`uniquesub`: a program for computing minimum unique substrings a manual

Stefan Kurtz
Center for Bioinformatics,
University of Hamburg

August 6, 2012

1 The program `uniquesub`

The program `uniquesub` is called as follows:

`uniquesub` [*options*] `-query` *files* [*options*]

files is a white space separated list of at least one filename. Any sequence occurring in any file specified in *files* is called *unit* in the following. In addition to the mandatory option `-query`, the program must be called with either option `-pck` or `-esa` which specify to use a packed index or an enhanced suffix array for a given set of subject sequences.

`uniquesub` computes for all positions i in each unit, say s of length n, the length $mup(s, i)$ of the minimum unique prefix at position i, if it exists. Uniqueness always refers to all substrings represented by the index. $mup(s, i)$ is defined by the following two statements:

- If $s[i..n-1]$ is not unique in the index, then $mup(s, i) = \perp$. That is, it is undefined.
- If $s[i..n-1]$ is unique in the index, then $mup(s, i) = m$, where m is the smallest value such that $i + m - 1 \leq n - 1$ and $s[i..i + m - 1]$ occurs exactly once as a substring in the index.

Note that it is possible that for all $i \in [0, n-1]$ we have $mup(s, i) = \perp$, which means that unit s does not contain any unique substring. In this case, the program reports nothing for the corresponding unit. The program was developed for designing whole genome tiling arrays. The corresponding publication is [1].

The following options are available in `uniquesub`:

`-esa` *indexname*
: Use the given enhanced suffix array to compute the matches.

`-pck` *indexname*
: Use the packed index (an efficient representation of the FMindex) to compute the matches.

`-query` *files*
: Specify a white space separated list of query files containing the units. At least one query file

must be given. The files may be in gzipped format, in which case they have to end with the suffix `.gz`.

`-min` ℓ
: Specify the minimum length ℓ of the minimum unique prefixes. That is, for each unit s and each positions i in s, the program reports the values i and $mup(s, i)$ whenever $mup(s, i) \geq \ell$.

`-max` ℓ
: Specify the maximum length ℓ of the minimum unique prefixes. That is, for each unit s and each positions i in s, the program reports the values i and $mup(s, i)$ whenever $mup(s, i) \leq \ell$.

`-output` (`querypos`|`sequence`)
: Specify what to output. At least one of the two keys words `querypos` and `sequence` must be used. Using the keyword `querypos` shows the query position. Using the keyword `sequence` shows the sequence content of the match.

`-help`
: Show a summary of all options and terminate with exit code 0.

The following conditions must be satisfied:

1. Either option `-min` or option `-max` must be used.
2. If both options `-min` and `-max` are used, then the value specified by option `-min` must be smaller than the value specified by option `-max`.
3. Either option `-pck` or `-esa` must be used. Both cannot be combined.

2 Examples

Suppose that in some directory, say `homo-sapiens`, we have 25 gzipped fasta files containing all 24 human chromomsomes plus one file with mitrochondrial sequences. These may have been downloaded from `ftp://ftp.ensembl.org/pub/current_fasta/homo_sapiens_47_36i/dna`.

In the first step, we construct the packed index for the entire genome:

```
gt packedindex mkindex -dna -dir rev -parts 15 -bsize 10 -locfreq 32
                       -indexname human-all -db homo-sapiens/*.gz
```

The program runs for almost two hours and delivers an index `human-all` consisting of three files:

```
ls -lh human-all.*
-rw-r----- 1 kurtz gistaff   37 2008-01-24 00:47 human-all.al1
-rw-r----- 1 kurtz gistaff 1.9G 2008-01-24 02:37 human-all.bdx
-rw-r----- 1 kurtz gistaff 3.4K 2008-01-24 02:37 human-all.prj
```

This is used in the following call to the program `uniquesub`:

```
gt uniquesub -output querypos -min 20 -max 30 -query queryfile.fna
             -pck human-all
unit 0 (Mus musculus, chr 1, complete sequence)
1007 20
1010 22
1011 22
1012 21
1013 21
...
```

For all units s in the multiple FASTA file `queryfile.fna`, a line is shown, reporting the number of the unit and the original fasta header. Also, all for positions i in s satisfying $20 \leq mup(s, i) \leq 30$, i and $mup(s, i)$ is reported.

The first column is the relative position in the unit sequence (counting from 0). The second column shows the length value.

To additionally report the sequence content of the minimum unique prefixes we add the keyword `sequence` to option `-output`:

```
gt uniquesub -output querypos sequence -min 20 -max 30
             -query queryfile.fna -pck human-all
unit 0 (Mus musculus, chr 1, complete sequence)
1007 20 ctgacagttttttttttta
1010 22 acagttttttttttttacttta
1011 22 cagttttttttttttacttttat
1012 21 agttttttttttttacttttat
1013 21 gttttttttttttacttttata
...
```

References

[1] S. Gräf, F.G.G. Nielsen, S. Kurtz, M.A. Huynen, E. Birney, H. Stunnenberg, and P. Flicek. Optimized design and assessment of whole genome tiling arrays. *Bioinformatics*, 23 ISMB/ECCB 2007:i195–i204, 2007.