NAME

MakeArray − Suffix Array Implementation for GeneSeqer

SYNOPSIS

MakeArray [*dbest*]

DESCRIPTION

MakeArray is an implementation of the Manber and Myer suffix array algorithm (Reference 1) for applications in conjunction with the **GeneSeqer** program (Reference 2). The distributed version is an implementation by Fred Goodman and George Juras in collaboration with VisualMetrics Corporation (unpublished).

Information on program availability may be obtained at http://bioinformatics.iastate.edu/bioinformatics2go/. Correspondence relating to **MakeArray**/**GeneSeqer** should be addressed to

Volker Brendel
Department of Genetics, Development and Cell Biology
Iowa State University
2112 Molecular Biology Building
Ames IA 50011-3260
U.S.A.
phone: (515) 294-9884; fax: (515) 294-6755
email: vbrendel@iastate.edu

REFERENCES

1. Manber, U. & Myers, G. (1993)
Suffix arrays: a new method for on-line search.
SIAM J. Comput. 22, 938-945.

2. Usuka, J., Zhu, W. & Brendel, V. (2000)
Optimal spliced alignment of homologous cDNA to a genomic DNA template.
Bioinformatics 16, 203-211.

USAGE

Input file format

FASTA file format refers to raw sequence data separated by identifier lines of the form starting with ">" followed by the sequence name.

Examples:

```
>gi|sequence1 - upper case
ACGATTGGATCAAAATCCATGAAAGAGGGGAATCTATAGGCGGAATTGAG
CGCCAGCGACTGGCTGCCTTGGCGGGGGAGGCCTTGGCGGA
```

```
>SQ;sequence2 - upper case with numbering
       1  ACGATTGGAT CAAAATCCAT GAAAGAGGGG AATCTATAGG CGGAATTGAG
      51  CGCCAGCGAC TGGCTGCCTT GGCGGGGGAG GCCTTGGCGG A

>vb_sequence3 - lower case
acgattggatcaaaatccatgaaagaggggaatctataggcggaattgagcgccagcgac
tggctgccttggcgggggaggccttggcgga

>vb:sequence4 - mixed format
       1  ACGATTGGAT CAAAATCCAT GAAAGAGGGG AATCTATAGG GGGGGGATCT
cgccagcgac
        tggctgcct       tggcggggg       AGGCCTTGGCGGA
```

Output

Output to standard output gives statistics of the EST database, including number of sequences, total length, minimum and maximum sequence length, and length distribution.

In addition, the program produces four binary files that are read by the **GeneSeqer** program (**GeneSeqer** option: **-dD** *dbest*):

dbest.dat
: Data file.

dbest.ind
: Index file.

dbest.suf
: Suffix Array file.

dbest.tre
: Lcp Tree file.

COMPILATION OPTIONS

The minimal allowed sequence length is set by

#define MinSeqLen x

in GENESEQER/src/EstData.c (default value: x = 12).

FILES

GENESEQER/README
GENESEQER/bin
GENESEQER/data (examples)
GENESEQER/doc/MakeArray.1 (this file)
GENESEQER/include
GENESEQER/src

SEE ALSO

GeneSeqer(1), SplicePredictor(1).

NOTES

A hardcopy of this manual page is obtained by ‘man -t ./MakeArray.1 | lpr’.

AUTHOR

Volker Brendel <vbrendel@iastate.edu>