Pairwise Alignment on the Loop Structured Computer

Zhongmin Li gs98167@si.hirosaki-u.ac.jp
Yoshio Yoshioka slyoshi@aries.si.hirosaki-u.ac.jp
Toshio Shimizu slsimi@si.hirosaki-u.ac.jp
Department of Information Science, Graduate School of Science, Hirosaki University,
3 Bunkyo-cho, Hirosaki, Aomori 036-8561, Japan

1 Introduction

The Comparison of biological sequences (protein or DNA) is becoming more and more important in the field of new biology. One method of increasing the speed of the calculations is the parallel computing. In this work, we tried to calculate pairwise alignment by the Dynamic Programming on a data-flow parallel computer, Loop Structured Computer (LSC).

2 The parallel computer, LSC

The LSC is a 128-PE (processing element) data-flow parallel computer developed at Hirosaki University (Fig. 1).

The LSC is a MIMD (multiple instruction multiple data stream) architecture parallel computer, constructed by 127 computation PEs and a communication PE connected to a host computer by RS-232-C. And the PEs are connected with the one directional shifter register in a loop configuration.

The number of computation PEs is variable (numbered from 1 to 127). Each processor is constructed with a MC68B09 MPU with a small amount of local memory (program packet memory is disposed in each processor), and it operates independently and asynchronously. The LSC needs to broadcast MIMD instruction packets to each PEs. The instructions and data in the host computer are packaged into packets (16-byte long) to be transfered to the associated PEs.

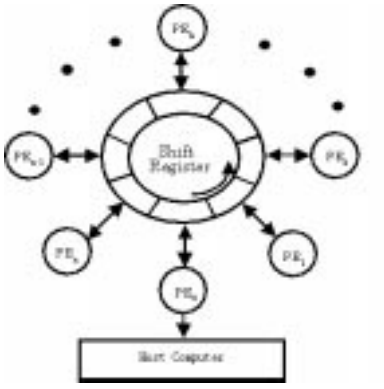
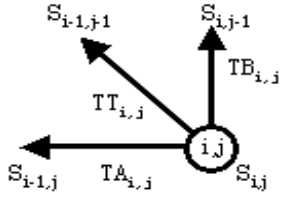
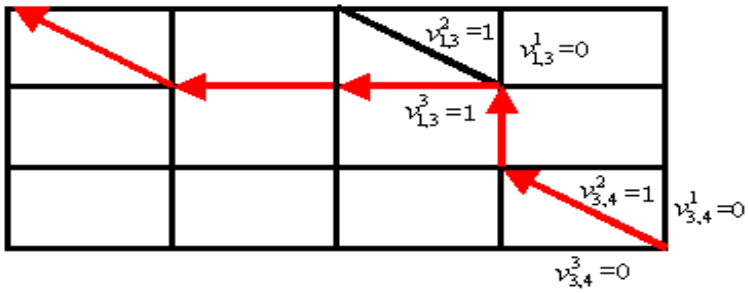


Figure 1: Architecture of the LSC.





3 The parallel algorithm

20 amino acids are represented with ASCII codes, e.g., Cysteine (C) is 67. Let us suppose that $A = a_1a_2a_3 \ldots a_n$ and $B = b_1b_2b_3 \ldots b_n$ are the sequences we want to align, providing the gap penalty d=3.

(1) The maximum similarity between those two sequences can be computed according to the following equations:

$$\begin{aligned} S_{0,0} &= 0; \\ S_{i,0} &= id \ (i = 1 \ to \ n); \\ S_{0,j} &= jd \ (j = 1 \ to \ m); \\ S_{i,j} &= \max \{S_{i-1,j-1} + W_{i,j}, \ S_{i-1,j} + d, \ S_{i,j-1} + d\}; \end{aligned}$$

where $W_{i,j}$ represents the similarity between elements a_i and b_j. $W_{i,j} = 1$ if $b_j = a_i$, and $W_{i,j} = -2$ if $b_j \neq a_i$.And the value $S_{i,j}$ represents the maximum similarity.

(2) Traceback: There are three direction: vertical, horizontal and diagonal for node(i, j) to back to node$(i-1, j)$, node$(i-1, j-1)$, node$(i, j-1)$, which can be determined by the following of equations: $TA_{i,j} = S_{i,j} - d$, $TT_{i,j} = S_{i,j} - W_{i,j}$, $TB_{i,j} = S_{i,j} - d$, and compared $TA_{i,j}$ with $S_{i-1,j}$, $TT_{i,j}$ with $S_{i-1,j-1}$, $TB_{i,j}$ with $S_{i,j-1}$:

$$\begin{array}{llll} \text{if } (TA_{i,j} = S_{i-1,j}) & V^1_{i,j}{=}1; & \text{else} & V^1_{i,j}{=}0; \\ \text{if } (TT_{i,j} = S_{i-1,j-1}) & V^2_{i,j}{=}1; & \text{else} & V^2_{i,j}{=}0; \\ \text{if } (TB_{i,j} = S_{i,j-1}) & V^3_{i,j}{=}1; & \text{else} & V^3_{i,j}{=}0; \end{array}$$

where $V_{i,j}$ represents the judge condition. if the product of all $V_{i,j}$s is 1, the best route is determined. And if there are many products which are 1, the horizontal direction is first.

4 Results and Discussion

$$\begin{array}{llll} \text{a[3]} = \text{"G} & \text{C} & \text{T"} & \\ \text{b[4]} = \text{"G} & \text{A} & \text{C} & \text{T"} \end{array} \quad \Longrightarrow \quad \begin{array}{llllllll} \text{G} & * & \text{C} & \text{T} & 71 & 100 & 67 & 84 \\ \text{G} & \text{A} & \text{C} & \text{T} & 71 & 65 & 67 & 84 \end{array}$$

Figure 2: The result of pairwise alignment.

There is a simple pairwise alignment (Fig. 2) which has been written with the special language of LSC. This is the first time that the letter sequences were solved on the data-flow parallel computer LSC. However, this pairwise alignment is too simple that is just 3 or 4 alphabet long. It is necessary to solve more general problems, e.g., longer sequences with consecutive gaps, and so on.

In order to valuate the processoring performance, the "speed-up rate" diagram was made. The result shows the nature of parallel processing is not so high at this stage. It seems necessary to rewrite our program to attain a higher parallel algorithm.

References

[1] Yoshio Yoshioka, Data-flow parallel computer LSC: Loop Structured Computer, *Fuji Soft Education Press*, 1995.

[2] Ishikawa, M., *Alignment Analysis of Protein Sequences by Parallel Computer*, 1994.

[3] Alexandrov, N. and Luethy, R., Alignment algorithm for homology modeling and threading, *Protein Science*, 7(2):254–258, 1998.