Bioinformatics, 32(7), 2016, 1118–1119
doi: 10.1093/bioinformatics/btv720
Advance Access Publication Date: 31 December 2015
Letter to the Editor

OXFORD

Sequence analysis

Authors' response to 'Comment on: ERGC: An efficient Referential Genome Compression Algorithm'

Subrata Saha* and Sanguthevar Rajasekaran*

Department of Computer Science and Engineering, University of Connecticut, Storrs, CT, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on 2 December 2015; revised on 2 December 2015; accepted on 3 December 2015

Abstract

Contact: subrata.saha@engr.uconn.edu or rajasek@engr.uconn.edu

1 Response

- Please note that the datasets we have used in Saha *et al.* (2014) have been used in several prior works as benchmark datasets. Please see Pinho *et al.* (2012), Ochoa *et al.* (2014), and many others. These datasets were not chosen to skew the results. To ensure fairness it is a standard practice to use the same benchmark datasets to compare different algorithms. This is exactly what we have done.
- For datasets D1 and D5, Deorowicz *et al.* claim that the results reported in Saha *et al.* (2014) are incorrect. To ensure the correctness we have rerun our program for the aforementioned datasets and got exactly the same results we have reported in Saha *et al.* (2014). Thus the claim of incorrectness is inappropriate. We have uploaded HG18 and YH (please see Section 2) datasets so that anyone can verify the correctness of the reported results.
- Deorowicz *et al.* say: 'First, the cited paper incorrectly reported 'NA' for the GDC algorithm in these two cases.' Here Deorowicz *et al.* refer to datasets D1 and D2. The fact is that GDC was not able to compress the single chromosome of the D1 and D2 datasets within 2 h. We stopped GDC after 2 h. Here again the use of the word 'incorrectly' is not appropriate. We have uploaded the programs GDC by Deorowicz *et al.* (2011) and iDoComp by Ochoa *et al.* (2014) that we have used to compare with ERGC (please see Section 2).
- The use of KOREAN genomes as the references was not intentional. For instance we have used hg18 in the D1 dataset as the reference.
- Deorowicz *et al.* claim that 'The KOREAN genomes differ from the other ones in that they contain both lower and upper case letters.' It is not the case. For example in D4 and D5 both the

Table 1. Performance evaluation of different algorithms using compressed size metric

			ERGC		GDC	
Dataset	Target	Reference	Chromosome 10	Chromosome 20	Chromosome 10	Chromosome 20
A1	HG17	HG18	470 015	243 018	**444 803**	**225 404**
A2	HG18	HG17	525 850	243 141	**444 803**	**225 404**
A3	HG18	HG19	13 314 010	**243 326**	**1 152 786**	596 112
A4	HG19	HG18	**580 896**	**250 493**	1 152 786	596 112
A5	HG19	HG38	27 360 488	6 857 334	**1 553 333**	**620 751**
A6	HG38	HG19	22 797 278	8 702 005	**1 553 333**	**620 751**
A7	KO131	KO224	**250 900**	**97 951**	474 731	165 916
A8	KO224	KO131	**206 609**	**78 295**	474 731	165 916

Best values are shown in boldface.

references and the targets contain both upper case and lower case letters.

- Deorowicz *et al*. have rerun the programs by changing all the characters to upper case. The results are shown in Table 3. When it comes to (lossless) compression, we don't have the option of making such changes. If we do so, we lose information and such an algorithm will be a lossy compression algorithm!
- Sequencers employing the state-of-the-art sequencing technology produce genomic sequences that contain both lower and upper case letters. Cases of letters carry important information in the context of molecular biology. In the UCSC genome database, genomes HG17 to HG38 all contain both upper case and lower case letters. Moreover ERGC is not restricted to *A, C, G, T* and *N* characters. There are several other valid characters that are used in clones to indicate ambiguity about the identity of certain bases in the sequence. It is not uncommon to see these wobble codes at polymorphic positions in DNA sequences. ERGC can handle virtually every character in the genomic sequences. Furthermore, ERGC is an order of magnitude faster than GDC or iDoComp on an average.
- We have run our software tool on different datasets (please, see Table 1). In A1 and A2 datasets ERGC is comparable with GDC. In A3, A4, A7 and A8 datasets ERGC performs better than GDC. GDC outperformed ERGC heavily on A5 and A6 datasets. Please note that both HG38 and HG19 contain both upper case and lower case letters. Moreover, GDC automatically selects the best reference genome among the sequences.
- Note that we have downloaded GDC from: sun.aei.polsl.pl/REFRESH/gdc/downloads/0.3/gdc on 10/31/2014; iDoComp was downloaded from: www.stanford.edu/~iochoa/iDoComp.html on 11/01/2014.

2 Datasets

The datasets and programs we have used can be found in the following sites:

HG18:
drive.google.com/folderview?id=0B4boAFd04Xu1UjZyTWZneGdWWEk&usp=sharing
YH:
drive.google.com/folderview?id=0B4boAFd04Xu1bjBFRlQzc3VyWXc&usp=sharing
Programs:
drive.google.com/folderview?id=0B4boAFd04Xu1SlZBLUQtQ1dwbUE&usp=sharing

References

Pinho,A.J. *et al*. (2012) GReEn: a tool for efficient compression of genome resequencing data. *Nucleic Acids Res*., **40**, e27

Ochoa,I. *et al*. (2014) iDoComp: a compression scheme for assembled genomes. *Bioinformatics*, **31**, 626–633.

Saha,S. and Rajasekaran,S. (2015) ERGC: an efficient referential genome compression algorithm. *Bioinformatics*, **31**, 3468–3475.

Deorowicz,S. and Grabowski,S. (2011) Robust relative compression of genomes with random access. *Bioinformatics*, **27**, 2979–2986.