View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Ghent University Academic Bibliography

biblio.ugent.be

The UGent Institutional Repository is the electronic archiving and dissemination platform for all UGent research publications. Ghent University has implemented a mandate stipulating that all academic publications of UGent researchers should be deposited and archived in this repository. Except for items where current copyright restrictions apply, these papers are available in Open Access.

This item is the archived peer-reviewed author-version of:

Leveraging CABAC for No-Reference Compression of Genomic Data with Random Access Support

Tom Paridaens, Jens Panneel, Wesley De Neve, Peter Lambert, and Rik Van de Walle

In: Data Compression Conference (DCC) Proceedings, 625, 2016.

To refer to or to cite this work, please use the citation to the published version:

Paridaens, T., Panneel, J., De Neve, W., Lambert, P., and Van de Walle, R. (2016). Leveraging CABAC for No-Reference Compression of Genomic Data with Random Access Support. *Data Compression Conference (DCC) Proceedings* 625.

Leveraging CABAC for no-reference compression of genomic data with random access support

Tom Paridaens*, Jens Panneel*, Wesley De Neve*†,
Peter Lambert*, and Rik Van de Walle*

*Data Science Lab
iMinds-Ghent University
Sint-Pietersnieuwstraat 41 B2
Ghent, 9000, Belgium
tom.paridaens@ugent.be

†Center for Biotech Data Science
GUGC-K
Songdomunhwa-ro 119, Yeonsu-gu
Incheon, 305-701, South Korea
wesley.deneve@ugent.be

In previous work, the authors developed a modular no-reference framework [1] that compresses FASTA files by applying a predict-and-residue method, as used in video coding [2]. In the first stage, the nucleotides are concatenated and split into blocks of a fixed size (typically the size of individual reads). In the second stage, the most effective coding or prediction tool is selected for each block.

We extended this framework with support for Context-Adaptive Binary Arithmetic Coding (CABAC), while at the same time preserving random access functionality and offering support for the full IUB/IUPAC nucleic acid codes alphabet. CABAC is applied on all syntax parameters and the residue. For each of the syntax parameters, we developed a technique for binarisation and context modelling. The addition of CABAC entropy coding provided a compression gain of between 34.45% and 70.41%. This resulted in a bit cost of between 0.124 bits/base (for test files with high coverage or many genomes of one type of species) and 1.096 bits/base (for test files with low coverage), while maintaining support for random access.

	7-zip Ultra	No CABAC	CABAC	
NA12878_S1	0.244	0.532	0.292	-45.11%
9827_2#49	1.135	1.672	1.096	-34.45%
HCC1954.mix1.n80t20	0.435	0.708	0.380	-46.33%
MiSeq_Ecoli_DH10B_110721_PF	0.135	0.282	0.134	-52.48%
K562_cytosol_LID8465_TopHat_v2	0.115	0.419	0.124	-70.41%

Table 1: Compression results for 7-zip and the proposed solution in bits/base.

References

[1] T. Paridaens, Y. Van Stappen, W. De Neve, P. Lambert, and R. Van de Walle, "Towards Block-based Compression of Genomic Data with Random Access Functionality," in *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on*, Dec 2014, pp. 1360–1363.

[2] M. Wien, *High Efficiency Video Coding*. Springer-Verlag Berlin Heidelberg, 2015.