Bi-Directional Context Modeling with Combinatorial Structuring for Genome Sequence Compression

Wenrui Dai* and Hongkai Xiong*

*Department of Electronic Engineering
Shanghai Jiao Tong University, Shanghai 200240, China
{daiwenrui,xionghongkai}@sjtu.edu.cn

This paper proposes a bi-directional context modeling (BCM) technique for reference-free DNA sequence compression, which constructs its contexts by combining arbitrary predicted symbols in two directions corresponding to approximate repeats and non-repeat regions. Thus, BCM can sequentially predict DNA sequences with weighted conditional probabilities that simultaneously exploit the correlations among matched approximate repeats and fit the variable-order statistics in non-repeat regions. Moreover, BCM eliminates the overhead of pointer information for specifying approximate repeats, as it is synchronized in both encoder and decoder.

To be concrete, each nucleotide x_t of sequence x_1^N is predicted with a weighted probability conditioned on its bi-directional contexts $\mathbf{s} = (s_1, s_2)$. Denote s_1 and s_2 the contexts for x_t extracted from buffered approximate repeats and non-repeat regions, respectively. Consequently, s_1 adopts combinatorial structuring of partially matched subsequences to represent the approximate repeats with insertion, deletion, and substitution. While s_2 is constructed by combining arbitrary predicted nucleotides in non-repeat regions. Given weights $\mathbf{w}_1^{(i)}$ and $\mathbf{w}_2^{(j)}$ for the two configurations $s_1^{(i)}$ and $s_2^{(j)}$ of s_1 and s_2, the weighted conditional probability $P_w(x_t|\mathbf{s})$ for x_t is obtained by

$$P_w(x_t|\mathbf{s}) = \gamma_1 \sum_i \mathbf{w}_1^{(i)} P_e(x_t|s_1^{(i)}) + \gamma_2 \sum_j \mathbf{w}_2^{(j)} P_e(x_t|s_2^{(j)}). \quad (1)$$

Eq. (1) implies that $P_w(x_t|\mathbf{s})$ is *de facto* obtained by weighting over all possible contexts from matched approximate repeats and predicted non-repeat regions. In practical coding, $\{\mathbf{w}_1^{(i)}\}$ and $\{\mathbf{w}_2^{(j)}\}$ are updated with gradient descent for each x_t to adaptively fit the statistics of x_1^N.

In theory, we show that upper bounds of excess model redundancy led by BCM vanish with the growth of sequence size. Experimental results show that BCM outperforms the state-of-the-art reference-free compressors FCM [1] and CTW+LZ [2].

References

[1] A. J. Pinho, A. Neves, C. Bastos, and P. Ferreira, "DNA coding using finite-context models and arithmetic coding," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, Apr. 2009, pp. 1693-1696.

[2] T. Matsumoto, K. Sadakane, and H. Imai, "Biological Sequence Compression Algorithms," *Genome Informatics*, vol. 11, pp. 43-52, Dec. 2000.

The work was supported in part by the NSFC under Grants 61425011, 61271218, and U1201255, and in part by the "Shu Guan" project under Grants 13SG13.

1068-0314/15 $31.00 © 2015 IEEE
DOI 10.1109/DCC.2015.67