`matstat`: a program for computing matching statistics a manual

Stefan Kurtz
Center for Bioinformatics,
University of Hamburg

December 13, 2012

1 The program `matstat`

The program `matstat` is called as follows:

`matstat` [*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.

`matstat` computes the *matching statistics* for each unit. That is, for each position i in each unit, say s of length n, $ms(s,i) = (l,j)$ is computed. Here l is the largest integer such that $s[i..i+l-1]$ matches a substring represented by the index and j is a start position of the matching substring in the index. We say that l is the length of $ms(s,i)$ and j is the subject position of $ms(s,i)$.

The following options are available in `matstat`:

`-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 value ℓ for the length of the matching statistics. That is, for each unit s and each position i in s, the program reports all values i and $ms(s,i)$ if the length of $ms(s,i)$ is at least ℓ.

`-max` ℓ

Specify the maximum length ℓ for the length of the matching statistics. That is, for each unit s

and each positions i in s, the program reports the values i and $ms(s, i)$ if the length of $ms(s, i)$ is at most ℓ.

`-output (subjectpos|querypos|sequence)`

Specify what to output. At least one of the three keys words `subjectpos`, `querypos`, and `sequence` must be used. Using the keyword `subjectpos` shows the subject position of the matching statistics. Using the keyword `querypos` shows the query position. Using the keyword `sequence` shows the sequence content

`-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 `matstat`:

```
gt matstat -output subjectpos querypos sequence -min 20 -max 30
           -query queryfile.fna -pck human-all
unit 0 (Mus musculus, chr 1, complete sequence)
22 20 390765125 actgtatctcaaaatataaa
253 21 258488266 gggaataaacatgtcattgag
254 20 258488267 ggaataaacatgtcattgag
275 20 900483549 taattctatttttctttctt
480 20 1008274536 gcttgaagatcatgatccag
..
```

Here, the first column shows the relative positions in unit 0 for which the length of the matching statistics is between 20 and 30. The second column is the corresponding length value. The third column shows position of the matching sequence in the index, and the fourth shows the sequence content.