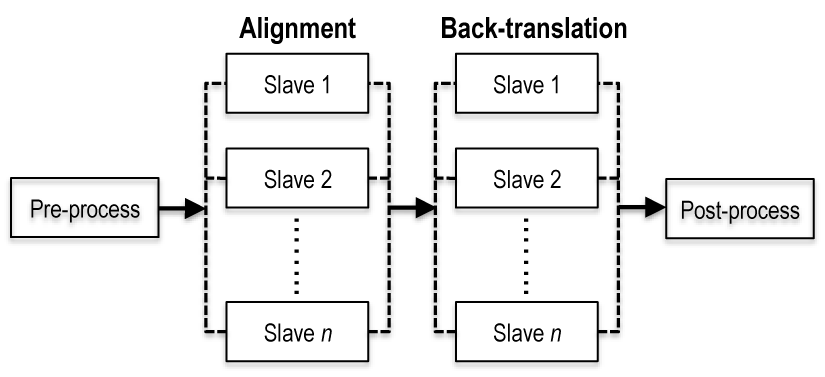


Fig. 2

