Fast filtering algorithms

This exposition is based on the following sources, which are all recommended reading:

1. Flexible Pattern Matching in Strings, Navarro, Raffinot, 2002, chapter 6.5, pages 162ff.

2. Burkhardt et al.: q-gram Based Database Searching Using a Suffix Array (QUASAR), RECOMB 99

We will present the hierarchical filtering approach of Navarro and Baeza-Yates and and the simple QUASAR idea.

Filtering algorithms

The idea behind filtering algorithms is that it might be easier to check that a text position does *not* match a pattern string that to verify that it does.

Filtering algorithms *filter out* portions of the text that cannot possibly contain a match, and, at the same time, find positions that can possibly match.

These potential match positions then need to be *verified* with another algorithm like for example the bit-parallel algorithm of Myers (BPM).

Filtering algorithms (2)

Filtering algorithms are very sensitive to the *error level* $\alpha := k/m$ since this normally affects the amount of text that can be discarded from further consideration. (m = pattern length, k = errors.)

If most of the text has to be verified, the additional filtering steps are an overhead compared to the strategy of just verifying the pattern in the first place.

On the other hand, if large portions of the text can be discarded quickly, then the filtering results in a faster search.

Filtering algorithms can improve the average-case performance (sometimes dramatically), but not the worst-case performance.

PEX

The pidgeonhole principle

The idea behind the presented filtering algorithm is very easy. Assume that we want to find all occurrences of a pattern $P = p_1, \ldots, p_m$ in a text $T = t_1, \ldots, t_n$ that have an edit distance of at most k.

If we *divide* the pattern into $k + 1$ pieces $P = p^1, \ldots, p^{k+1}$, then at least *one* of the pattern pieces has to match *without error*.

The pidgeonhole principle (2)

There is a more general version of this principle first formalized by Myers in 1994:

Lemma 1. *Let Occ match P with k errors, $P = p^1, \ldots, p^j$ be a concatenation of subpatterns, and $a_1, \ldots, a_j$ be nonnegative integers such that $A = \sum_{i=1}^{j} a_i$. Then, for some $i \in 1, \ldots, j$, Occ includes a substring that matches p^i with $\lfloor a_i k / A \rfloor$ errors.*

Proof: *Exercise.*

The pidgeonhole principle (3)

So the basic procedure is:

1. *Divide:* Divide the pattern into $k+1$ pieces of approximately the same length.

2. *Search:* Search all the pieces simultaneously with a multi-pattern string matching algorithm. According to the above lemma, each possible occurrence will match at least one of the pattern pieces.

3. *Verify:* For each found pattern piece, check the neighborhood with a verification algorithm that is able to detect an occurrence of the whole pattern with edit distance at most k. Since we allow indels, if $p_{i_1} \ldots p_{i_2}$ matches the text $t_j \ldots t_{j+i_2-i_1}$, then the verification has to consider the text area $t_{j-(i_1-1)-k} \ldots t_{j+(m-i_1)+k}$, which is of length $m+2k$.

An example

Say we want to find the pattern *annual* in the texts

t_1 = *any_annealing* and

t_2 = *an_unusual_example_with_numerous_verifications*

with at most 2 errors.

An example (2)

1. *Divide:* We divide the pattern *annual* into $p^1 = an$, $p^2 = nu$, and $p^3 = al$. One of these subpattern has to match with 0 errors.

2. *Search:* We search for all subpatterns:

```
1: searching for an:  in t_1:  find positions 1, 5
                      in t_2:  find position  1
2: searching for nu:  in t_1:  find no positions
                      in t_2:  find positions 5, 25
3: searching for al:  in t_1:  find position 9
                      in t_2:  find position 9
```

3. *Verification:* We have to verify 3 positions in t_1, and 4 positions in t_2, to find 3 occurrences at positions (indexed by the last character) 9, 10, 11 in t_1 and *none* in t_2.

Hierarchical verification

The toy example makes clear that *many* verifications can be triggered that are unsuccesssful and that many subpatterns can trigger the *same* verification. Repeated verfications can be avoided by carefully sorting the occurrences of the pattern (exercise).

It was shown by Baeza-Yates and Navarro that the running time is dominated by the multipattern search for error levels $\alpha = k/m$ below $1/(3\log_{|\Sigma|} m)$. In this region, the search cost is about $O(kn\frac{\log_{|\Sigma|} m}{m})$. For higher error levels, the cost for verifications starts to dominate, and the filter efficiency deteriorates abruptly.

Baeza-Yates and Navarro introduced the idea of hierarchical verification to reduce the verification costs, which we will explain next. Then we will work out more details of the three steps.

Hierarchical verification (2)

Navarro and Baeza-Yates use Lemma **??** for a *hierarchical verification*. The idea is that, since the verification cost is high, we pay too much for verifying the *whole* pattern *each* time a small piece matches. We could possibly reject the occurrence with a cheaper test for a shorter pattern.

So, instead of directly dividing the pattern into $k + 1$ pieces, we do it hierarchically. We split the pattern first in two pieces and search for each piece with $\lfloor k/2 \rfloor$ errors, following Lemma **??**. The halves are then recursively split and searched until the error rate reaches zero, i. e. we can search for exact matches.

With hierarchical verification the area of applicability of the filtering algorithm grows to $\alpha < 1/\log_{|\Sigma|} m$, an error level three times as high as for the naive paritioning and verification. In practice, the filtering algorithm pays off for $\alpha < 1/3$ for medium long patterns.

Hierarchical verification (3)

Example. Say we want to find the pattern P = `aaabbbcccddd` in the text T = `xxxbbbxxxxxx` with at most $k = 3$ differences. The pattern is split into four pieces p^1 = `aaa`, p^2 = `bbb`, p^3 = `ccc`, p^4 = `ddd`. We search with $k = 0$ errors in level 2 and find *bbb*.

```
level 0              aaabbbcccddd              with k=3 errors
                     /          \
level 1          aaabbb          cccddd        with k=1 errors
                 /    \          /    \
level 2        aaa    bbb      ccc    ddd      with k=0 errors
```

Hierarchical verification (4)

Now instead of verifying the complete pattern in the complete text (at level 0) with $k = 3$ errors, we only have to check a slightly bigger pattern (`aaabbb`) at level 1 with one error. This is much cheaper. In this example we can decide that the occurrence `bbb` cannot be extended to a match.

```
level 0              aaabbbcccddd              with k=3 errors
                     /          \
level 1          AAABBB          cccddd        with k=1 errors
                 /    \          /    \
level 2        aaa    BBB      ccc    ddd      with k=0 errors
```

The PEX algorithm

Divide: Split pattern into $k+1$ pieces, such that each piece has equal probability of occurring in the text. If no other information is available, the uniform distribution is assumed and hence the pattern is divided in pieces of equal length.

The PEX algorithm (2)

Build Tree: Build a tree of the pattern for the hierarchical verification. If $k + 1$ is not a power of 2, we try to keep the binary tree as balanced as possible.

Each node has two members *from* and *to* indicating the first and the last position of the pattern piece represented by it. The member *err* holds the number of allowed errors. A pointer *myParent* leads to its parent in the tree. (There are no child pointers, since we traverse the tree only from the leafs to the root.) An internal variable *left* holds the number of pattern pieces in the left subtree. *idx* is the next leaf index to assign. *plen* is the length of a pattern piece.

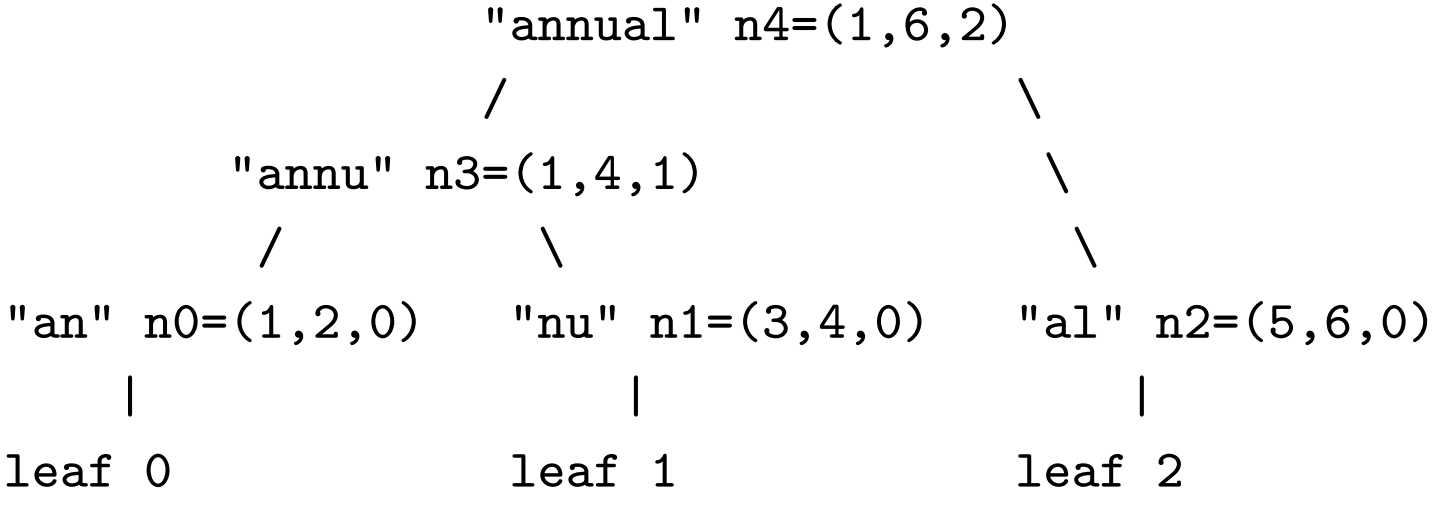
Algorithm CreateTree generates a hierarchical verification tree for a single pattern. (Lines **??** and **??** are justified by Lemma **??**.)

The PEX algorithm (3)

```
1  CreateTree( p = p_i p_{i+1} ... p_j,  k,  myParent,  idx,  plen )
2  // Note: the initial call is: CreateTree ( p,  k,  nil,  0,  ⌊m/(k+1)⌋ )
3  Create new node node
4  from(node) = i
5  to(node) = j
6  left = ⌈(k+1)/2⌉
7  parent(node) = myParent
8  err(node) = k
9  if  k = 0
10    then  leaf_idx = node
11    else
12          lk = ⌊(left · k)/(k+1)⌋
13          CreateTree( p_i ... p_{i+left·plen−1},  lk,  node,  idx,  plen )
14          rk = ⌊((k+1−left) · k)/(k+1)⌋
15          CreateTree( p_{i+left·plen} ... p_j,  rk,  node,  idx+left,  plen )
16 fi
```

The PEX algorithm (4)

Example: Find the pattern P = `annual` in the text T = `annual_CPM_anniversary` with at most $k = 2$ errors. First we build the tree with $k + 1 = 3$ leaves. Below we write at each node n_i the variables (*from*, *to*, *error*) .



The PEX algorithm (5)

Search: After constructing the tree, we have $k+1$ leafs $leaf_i$. The $k+1$ subpatterns

$$\{\ p_{from(n)}, \ldots, p_{to(n)},\ n = leaf_i,\ i \in \{0, \ldots, k\}\ \}$$

are sent as input to a multi-pattern search algorithm (e. g. Aho-Corasick, Wu-Manbers, or SBOM). This algorithm gives as output a list of pairs (pos, i) where pos is the text position that matched and i is the number of the piece that matched.

The PEX algorithm performs verifications on its way upward in the tree, checking the presence of longer and longer pieces of the pattern, as specified by the nodes.

The PEX algorithm (6)

```
1  Search phase of algorithm PEX
2  for (pos, i) ∈ output of multi-pattern search do
3      n = leaf_i; in = from(n); n = parent(n);
4      cand = true;
5      while cand = true and n ≠ nil do
6              p_1 = pos − (in − from(n)) − err(n);
7              p_2 = pos + (to(n) − in) + err(n);
8              verify text t_{p_1} ... t_{p_2} for pattern piece p_{from(n)} ... p_{to(n)}
9                  allowing err(n) errors;
10             if pattern piece was not found
11               then cand = false;
12               else n = parent(n);
13             fi
14     od
15     if cand = true
16       then report the positions where the whole p was found;
17     fi
18 od
```

The PEX algorithm (7)

We search for `annual` in `annual_CPM_anniversary`. We constructed the tree for `annual`. A multi-pattern search algorithm finds: $(1,1)$, $(12,1)$, $(3,2)$, $(5,3)$. (Note that leaf i corresponds to pattern p^{i+1}). For each of these positions we do the hierarchical verification:

```
Initialization for (1,1);
n=n0; in=1; n=n3; cand=true;
While loop;
  a) p1=1-(1-1)-1=0;  p2=1+(4-1)+1=5;
  verify pattern annu in text annua with 1 error => found !
  b) p1=1-(1-1)-2=-1; p2=1+(6-1)+2=8;
  verify pattern annual in text annual_C => found !
  c) report end positions (6,7,8)
```

The PEX algorithm (8)

```
Initialization for (3,2);
n=n1; in=3; n=n3; cand=true;
While loop;
  a) p1=3-(3-1)-1=0;  p2=3+(4-3)+1=5;
  verify pattern annu in text annua with 1 error => found !
  b) p1=3-(3-1)-2=-1; p2=3+(6-3)+2=8;
  verify pattern annual in text annual_C => found !
  c) report end positions (6,7,8)
```

The PEX algorithm (9)

```
Initialization for (12,1);
n=n0; in=1; n=n3; cand=true;
While loop;
  a)  p1=12-(1-1)-1=11; p2=12+(4-1)+1=16;
  verify pattern annu in text _anniv with 1 error => found !
  b)  p1=12-(1-1)-2=10; p2=12+(6-1)+2=19;
  verify pattern annual in text M_annivers => NOT found !
```

Summary

- Filtering algorithms prevent a large portion of the text from being looked at.
- The larger $\alpha = k/m$, the less efficient filtering algorithms become.
- Filtering algorithms based on the pidgeonhole principle need an exact, multi-pattern search algorithm and a verification capable approximate string matching algorithm.
- The PEX algorithm starts verification from short exact matches and considers longer and longer substrings of the pattern as the verification proceeds upward in the tree.

QUASAR - q-gram based database searching

This exposition has been developed by Knut Reinert. It is based on the following sources, which are all recommended reading:

1. Burkhardt et al. (1999) *q-gram Based Database Searching Using a Suffix Array (QUASAR)*, Proc. RECOMB 99.

2. Burkhardt and Kärkkäinen (2001) *Better Filtering with Gapped q-grams*, Proc. CPM 01.

The tool *QUASAR* aims at aligning a query $S = s_1, \ldots, s_m$ in a text, also called database $D = d_1, \ldots, d_n$. It can be seen as an efficient filter that uses exact matches. In contrast to online filtering algorithms, QUASAR uses a suffix array as indexing structure for the database.

Quasar

Quasar

QUASAR, or "Q-gram Alignment based on Suffix ARrays", is a filtering approach. QUASAR finds all *local* approximate matches of a *query sequence* S in a *database* $D = \{d, ...\}$. The verification is performed by other means.

Definition. A sequence d is *locally similar* to S, if there exists at least one pair $(S_{i,i+w-1}, d')$ of substrings such that:

1. $S_{i,i+w-1}$ is a substring of length w and d' is a substring of D, and
2. the substrings d' and $S_{i,i+w-1}$ have edit distance at most k.

We call this the *approximate matching problem with k differences and window length w*.

For simplicity, we assume that the database consists of only one sequence, i. e. $D = \{d\}$.

The q-gram lemma

A short subsequence of length q is called a *q-gram*. In the following we start by considering the first w letters of S. The algorithm uses the following lemma:

Lemma 2. *Let P and S be strings of length w with at most k differences. Then P and S share at least $w + 1 - (k + 1)q$ common q-grams.*

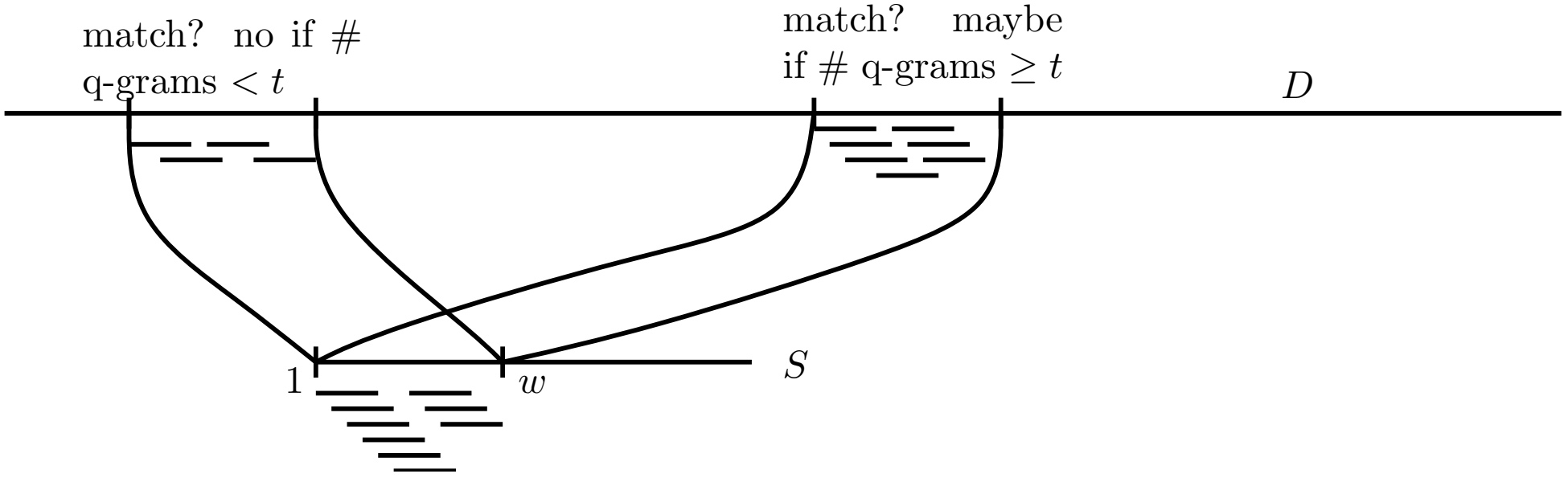
In our case, this means:

Lemma 3. *Let an occurrence of $S_{1,w}$ with at most k differences end at position j in D. Then at least $w + 1 - (k + 1)q$ of the q-grams in $S_{1,w}$ occur in the substring $D_{j-w+1,j}$.*

Proof: Exercise. . . .

That means that as a necessary condition for an approximate match, at least $t = w + 1 - (k + 1)q$ of the q-grams contained in $S_{1,w}$ occur in a substring of D with length w. For example the strings ACAGCTTA and ACACCTTA have $8 + 1 - (1 + 1)3 = 3$ common 3-grams, namely ACA, CTT and TTA.

The q-gram lemma (2)



q-gram index

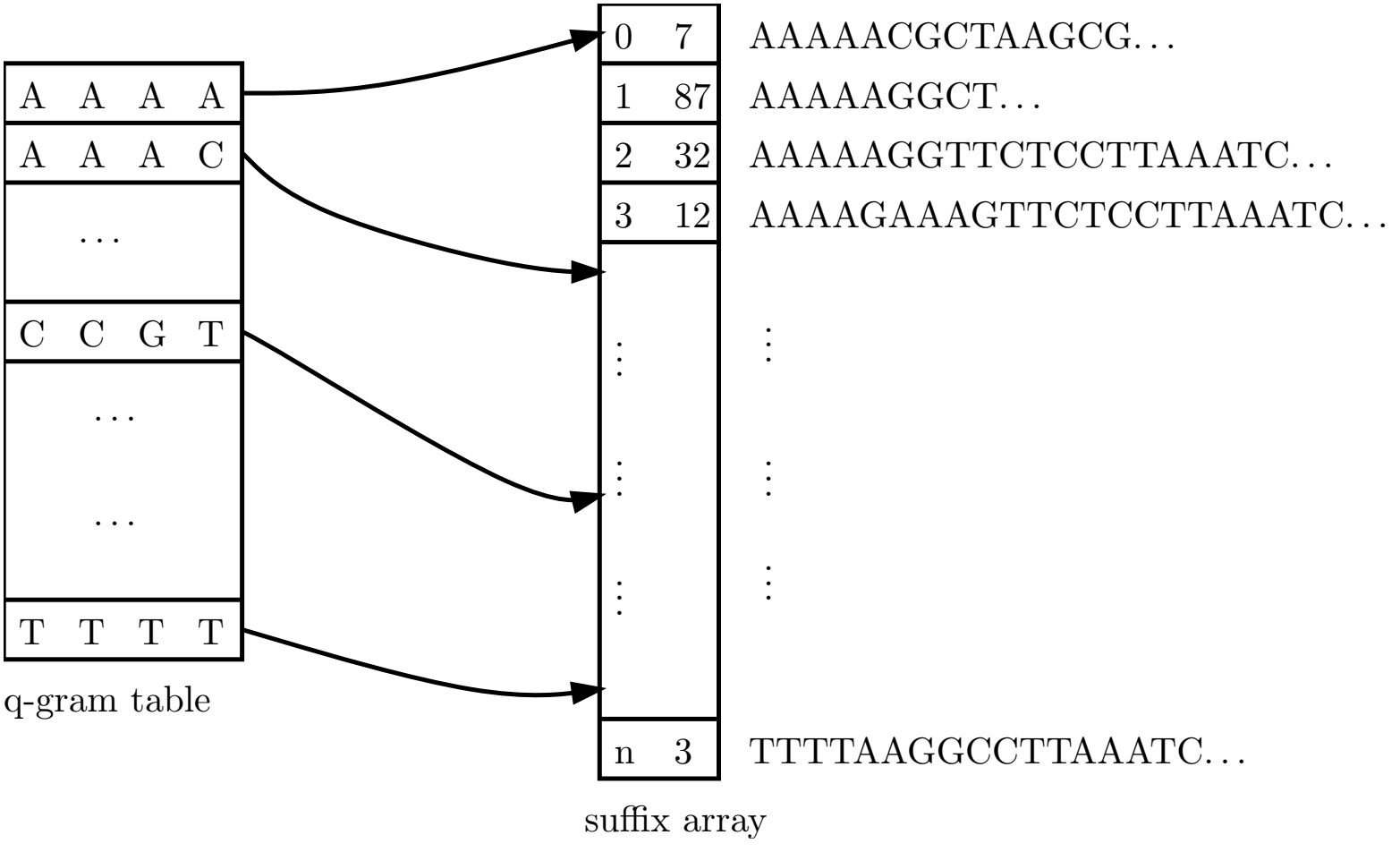
The algorithm builds in a first step an indexing structure as follows:

1. Build a suffix array A over D.

2. Given q, compute for all possible $|\Sigma|^q$ q-grams the start position of the hitlist. This allows to lookup a q-gram in constant time.

3. If another q is specified, A is used to recompute the above table.

q-gram index (2)



Counting q-grams

Now we have to find all approximate matches between $S_{1,w}$ and D, that means we have to find all substrings in D that share at least t q-grams with $S_{1,w}$. The algorithm proceeds in the following basic steps on which we will elaborate:

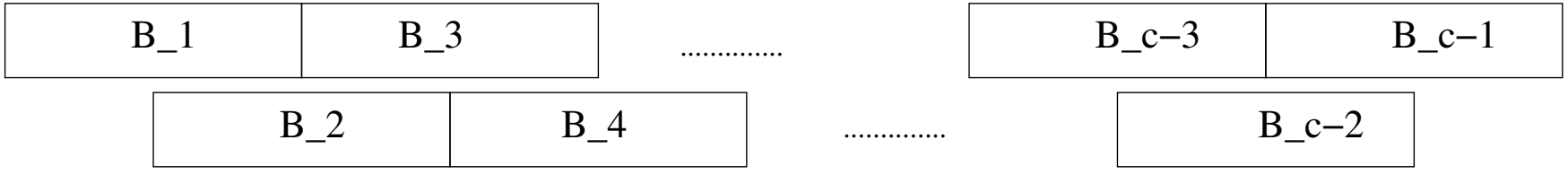
1. Define two arrays of *non-overlapping* blocks of size $b \geq 2w$. The first array is shifted by $b/2$ against the other.

2. Process all q-grams in $S_{1,w}$ and increment the counters of the corresponding blocks.

3. All blocks containing approximate matches will have a counter of at least t. (The reverse is not true).

4. Shift the search window by one. Now we consider $S_{2,w+1}$.

Blocking



Since we want to count the q-grams that are in common between the query and the database, we use counters. Ideally we would use a counter of size w for each substring of this size. Since this uses too much memory, we build larger, non-overlapping blocks. While this decreases the memory usage, it also decreases the specificity.

Since the blocks are not overlapping we might miss q-grams that cross the block boundary. As a remedy, we use a second, shifted array of blocks.

Window Shifting

We started the search for approximate matches of window length w with the first w-mer in S, namely $S_{1,w}$. In order to determine the approximate matches for the next window $S_{2,w+1}$, we only have to discard the old q-gram $S_{1,q}$ and consider the new q-gram $S_{w-q+2,w+1}$.

To do that we decrement the counters of all blocks that contain $S_{1,q}$ that have not reached the threshold t. However, if the counter has already reached t it stays at this value to indicate a match for the extension phase.

For the new block we use the precomputed index and the suffix array to find the occurrences of the new q-gram and increment the corresponding block counters (at most two).

Alignment

After having computed the list of blocks, QUASAR uses BLAST to actually search the blocks. Here are some results from the inital implementation. QUASAR was run with $w = 50$, $q = 11$, and t such that windows with at most 6% differences are found. Reasonable values for the block size are 512 to 4096.

DB size	query	id. res.	filtr. ratio	QUASAR	BLAST
73.5 Mb	368	91.4%	0.24%	0.123 s	3.27 s
280 Mb	393	97.1%	0.17%	0.38 s	13.27 s

"A database in BLAST format is built in main memory which is then passed to the BLAST search engine. The construction of this database requires a significant amount of time and introduces unnecessary overhead."

Gapped q-grams

Gapped q-grams

In order to achieve a high filtration rate, we would like to choose q as large as possible, since the number of hits decreases exponentially in q. On the other hand, the threshold $t = w - q - qk + 1$ also decreases with increasing q thereby reducing the filtering efficiency. The question is whether we could increase the length of the q-grams somehow, such that the threshold t stays high.

This can indeed be achieved by using *gapped q-grams*. For example the 3-grams with the *shape* `##.#` in the string `ACAGCT` are `AC.G`, `CA.C`, and `AG.T`:

```
 ACAGCT
------
AC G
 CA C
  AG T
```

Next we define the concept formally.

Gapped q-grams (2)

Definition 4.

- A *shape* Q is a set of non-negative integers containing 0.
- The *size* of Q, denoted by $|Q|$, is the cardinality of the set.
- The *span* of Q is $s(Q) = \max Q + 1$.
- A shape of size q and span s is called *(q, s)-shape*.
- For any integer i and shape Q, the *positioned shape Q_i* is the set $\{i + j \mid j \in Q\}$.
- Let $Q_i = \{i_1, i_2, \ldots, i_q\}$, where $i = i_1 < i_2 < i_3 < \cdots < i_q$, and let $S = s_1 s_2 \ldots s_m$ be a string. For $1 \leq i \leq m - s(Q) + 1$, the *Q-gram at position i in S*, denoted by $S[Q_i]$, is the string $s_{i_1} s_{i_2} \ldots s_{i_q}$.
- Two strings P and S *have a common Q-gram at position i* if $P[Q_i] = S[Q_i]$.

Gapped q-grams (3)

Example 5. Let $Q = \{0, 1, 3, 6\}$ be a shape. Using the graphical representation it is the shape `##.#..#`. Its size is $|Q| = 4$ and its span is $s(Q) = 7$. The string *ACGGATTAC* has three Q-grams: $S[Q_1] = s_1 s_2 s_4 s_7 = ACGT$, $S[Q_2] = CGAA$, and $S[Q_3] = GGTC$.

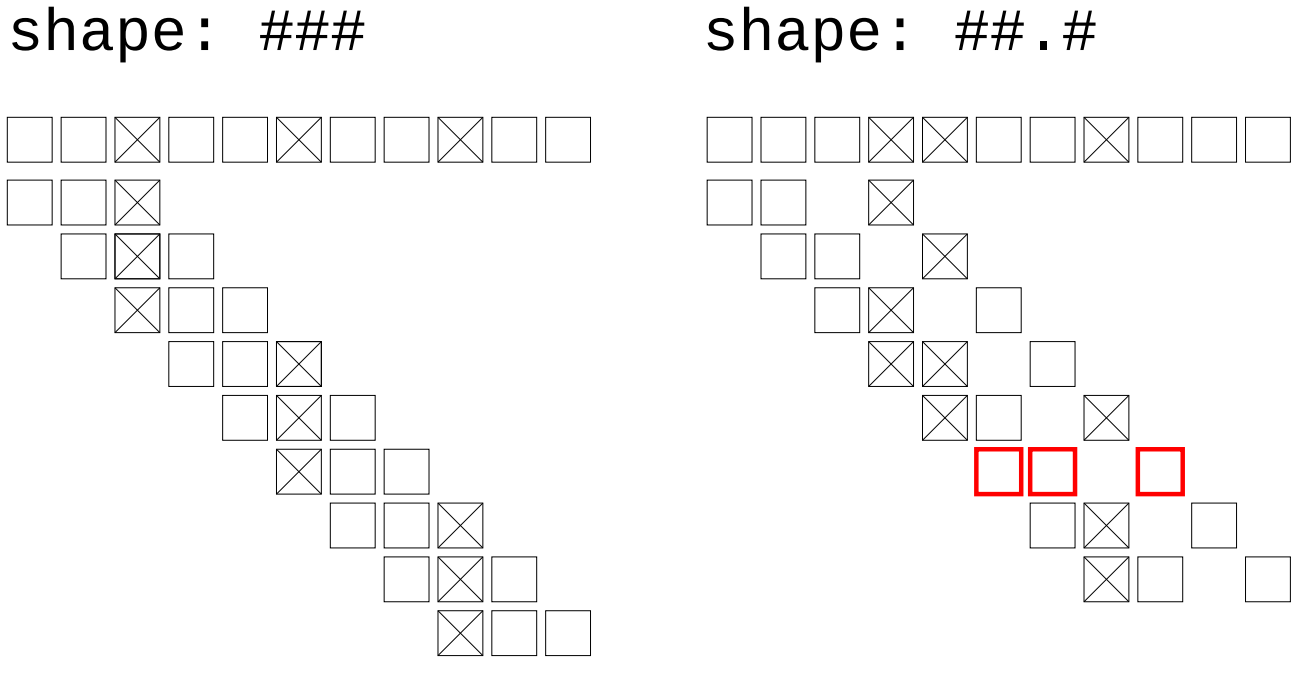
The q-gram lemma can be extended for gapped q-grams. A generalization gives

$$t = w - s(Q) - |Q|k + 1.$$

However it is not tight anymore (we will prove this).

Gapped q-grams (4)

Example 6. Let $w = 11$ and $k = 3$ and consider the 3-shapes `###` and `##.#`. The above threshold for the two shapes is $0 = 11 - 3 \cdot 4 + 1$ and $-1 = 11 - 4 - 3 \cdot 3 + 1$ respectively. Thus neither shape would be useful for filtering. However, the real threshold for `##.#` is 1. This can be checked by a full enumeration of all combinations of 3 mismatches.



Worst-case mismatch positions

New threshold

What is the (tight) threshold for arbitrary Q-shapes?

Let $P = p_1, \ldots, p_w$ and $S = s_1, \ldots, s_w$ be two strings of length w. Let $R(P, S)$ be the set of positions where S and P do not match. Then $|R(S, P)|$ is the Hamming distance of P and S.

To determine the common Q-grams of P and S only the mismatch set is needed: It holds that

$$P[Q_i] = S[Q_i] \quad \text{if and only if} \quad Q_i \cap R(P, S) = \emptyset.$$

Using this notation we can define the threshold of a shape Q for a pattern of length w and Hamming distance k as:

$$t(Q, w, k) := \min_{R \subseteq \{1,\ldots,w\}, |R|=k} \left| \{ i \in \{1, \ldots, w - s(Q) + 1\} \mid Q_i \cap R = \emptyset \} \right|$$

New threshold (2)

From the above discussion we get the following tight form of the q-gram lemma for arbitrary shapes:

Lemma 7. *Let Q be a shape. For any two strings P and S of length w with Hamming distance k, the number of common Q-grams of P and S is at least $t(Q, w, k)$. Furthermore, there exist two strings P and S of length w and Hamming distance k, for which the number of common Q-grams is exactly $t(Q, w, k)$.*

New threshold (3)

It is easy to see that this bound is as least as tight as the lower bound we already introduced:

Lemma 8.

$$t(Q, w, k) \geq \max\{0, w - s(Q) - |Q|k + 1\}$$

Proof: Let R be the set minimizing the expression in the definition of $t(Q, w, k)$. For each $j \in R$ there are exactly $|Q|$ integers i such that $j \in Q_i$. Therefore, at most $k|Q|$ of the positioned shapes Q_i, $i \in \{1, \ldots, w-s(Q)+1\}$, intersect with R, and at least $w - s(Q) - k|Q| + 1$ do not intersect with R.

New threshold (4)

T The above lemma gives indeed the exact threshold for ungapped q-grams.

Lemma 9. *Let Q be a contiguous shape,i. e., $Q = \{0, \ldots, q-1\}$. Then*

$$t(Q, w, k) = \max\{0, w - s(Q) - |Q|k + 1\} = \max\{0, w - q(k+1) + 1\}.$$

Proof: The lower bound is shown by Lemma **??**. For the upper bound we choose $R = \{q, 2q, \ldots, kq\}$. Then Q_i intersects with R if and only if $i \in \{1, \ldots, kq\}$, and thus does not intersect with R if $i \in \{kq+1, \ldots, w-q+1\}$. Hence for this R we have only $w-q+1-kq-1+1 = w-(k+1)q+1$ common q-grams.

New threshold (5)

The following table gives the exact thresholds for all shapes for $w = 50$ and $k = 5$. One can see that in many cases, especially for higher values of q, best gapped shapes have higher thresholds than contiguous shapes of the same or even smaller size.

$s\downarrow$: $q\rightarrow$	4	5	6	7	8	9	10
5	26	21	–	–	–	–	–
6	25	20	15	–	–	–	–
7	24	19	14	9	–	–	–
8	23	18	13	8	3	–	–
9	22	**18** $>$ 17	**14** $>$ 12	**9** $>$ 7	**5** $>$ 2	0	–
10	21	**18** $>$ 16	**13** $>$ 11	**10** $>$ 6	**6** $>$ 1	**3** $>$ 0	0
11	20	**16** $>$ 15	**13** $>$ 10	**10** $>$ 5	**7** $>$ 0	**4** $>$ 0	**2** $>$ 0
12	19	**16** $>$ 14	**12** $>$ 9	**9** $>$ 4	**7** $>$ 0	**4** $>$ 0	**2** $>$ 0

New threshold (6)

It has to be noted that it does not suffice to put in gaps somewhere; the gaps have to be choosen carefully. For example in the above table ($w = 50$, $k = 5$, and $q = 12$) there are only two shapes with a positive threshold, namely `###.#..###.#..###.#` and `#.#.#...#.....#.#.#...#.....#.#.#...#` and their mirror images.

Minimum coverage

The filtering efficiency of a Q-gram clearly depends on the threshold $t(Q, w, k)$. However there is also another factor that influences it. This factor is called *minimum coverage*.

Before we define it formally lets have a look at an example.

Minimum coverage (2)

Example 10. Let $w = 13$ and $k = 3$. Then both shapes `###` and `##.#` have a threshold of two. If two strings have four consecutive characters then they have two common 3-grams of shape `###`. In contrast, in order to have two common 3-grams of shape `##.#`, two strings need at least 5 matching characters.

This means, that the gapped 3-gram would have a lower count of common q-grams on strings that have only four consecutively matching characters although it has the same threshold.

Minimum coverage (3)

Definition 11. Let Q be a shape and t be a non-negative integer. The *minimum coverage* of Q for threshold t is:

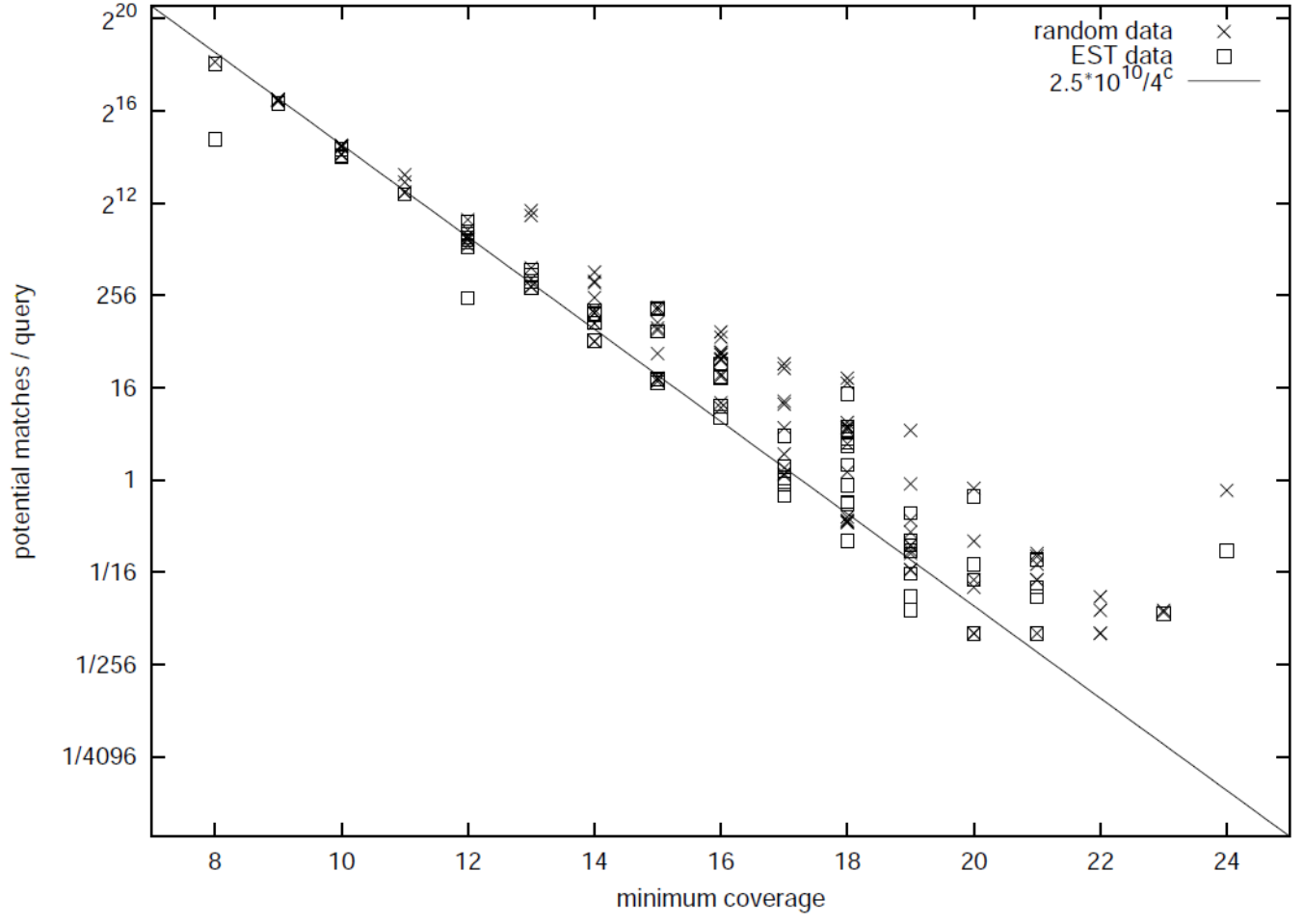
$$c(Q, t) = \min_{C \subset \mathbb{N}, |C|=t} | \cup_{i \in C} Q_i | .$$

Hence the minimum coverage is the minimum number of characters that need to match between a pattern and a text substring for there to be t matching Q-grams.

Whenever possible, gapped Quasar chooses the highest minimum coverage, since it makes it more unlikely that a random string matches t Q-grams. This improves the filter efficency.

Minimum coverage (4)

Computational experiments indicate that there is a strong correlation between the minimum coverage $c(Q, t(Q, m, k))$ and the filter efficiency.



Correlation between expected and actual number of potential matches.

Best shapes (5)

The following table shows different shapes for $k = 5$. The column *best* shows the shape with the highest minimum coverage (ties are broken using the threshold). The column *median* shows the median shape ordered by minimum coverage. If one chooses a random shape, the chance is 50% to be better (or worse) than this one. The last column show the best *one-gapped* shape. (The details of the tie breaking used here can be read in the paper.)

q	best	median	1-gapped
6	##.....#..#..#.#	#.###.....#.#	#####....#
7	#.##.....##..#.#	##..#..#..##	#####...##
8	##..#.#.............##..#.#	#.#..####...#.#	######...##
9	###..#..#.#...#.##	######..#.#.#	########..##
10	###..#..#.#..###.#	##.##..#.#.###.#	########

Index structure

It is not necessary to use a suffix array for ungapped q-grams, and it is not possible anymore to use a suffix array for the gapped Q-grams. Instead, the database is scanned twice. The first time the number of occurrences of all Q-grams is counted.

In the second scan, the positions at which a q-gram starts are recorded in an array of size n. During that scan, the index points to the start of the respective list.

The detail shall be worked out as an *exercise*.

Extension to Levenshtein distance

Note that the q-gram method presented so far can only be used to find local approximate matches with the *Hamming* distance.

The q-gram method can be generalized to the *Levenshtein* distance. Burkhardt and Kärkkäinen have described an extension that uses 'one-gapped q-grams'.

The idea is to model insertions and deletions by additional Q-grams. For example, with the basic shape `##-#` applied the text, we would use `##-#`, `##--#`, and `###` for the pattern.

The filter then compares all three shapes in the pattern to the q-grams of the basic shape in the text. Thus matching q-grams are even found in the presence of indels.

Otherwise he algorithm stays essentially unchanged, except that the threshold computation is slightly different.

Summary

- Filtering based on q-grams using a suffix array with an index is an efficient filtering method.
- In the gapless case, filtering efficiencies of $\approx 0.2\%$ were observed for genomic sequences.
- Gapped Q-grams improve the filtering efficency further (by orders of magnitude).
- The threshold t and the minimum coverage both influence the filter efficency.
- No closed formula is known for computing t for gapped Q-grams.

Q-gram filters for ε-matches

This exposition was developed by Clemens Gröpl. It is based on:

- Kim R. Rasmussen, Jens Stoye, Eugene W. Myers: *Efficient q-Gram Filters for Finding All ϵ-Matches over a Given Length*, Journal of Computational Biology, Volume 13, Number 2, 2006, pages 296–308. (Originally presented at GCB 2004 and RECOMB 2005.) [RSM06]

Motivation

Comparison of large genomic sequences can be speeded up a lot if *filtering techniques* are applied. The key observation is that a local alignment of high sequence similarity must contain at least a few short exact matches.

The idea of using q-grams for fast filtering is not new. A q-gram is a substring of length q. Programs like BLAST use q-grams which occur in both sequences as *seeds* for a local alignment search.

It has also been observed that combining the idea of seeds with a combinatorial argumentation based on some form of the *pigeon hole principle* can be used to discard large parts of the input sequences from further consideration, because they cannot contain a good local alignment.

Motivation (2)

We can distinguish three kinds of algorithms.

When applied for finding highly similar regions, the classical *exact* algorithms (e. g. Smith-Waterman) will spend most of the time verifying that there is no match between a given pair of regions. The running times (typically the product of sequence lengths) are infeasible for genome size sequences.

Heuristics like BLAST typically employ a q-gram index to locate seeds and perform a verification for the candidate regions located in this way. However, BLAST might fail to recognize an existing match, unless the filtering parameters are set very stringent. Thus one has to trade off sensitivity against speed.

A *filter* is an algorithm that allows us to discard large parts of the input, but is guaranteed not to loose *any* significant match. The trade-off to be considered for filtering algorithms is thus only whether the additional effort is payed off by the saving of time spent for verifications.

Motivation (3)

In this lecture, we will consider the problem of finding matches of low *error rate* ε and a given *minimum length* n_0.

The cost measure will be the *edit distance* (Levenshtein distance). That is, the distance between two strings is the number of insertions, deletions, and substitutions needed to transform one into the other.

The SWIFT algorithm is an improvement of the QUASAR algorithm by Burkhardt et. al.. Note, however, that QUASAR uses an absolute error threshold rather than an error rate. Using an error rate is more appropriate since the length of a local alignment is not known in advance.

The filter has been successfully applied for the *fragment overlap* computation in sequence assembly and for *BLAST-like searching* in EST sequences.

Definitions

As usual, let A and B denote strings over a finite alphabet Σ, let $|A|$ be the length of A, let $A[i]$ be the i-th letter of A, and let $A[p..q]$ be the substring starting at position p and ending with position q of A, thus $A[i..i]$ consists of the letter $A[i]$. A substring of length $q > 0$ of A is a *q-gram* of A.

The *(unit cost) edit distance* between strings A and B is the minimum number of edit operations (insertion, deletion, substitution) in an alignment of A and B. It is denoted by $\mathrm{dist}(A, B)$.

The edit distance can be computed by the well-known Needleman-Wunsch algorithm. It computes in $O(|A||B|)$ time an *edit matrix* $E(i, j) := \mathrm{dist}(A[1..i], B[1..j])$. The letter $A[i]$ corresponds to the step from row $i - 1$ to i, so it is natural to visualize the letters *between* the rows and columns of the edit matrix, etc..

An *ε-match* is a local alignment for substrings (α, β) with an *error rate* of at most ε. That is, $\mathrm{dist}(\alpha, \beta) \leq \varepsilon|\beta|$. (Note the 'asymmetry' in the definition of error rate.)

Definitions (2)

The problem can now be formally stated as follows:

> Given a *target* string A and a *query* string B, a *minimum match length* n_0 and a *maximum error rate* $\varepsilon > 0$;
>
> Find all ε-matches (α, β) where α and β are substrings of A and B, respectively, such that
>
> 1. $|\beta| \geq n_0$ and
> 2. $\text{dist}(\alpha, \beta) \leq \lfloor \varepsilon |\beta| \rfloor$.

q-gram filters for ε-matches

A *q-hit* is a pair (i, j) of indices such that $A[i..i + q - 1] = B[j..j + q - 1]$.

The basic idea of the q-gram method is as follows:

1. Find (enumerate) all q-hits between the query and the target strings.
2. Identify regions (in the Cartesian product of the strings) that have "enough" hits.
3. Such candidate regions are then subjected to a closer examination.

The concrete methods differ in the shape and the size of the regions.

q-gram filters for ε-matches (2)

The following lemma relates ε-matches (α, β) to *parallelograms* of the edit matrix. For a moment, we assume that the length of β is known, so that we can work with an absolute bound on the distance.

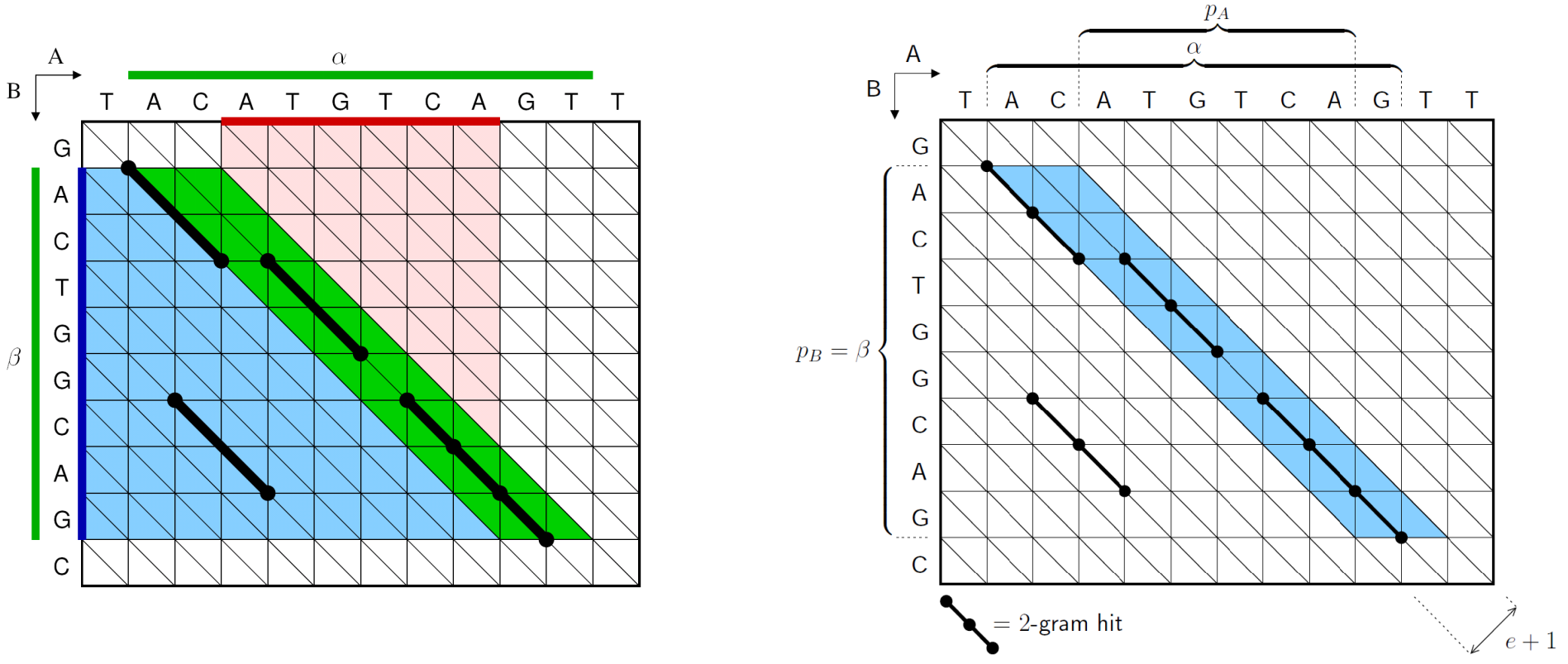
An $n \times e$ *parallelogram* of the edit matrix consists of entries from $n+1$ consecutive rows and $e+1$ consecutive diagonals.

Lemma 1. Let α and β be substrings of A and B, respectively, and assume that $|\beta| = n$ and $\text{dist}(\alpha, \beta) \leq e$. Then there exists an $n \times e$ parallelogram P such that

1. P contains at least $T(n, q, e) := (n+1) - q(e+1)$ q-hits,
2. the B-projection of the parallelogram is $p_B(P) = \beta$,
3. the A-projection $p_A(P)$ of the parallelogram is contained in α.

The A- and B-projections are defined as illustrated below.

q-gram filters for ε-matches (3)



The *A-projection* $p_A(P)$ of a parallelogram P is defined as the substring of A between the last column of the first row of P and the first column of the last row of P.

The *B-projection* $p_B(P)$ of a parallelogram P is defined as the substring of B between the first and the last row of P.

q-gram filters for ε-matches (4)

(Note: these figures are taken from the RECOMB and GCB version, which uses the transposed matrix of the JCB article.)

q-gram filters for ε-matches (5)

Clearly, a q-hit (i, j) corresponds to $q+1$ consecutive entries of the edit matrix along the diagonal $j - i$. A q-hit is *contained* in a parallelogram if its corresponding matrix entries are.

The proof of Lemma 1 is straightforward: Consider the path of an optimal alignment of α and β. At each row except for the last q ones, we have a q-gram unless there is an edit operation among the next q edges. Each edit operation can 'destroy' at most q q-hits.

So the case where $|\beta|$ is fixed was easy. Next we consider ϵ-matches for $|\beta| \geq n_0$. The following lemma is the combinatorial foundation of the SWIFT algorithm.

q-gram filters for ε-matches (6)

Lemma 2. Let α and β be substrings of A and B, respectively, and assume that $|\beta| \geq n_0$ and $\mathrm{dist}(\alpha, \beta) \leq \varepsilon|\beta|$. Let $U(n, q, \varepsilon) := T(n, q, \lfloor \varepsilon n \rfloor) = (n+1) - q(\lfloor \varepsilon n \rfloor + 1)$ and assume that the q-gram size q and the threshold τ have been chosen such that

$$q < \lceil 1/\varepsilon \rceil \qquad \text{and} \qquad \tau \leq \min\{U(n_0, q, \varepsilon), U(n_1, q, \varepsilon)\},$$

where $n_1 := \lceil (\lfloor \varepsilon n_0 \rfloor + 1)/\varepsilon \rceil$.

Then there exists a $w \times e$ parallelogram P such that:

1. P contains at least τ q-hits whose projections intersect α and β,
2. $w = (\tau - 1) + q(e + 1)$,
3. $e = \left\lfloor \dfrac{2\tau + q - 3}{1/\varepsilon - q} \right\rfloor$,
4. if $|\beta| \leq w$, then $p_B(P)$ contains β, otherwise β contains $p_B(P)$.

q-gram filters for ε-matches (7)

The purpose of Lemma 2 is as follows. Given parameters ε and n_0, we can choose suitable values for q, τ, w, and e using Lemma 2. Then we enumerate all parallelograms P with enough hits according to these parameters. All relevant ε-matches can be found in these regions.

q-gram filters for ε-matches (8)

Proof of Lemma 2. The lemma is proven in three steps:

1. Assuming there is an ε-match (α, β) of length $|\beta| = n \geq n_0$, show that there are at least τ q-hits in the surrounding $n \times \lfloor \varepsilon n \rfloor$ parallelogram.

2. Argue that there is a $w \times e$ parallelogram that contains at least τ q-hits, where w and e do not depend on $n \geq n_0$.

3. Determine the dimensions w and e of such a parallelogram.

q-gram filters for ε-matches (9)

. . . details omitted . . .

q-gram filters for ε-matches (10)

TABLE 1. FILTER PARAMETERS FOR $\epsilon = 0.05$, BY LEMMA 2 AND COROLLARY 1, RESPECTIVELY

	$q = 7$			$q = 9$			$q = 11$		
n_0	*30*	*50*	*100*	*30*	*50*	*100*	*30*	*50*	*100*
w	37	64	128	39	68	136	40	71	133
e	2	4	9	2	4	9	2	4	8
τ	17	30	59	13	24	47	8	17	35

	$q = 11$								
τ	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
n_0	28	29	41	42	43	44	45	46	47
w	39	40	52	53	54	55	67	68	69
e	2	2	3	3	3	3	4	4	4

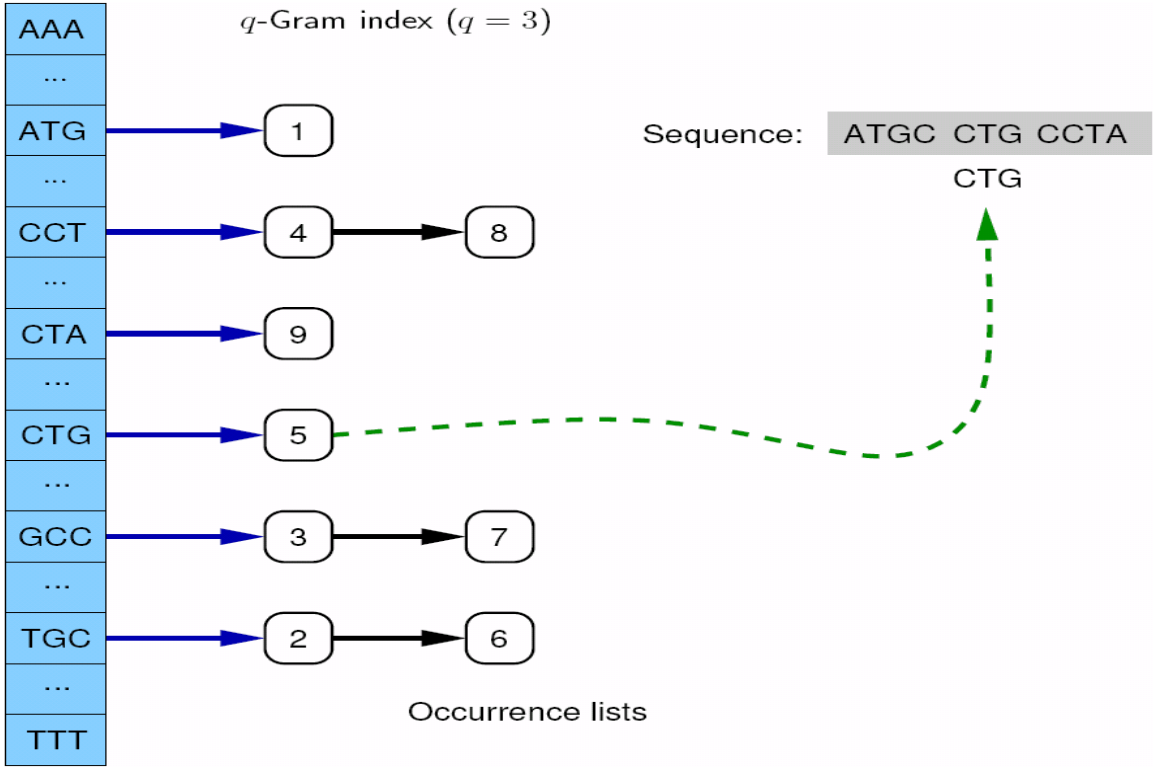
Algorithm

The SWIFT algorithm relies on the q-gram filter for ϵ-matches of length n_0 or greater. Using the parameters obtained from Lemma 2, it searches for all $w \times e$ parallelograms which contain a sufficient number of q-grams.

Algorithm (2)

In the preprocessing step, we construct a q-gram index for the target sequence A. The index consists of two tables:

1. The *occurrence table* is a concatenation of the lists $L(G) := \{\, i \mid A[i..i+q-1] = G \,\}$ for all q-grams $G \in \Sigma^q$ in A.
2. The *lookup table* is an array indexed by the natural encoding of G to base $|\Sigma|$, giving the start of each list in the occurrence table.



Algorithm (3)

Once the q-gram index is built, the $w \times e$ parallelograms containing τ or more q-hits can be found using a simple sliding window algorithm.

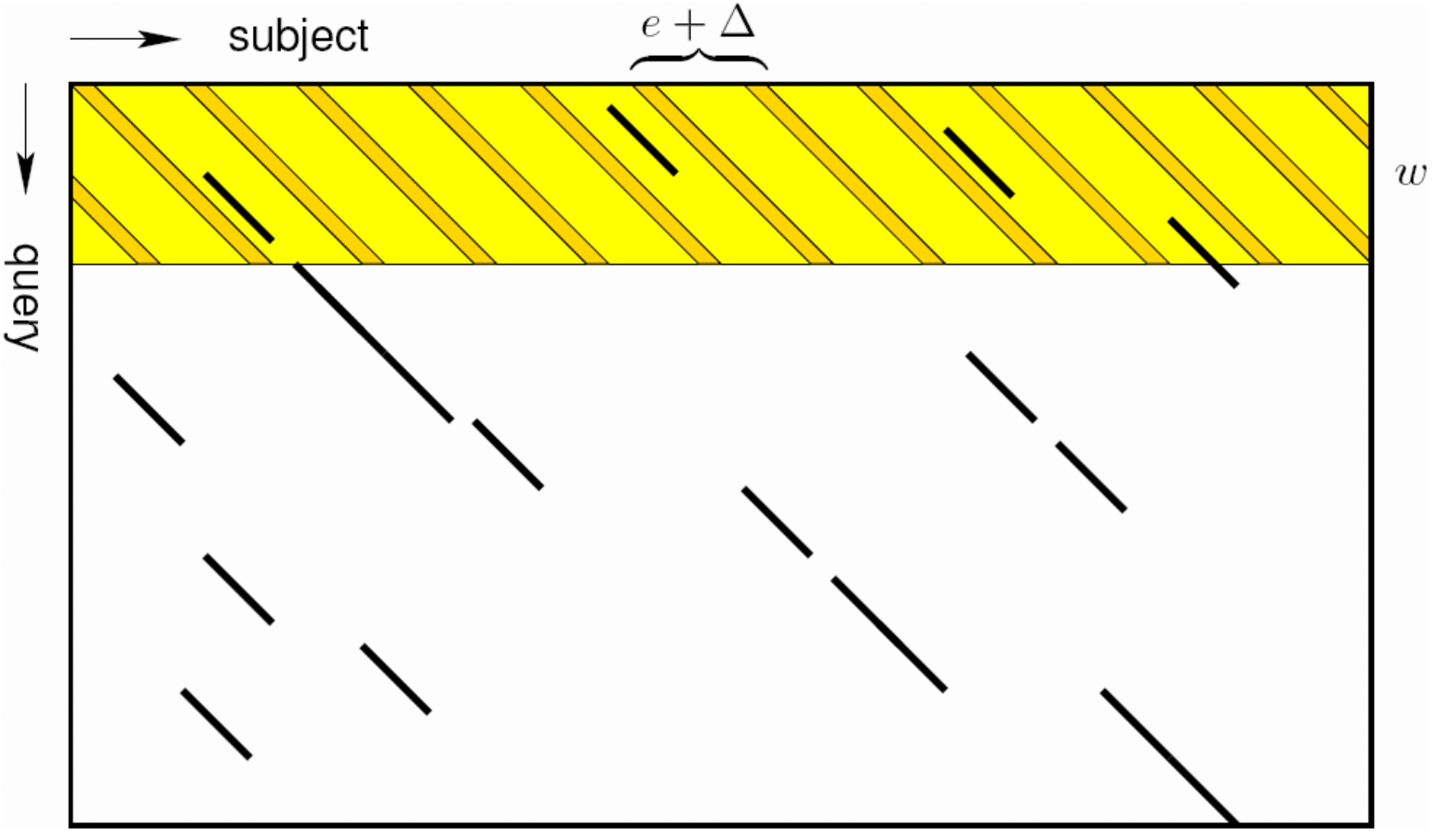
The idea is to split the (fictitious) edit matrix into overlapping *bins* of $e+1$ diagonals. For each bin we count the number of q-hits in the $w \times e$ parallelogram that is the intersection of the diagonals of the corresponding bin and the rows of the sliding window $W_j := B[j..j+q-1]$.

As the sliding window proceeds to W_{j+1}, the bin counters are updated to reflect the changes due to the q-grams leaving and entering the window.

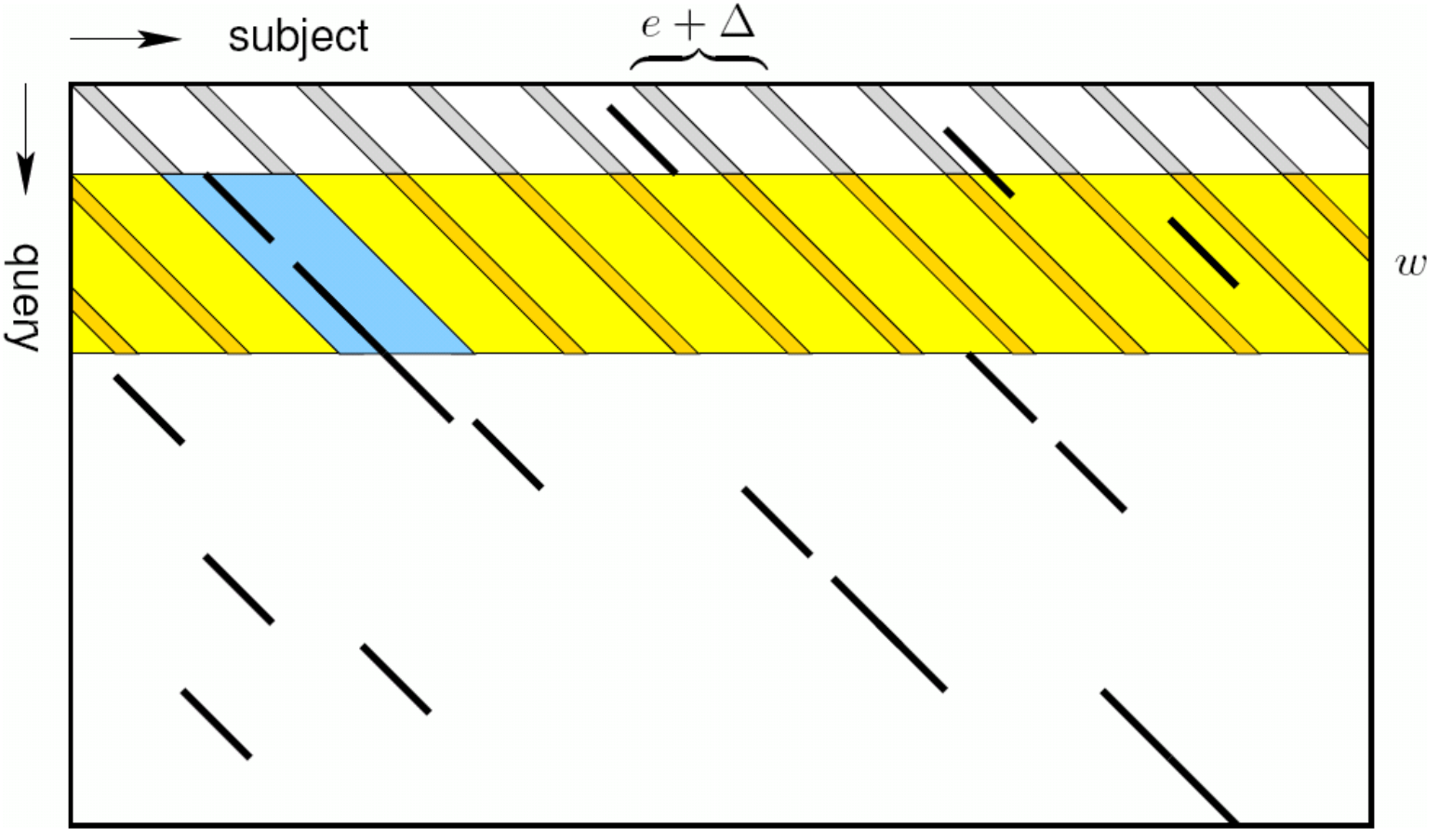
Whenever a bin counter reaches τ, the corresponding parallelogram is reported. Overlapping parallelograms can be merged on the fly.

The space requirement for the bins is reduced by searching for somewhat larger parallelograms of size $w \times (e+\Delta)$. Then each bin counts for $e+\Delta+1$ diagonals, and successive bins overlap by e diagonals. While this will lead to more verifications, it reduces the number of bins which have to be maintained. In practice, Δ is set to a power of 2, and bin indices are computed with fast bit-operations.

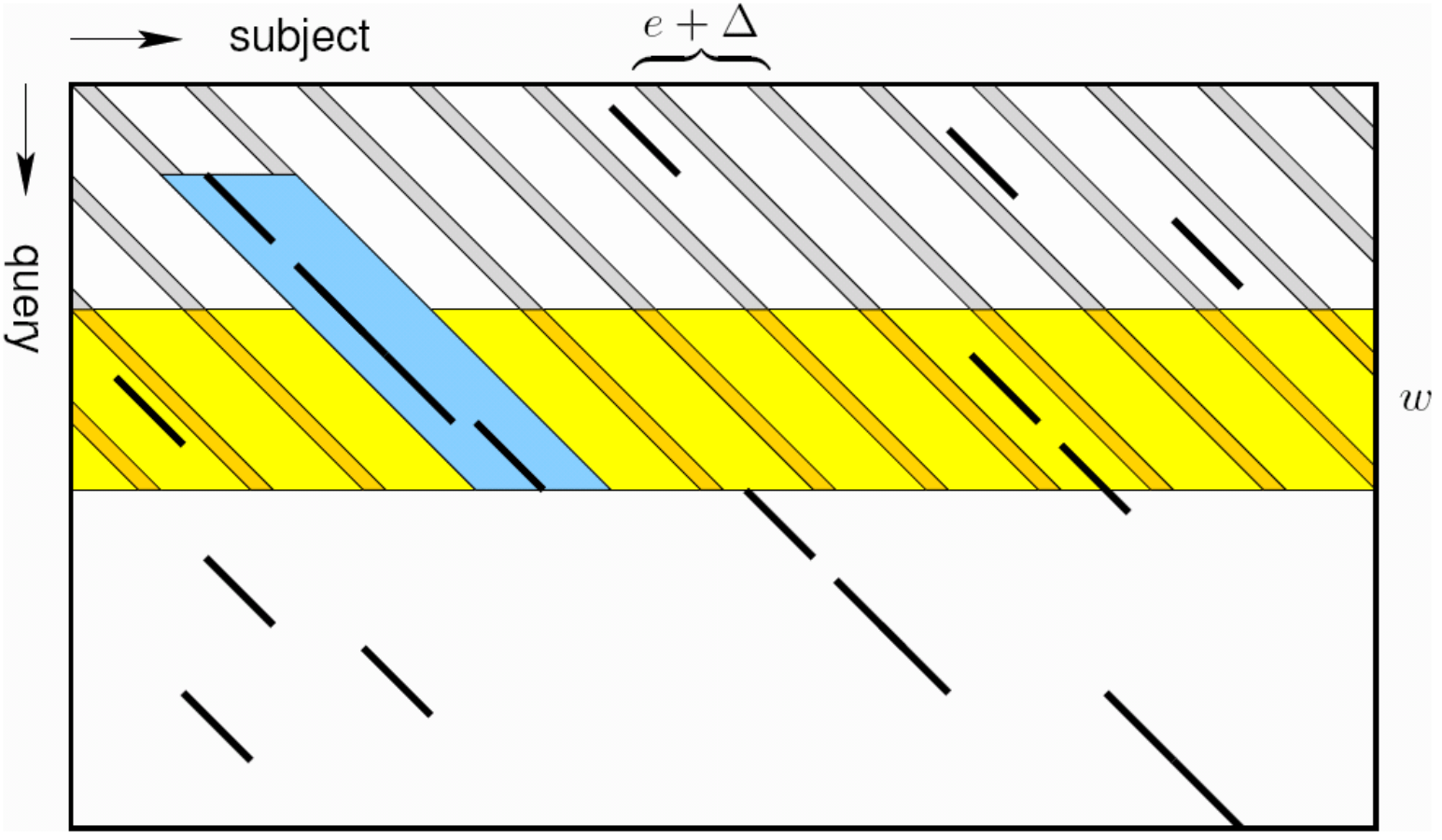
Algorithm (4)



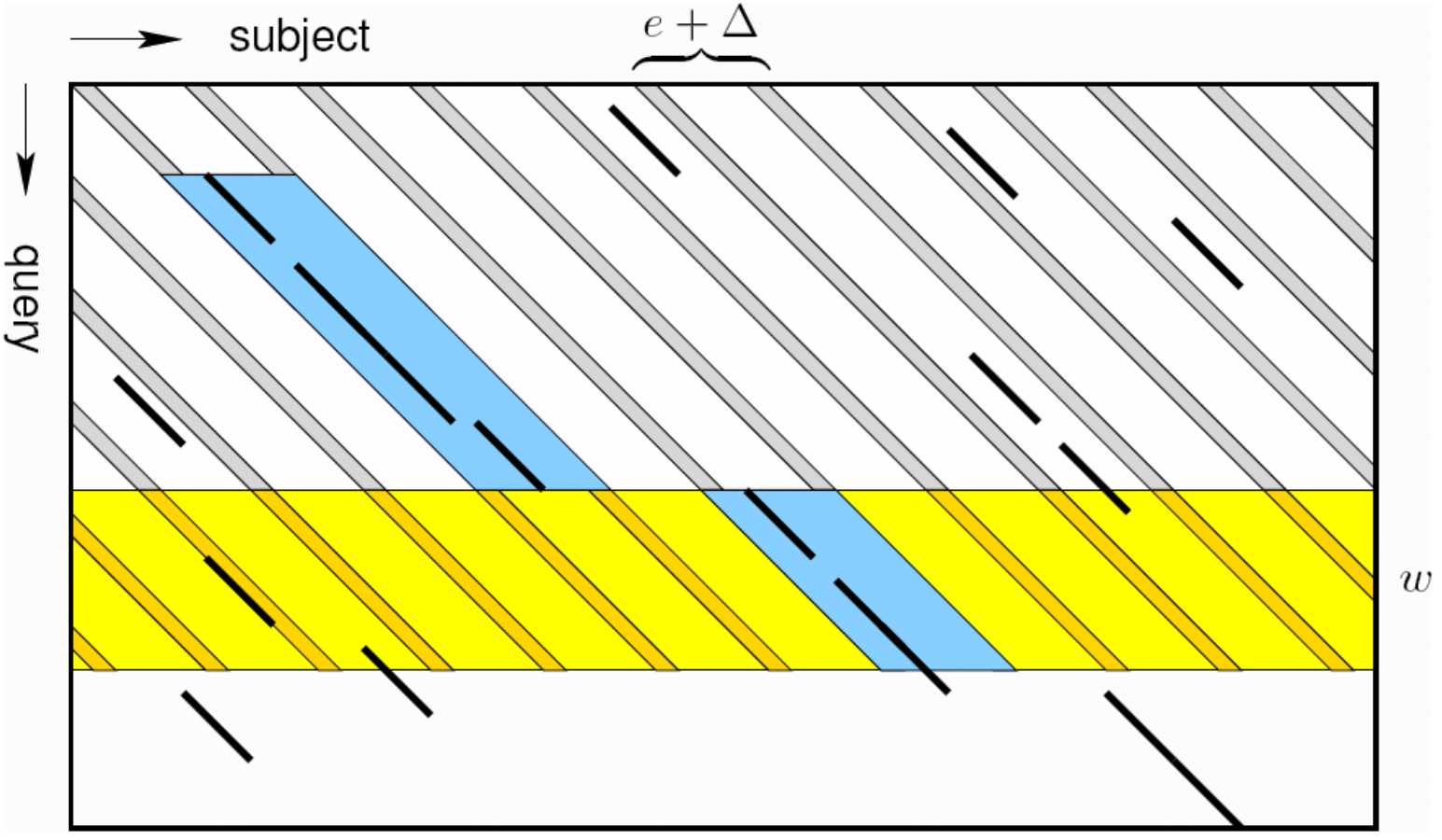
Algorithm (5)



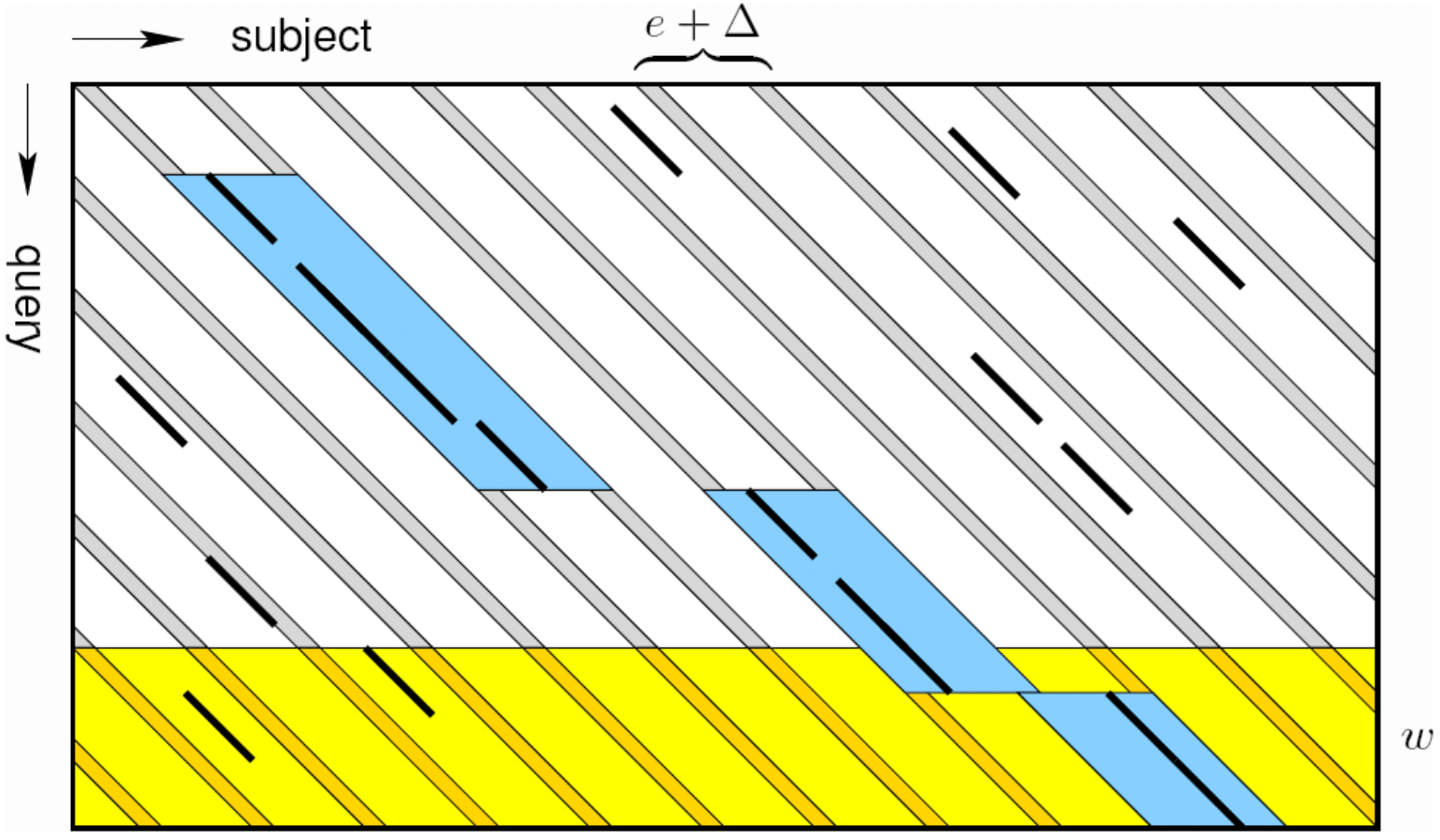
Algorithm (6)



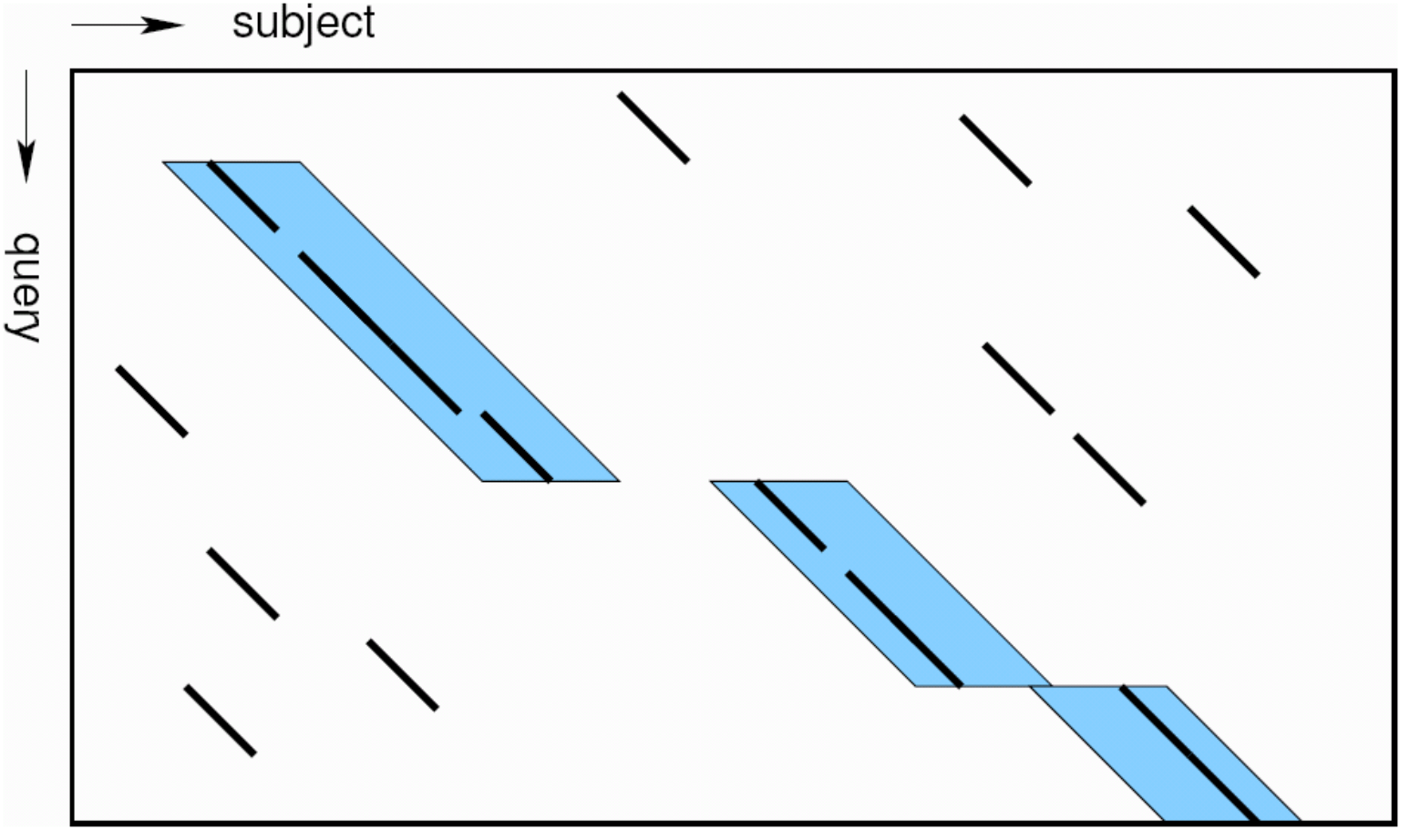
Algorithm (7)



Algorithm (8)

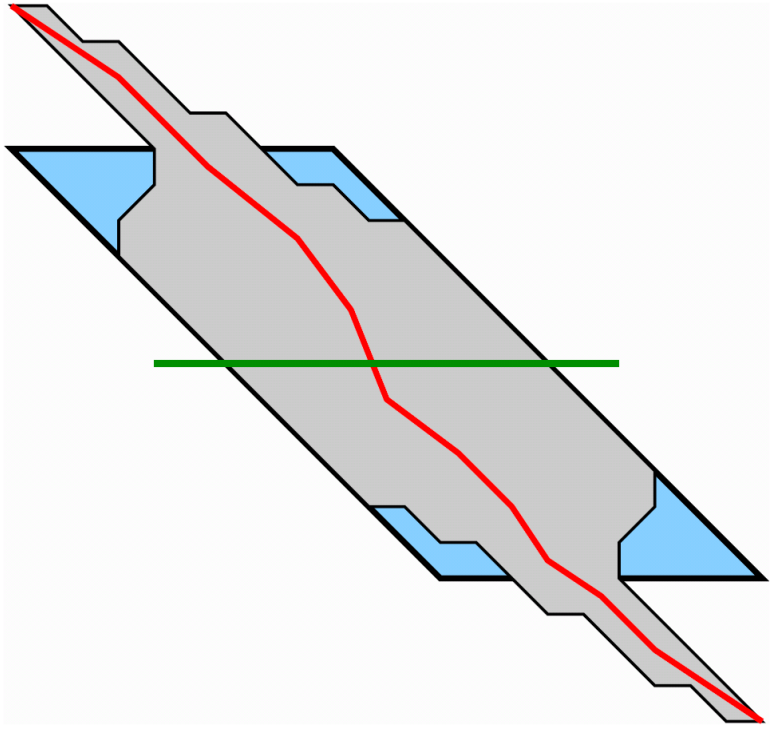


Algorithm (9)



Algorithm (10)

Each ‘candidate’ parallelogram must be checked for the presence of an ε-match. This can be done trivially by dynamic programming. Alternatively, one can use the knowledge about the q-grams in the ε-match to construct an alignment by sparse dynamic programming.



Algorithm (11)

Algorithm 2: Filter for identifying parallelograms for ϵ-matches

Input : Query B; q-gram index I for target A; parameters w, e, τ; and $\Delta = 2^z$
Output: Set of parallelograms P

```
1  Allocate and initialize array of bin records Bins
2  P ← ∅
3  for j ← 0 to |B| − q do
4      G ← B[j, j + q − 1]
5      L(G) ← lookup occurrence list for G in I
6      foreach i ∈ L(G) do
7          d ← |A| + j − i
8          b_0 ← d ≫_bit z
9          b_m ← b_0 mod |Bins|
10         P ← P ∪ UpdateBin(Bins[b_m], j, b_0 ≪_bit z)
11         if (d &_bit (Δ − 1)) < e then
12             b_m ← (b_m + |Bins| − 1) mod |Bins|
13             P ← P ∪ UpdateBin(Bins[b_m], j, (b_0 − 1) ≪_bit z)
14         if (j − e) mod (Δ − 1) = 0 then
15             b_0 ← (j − e) ≫_bit z
16             b_m ← b_0 mod |Bins|
17             /* CheckAndResetBin is similar to lines 3–8 of UpdateBin */
18             P ← P ∪ CheckAndResetBin(Bins[b_m], j, b_0 ≪_bit z)
19 P ← P ∪ { remaining parallelograms in Bins }
```

Algorithm (12)

Algorithm 1: `UpdateBin`(r, j, d)

Input : Bin record r; q-hit position j in sequence B; and offset bin diagonal d.
Output: Empty or singleton parallelogram set P.

```
1  P ← ∅
2  if j − w + q > r.max then
3      if r.count ≥ τ then
4          p.left ← |A| − d
5          p.top ← r.max + q
6          p.bottom ← r.min
7          P ← { p }
8      r.count ← 0
9  if r.count = 0 then
10     r.min ← j
11 if r.max < j then
12     r.max ← j
13     r.count ← r.count + 1
14 return P
```

Results

TABLE 3. FILTRATION RATIOS AND TIMES FOR EST ALL-AGAINST-ALL COMPARISON

	SWIFT		*QUASAR*			
	Filtration		*Filtration, best ratio*		*Filtration, best time*	
(ϵ, n_0)	*Ratio*	*Time (s)*	*Ratio*	*Time (s)*	*Ratio*	*Time (s)*
(0.05, 50)	$6.5 \cdot 10^{-6}$	6.0	$4.5 \cdot 10^{-4}$	36.1	$2.1 \cdot 10^{-3}$	4.2
(0.04, 30)	$4.5 \cdot 10^{-6}$	5.0	$4.0 \cdot 10^{-4}$	69.0	$3.1 \cdot 10^{-3}$	4.4
(0.05, 30)	$5.4 \cdot 10^{-6}$	6.1	$4.3 \cdot 10^{-4}$	68.5	$3.5 \cdot 10^{-3}$	4.4

TABLE 2. RUNNING TIMES FOR EST ALL-AGAINST-ALL COMPARISON[a]

	SWIFT			*BLAST*	*SSEARCH*
(ϵ, n_0)	(0.05, 50)	(0.04, 30)	(0.05, 30)	—	—
Running time	18 s	29 s	35 s	773 s	8 h

[a]The time for the database formatting and preprocessing in BLAST (3 s) and SWIFT (12 s) is not included.