RLPBWT con bitvectors

Algorithm 1 Algoritmo per la costruzione di una colonna della RLPBWT con bitvectors

```
1: function BUILD(col, pref, div)
         c \leftarrow 0, \ u \leftarrow 0, \ v \leftarrow 0, \ u' \leftarrow 0, \ v' \leftarrow 0, \ curr_{lcs} \leftarrow 0, \ tmp_{thr} \leftarrow 0, \ tmp_{beg} \leftarrow 0
3:
          start \leftarrow \top, \ beg_{run} \leftarrow \top, \ push_{zero} \leftarrow \bot, \ push_{one} \leftarrow \bot
4:
         for every k \in [0, height) do
5:
              if k = 0 \land col[pref[k]] = 1 then
6:
                  start \leftarrow \bot
7:
              if col[k] = 0 then
8:
                  c \leftarrow c + 1
9:
         runs \leftarrow [0..0]
                                                                                            \triangleright sparse bit
vector for runs of length height+1
10:
          thrs \leftarrow [0..0]
                                                                                         \triangleright sparse bit
vector for thresholds of length height
          zeros \leftarrow [0..0]

ones \leftarrow [0..0]
11:
                                                                                                        \triangleright sparse bit
vector for zeros of length c
12:
                                                                                            \triangleright sparse bitvector for ones of length height-c
          samples_{beg} \leftarrow [], samples_{end} \leftarrow [] if start then
13:
                                                                                                   \triangleright couple of vectors for samples of length r
14:
15:
              push_{one} \leftarrow \top
16:
          else
17:
              push_{zero} \leftarrow \top
18:
          for every k \in [0, height) do
19:
               if beg_{run} then
20:
                    u \leftarrow u', \ v \leftarrow v', \ tmp_{beg} \leftarrow pref[k]
21:
                    beg_{run} \leftarrow \bot
22:
               if col[pref[k]] = 1 then
23:
                   v' \leftarrow v' + 1
24:
               else
25:
                    u' \leftarrow u' + 1
26:
               if k = 0 \lor col[pref[k]] \ne col[pref[k-1]] then
27:
                    curr_{lcs} \leftarrow div[k], \ tmp_{thr} \leftarrow k
28:
               if div[k] < curr_{lcs} then
29:
                   curr_{lcs} \leftarrow div[k], \ tmp_{thr} \leftarrow k
               if k = height - 1 \lor col[pref[k]] \ne col[pref[k+1]] then
30:
31:
                    runs[k] \leftarrow 1
32:
                    if k \neq height - 1 \wedge div[k+1] < div[tmp_{thr}] then
33:
                        thrs[k] \leftarrow 1
34:
                    _{
m else}
35:
                        thrs[tmp_{thr}] \leftarrow 1
                    push(samples_{beg}, tmp_{beg})
36:
37:
                    push(samples_{end}, pref[k])
38:
                    if push_{one} then
39:
                        if v \neq 0 then
40:
                             ones[k-1] = 1
41:
                        swap(push_{zero}, push_{one})
42:
                    else
43:
                        if u \neq 0 then
44:
                             zeros[k-1] = 1
45:
                        swap(push_{zero}, push_{one})
46:
                    beg_{run} \leftarrow \top
47:
          if |zeros| \neq 0 then
48:
               zeros[|zeros|-1] \leftarrow 1
49:
          if |ones| \neq 0 then
50:
               ones[|ones|-1] \leftarrow 1
51:
           build rank/select for the four bitvectors
          \mathbf{return}\ (\mathit{start},\ c,\ \mathit{runs},\ \mathit{zeros},\ \mathit{ones},\ \mathit{samples}_{\mathit{beg}},\ \mathit{samples}_{\mathit{end}},\ \mathit{div})
```

Algorithm 2 Algoritmo per estrazione simbolo da una run in una colonna

```
1: function GET_SYMBOL(s, r) 
ightharpoonup s = 	op 	ext{ iff column start with } 0, r 	ext{ run index}
2: if s then
3: if r 	ext{ mod } 2 = 0 then return 0 else return 1
4: else
5: if r 	ext{ mod } 2 = 0 then return 1 else return 0
```

Algorithm 3 Algoritmo per uvtrick

```
1: function uvtrick(k, i)
                                                                                \triangleright k is column index, i row index
 2:
         if i = 0 then
 3:
             return (0, 0)
         run \leftarrow rank_h^k(i)
 4:
         if run = 0 then
 5:
 6:
             if start^k then
                 return (index, 0)
 7:
 8:
             else
 9:
                 return (0, index)
10:
         else if run = 1 then
              if start^k then
11:
                  return (select_h^k(run) + 1, i - (select_h^k(run) + 1))
12:
13:
                  return (i - (select_h^k(run) + 1), select_h^k(run) + 1)
14:
15:
         else
              if run \mod 2 = 0 then
16:
17:
                  pre_u \leftarrow select_u^k(\frac{run}{2}) + 1
                  pre_v \leftarrow select_v^{\tilde{k}}(\frac{r\tilde{u}n}{2}) + 1
18:
                  offset \leftarrow i - (se\tilde{l}ect_h^k(run) + 1)
19:
                  if start^k then
20:
21:
                      return (pre_u + offset, pre_v)
22:
                  else
                      return (pre_u, pre_v + offset)
23:
24:
              else
                  run_u \leftarrow \left(\frac{run}{2}\right) + 1
25:
                  run_v \leftarrow \frac{ru\tilde{n}}{2}
26:
                  if \neg start^{k^2} then
27:
28:
                      swap(run_u, run_v)
29:
                  pre_u \leftarrow select_u^k(run_u) + 1
30:
                  pre_v \leftarrow select_v^k(run_v) + 1
                  offset \leftarrow i - (select_h^k(run) + 1)
31:
                  if start^k then
32:
33:
                      return (pre_u, pre_v + offset)
34:
                  else
                      return (pre_u + offset, pre_v)
35:
```

Algorithm 4 Algoritmo per lf-mapping

```
1: function LF(k, i, s) \triangleright k is column index, i row index, s symbol 2: c \leftarrow rlpbwt[k].c 3: (u, v) \leftarrow uvtrick(k, i) 4: if s = 0 then 5: return u 6: else 7: return c + v
```

${\bf Algorithm}~{\bf 5}$ Algoritmo per lf-mapping inverso

```
1: function REVERSE_LF(k, i)
                                                                   \triangleright k is column index, i row index
         if k = 0 then
                                                                                                ⊳ by design
 2:
              return 0
 3:
         k \leftarrow k - 1
 4:
         c \leftarrow rlpbwt[k].c
 5:
         if i < c then
 6:
              if start^k then
 7:
                   run \leftarrow rank_u^k(i) \cdot 2
 8:
              else
 9:
                   run \leftarrow rank_{n}^{k}(i) \cdot 2 + 1
10:
              i_{run} \leftarrow 0
11:
              if run \neq 0 then
12:
                   i_{run} \leftarrow select^k_h(run) + 1
13:
              (prev_0, \_) \leftarrow uvtrick(k, i_{run})
14:
              return i_{run} + (i - prev_0)
15:
         else
16:
              if start^k then
17:
                   run \leftarrow rank_v^k(i) \cdot 2 + 1
18:
19:
              else
                   run \leftarrow rank_v^k(i) \cdot 2
20:
              i_{run} \leftarrow 0
21:
22:
              if run \neq 0 then
                   i_{run} \leftarrow select_h^k(run) + 1
23:
              (\_, prev_1) \leftarrow uvtrick(k, i_{run})
24:
              return i_{run} + (i - (c + prev_1))
25:
```

Algorithm 6 Algoritmo per match con aplotipo esterno con panel $width \times height$

```
1: function EXTERNAL_MATCHES(z)
                                                                                           \triangleright assuming |z| = rlpbwt.width
          f \leftarrow 0, \ f_{run} \leftarrow 0, \ f' \leftarrow 0
 2:
          g \leftarrow 0, \ g_{run} \leftarrow 0, \ g' \leftarrow 0
 3:
          e \leftarrow 0, l \leftarrow 0
 4:
 5:
          for every k \in [0, |z|) do
               f_{run} \leftarrow rank_h^k(f), \ g_{run} \leftarrow rank_h^k(g)
 6:
               f' \leftarrow lf(k, f, z[k]), g' \leftarrow lf(k, g, z[k])
 7:
 8:
              l \leftarrow g - f
              if f' < g' then
 9:
10:
                    f \leftarrow f', \ g \leftarrow g'
               else
11:
12:
                    if k \neq 0 then
13:
                        report matches in [e, k-1] with l haplotypes
                    if f' = |lcp^{k+1}| then
14:
15:
                        e \leftarrow k + 1
16:
                    else
                         e \leftarrow lcp^{k+1}[f']
17:
18:
                    if (z[e] = 0 \land f' > 0) \lor f' = height then
19:
                         f' \leftarrow g' - 1
20:
                        if e > 1 then
                             f_{rev} \leftarrow f', \ k' \leftarrow k+1
21:
                             while k' \neq e - 1 do
22:
                                  f_{rev} \leftarrow reverse\_lf(k', f_{rev}), k' \leftarrow k' - 1
23:
                             run \leftarrow rank_h^{k'}(f_{rev}), \ symb \leftarrow get\_symbol(start^{k'}, run)
24:
                             while e > 0 \land z[e-1] = symb do
25:
26:
                                  f_{rev} \leftarrow reverse\_lf(e, f_{rev})
                                  run \leftarrow rank_h^{e-1}(f_{rev})
27:
28:
                                  symb \leftarrow get\_symbol(start^{e-1}, run)
                        while f' > 0 \wedge (k+1) - lcp^{k+1}[f] \le e do e \leftarrow e-1
29:
30:
                         f \leftarrow f', \ g \leftarrow g'
31:
                    else
32:
                         g' \leftarrow f' - 1
33:
                        if e \ge 1 then
                             f_{rev} \leftarrow f', \ k' \leftarrow k+1
34:
                             while k' \neq e - 1 do
35:
                                  f_{rev} \leftarrow reverse\_lf(k', f_{rev}), k' \leftarrow k' - 1
36:
                             run \leftarrow rank_h^{k'}(f_{rev}), \ symb \leftarrow get\_symbol(start^{k'}, run)
37:
                             while e > 0 \wedge z[e-1] = symb \ \mathbf{do}
38:
39:
                                  f_{rev} \leftarrow reverse\_lf(e, f_{rev})
                                  run \leftarrow rank_h^{e-1}(f_{rev})
40:
                                  symb \leftarrow get\_symbol(start^{e-1}, run)
41:
                        while e < height \land (k+1) - lcp^{k+1}[e] \le e do e \leftarrow e+1
42:
43:
                         f \leftarrow f', g \leftarrow g'
          if f < g then
44:
45:
               report matches in [e, |z|-1] with l haplotypes
46:
```

Algorithm 7 Algoritmo per match con matching-statistics (MS) e thresholds

```
1: function MATCHES_MS(z)
                                                                                                 \triangleright ms vectors with row and len of length |z|
         ms_{row} \leftarrow [0..0], \ ms_{len} \leftarrow [0..0]
3:
         curr_{row} \leftarrow rlpbwt[0].samples_{end}[|rlpbwt[0].samples_{end}|-1]
4:
         curr_{index} \leftarrow curr_{row}
 5:
         curr_{run} \leftarrow rank_h^0(curr_{index})
6:
7:
         symb \leftarrow get\_symbol(start^0, curr_{run})
                                                                                                           ▶ build matching statistics row
         for every k \in [0, |z|) do
8:
              if z[i] = symb then
9:
                  ms_{row}[k] \leftarrow curr_{row}
10:
                   if k \neq |z| - 1 then
11:
                        (curr_{index},\ curr_{run},\ symb) \leftarrow UPDATE(k, curr_{index}, z)
12:
               else
13:
                   curr_{thr} \leftarrow rank_t^k(curr_{index})
                   force_{down} \leftarrow \top iff we are over a threshold not at the end of a run force_{down} \leftarrow \top iff we are over a threshold at the end of a run and DOWN function is \top if |samples_{beg}^k| = 1 then
14:
15:
16:
17:
                        ms_{row}[k] \leftarrow height
18:
                        if k \neq |z| - 1 then
19:
                            curr_{row} \leftarrow rlpbwt[k+1].samples_{end}[|rlpbwt[k+1].samples_{end}|-1]
                            \begin{array}{l} curr_{index} \leftarrow height-1 \\ curr_{run} \leftarrow rank_h^{k+1}(curr_{index}) \end{array}
20:
21:
                            symb \leftarrow get\_symbol(start^{k+1}, curr_{run})
22:
23:
                   else if (curr_{run} \neq 0 \land curr_{run} = curr_{thr} \land \neg down) \lor curr_{run} = |samples_{bea}^{k}| - 1 then
24:
                        curr_{index} \leftarrow select_h^k(curr_{run})
25:
                        curr_{row} \leftarrow samples_{end}^{k}[curr_{run} - 1]
                        ms_{row}[k] \leftarrow curr_{row}

if k \neq |z| - 1 then
26:
27:
28:
                            (curr_{index}, curr_{run}, symb) \leftarrow UPDATE(k, curr_{index}, z)
29:
30:
                        curr_{index} \leftarrow select_h^k(curr_{run} + 1) + 1
                        curr_{row} \leftarrow samples_{beg}^{k}[curr_{run} + 1]
31:
                        ms_{row}[k] \leftarrow curr_{row}
if k \neq |z| - 1 then
32:
33:
                            (curr_{index},\ curr_{run},\ symb) \leftarrow UPDATE(k, curr_{index}, z)
\Rightarrow build matching statistics len
34:
35:
          for every k \in [0, |ms_{row}|) do
36:
               if ms_{row}[k] = height then
37:
                   ms_{len}[k] \leftarrow 0
38:
               else if k \neq 0 \land ms_{row}[i] = ms_{row}[i-1] \land ms_{len}[i-1] \neq 0 then
39:
                   ms_{len}[i] \leftarrow ms_{len}[i-1] + 1
40:
                                                                 \triangleright ra is a data structure for random access over the originale panel
41:
                   tmp_{index} \leftarrow i, \ tmp_{len} \leftarrow 0
42:
                   while tmp_{index} \ge 0 \land z[tmp_{index}] = ra(ms_{row}[k], tmp_{index}) do
43:
                        tmp_{index} \leftarrow tmp_{index} - 1, \ tmp_{len} \leftarrow tmp_{len} + 1
44:
                   ms_{len}[k] \leftarrow tmp_{len}
          for every k \in [0, |ms_{row}|) do
45:
                                                                                                    ▶ build matching statistics matches
46:
               if (ms_{len}[k] > 1 \land ms_{len}[k] \ge ms_{len}[k+1]) \lor (k=|z|-1 \land ms_{len}[k] \ne 0 then
         report match ending in k, with length ms_{len}[k], with at least row ms_{row}[k] in case extend the matches
47:
```

function DOWN(pos, prev, next)

using LCE queries or random access check the longest common prefix between pos and prev and between pos and next

if the latter is greater or equal return \top , else \bot

Algorithm 8 Algoritmo per match con matching-statistics (MS) e LCE

```
1: function MATCHES_MS_LCE(z)
         ms_{row} \leftarrow [0..0], \ ms_{len} \leftarrow [0..0]
                                                                                            \triangleright ms vectors with row and len of length |z|
<u>3</u>:
        curr_{row} \leftarrow rlpbwt[0].samples_{end}[|rlpbwt[0].samples_{end}|-1]
4:
         curr_{index} \leftarrow curr_{row}, \ curr_{run} \leftarrow rank_h^0(curr_{index})
5:
         symb \leftarrow get\_symbol(start^0, curr_{run})
                                                                                                     ▶ build matching statistics row
6:
7:
         for every k \in [0, |z|) do
             if z[i] = symb then
8:
                 ms_{row}[k] \leftarrow curr_{row}
9:
                 if k = 0 then
10:
                       ms_{len}[k] \leftarrow 1
11:
                  _{
m else}
12:
                       ms_{len}[k] \leftarrow ms_{len}[k-1] + 1
13:
                  if k \neq |z| - 1 then
14:
                       (curr_{index}, curr_{run}, symb) \leftarrow UPDATE(k, curr_{index}, z)
15:
              else
16:
                  if |samples_{beg}^k| = 1 then
17:
                       ms_{row}[k] \leftarrow height
18:
                       ms_{len}[k] \leftarrow 0
19:
                       if k \neq |z| - 1 then
20:
                           curr_{row} \leftarrow rlpbwt[k+1].samples_{end}[|rlpbwt[k+1].samples_{end}|-1]
21:
                           \begin{array}{l} curr_{index} \leftarrow height-1 \\ curr_{run} \leftarrow rank_h^{k+1}(curr_{index}) \end{array}
22:
                           symb \leftarrow get\_symbol(start^{k+1}, curr_{run})
23:
24:
                  else
25:
                       if curr_{run} = |samples_{beg}^k| - 1 then
26:
                           curr_{index} \leftarrow select_h^k(curr_{run}), \ prev_{row} \leftarrow samples_{end}^k[curr_{run} - 1]
27:
                           lce \leftarrow LCE(k, curr_{row}, prev_{row})
28:
                           ms_{row}[k] \leftarrow prev_{row}, \ curr_{row} \leftarrow prev_{row}
29:
                           if k = 0 then
30:
                               ms_{len}[k] \leftarrow 1
31:
                           _{
m else}
32:
                               ms_{len}[k] \leftarrow min(ms_{len}[k-1], lce_{len}) + 1
33:
                           if k \neq |z| - 1 then
                               (curr_{index},\ curr_{run},\ symb) \leftarrow UPDATE(k, curr_{index}, z)
34:
35:
                       else if curr_{run} = 0 then
36:
                           curr_{index} \leftarrow select_h^k(curr_{run} + 1) + 1, \ next_{row} \leftarrow samples_{beg}^k[curr_{run} + 1]
                           lce \leftarrow LCE(k, curr_{row}, next_{row})
37:
38:
                           ms_{row}[k] \leftarrow next_{row}, \ curr_{row} \leftarrow next_{row}
39:
                           if k = 0 then
40:
                               ms_{len}[k] \leftarrow 1
41:
                           else
42:
                               ms_{len}[k] \leftarrow min(ms_{len}[k-1], lce_{len}) + 1
43:
                           if k \neq |z| - 1 then
                               (curr_{index}, curr_{run}, symb) \leftarrow UPDATE(k, curr_{index}, z)
44:
45:
46:
                           prev_{row} \leftarrow samples_{end}^{k}[curr_{run} - 1], \ next_{row} \leftarrow samples_{beg}^{k}[curr_{run} + 1]
47:
                           lce \leftarrow \max_{len}(LCE(k, curr_{row}, prev_{row}), LCE(k, curr_{row}, next_{row}))
48:
                           curr_{row} \leftarrow lce_{row}
49:
                           ms_{row}[k] \leftarrow curr_{row}
50:
                           if k = 0 then
51:
                               ms_{len}[k] \leftarrow 1
52:
                           else
53:
                               ms_{len}[k] \leftarrow min(ms_{len}[k-1], lce_{len}) + 1
54:
                           if k \neq |z| - 1 then
55:
                               (curr_{index}, \ curr_{run}, \ symb) \leftarrow UPDATE(k, curr_{index}, z)
56:
          for every k \in [0, |ms_{row}|) do
                                                                                               ▶ build matching statistics matches
57:
              if (ms_{len}[k] > 1 \land ms_{len}[k] \ge ms_{len}[k+1]) \lor (k=|z|-1 \land ms_{len}[k] \ne 0 then
58:
                   report match ending in k, with length ms_{len}[k], with at least row ms_{row}[k]
```

Algorithm 9 Algoritmo per l'update usando le matching statistics

```
1: function UPDATE(k, curr_{index}, z)

2: curr_{index} \leftarrow lf(k, curr_{index}, z[k])

3: curr_{run} \leftarrow rank_h^{k+1}(curr_{index})

4: symb \leftarrow get\_symbol(start^{k+1}, curr_{run})

5: return (curr_{index}, curr_{run}, symb)
```

Algorithm 10 Algoritmo per la costruzione della struttura per φ e φ^{-1}

```
1: function Build_Phi(cols, panel, prefix)
                                                                                                   \triangleright prefix is the last prefix array
           \varphi \leftarrow [[0..0]..[0..0]], \ \varphi^{-1} \leftarrow [[0..0]..[0..0]]
                                                                                     \triangleright sparse bit vector panels for \varphi and \varphi^{-1}
 2:
           \varphi_{supp} = [], \ \varphi_{supp}^{-1} = []
                                                                                              \triangleright vectors for \varphi and \varphi^{-1} row values
 3:
           for every k \in [0, |cols|) do
 4:
                for every i \in [0, |samples_{beg}|) do
 5:
                     \varphi[sample_{beg}^{k}[i]][k] \leftarrow 1
 6:
                      if i = 0 then
 7:
 8:
                           push(\varphi_{supp}[sample_{beq}^{k}[i]], panel_{height})
 9:
                      else
                           push(\varphi_{supp}[sample_{beq}^{k}[i]], sample_{end}^{k}[i-1])
10:
                     \begin{split} \varphi^{-1}[sample_{end}^{k}[i]][k] \leftarrow 1 \\ \textbf{if } i = |sample_{beg}^{k}| - 1 \textbf{ then} \end{split}
11:
12:
13:
                           push(\varphi_{supp}^{-1}[sample_{end}^{k}[i]], panel_{height})
14:
                      else
                           push(\varphi_{supp}^{-1}[sample_{end}^{k}[i]], sample_{beq}^{k}[i+1])
15:
16:
           for every k \in [0, |prefix|) do
                 if \varphi[k][|\varphi[k]| - 1] = 0 then
17:
18:
                      \varphi[k][|\varphi[k]| - 1] \leftarrow 1
                      if k = 0 then
19:
                           push(\varphi_{supp}[prefix^k], panel_{height})
20:
21:
                           push(\varphi_{supp}[prefix^k], prefix^k[i-1])
22:
                 if \varphi^{-1}[k][|\varphi[k]| - 1] = 0 then
23:
                      \varphi^{-1}[k][|\varphi[k]| - 1] \leftarrow 1
24:
25:
                      if k = |prefix| - 1 then
                           push(\varphi_{supp}^{-1}[prefix^k], panel_{height})
26:
27:
                      else
                           push(\varphi_{supp}^{-1}[prefix^k], prefix^k[i+1])
28:
29:
           build rank/select for every sparse bitvector in \varphi and \varphi^{-1}
```

Algorithm 11 Algoritmi per le query a φ e φ^{-1}

```
1: function \varphi(prefix_{value}, col)
         res \leftarrow \varphi_{supp}^{prefix_{value}}[rank_{\varphi}^{prefix_{value}}(col)]
2:
3:
         if res = panel_{height} then
               return null
4:
         else
5:
               return res
6:
1: function \varphi^{-1}(prefix_{value}, col)
         res \leftarrow \varphi_{supp}^{-1} \stackrel{prefix_{value}}{=} [rank_{\varphi^{-1}}^{prefix_{value}}(col)]
2:
         if res = panel_{height} then
3:
4:
               return null
         else
5:
               return res
6:
```

Algorithm 12 Algoritmo per estendere un match in *col* usando φ , φ^{-1} e MS

```
1: function EXTEND_MATCHES(col, row, len)
         check_{down} \leftarrow \top, \ check_{up} \leftarrow \top
 2:
 3:
         while check_{down} do
              down_{row} \leftarrow \varphi^{-1}(row, col)
 4:
              if lce\_bounded(col, row, down_{row}, len) then
 5:
 6:
                   push(haplos, down_{row})
                  row \leftarrow down_{row}
 7:
              else
 8:
 9:
                   check_{down} \leftarrow \bot
         while up_{down} do
10:
              up_{row} \leftarrow \varphi(row, col)
11:
              if lce\_bounded(col, row, up_{row}, len) then
12:
                  push(haplos, up_{row})
13:
                  row \leftarrow up_{row}
14:
              else
15:
                   check_{up} \leftarrow \bot
16:
         return haplos
17:
```