International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 12 6503 - 6506

Application of Reference - Based Lossless Genome Compression

Heba Afify
Department of Systems and Biomedical Engineering, MTI University,
Egypt, Cairo
hebaaffify@yahoo.com

Abstract- Genomic data technology has advanced by using many algorithms that not only facilitate a meaningful analysis of these data but also aid in efficient compression, storage, retrieval, updating, and transmission of huge volumes of the generated data. This has necessitated the development of novel bioinformatics approaches and generic compression tools. In recent years, many efforts have been madeto use a reference genome for storage that based on encode the differences between sequence and the reference genome. We used the difference compression to update the compressed set of similar sequences.In addition, we found that there is similarity degree between different organisms, so we used difference compression to compress data set from two different species. It used to determine which species can compress related to another species, and which reference is appropriate for data set.Results show that the entropy, which is an indicator of the compression efficiency, and a measure of relatedness, is much lower with variable reference that based on minimum entropy than that with the single fixed Cambridge reference sequence. It noted that execution time for encoding huge data set by using Cambridge reference less rather execution time for data set by using entropy to select reference.

Keywords *–Bioinformatics, Difference compression, Cambridge reference.*

*****

I. INTRODUCTION

The increasing volume of genomic data collected in recent years has prompted increasing demand for bioinformatics tools for genomic and proteomic data analysis and compression [1]. Web based bioinformatics application platforms have become one of the popular tools for genomic data analysis among the bioscience community [2]. However, these application platforms utilize different stand-alone bioinformatics applications and they use different data sources in different formats.

Since 1995s onward, the term differentialcompression used to refer to data differencing [3]. Generally, differencing algorithms achieve compression by finding common sequences between two versions of the same data that encoded by using a copy reference. A differencing algorithm is an algorithm that finds and outputs the changes made between two versions of the same data by locating common sequences to be copied and unique sequences to be addedexplicitly.

The differential compression of DNA sequences [4] remains a challenging problem, with profound implications in biology and with important technological impact when the use of genomic data will become a daily practice in health and medicine. As such, it will certainly be investigated further due to several reasons: (1) benefits when storing and updating the genome files, (2) possibilities for comparison of entire genomes, and (3) discovering statistically significant relationships among various organisms. Lossless differential compression for DNA [5] is the efficient compression of entire databases of sequences. This is particularly true if the sequences to be analyzed are very large and do not change. DNAzip [5] was first algorithm introducing the important idea of only storing differences to a reference sequence, but in this case for storing an entire, assembled genome as a series of difference. Brandon et al [6] found that selecting thereference sequence is important for having an effective compression of dataset. On the other hand, C. Wang [4] implemented a generic tool, GRS, for de novo compression of genome sequencingdata, which does not need the reference SNPs map. DNAEncodeWG [7] also presented how to compress DNA sequence data using the whole genome sequence of an organism to identify differences between DNA sequences if a repository of the whole genome sequence of the organism is accessible through the Web. Kozanitis [8] focused on fragment compression as opposed to sequence compression by using SLIMGENE. H. Afify [9] presented another algorithm in which for each pair of similar sequences, a third sequence can be generated; representing the difference between them, and the entropy of the generated difference sequence can be estimated. Difference sequence can help in building phylogenic tree, while the entropy can help in selecting appropriate compression reference for short dataset. H. Afify [10] presented a differential compression algorithm that based on production of difference sequences according to op-code table in order to optimize the compression of homologous sequences in dataset. Therefore, the stored data are composed of reference sequence, the set of differences, and differences locations, instead of storing each sequence individually. This algorithm does not require a priori knowledge about the statistics of the sequence set. It was applied to three different datasets of genomic sequences, it achieved up to 195-fold compression rate corresponding to 99.4% space saving.

In this study, we describe two applications that based on the differential compression of DNA sequence dataset. Frist, updating the differentially compressed set of similar sequences. We used this application, when a new similar sequence is available. Second, compression of data set from two different species. We used this application, when DNA similarity degree between different organisms is available e.g. Human and mouse [11]. It based on the evolution history between various organisms. This application may be used when we need to compress unknown data set from different organisms.To evaluate the suggested applications, we used two methods to compress the data set: by using Cambridge signal sequence as the reference of data set, on the other hand by selecting reference based on minimum entropy [9].

II. MATERIAL AND METHOD

A. *Data Extraction*

The data used in proposed work consists of huge data consists of two different data sets of genomic sequences including 3615 human mitochondrial genomic sequences, and 100 mouse sequences MusMusculusDmesticus. Mitochondrial data set takes 56MB size in GenBank, and is downloadable from the GenBank database, HapMap web site and the MITOMAO database [12]. Mouse datasets takes 106KB in GenBank, and is downloaded from Mouse Genome Database [13].References were selected as follows: Cambridge sequence NC_012920 sequence for human mitochondrial data set which is about 16569bp long, HM17663 for virus data set which is about 1714bp long and AJ843867 for mouse data set which is about 1009bp long.

We divided huge data set into N separate small data sets randomly which combined to form huge data set for updating. Therefore, data set divided into two types of data; learning data and testing data. Testing data sequences presented to our algorithm as new sequences that require adding to the learning data set.

B.Update the Differentially Compressed set of similar sequences

We choose some sequences as testing data and other sequences as learning data for both types of data set. We used this application, when a new similar sequence is available. We used two methods to update the compressed data set as the following.

1) First method: by using Cambridge sequence (single) as the reference of learning data set, then we calculated difference sequences between reference sequence and each sequence in the testing data set.
2) Second method: by selecting variable reference based on minimum entropy [9].

Then, we compared average entropy and number of differences for each method to determine which method is more efficient to update the compressed data set.It more difficult to use this method for updating the huge data because it takes more time; in turn this new sequence need to determine which sequence from testing data is valid to be reference for this new sequence. Note that the reference that is valid for original data is not necessary to be valid for any external sequence that adds to the original data in the future. In summary, changing of references will occur for any new addition of sequences.

C.Compression of data set from two different species

We found that there is DNA similarity degree between different organisms e.g. Human and mouse. The overall distribution of local (G+C) content is significantly similar between the mouse and human genomes as in Fig.1. Mouse has a higher mean (G+C) content than human (42% compared with 41%), but human has a larger fraction of windows with either high or low (G+C) content. The distribution was determined using the unmasked genomes in 20-kb non-overlapping windows, with the fraction of windows (y axis) in each percentage bin (x axis) plotted for both human and mouse [14].

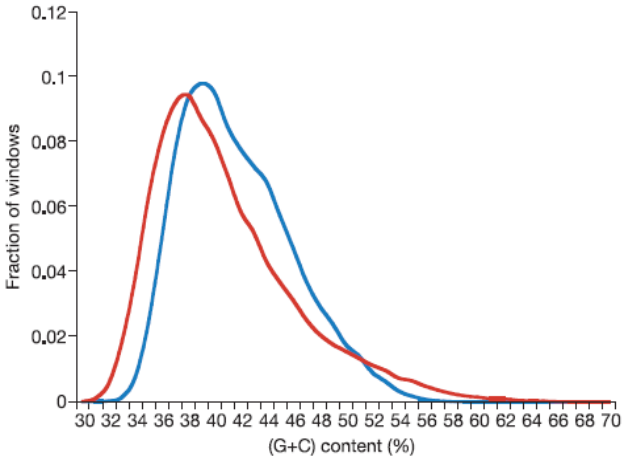

Figure 1. Distribution of (G+C) content in the mouse (blue) and human (red) genomes

The evolutionary tree shows the position of mammals whosegenomes have been sequenced and analyzed or are being sequenced.Note also the fast evolutionary rate of rats and mice compared with humans [15]. Moreover, as scientists begin to understand the common elements shared among species, it may also become possible to approach the even harder challenge of difference compression to identify and understand the differences that make each species unique.

In this part, the data set consists of the collection between two species to determine which species can compress related to another species, and which reference is appropriate for data set. We used this application, when differentially compressing one set of similar sequences in terms of another set of similar sequences. It depends on study the evolution history between various organisms. We used two methods to compress the data set as the following.

1) First method: by using Cambridge signal sequence as the reference of data set, then we calculated difference sequences between reference sequence and each sequence.
2) Second method: by selecting reference based on minimum entropy [9].

Each species has two references based on two methods. So, we have applied four references to data set for comparing the compression ratio of each method.

III. RESULTS AND DISCUSSION

Compression algorithms were run against data sets in order to establish the viability of differential compression for operating system applications such as updating new data, file system backup and restore, selection of best reference for compression, and compress unknown data set from different organisms. We discuss the results of three parts as the following:

A. Update the compressed data set

Updating the data set with a group of sequences, similar to the differentially compressed sequences,is considered. In

this application, the original differentially encoded data set was build using some sequences of the above set of small genomic set, while we randomly selected other sequences as testing data, were presented as new sequences that require adding to the original data set for updating the set. Results show that the entropy is much lower with variable reference that based on minimum entropy than that with the single fixed Cambridge reference sequence. Results of average entropy and number of differences between the compressed and reference sequences are shown in Table 1.

Table I: COMPARISION BETWWEN TWO METHODS FOR UPDATING NEW DATA SET

Methods	Average Entropy	No. of Difference	Compression ratio
By selecting reference	**0.0039**	**236**	**0.0031**
By using Cambridge sequence	0.0057	276	0.0039

This application is used to update the set with new similar sequences whose differences do not obey the same statistical model as that of differences of the original set. When we update the compressed data set, we need to reorder all sequences based on reference selection.For updating, we found that method based on entropy theory to select reference is more efficient than method based on Cambridge reference. On the other hand, this application takes a long time for huge data by using the method based on entropy theory rather than method based on Cambridge reference.By using entropy theory for reference selection, we search for good reference that closes to each sequence. So, reference that suitable for compressed data set is not always appropriate for updating the compressed data set.For the same data set, there are variable references for each new updating.

B. Compression data set from two different species

Another application, the data set used consists of collection between Human &Mouse. We used two methods to compress the data set by changing the reference sequence as in Table 2. It means that compression of sequence related to another sequence from different class is available, on condition that there is similarity between different classes or sequence close to another sequence according to evolution tree.

Table II.COMPARISION BETWWEN TWO METHODS FOR DATA SET

References	Differences size	Locations size	Compression ratio	Space saving
'NC_012920'	1.90 KB	12.9KB	0.086	91.3%
'AJ843867'	40.9 KB	12.5KB	0.311	68.8%
'DQ779930'	1.87KB	12.7KB	**0.084**	**91.6%**
'MMU47467'	40.4KB	12.3KB	0.307	69.3%

We found that size of data set before compression = 171.9KB (1408808 bits). By using 'NC_012920' (Cambridge reference of human data) or 'DQ779930'(selected reference of human data based on entropy) as reference for data set, we found that compression of differences by ZLIB Deflator algorithm[16] is better than Huffman [17]. We also found that compression of location distances is better than locations only by ZLIB Deflator algorithm [16].By using 'AJ843867' (Cambridge reference of mouse data) or 'MMU47467' (selected reference of mouse data based on entropy) as reference for data set, we found that compression of differences by Huffman is better than ZLIB Deflator algorithm. We also found that compression of location distances is better than locations only by ZLIB Deflator algorithm.

By using 'DQ779930' as reference for data set, we found that this reference achieve the best of compression ratio than the others. Therefore, space saving of data set by using this reference is 91.6%. It means that 'DQ779930' as reference is suitable for data compression that consists of collection between human and mouse rather than the others.

We noted that method of reference selection for human data set by entropy theory collection between human and mouse, but mouse sequence is used only to compress data set of the same class. This meaning confirms the idea of evolution history.

From above results, we found that compression of sequence related to another sequence from the same class is more efficient rather than this application because of high similarity between sequences from the same class. However, this application may be used when we need to compress unknown data set from different organisms

C. Execution Time

The computational complexity of the two application is related to computation time to obtain the solution for genomic updating and compression unknown dataset.It noted that execution time for encoding huge data set by using Cambridge reference less rather execution time for data set by using entropy to select reference. Furthermore, it allows the users to increase the accuracy of the compression of huge data sets by allowing selection of reference based on entropy with a long time. Given the Cambridge reference sequence, compression can be performed with virtually no noticeable delay.

Fig.2 shows the required time to compress huge data set by using three approaches. First, ZLIB deflator algorithm was used to store the differences and locations. Second, Huffman coding was applied to the same data set. Third, Arithmetic coding was used to compress the same data set. As expected, Huffman coding performs well for small size of sequences. With increasing size, the ZLIB deflator algorithm performs better than either the Arithmetic coding or Huffman coding.

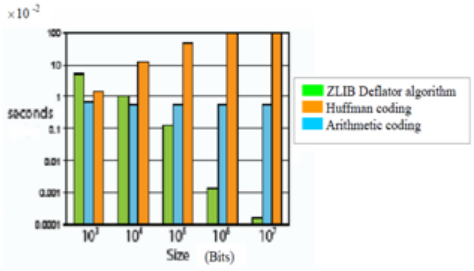

Figure 2.Time comparison of three coding algorithm

IV. CONCLUSIONS

Reference based compression relying on entropy estimation can be successfully applied to phylogenetic research, as shown in [18]. The proposed challenge is to develop a differential compression scheme for huge data by using two applications. First, updating new data set with a group of sequences. Second, compression data set from two different species. Single Cambridge sequence is external reference that outside the original data set. However, it achieves the rapid compression rather than variable reference based on entropy estimation that is inside the original data set. Variable reference depend on reordering the huge sequences for updating new data. Therefore, variable reference is more efficient for compression unknown data set from different species to compress one set of similar sequences in terms of another set of similar sequences e.g. mouse in terms of human.

Finally, differential compression is still largely an art of science and to gain proficiency in an art we need to get a feel for the process. Selecting the reference sequence is not depends on fixed or static rules; each data has a special treatment for compression.

REFERENCES

[1] B.A.Gata,"Database Similarity Searching Using BLAST and FastA", AustralasianBiotechnology, vol. 5, pp.282-290, 1995.
[2] J. Cohen, "Computer Science and Bioinformatics," Communications of the ACM,vol. 48, no.3, 2005.
[3] D.G. Korn, K.P. Vo, Vdelta, "Differencing and Compression", Practical ReusableUnix Software, Editor B. Krishnamurthy, John Wiley & Sons, Inc., 1995.
[4] C. Wang and D. Zhang, "A Novel Compression Tool for Efficient Storage of GenomeResequencing data," *Nucleic Acids Research*, doi:10.1093/nar/gkr009, January 2011.
[5] S. Christley, Y.Lu, C. Li, and X. Xie, "Human Genomes as Email Attachments," Bioinformatics,vol. 25, no. 2, pp. 274-275, 2009.
[6] C.Brandon, D. C. Wallace and P. Baldi, "Data Structures and Compression algorithms forGenomic Sequence data," *Bioinformatics*, vol. 25, no. 14, pp. 1731-1738, 2009.
[7] H. Do Kim, Ju-Han Kim, "DNA Data Compression Based on the Whole Genome Sequence,"*Journal of Convergence Information Technology*, vol. 4, no. 3, 2009.
[8] C. Kozanitis, C.Saunders, S.Kruglyak, V.Bafna, and G.Varghese, "Compressing GenomicSequence Fragments Using SLIMGENE," *RECCOMB*, 2010.
[9] H. Afify, M. Islam, M. Abdel Wahed, and Y. M. Kadah, "Genomic Sequences DifferentialCompression Model," *27thNational Radio Science Conference, NRSC'2010*, Menouf, Egypt,March 16-18, 2010.
[10] H. Afify, M. Islam and M.AbdelWahed," DNA Lossless Differential Compression Algorithms based on similarity of genomic sequences database", International Journal of Computer Science & Information Technology (IJCSIT) Vol 3, No 4, August 2011
[11] Cheng Y, Ma Z, Kim BH, Wu W, Cayting P, Boyle AP, Sundaram V, Xing X, Dogan N, Li J," Principles of regulatory information conservation between mouse and human",*Nature*. 20;515(7527):371-5. doi: 10.1038/nature13985. PMID: 25409826, Nov 2014.
[12] http://www.fludb.org/brc/home.do?decorator=influenza.
[13] http://www.informatics.jax.org
[14] Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P et al." Initial sequencing and comparative analysis of the mouse genome, Nature 420:520-562, jun.2003.
[15] Kerstin Lindblad-Toh, "Genome sequencing", Nature 428, 475-476, April 2004.
[16] Gailly, Jean-loup, Adler, Mark, "zlib Applications", 18-04-2002.
[17] Khalid Sayood, "Introduction to Data Compression, 3rd ed.," University of Nebraska, 2006.
[18] J. Hagenauer, Z. Dawy, B. Goebel, P. Hanus, and J. C. Mueller, "Genomic analysis using methods from information theory," in Proc. of the ITW 2004, pp. 55–59, Oct. 2004.