

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,
Gusfield et al, Chapter 2 - Section 2.2)



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)):
            if t[i+j] != p[j]:
                match = False
                break
        if match:
            occurrences.append(i)
    return occurrences
```

all chars matched; record

mismatch; reject alignment

compare characters

Loop over characters

Loop over alignments



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



Inspiring Excellence

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

P: word

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



Inspiring Excellence

Boyer-Moore Algorithm - Bad Character Rule

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

Step 1: $T: GCTT\textcolor{red}{C}TGCTACCTTTGCGCGCGCGGGAA$
 $P: \textcolor{red}{C}\textcolor{blue}{C}TT\textcolor{red}{T}\textcolor{green}{T}GC$

Step 2: $T: GCTTCTGCT\textcolor{red}{A}\textcolor{green}{C}CTTTGCGCGCGCGGGAA$
 $P: \textcolor{red}{C}\textcolor{blue}{C}TTTT\textcolor{red}{G}\textcolor{green}{C}$

Step 3: $T: GCTTCTGCTA\textcolor{green}{C}CTTTGCGCGCGCGGGAA$
 $P: \textcolor{green}{C}CTTTG\textcolor{blue}{C}$



Inspiring Excellence

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

Step 1: $T: \text{CGTGCCTAC}TTACTTACTTAC\text{TACGCGAA}$
 $P: \text{CTTAC}T\text{AC}$

Step 2: $T: \text{CGTGCCTAC}T\text{ACTTAC}TTACTTACTTAC\text{TACGCGAA}$
 $P: \text{CTTAC}TTAC$

Step 3: $T: \text{CGTGCCTACTTAC}TTAC\text{TACGCGAA}$
 $P: \text{CTTAC}TTAC$



Inspiring Excellence

Boyer-Moore algorithm - Put Together

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

Step 1: $T: GTTATAGCTGATCGCGCGTAGCGCGAA$
 $P: GTAGCGGGC$ 

bc: 6, gs: 0 bad character

Step 2: $T: GTTATAGCTGATCGCGCGTAGCGCGAA$
 $P: GTAGCGGGC$ 

bc: 0, gs: 2 good suffix

Step 3: $T: GTTATAGCTGATCGCGCGTAGCGCGAA$
 $P: GTAGCGGGC$ 

bc: 2, gs: 7 good suffix

Step 4: $T: GTTATAGCTGATCGCGCGTAGCGCGAA$
 $P: GTAGCGGGC$

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



Boyer-Moore algorithm - Put Together Example

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

11 characters of T we ignored



Step 1: $T: \text{GTTATAGCTGATCGCGCGTAGCGCGAA}$
 $P: \text{GTAGCGGC}\text{G}$

Step 2: $T: \text{GTTATAGCTGATCGCGCGCGTAGCGCGAA}$
 $P: \text{GTAGC}\text{GGCG}$

Step 3: $T: \text{GTTATAGCTGATCGCGCGCGTAGCGCGAA}$
 $P: \text{GTAGCGGC}\text{G}$

Step 4: $T: \text{GTTATAGCTGATCGCGCGTAGCGCGAA}$
 $P: \text{GTAGCGGC}\text{G}$



Skipped 15 alignments



ing Excellence

Indexing the Genome

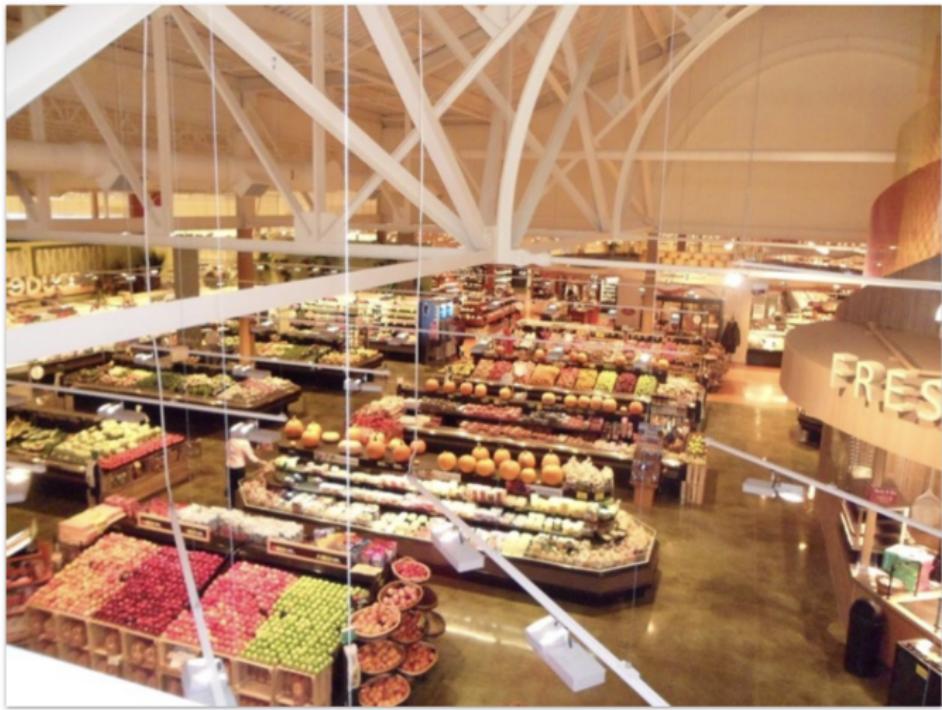
- nest site hunting, 482–87
- honeypot ants, *see* *Myrmecocystus*
- hormones, 106–9
 - see also* exocrine glands
- house (nest site) hunting, 482–92
- Hymenoptera (general), xvi
 - haplodiploid sex determination, 20–22
- Hypoponera* (ants), 194, 262, 324, 388
 - inclusive fitness, 20–23, 29–42
 - information 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), replete, 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



Inspiring Excellence

Indexing the Genome



Grocery store items grouped into aisles

Indexing the Genome

Index of T

$T:$ CGTGC GTGCTT



Inspiring Excellence

Indexing the Genome

Index of T

CGTGC: 0

$T:$ CGTGCGTGCTT



Inspiring Excellence

Indexing the Genome

Index of T

CGTGC: 0

GTGCG: 1

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



Inspiring Excellence

Indexing the Genome

Index of T

CGTGC: 0
GTGCG: 1
TGCCT: 2

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



Inspiring Excellence

Indexing the Genome

Index of T

CGTGC: 0

GCGTG: 3

GTGCC: 1

TGCCT: 2



$T: \text{C} \text{G} \text{T} \text{G} \text{C} \text{G} \text{T} \text{G} \text{C} \text{T} \text{T}$



Inspiring Excellence

Indexing the Genome

Index of T

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

$T: \text{C G T G C} \underline{\text{G T G C T T}}$



Inspiring Excellence

Indexing the Genome

Index of T

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

$T: C G T G C G \underline{T} G C T T$



Inspiring Excellence

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



Inspiring Excellence

Indexing the Genome

k-mer: substring
of length k

	<i>Index of T</i>
CGTGC	0, 4
GCGTG	3
GTGCC	1
GTGCT	5
TGCCT	2
TGCTT	6

5-mer index

T: CGTGC G T G C T T



Inspiring Excellence

Indexing the Genome

<i>Index of T</i>	
CGTGC:	0, 4
GCGTG:	3
GTGCC:	1
GTGCT:	5
TGCCT:	2
TGCTT:	6

$T: \text{C G T G C G T G C T T}$

$P: \underline{\text{G C G T G C}}$



Inspiring Excellence

Indexing the Genome

Index of T

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

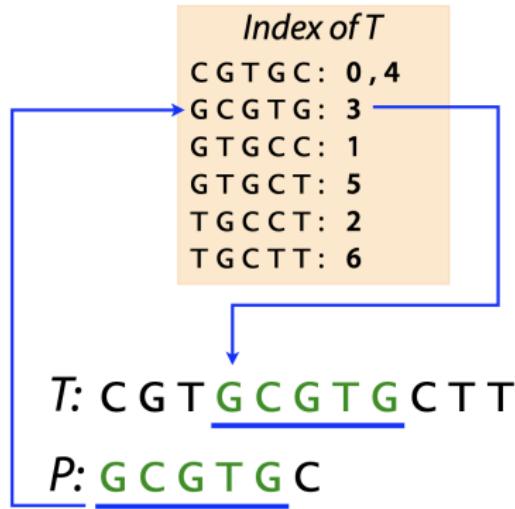
$T: \text{CGTGC GTGCTT}$

$P: \underline{\text{GCGTG}}$



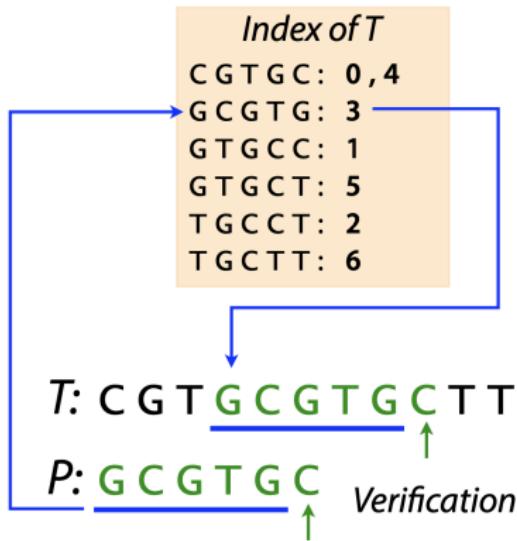
Inspiring Excellence

Indexing the Genome



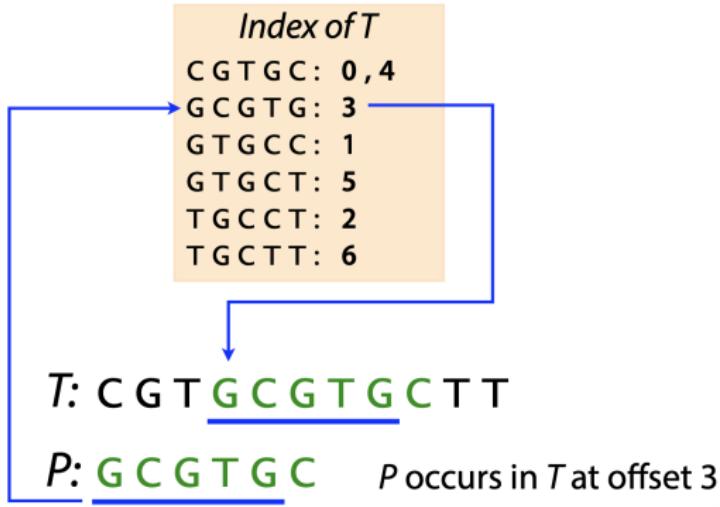
Inspiring Excellence

Indexing the Genome



Inspiring Excellence

Indexing the Genome



Inspiring Excellence

Indexing the Genome

<i>Index of T</i>	
CGTGC:	0, 4
GCGTG:	3
GTGCC:	1
GTGCT:	5
TGCCT:	2
TGCTT:	6

?

T: CGTGC GTGCTT

P: GCGTGC



Inspiring Excellence

Indexing the Genome

Index of T

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

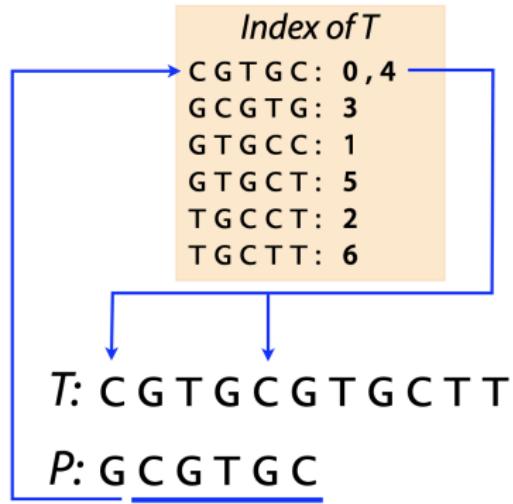
T : CGTGC GTGCTT

P : GCGTGC



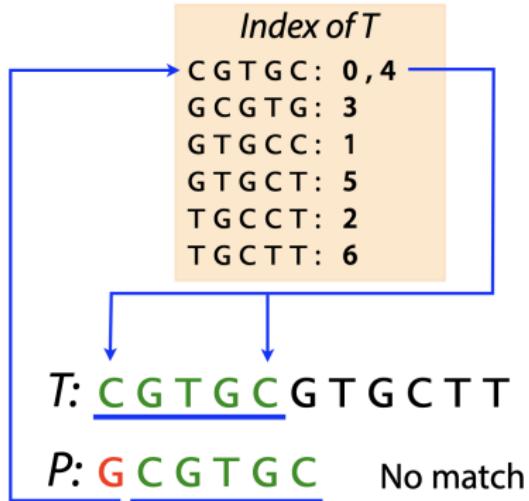
Inspiring Excellence

Indexing the Genome



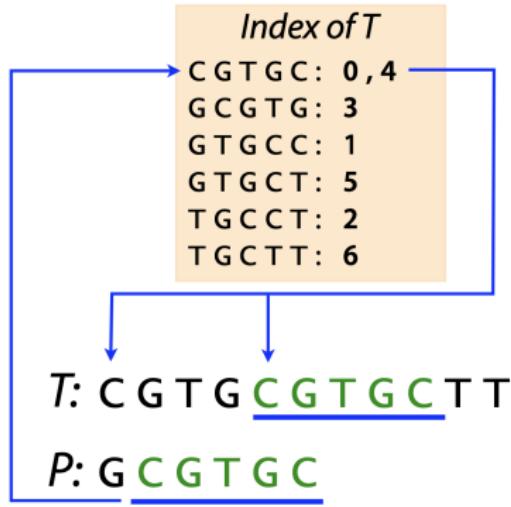
Inspiring Excellence

Indexing the Genome



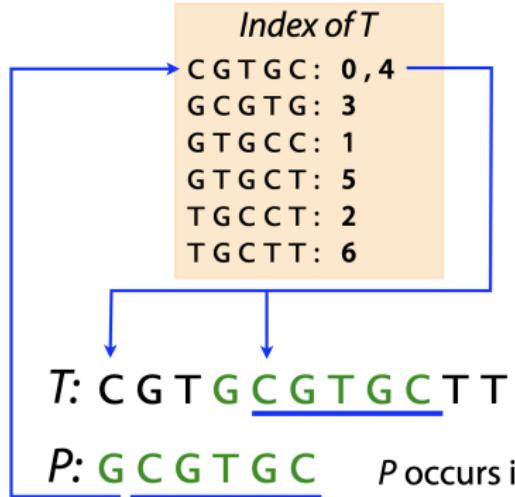
Inspiring Excellence

Indexing the Genome



Inspiring Excellence

Indexing the Genome



Inspiring Excellence

Indexing the Genome

Index of T

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

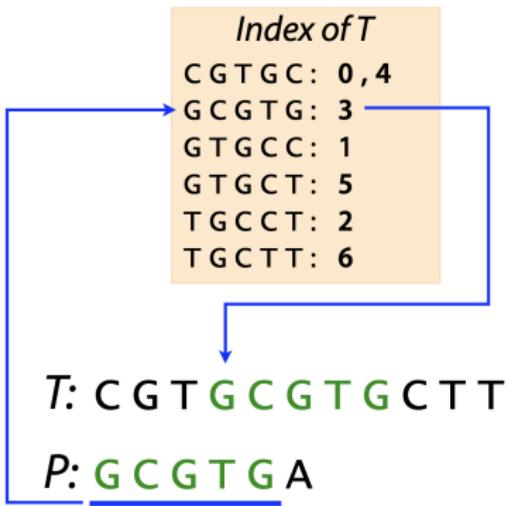
$T: C G T G C G T G C T T$

$P: \underline{G C G T G A}$



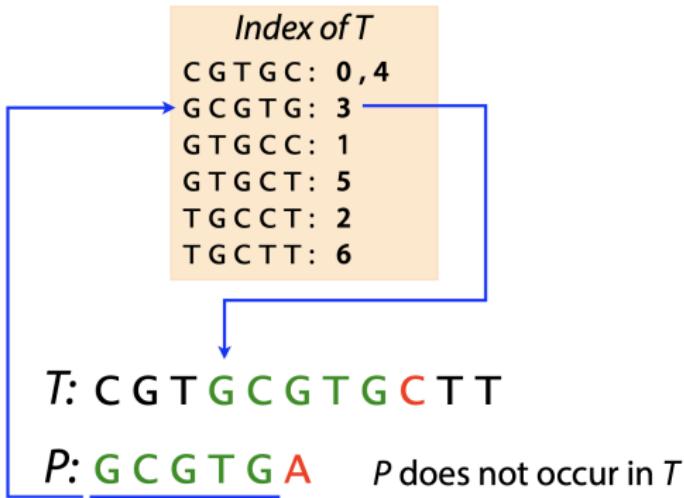
Inspiring Excellence

Indexing the Genome



Inspiring Excellence

Indexing the Genome



Inspiring Excellence

Indexing the Genome

Index of T

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

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

$P: \underline{G C G T A C}$



Inspiring Excellence

Indexing the Genome

Index of T

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

→ X

$T: \text{CGTGC} \text{GTGCTT}$

$P: \underline{\text{GCGTA}}\text{C}$ P does not occur in T



Inspiring Excellence

Indexing the Genome

Index of T

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

1 index hit

$T: \text{CGTGC GTGCTT}$

$P: \underline{\text{GCGTG C}}$



Inspiring Excellence

Indexing the Genome

Index of T

CGTGC: 0, 4

GCGTG: 3

GTGCC: 1

GTGCT: 5

TGCCT: 2

TGCTT: 6

2 index hits

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

$P: G \underline{C} G T G C$



Inspiring Excellence

Searching in a Multimap

Index of T

C GT GC: 0, 4
G CGT G: 3
GT GCC: 1
GT GCT: 5
T GC CT: 2
TG CTT: 6

Multimap

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



Inspiring Excellence

Searching in a Multimap

size homeostasis, 482–487

homeostatic or homeostatic
regulation, 482–487

or also exocrine glands

hormone (size homeostasis), 482–492

Hormone system

- supplements air deterioration, 20–22
- Hypogastrine* (ants), 194, 262, 326, 389

inclusive fitness, 20–23, 39–42

information measurement, 251–252

interactions, 388–389

- or also ergophagy, ergonal quinone, gastrulation*

infraspecific competition, 266, 280, 288, 321

insects, or termites

juvenile hormone, case, 106–9, 372

kin recognition, 293–298

kin selection, 18–19, 23–24, 28–42, 299

laboratory experiments, 298–300

language, 20–21

large-scale selection, 298

mass communication, 62–63, 214–218

maternal care, 159

maternal influence, 119

Magnification (ants), 457

Majusculina (ants), or *Polyergus* (ants), 129

Melipona (ants), or species, 257

memory, 117–179, 213

Möser (Brauerstein ant), 212, 232

mist, 317

Mutualism, 127, 121, 214, 216–17,

292

motor displays, 235–47

mouse hunting, 202

multilevel selection, 7, 7–13, 24–29

mutation, visual, 366–73

mutualism, or symbiosis, 202

Mutillia (ants), or species, 318

Mutillia (ants), 320



Multimap

T: CGTGC GTGCTT



Inspiring Excellence

Searching in a Multimap



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



Inspiring Excellence

Searching in a Multimap

G T G	0
T G C	1

$T: \underline{G T G C} G T G T G G G G G$



Inspiring Excellence

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 G



Inspiring Excellence

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 G



Searching in a Multimap

Alphabetical by k-mer



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

$T: GTGCGTGTGGGGG$



Inspiring Excellence

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

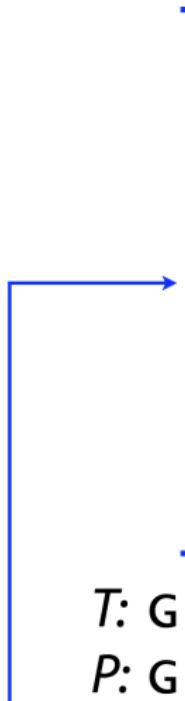
inclusive fitness, 20–23, 29–42
information measurement, 251–52
intercastes, 388–89
 see also ergatogynes; ergatoid queens;
 gamergates
Iridomyrmex (ants), 266, 280, 288, 321
Isoptera, *see termites*

juvenile hormone, caste, 106–9, 372

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

Macrotelmes (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

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

$T: GTGC$ G

$P: GCG$ TGG

Searching in a Multimap

T G G > G T G

→

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 G G G G
P: G C G T G G*

Searching in a Multimap

After 1st bisection

T G G > T G C

↓

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

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

Searching in a Multimap

After 2nd bisection

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

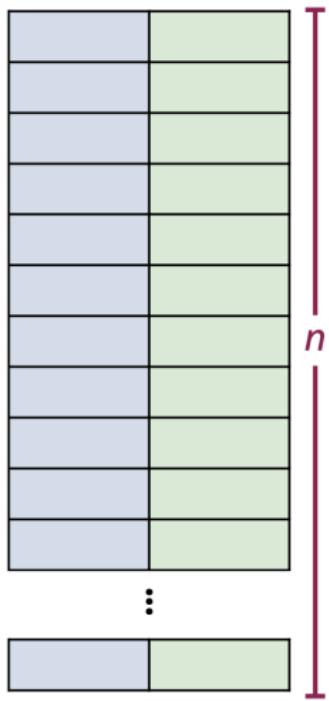
T G G = T G G

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

P: G C G T G G



Searching in a Multimap

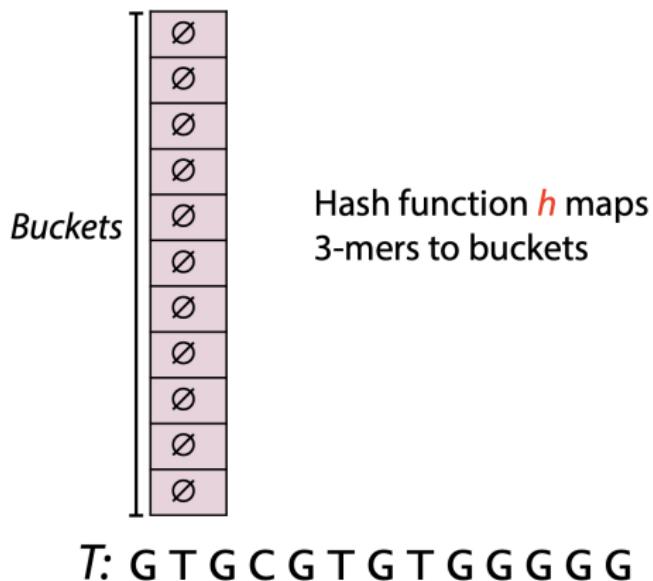


$\sim \log_2(n)$ bisections
per query



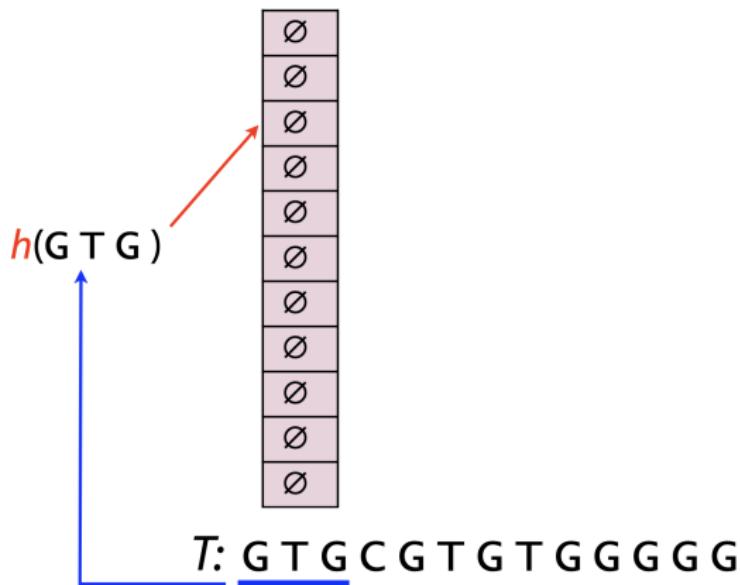
Inspiring Excellence

Searching in a Hash Table



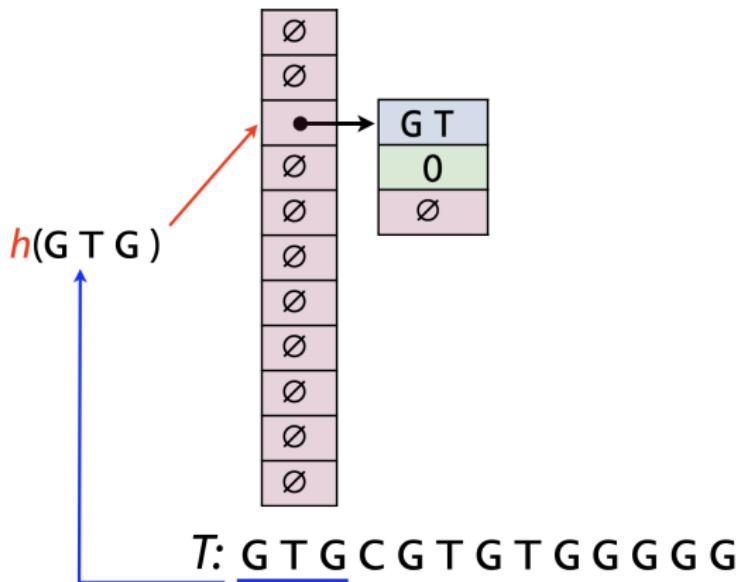
Inspiring Excellence

Searching in a Hash Table



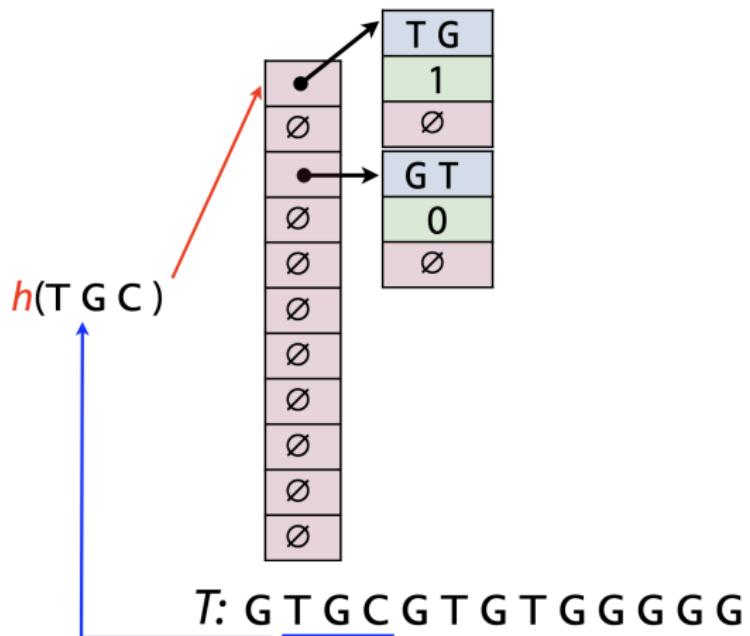
Inspiring Excellence

Searching in a Hash Table



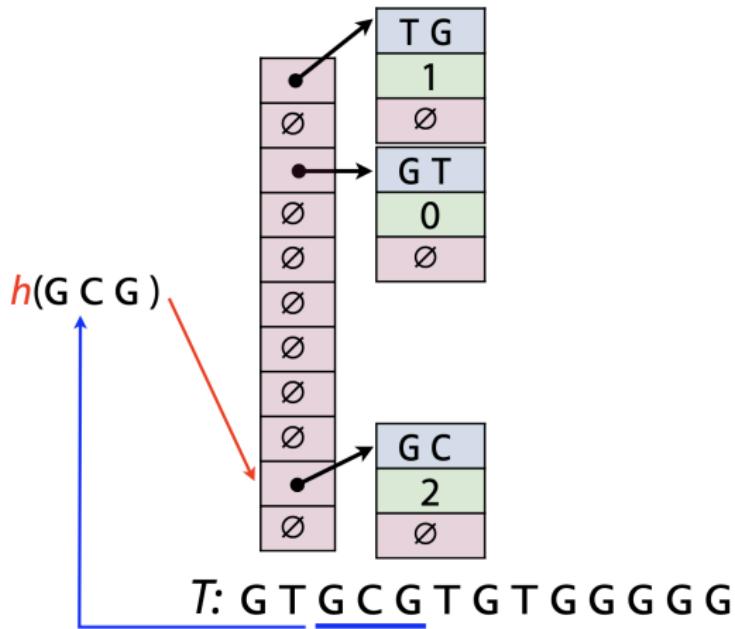
Inspiring Excellence

Searching in a Hash Table



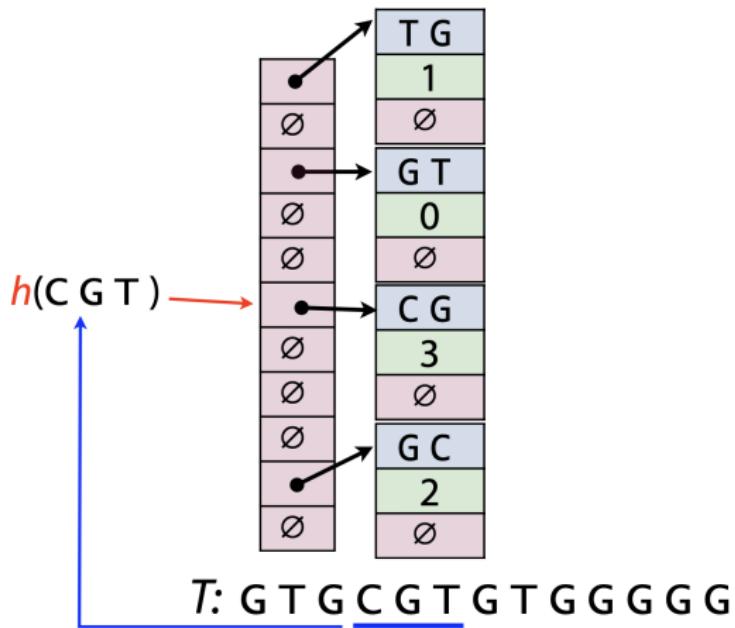
Inspiring Excellence

Searching in a Hash Table



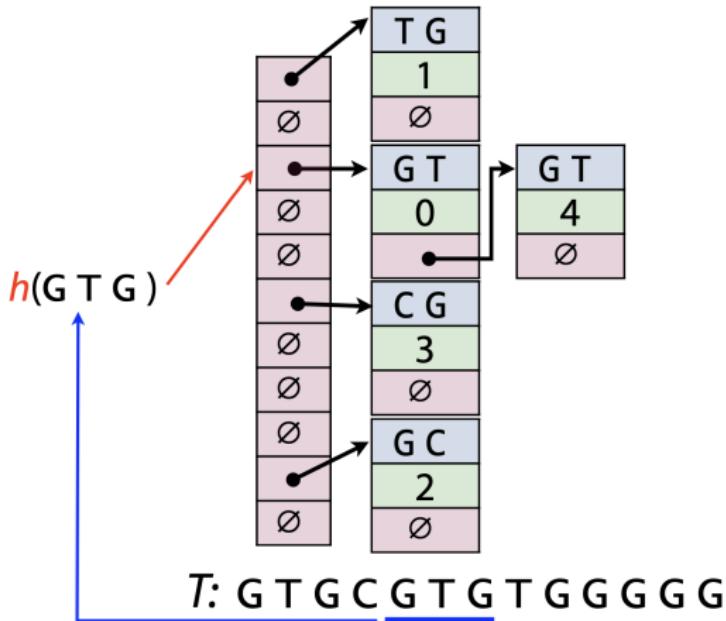
Inspiring Excellence

Searching in a Hash Table



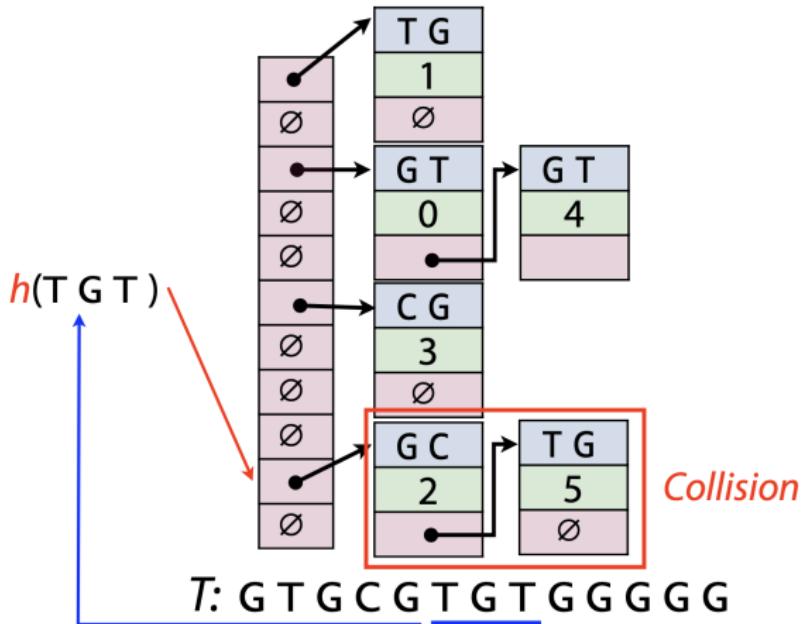
Inspiring Excellence

Searching in a Hash Table



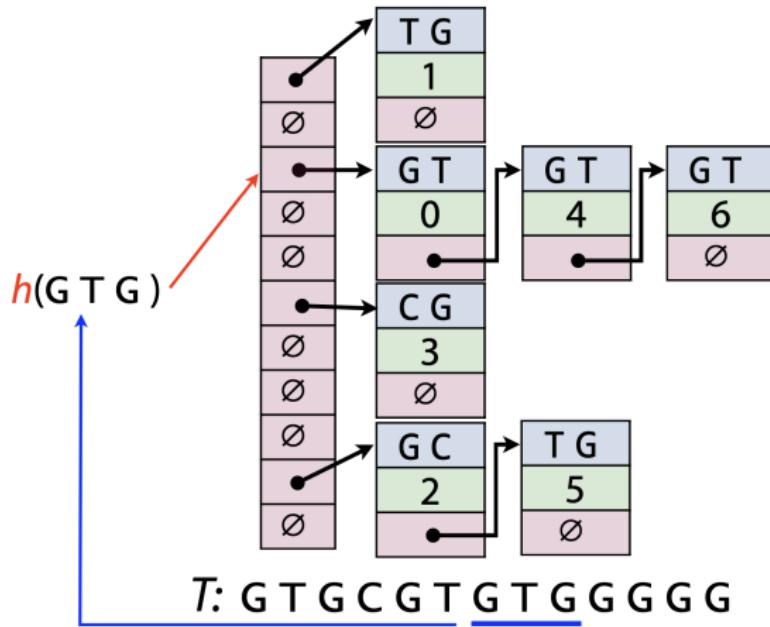
Inspiring Excellence

Searching in a Hash Table



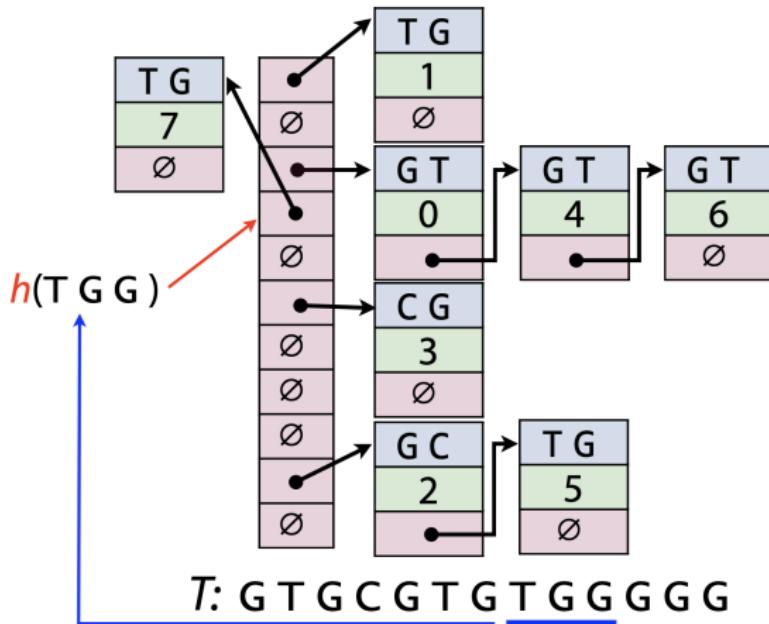
Inspiring Excellence

Searching in a Hash Table



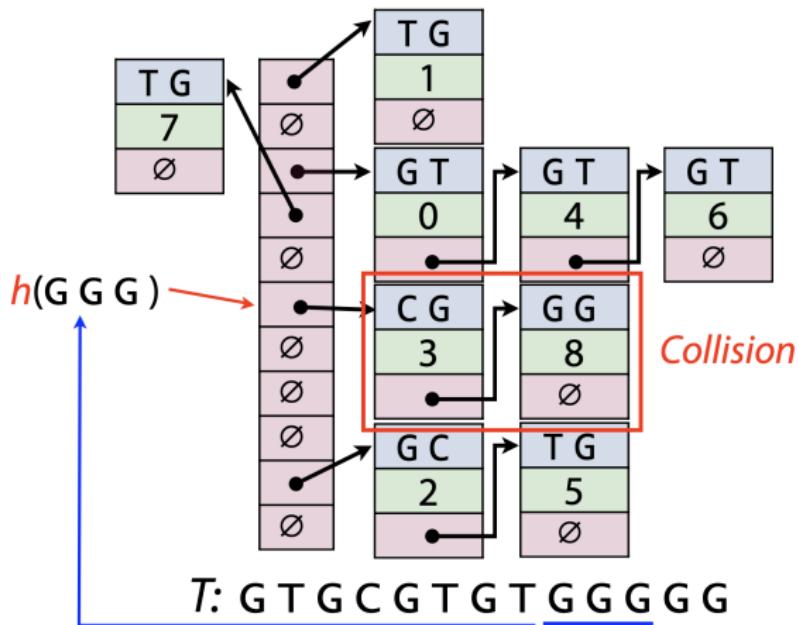
Inspiring Excellence

Searching in a Hash Table



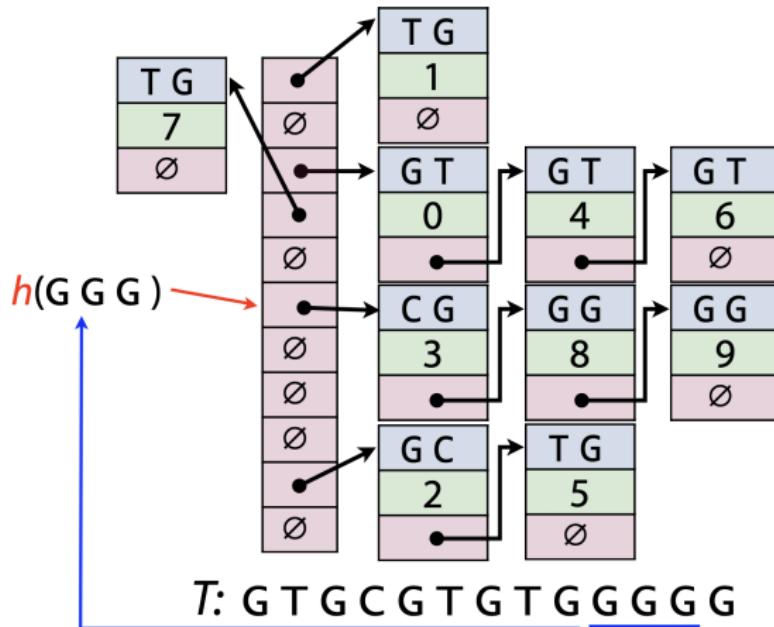
Inspiring Excellence

Searching in a Hash Table



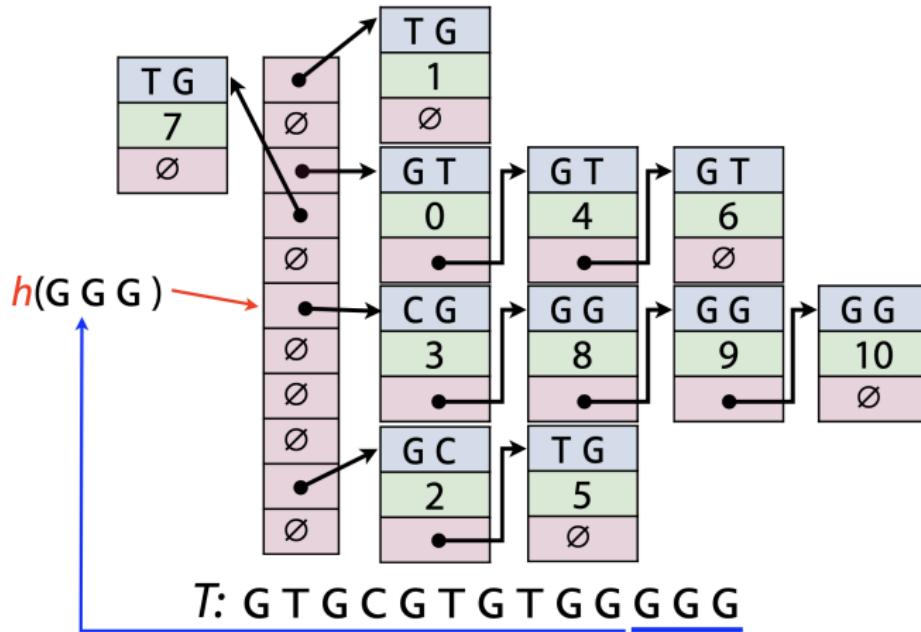
Inspiring Excellence

Searching in a Hash Table



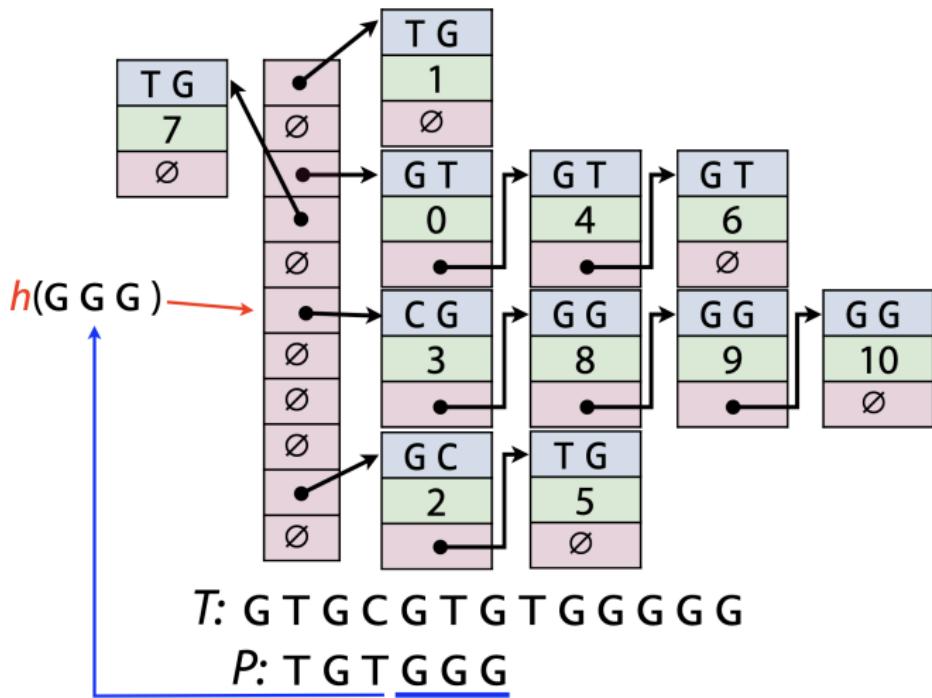
Inspiring Excellence

Searching in a Hash Table



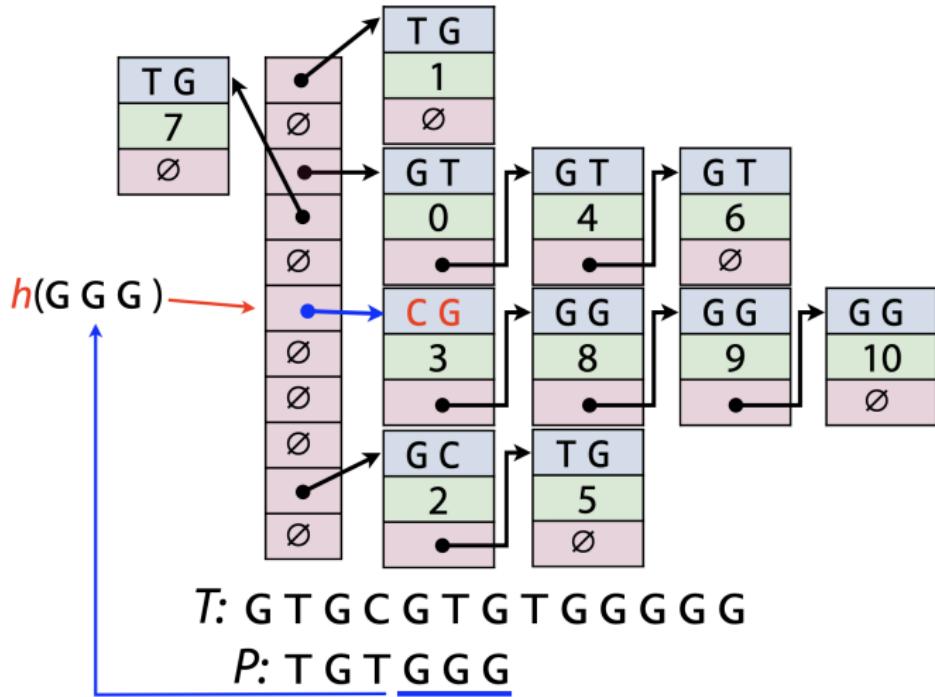
Inspiring Excellence

Searching in a Hash Table



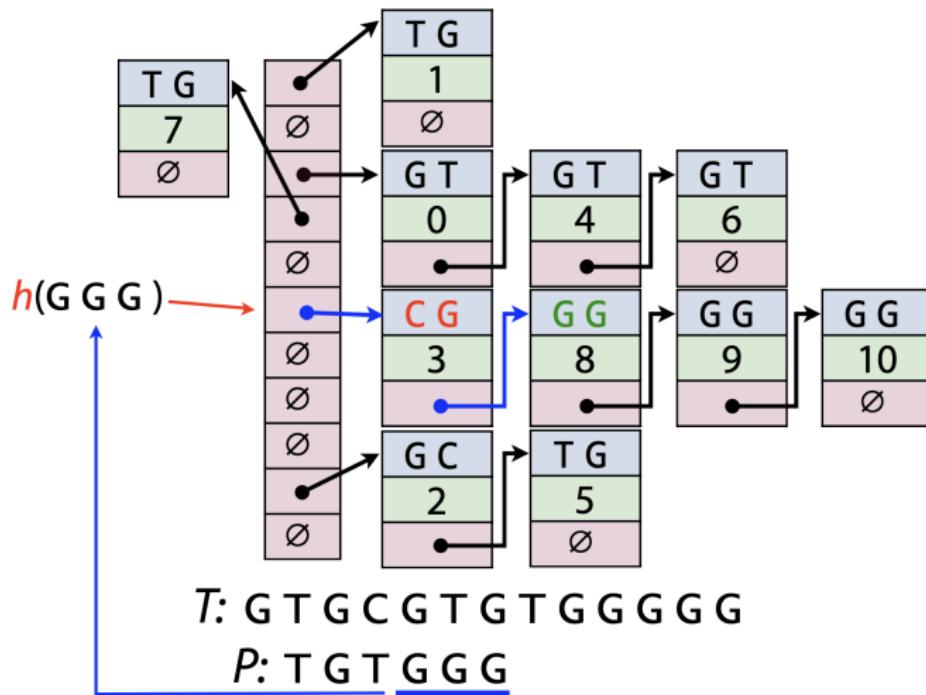
Inspiring Excellence

Searching in a Hash Table



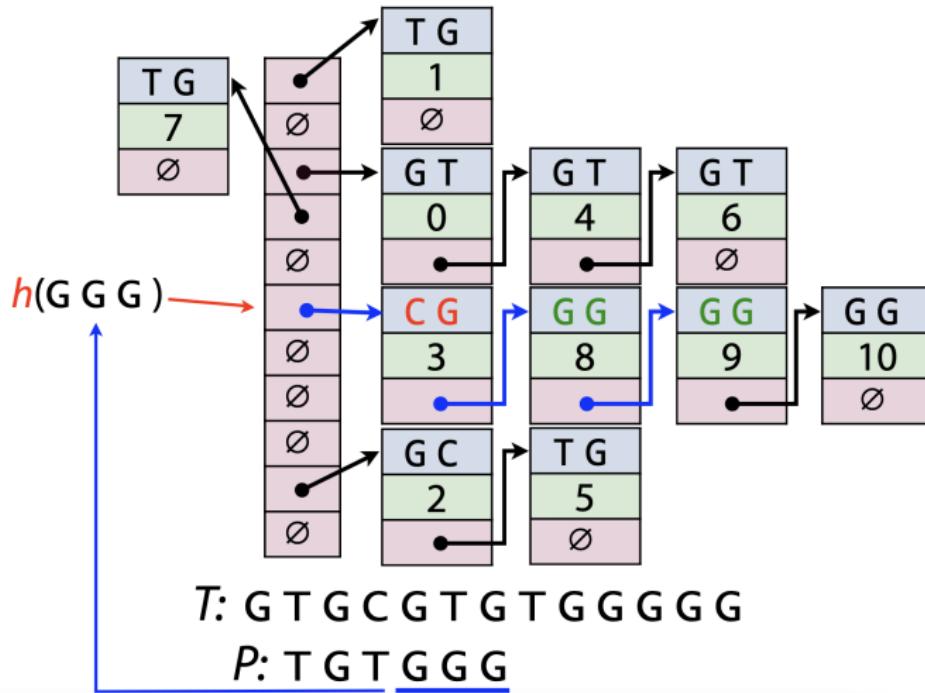
Inspiring Excellence

Searching in a Hash Table



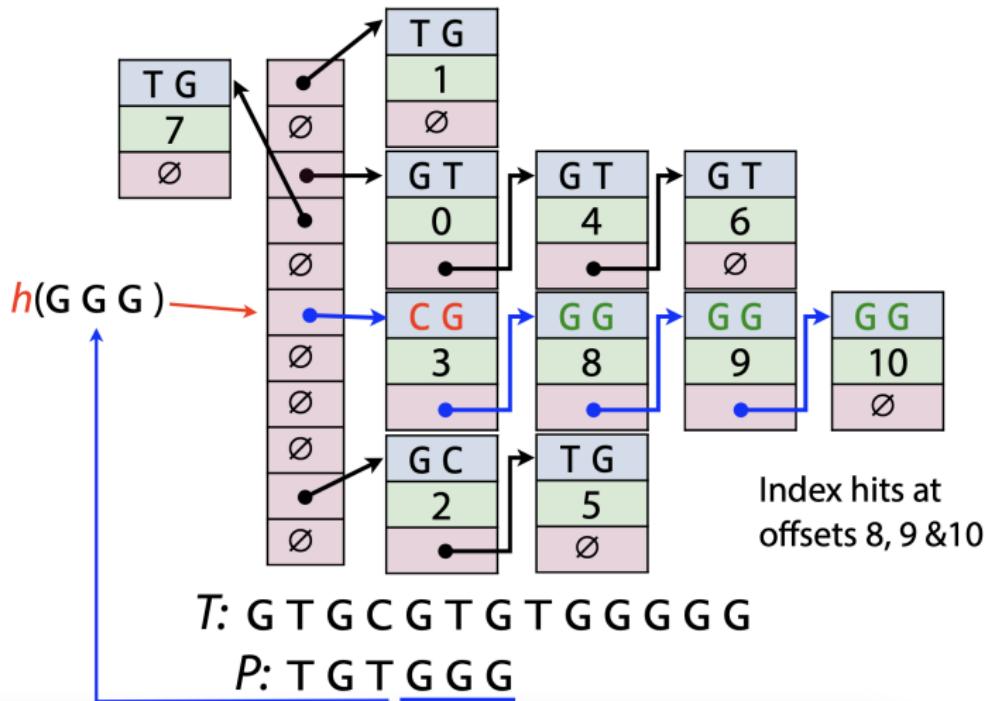
Inspiring Excellence

Searching in a Hash Table



Inspiring Excellence

Searching in a Hash Table



Inspiring Excellence

Other Variants of Index Search

Index of T

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

T: CGTGC GTGCTT



Inspiring Excellence

Other Variants of Index Search

Index of T

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

$T: \text{CGTGC GTGCTT}$



Inspiring Excellence

Other Variants of Index Search

Index of T

CGTGC: 0,4
TGCCT: 2
TGCTT: 6

$T:$ CGTGC GTGCTT



Inspiring Excellence

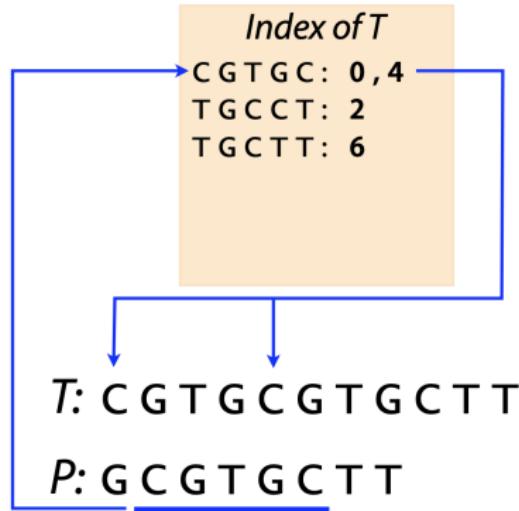
Other Variants of Index Search

<i>Index of T</i>	
CGTGC:	0,4
TGCCT:	2
TGCTT:	6

$T: \text{C G T G C G T G C T T}$

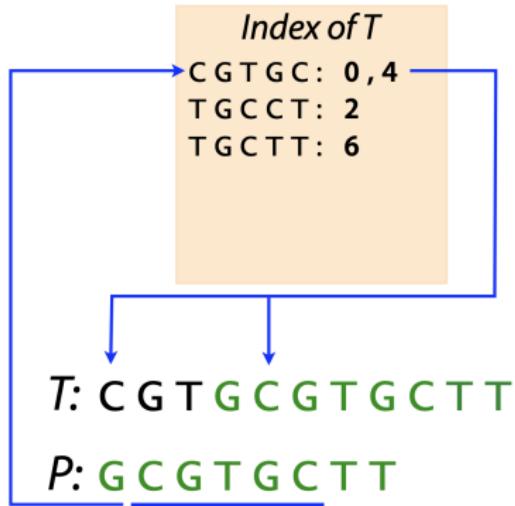
$P: \underline{\text{G C G T G C T T}}$

Other Variants of Index Search



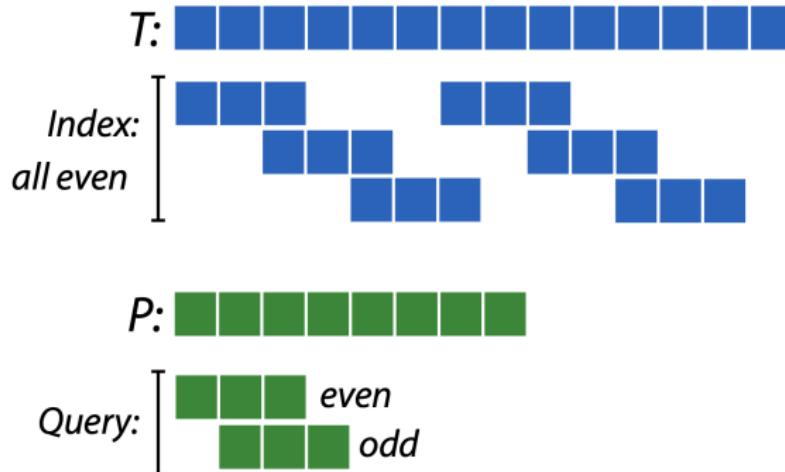
Inspiring Excellence

Other Variants of Index Search



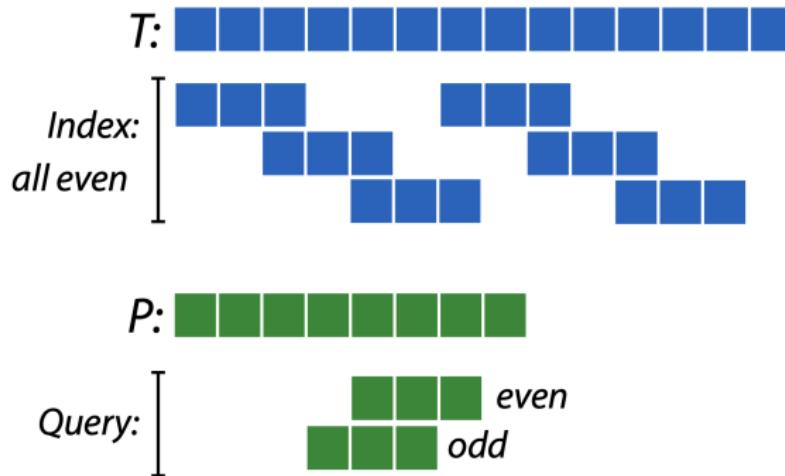
Inspiring Excellence

Other Variants of Index Search



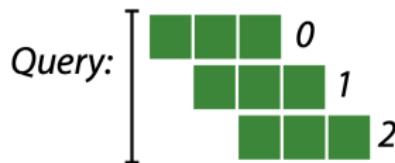
Inspiring Excellence

Other Variants of Index Search



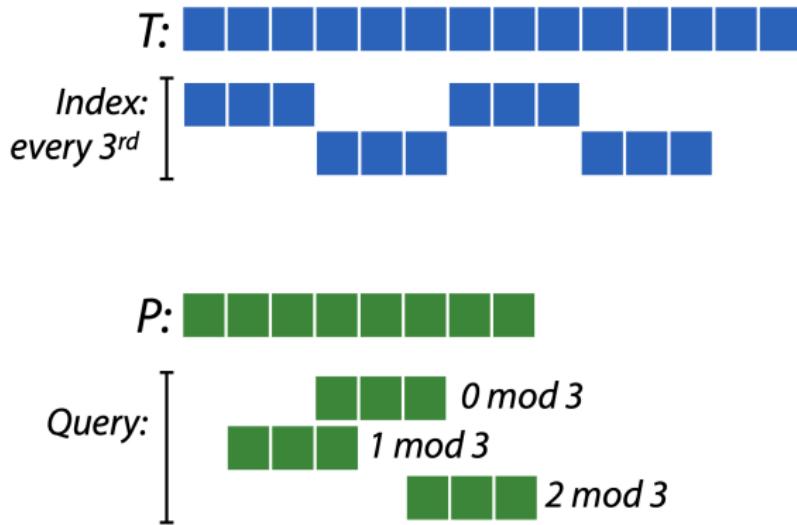
Inspiring Excellence

Other Variants of Index Search



Inspiring Excellence

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



Inspiring Excellence

Other Variants of Index Search

Index of T

CGGGT: 0

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



Inspiring Excellence

Other Variants of Index Search

Index of T

CGGGT: 0

GTCTG: 1

$T: \underline{C} \underline{G} \underline{T} \underline{G} \underline{C} \underline{G} \underline{T} \underline{G} C T T$



Inspiring Excellence

Other Variants of Index Search

Index of T

CGGGT: 0
GTCTG: 1
TGGGC: 2

$T: C \underline{G} T \underline{G} C G \underline{T} G \underline{C} T T$



Inspiring Excellence

Other Variants of Index Search

Index of T

C G G G T:	0
C G G T T:	4
G C T C T:	3
G T C T G:	1
T G G G C:	2

$T: \text{C G T G C G T G C T T}$



Inspiring Excellence

Other Variants of Index Search

Index of T

C GG GT: 0
C GG TT: 4
G CT CT: 3
G T C TG: 1
T G GG C: 2

T: CGTGC GTGCTT

P: G C G T A C T

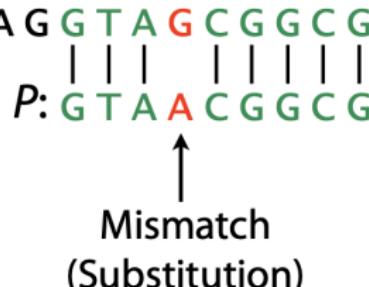


Inspiring Excellence

Approximate Matching - Hamming Distance

Differences between read and reference occur because of:

- Sequencing error
- Natural variation

T: G G A A A A A G A G G T A G C G G G C G T T T A A C A G T A G
P: G T A A C G G G C G

Mismatch
(Substitution)

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

X: G A G G T A G C G G C G T T T A A C Hamming distance = 3
Y: G T G G T A A C G G G G T T T A A C

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)):
            if t[i+j] != p[j]:
                nmm += 1
                if nmm > maxDistance:
                    break
            if nmm <= maxDistance:
                occurrences.append(i)
    return occurrences
```

Loop over characters
compare characters
mismatch
exceeded max hamming dist
approximate match

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



Inspiring Excellence

Other Variants/Errors

T: GGAAAAAAGAGG **G T A G C - G C G** TTTAACAGTAG

P: **G T A G C G G C G**

↑
Insertion

T: GGAAAAAAGAGG **G T A G C G G C G** TTTAACAGTAG

P: **G T - G C G G C G**

↑
Deletion



Inspiring Excellence

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

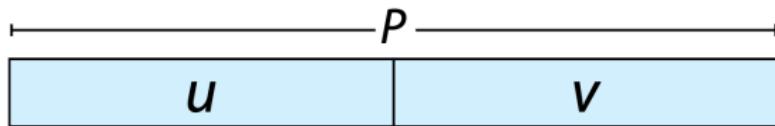
X : G C G T A T G C G G C T A - A C G C
| | | | | | | | | | | | | | | | | | |
 Y : G C - T A T G C G G C T A T A C G C

Edit distance = 2



Inspiring Excellence

Approximate Matching - Pigeonhole Principle

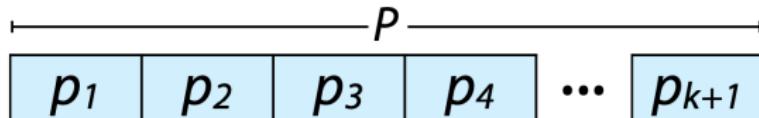


If P occurs in T with 1 edit, then u or v appears with no edits



Inspiring Excellence

Approximate Matching - Pigeonhole Principle

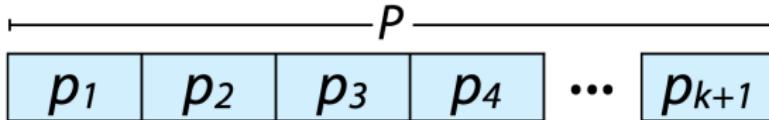


If P occurs in T with up to k edits, then at least one of p_1, p_2, \dots, p_{k+1} must appear with 0 edits



Inspiring Excellence

Approximate Matching - Pigeonhole Principle



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



Inspiring Excellence

Approximate Matching - Pigeonhole Principle



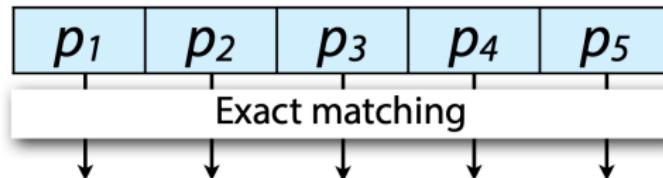
Inspiring Excellence

Approximate Matching - Pigeonhole Principle



Inspiring Excellence

Approximate Matching - Pigeonhole Principle

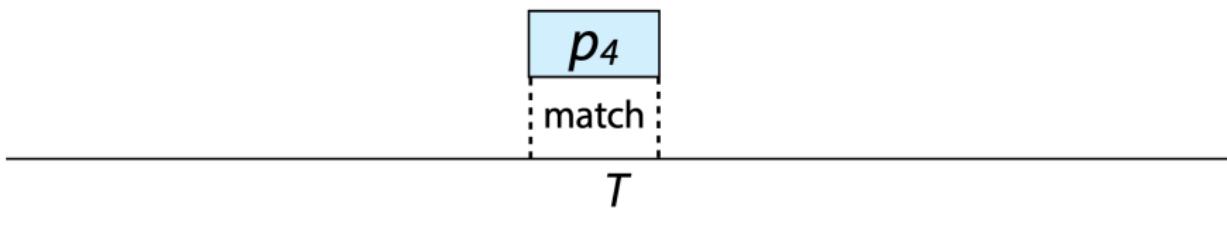


T



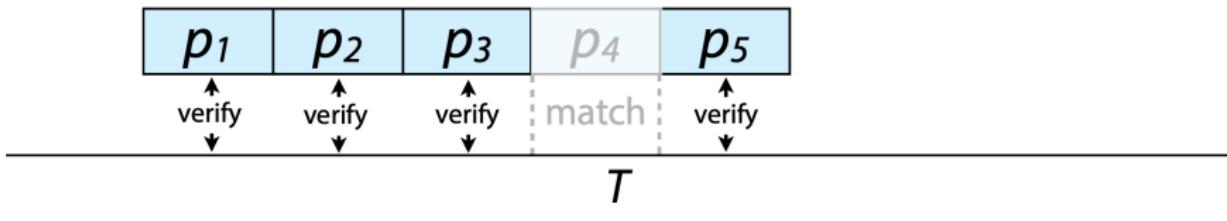
Inspiring Excellence

Approximate Matching - Pigeonhole Principle



Inspiring Excellence

Approximate Matching - Pigeonhole Principle



Inspiring Excellence

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



Inspiring Excellence