Genome Compression Based on Hilbert Space Filling Curve

HongYi Guo[1, a], Min Chen[1,b *], Xi Liu[1,c] and MengMeng Xie[1,d]

[1]Information Security College, Yunnan Police Officer Academy, Kunming, Yunnan, China

[a]614497679@qq.com, [b]minkeychen@sina.cn, [c]1653554425@qq.com , [d]1197473983@qq.com

* corresponding author: Min Chen, minkeychen@sina.cn

Keywords: Genome sequence compression; Hilbert space filling; Context weighting; Description length

Abstract. The genome compression algorithm based on the Hilbert space-filling curve is proposed in this paper. In order to utilize the correlations among the bases in the genome sequence, our algorithm firstly use the Hilbert space filling curve to map the genome sequence into a new 2-D image. Then the image obtained is encoded by the context weighting modeling technology. For context weighting, the values of weights are corresponding to the description length of the context model which is corresponding to their weights respectively. When the receiver obtain the mapping image, the reverse Hilbert space filling matrix is used to help the decoding procedure. The experiments results indicate that the final compression results by our algorithm are better than the results by other similar algorithms, although the useless area which will lead to the reduce of the compression efficiency is resulted in the mapping image by our algorithm.

Introduction

The genome compression can reduce the genome storage cost to improve the compression efficiency. Although the efficient compression results can be achieved by using the substitution algorithms[1-3]. However, when the size of the dictionary becomes larger and larger, the costs to code the item indexes will become too long, which actually reduce the compression efficiency. Meanwhile, in order to reduce the coding cost in the substitution algorithms, the dictionary can be initialized. But in this way, the cost to code the dictionary should not be ignored. Especially in [3], although the highest compression efficiency can be achieved by using this algorithm, however, the cost to code the dictionary which is used in this algorithm is not included in its code length. If the receiver has no this dictionary, the decoding process is not executed. It implies that the cost to describe the dictionary can not be ignored in the substitution algorithms.

On the other hand, the entropy coding based on the context modeling technology is used for the genome sequence compression. These algorithms can obtain the high compression efficiency by using the correlation among the bases in the genome sequence to construct the conditional probability distributions to drive the arithmetic encoder. However, due to the indels in the genome sequence, the context modeling by using the past bases directly can not obtain the context model with high coding efficiency. In[4], Pihno concluded that the compression results by using the context model with finite orders will be better than the results by using the context model with higher order. In order to sufficiently utilize the correlation among bases but not to lead to the high model order, the context weighting is employed. The similar algorithm[5-7]are also use context modeling method for genome compression. However, all of these algorithms fall into that the initial genome sequence are not deal for ensuring more correlation can be used.

In this paper, the Hilbert spacing filling curve is suggested to map the genome sequence into 2-D images in order to achieve higher compression efficiency. The details will be discussed in this paper.

Context modeling

Let $x_n,\ldots,x_0$ denote a genome sequence and each x_t denotes the current base to be coded. The codelength L for coding this genome sequence can be described as:

$$L = -\log P(x_n,\ldots,x_0) = -\log \prod_{t=1}^{n} P(x_t \mid x_{t-1},\ldots,x_0) \tag{1}$$

Where $P(x_n,\ldots,x_0)$ denotes the joint probability distribution of $x_n,\ldots,x_0$. However, it is not easy to estimate these conditional probability distributions $P(x_t \mid x_{t-1},\ldots,x_0)$ in practice. One intuitive method is to estimate distributions $P(x_t \mid x_{t-1},\ldots,x_{t-K})$. Thus, L can be described approximately as (2)

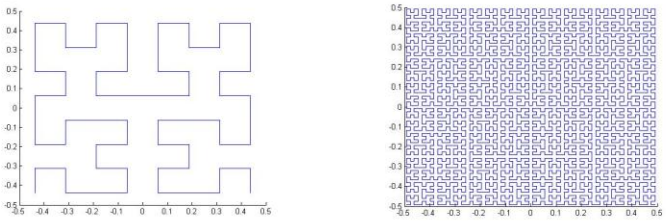
$$L = -\log \prod_{t=1}^{n} P(x_t \mid x_{t-1},\ldots,x_{t-K}) \tag{2}$$

Where K is the order of the context model. Theoretically, the entropy of $P(x_t \mid x_{t-1},\ldots,x_{t-K})$ is larger than the entropy of $P(x_t \mid x_{t-1},\ldots,x_0)$, namely, this approximate estimation may increase the codelength. However, for some memoryless source with order K, it is satisfied that $P(x_t \mid x_{t-1},\ldots,x_{t-K}) = P(x_t \mid x_{t-1},\ldots,x_0)$, or $H(x_t \mid x_{t-1},\ldots,x_{t-K}) \approx H(x_t \mid x_{t-1},\ldots,x_{t-K-1})$. It implies that increasing the number of bases which are conditioned on the current base x_t does not always reduce the entropy furthermore. To this way, in purpose of achieving better estimation, these K bases should be mostly correlated to x_t. For some sources, such as images, the current symbol and its neighboring symbols are correlated with each. Thus, in context modeling for those sources, the past K symbols are used as conditions. However, for genome sequence, due to the existences of indels and fracture sub sequences, correlations among these neighboring bases are weaken. In context modeling for genome sequence, conditional probability distributions with high compression efficiency can not be obtained by directly using past K bases for construction.

On the other hands, any operations of reducing dimensions can weak the correlations among neighboring symbols and verse vice. Namely, if the genome sequence can be mapped into signals with two dimensions, the more correlations among bases could be used for context modeling. In this paper, Hilbert space filling technology is employed to do this mapping operation.

Hilbert space filling

One signal in dimension 1 can be mapped into dimension 2 by using Hilbert space filling curve. When the order of Hilbert curve is given, the curve is extended to 2-dimension space and the corresponding space can be filled sufficiently with the increment of orders. The Hilbert space filling curves with order 3 and 6 are shown in Fig. 1.



(a) order 3 (b) order 6

Fig. 1 Two Hilbert space filling curves with order 3 and 6

In practice, Hilbert space filling curve is used to map one genome sequence into 2 dimension. In order to implement this mapping operation, the Hilbert space filling matrix should be obtained firstly.

One Hilbert space filling matrix can be obtained by recursion operation which is described as (3)

$$\mathbf{S}^{k'} = \begin{cases} \begin{bmatrix} \mathbf{S}^{k'-1} & 4^{k'}\mathbf{E}^{k'-1} + \mathbf{S}^{k'-1} \\ 4^{(k'+1)}\mathbf{E}^{k'-1} - \mathbf{S}^{k'-1} & (3\times 4^{k'}+1)\mathbf{E}^{k'-1} - (\mathbf{S}^{k'-1})^T \end{bmatrix} & k' \text{ is odd number} \\ \begin{bmatrix} \mathbf{S}^{k'-1} & (4^{k'+1}+1)\mathbf{E}^{k'-1} - \mathbf{S}^{k'-1} \\ 4^{k'}\mathbf{E}^{k'-1} + (\mathbf{S}^{k'-1})^T & (3\times 4^{k'+1})\mathbf{E}^{k'-1} - (\mathbf{S}^{k'-1})^T \end{bmatrix} & k' \text{ is even number} \end{cases} \quad (3)$$

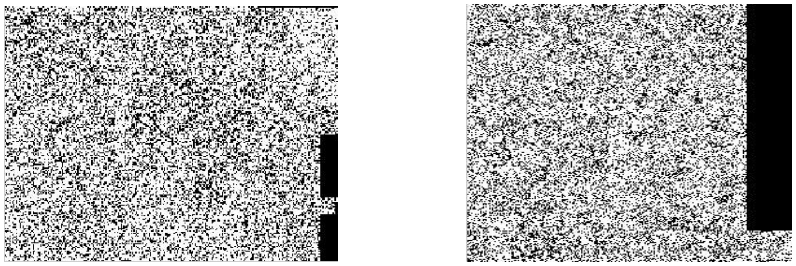
And its initial input could be

$$\begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix} \text{ or } \begin{bmatrix} 2 & 3 \\ 1 & 4 \end{bmatrix} \quad (4)$$

After iterations, the Hilbert space filling matrix with different orders can be obtained. For instance, the matrix with order 2 is given in (5)

$$\begin{bmatrix} 1 & 2 & 15 & 16 \\ 4 & 3 & 14 & 13 \\ 5 & 8 & 9 & 12 \\ 6 & 7 & 10 & 11 \end{bmatrix} \quad (5)$$

In (5), the value n of the item in the location (x,y) means that the n_{th} base in the genome sequence should filled on this location. For instance, the value of the item (2,2) is 3, the third bases in the genome sequence is 'A' (value 0). In the filling process, this base will be filled on the location (2,2) by using value 0 to instead the value 3 in this location. When each base is filled in the matrix, an image which represents the spacial distribution of the genome sequence can be obtained. In Fig. 2, two images of two genome sequences(NC_010162 and NC_013595) are given.



(a) NC_010162 (b)NC_013595

Fig. 2 Two images of two genome sequences mapping

This procedure can be describes as follows: Obtaining the corresponding Hilbert space filling matrix, put each base on its corresponding location of the matrix. For in

In context-based entropy coding technology, the conditional probability distributions $P(x_t \mid x_{t-1},\ldots,x_{t-K})$ are constructed to instead the true distribution $P(x_t \mid x_{t-1},\ldots,x_0)$

Details of the proposed algorithm

Firstly, we make the genome sequence changing to two binary sequences by using (3), then we use HSFC to map these two binary sequence into two bilevl images. For compressing these two images, two context models are constructed separately. Due to the binary source, more conditions may not always lead to higher model cost. In this work, 10 previous bases of the current base are used for conditioning, i.e. The order of our context models is 10. To select bases, the conditions template is given in Fig. 3

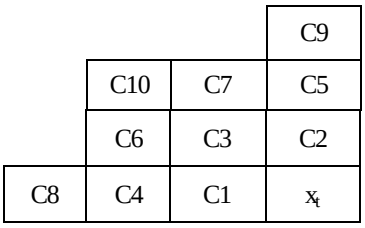


Fig. 3 The template of conditions used in our algorithm

By using this template, the context model can be constructed easily, which contains different correlated bases. In coding process, every coded base are used to join into its corresponding counts vector to estimated the conditional probability distribution for coding next base. However, when HSFC is employed, there are no-base area needed to be handled. Actually, the structure of the mapped image is known for both sender and receiver, which make this problem more easy. In coding process, if the current base is in the no-base area, its coding is ignored and on the receiver side, the value 'X' is filled into this corresponding location. Namely, we can determine the structure of the image aforehand, then we can just assign the codewords for each base without coding those no-base symbols.

Experiments and Results

Three genome sequences NC_013595, NC_013131 and NC_010162 are used as the source sequence to testify our compression algorithm. For comparison, the results from [5-7] are also listed in the table. All of these results are given in Table 1.

Table 1 The comparison of compression results

sequence	size	results (bit/base)			
		algorithm[5]	algorithm[6]	algorithm[7]	proposed
NC_013595	10341314	1.796	1.759	1.742	1.740
NC_013131	10468872	1.817	1.779	1.770	1.765
NC_010162	13033779	1.755	1.743	1.732	1.725

It is obviously that the proposed algorithm can perform better results than other previous algorithms. It comes from the deal of the genome sequence into 2-D images, which leads more correlation can be used for compression.

Conclusions

The genome compression algorithm based on the Hilbert space-filling curve is proposed in this paper. In order to utilize the correlations among the bases in the genome sequence, our algorithm firstly use the Hilbert space filling curve to map the genome sequence into a new 2-D image. Then the image obtained is encoded by the context weighting modeling technology. For context weighting, the values of weights are corresponding to the description length of the context model which is corresponding to their weights respectively. When the receiver obtain the mapping image, the reverse Hilbert space filling matrix is used to help the decoding procedure. The experiments results indicate that the final compression results by our algorithm are better than the results by other similar algorithms, although the useless area which will lead to the reduce of the compression efficiency is resulted in the mapping image by our algorithm.

References

[1]. T. Matsumoto, K. Sadakane, and H. Imai Biological sequence compression algorithms., GenomeInformatics, Vols.11, pp.43–52, 2000.

[2]. S.Deorowicz etc, Robust relative compression of genomes with random access, Bioinformatics, Vol.27, No.21, pp.2979–2986, 2011.

[3]. S. Deorowicz etc, Genome compression: a novel approach for large collections, Bioinformatics, Vol.29, no.20: pp.2572-2578, 2013.

[4]. A. J. Pinho, etc, DNA coding using finite-context models and arithmetic coding, proceeding of ICASSP-2009, Taipei, Tai-wan, Apr. 2009.

[5]. M. D. Cao, T. I. Dix, L. Allison, and C. Mears, A simple statistical algorithm for biological sequence compression, in Proc. of the Data Compression Conf, DCC-2007, Snowbird, Utah, 2007.

[6]. Armando J. Pinho etc, Bacteria DNA Sequence Compression using a mixture of finite-context models, IEEE Statistical Signal Processing Workshop, Portugal, pp.125-128, 2011.

[7]. X. Wu, G. Zhai, Adaptive Sequential Prediction of Multidimensional Signals with Applications to Lossless Image Coding, IEEE Trans. Image Processing, Vol. 20, NO. 1,pp.36-42, 2011.