Bioinformatics: String Matching for Read Mapping

Swakkhar Shatabda

Department of Computer Science and Engineering
BRAC University

Book Reference

Boyer-Moore Algorithm (Algorithms on Strings, Trees, and Sequences, Gusfiled et al, Chapter 2 - Section 2.2)

BRAC UNIVERSITY
Inspiring Excellence

Exact Match - Naive Algorithm

```
def naive(p, t):
    occurrences = []
    for i in range(len(t) - len(p) + 1):  # loop over alignments
        match = True
        for j in range(len(p)):           # loop over characters
            if t[i+j] != p[j]:            # compare characters
                match = False             # mismatch; reject alignment
                break
        if match:
          occurrences.append(i)           # all chars matched; record
    return occurrences
```

BRAC UNIVERSITY
Inspiring Excellence

Boyer-Moore Algorithm - Idea

P: word

T: There would have been a time for such a word

word

u doesn't occur in *P*, so we can skip next two alignments

P: word

T: There would have been a time for such a word

word

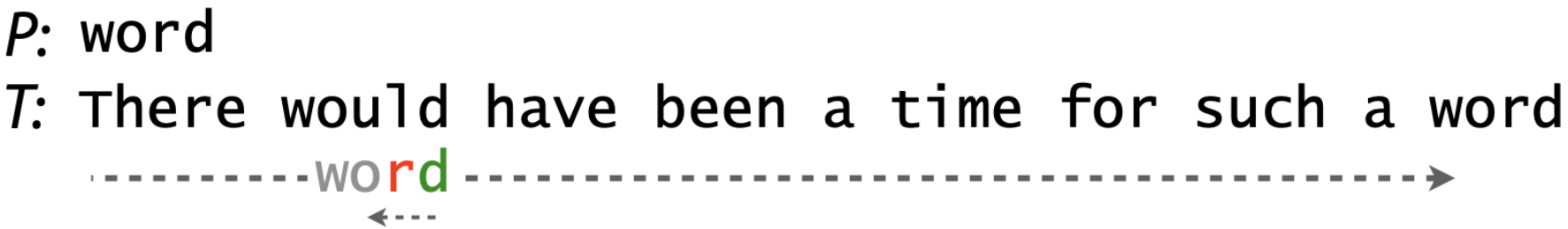
word skip!

word skip!

word

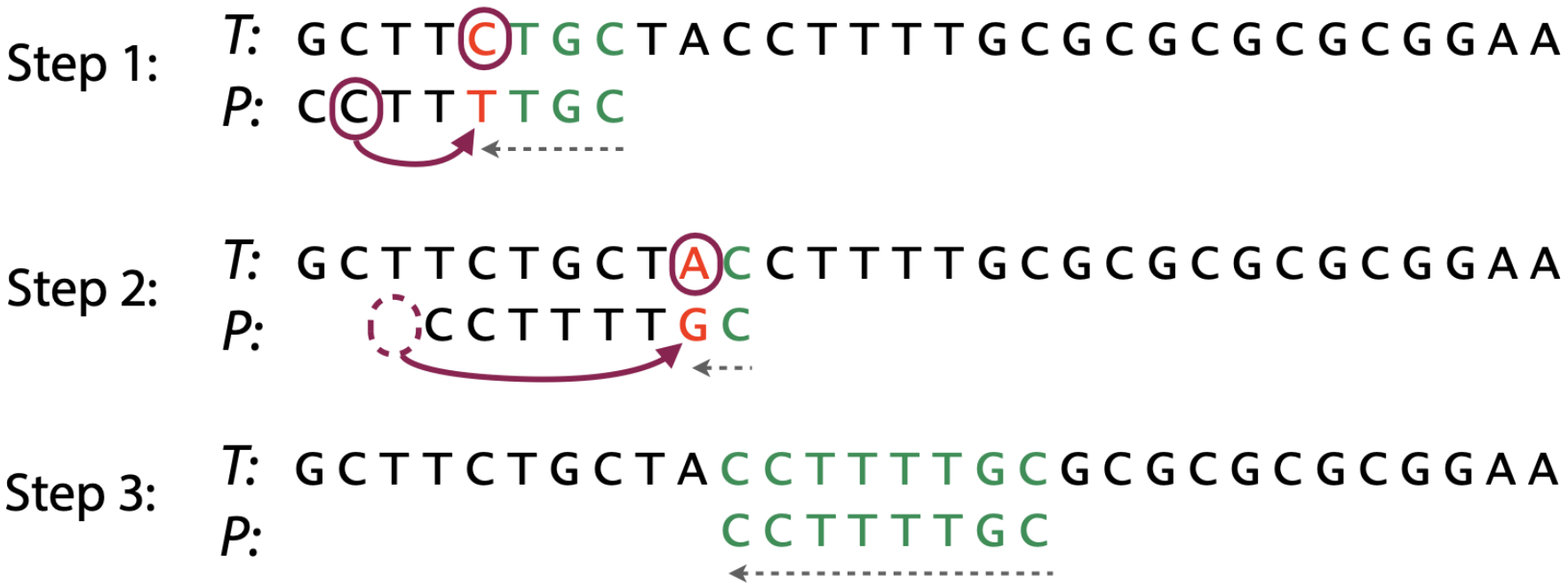
Boyer-Moore Algorithm - Idea

- Learn from character comparisons to skip pointless alignments
- Try alignments in left-to-right order, and try character comparisons in right-to-left order



Boyer-Moore Algorithm - Bad Character Rule

- Upon mismatch, skip alignments until (a) mismatch becomes a match, or (b) P moves past mismatched character

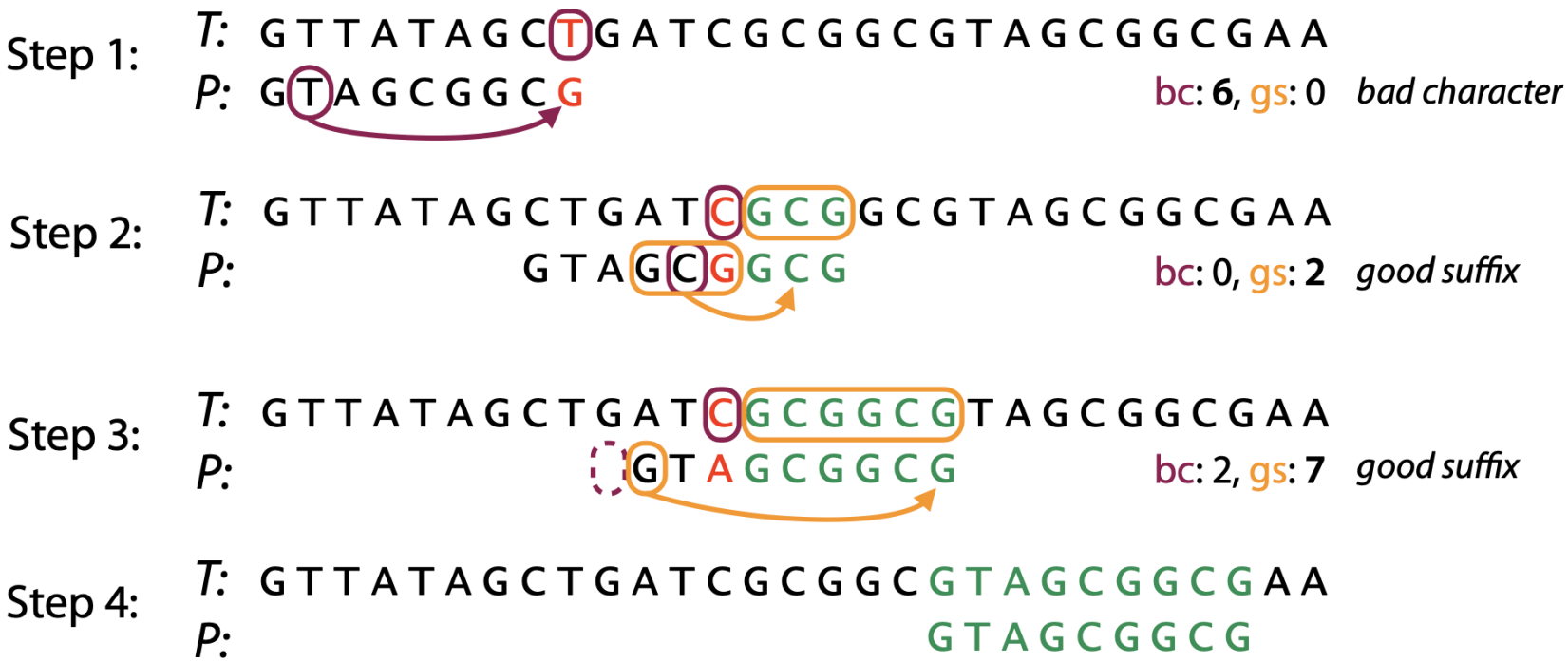


Boyer-Moore Algorithm - Good Suffix Rule

- Let t = substring matched by inner loop; skip until (a) there are no mismatches between P and t or (b) P moves past t

Boyer-Moore algorithm - Put Together

- Use bad character or good suffix rule, whichever skips more

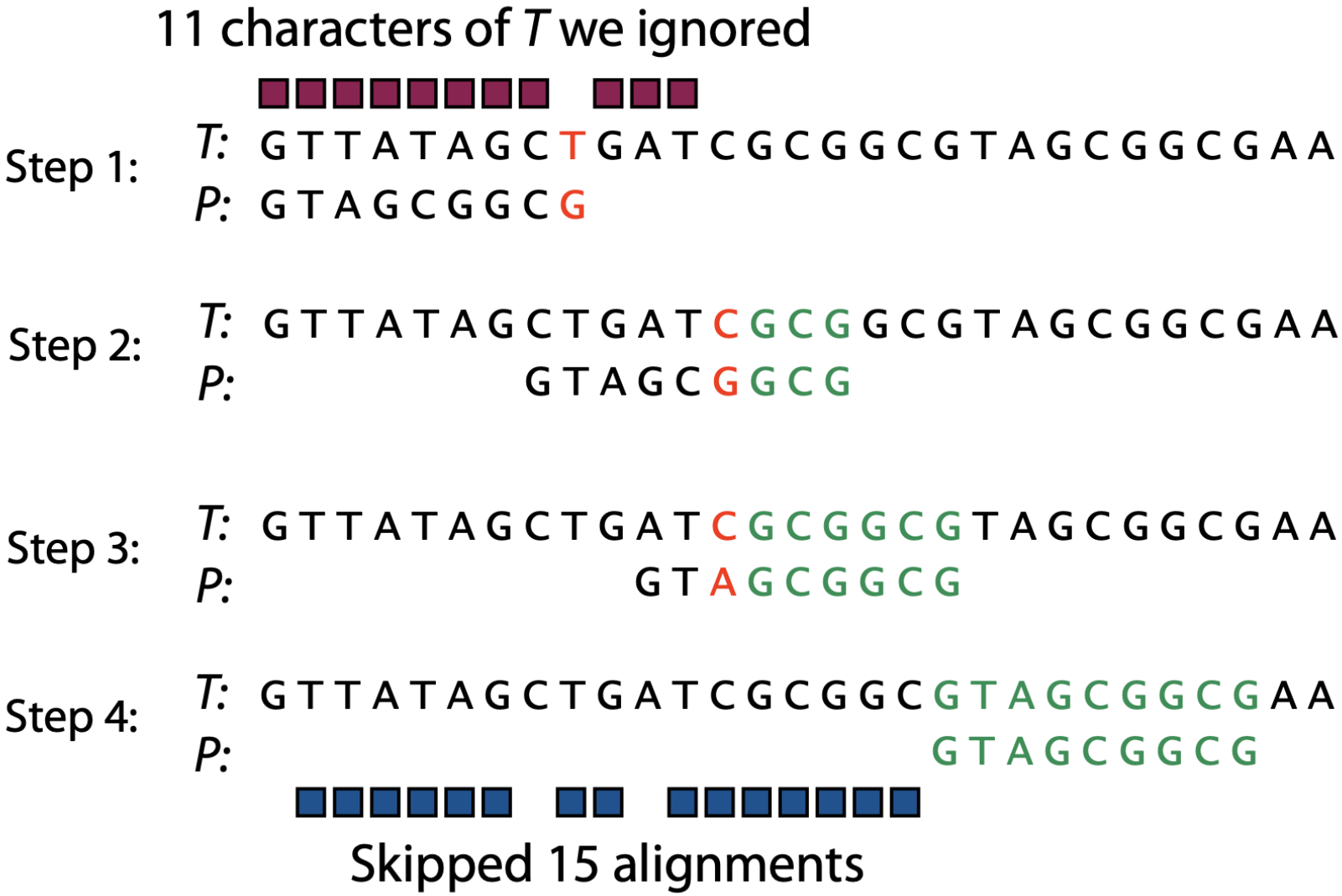


https://colab.research.google.com/drive/1XFYYu7PmSTrU2mdRxEYHGEIKvs41H-WN?usp=sharing

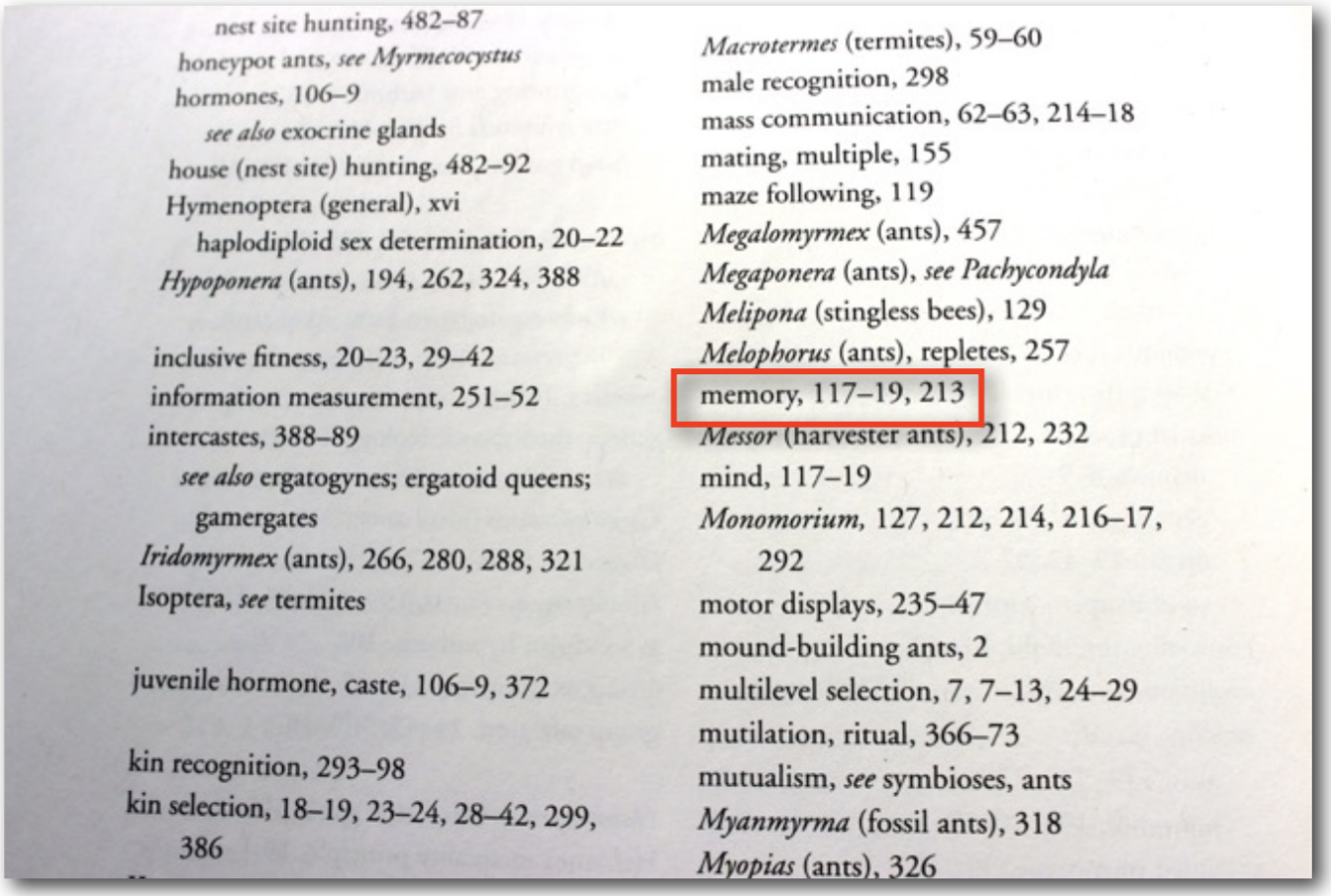
BRAC UNIVERSITY
Inspiring Excellence

Boyer-Moore algorithm - Put Together Example

- Use bad character or good suffix rule, whichever skips more



Indexing the Genome



nest site hunting, 482–87
honeypot ants, *see Myrmecocystus*
hormones, 106–9
see also exocrine glands
house (nest site) hunting, 482–92
Hymenoptera (general), xvi
haplodiploid sex determination, 20–22
Hypoponera (ants), 194, 262, 324, 388

inclusive fitness, 20–23, 29–42
information measurement, 251–52
intercastes, 388–89
see also ergatogynes; ergatoid queens; gamergates
Iridomyrmex (ants), 266, 280, 288, 321
Isoptera, *see* termites

juvenile hormone, caste, 106–9, 372

kin recognition, 293–98
kin selection, 18–19, 23–24, 28–42, 299, 386

Macrotermes (termites), 59–60
male recognition, 298
mass communication, 62–63, 214–18
mating, multiple, 155
maze following, 119
Megalomyrmex (ants), 457
Megaponera (ants), *see Pachycondyla*
Melipona (stingless bees), 129
Melophorus (ants), repletes, 257
memory, 117–19, 213
Messor (harvester ants), 212, 232
mind, 117–19
Monomorium, 127, 212, 214, 216–17, 292
motor displays, 235–47
mound-building ants, 2
multilevel selection, 7, 7–13, 24–29
mutilation, ritual, 366–73
mutualism, *see* symbioses, ants
Myanmyrma (fossil ants), 318
Myopias (ants), 326

Key terms ordered alphabetically, with associated page #s

BRAC UNIVERSITY
Inspiring Excellence

Indexing the Genome

Grocery store items grouped into aisles

Indexing the Genome

T: C G T G C G T G C T T

Indexing the Genome

Index of T

C G T G C : 0

T: C G T G C G T G C T T

Indexing the Genome

Index of T

CGTGC: 0
GTGCG: 1

T: C G T G C G T G C T T

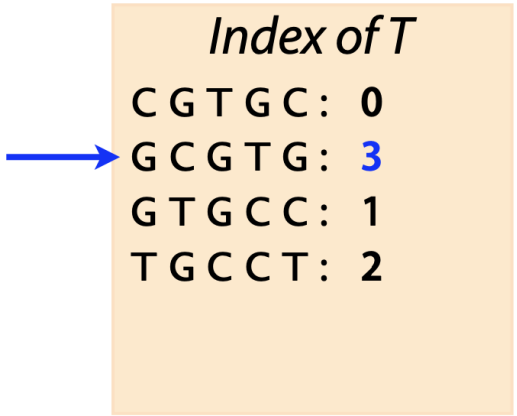
Indexing the Genome

Index of T

CGTGC: 0
GTGCG: 1
TGCGT: 2

T: C G T G C G T G C T T

Indexing the Genome



T: CGTGCGTGCTT

Indexing the Genome

T: C G T G C G T G C T T

Indexing the Genome

Index of T

CGTGC: 0,4
GCGTG: 3
GTGCC: 1
GTGCT: 5
TGCCT: 2

T: C G T G C G T G C T T

Indexing the Genome

Index of T

CGTGC: 0,4
GCGTG: 3
GTGCC: 1
GTGCT: 5
TGCCT: 2
TGCTT: 6

T: C G T G C G T G C T T

Indexing the Genome

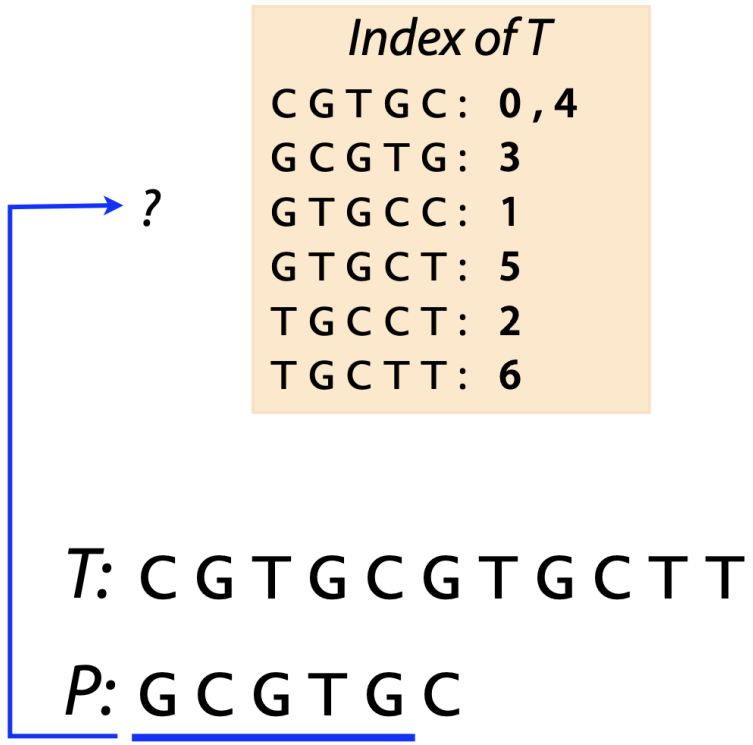
k-mer: substring of length k

Index of T	
C G T G C:	0, 4
G C G T G:	3
G T G C C:	1
G T G C T:	5
T G C C T:	2
T G C T T:	6

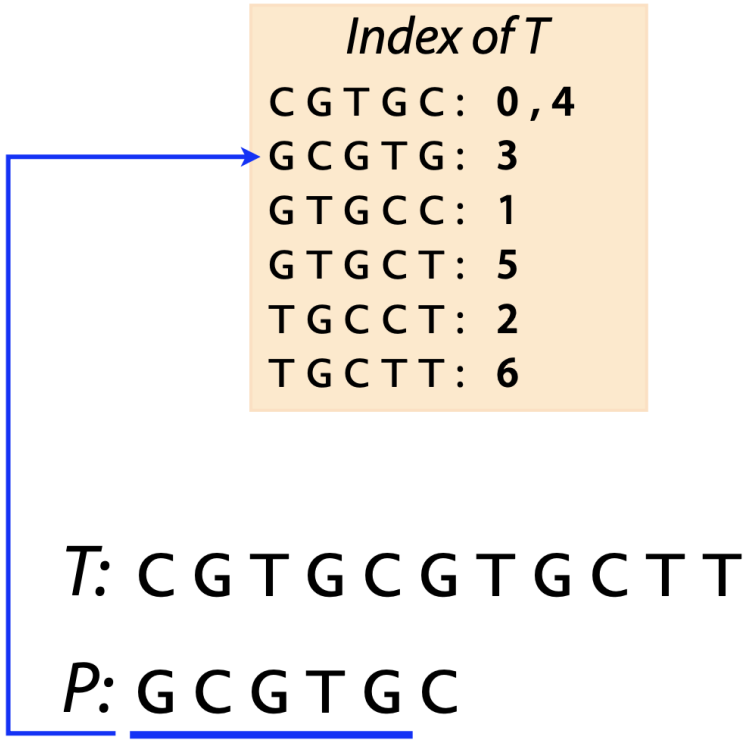
5-mer index

T: C G T G C G T G C T T

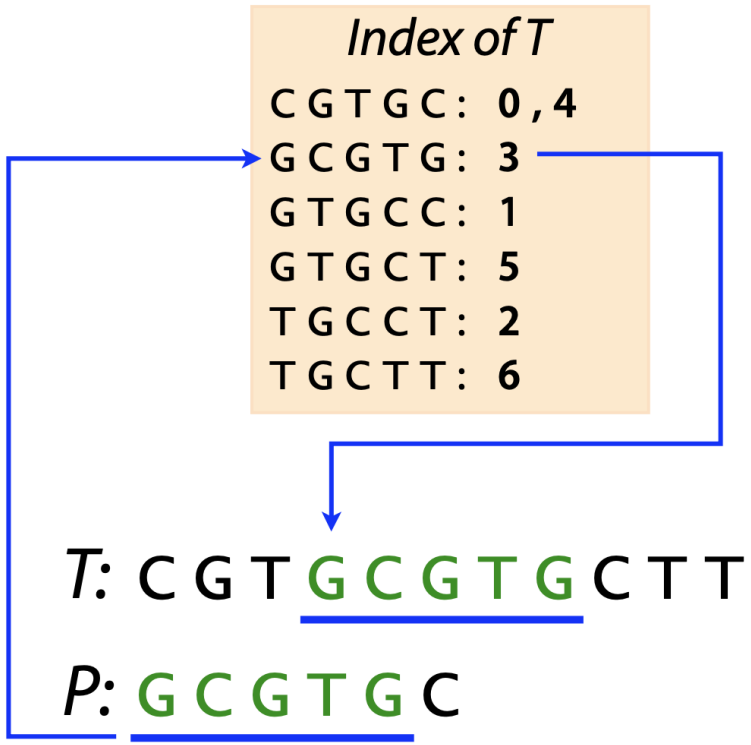
Indexing the Genome



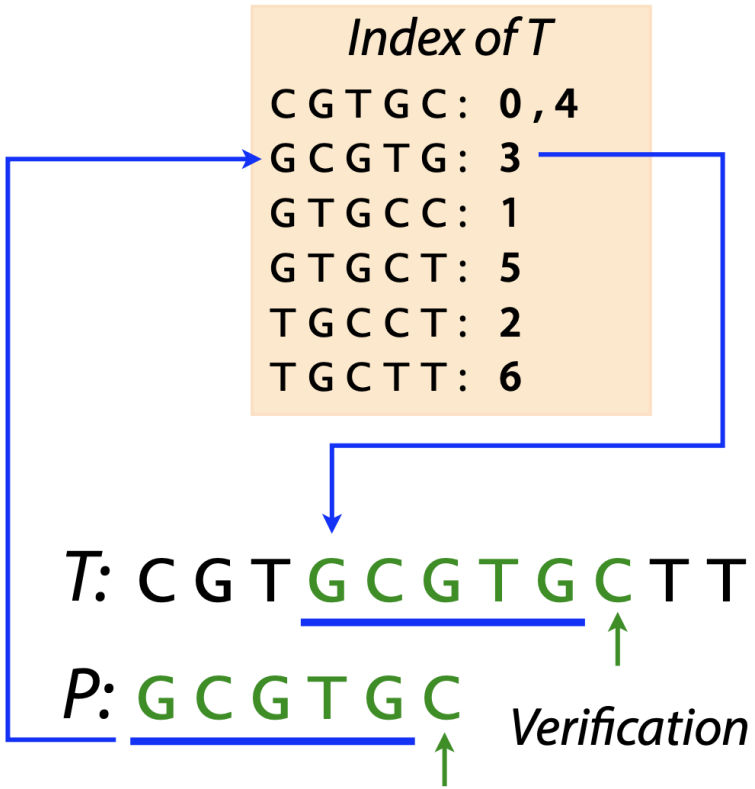
Indexing the Genome



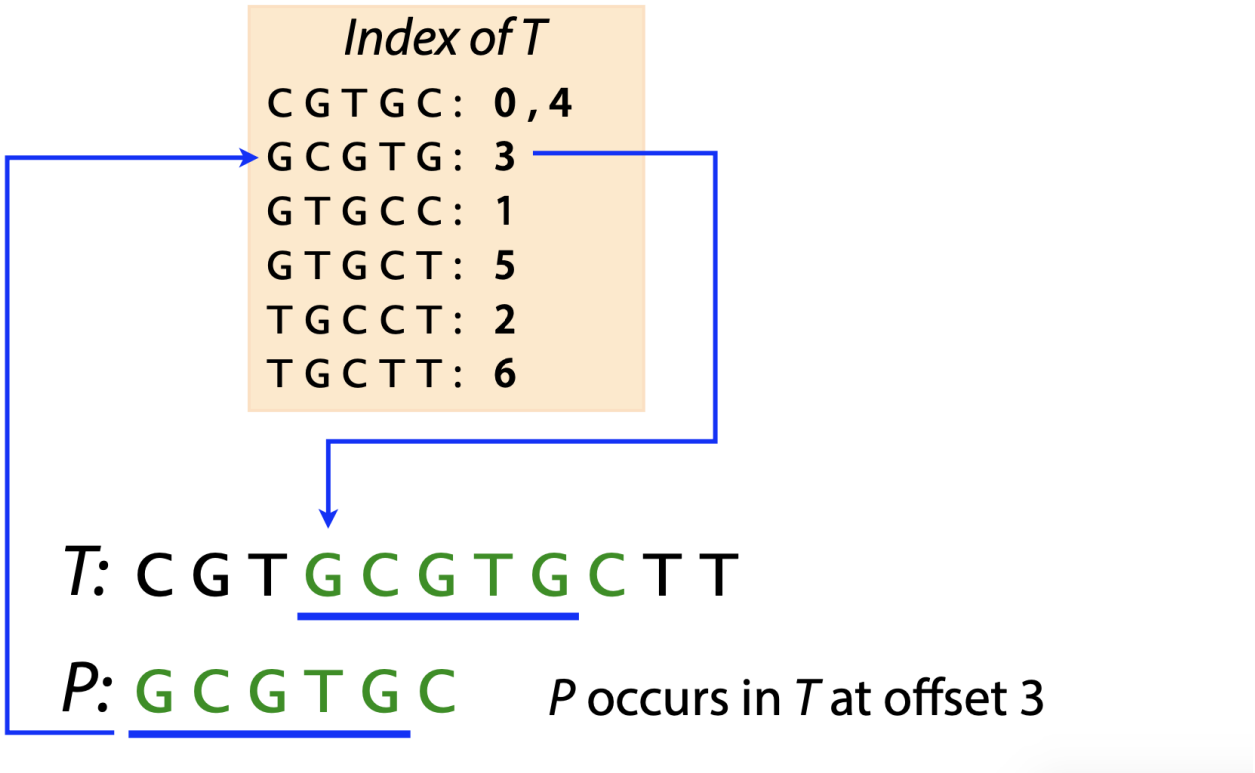
Indexing the Genome



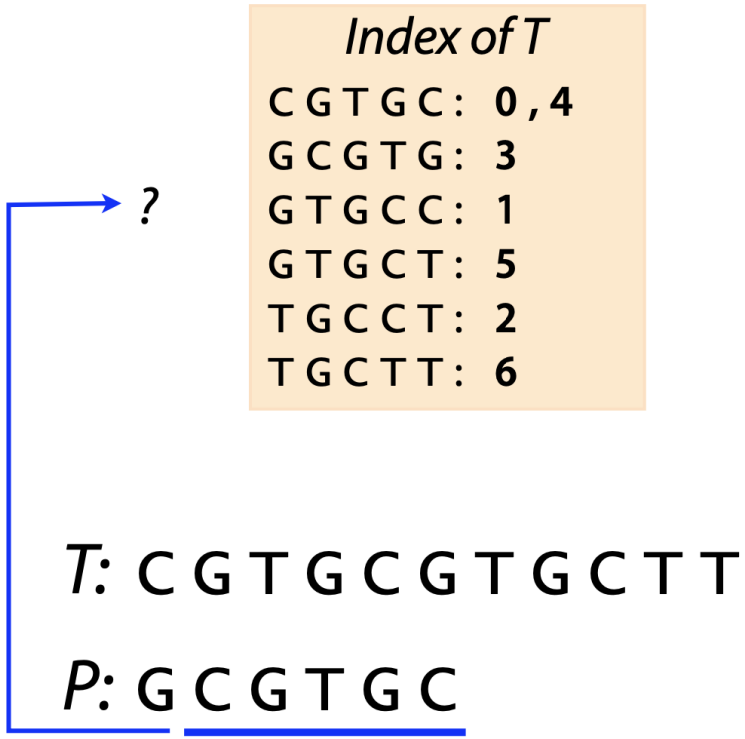
Indexing the Genome



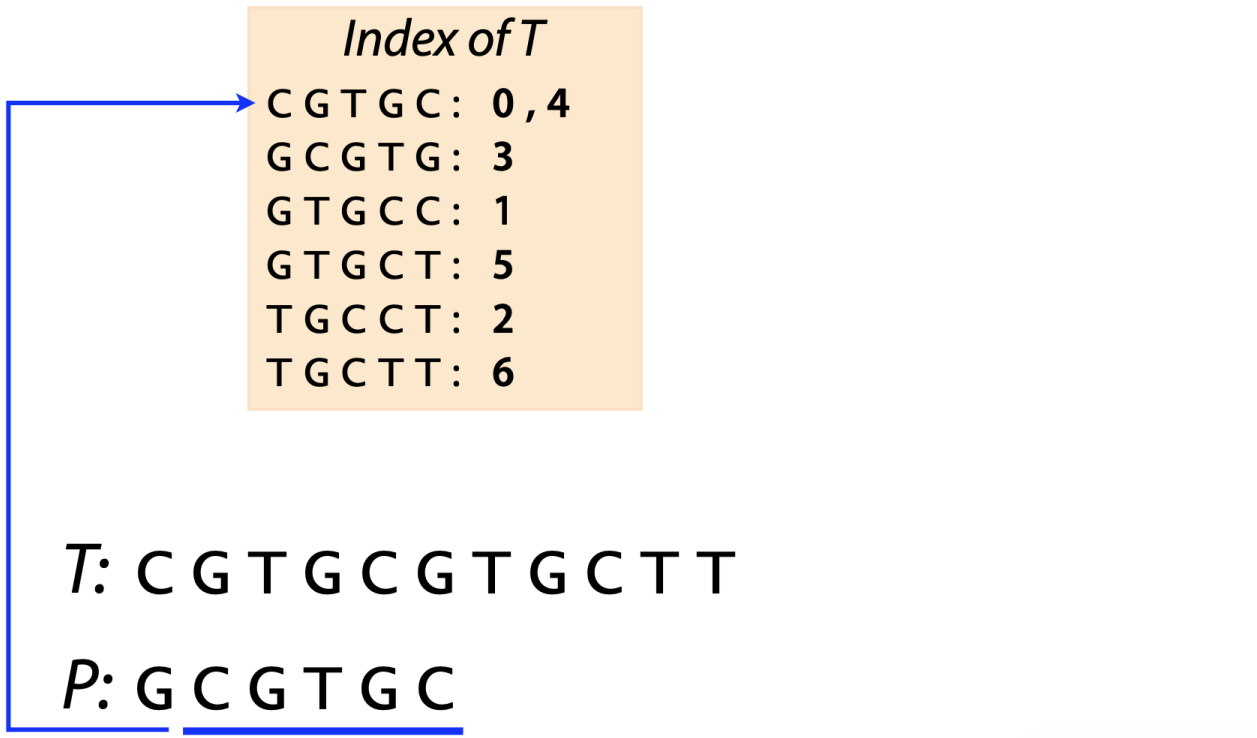
Indexing the Genome



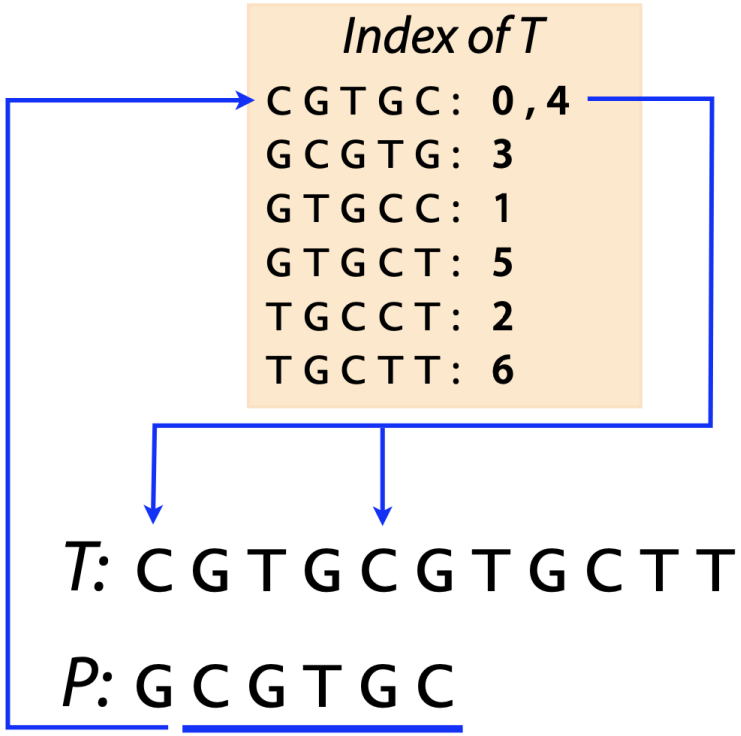
Indexing the Genome



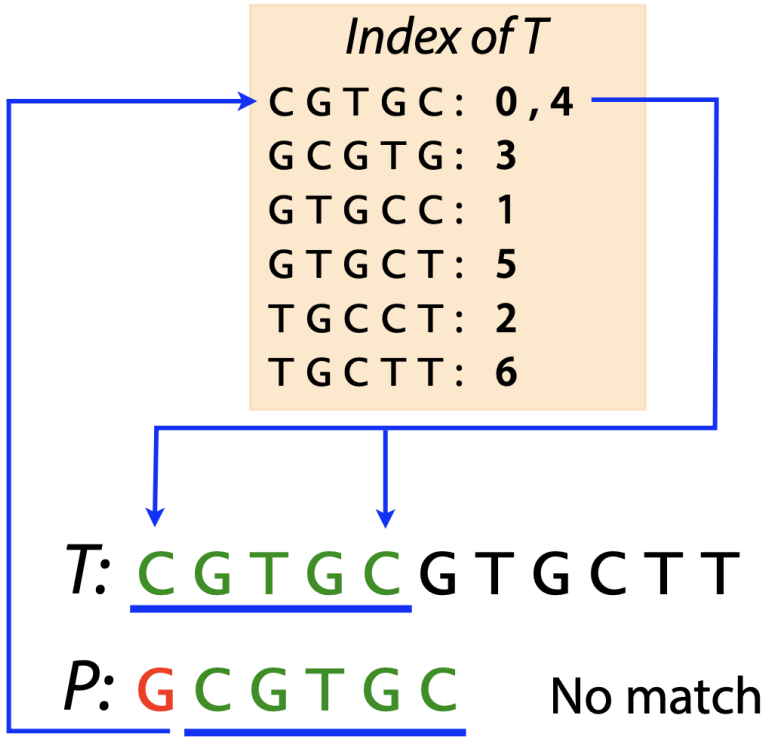
Indexing the Genome



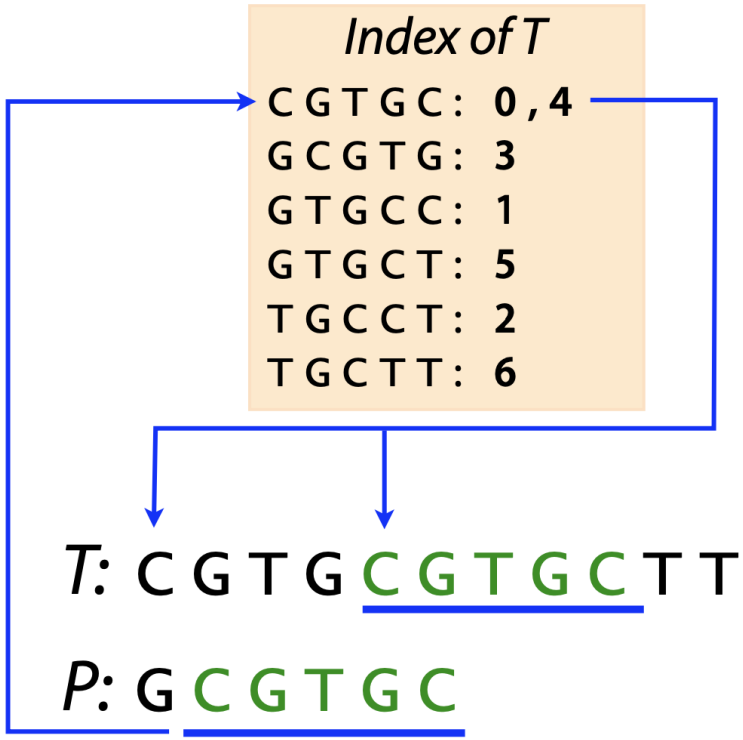
Indexing the Genome



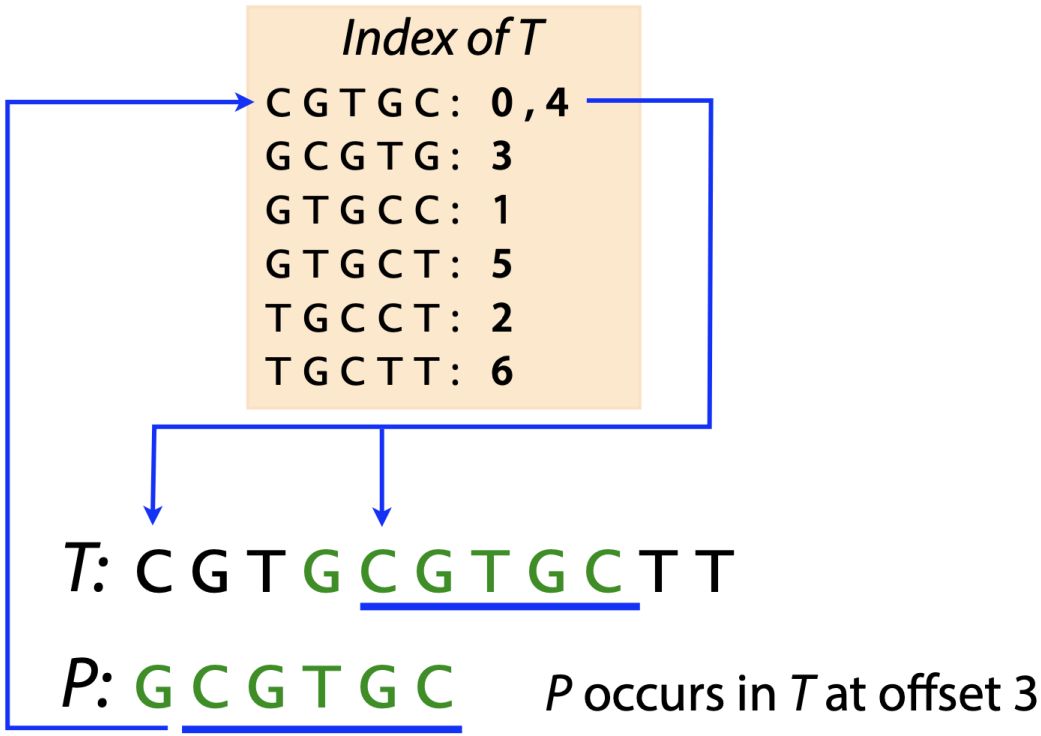
Indexing the Genome



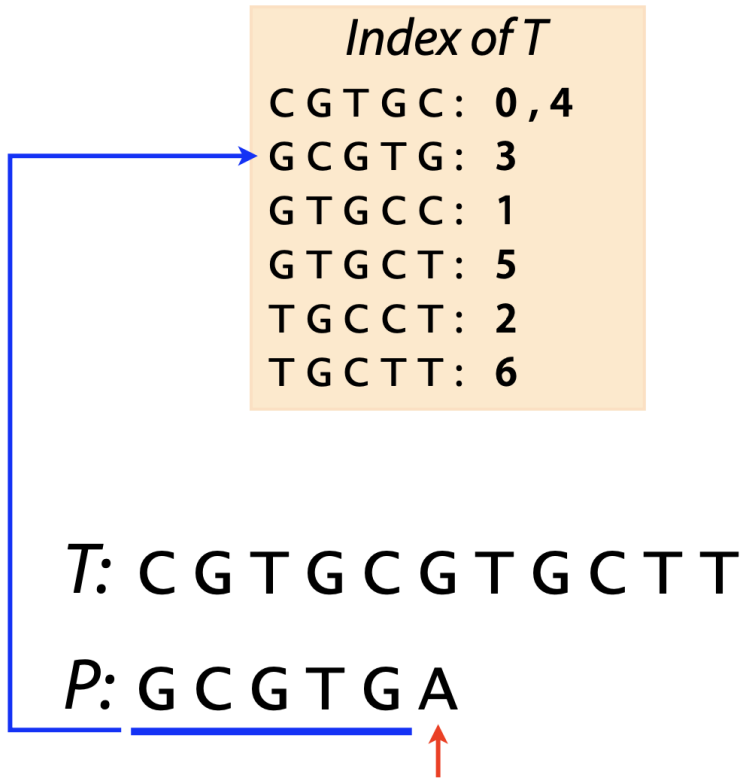
Indexing the Genome



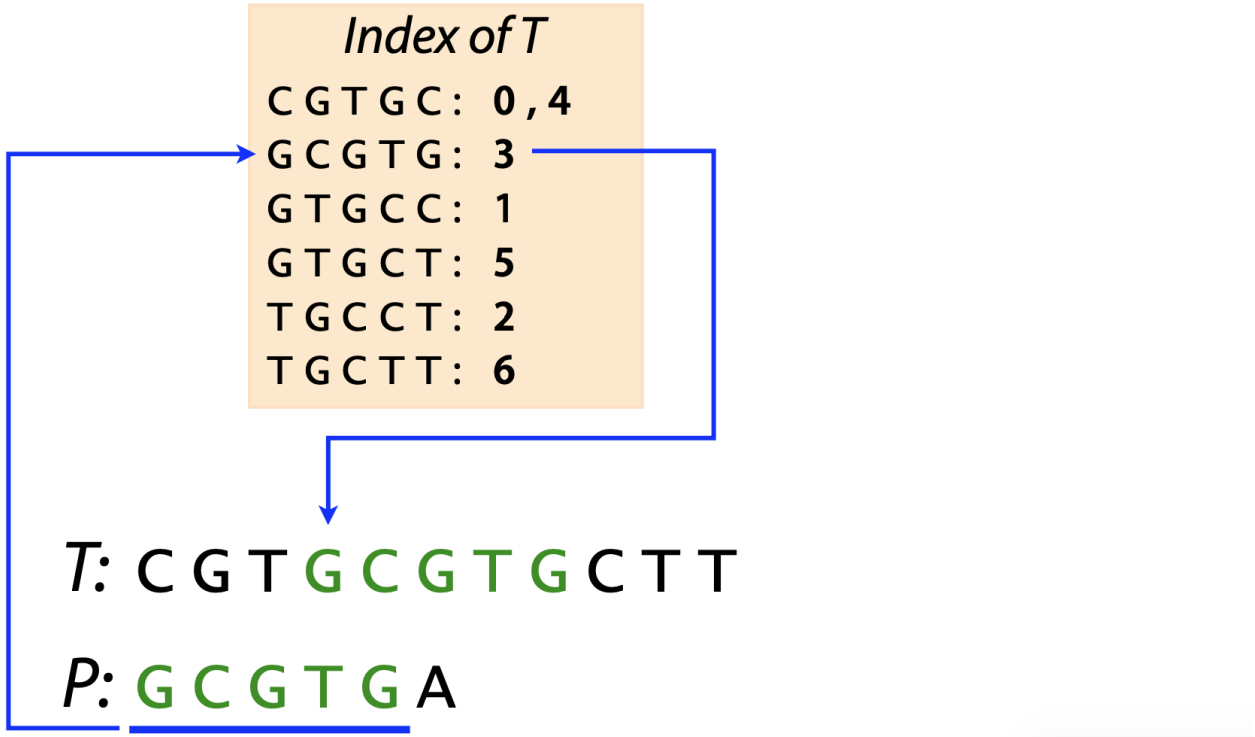
Indexing the Genome



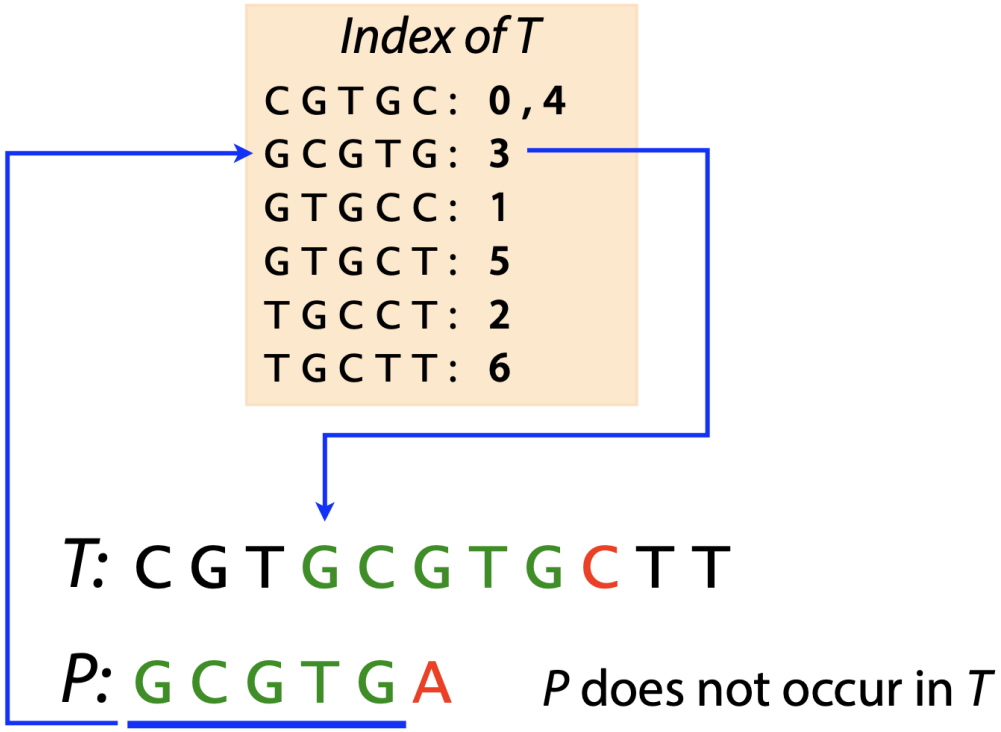
Indexing the Genome



Indexing the Genome



Indexing the Genome



Indexing the Genome

Index of T

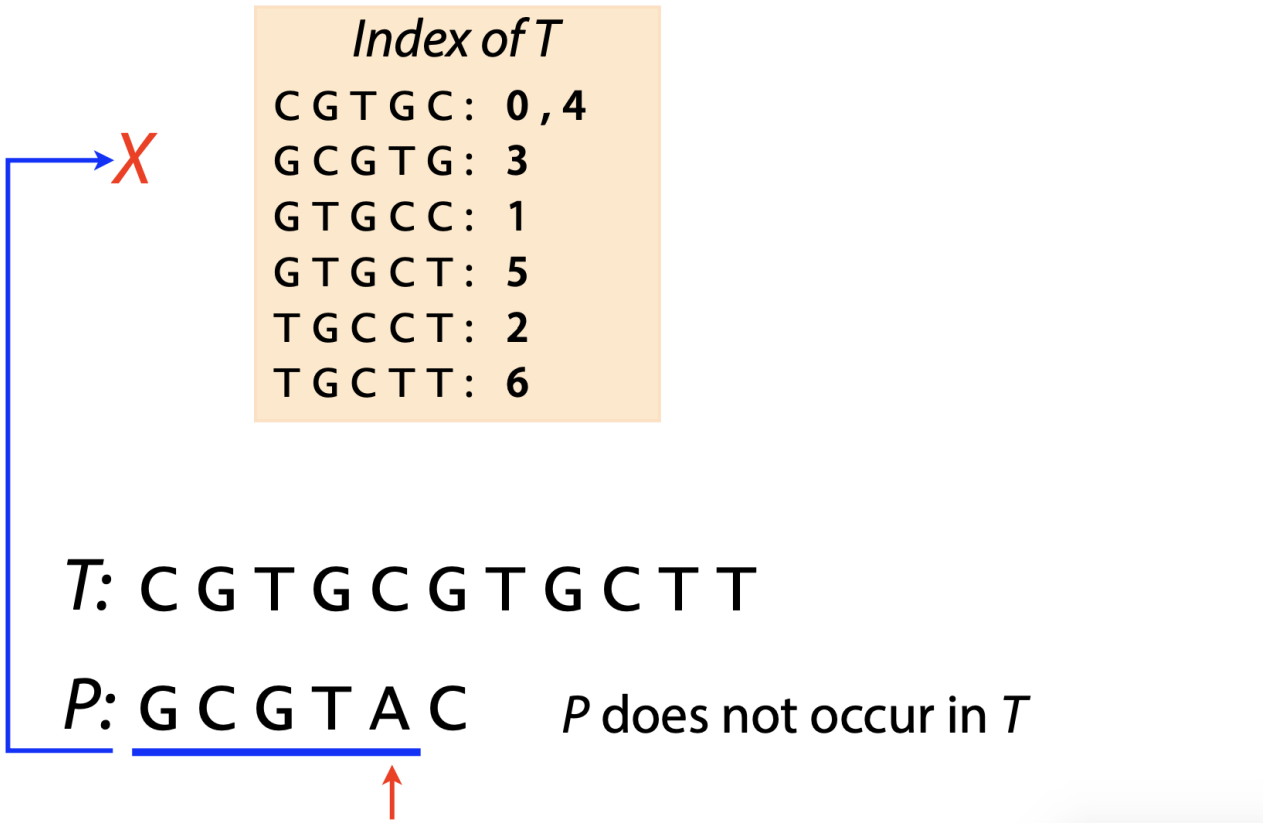
CGTGC: 0,4
GCGTG: 3
GTGCC: 1
GTGCT: 5
TGCCT: 2
TGCTT: 6

T: C G T G C G T G C T T

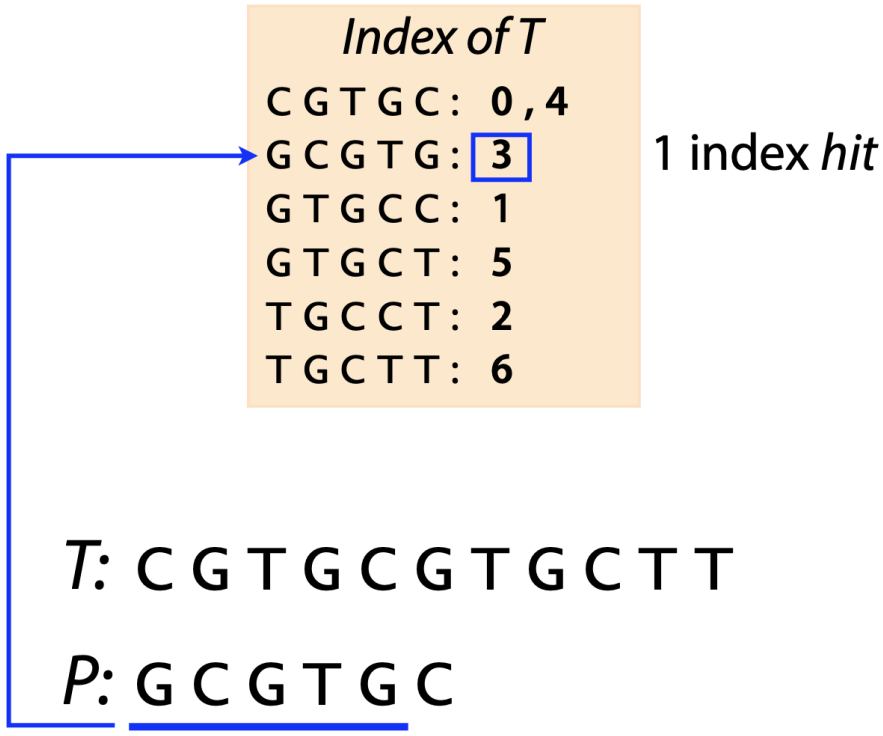
P: G C G T A C

BRAC UNIVERSITY
Inspiring Excellence

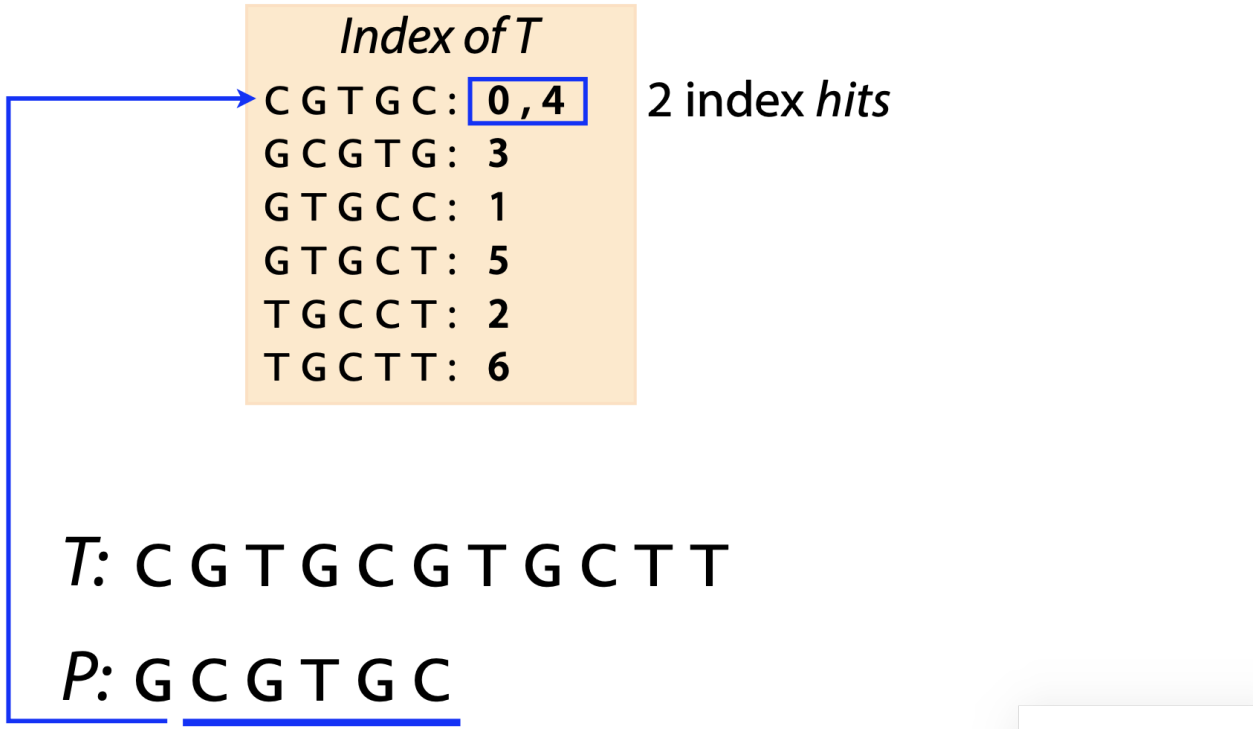
Indexing the Genome



Indexing the Genome



Indexing the Genome



Searching in a Multimap

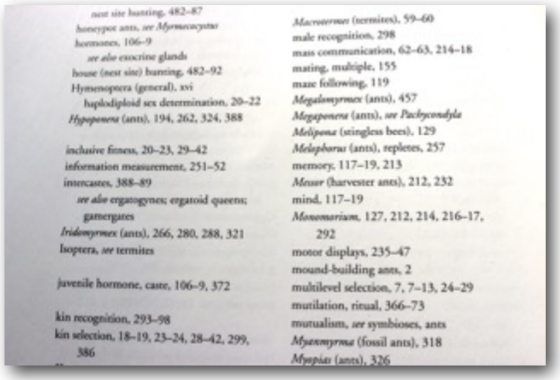
Index of T

CGTGC: 0,4
GCGTG: 3
GTGCC: 1
GTGCT: 5
TGCCT: 2
TGCTT: 6

Multimap

T: C G T G C G T G C T T

Searching in a Multimap



Multimap

T: C G T G C G T G C T T

Searching in a Multimap

GTG	0

T: G T G C G T G T G G G G

Searching in a Multimap

G T G	0
T G C	1

T: G T G C G T G T G G G G

Searching in a Multimap

G T G	0
T G C	1
G C G	2

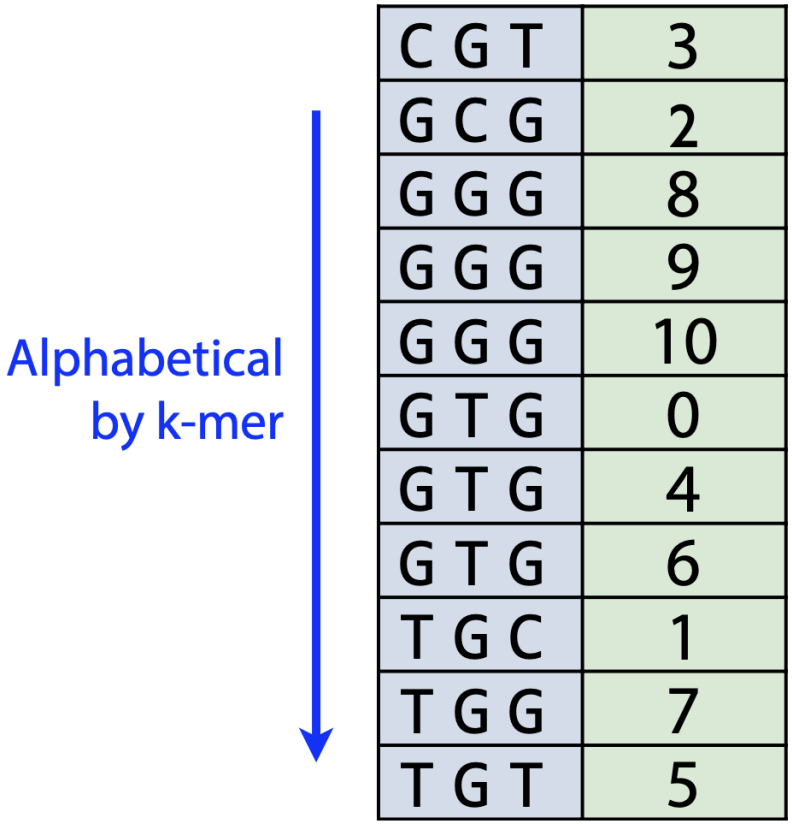
T: G T G C G T G T G G G G

Searching in a Multimap

G T G	0
T G C	1
G C G	2
C G T	3
G T G	4
T G T	5
G T G	6
T G G	7
G G G	8
G G G	9
G G G	10

T: G T G C G T G T G G G G G

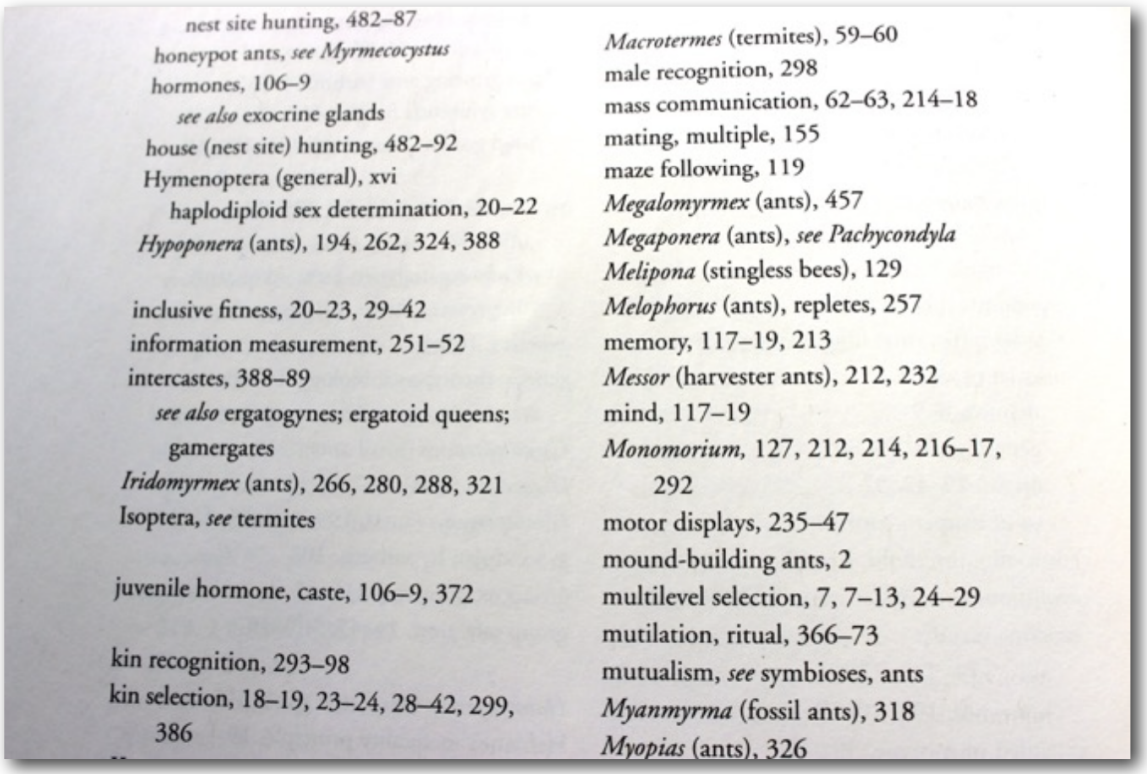
Searching in a Multimap



C G T	3
G C G	2
G G G	8
G G G	9
G G G	10
G T G	0
G T G	4
G T G	6
T G C	1
T G G	7
T G T	5

T: G T G C G T G T G G G G G

Searching in a Multimap



nest site hunting, 482–87
honeypot ants, *see Myrmecocystus*
hormones, 106–9
see also exocrine glands
house (nest site) hunting, 482–92
Hymenoptera (general), xvi
haplodiploid sex determination, 20–22
Hypoponera (ants), 194, 262, 324, 388

inclusive fitness, 20–23, 29–42
information measurement, 251–52
intercastes, 388–89
see also ergatogynes; ergatoid queens; gamergates
Iridomyrmex (ants), 266, 280, 288, 321
Isoptera, *see* termites

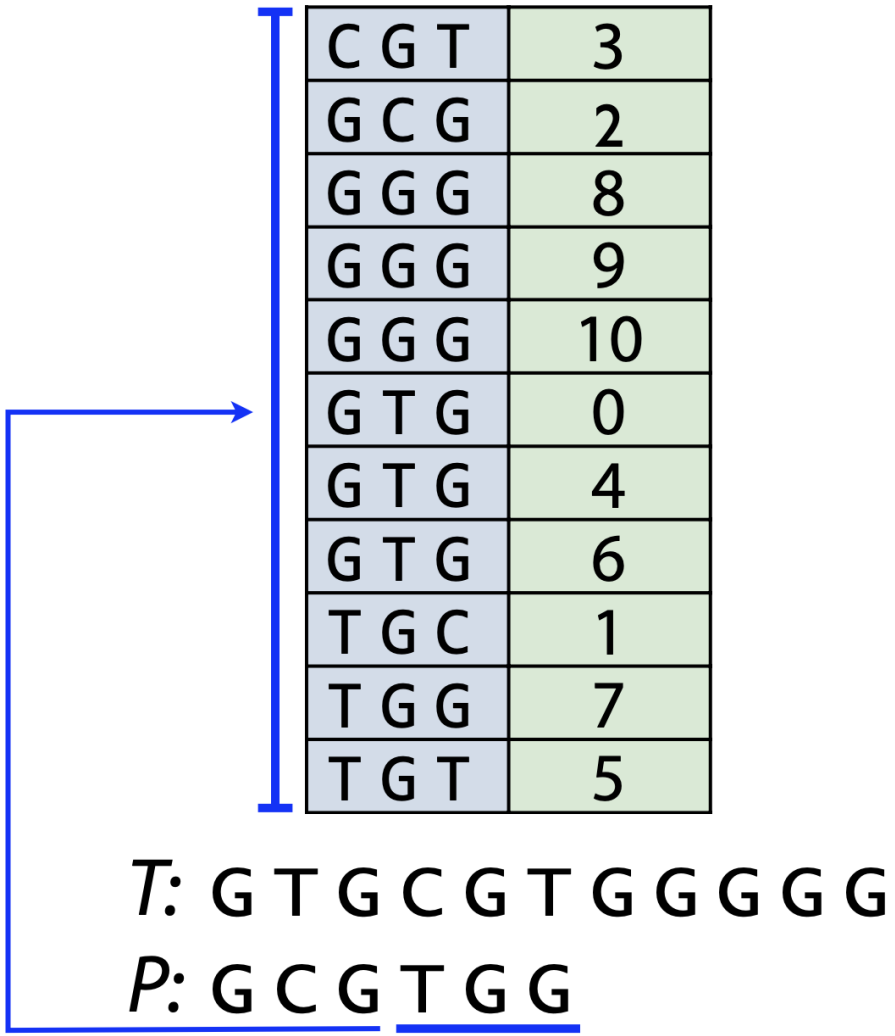
juvenile hormone, caste, 106–9, 372

kin recognition, 293–98
kin selection, 18–19, 23–24, 28–42, 299, 386

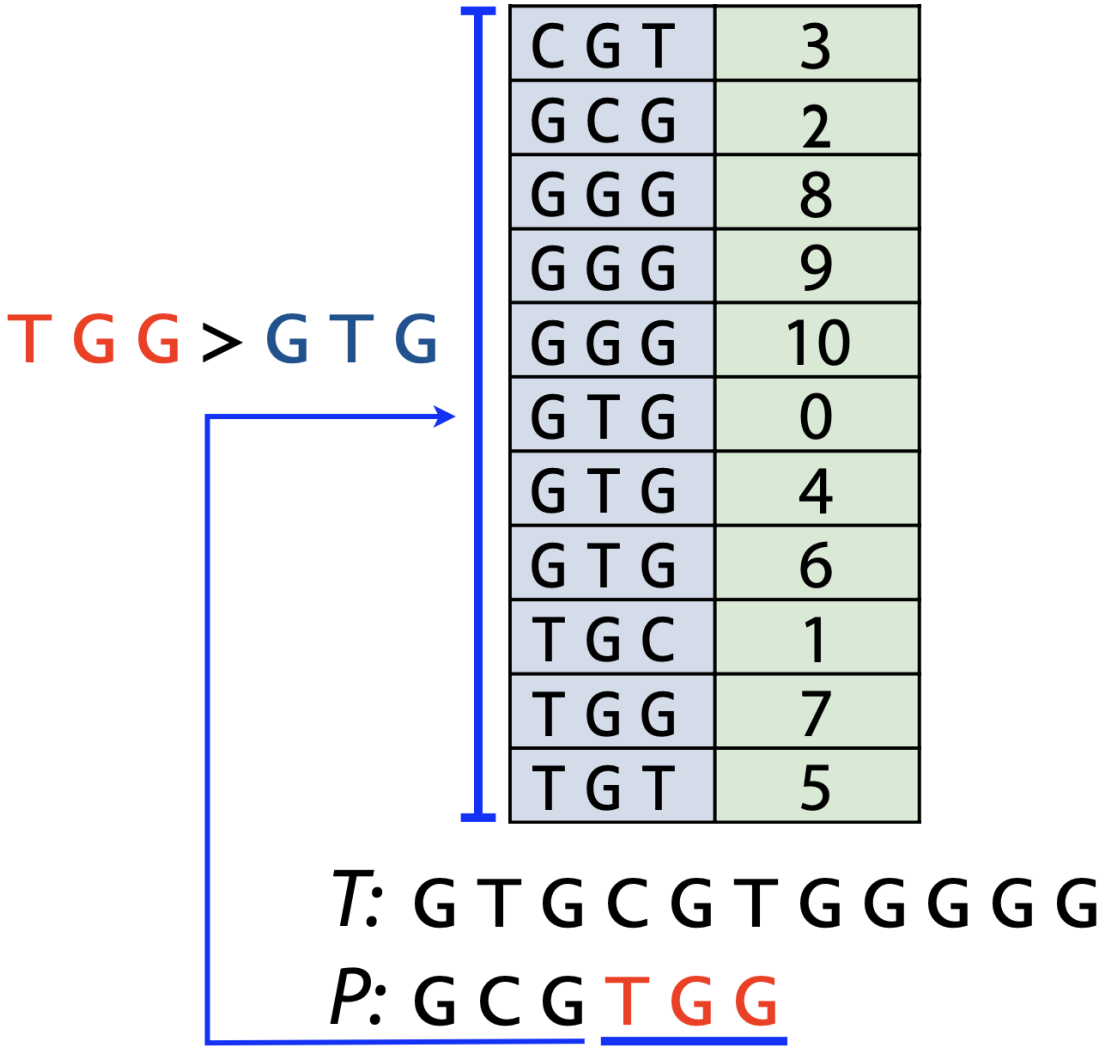
Macrotermes (termites), 59–60
male recognition, 298
mass communication, 62–63, 214–18
mating, multiple, 155
maze following, 119
Megalomyrmex (ants), 457
Megaponera (ants), *see Pachycondyla*
Melipona (stingless bees), 129
Melophorus (ants), repletes, 257
memory, 117–19, 213
Messor (harvester ants), 212, 232
mind, 117–19
Monomorium, 127, 212, 214, 216–17, 292
motor displays, 235–47
mound-building ants, 2
multilevel selection, 7, 7–13, 24–29
mutilation, ritual, 366–73
mutualism, *see* symbioses, ants
Myanmyrma (fossil ants), 318
Myopias (ants), 326

BRAC UNIVERSITY
Inspiring Excellence

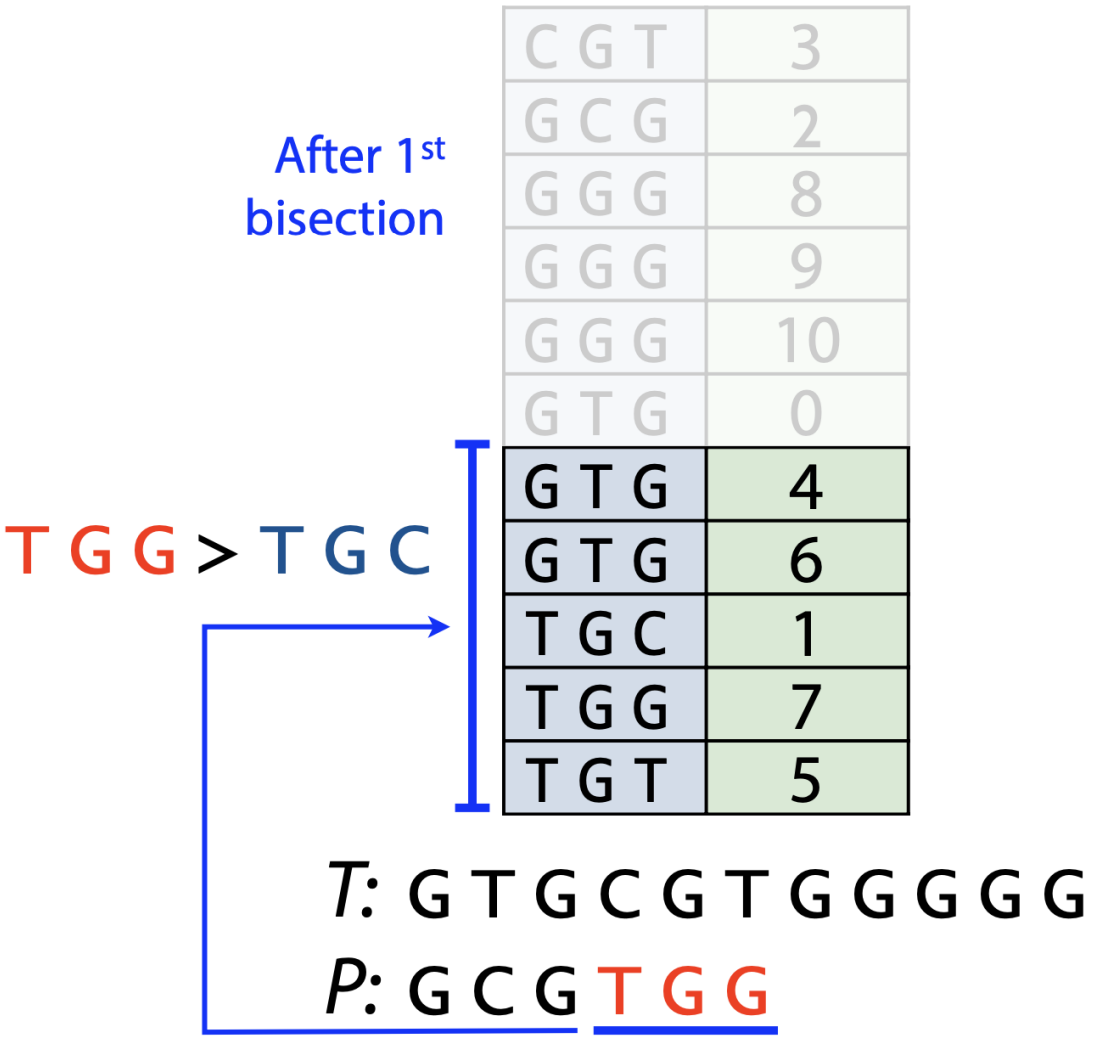
Searching in a Multimap



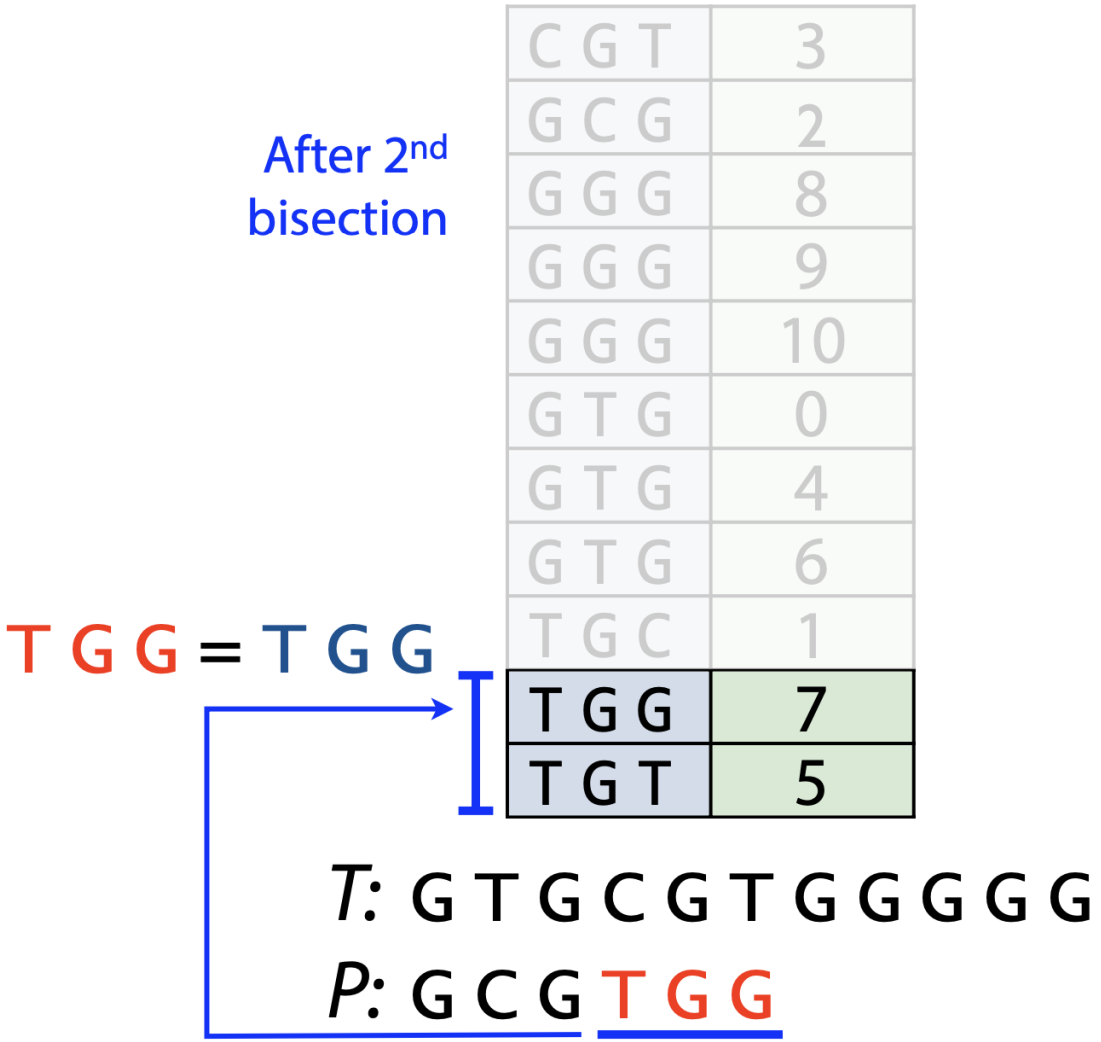
Searching in a Multimap



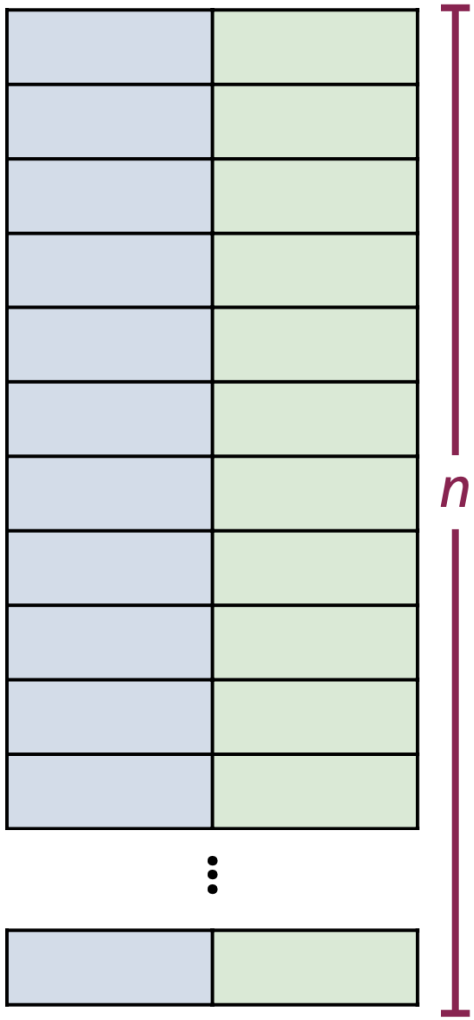
Searching in a Multimap



Searching in a Multimap

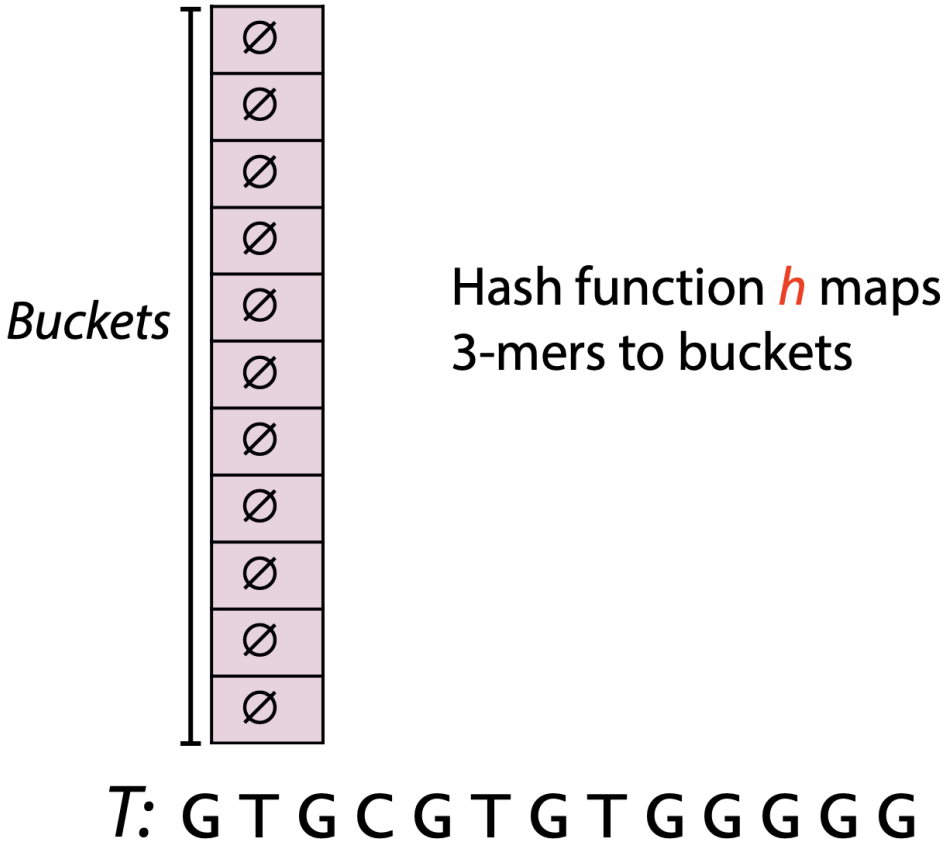


Searching in a Multimap

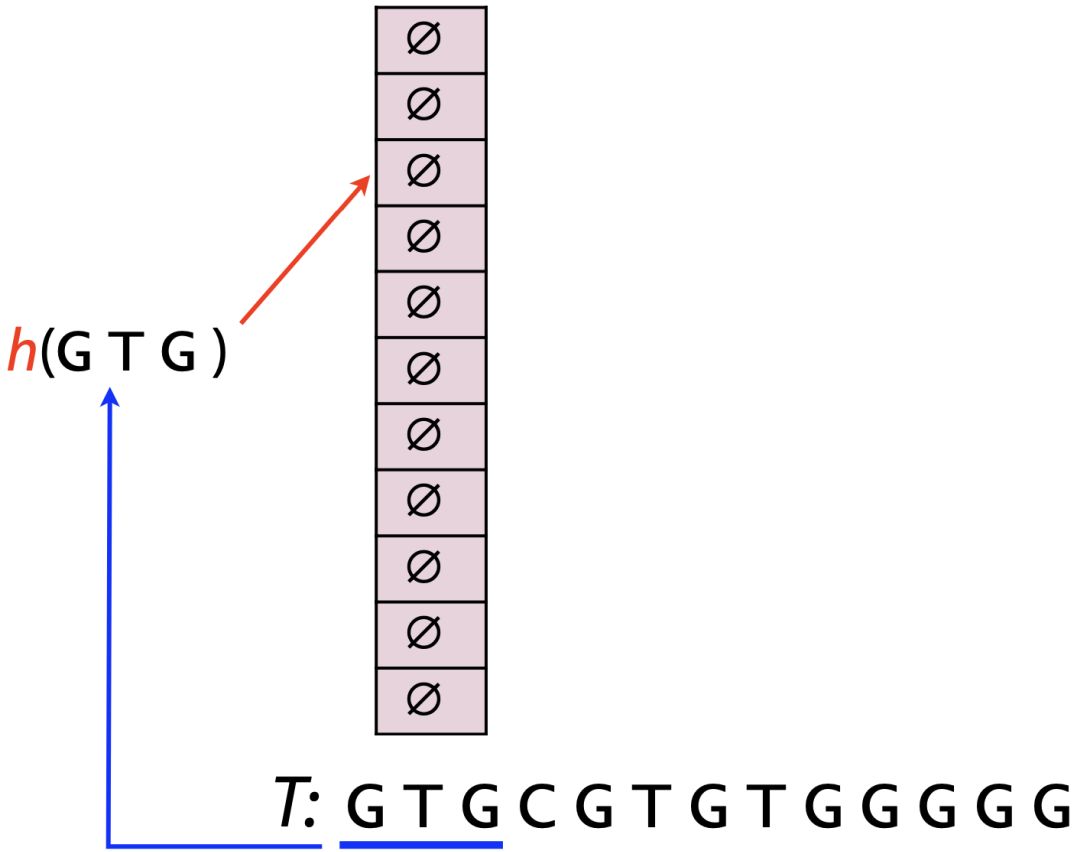


~ $\log_2(n)$ bisections per query

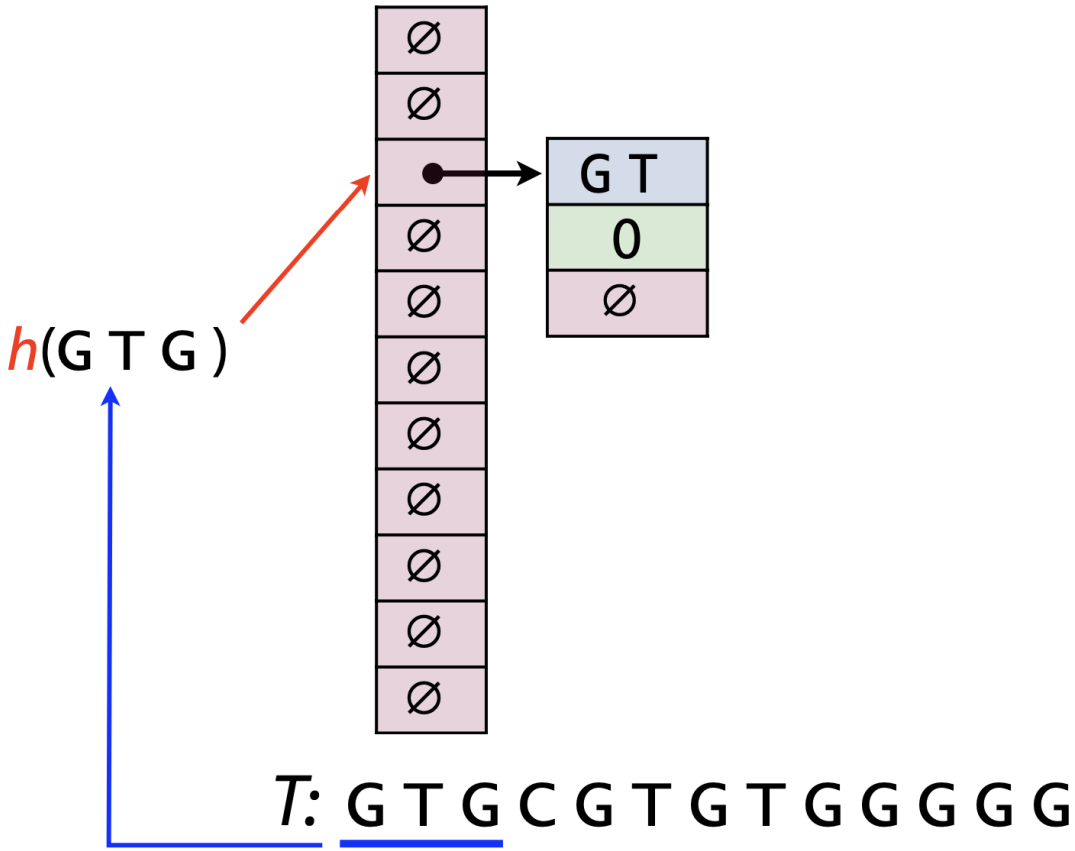
Searching in a Hash Table



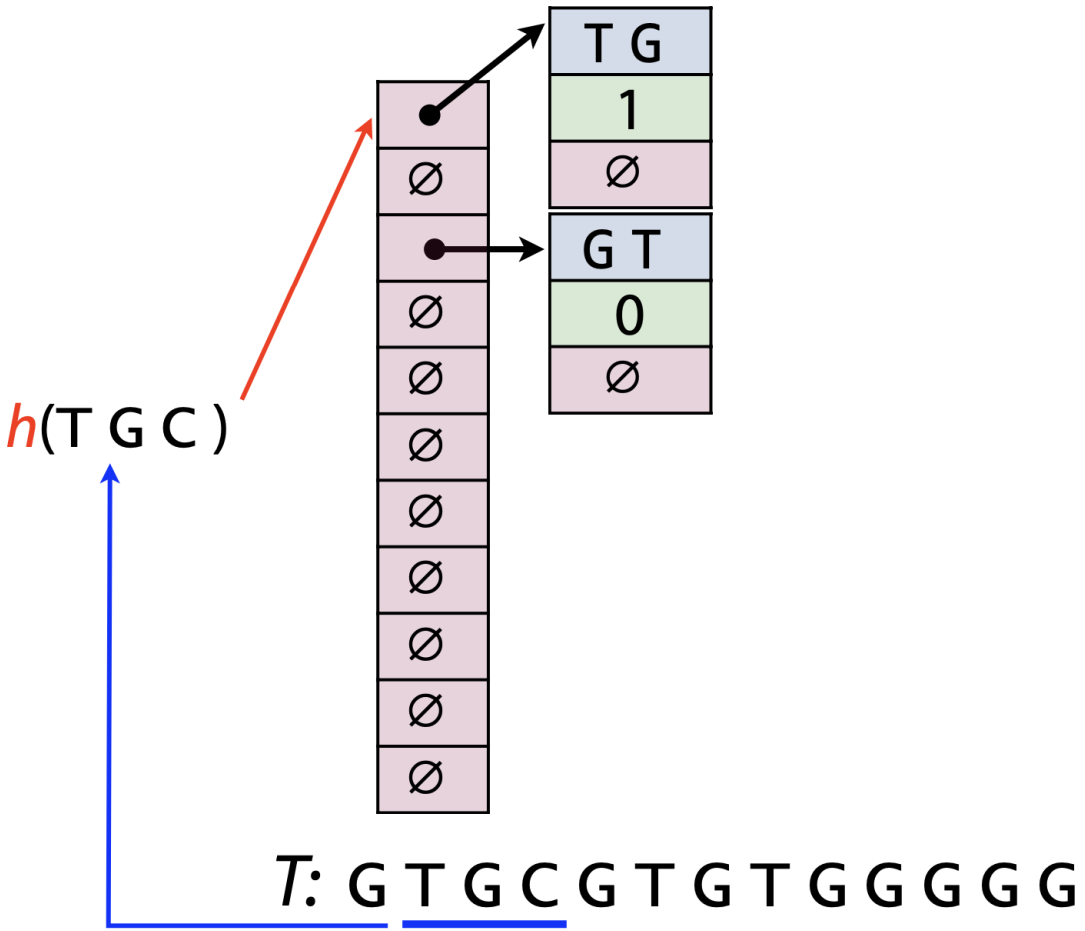
Searching in a Hash Table



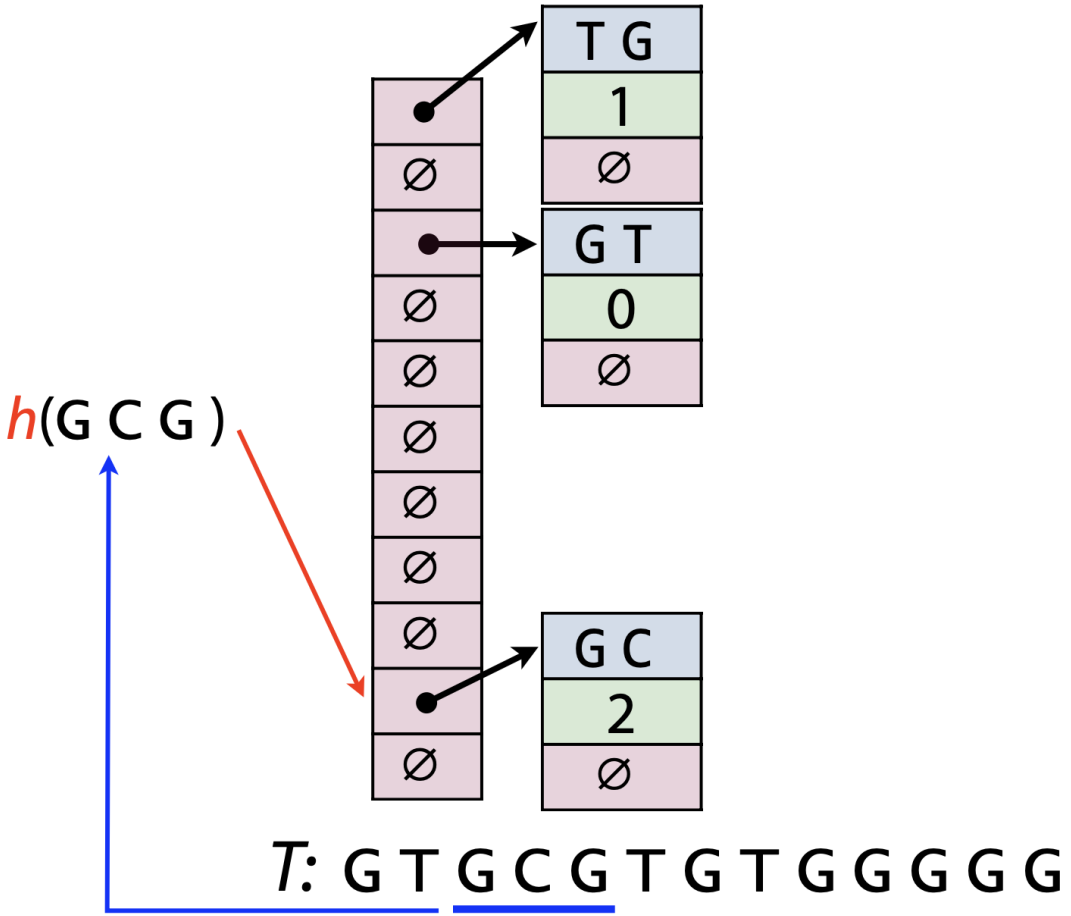
Searching in a Hash Table



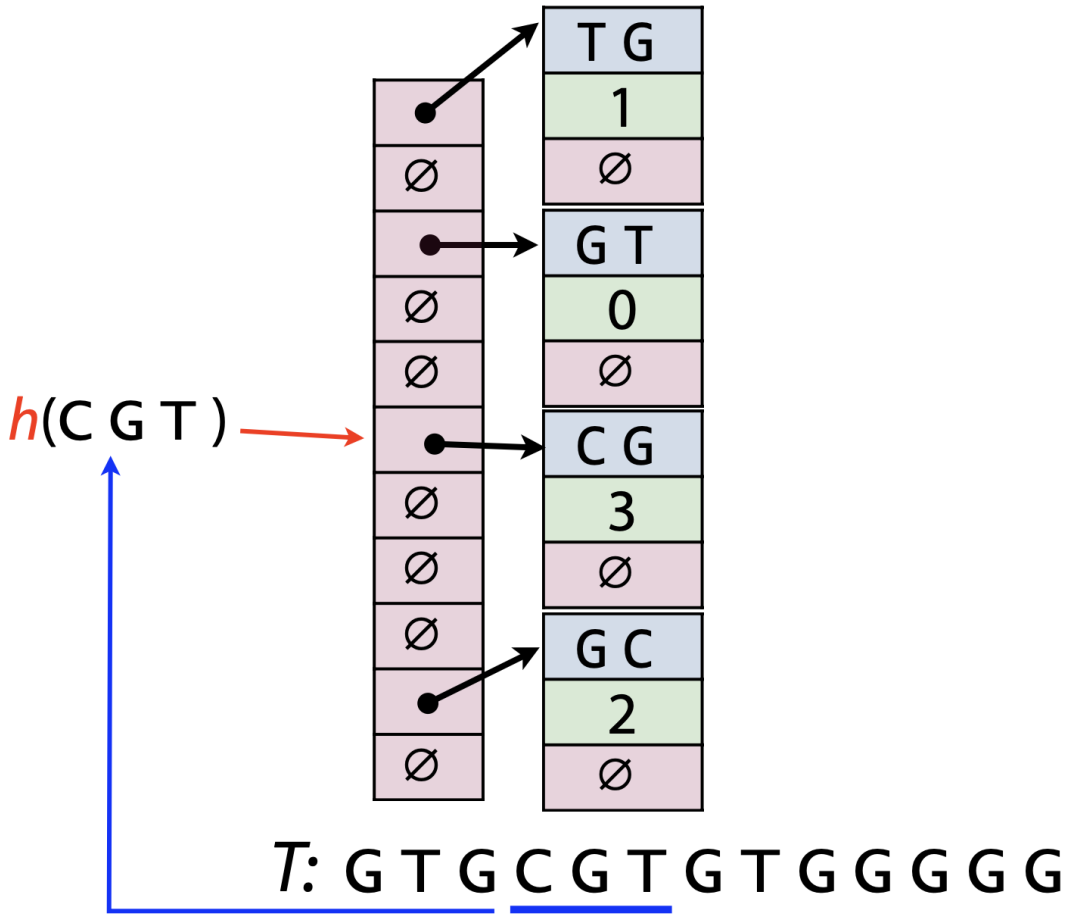
Searching in a Hash Table



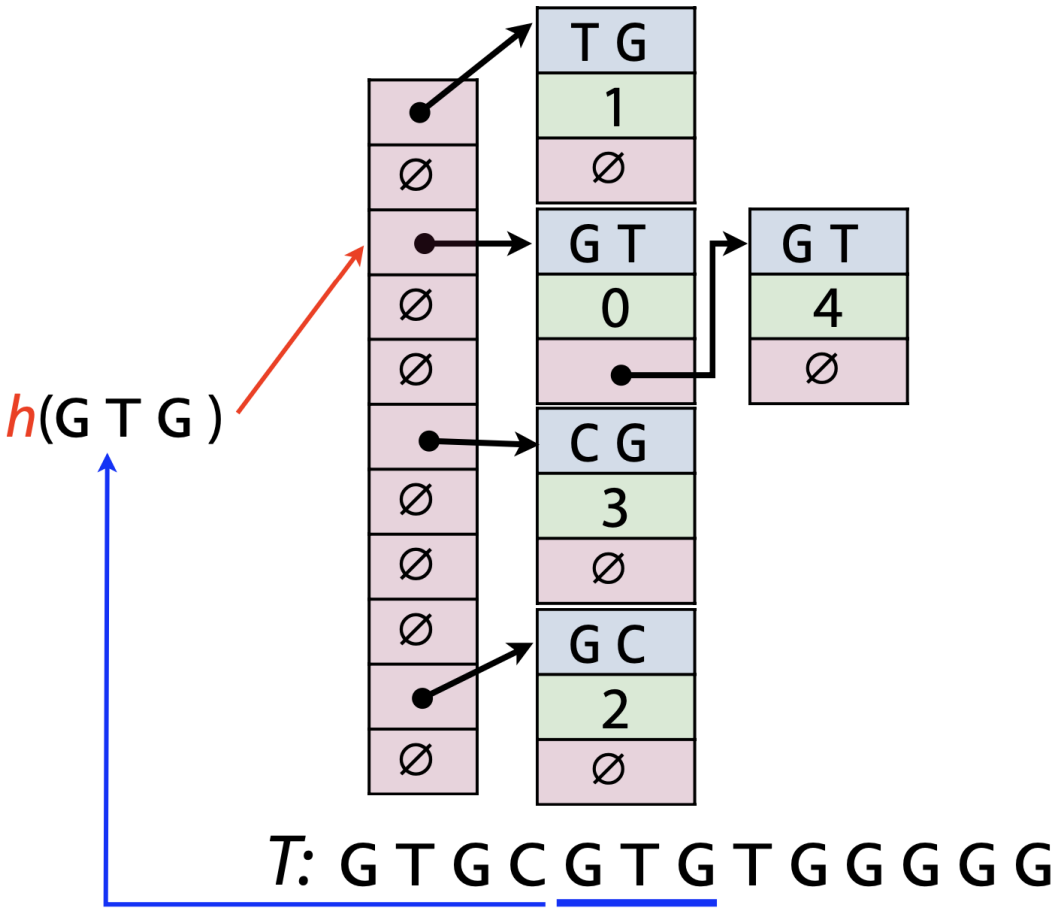
Searching in a Hash Table



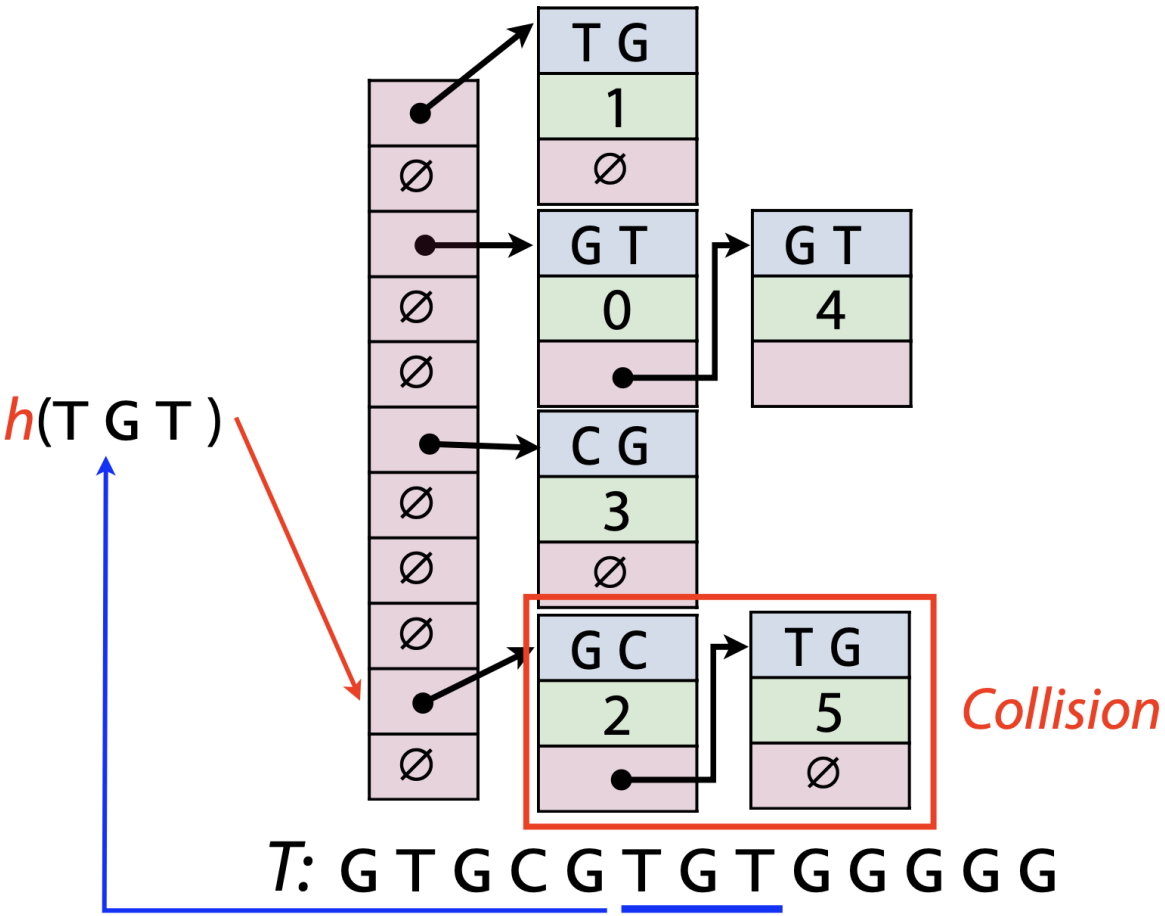
Searching in a Hash Table



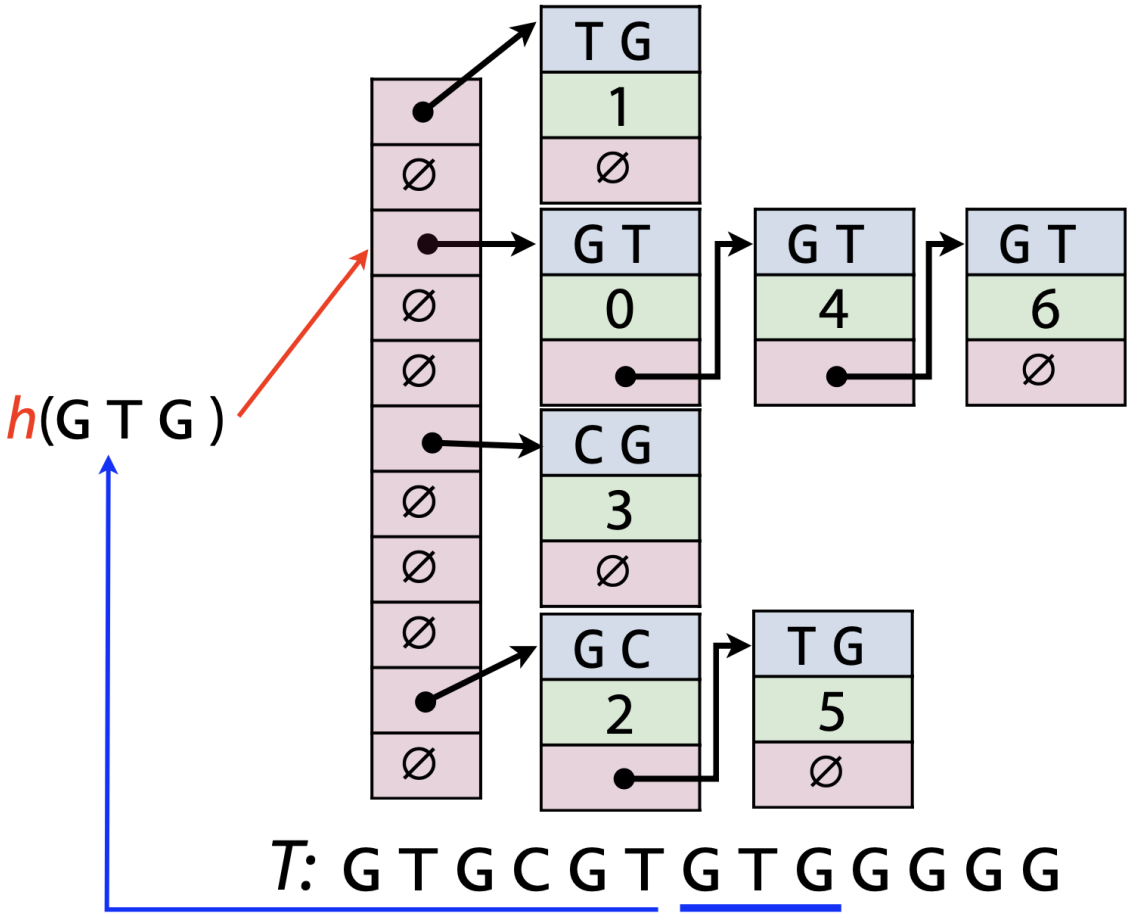
Searching in a Hash Table



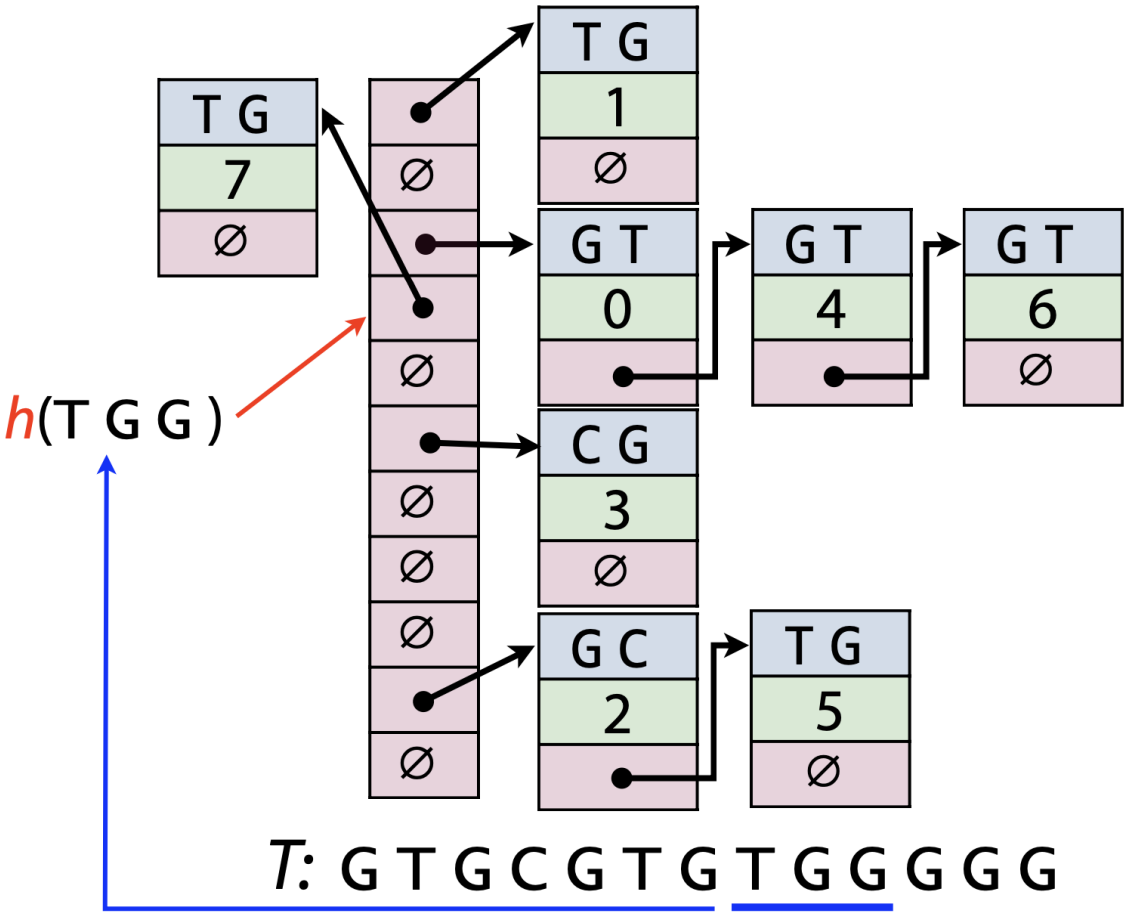
Searching in a Hash Table



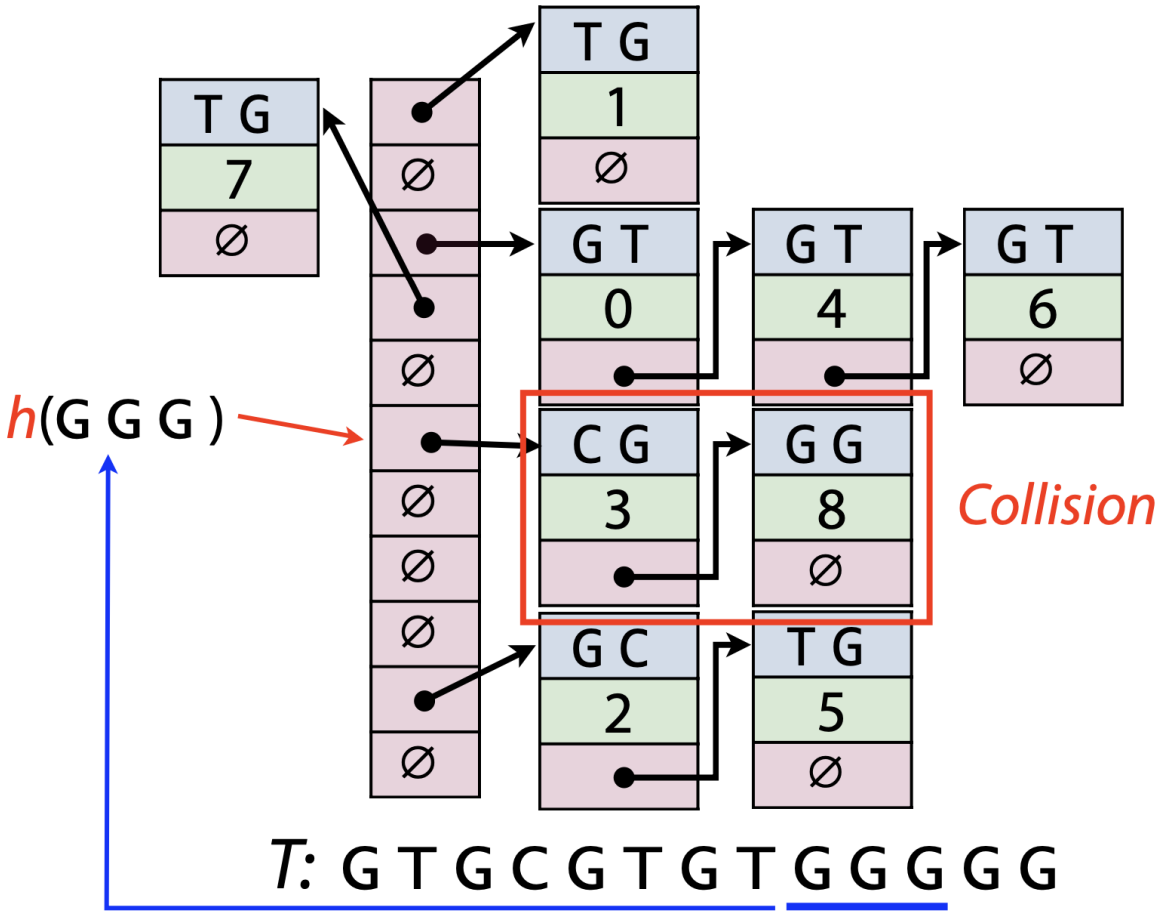
Searching in a Hash Table



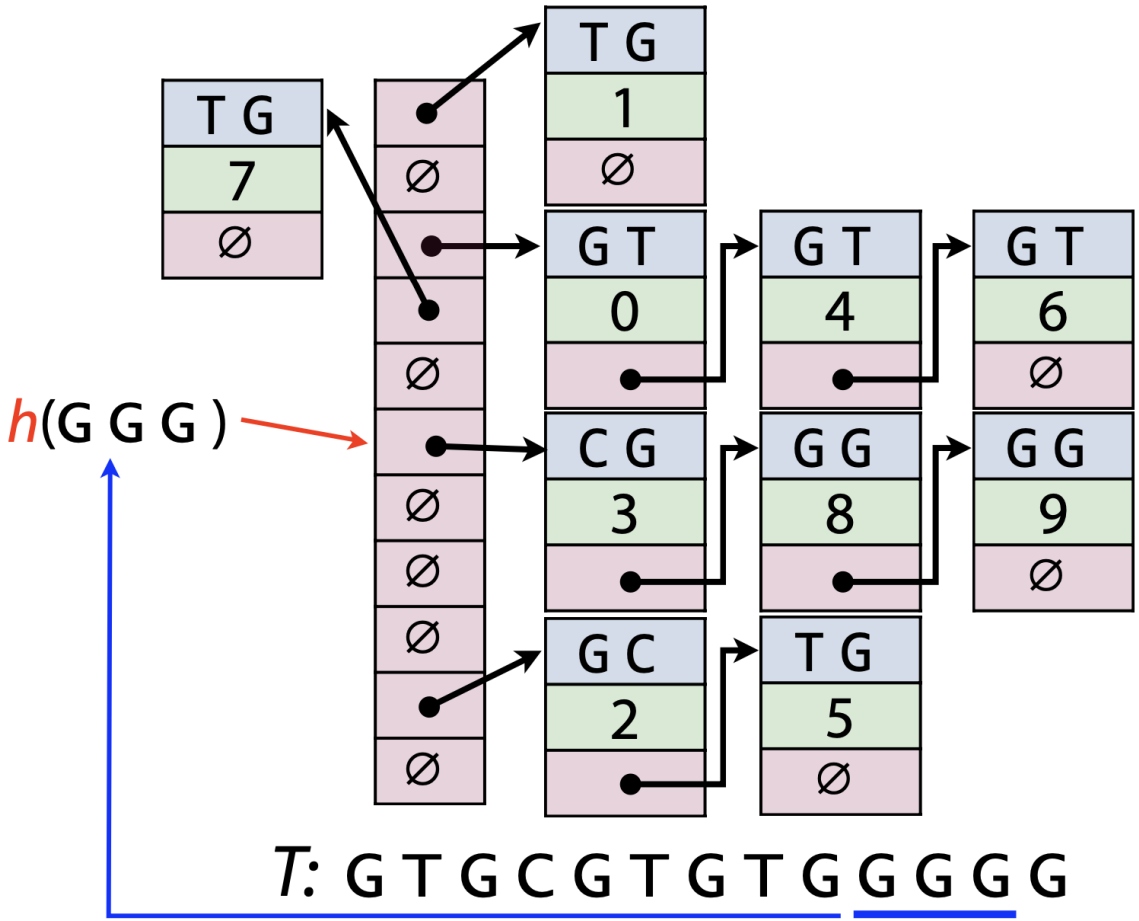
Searching in a Hash Table



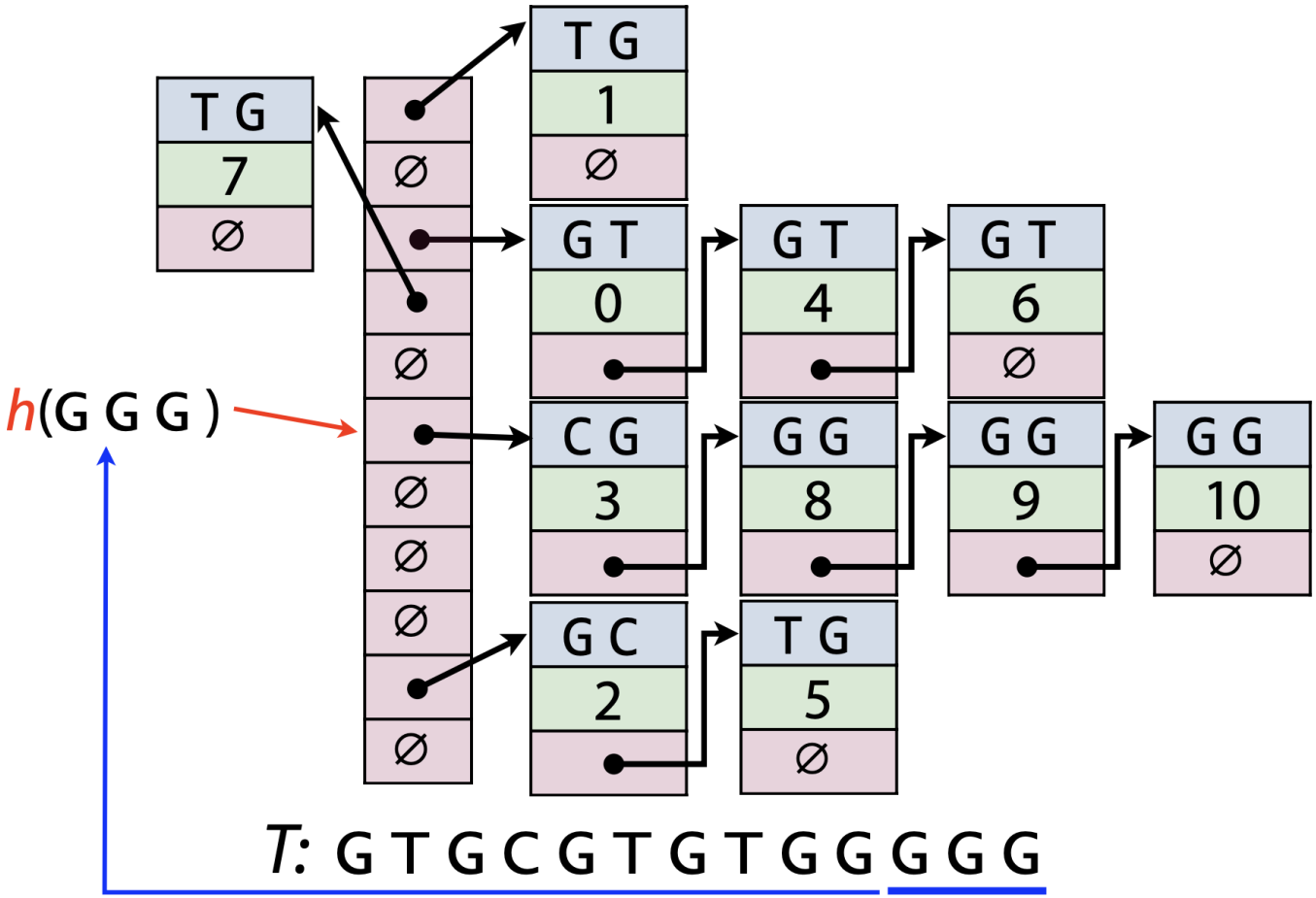
Searching in a Hash Table



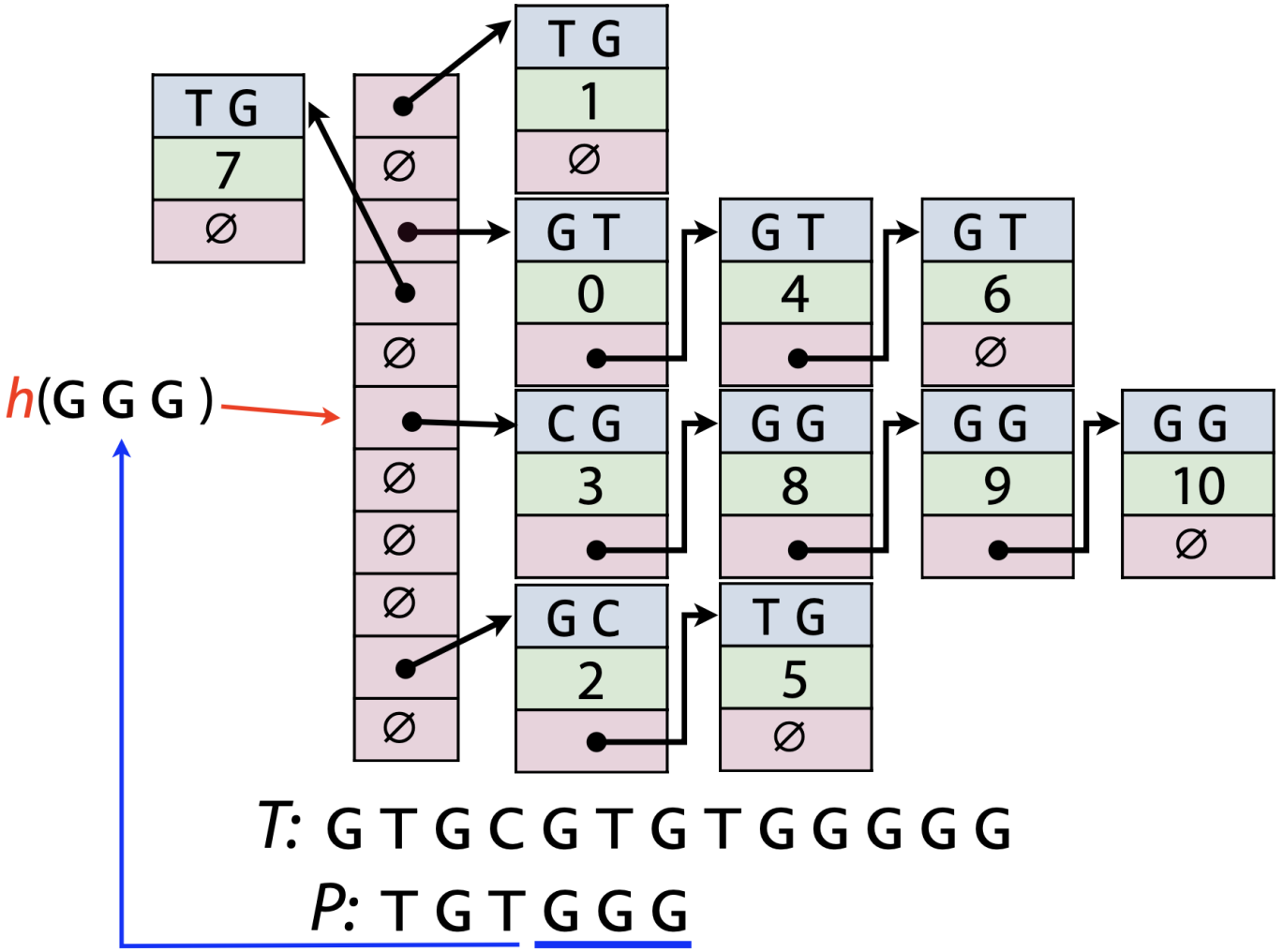
Searching in a Hash Table



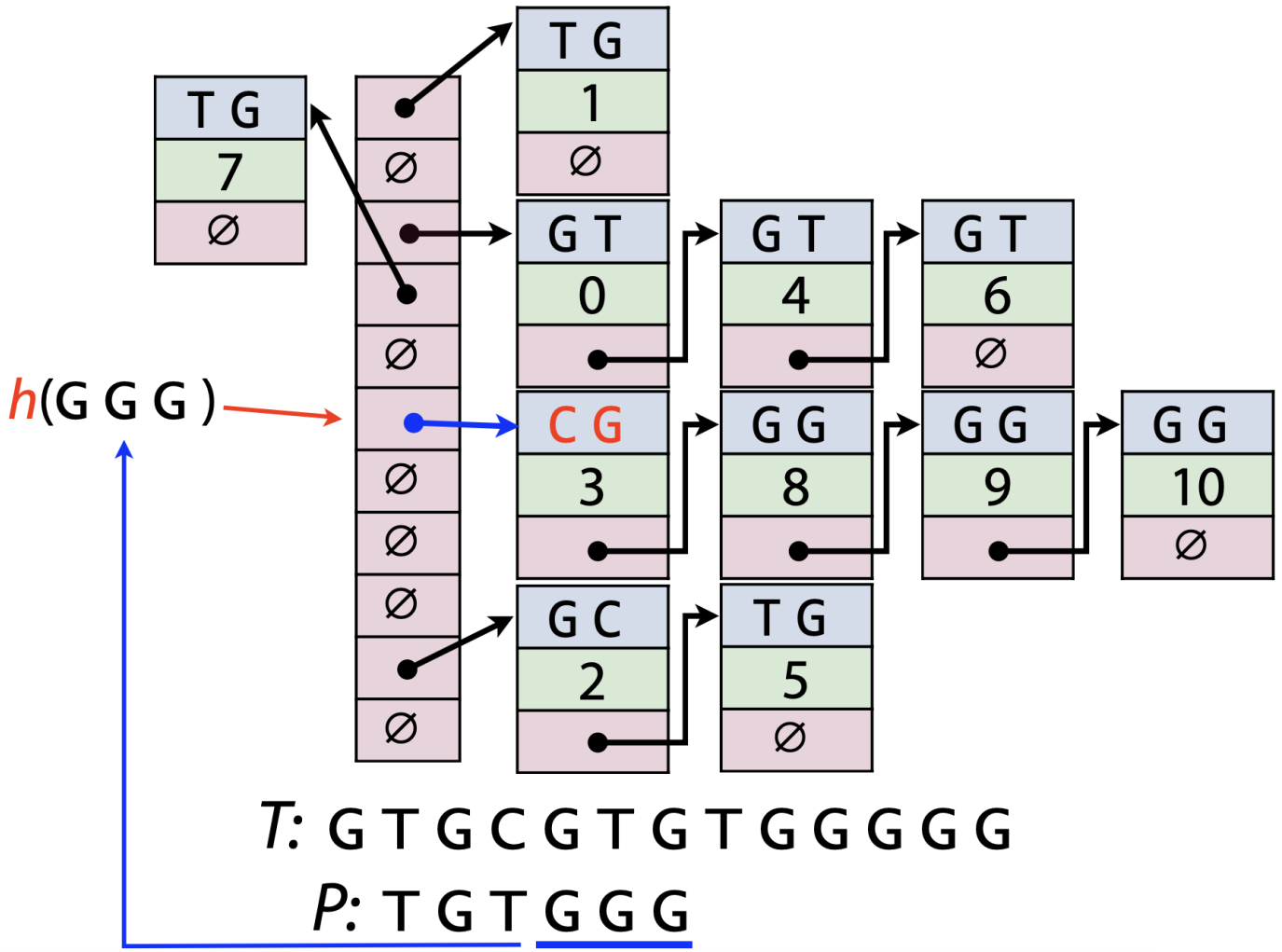
Searching in a Hash Table



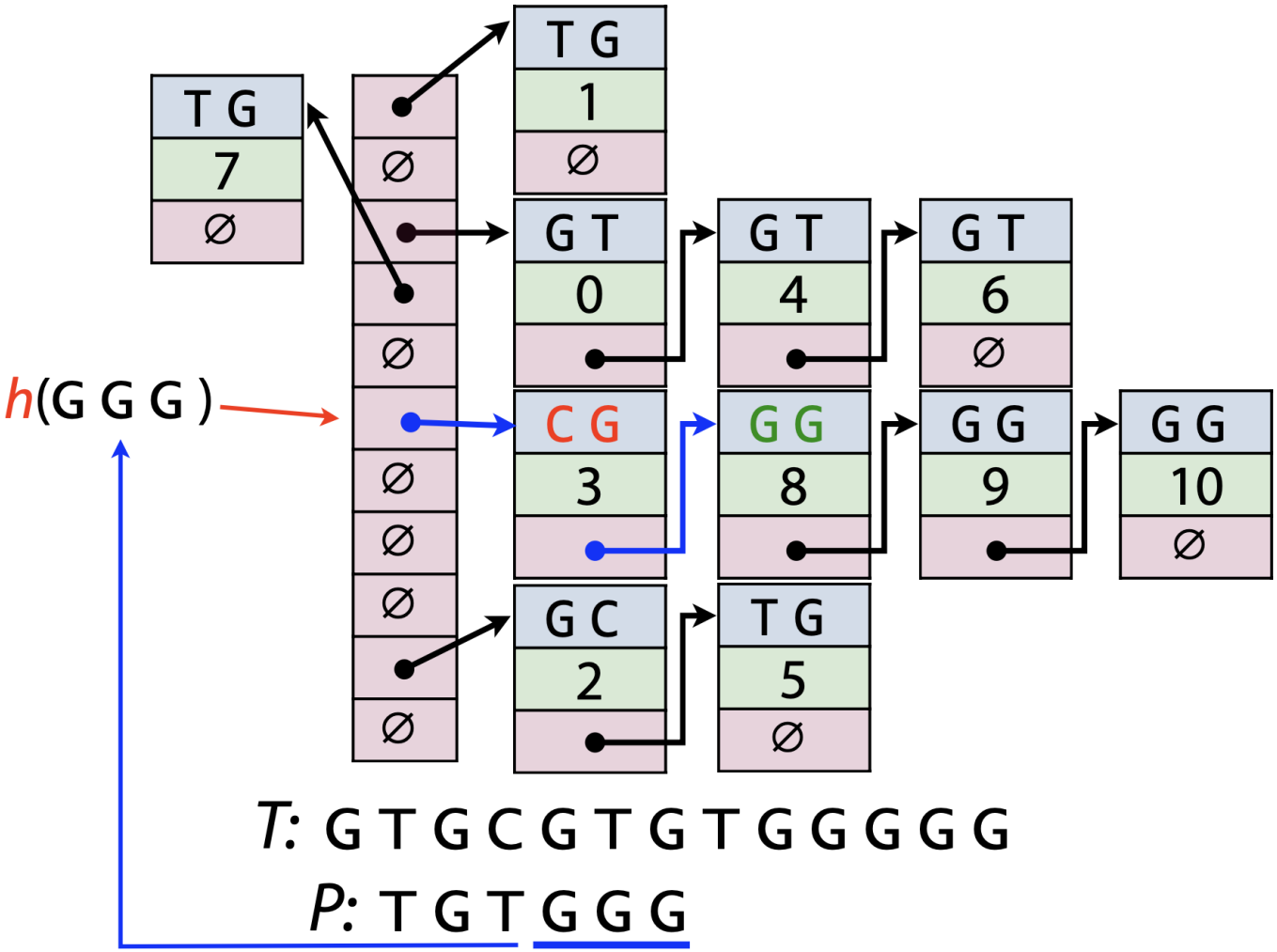
Searching in a Hash Table



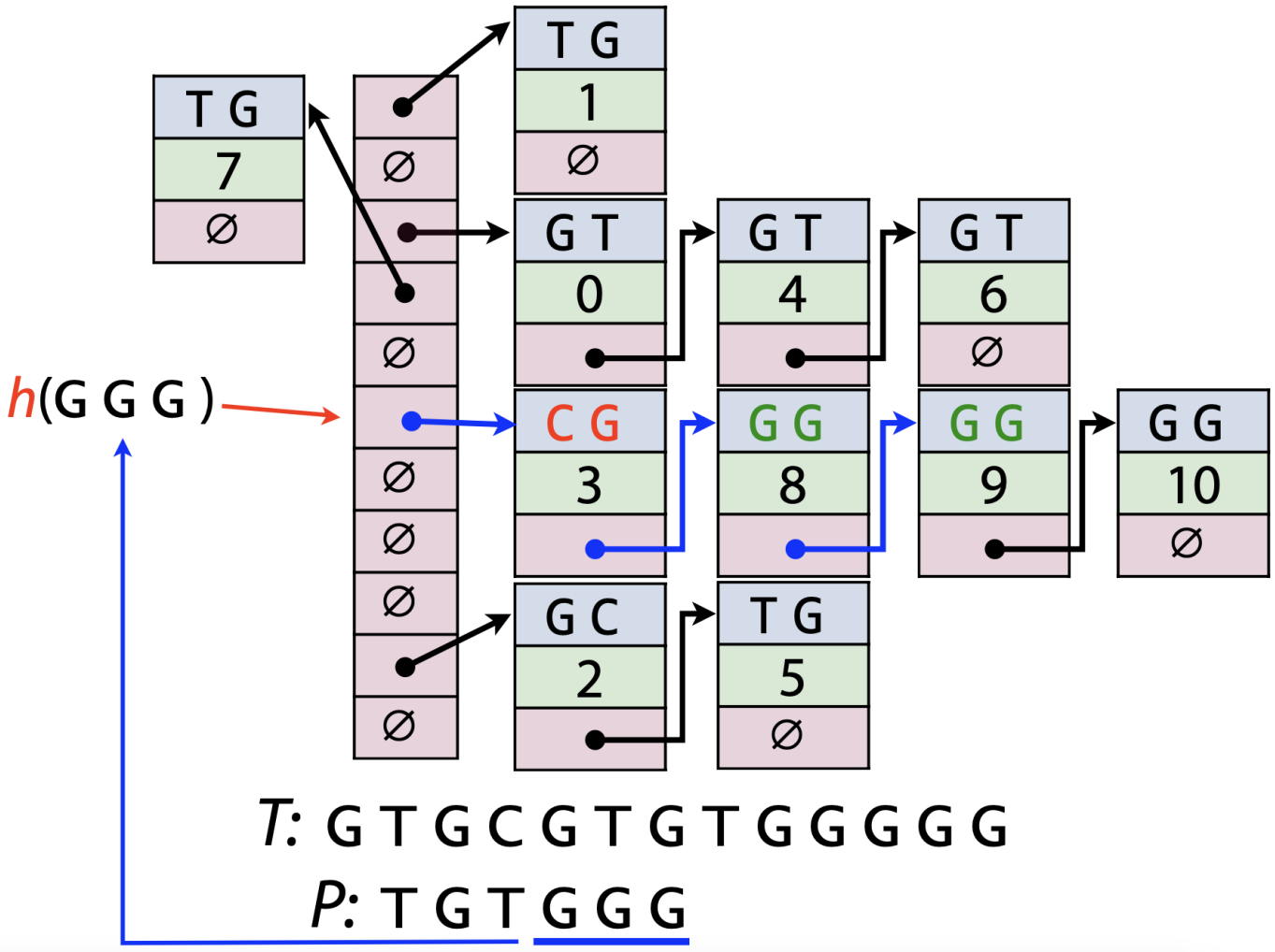
Searching in a Hash Table



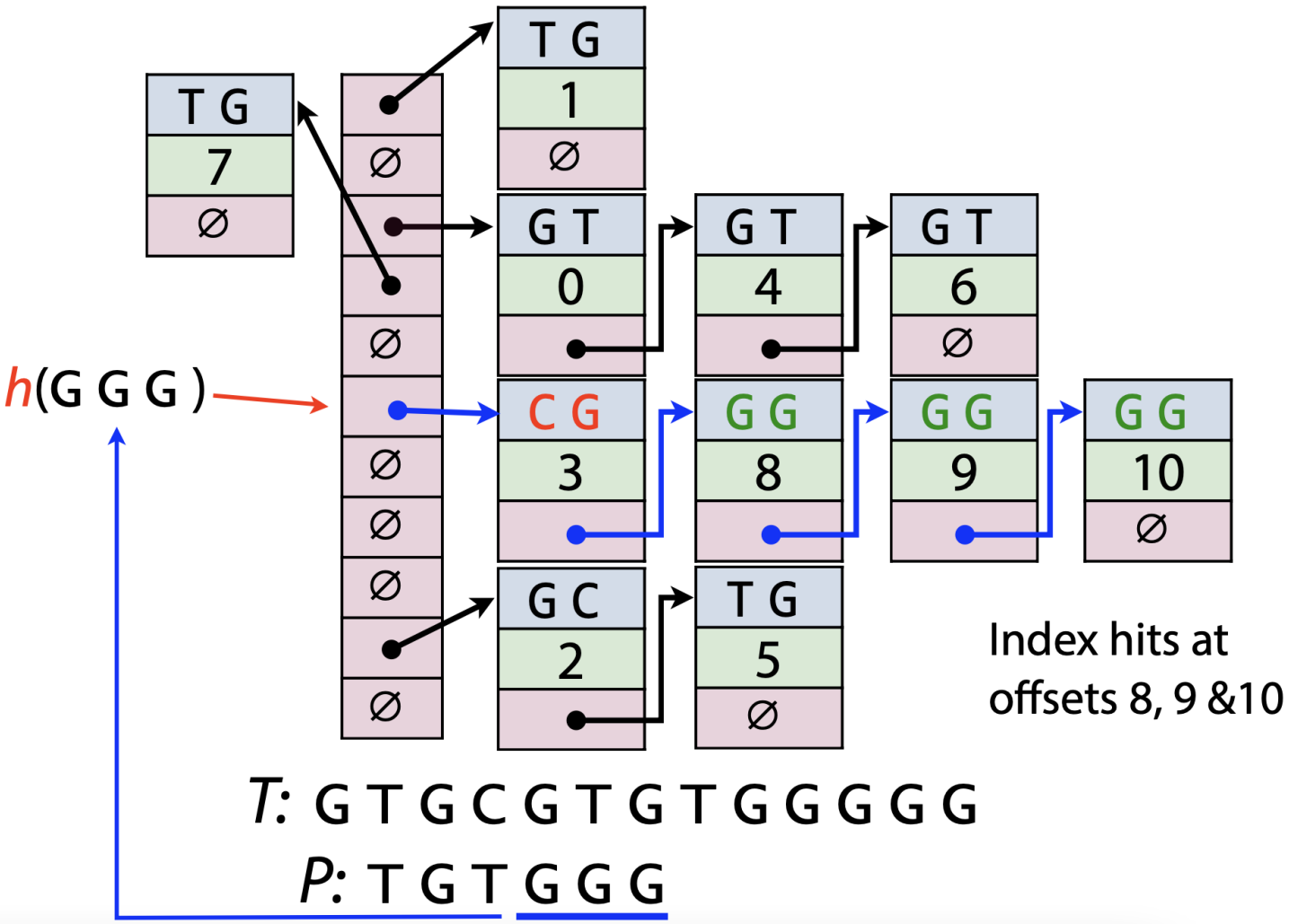
Searching in a Hash Table



Searching in a Hash Table



Searching in a Hash Table



Other Variants of Index Search

Index of T

CGTGC: 0,4
GCGTG: 3
GTGCC: 1
GTGCT: 5
TGCCT: 2
TGCTT: 6

T: C G T G C G T G C T T

Other Variants of Index Search

Index of T

CGTGC: 0,4
GCGTG: ~~3~~
GTGCC: ~~1~~
GTGCT: ~~5~~
TGCCT: 2
TGCTT: 6

T: C G T G C G T G C T T

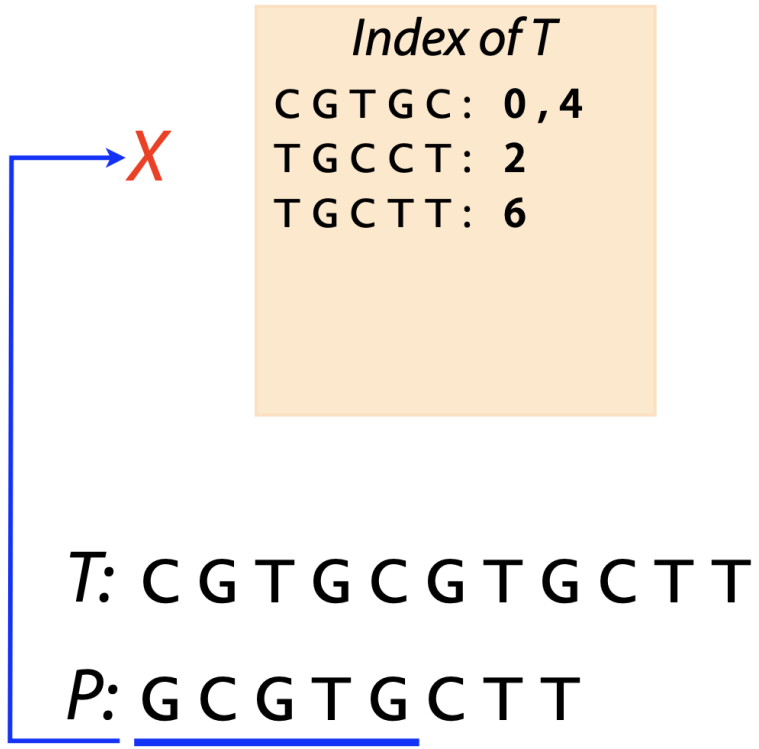
Other Variants of Index Search

Index of T

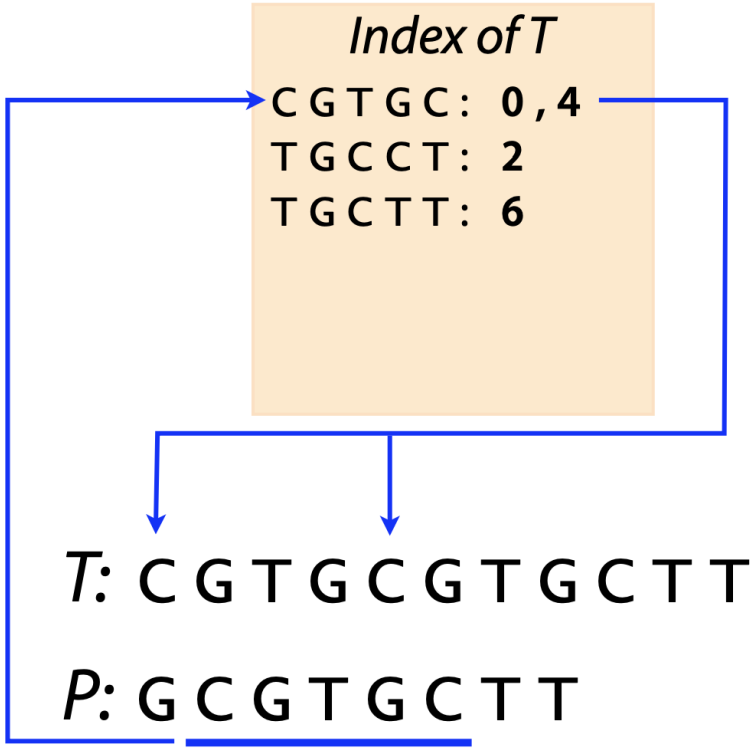
CGTGC: 0,4
TGCCT: 2
TGCTT: 6

T: C G T G C G T G C T T

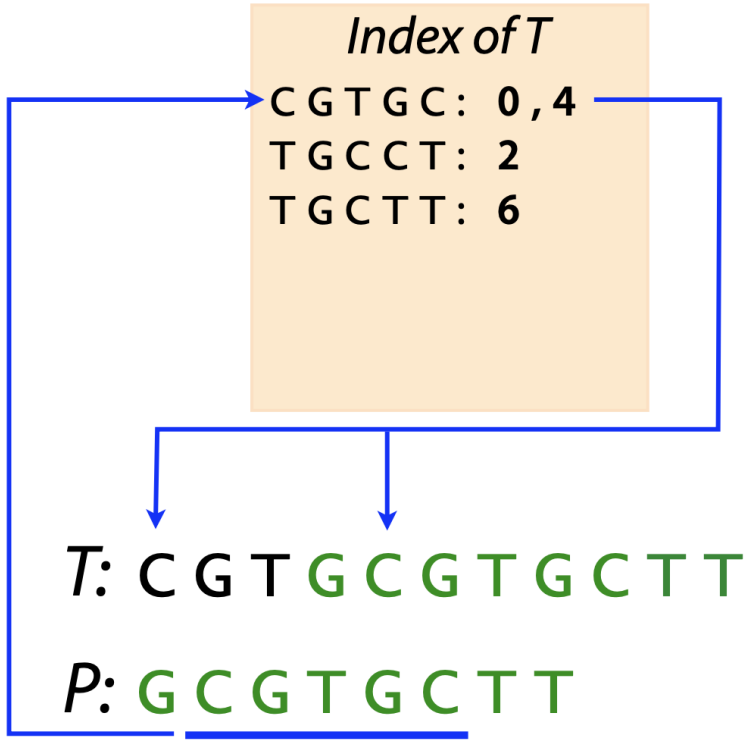
Other Variants of Index Search



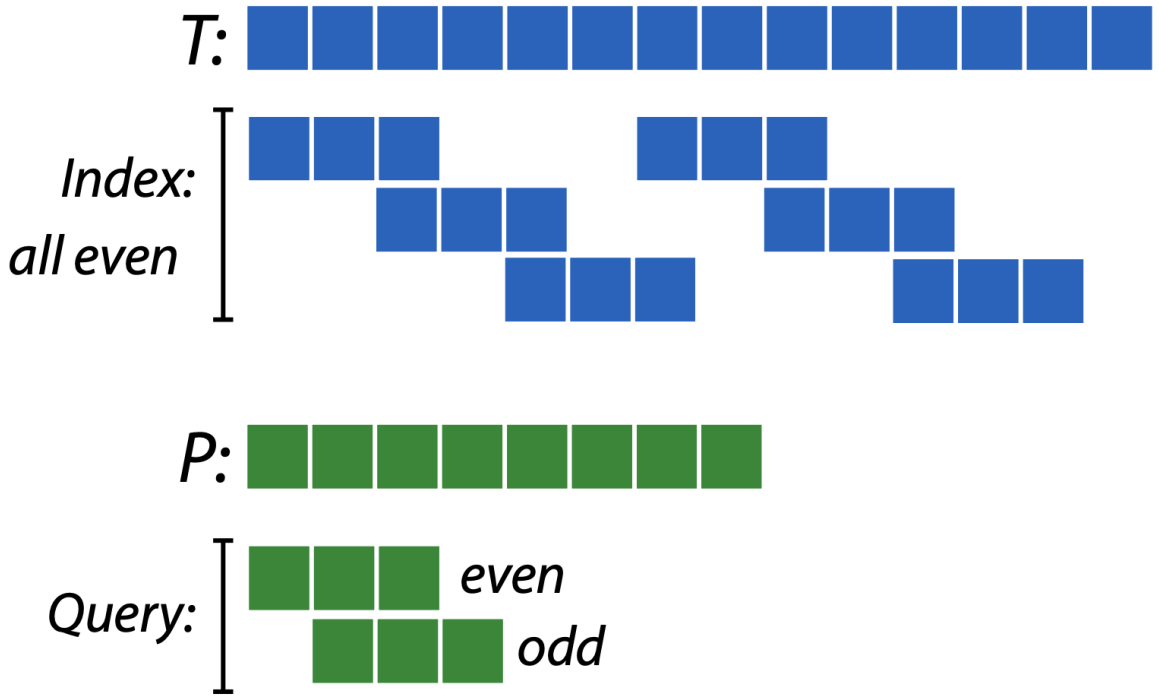
Other Variants of Index Search



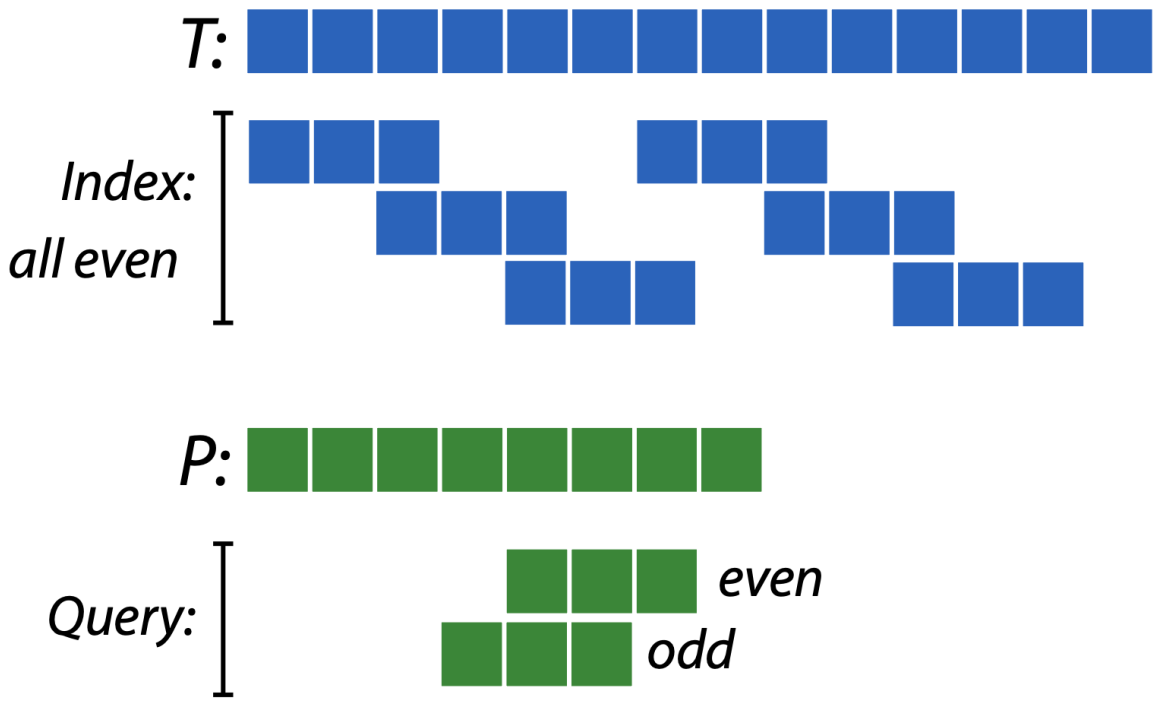
Other Variants of Index Search



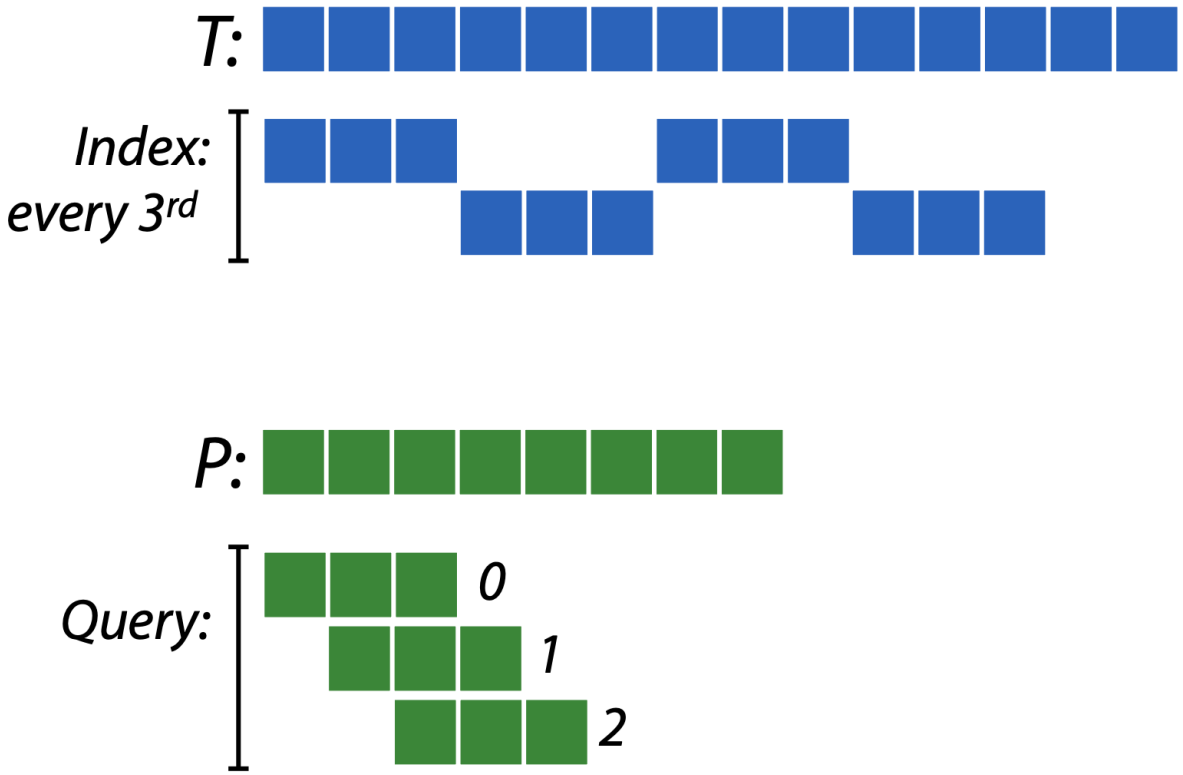
Other Variants of Index Search



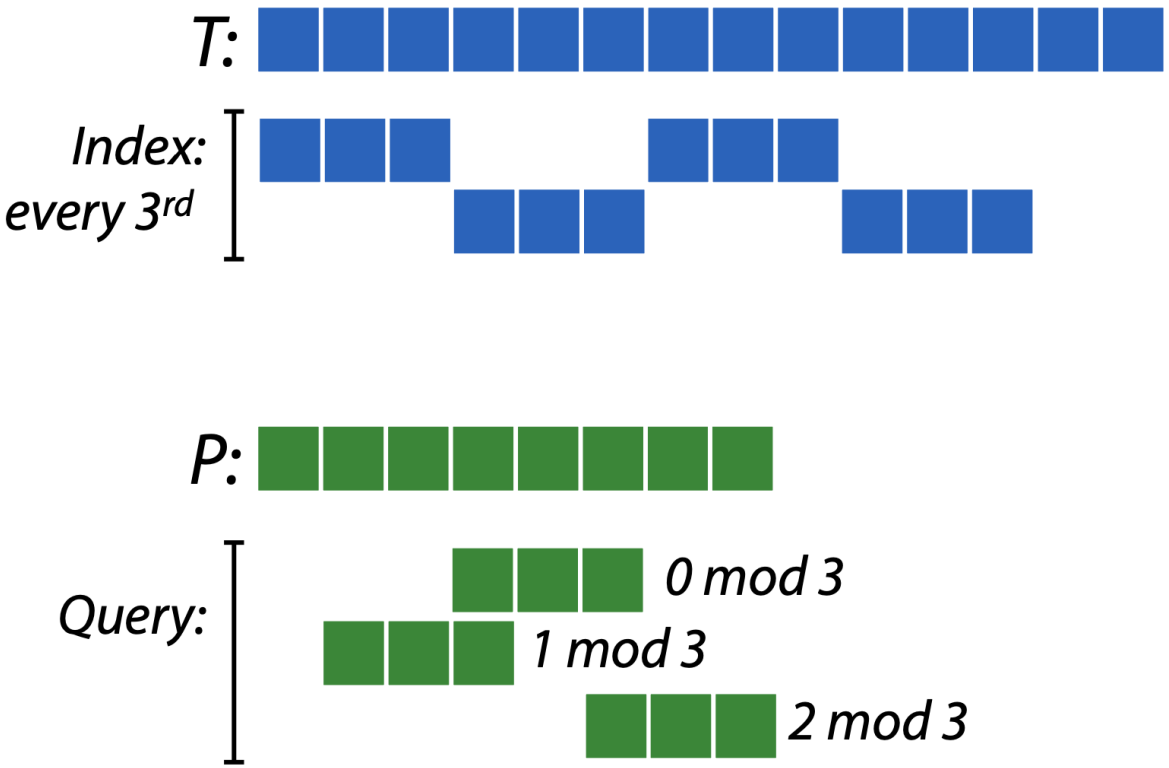
Other Variants of Index Search



Other Variants of Index Search



Other Variants of Index Search



Other Variants of Index Search

Subsequence of *S*: string of characters also occurring in *S* in the same order

```
>>> seq = 'AACCGGTT'
>>> seq[0] + seq[1] + seq[5] + seq[7]
'AAGT'  # subsequence
>>> seq.find('AAGT')
-1  # not a substring
```

Substrings are also subsequences, subsequences are not necessarily substrings

BRAC UNIVERSITY
Inspiring Excellence

Other Variants of Index Search

Index of T

C G G G T : 0

T: C G T G C G T G C T T

Other Variants of Index Search

Index of T

C G G G T : **0**

G T C T G : 1

T: C G T G C G T G C T T

Other Variants of Index Search

Index of T

CGGGT: **0**
GTCTG: **1**
TGGGC: **2**

T: C G T G C G T G C T T

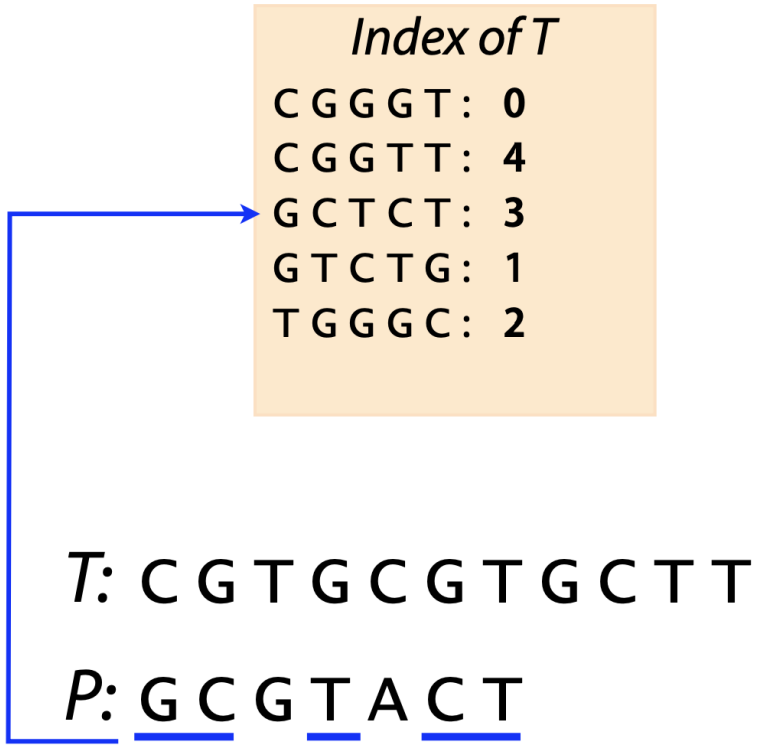
Other Variants of Index Search

Index of T

CGGGT: 0
CGGTT: 4
GCTCT: 3
GTCTG: 1
TGGGC: 2

T: C G T G C G T G C T T

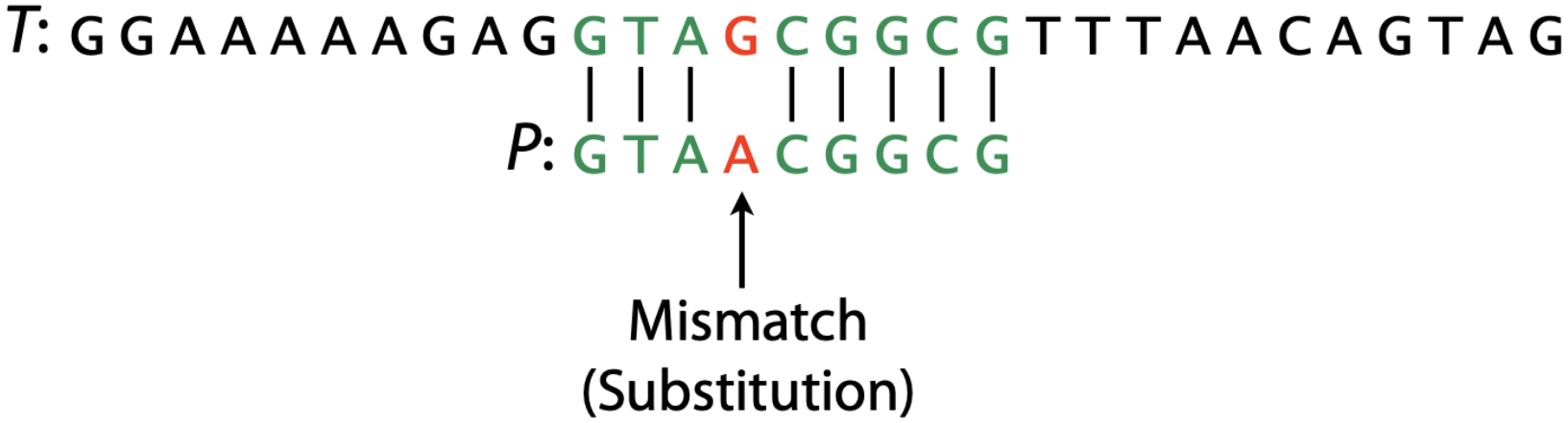
Other Variants of Index Search



Approximate Matching - Hamming Distance

Differences between read and reference occur because of:

- Sequencing error
- Natural variation



For X & Y where $|X| = |Y|$, *hamming distance* = minimum # substitutions needed to turn one into the other

Hamming distance = 3

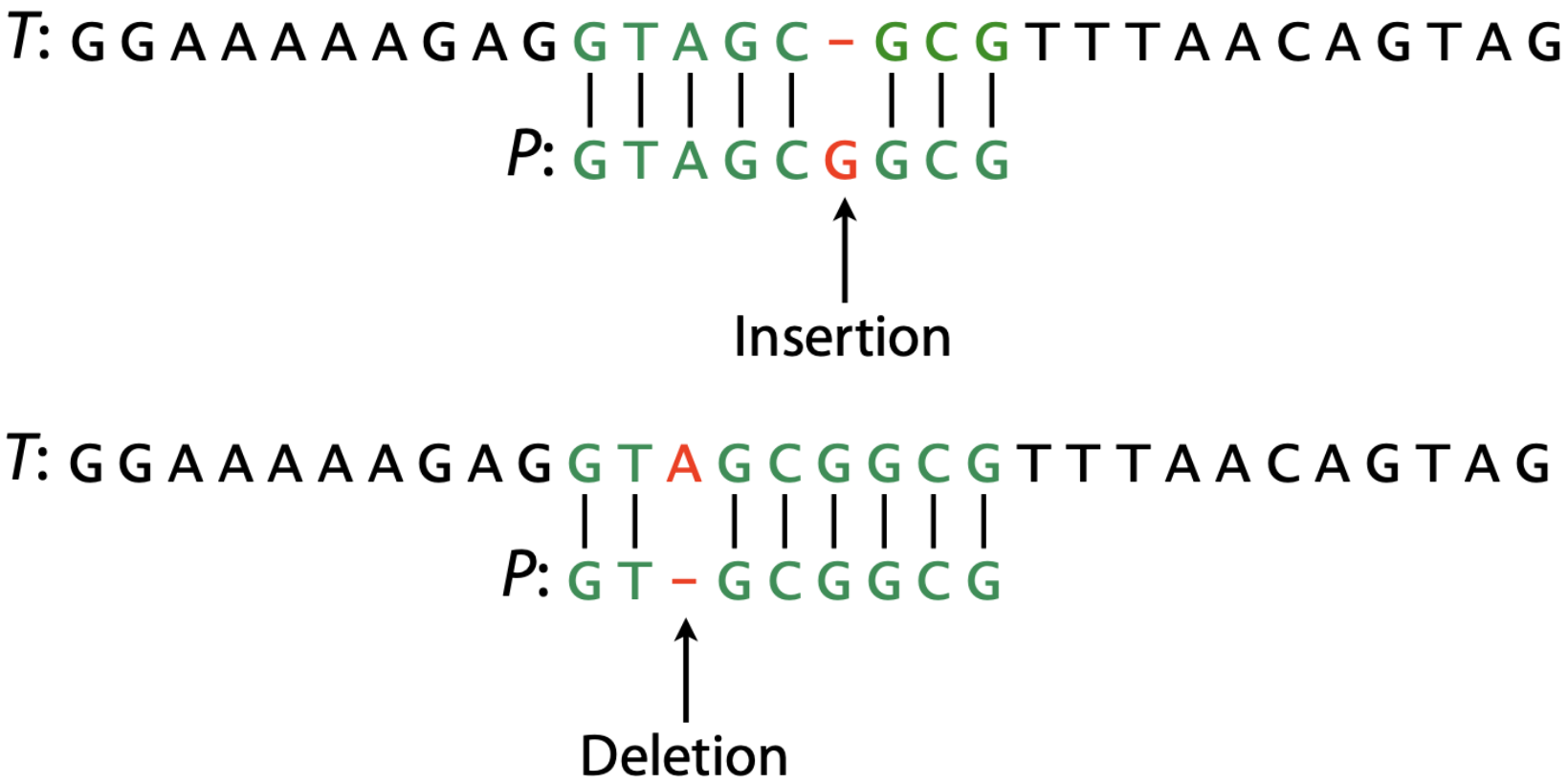
Naive Approximate Matching

```
def naiveHamming(p, t, maxDistance):
    occurrences = []
    for i in xrange(len(t) - len(p) + 1): # loop over alignments
        nmm = 0
        match = True
        for j in xrange(len(p)):          # loop over characters
            if t[i+j] != p[j]:            # compare characters
                nmm += 1                  # mismatch
                if nmm > maxDistance:
                    break                 # exceeded max hamming dist
        if nmm <= maxDistance:
          occurrences.append(i)           # approximate match
    return occurrences
```

- Not all types of errors are covered
- Need faster algorithms.

BRAC UNIVERSITY
Inspiring Excellence

Other Variants/Errors



Edit Distance aka Levenshtein Distance

For *X* & *Y*, *edit distance* = minimum # edits (substitutions, insertions, deletions) needed to turn one into the other

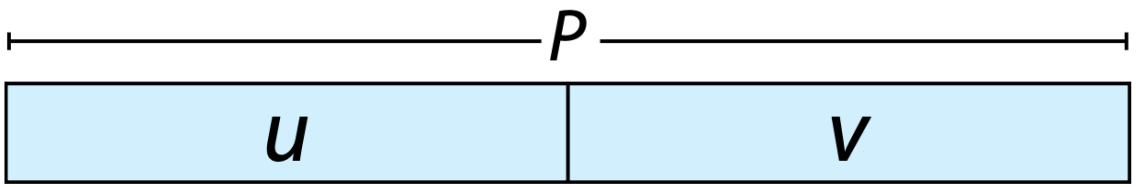
```
X: T G G C C G C G C A A A A A C A G C
   | |   | | | | | | | | | |   | | | |
Y: T G A C C G C G C A A A A - C A G C
```

Edit distance = 2

```
X: G C G T A T G C G G C T A - A C G C
   | |   | | | | | | | | | |   | | |
Y: G C - T A T G C G G C T A T A C G C
```

Edit distance = 2

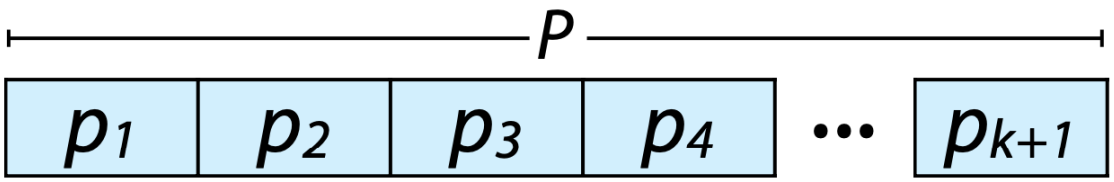
Approximate Matching - Pigeonhole Principle



If P occurrs in T with 1 edit, then u or v appears with no edits

BRAC UNIVERSITY
Inspiring Excellence

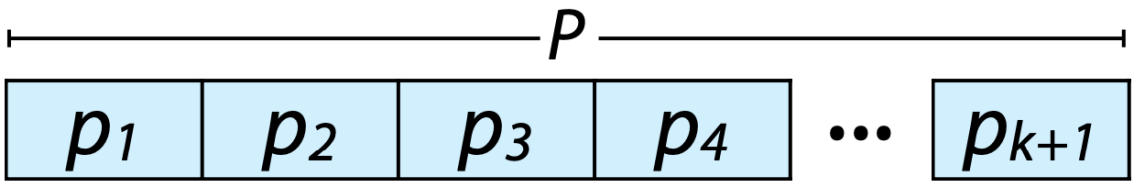
Approximate Matching - Pigeonhole Principle



If P occurs in T with up to k edits, then at least one of $p_1, p_2, ..., p_{k+1}$ must appear with 0 edits

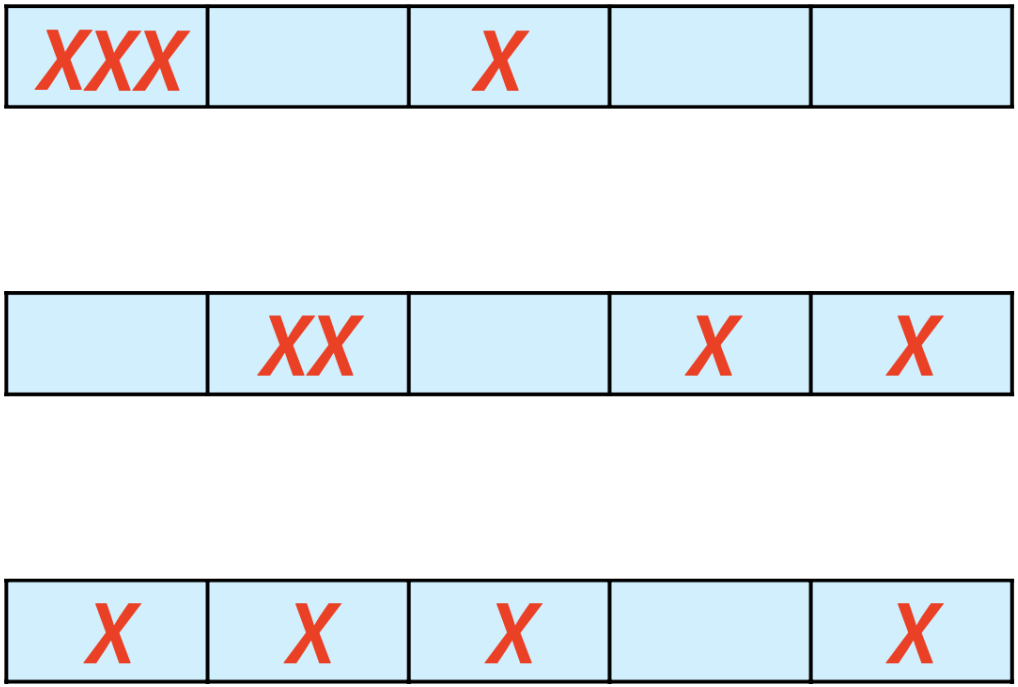
BRAC UNIVERSITY
Inspiring Excellence

Approximate Matching - Pigeonhole Principle



If P occurrs in T with up to k edits, at least one of $p_1, p_2, ..., p_{k+1}$ must appear with 0 edits

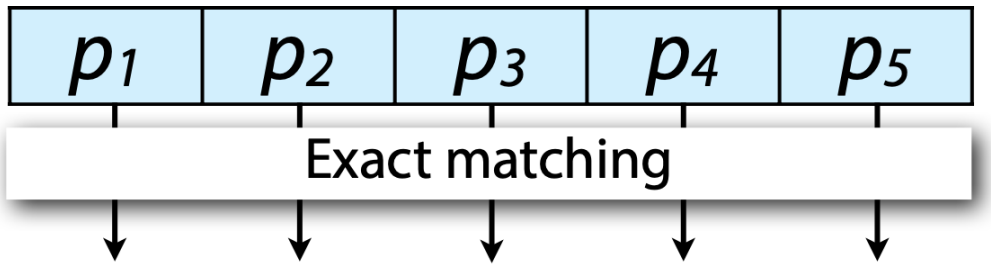
Approximate Matching - Pigeonhole Principle



Approximate Matching - Pigeonhole Principle

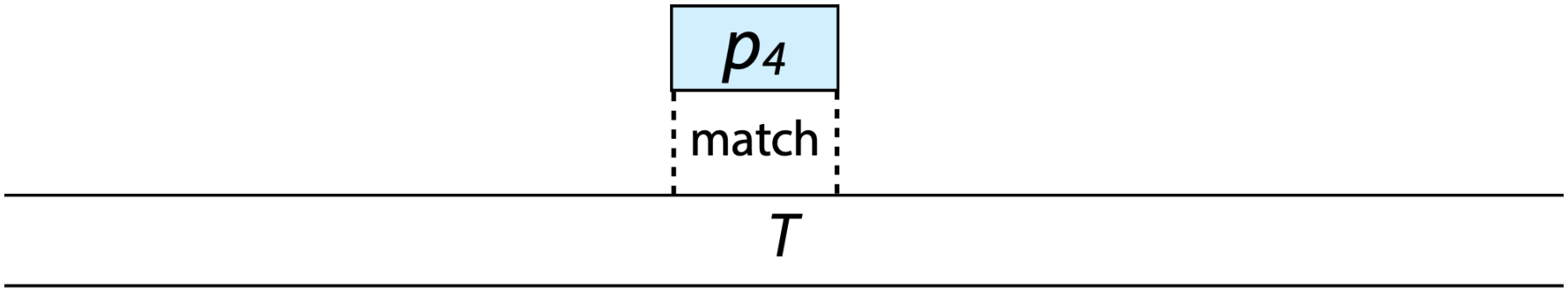
BRAC UNIVERSITY
Inspiring Excellence

Approximate Matching - Pigeonhole Principle

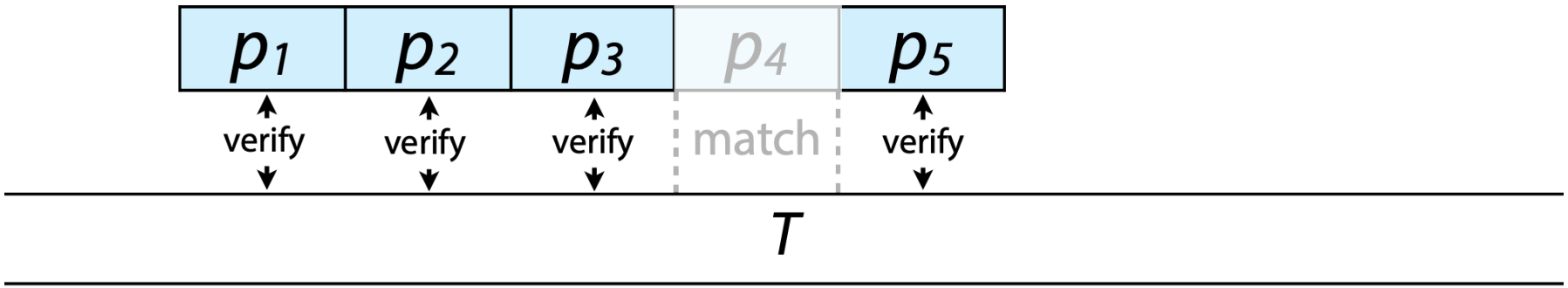


T

Approximate Matching - Pigeonhole Principle



Approximate Matching - Pigeonhole Principle



Approximate Matching - Pigeonhole Principle

	Boyer-Moore, exact			Boyer-Moore, ≤1 mismatch with pigeonhole			Boyer-Moore, ≤2 mismatches with pigeonhole		
	# character comparisons	wall clock time	# matches	# character comparisons	wall clock time	# matches	# character comparisons	wall clock time	# matches
P: "tomorrow" **T**: Shakespeare's complete works	786 K	1.91s	17	3.05 M	7.73 s	24	6.98 M	16.83 s	382
P: 50 nt string from Alu repeat* **T**: Human reference (hg19) chromosome 1	32.5 M	67.21 s	336	107 M	209 s	1,045	171 M	328 s	2,798

BRAC UNIVERSITY
Inspiring Excellence