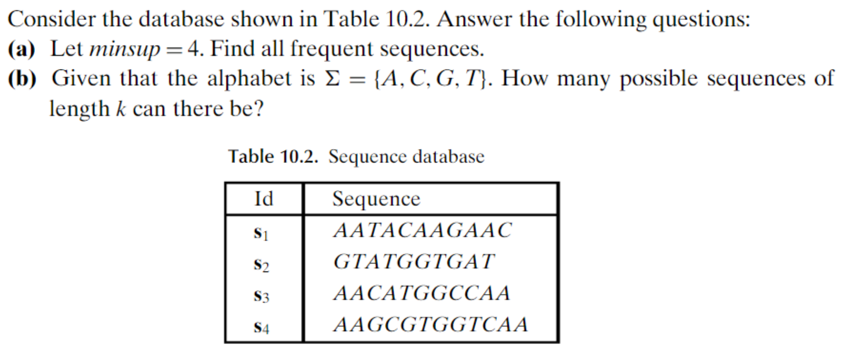
Data Mining (CSE542)

Homework 05

ID: ____ Name: ___조원석__ Date:___2023.04.17__

Task-1



Consider the database shown in Table 10.2. Answer the following questions:

(a) Let *minsup* = 4. Find all frequent sequences.

(b) Given that the alphabet is $\Sigma = \{A, C, G, T\}$. How many possible sequences of length k can there be?

Table 10.2. Sequence database

Id	Sequence
s_1	*AATACAAGAAC*
s_2	*GTATGGTGAT*
s_3	*AACATGGCCAA*
s_4	*AAGCGTGGTCAA*

(a) Let minsup = 4. Find all frequent sequences.

A−4, G−4, T−4

AA −4, AG −4, AT −4, GA −4, TA −4, TG −4

AAT −4, AGA −4, ATA −4, ATG −4, GAA −4, TAA −4, TGA−4,

AATA – 4, ATGA – 4

(b) Given that the alphabet is = {A,C,G,T}. How many possible sequences of length k can there be?

4^k

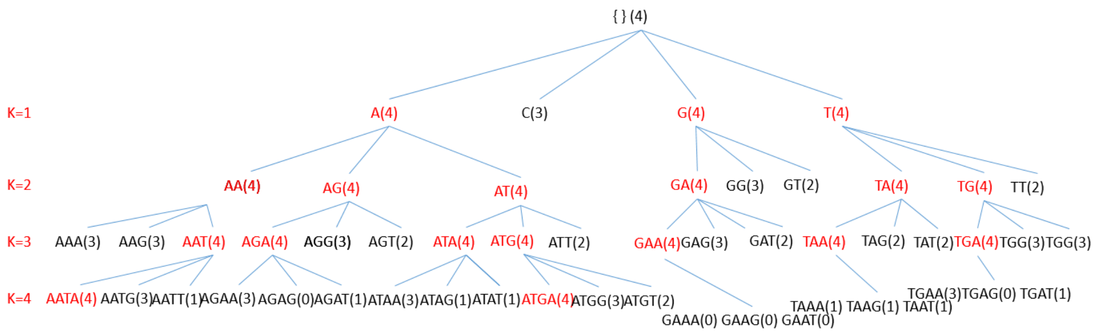
$k = 1, 4^1 = 4$

$k = 2, 4^2 = 16$

$k = 3, 4^3 = 64$

$k = 4, 4^4 = 256$

(c) Show the steps of the PrefixSpan algorithm

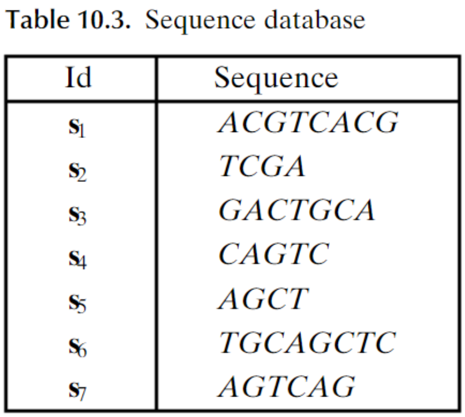


Task-2

(a) Let minsup = 4. Find all frequent sequences

(b) Show the steps of the PrefixSpan algorithm

Table 10.3. Sequence database



Id	Sequence
s_1	*ACGTCACG*
s_2	*TCGA*
s_3	*GACTGCA*
s_4	*CAGTC*
s_5	*AGCT*
s_6	*TGCAGCTC*
s_7	*AGTCAG*

(a) Let minsup = 4. Find all frequent sequences

A–7, C–7, G–7, T–7

AC –6, AG –6, AT –6, CA –6, CC –4, CG –6, CT –5,

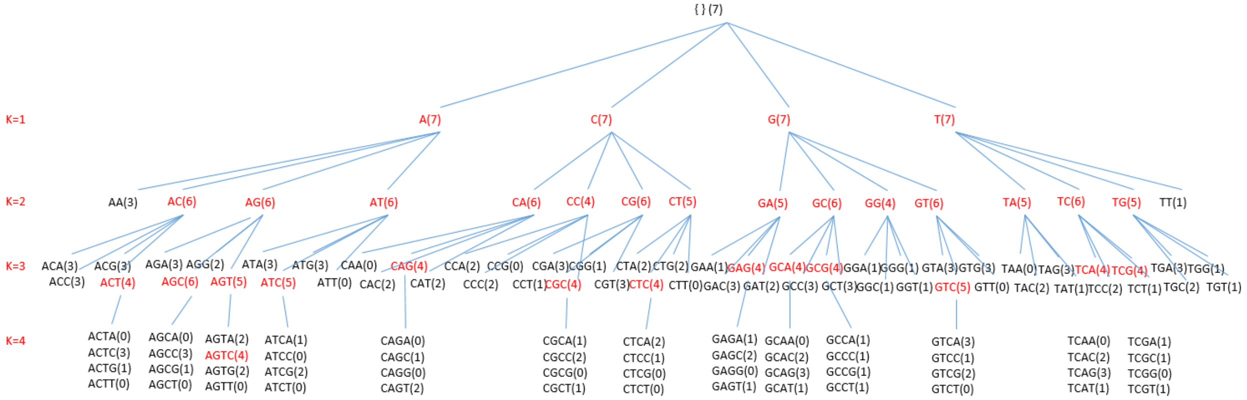
GA –5, GC –6, GG –4, GT –6, TA –5, TC –6, TG –5,

ACT –4, AGC–6, AGT –5, ATC–5, CAG –4, CGC –4, CTC –4,

GAG –4, GCA–4, GCG –4, GTC –5, TCA –4, TCG –4,

AGTC –4

(b) Show the steps of the PrefixSpan algorithm


{ } (7)
K=1
A(7)
C(7)
G(7)
T(7)
K=2
AA(3)
AC(6)
AG(6)
AT(6)
CA(6)
CC(4)
CG(6)
CT(5)
GA(5)
GC(6)
GG(4)
GT(6)
TA(5)
TC(6)
TG(5)
TT(1)
K=3
ACA(3)
ACC(3)
ACG(3)
ACT(4)
AGA(3)
AGG(2)
AGC(6)
AGT(5)
ATA(3)
ATC(5)
ATG(3)
ATT(0)
CAA(0)
CAC(2)
CAG(4)
CAT(2)
CCA(2)
CCC(2)
CCG(0)
CCT(1)
CGA(3)
CGC(4)
CGG(1)
CGT(3)
CTA(2)
CTC(4)
CTG(2)
CTT(0)
GAA(1)
GAC(3)
GAG(4)
GAT(2)
GCA(4)
GCC(3)
GCG(4)
GCT(3)
GGA(1)
GGC(1)
GGG(1)
GGT(1)
GTA(3)
GTC(5)
GTG(3)
GTT(0)
TAA(0)
TAC(2)
TAG(3)
TAT(1)
TCA(4)
TCC(2)
TCG(4)
TCT(1)
TGA(3)
TGC(2)
TGG(1)
TGT(1)
K=4
ACTA(0)
ACTC(3)
ACTG(1)
ACTT(0)
AGCA(0)
AGCC(3)
AGCG(1)
AGCT(0)
AGTA(2)
AGTC(4)
AGTG(2)
AGTT(0)
ATCA(1)
ATCC(0)
ATCG(2)
ATCT(0)
CAGA(0)
CAGC(1)
CAGG(0)
CAGT(2)
CGCA(1)
CGCC(2)
CGCG(0)
CGCT(1)
CTCA(2)
CTCC(1)
CTCG(0)
CTCT(0)
GAGA(1)
GAGC(2)
GAGG(0)
GAGT(1)
GCAA(0)
GCAC(2)
GCAG(3)
GCAT(1)
GCCA(1)
GCCC(1)
GCCG(1)
GCCT(1)
GTCA(3)
GTCC(1)
GTCG(2)
GTCT(0)
TCAA(0)
TCAC(2)
TCAG(3)
TCAT(1)
TCGA(1)
TCGC(1)
TCGG(0)
TCGT(1)