SuperCon2003

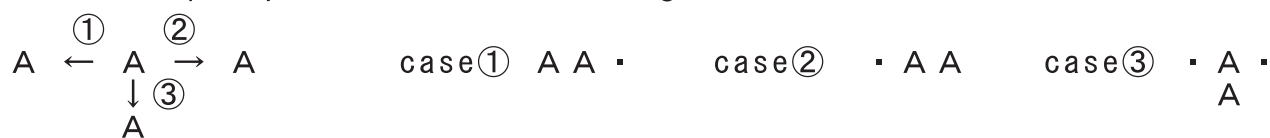**Sponsored by GSIC, Tokyo Institute of Technology**

SUPER COMPUTER CONTEST 2003
FOR HIGH SCHOOL STUDENTS

Entry Exercise: "Longest Chain Problem"

Imagine a square filled with 9 different kinds of symbols. We define the symbol-chain as the same symbol sequence. This sequence should be made up only to the direction of left, right and downward.

```
        ①    ②
A  ←  A  →  A            case①  A A ·      case②  · A A      case③  · A ·
      ↓ ③                                                             A
      A
```

Given a symbol square, write a program which finds the longest symbol-chain. The following example illustrates the longest chain of symbol A.

```
C C C D D D C D D D          · · · · · · · · · ·
B C B A D B C B A D          · · · A · · · · · ·
D B B A A A A A E E          · · · A A A · · · ·
C D B B B A A E A E          · · · · · A · · · ·
B C A A B B A A A E          · · · · · A A A A ·
B C B A D B C B A D          · · · · · · · · A ·
C D B B C C C B A D          · · · · · · · · A ·
C B A D D B C C D B          · · · · · · · · · ·
C C C D D D C D D D          C C C D D D C D D D
B C B A D B C B A D          B C B A D B C B A D
```

Contest Exercise: "Genome Analysis: Search for a Largest Protein Family"

A protein is a sequence of amino acids. Similar proteins (amino acid sequences) are considered to have similar functions. Thus, for identifying a type of proteins characterizing a given species, it is important to know a set of similar protein (which we call a "protein family").

There are 20 amino types and symbolized as A, B, C, ..., S,T here. Proteins, for example X and Y, are expressed as follows(Fig. 1):

```
X: CBADDKDDBBTPPEP  (length |X| = 15)
Y: CGAEEDDKDDBBTEEFGKKK   (length |Y| = 20)
    *      *      *
```
Fig.1

```
X: CBA--DDKDDBBTPPEP
Y: CGAEEDDKDDBBT--EEFGKKK
    *  *  *******  *
```
Fig.2

We define matching score is the (max) number of places where X and Y match, and sequence identity of X and Y is (matching score / min(|X|,|Y|)) *100. You can rearrange the original protein sequences to gain more matches(3 in Fig.1 and 11 in Fig.2). For a given set of proteins, write a program which finds a large (lager is better) protein family, which seq.id >= 25％.

The Contest was performed with the following hardware/software conditions:
SGI Origin2000(400MHz, 256 processors, 256GB memory)
64 processors and 32GB memory are available for contestants

Results:

1 minor answer and 2 correct answers within 156sec.

Team shiba(Shibaura Kashiwa High School)
KUWADA Junya, SHIROISHITA Hiroshi

We are waiting for your challenge next year 2004.

TOKYO INSTITUTE OF TECHNOLOGY