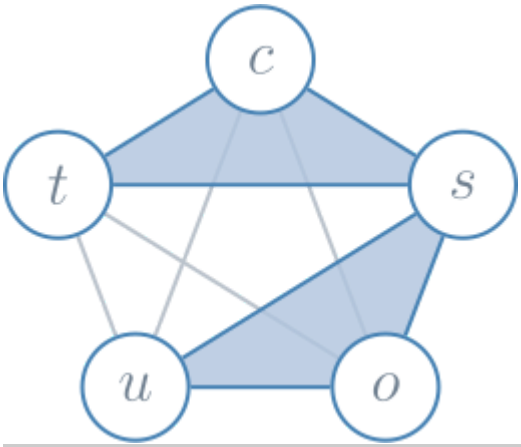

Theoretical Computer Science / Theoretische Informatik

Institut[e] f(o|ü)r Informati(cs|k), [Universität] Osnabrück [University]

Closest String and Closest Substring Problems

Here, you can find the benchmark sets used in our study of CSP and CSSP. Detailed descriptions can be found in the paper:

- M. Chimani, M. Woste, S. Böcker. **A Closer Look at the Closest String and Closest Substring Problem.** SIAM Workshop on Algorithm Engineering & Experiments 2011, San Francisco (ALENEX11), pp. 13-24, SIAM, 2011. (link to published version [http://www.siam.org/proceedings/alenex/2011/alx11_02_chimanim.pdf])

Instances

Type	Description	Filename	Solutions
CSP	Random	csp_rnd.tar.bz2 (135 MB)	results_csp_rnd.csv
CSP	McClure [1]	mcclure.tar.bz2 (2.7 KB)	results_mcclure.csv
CSP	Hufsky [2]	hufsky.tar.bz2 (175 KB)	results_hufsky.csv
CSSP	Random	cssp_rnd.tar.bz2 (119 KB)	results_cssp_rnd.csv

File format for CSP instances

The tar.bz2-files contain simple text files where each text file contains a single instance. The naming scheme for these files is as follows:

```
Random instances:  |Σ|-k-n-r-i.csp
McClure instances: McClure-p-|Σ|-k-n.csp
Hufsky instances:  Hufsky-k-n-i.csp
```

with

Symbol	Description
p	page of McClure paper
|Σ|	size of alphabet Σ
k	number of strings
n	length of strings
r	distance ratio such that α=n/r

i	instance number (10 instances are generated for each parameter combination for random instances, 100 for Hufsky instances)

The format of a text file is as follows:

```
<Alphabet-size (|Σ|)>
<Number of Strings (k)>
<Length of Strings (n)>
<1st Character of the Alphabet Σ>
...
<|Σ|-th Character of the Alphabet Σ>
<1st String>
...
<k-th String>
```

File format for CSSP instances

The tar.bz2-files contain simple text files where each text file contains a single instance. The naming scheme for these files is as follows:

```
{i,r}-|Σ|-k-n-l-d-{l,e}-i.cssp
```

with

Symbol	Description
{i,r}	i stands for implanted motif, r represents random instances
|Σ|	size of alphabet Σ
k	number of strings
n	length of strings
l	length of target string
d	distance
{l,e}	e stands for instances with E(l,d)≈1, l stands for instances with E(l,d)≈0.01, omitted in case of random instances
i	instance number (5 instances are generated for each parameter combination)

For implanted motif instances there are also files ending with .sup. These files contain supplemental information about the instances (see below).

The format of a .cssp file is as follows:

```
<Alphabet Size (|Σ|)>
<Number of Strings (k)>
<Length of Strings (n)>
<Length of Target String (l)>
<1st Character of the Alphabet Σ>
...
<|Σ|-th Character of the Alphabet Σ>
<1st String>
...
<k-th String>
```

The format of a .sup file is as follows:

```
<Implanted Master Substring>
<Distance (d) of the Implanted Master Substring>
<Position of Implanted Master Substring in the 1st String>
```

```
...
<Position of Implanted Master Substring in the k-th String>
```

File format for solution files

The solution files are plain text files in the CSV format with a semicolon as delimiter. The first line of a file serves as a column header. Each further line represents a specific instance. The column headers are as follows:

Header	**Description**
filename	the filename of the specific instance
formulation	only for CSSP instances: formulation used to solve the instance where d stands for δ
lb	lower bound of the solution value
ub	upper bound of the solution value
time	running time in seconds

---

1) The McClure instances originally appeared in [M. McClure, T. Vasi, and W. Fitch. Comparative analysis of multiple protein-sequence alignment methods. Molecular Biology and Evolution, 11(4):571-592, 1994]

2) The Hufsky instances were proposed in [F. Hufsky, L. Kuchenbecker, K. Jahn, J. Stoye, and S. Böcker. Swiftly computing center strings. In Proc. WABI 2010, volume 6293 of LNCS, pages 325-336. Springer, 2010]

research/csp_cssp.txt · Last modified: April 17, 2013 (22:00) by chimani