`tagerator`: a program for mapping short sequence tags a manual

Stefan Kurtz
Center for Bioinformatics,
University of Hamburg

26/08/2013

1 Preliminary definitions

By S let us denote the concatenation of all subject sequences. By $edist(u, v)$ we denote the unit edit distance of two strings u and v. Consider a sequence p of length m. An approximate match of p with up to k differences is a substring v of S such that $edist(p, v) \leq k$. An approximate prefix match of p with k differences and at most t occurrences is a substring v of S such that v occurs at most t times in S and $edist(v, p[1..i]) \leq k$ for some $i \in [1, m]$.

2 The program `tagerator`

The program `tagerator` is called as follows:

`gt tagerator` [*options*] `-q` *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 *short sequence tag* or *tag* for short. In addition to the mandatory option `-q`, the program must be called with either option `-pck` or `-esa`, which specify to use a packed index and an enhanced suffix array, respectively. Both indices are constructed from a given set of subject sequences.

`tagerator` maps each short sequence tag, say p of length m against the given index. The length of the tag is limited by the size of a pointer. If `gt` is a 32-bit binary, then m must be smaller or equal to 32. If `gt` is a 64-bit binary, m must be smaller or equal to 64. The program runs in three basic modes:

- In the *ms*-mode, it computes for all $i \in [1, m]$ the length ℓ of the longest prefix of $w[i..m]$ matching any substring of S. In addition, it reports start positions of such a prefix in S. As these values make up the well known matching statistics, we denote it by *ms*. The length value ℓ is the matching statistics length and the position values are the matching statistics positions. Both, the matching statistics length and the matching statistics position is uniquely determined. The matching statistics mode requires to specify a maximum number of occurrences, see option `-maxocc`.
- In the *cdiff*-mode, it computes all start positions of approximate matches with up to k differences in S. For each start position of an approximate match, say j, it reports the minimum integer ℓ such that $edist(p, S[j..j+\ell-1]) \leq k$. Since this mode matches the complete sequence p, we call this mode *cdiff*, for complete difference.

- In the *pdiff*-mode, it computes all start positions of approximate prefix matches with up to k differences and at most t occurrences in S. For each start position of an approximate prefix match, say j, it reports the minimum integers i and ℓ such that $edist(p[1..i], S[j..j+\ell-1]) \leq k$. Since this mode matches a prefix of p with some differences, we call this mode *pdiff*. This mode requires to specify a maximum number of occurrences, see option `-maxocc`.

The following options are available in `tagerator`:

`-q` *files*
: Specify a white space separated list of query files (in multiple FASTA format) containing the tags. 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`.

`-esa` *indexname*
: Use the given enhanced suffix array index to map the short sequence tags.

`-pck` *indexname*
: Use the packed index (an efficient representation of the FMindex) to map the short sequence tags.

`-e` *k*
: Specify the number of differences allowed. k must be a non-negative number. $k = 0$ means that no differences are allowed (exact matching) and $k > 0$ means a positive number of differences. If this option is not used, then the program runs in *ms*-mode, i.e. it computes the matching statistics for each short sequence tag.

`-nod`
: Do not compute direct matches, i.e. matches on the forward strand. If this option is not used, then matches are computed on the forward strand.

`-nop`
: Do not compute palindromic matches, i.e. matches on the reverse complemented strand. If this option is not used, then matches are computed on the reverse complemented strand.

`-maxocc` *t*
: Specify the maximum number of occurrences of exact prefix matches (in case of the matching statistics) or approximate prefix matches.

`-withwildcards`
: Output matches containing wildcard characters (e.g. N). This option cannot be used in any of the following cases:

  - with option `-pck`,
  - in the *ms*-mode,
  - for all modes with $k = 0$.

`-best`
: For each tag, only show matches for the smallest possible distance. That is, if a tag has exact matches in the input index, then only exact matches are shown. If there are no exact matches, but matches with distance 1, then only these are shown. If there are no matches matches with distance 1 (and hence no exact matches), but with distance 2, then only these are shown etc.

`-output` $key_1 \ldots key_q$

Use combination of the following keywords to specify output according to the following table:

keyword	shows the following
tagnum	show ordinal number of tag
tagseq	show tag sequence
dblength	show length of match in database
dbstartpos	show start position of match in database
abspos	show absolute value of dbstartpos
dbsequence	show sequence of match
strand	show strand
edist	show edit distance
tagstartpos	show start position of match in tag (only for `-maxocc`)
taglength	show length of match in tag (only for option `-maxocc`)
tagsuffixseq	show suffix tag involved in match (only for option `-maxocc`)

This option only has an effect when used in the cdiff-mode. If in cdiff mode this option is not used, then the output is such as if keywords `tagnum`, `tagseq`, `dblength`, `dbstartpos`, `strand` were used.

`-help`

Show a summary of all options and terminate with exit code 0.

The following conditions must be satisfied:

1. Option `-q` is mandatory.
2. Either option `-pck` or `-esa` must be used. Both cannot be combined.
3. If option `-e` is not used, then option `-maxocc` is required.

3 Examples

Suppose that in some directory, say `homo-sapiens`, we have 24 gzipped FASTA files containing all 24 human chromomsomes. 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 human genome:

```
gt packedindex mkindex -dna -dir rev -parts 12 -bsize 10 -sprank -locfreq 32
                       -tis -ssp -indexname pck-human -db homo-sapiens/*.gz
```

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

```
ls -lh human-all.*
-rw-r----- 1 kurtz gistaff   37 2008-07-13 17:00 pck-human.al1
-rw-r----- 1 kurtz gistaff 2.6G 2008-07-13 19:22 pck-human.bdx
-rw-r----- 1 kurtz gistaff 3.3K 2008-07-13 19:22 pck-human.prj
```

Suppose that the compressed file `Q1.gz` contains short sequence tags in multiple FASTA format. In a first call we run `tagerator` in *ms*-mode:

```
$ gt tagerator -q Q1.gz -maxocc 10 -pck pck-human
# computing matching statistics
# indexname(pck)=pck-human
# queryfile=Q1.gz
# for each match show: tagnum tagseq dblength dbstartpos strand taglength
# 0 15 aagcttgctgctgca
10 + 88693
9 + 88694 213545
8 + 25583 88695 213546 281700 325202 451235 747565
12 + 470064
11 + 470065
10 + 158315 363967 470066 501203 576660 729958
9 +
8 +
7 +
6 +
5 +
4 +
3 +
2 +
1 +
12 - 271321
11 - 271322
10 - 271323
9 - 217234 271324 762404
8 - 146318 216480 217235 271325 294145 762405
```

The first line shortly reports the kind of computation performed. The second and third line give the name of the index containing the subject sequences. It is reported whether it is a packed index or an enhanced suffix array. Then, for each tag its length, say m, and the tag p itself is shown, followed by a block of m lines each containing one integers. For all $i \in [1, m]$, the first column in the ith line is the matching statistics length, say l. This means that $p[i..i+l-1]$ is the maximum length prefix of $p[i..m]$ that occurs as a substring in the index. The second column gives the strand of the match: a + stands for the forward strand and a − for the reverse strand. If the number of occurrences of $p[i..i+l-1]$ is smaller than the maximum occurrence parameter (10 in the above case), then all matching statistics positions are reported in ascending order. If it is larger than the maximum occurrence parameter, no positions are shown. So, for example, the sequence `gcttgctg` of length 8 starting at tag position 3 is the longest prefix of `gcttgctgctgca` that occurs as a substring in the index. It occurs at the positions 25583, 88695, 213546, etc. Note that the matching statistics lines for the forward strand are followed by the matching statistics block on the reverse strand. Note however, that the matching statistics positions are always reported with respect to the forward strand.

In a second call, we run `tagerator` in *cdiff*-mode:

```
$ gt tagerator -q Q1.gz -pck pck-human -e 2
# computing complete matches with up to 2 differences
# indexname(pck)=pck-human
# queryfile=Q1.gz
# for each match show: tagnum tagseq dblength dbstartpos strand
# 0 aagcttgctgctgca
14 904 114 +
13 1439 224 +
13 1250 191 +
13 413 155 +
13 468 294 +
15 1158 74 +
14 227 311 +
```

```
13 1439 111 +
15 496 288 +
15 356 327 +
15 1690 336 +
13 1439 223 +
13 1250 190 +
14 1178 82 +
14 204 59 +
14 273 46 +
14 1156 136 +
14 803 343 +
15 1439 222 +
15 1250 189 +
15 803 342 +
16 1474 165 -
# 1 atttgggactgtatctca
```

The first line shortly reports the kind of computation performed. The second and third line are as in the previous example. Then, for each tag its length and the tag itself is shown, followed by a block of lines each containing three integers and one of the symbols + or −, denoting the strand of the match. Each such triple of integers reports the length of an approximate match, the subject sequence in which the match occurs, and the relative position of the match in this sequence. For each start position only the shortest length is reported. If there are no approximate matches, then no such line appears. Note that the match always refers to the complete pattern.

Our third example concerns the *pdiff*-mode:

```
$ gt tagerator -q Q1.gz -pck pck-human -e 1 -maxocc 5
# computing prefix matches with up to 1 differences and at most 5 occurrences in the subject sequences
# indexname(pck)=pck-human
# queryfile=Q1.gz
# for each match show: tagnum tagseq dblength dbstartpos strand taglength
# 0 15 aagcttgctgctgca
8 1264 122 + 8
8 549 197 + 8
8 734 280 + 8
9 699 24 + 8
9 1060 311 + 8
9 1594 18 + 9
9 1475 90 + 9
9 1022 445 + 9
9 718 87 + 9
9 913 382 + 9
9 835 315 + 9
9 855 408 + 9
8 1534 258 + 7
8 1264 121 + 7
8 1476 281 + 7
```

The output is similar as in the previous example, except that each match is reported by four integers with a sign between the first three and the last number. The first number reports the length of the approximate match, the second reports the subject sequence number in which the match occur, the third its relative position in the subject sequence. The last number reports the length of the prefix of the tag involved in the approximate match. Note that in some cases, the two length values are different.