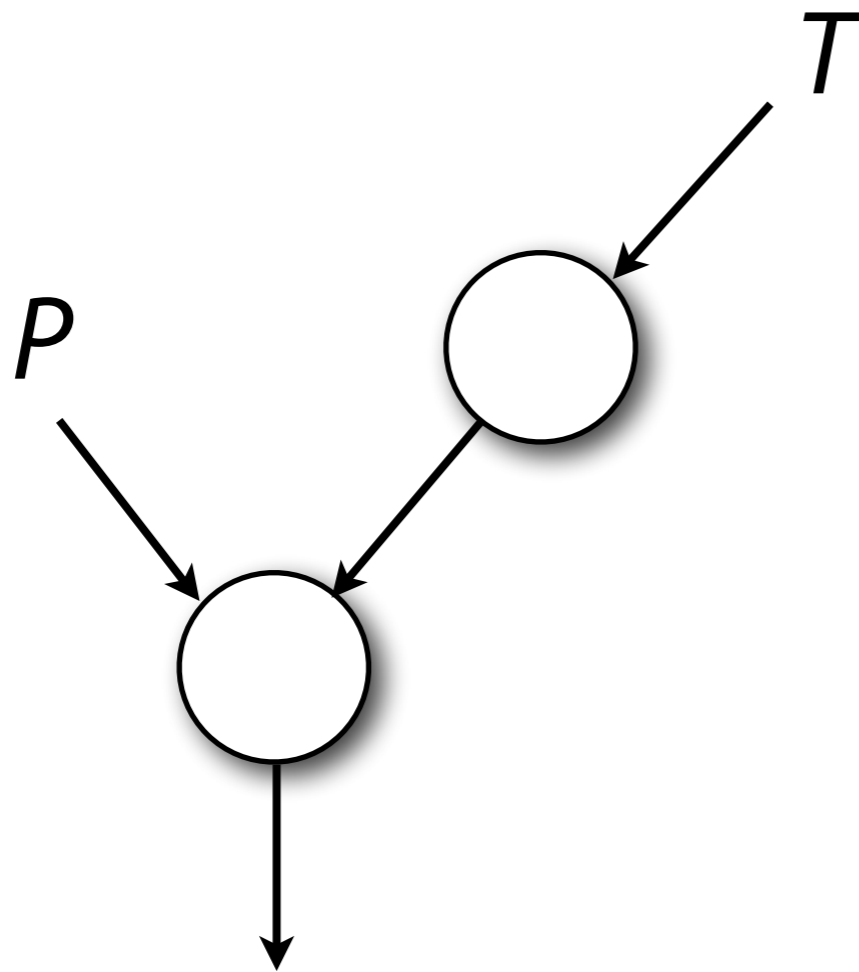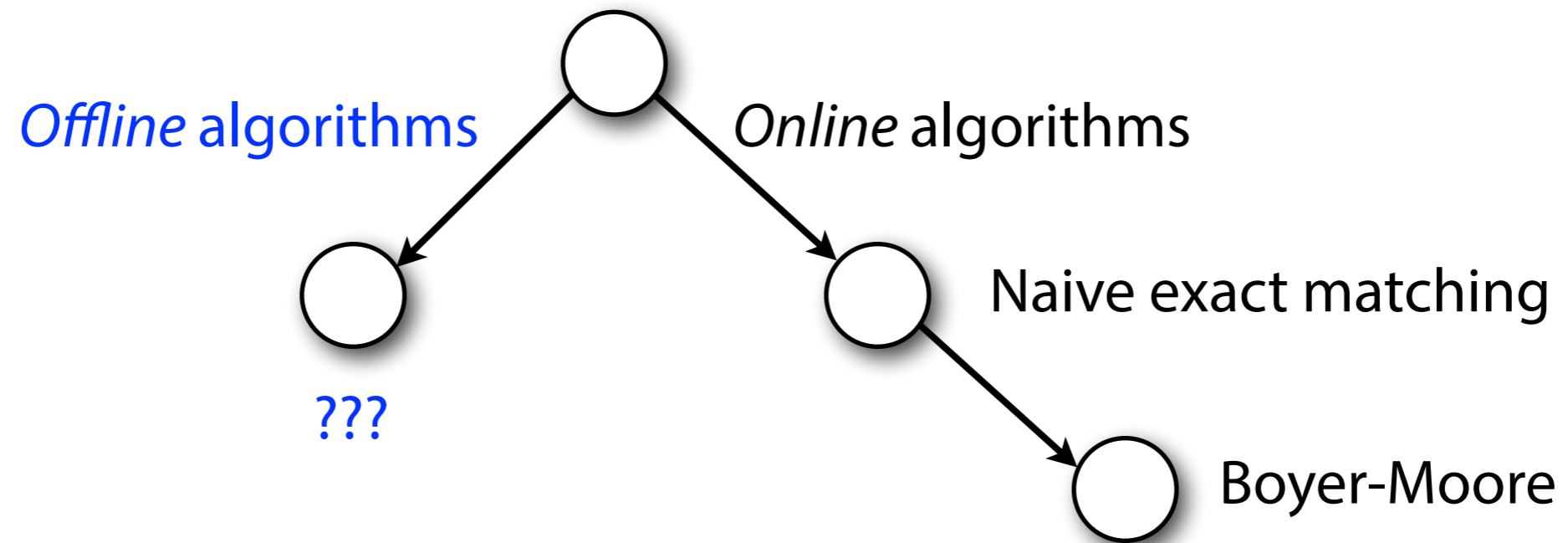# Preprocessing

Algorithm that preprocesses *T* is *offline*.
Otherwise, algorithm is *online*.

Online or offline?

- Naïve algorithm
- Boyer-Moore
- Web search engine
- Read alignment

Offline

Index

Key terms ordered alphabetically, with associated page #s

Index

Grocery store items grouped into aisles

Indexing DNA

Index of T

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

Indexing DNA

Index of T

C G T G C : **0**

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

Indexing DNA

Index of T

C G T G C : **0**
G T G C G : **1**

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

Indexing DNA

Index of T

C G T G C : **0**

G T G C G : **1**

T G C G T : **2**

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

Indexing DNA

Index of T

C G T G C : **0**

G C G T G : **3**

G T G C C : **1**

T G C C T : **2**

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

Indexing DNA

Index of T

CGTGC: **0**, **4**

GCGTG: **3**

GTGCC: **1**

TGCCT: **2**

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

Indexing DNA

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

Indexing DNA

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**

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

Indexing DNA

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

Querying the index

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**

?

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

P: G C G T G C

Querying the index

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 G C

Querying the index

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
```

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

P: G C G T G C

Querying the index

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**

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

P: G C G T G

Verification

Querying the index

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**

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

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

P occurs in *T* at offset 3

Querying the index

Index of T

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

?

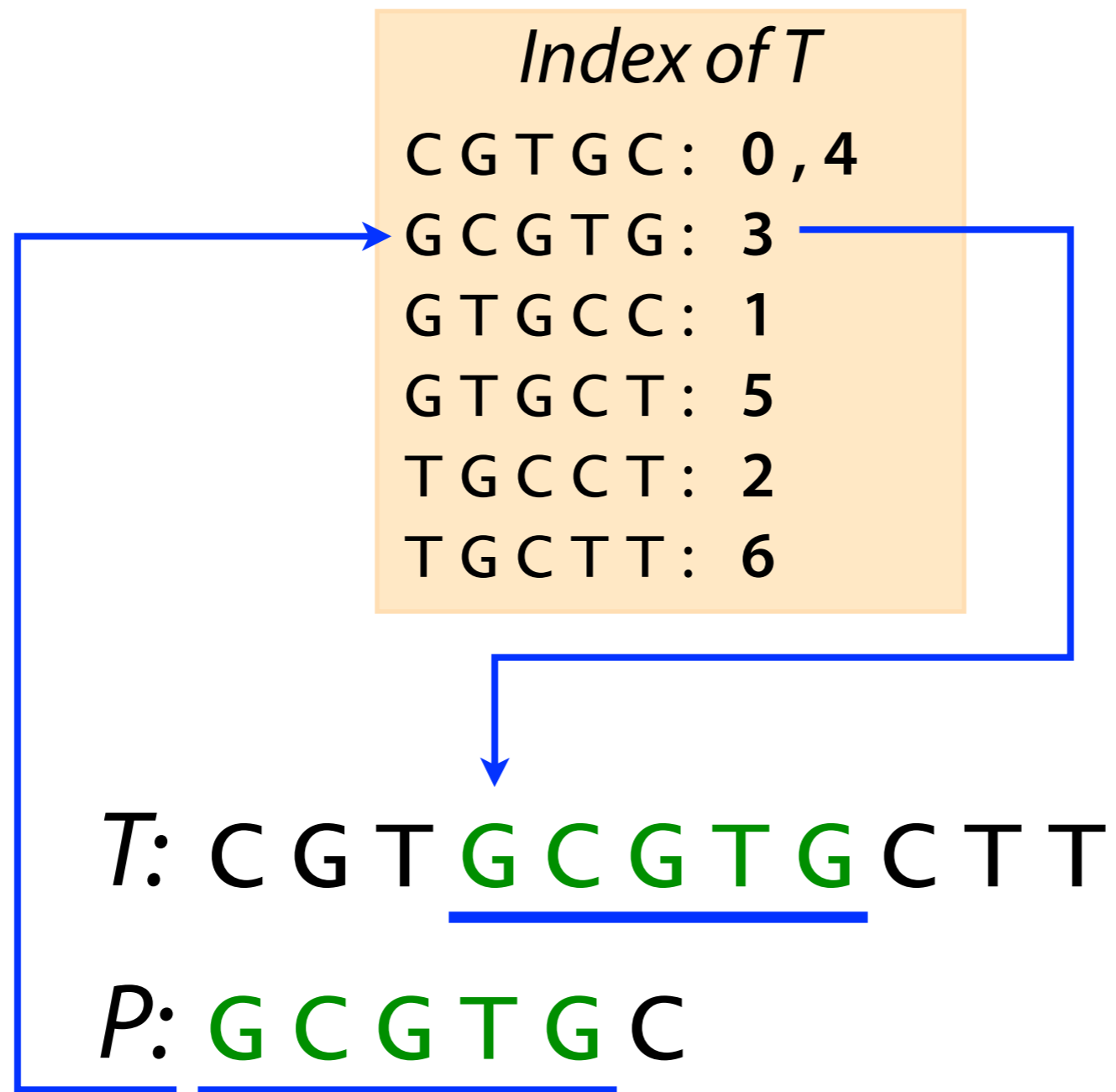T: C G T G C G T G C T T

P: G C G T G C

Querying the index

Index of T

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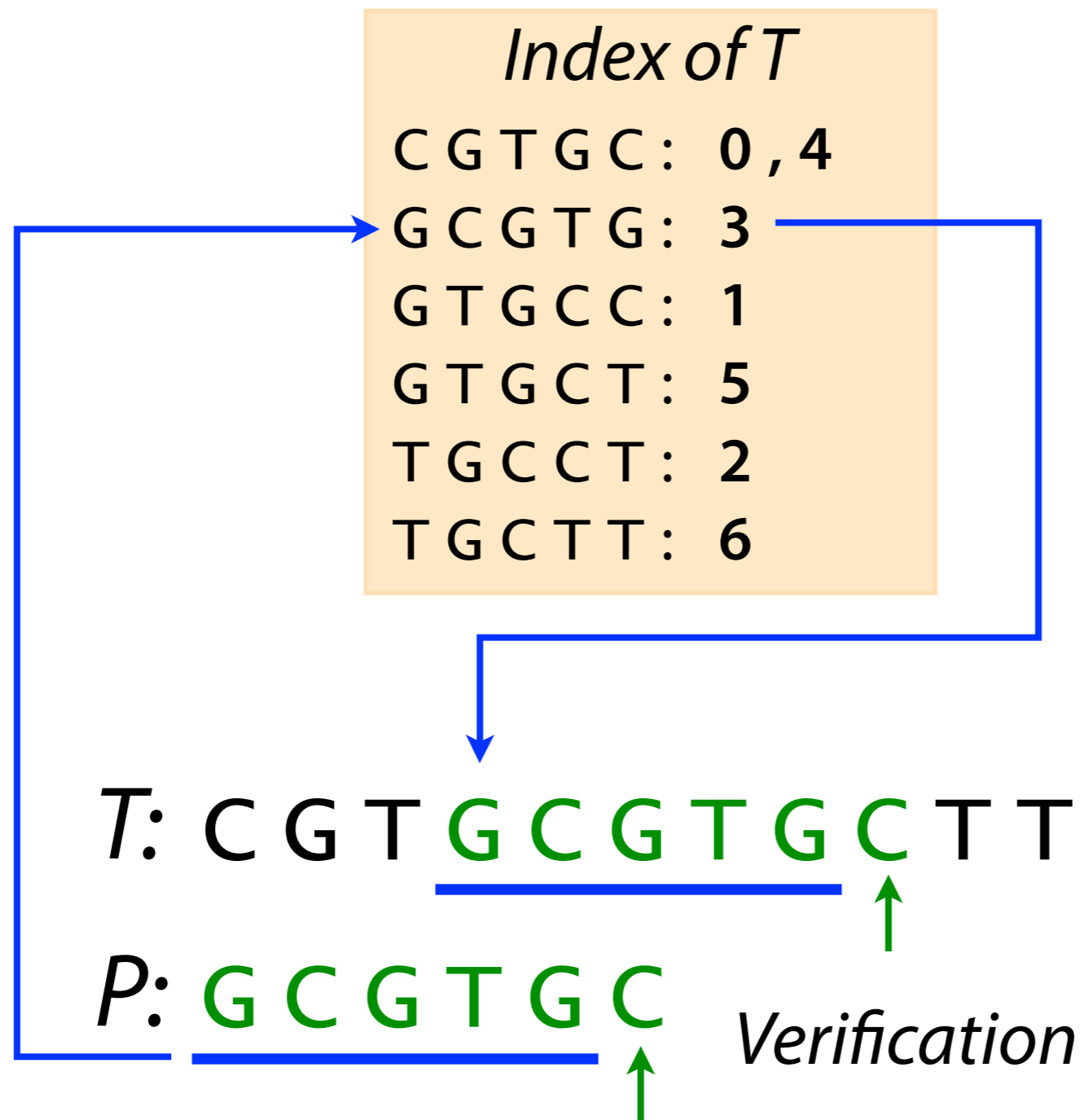GCGTG: 3
GTGCC: 1
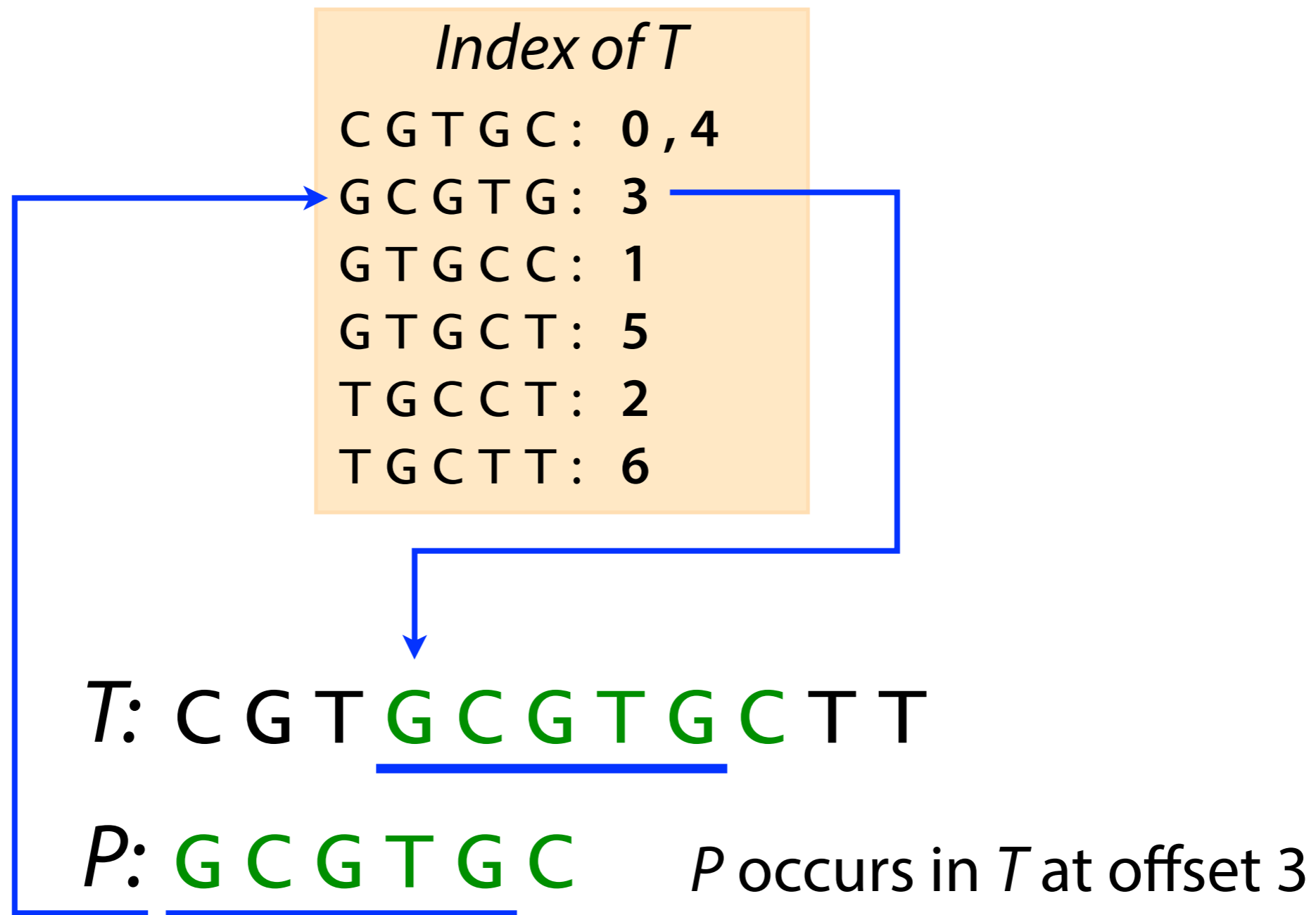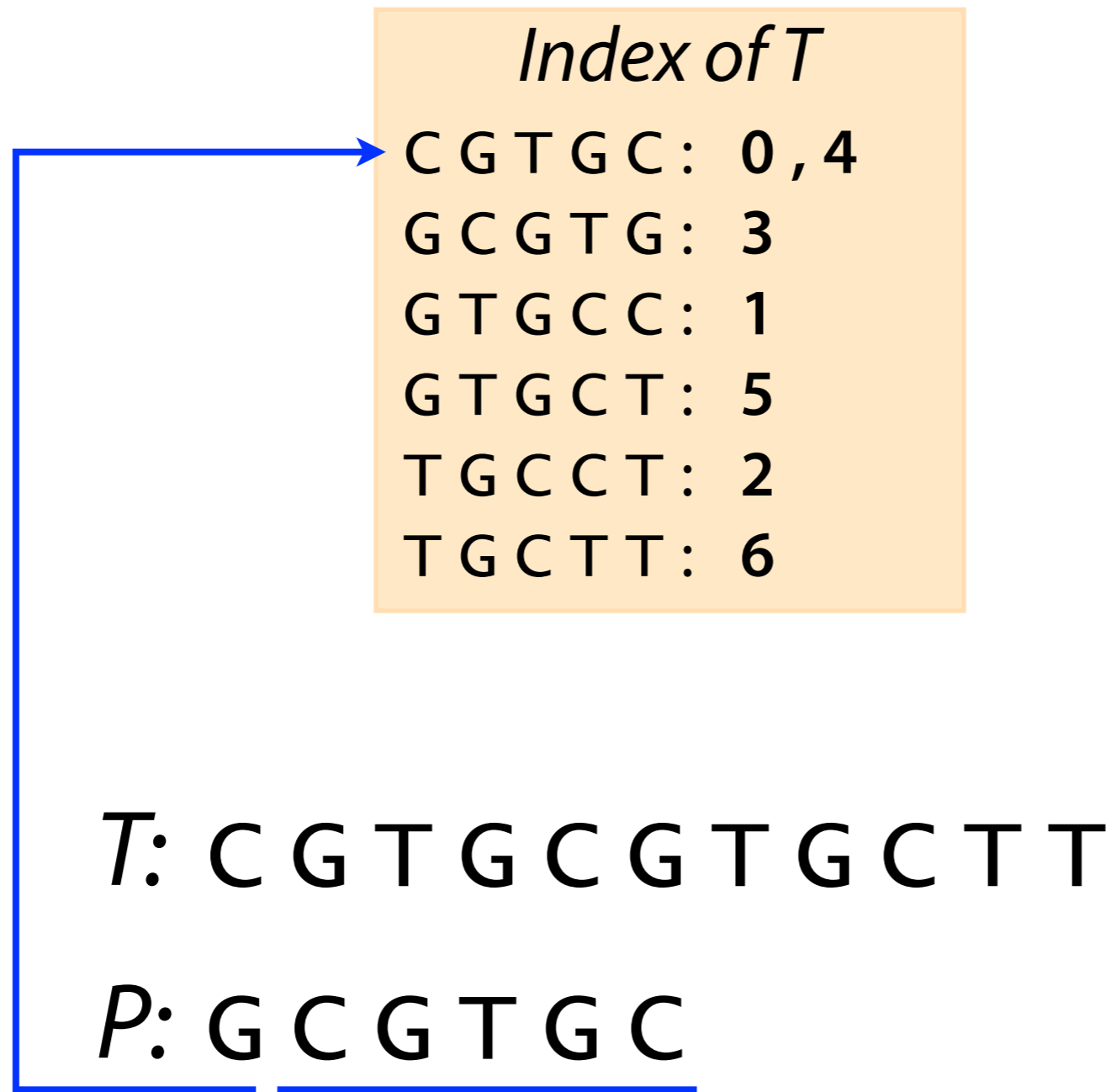GTGCT: 5
TGCCT: 2
TGCTT: 6

T: CGTGCGTGCTT

P: GCGTGC

Querying the index

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**

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

P: G C G T G C

Querying the index

Index of T

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

T: CGTGCGTGCTT

P: GCGTGC   No match

Querying the index

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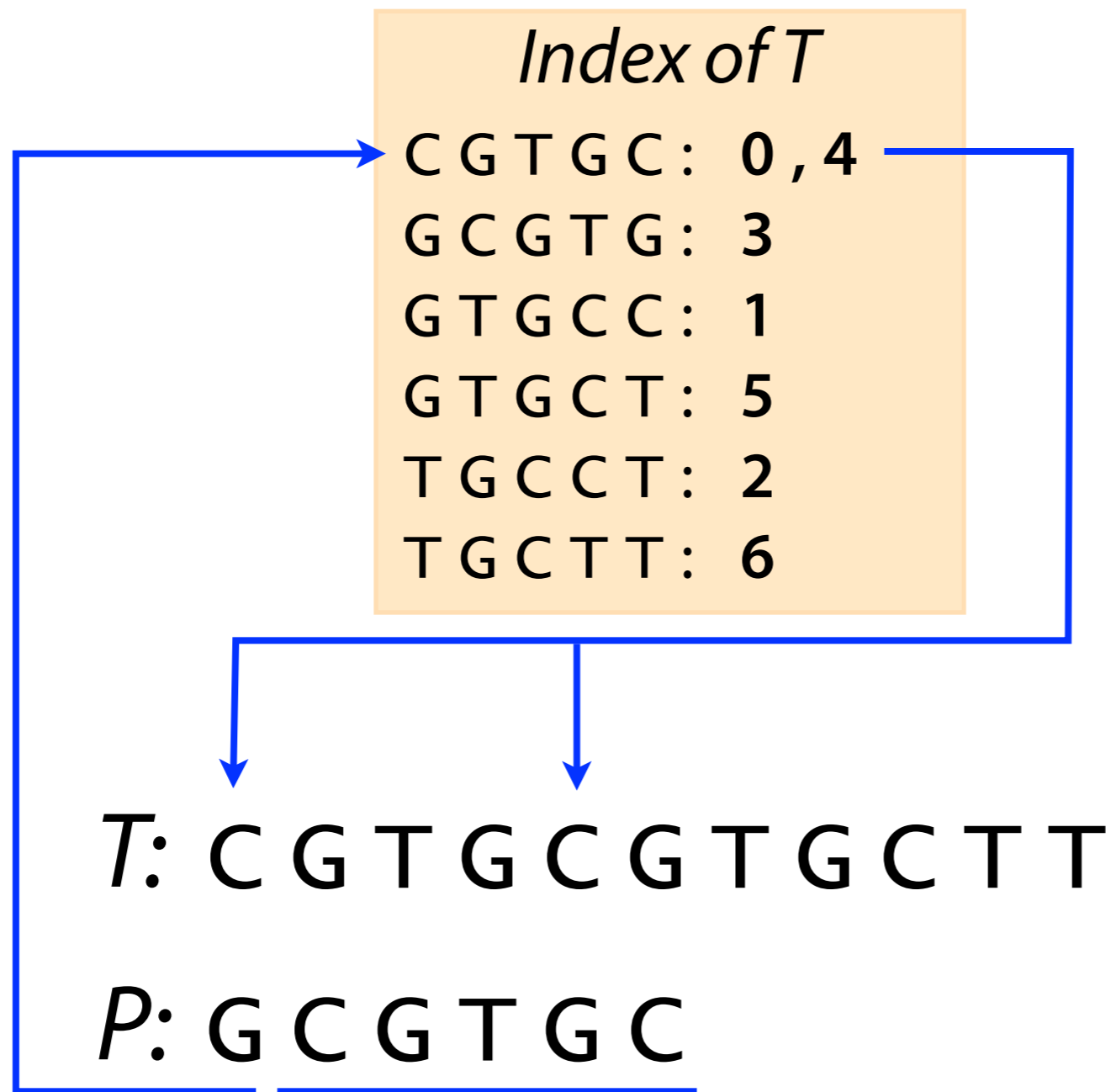G T G C T : **5**
T G C C T : **2**
T G C T T : **6**

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

P: G C G T G C

Querying the index

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**

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

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

P occurs in *T* at offset 3

Querying the index

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 G A

Querying the index
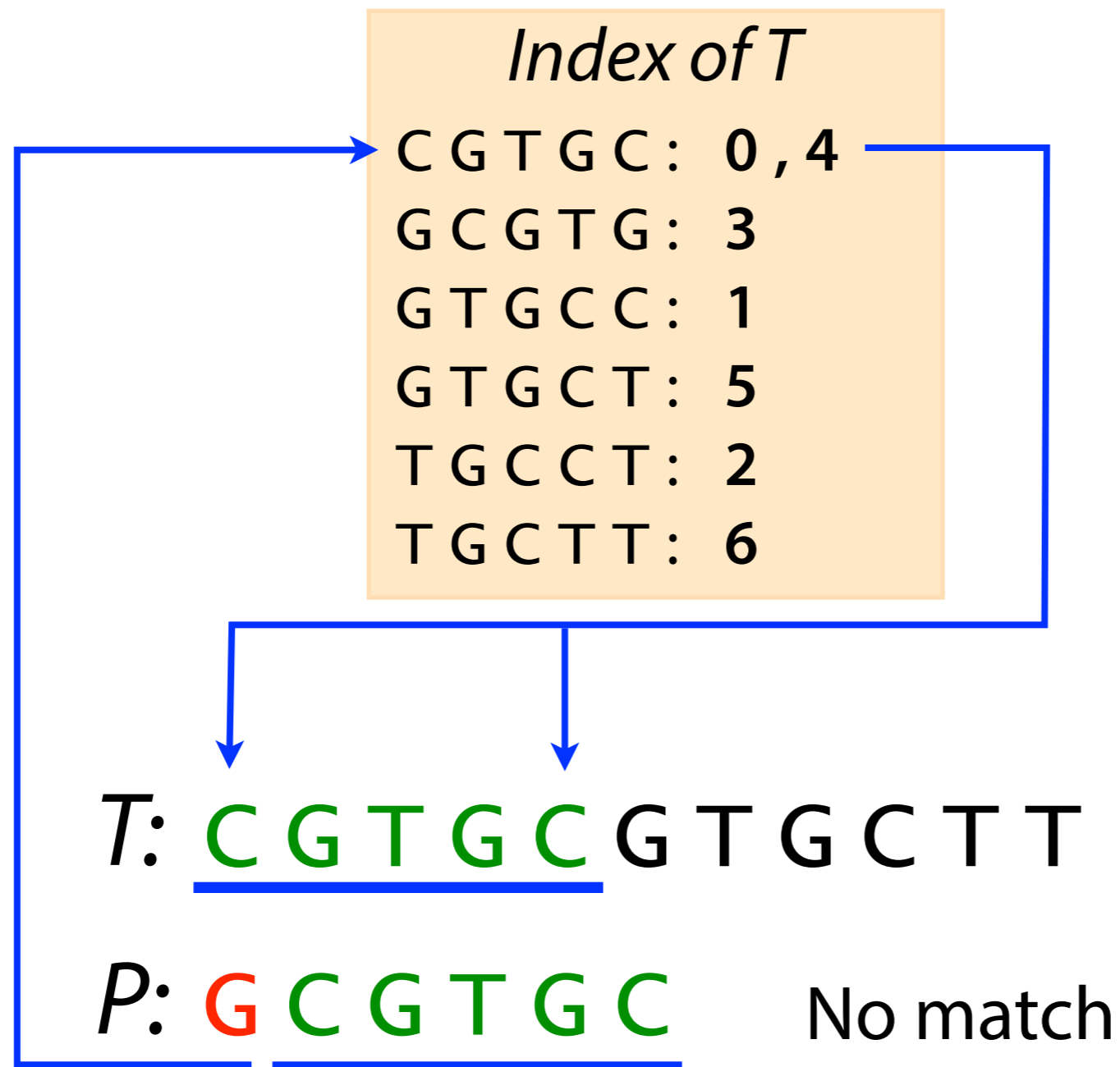

Index of T

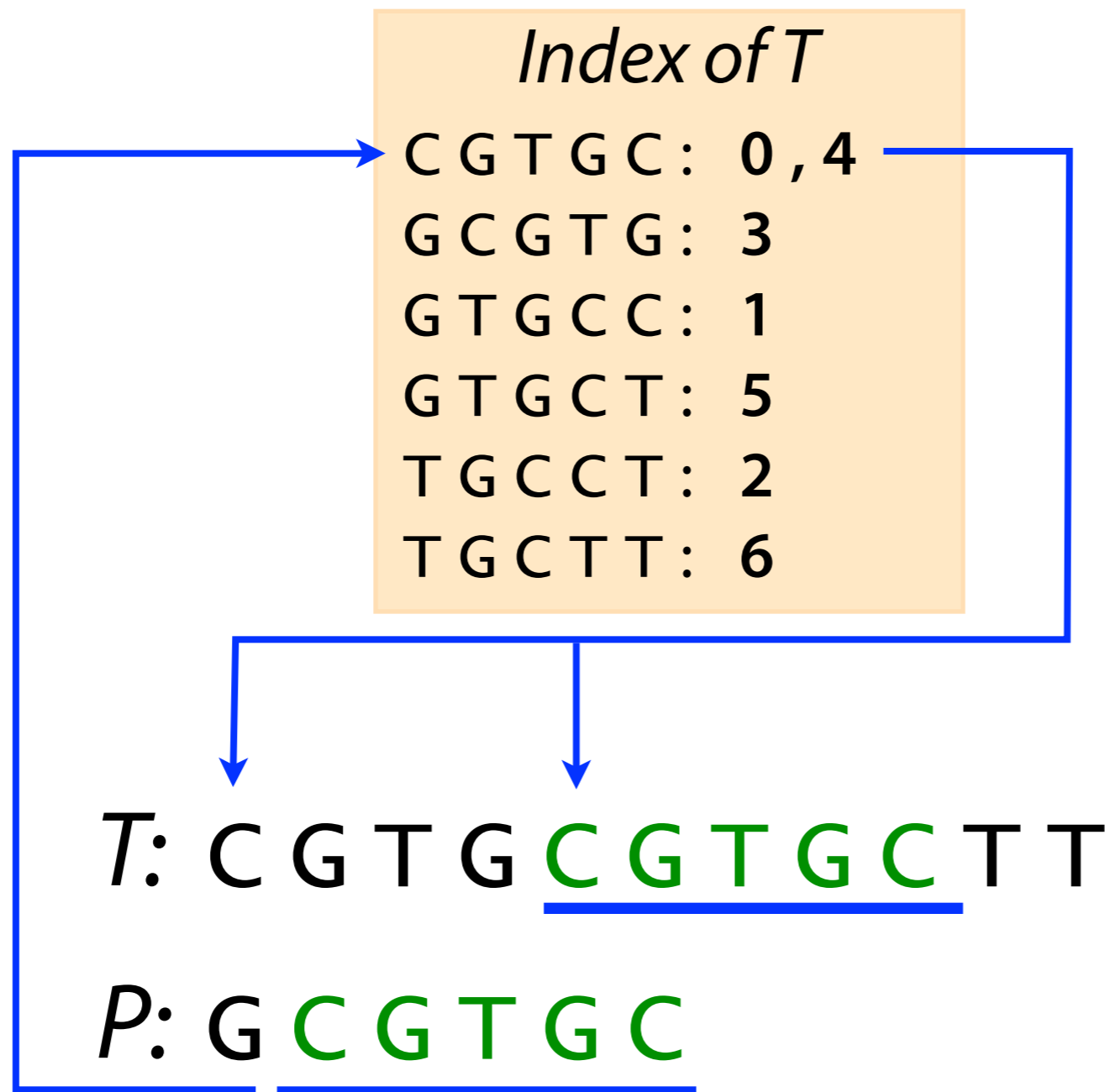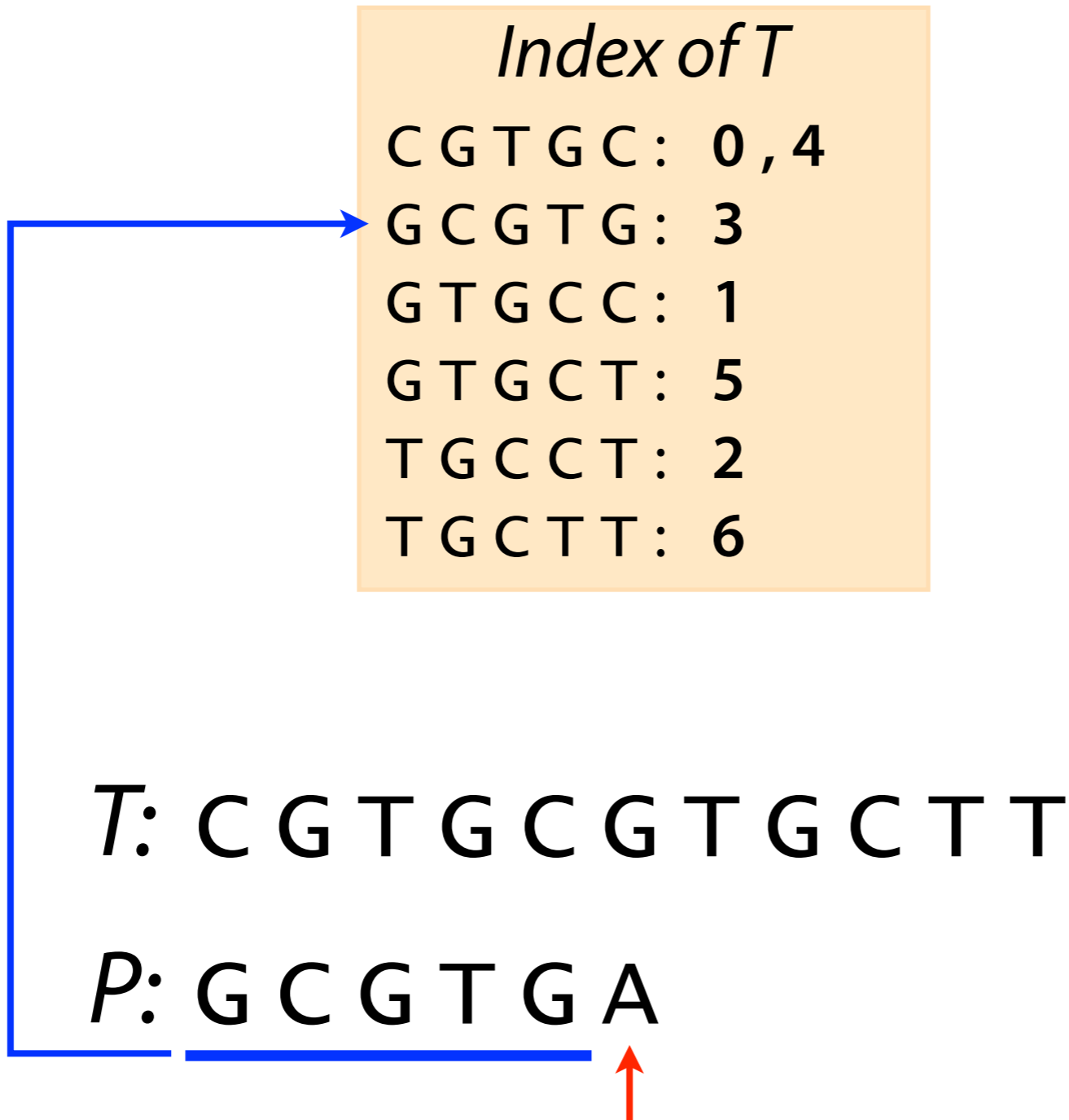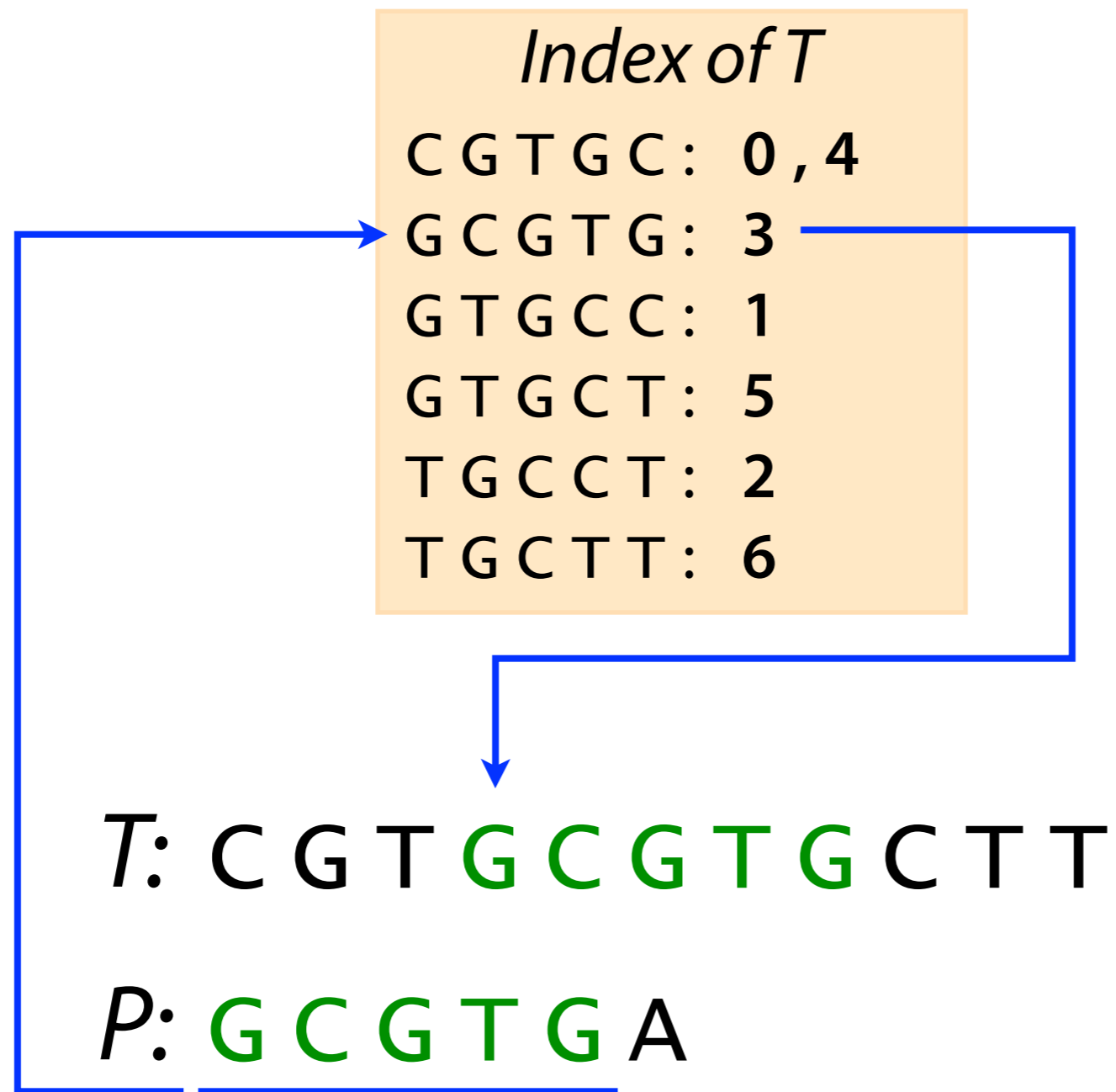C G T G C : **0 , 4**
G C G T G : **3**
G T G C C : **1**
G T G C T : **5**
T G C C T : **2**
T G C T T : **6**

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

P: G C G T G A

Querying the index

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**

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

P: G C G T G A   *P* does not occur in *T*

Querying the index

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**

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

P: G C G T A C

Querying the index

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

X

T: C G T G C G T G C T T
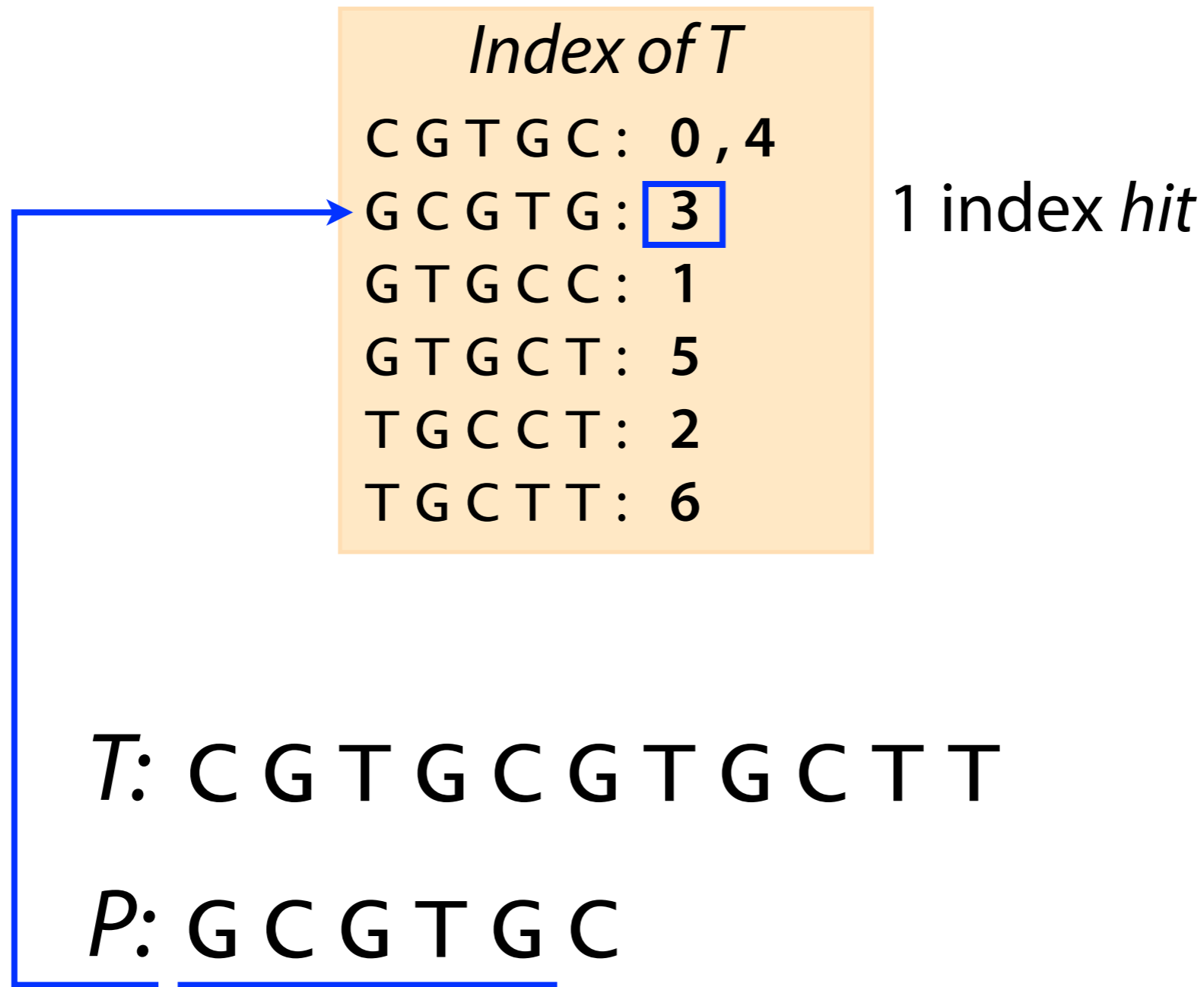
P: G C G T A C   P does not occur in T

Querying the index

Index of T

```
C G T G C :  0 , 4
G C G T G :  3      1 index hit
G T G C C :  1
G T G C T :  5
T G C C T :  2
T G C T T :  6
```

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

P: G C G T G C

Querying the index

Index of T

C G T G C : [**0 , 4**]   2 index *hits*
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**

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

P: G C G T G C