Compactação Trivial

Economizar espaço (virtual ou real) muitas vezes é importante. Usar menos espaço é mais eficiente, e é possível economizar dinheiro. Se você estivesse alugando um apartamento que fosse maior do que o necessário para suas coisas e a sua família, seria possível fazer um “downsize” para um lugar menor, mais barato. Se você paga por byte para armazenar seus dados em um servidor, talvez queira compactá-los para que a armazenagem tenha um custo menor. A compactação é o ato de tomar os dados e codificá-los (modificar o seu formato) de modo que ocupem menos espaço. A descompactação é o processo inverso, que faz com que os dados retornem ao seu formato original.

Considere os nucleotídeos que formam um gene no DNA. Cada nucleotídeo pode assumir apenas um entre quatro valores: A, C, G ou T. No entanto, se o gene for armazenado como uma string, que pode ser imaginada como uma coleção de caracteres Unicode, cada nucleotídeo será representado por um caractere, o qual, em geral, exige 8 bits para armazenagem. Em binário, apenas 2 bits são necessários para armazenar um tipo com quatro valores possíveis: 00, 01, 10 e 11 são os quatro valores diferentes que podem ser representados por 2 bits. Se atribuirmos o valor 00 a A, 01 a C, 10 a G e 11 a T, a área de armazenagem necessária para uma string de nucleotídeos poderá ser reduzida em 75% (de 8 bits para 2 bits por nucleotídeo).

Compactar a sequência de um gene:
"TAGGGATTAACCGTTATATATATATAGCCATGGATCGATTATATAGGGATTAACCGTTATATATATATAGCCATGGATCGATTATA"

ACGT -> 32bits

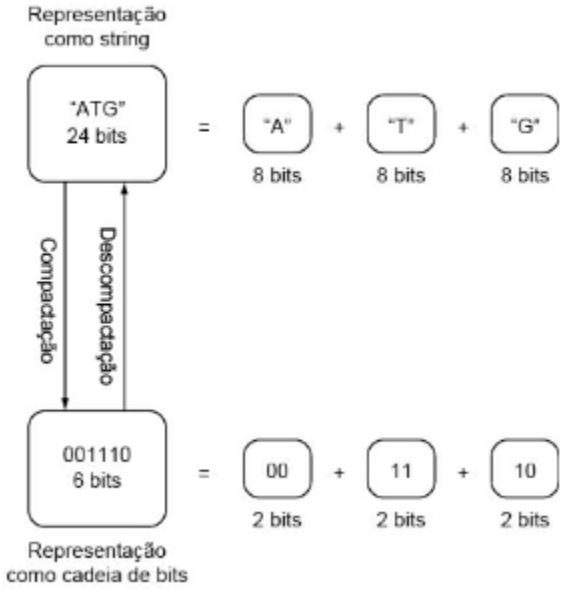
ACGT -> 8bits

2Bits: 00, 01, 10, 11 -> 2^2 -> 4

ABCDEFGH -> 64 bits

ABCDEFGH -> 24 bits

3Bits: 000,001,010,011,100,101,110,111 ->2^3 -> 8



Livro: Kopec, David. Problemas Clássicos de Ciência da Computação com Python (p. 36). Novatec Editora. Kindle Edition.

Sequência de um vírus:

ACCGCGCCTTCTTTTCCTGCGAGGGCCCGGTAGGGACCGAGCGCTTTGATTTAAAGCCTGGTTCTGCTTT
GTATGATTTATCTAAAGCAGCCCAATCTAAAGAAACCGGTCCCGGGCACTATAAATTGCCTAACAAGTGC
GATTCATTCATGGATCCACAGAACGCCCTGTATTATCAGCCGCGGGTACCCACAGCAGCTCCGACATCCG
GAGGAGTGCCGTGGAGTCGCGTAGGCGAGGTAGCTATTTTGAGCTTTGTTGCATTGATTTGCTTTTACCT
GCTTTACCTTTGGGTGCTGAGAGACCTTATCTTAGTTCTGAAGGCTCGACAAGGCAGATCCACGGAGGA
GCTGATATTTGGTGGACAAGCTGTGGATAGGAGCAACCCTATCCCTAATCTACCTTCACCACCAAGTCAG
GGCAATCCCGGGCCATTTGTTCCAGGCACGGGATAAGCAATCAGCCATGTCCACGTCCAAGAGGAAGCG
GGGAGATGATTCGAATTGGAATAAGCGGGTGCCTAAGAAGAAGCCATCTTCAGCTGGGCTGAAGAGGG
CTGGAAGCAAGGCCGATAGGCCATCCCTCCAAATCCAGACACTCCAGCATGCTGGGACCACCATGATAA
CTGTCCCATCCGGAGGAGTATGTGACCTCATCAACACCTATGCCCGAGGATCTGACGAGGGCAACCGCC
ACACCAGCGAGACTCTGACGTACAAGATCGCCGTCGACTACCACTTCGTTGCCGACGCGGCTGCCTGCCG
CTACTCCAACACCGGAACCGGTGTAATGTGGCTGGTGTATGACACCACTCCCGGCGGACAAGCTCCGACC
CCGCAAACTATATTTGCCTACCCTGACACGCTAAAAGCGTGGCCGGCCACATGGAAAGTGAGCCGGGAG
CTGTGTCATCGCTTCGTGGTGAAACGGCGATGGTTGTTCAACATGGAGACCGACGGTCGGATTGGTTCG
GATATCCCTCCCTCGAATACAAGTTGGAAGCCTTGCAAGCGCAACATCTACTTCCACAAGTTCACGAGTG
GGTTGGGAGTGAGAACGCAGTGGAAGAATGTAACGGACGGAGGAGTTGGTGCCATCCAGAGAGGAGC

TCTGTACATGGTCATTGCCCCAGGCAATGGCCTTACTTTTACTGCCCATGGGCAGACCCGTCTGTACTTTA
AGAGTGTTGGCAACCAGTAATGAATAAAAACTCCCGTTTTATTATATTTGATGAATGCTGAAAGCTTACAT
TAATATGTCGTGCGATGGCACGAAAAAACACACGCAAACAATACAGGGGGGTAGTCGGCGGGCGGCTA
AGGGTGGTGCTCGGCGGGCAGAACATCGAAAAATCAAGATCTATATGAATTACACTTCCTCCGTAGGAG
GAAGCACAGGGGGAGAATACCACTTCTCCCCCGGCGACATAATGTAAATGACGCAGTTTGCCTCGAAAT
ACTCCAGCTGCCCTGGAGTCATTTCCTTCATCCAATCTTCATCCGAGTTGGCGAGGATTATTGTAGGCTTA
GACTTCTTCTGCACCTTTTTCTTCTTACCATACTTGGGGTTTACAATGAAATCCCTCTGACAGCCAACTAAC
TGTTTCCAACAAGGACAGAATTTAAACGGAATATCATCTACGATGTTGTAGATTGCGTCTTCGTTGTATGA
AGACCAATCAACATTATTTTGCCAGTAATTATGAACCCCTAGGCTTCTGGCCCAAGTAGATTTTCCGGTTC
TTGTTGGGCCGACGATGTAGAGGCTCTGCTTTCTTGATCTTTCATCTGATGACTGGATACAGAATCCATCC
ATTGGAGGTCAGAAATTGCATCCTCGAGGGTATAACAGGTAGGTTGAAGGAGCATGTAAGCTTCGGGAC
TAACCTGGAAGATGTTAGGCTGGAGCCAATCGTTGATTGACTCATTACAAAGTAAATCAGGTGAGGAGG
GTGGATGAGGATTGGTGAACTCTTCCTGAATCTCAGGAAAAAGCTTATTTGCAGAGTATTCAAAATACTG
CAATTTTGTGGACCAATCAAAGGGGAGCTCTTTCTGGATCATGGAGAGGTACTCTTCTTTGGAGGTAGCG
TGTGAAATAATGTCTCGCATTATTTCATCTTTAGAAGGCTTTTTTTCCTTTACCTCTGAATCAGATTTTCCTA
GGAAGGGGGACTTCCTAGGAATGAAAGTACCTCTCTCAAACACAGCCAGAGGTTCCTTGAGAATGTAAT
CCCTCACTCTGTTAACTGACTTGGCACTCTGAATATTTGGGTGAAACCCATTTATATCAAAGAACCTTGAG
TCAGATATCCTTATCGGCTTCTCTGGCTGAAGCAATGCATGTAAATGCAAACTTCCATCTTTATGTGCCTCT
CGGGCACATAGAATATATTTGGGAATCCAACGAACGACGAGCTCCCAGATCATCTGACAGGCGATTTCA
GGATTTTCTGGACACTTTGGATAGGTTAGGAACGTGTTAGCGTTCCTGTGTGAGAACTGACGGTTGGATG
AGGAGGAGGCCATAGCCGACGACGGAGGTTGAGGCTGAGGGATGGCAGACTGGGAGCTCCAAACTCT
ATAGTATACCCGTGCGCCTTCGAAATCCGCCGCTCCATTGTCTTATAGTGGTTGTAAATGGGCCGGACCG
GGCCGGCCCAGCAGGAAAAGAAGGCGCGCACTAATAT