

Komprimierung von Daten

- Mit der Huffman-Codierung ist es verlustfreie Komprimierung von Daten möglich
 - Der Algorithmus arbeitet auf Basis von Zeichenhäufigkeiten
- Der zum Codieren benutzte Codebaum muss auch beim Decodieren bekannt sein
 - Will man nicht jedes Mal den “passenden” Codebaum mitliefern, muss man sich vorab auf einen einigen, der für möglichst viele Anwendungsfälle “gut” passt

- Benannt nach seinen Erfindern Abraham Lempel, Jacob Ziv, Terry A. Welch
- Verlustfreies Kompressionsverfahren
- Idee: Erstelle ein Wörterbuch, das nicht nur Zeichen, sondern auch Zeichenfolgen einen Code fester Länge zuweist
 - Häufig vorkommende Zeichenfolgen lassen sich dann kürzer codieren
- Vorteil:
 - Der Algorithmus arbeitet so, dass das Wörterbuch nicht explizit mitgeliefert werden muss, sondern beim Decodieren (bzw. Dekomprimieren) rekonstruiert werden kann

LZW-Verfahren: Komprimierung

Initialisiere Wörterbuch mit den vorkommenden Einzelzeichen

Muster = Erstes Eingabezeichen

solange Eingabe nicht leer:

 Z = nächstes Eingabezeichen

 falls (Muster+Z) im Wörterbuch:

 Muster = (Muster+Z)

 ansonsten:

 gib den Code von Muster aus

 füge (Muster+Z) dem Wörterbuch hinzu

 Muster = Z

gib den Code von Muster aus

Beispiel Komprimierung - Festlegungen

- Wir betrachten eine DNA-Sequenz, bestehend aus einer Folge der vier DNA-Basen Adenin (A), Guanin (G), Cytosin (C) und Thymin (T)
 - GATAAATCTGGTCTTATTCC
- Annahme:
 - Unkomprimierte Codierung mit 2 Bit pro Zeichen
- Vereinfachung:
 - LZW-Codierung mit fester Länge: 4 Bit pro Wörterbucheintrag

Beispiel Komprimierung - Initialisierung

GATAAATCTGGTCTTATTTCC

Wörterbuch:

A = 0000

G = 0001

C = 0010

T = 0011

Muster = "G"

Beispiel Komprimierung - 1. Durchlauf

GATAAATCTGGTCTTATTTC

Wörterbuch:

A = 0000

G = 0001

C = 0010

T = 0011

Muster = "G"

Z = 'A'

(Muster+Z) = "GA" (nicht im Wörterbuch)

Ausgabe 0001

"GA" ins Wörterbuch einfügen

Muster = "A"

Beispiel Komprimierung - 2. Durchlauf

GATAAATCTGGTCTTATTTC

Wörterbuch:

A = 0000

G = 0001

C = 0010

T = 0011

GA = 0100

Muster = "A"

Z = 'T'

(Muster+Z) = "AT" (nicht im Wörterbuch)

Ausgabe 0000

"AT" ins Wörterbuch einfügen

Muster = "T"

Beispiel Komprimierung - 3. Durchlauf

GATAAATCTGGTCTTATTTC

Wörterbuch:

A = 0000

G = 0001

C = 0010

T = 0011

GA = 0100

AT = 0101

Muster = "T"

Z = 'A'

(Muster+Z) = "TA" (nicht im Wörterbuch)

Ausgabe 0011

"TA" ins Wörterbuch einfügen

Muster = "A"

Beispiel Komprimierung - 4. Durchlauf

GATAAATCTGGTCTTATTTC

Wörterbuch:

A = 0000

G = 0001

C = 0010

T = 0011

GA = 0100

AT = 0101

TA = 0110

Muster = "A"

Z = 'A'

(Muster+Z) = "AA" (nicht im Wörterbuch)

Ausgabe 0000

"AA" ins Wörterbuch einfügen

Muster = "A"

Beispiel Komprimierung - 5. Durchlauf

GATAAATCTGGTCTTATTTC

Wörterbuch:

A = 0000

G = 0001

C = 0010

T = 0011

GA = 0100

AT = 0101

TA = 0110

AA = 0111

Muster = "A"

Z = 'A'

(Muster+Z) = "AA" (im Wörterbuch)

Muster = "AA"

Beispiel Komprimierung - 6. Durchlauf

GATAAATCTGGTCTTATTTC

Wörterbuch:

A = 0000

G = 0001

C = 0010

T = 0011

GA = 0100

AT = 0101

TA = 0110

AA = 0111

Muster = "AA"

Z = 'T'

(Muster+Z) = "AAT" (nicht im Wörterbuch)

Ausgabe 0111

"AAT" ins Wörterbuch einfügen

Muster = "T"

Beispiel Komprimierung - Zustand nach dem 6. Durchlauf

GATAAATCTGGTCTTATTTC

Wörterbuch:

A = 0000

G = 0001

C = 0010

T = 0011

GA = 0100

AT = 0101

TA = 0110

AA = 0111

AAT = 1000

Bisherige Ausgabe

0001 0000 0011 0000 0111

Aufgabe 1

- Führe die Kompression der Zeichenfolge **GATAAATCTGGTCTTATTTC** mit den genannten Bedingung von Hand vollständig durch

LZW-Verfahren: Dekomprimierung

Initialisiere Wörterbuch mit den vorkommenden Einzelzeichen

Code = Erstes Eingabezeichen

gib Muster(Code) aus

AlterCode = Code

solange Eingabe nicht leer:

Code = nächstes Eingabezeichen

falls Code im Wörterbuch:

M = Muster(Code)

gib M aus

füge (Muster(AlterCode)+M[0]) dem Wörterbuch hinzu

ansonsten:

Tmp = Muster(AlterCode)

Tmp = Tmp + Tmp[0]

gib Tmp aus

füge Tmp dem Wörterbuch hinzu

AlterCode = Code

Beispiel Dekomprimierung - Initialisierung

0001 0000 0011 0000 0111

Wörterbuch:

A = 0000

G = 0001

C = 0010

T = 0011

Code = 0001

Ausgabe "G"

AlterCode = 0001

Beispiel Dekomprimierung - 1. Durchlauf

0001 0000 0011 0000 0111

Wörterbuch:

A = 0000

G = 0001

C = 0010

T = 0011

Code = 0000 (im Wörterbuch)

M = Muster(0000) = "A"

Ausgabe "A"

füge "GA" dem Wörterbuch hinzu

AlterCode = 0000

Beispiel Dekomprimierung - 2. Durchlauf

0001 0000 0011 0000 0111

Wörterbuch:

A = 0000

G = 0001

C = 0010

T = 0011

GA = 0100

Code = 0011 (im Wörterbuch)

M = Muster(0011) = "T"

Ausgabe "T"

füge "AT" dem Wörterbuch hinzu

AlterCode = 0011

Beispiel Dekomprimierung - 3. Durchlauf

0001 0000 0011 0000 0111

Wörterbuch:

A = 0000

G = 0001

C = 0010

T = 0011

GA = 0100

AT = 0101

Code = 0000 (im Wörterbuch)

M = Muster(0000) = "A"

Ausgabe "A"

füge "TA" dem Wörterbuch hinzu

AlterCode = 0000

Beispiel Dekomprimierung - 4. Durchlauf

0001 0000 0011 0000 0111

Wörterbuch:

A = 0000

G = 0001

C = 0010

T = 0011

GA = 0100

AT = 0101

TA = 0110

Code = 0111 (nicht im Wörterbuch)

Tmp = Muster(0000) = "A"

Tmp = "A" + 'A' = "AA"

Ausgabe "AA"

füge "AA" dem Wörterbuch hinzu

AlterCode = 0111

Beispiel Dekomprimierung - Zustand nach dem 4. Durchlauf

0001 0000 0011 0000 0111

Wörterbuch:

Bisherige Ausgabe

A = 0000

GATAAA

G = 0001

C = 0010

T = 0011

GA = 0100

AT = 0101

TA = 0110

AA = 0111