Linear Time Suffix Array Computation Tool

Robert Bakarić
rbakaric@irb.hr
bakaric@evolbio.mpg.de
01.09.2015
SuffixArray-1.0

Abstract

Suffix array is a data structure used commonly for full text indices, data compression algorithms and in various matching algorithms in bioinformatics and computational biology. It is an array of indices each corresponding to a starting position of a suffix in a given string. This software is a C++ encapsulation of sais-light-2.4.1 library written by Yuta Mori [1] implementing the induced sorting [2] based suffix array construction algorithm.

Contents

1 Installation

The simplest way to compile this program is to:

1. Unpack the SuffixArray package (`suffixarray-XXX.tar.gz`):

   ```
   tar -xvzf suffixarray-XXX.tar.gz
   ```

2. Change the current directory to `suffixarray-XXX`:

   ```
   cd suffixarray-XXX/
   ```

3. Configure the program for your system (–bindir is optional):

   ```
   ./configure --bindir=/absolute/directory/path/suffixarray-xxx/bin
   ```

4. Compile the program:

   ```
   make
   ```

5. Install the program:

   ```
   make install
   ```

Your binaries should be located in your local bin directory if `--bindir` option has been set. Otherwise installation needs to be carried out with root privileges in order to be installed into `/usr/local/bin` directory.

2 Input files

The SuffixArray program takes a simple ASCII txt file and computes an array of indexes corresponding to lexicographically ordered text suffixes. An example of the input file can be found in `./suffixarray-xxx/demo` and it should look like this:

`TheArtOfWar`:

```
The Project Gutenberg eBook, The Art of War, by Sun Tzu


This eBook is for the use of anyone anywhere at no cost and with
almost no restrictions whatsoever.  You may copy it, give it away or
re-use it under the terms of the Project Gutenberg License included
with this eBook or online at www.gutenberg.org


Title: The Art of War
Subtitle: Text Only, no Commentary

Author: Sun Tzu

Translator: Lionel Giles

Release Date: December 28, 2005  [eBook #17405]
[Last updated: January 14, 2012]
```

```
Language: English

Character set encoding: ISO-646-US (US-ASCII)
...
```

3 Program options

I order to see program options type:

```
./bin/SuffixArray -h
```

Expected output:

```
Usage: ./program [options]

*******************************************************************************
                     SuffixArray - Suffix array computation tool
                                        by
                                   Robert Bakaric

CONTACT:
 Code written and maintained by Robert Bakaric,
 email: rbakaric@irb.hr , bakaric@evolbio.mpg.de

ACKNOWLEDGEMENT:
         1. Yuta Mori, 2010. https://sites.google.com/site/yuta256/sais

         2. Ge Nong, Sen Zhang and Wai Hong Chan, Two Efficient Algorithms for
            Linear Suffix Array Construction, 2008.

LICENSE:
 The program is distributed under the GNU General Public License. You should have
 received a copy of the licence together  with this software. If not, see
 http://www.gnu.org/licenses/
*******************************************************************************


Allowed options:
  -h [ --help ]           produce help message
  -v [ --version ]        print version information
  -i [ --input-file ] arg input file
```

4 Functions and classes

SuffixArray class:

`SuffixArray` :	SuffixArray class.
`make` :	Explicit constructor.
`destroy` :	Explicit destructor.
`ComputeSuffixArray` :	Function computes a suffix array for a given string.
`GetSuffixArray` :	Function returns index values of lexicographically ordered suffixes as a vector of template integers.

5 Example

A minimal example demonstrating the usage of SuffixArray demo program:

```
./bin/SuffixArray -i demo/TheArtOfWar

86385 86384 1076 67263 65932 1077 820 616 4922 8584 31080 25742
```

```
61393 1096 46956 15617 12656 58295 33735 40930 19592 67264 65992
65933 1078 4905 8557 12628 33706 80064 67755 81034 83917 66057
55 306 66164 15605 19563 25723 31051 40917 46930 58269 61368 65885
821 617 721 85658 4923 8585 31081 25743 61394 1097 46957 15618
12657 58296 33736 40931 19593 33924 15781 41177 25829 8916 61831
58589 1173 19802 12824 47251 31217 5441 34042 25981 5651 19931
62148 1328 31495 58732 9084 15951 47336 41259 47439 41466 13092
...
```

5.1 SuffixArray.hpp

Adding the `SuffixArray.hpp` header file to your program will allow you to include all the functions described in section 4. A minimal example:

```
    #include<vector>
    #include<SuffixArray.hpp>

    string text("This is my text");


/* Make SuffixArray */

    /* Construction */
    SuffixArray<int|long|unsigned|double> SA(text);
    /*  OR  */
    SuffixArray<int|long|unsigned|double> SA;
    SA.make(text);

/* Get Array */

   vector<int|long|unsigned|double>Array = SA.GetSuffixArray();
                                        // [15 4 7 10 0 12 1 2 5 8 3 6 14 11 13 9]
```

6 Acknowledgement

1. Yuta Mori, 2010. https://sites.google.com/site/yuta256/sais
2. Ge Nong, Sen Zhang and Wai Hong Chan, Two Efficient Algorithms for Linear Suffix Array Construction, 2008.

7 Future work