applied sciences

MDPI

Article

BAQALC: Blockchain Applied Lossless Efficient Transmission of DNA Sequencing Data for Next Generation Medical Informatics

Seo-Joon Lee [1], Gyoun-Yon Cho [2], Fumiaki Ikeno [3] and Tae-Ro Lee [4,*]

1 ALFO Research Laboratory, Curaum Inc., Seoul 03997, Korea; richardlsj@korea.ac.kr
2 Research Institute of Health Science, Korea University, Seoul 02841, Korea; gycho@korea.ac.kr
3 Division of Cardiovascular Medicine, School of Medicine, Stanford University, 300 Pasteur Drive Stanford, California, CA 94305, USA; fikeno@stanford.edu
4 School of Health Policy & Management, Korea University, Seoul 02841, Korea
* Correspondence: trlee@korea.ac.kr

Received: 22 July 2018; Accepted: 23 August 2018; Published: 27 August 2018

Abstract: Due to the development of high-throughput DNA sequencing technology, genome-sequencing costs have been significantly reduced, which has led to a number of revolutionary advances in the genetics industry. However, the problem is that compared to the decrease in time and cost needed for DNA sequencing, the management of such large volumes of data is still an issue. Therefore, this research proposes Blockchain Applied FASTQ and FASTA Lossless Compression (BAQALC), a lossless compression algorithm that allows for the efficient transmission and storage of the immense amounts of DNA sequence data that are being generated by Next Generation Sequencing (NGS). Also, security and reliability issues exist in public sequence databases. For methods, compression ratio comparisons were determined for genetic biomarkers corresponding to the five diseases with the highest mortality rates according to the World Health Organization. The results showed an average compression ratio of approximately 12 for all the genetic datasets used. BAQALC performed especially well for lung cancer genetic markers, with a compression ratio of 17.02. BAQALC performed not only comparatively higher than widely used compression algorithms, but also higher than algorithms described in previously published research. The proposed solution is envisioned to contribute to providing an efficient and secure transmission and storage platform for next-generation medical informatics based on smart devices for both researchers and healthcare users.

Keywords: blockchain; DNA sequence; lossless compression; medical informatics

1. Introduction

Due to the development of high throughput DNA sequencing technology, genome sequencing costs have been significantly reduced, which has led to a number of revolutionary advances in the genetics industry [1]. Next-generation sequencing (NGS) allows significant amounts of DNA to be sequenced in parallel, and minimizes the need for comparatively inefficient fragment-cloning methods that are usually used in Sanger sequencing technologies [2].

However, the management of such large volumes of data is still an issue, and has thus been of interest to researchers [3,4]. Management refers to the transmission and storage of the sequence data. For example, it is still very common in the field of biomedical research to wait hours or sometimes days for DNA data transmission requests to be completed. Also, security and reliability issues remain obstacles in public sequence databases [5], which is even more discouraging to researchers.

This is also expected to be a problem for next-generation medical informatics, such as personal health record (PHR) systems [6], where healthcare consumers own their entire health data, and in

m-Health systems, where health data is freely transmitted between servers and smart devices [7]. DNA sequence data is not an exception for future consideration. In fact, it has already been emphasized by prior research [8–10] that protective approaches are required to defend against genome data attacks.

Data compression is considered a conventional solution that is often used to attempt to reduce dependency on storage, and at the same time, to reduce the resources needed for DNA transmission. Efficient sequencing data compression methods are in high demand, not only by researchers [11–13], but also by industry, considering the fact that Aspera (Everyville, California CA, USA), a high-performance file transfer (including DNA) software company, was acquired by IBM (Armonk, New York NY, USA) in December 2013. This is because data compression not only reduces transmission time, but also reduces the need for disk space.

Therefore, this research proposes Blockchain Applied FASTQ and FASTA Lossless Compression (BAQALC), a lossless compression algorithm that allows for efficient transmission and storage of the immense amount of DNA sequence data that are being generated by NGS within a reliable blockchain network [14]. The results showed comparatively high-performance, not only compared to widely used compression algorithms, but also compared to algorithms described in prior research. The proposed solution is envisioned to contribute to providing an efficient transmission and storage platform for next-generation medical informatics based on smart devices [15] such as PHR, telemedicine [16], e-Health [17], or m-Health.

2. Background

2.1. DNA Data Composition

Like many other viral sequences that rely on synthetic biology, major outbreaks, exobiology, and so on, human biological information is stored in DNA. This information is theoretically stored as a code made up of four chemical bases: Adenine (A), Cytosine (C), Guanine (G), and Thymine (T) as shown in Figure 1a [18]. It is stored in such a way that the order or sequence determines the layout and maintenance of organisms. A-T and C-G are pairs, and they are attached to the vertical sidepieces of a spiral ladder of a sugar phosphate backbone.

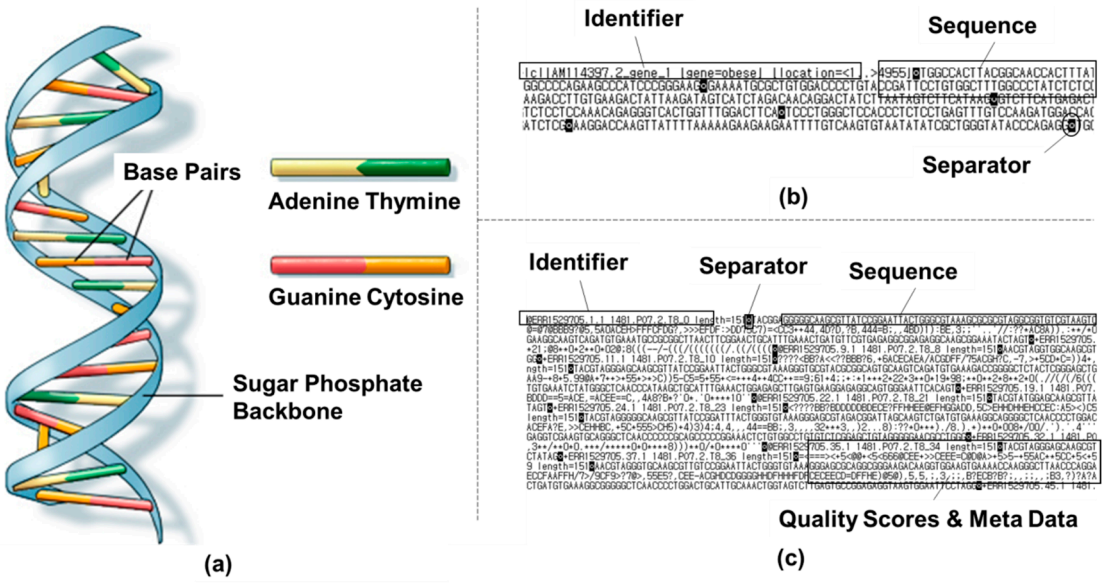

Figure 1. (**a**) DNA Base Pairs Attached to Sugar Phosphate Backbone and DNA Data File Formats of (**b**) FASTA and (**c**) FASTQ.

There are several types of DNA data file format, with "FASTA" and "FASTQ" files being most representative. The FASTA file format, which is shown in Figure 1b, is a relatively simple format

compared to FASTQ because it consists only of identifiers, sequences, and separators. Identifiers are names of a certain DNA sequence, such as identification codes of leptin, insulin, brain natriuretic peptide etc. Separators simply separate DNA sequences, and sequences are A, C, G, and T bases. In the FASTA format, DNA sequences account for a substantial portion of the file, making it easy to analyze or process. This will be more thoroughly discussed in Section 3.1.

On the other hand, FASTQ files not only contain identifiers, sequences, and separators, but also other diverse data such as quality scores and metadata. In this case, quality scores usually account for more than half of the file [19], and it is no doubt noisier than the DNA sequence alone. The complexity of FASTQ [20] makes it harder to analyze or process; this will also be more thoroughly discussed in Section 3.2.

2.2. Data Compression

Digital vital signal compression has been well established by our research team, from very redundant sequence signals such as Electrocardiography (ECG) [21] to much less diverse (non-redundant) sequence signals such as Electromyography EMG [22]. Redundant data such as ECG has been the primary target for signal processing researchers because of its high redundancy [23,24]. On the other hand, EMG data compression has been particularly challenging due to its extreme irregularity. Lack of redundancy is a set-back for dictionary-based approaches [25] founded on algorithms such as Lempel-Ziv (LZ) [26,27], or Huffman and its variants. Accordingly, in the proposed literature's case of handling DNA, our hypothesis is that DNA would also cause some difficulties due to irregular quality scores and metadata.

Although there have been lossy compression methods for the efficient compression of DNA [25,28], our research adhered to lossless methods. However, our logic is that even if quality scores and other metadata are noisy, lossless approaches should be adopted under the condition that the reconstructed data is not distorted. Especially, in terms of healthcare, handling important vital signs of human beings should be approached with caution; thus, a lossless approach is deemed optimal.

2.3. Prior Research on DNA Compression

Most biomedical research has focused on the investigation of biological DNA markers that affect metabolic health [29,30]. However, there are some, but not many research articles that are classified as biomedical research but that concentrate on the processing of biomedical signals such as DNA. Relatively recent peer-reviewed research on DNA compression is summarized in Table 1.

Table 1. Summary of Related Research.

Solution	Contents
LW-FQZip 2 [31]	Parallelized reference-based compression of FASTQ files. Tool for archival or space-sensitive (quality scores, metadata, and nucleotide bases) applications of high-throughput DNA sequence data.
QUIP [11]	Lossless compression algorithm. Adopted a reference-based compression, and known as the first assembly-based compressor that uses a novel de novo assembly algorithm.
DSRC 2 [28]	In this algorithm, a single thread reads the input FASTQ file in blocks (typically of tens of MBs size) and puts them into an input queue. Several threads perform the compression of the blocks, storing the results in an output queue. Finally, a single thread writes the compressed blocks in a single file.
CRAM [32]	Reference-based compression method. Targets well-studied genomes. Aligns new sequences to a reference genome and then encodes the differences between the new sequence and the reference genome for storage.
LFQC [33]	Lossless non-reference based FASTQ compression algorithm by generating identifier fields systematically.
SCALCE [25]	Boosting scheme based on locally consistent parsing technique, which reorganizes the DNA reads in a way that results in a higher compression speed and compression rate, independent of the compression algorithm in use and without using a reference genome.
UHT [34]	DNA compression based on using Huffman coding. Unbalanced Huffman encoding/Tree, forcing the Huffman tree to be unbalanced to be better than the standard Huffman.
UHTL [34]	DNA compression based on using Huffman coding. Developed version of UHT that prioritizes encoding the k-mers that contain the least frequent base.
MUHTL [34]	DNA compression based on using Huffman coding. Developed version of UHTL that allows more k-mers to be encoded to apply multiple Standard Huffman Tree (SHT)/UHT/UHTL coding.
FQZCOMP [19]	Proposed Fqzcomp and Fastqz that both accept FASTQ files as input, with the latter also taking an optional genome sequence to perform reference-based compression. Additionally proposes Samcomp, which also performs reference-based compression but requires previously aligned data in the SAM format instead.

Based on the summary of the selected related research, in depth compression efficiency comparisons will be discussed in the evaluation Section 4.2. In addition, some important, widely-used algorithms (other than peer-reviewed research) were also selected, including SHT, Lempel-Ziv Welch (LZW), gzip [35], and bzip 2. The algorithm gzip is particularly widely used in DNA compression currently. Other FASTA specialized compression algorithms were introduced by Pinho et al. [36], and Mohammed et al [37], which also provided some decent compression results.

2.4. Blockchain Technology

Among the many values of blockchain technology, trust-enforced security can be considered as the main one. Blockchain's close-to-infinite chained block architecture makes the algorithm's security [14] difficult to penetrate by attackers under sufficient user-supported circumstances. Also, distributed transaction ledgers among network users provide highly reliable data transactions which are highly resistant to fraud or fabrication. Prior research in data transmission networks has already made some preliminary attempts to apply blockchain technology for the transparency of public data verification [38], which has high implications in public DNA sequence data, where some sequences are unreliably modified or shared.

3. System Description

3.1. Overall Architecture

The overall system architecture in which the proposed BAQALC solution should be embedded is depicted in Figure 2. The overall architecture includes a basic blockchain-based scheme where biomedical researchers transmit/request DNA data which is ensured by distributed ledgers. Also in this scheme, healthcare users may request to receive fast DNA analysis results or data in future medical informatics systems, which is an innovative field mentioned in prior research [39]. Note that the BAQALC solution is developed in C to be compatible with a wide range of platforms; this will also be discussed in Section 4.1.

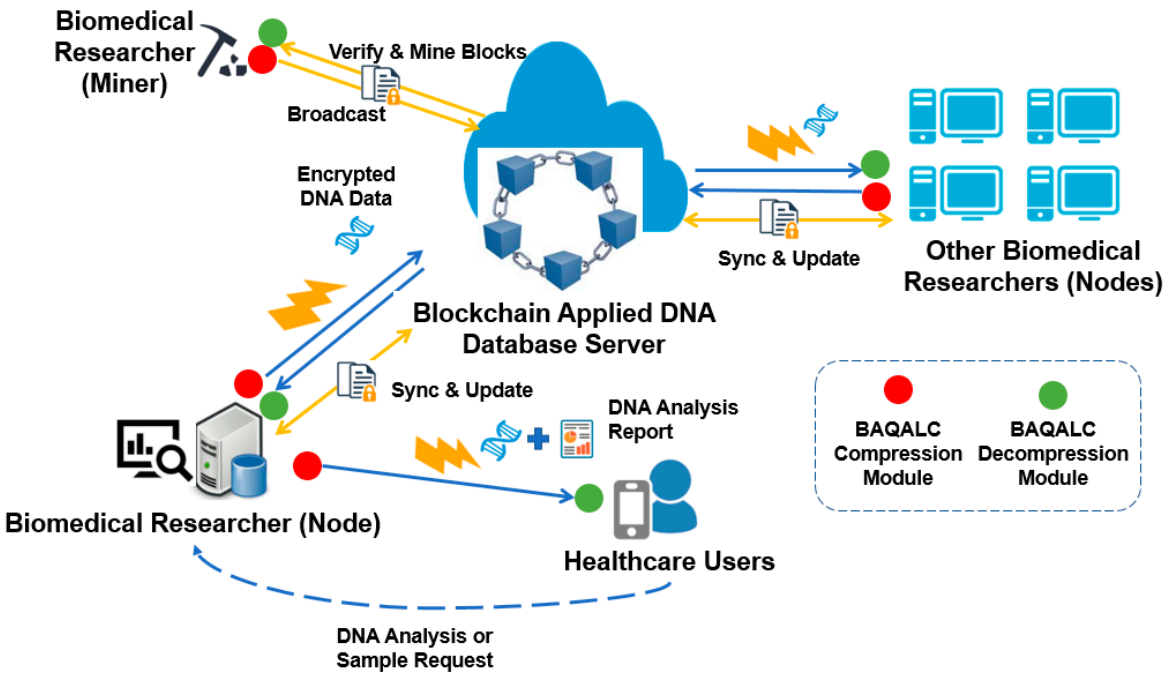

Figure 2. Overall System Architecture of BAQALC Modules.

When a biomedical researcher (node) discovers a nucleotide/peptide sequences, the researcher may transmit DNA data using the BAQALC compression module, which is depicted in a red circle in Figure 2. Note that the BAQALC decompression module is depicted in green circles. The transmitted compressed DNA data is sent to a public DNA database server, for example the National Center for Biotechnology Information's (NCBI), which is the most widely used DNA database world-wide. This creates a transaction.

Then, another biomedical researcher (miner) verifies whether the transaction is valid. As a miner, the researcher decodes the transaction sequence for verification. Verification includes filtering basic anomaly transactions such as uploading by researchers with unreliable histories, sequences that are too short or long compared to reference sequences, chimera detection, and many other errors. If the transaction is valid, the miner broadcasts the updated block chain into the blockchain network. Broadcasted updates are synced to all biomedical researcher nodes. This mechanism prevents any fraud or maliciously-fabricated updates by the nodes, creating a reliable database overall.

The DNA database is stored and kept in the database in an encrypted (compressed) block chained form. Other biomedical researchers may access the database and request the DNA data. Note that simple requesting only calls the blockchain data by reference, so no transaction is triggered. Once the request is accepted, DNA data is accessed, and when it reaches the requesting biomedical researchers' servers, the data is decoded using a BAQALC decompression module. If other biomedical researchers wish to upload DNA discoveries of their own, they may also transmit DNA data using the BAQALC compression module, which will trigger more transactions for verification by miners.

In addition to researchers, future healthcare users can subscribe to their own DNA data for medical informatics systems. Users can request DNA analysis results or even raw data of their health samples from medical researchers. Researchers analyze samples, and report results are transmitted to users. If the user has requested their own raw DNA data, the data is simply transmitted via BAQALC. After being decompressed by BAQALC, healthcare users can store and view their DNA through their mobile electronic devices.

3.2. Proposed Solution: BAQALC Algorithm

BAQALC, the proposed solution for DNA data compression, is explained in this section. The BAQALC solution comprises a compression module and decompression module. The specific flow of each module is shown in Figure 3.

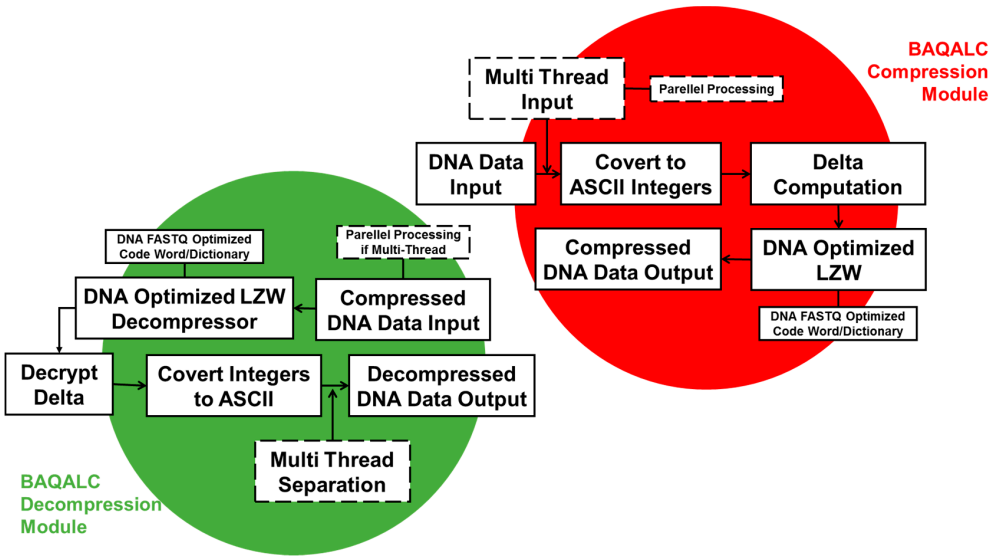

Figure 3. BAQALC Compression and Decompression Modules.

In the BAQALC compression module, DNA data is initially input. If it is a multi-thread input, the algorithm is subsequently processed in parallel. As discussed in Section 2.1, DNA FASTQ files are in ASCII characters, which are converted to the according integers. That is, alphabets or characters are converted. Integers are then delta computed, before finally being put into the DNA-optimized LZW algorithm. Here, the delta computation is the difference between the previous and the next integer. As a result, only the differences between the integer samples remain, which contributes to the reduced integer size (except for some exceptions, if the differences are too distant).

This final step is based on DNA FASTQ-optimized code and a dictionary that has been developed by the authors. In this step we have also developed a bit allocation scheme according to dictionary size. Although many of the specifics cannot be revealed, as an example, one of the randomly selected samples that we have used in the evaluation section could be processed with a dictionary size of 46. This can be bit-allocated within 8 bits into substitute integers from -32 to 31, in which this coverage helped predict the level of redundancy. After this final step, the compressed DNA data output file is constructed.

The BAQALC decompression module follows the inverse steps of the compression process. After the compressed DNA data is input, it is decoded by the DNA-optimized LZW decompressor (parallel processing if multi-thread). Then, delta is decoded and all converted integers are converted back to their original ASCII characters. If this occurs in a multi-thread, then channels are separated, otherwise not. Lastly, the decompressed DNA data is output losslessly.

4. Results

4.1. Materials and Methods

The FASTA format is a text-based format that represents nucleotide or peptide sequences. Nucleotides or amino acids are represented using single-letter codes. The FASTQ format is not limited to nucleotide and peptide sequences, but also includes other information such as metadata or quality scores. In this study, we use formats collected from the NCBI Sequence Read Archive (SRA) [40]. Pearson's correlation method was used for statistical analysis.

Selected DNA samples in NCBI SRA were five of each from five groups of insulin (SRX2528043, SRX2528042, SRX2528041, SRX2528040, and SRX2528039), leptin (ERX1600597, ERX1600596, ERX1600595, ERX1600594, and ERX1600593), Epstein-Barr virus (SRX2776811, SRX2776810, SRX2776809, SRX2776808, and SRX2776807), cystic fibrosis (SRX2935589, SRX2935588, SRX2935587, SRX2935586, and SRX2935585), and lung cancer (SRX2388195, SRX2388194, SRX2388193, SRX2388192, and SRX2388191). These DNA datasets were selected because they are genetic biomarkers for the top five causes of death as shown in Figure 4, provided by the World Health Organization (WHO) [41].

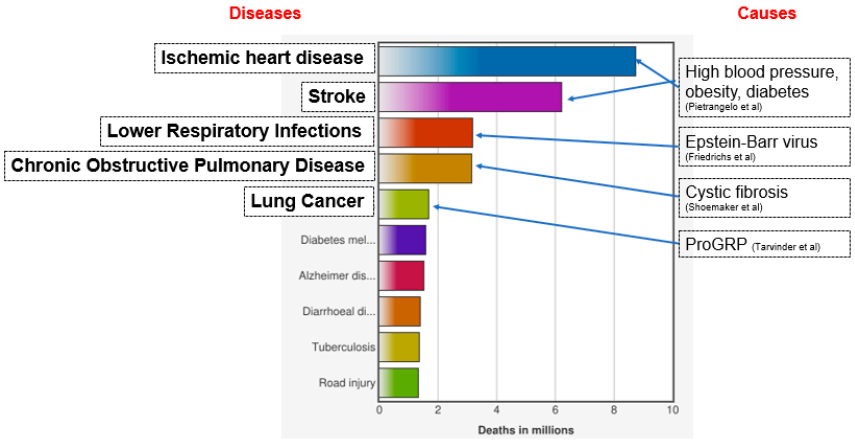

Figure 4. Top five Causes of Death (Highlighted) and Their Related DNA Markers.

Insulin and leptin-related problems are known causes for ischemic heart disease and stroke [42], whereas Epstein-Barr virus is a known marker for lower respiratory infections [43]. In addition, cystic fibrosis is a marker for chronic obstructive pulmonary disease [44], and lung cancer has been proven to have a genetic cause [45].

4.2. Comparison between Datasets

The Compression ratio (CR) was calculated using the following Equation (1), where US is the uncompressed size and CS is the compressed size. All results were rounded to the second decimal place. Average CRs were used to statistically explain the overall compression performance of the solution, and standard deviation (SD) CRs were used to explain the performance stability of the solution. Note that the lower the SD value, the higher the stability.

$$\mathrm{CR} = \frac{\mathrm{US}}{\mathrm{CS}} \tag{1}$$

Six DNA FASTQ samples were used for insulin, leptin, Epstein-Barr virus, cystic fibrosis, and lung cancer samples. In summary, a total of 25 FASTQ DNA samples from 5 different disease classifications were used. The results of the proposed algorithm's performance (average and standard deviation) for the 5 database groups are shown in Table 2.

Table 2. CR Results of BAQALC for Five Disease Datasets.

Database	CR	Used Data	CR Average ± SD
Insulin	10.98	SRX2528043	11.17 ± 0.11
	11.24	SRX2528042	
	11.22	SRX2528041	
	11.25	SRX2528040	
	11.18	SRX2528039	
Epstein-Barr	14.7	SRX2776811	13.65 ± 0.7
	13.79	SRX2776810	
	13.63	SRX2776809	
	13.36	SRX2776808	
	12.78	SRX2776807	
Lung	17.01	SRX2388195	17.01 ± 0.06
	16.98	SRX2388194	
	17.08	SRX2388193	
	17.06	SRX2388192	
	16.94	SRX2388191	
Leptin	10	ERX1600597	10.61 ± 0.56
	10.25	ERX1600596	
	10.44	ERX1600595	
	11	ERX1600594	
	11.38	ERX1600593	
Cystic Fibrosis	10.06	SRX2935589	10.09 ± 0.15
	10.13	SRX2935588	
	9.85	SRX2935587	
	10.17	SRX2935586	
	10.26	SRX2935585	
Total CR	All Data		12.51 ± 2.64

BAQALC CR performance results for the selected 25 samples showed an average and standard deviation of 11.17 ± 0.11, 10.61 ± 0.56, 13.65 ± 0.7, 10.09 ± 0.15, and 17.01 ± 0.06 for insulin, leptin, Epstein-Barr virus, cystic fibrosis, and lung cancer, respectively. The total average and SD of BAQALC

was 12.51 ± 2.64. BAQALC CR was the highest for lung cancer (average 17.01). BAQALC performance also showed the highest stability for lung cancer (SD 0.06).

4.3. Compression Ratio Comparison to Widely Used Algorithms

The results of the CR comparison to widely used algorithms are shown in Table 3. SHT, LZ, gzip, and bzip 2 were selected as widely used algorithms for compression ratio comparisons. Fifteen data were randomly chosen for assessment. In cases where the data origin was not explicit in the literature, disease marker classifications were shown.

Table 3. CR Comparison between BAQALC and Widely Used Algorithms SHT, LZW, Gzip, and Bzip 2.

Solutions	CR	Used Data	CR Average ± SD
SHT	2.22	SRX2935589	2.67 ± 0.72
	2.22	SRX2935588	
	2.22	SRX2935587	
	2.22	SRX2935586	
	2.22	SRX2935585	
	2.2	SRX2528043	
	2.21	SRX2528042	
	2.21	SRX2528041	
	2.21	SRX2528040	
	2.21	SRX2528039	
	3.44	Cholerae	
	3.44	Abscessus	
	3.34	S. cerevisiae	
	3.27	N. crassa	
	4.45	Chr22	
Gzip	3.38	SRR2916693	3.41 ± 0.87
	2.92	SRR2994368	
	2.35	SRR3211986	
	2.2	ERR739513	
	3.39	SRR3190692	
	5.59	ERR385912	
	3.85	ERR386131	
	2.71	SRR034509	
	3.15	ERR174310	
	4.24	ERR194147	
	3.29	Cholerae	
	3.31	Abscessus	
	3.1	S. cerevisiae	
	3.12	N. crassa	
	4.57	Chr22	
BAQALC	11.22	SRX2528041	12.51 ± 2.58
	11.25	SRX2528040	
	11.18	SRX2528039	
	10.44	ERX1600595	
	11	ERX1600594	
	11.38	ERX1600593	
	13.63	SRX2776809	
	13.36	SRX2776808	
	12.78	SRX2776807	
	9.85	SRX2935587	
	10.17	SRX2935586	
	10.26	SRX2935585	
	17.08	SRX2388193	
	17.06	SRX2388192	
	16.94	SRX2388191	
LZW	3.24	SRX2935588	3.24 ± 0.12
	3.23	SRX2935587	
	3.24	SRX2935586	
	3.24	SRX2935585	
	3.21	SRX2528043	
	3.21	SRX2528042	
	2.88	SRX2528041	
	3.21	SRX2528040	
	3.2	SRX2528039	
	3.2	ERX1600597	
	3.38	ERX1600596	
	3.35	ERX1600595	
	3.34	ERX1600594	
	3.32	ERX1600593	
	3.31	SRX2388195	

Table 3. *Cont.*

Solutions	CR	Used Data	CR Average ± SD
Bzip 2	4.13	SRR2916693	3.99 ± 1.15
	3.51	SRR2994368	
	2.75	SRR3211986	
	2.52	ERR739513	
	4.1	SRR3190692	
	7.19	ERR385912	
	4.65	ERR386131	
	3.17	SRR034509	
	3.82	ERR174310	
	5.08	ERR194147	
	3.52	Cholerae	
	3.57	Abscessus	
	3.41	S. cerevisiae	
	3.43	N. crassa	
	5.04	Chr22	

Comparison between BAQALC and widely used algorithm results showed that the average and SD CRs were 2.67 ± 0.72, 3.24 ± 0.12, 3.41 ± 0.87, 3.99 ± 1.15, and 12.51 ± 2.58, for SHT, LZW, gzip, bzip 2, and BAQALC, respectively. The proposed BAQALC algorithm showed the highest CR performance (average 12.52). However, BAQALC's stability was the lowest (SD 2.58), whereas LZW's stability was the highest (SD 0.12).

4.4. Compression Ratio Comparison to Related Research

The CR performance of BAQALC compared with systematically selected, related research is shown in this section. Note that for related research, CRs were averaged according to their best performance shown in their literature. In addition, most prior published results were reported in percentages. These were translated to CRs by putting 100% into Uncompressed Size (US), and percentages into Compressed Size (CS).

The results are shown in Table 4. No lossy algorithm solutions were selected because only lossless algorithms were considered for this research. In cases where the data origin was not explicit in the literature, disease marker classifications were shown. Also, in the case of the proposed solution BAQALC, only two sets from each database, i.e., from insulin, leptin, Epstein-Barr virus, cystic fibrosis, and lung cancer (a total of ten), were chosen to match the overall data numbers of other research to minimize bias (stability tends to increase when the number of data increase).

Table 4. CR Comparison between BAQALC and Related Researches.

Researches	CR	Used Data	CR Average ± SD
LW-FQZip 2	6.54	SRR2916693	6.99 ± 4.9
	6.25	SRR2994368	
	3.1	SRR3211986	
	2.87	ERR739513	
	8.55	SRR3190692	
	20.0	ERR385912	
	6.25	ERR386131	
	4.41	SRR034509	
	4.98	ERR174310	
	6.99	ERR194147	
DSRC	4.95	SRR2916693	5.83 ± 2.89
	4.31	SRR2994368	
	4.93	SRR3190692	
	12.82	ERR385912	
	5.95	ERR386131	
	3.83	SRR034509	
	4.95	ERR174310	
	4.93	ERR194147	
LFQC	7.87	SRR2916693	6.96 ± 5.4
	3.1	SRR3211986	
	2.87	ERR739513	
	17.24	ERR385912	
	6.45	ERR386131	
	4.22	SRR034509	

Table 4. *Cont.*

Researches	CR	Used Data	CR Average ± SD
UHT	3.69	Cholerae	3.91 ± 0.5
	3.74	Abscessus	
	3.66	S. cerevisiae	
	3.68	N. crassa	
	4.8	Chr22	
MUHTL	3.85	Cholerae	4.09 ± 0.55
	3.91	Abscessus	
	3.83	S. cerevisiae	
	3.8	N. crassa	
	5.07	Chr22	
Quip	4.78	SRR2916693	5.82 ± 3.16
	4.98	SRR2994368	
	3.0	SRR3211986	
	6.06	SRR3190692	
	13.89	ERR385912	
	5.65	ERR386131	
	3.98	SRR034509	
	5.0	ERR174310	
	5.0	ERR194147	
CRAM	4.57	SRR2916693	3.74 ± 0.68
	3.79	SRR2994368	
	2.95	SRR3211986	
	2.81	ERR739513	
	4.48	SRR3190692	
	3.92	ERR386131	
	3.65	SRR034509	
SCALCE	5.81	SRR2916693	6.59 ± 3.5
	5.78	SRR2994368	
	2.99	SRR3211986	
	7.87	SRR3190692	
	15.15	ERR385912	
	6.02	ERR386131	
	4.08	SRR034509	
	5.1	ERR174310	
	6.49	ERR194147	
UHTL	3.84	Cholerae	4.08 ± 0.55
	3.89	Abscessus	
	3.83	S. cerevisiae	
	3.8	N. crassa	
	5.06	Chr22	
FQZCOMP	4.69	SRR003177	5.95 ± 1.31
	7.04	SRR007215_1	
	4.6	SRR027520_1	
	7.46	SRR065390_1	
BAQALC	10.98	SRX2528043	12.51 ± 2.86
	11.24	SRX2528042	
	10	ERX1600597	
	10.25	ERX1600596	
	14.7	SRX2776811	
	13.79	SRX2776810	
	10.06	SRX2935589	
	10.13	SRX2935588	
	17.01	SRX2388195	
	16.98	SRX2388194	

LW-FQZip 2 had a CR of 6.99 (SD 4.9 according to short/long reads) regardless of long reads and short reads. QUIP's CR was 5.82 on average (SD 3.16 according to short/long reads). DSRC 2's average and SD were 5.83 and 2.89. CRAM showed a CR of 3.74 on average and 0.68 SD. LFQC showed a CR of 6.96 on average and SD of 5.4. In the case of SCALCE, the average and SD CRs were 6.59 and 3.5. UHT's average and SD CRs were 3.91 and 0.5. UHTL's average and SD CRs were 4.08 and 0.55. MUHTL's average and SD CRs were 4.09 and 0.55, and finally, FQZCOMP's were 5.95 and 1.31.

BAQALC showed the highest compression performance between the related algorithms, with a CR of 12.51. The poorest stability was shown by LFQC, with an SD of 5.1, although the compression performance was relatively impressive. UHT showed highest stability with an SD of 0.5, but its compression ranking was the worst, with 3.91.

4.5. Overall CR and Stability Comparison

In this section, overall CR and stability performance comparison of all widely used algorithms and related researches were conducted. Results are shown in Figure 5.

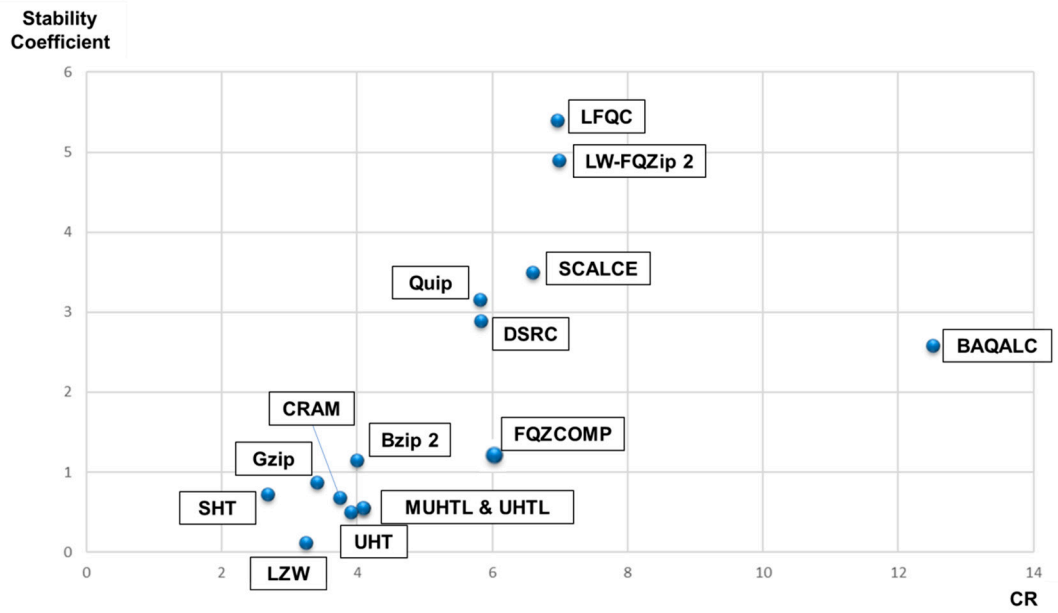

Figure 5. CR and Stability Comparison of All Solutions.

Statistical results showed that Pearson's correlation coefficient between CR and stability was 0.62, with a *p*-value of 0.018 (statistically significant under 95% confidence interval). In other words, the higher the compression ratio, the lower the stability of the algorithms.

5. Discussion and Conclusions

Among all compression algorithms considered, the proposed BAQALC algorithm showed the highest compression performance. Although BAQALC's stability was not the best, it was not significantly different from the highest stability performers, e.g., LZW or UHT.

An overall trend of DNA compression solutions can be inferred from this research. There was a trade-off trend between compression performance and stability performance among related research in DNA data compression. As noted in Section 4.2., the higher the stability coefficient (SD), the lower the stability.

BAQALC displayed the best performance with lung cancer FASTQ data. This should be taken into consideration for future research, because there is a possibility to improve the performance according to disease type.

In addition, this solution is designed to operate within a blockchain network, making it immune to hacking and unreliable uploading. The proposed solution is envisioned to contribute to providing an efficient and secure transmission and storage platform for next-generation medical informatics systems regarding DNA.

The proposed BAQALC solution specifically counters UHT, UHTL, and MUHTL solutions provided by Al-Okaily et al., which are methods based on Huffman logic. These two studies are based on the same variants of dictionary coding methods. Although Huffman is a powerful compression method when modified, our results showed that LZW modifications could better serve as compression solutions for DNA FASTQ files.

This research proposes BAQALC, a lossless compression algorithm that efficiently and securely transmits and stores the immense amount of all DNA sequence data types. Through BAQALC, efficient and reliable transmission and storage of the immense amounts of DNA sequence data will be easier and more reliable, even for complex data such as FASTQ that are being generated by NGS for genomes.

The limitation of this research is that the stability was low compared to widely-used algorithms. Future research should include reducing the process overload or instability of the algorithm.

Furthermore, more specific algorithm modes that are optimized according to disease types should be researched. Lastly, in terms of service rather than research, a reward system should be adopted by public databases in order to facilitate motivation among miners.

Author Contributions: S.-J.L. designed the entire research, developed, analyzed, and evaluated the research. G.-Y.C. mainly contributed to the development of this research. T.-R.L. managed the overall research as project manager.

Funding: Korea Technology and Information Promotion Agency: S2601135, Curaum Inc.: CISKRSCGA01.

Acknowledgments: This research was government funded by TIPA (Technology and Information Promotion Agency), grant number S2601135. This research was also funded by the Curaum Research Scholarship Grant, grant number CISKRSCGA01.

Conflicts of Interest: There are no conflicts of interest.

References

1. Van Dijk, E.L.; Auger, H.; Jaszczyszyn, Y.; Thermes, C. Ten years of next-generation sequencing technology. *Trends Genet.* **2014**, *30*, 418–426. [CrossRef] [PubMed]
2. Chen, L.; Mei, X.P.; Gao, C.J.; Zhang, G.H.; Sun, X.D. Histologic Distribution, Fragment Cloning, and Sequence Analysis of G Protein Couple Receptor 30 in Rat Submaxillary Gland. *Anat. Rec. Integr. Anat. Evol. Biol.* **2011**, *294*, 706–711. [CrossRef] [PubMed]
3. Sardaraz, M.; Tahir, M.; Ikram, A.A. Advances in high throughput DNA sequence data compression. *J. Bioinf. Comput. Biol.* **2016**, *14*, 1630002. [CrossRef] [PubMed]
4. Zhu, Z.; Zhang, Y.; Ji, Z.; He, S.; Yang, X. High-throughput DNA sequence data compression. *Briefings Bioinf.* **2015**, *16*, 1–15. [CrossRef] [PubMed]
5. Nilsson, R.H. Taxonomic reliability of DNA sequences in public sequence databases: A fungal perspective. *PLoS ONE* **2006**, *1*, e59. [CrossRef] [PubMed]
6. Showell, C. Barriers to the use of personal health records by patients: A structured review. *PeerJ* **2017**, *5*, e3268. [CrossRef] [PubMed]
7. Lee, S.J.; Cho, G.Y.; Song, S.H.; Jang, J.S.; Lee, K.I.; Lee, T.R. Solution for Efficient Vital Data Transmission and Storing in m-Health Environment. *J. Digit. Converg.* **2015**, *13*, 227–235. [CrossRef]
8. Bouillaguet, C.; Derbez, P.; Dunkelman, O.; Keller, N.; Rijmen, V.; Fouque, P.A. Low-data complexity attacks on AES. *IEEE Trans. Inf. Theory* **2012**, *58*, 7002–7017. [CrossRef]
9. Zhang, L.Y.; Liu, Y.; Wang, C.; Zhou, J.; Zhang, Y.; Chen, G. Improved known-plaintext attack to permutation-only multimedia ciphers. *Inf. Sci.* **2018**, *430–431*, 228–239. [CrossRef]
10. Hosseini, M.; Pratas, D.; Pinho, A.J. Cryfa: A secure encryption tool for genomic data. *Bioinformatics* **2018**, bty645. [CrossRef] [PubMed]
11. Jones, D.C.; Ruzzo, W.L.; Peng, X.; Katze, M.G. Compression of next-generation sequencing reads aided by highly efficient de novo assembly. *Nucleic Acids Res.* **2012**, *40*, e171. [CrossRef] [PubMed]
12. Tembe, W.; Lowey, J.; Suh, E. G-SQZ: Compact encoding of genomic sequence and quality data. *Bioinformatics* **2010**, *26*, 2192–2194. [CrossRef] [PubMed]
13. Hach, F.; Numanagic, I.; Sahinalp, S.C. DeeZ: Reference-based compression by local assembly. *Nat. Methods* **2014**, *11*, 1082–1084. [CrossRef] [PubMed]
14. Khan, M.A.; Salah, K. IoT security: Review, blockchain solutions, and open challenges. *Future Gener. Comput. Syst.* **2018**, *82*, 395–411. [CrossRef]
15. Lee, S.J.; Rho, M.J.; Yook, I.H.; Park, S.H.; Jang, K.S.; Park, B.J.; Lee, O.; Lee, D.J.; Choi, I.Y. Design, Development and Implementation of a Smartphone Overdependence Management System for the Self-Control of Smart Devices. *Appl. Sci.* **2016**, *6*. [CrossRef]
16. Doolittle, G.C.; Spaulding, A.O.; Williams, A.R. The Decreasing Cost of Telemedicine and Telehealth. *Telemed. J. E Health* **2011**, *17*, 671–675. [CrossRef] [PubMed]
17. Chen, M.; Gonzalez, S.; Leung, V.; Zhang, Q.; Li, M. A 2G-RFID-Based E-Healthcare System. *IEEE Wirel. Commun.* **2010**, *17*, 37–43. [CrossRef]
18. What is DNA?–Genetics Home Reference–NIH. Available online: https://ghr.nlm.nih.gov/primer/basics/dna (accessed on 23 August 2018).

19. Bonfield, J.K.; Mahoney, M.V. Compression of FASTQ and SAM format sequencing data. *PLoS ONE* **2013**, *8*, e59190. [CrossRef] [PubMed]
20. Guerra, A.; Lotero, J.; Isaza, S. Performance comparison of sequential and parallel compression applications for DNA raw data. *J. Supercomput.* **2016**, *72*, 4696–4717. [CrossRef]
21. Cho, G.Y.; Lee, S.J.; Lee, T.R. An optimized compression algorithm for real-time ECG data transmission in wireless network of medical information systems. *J. Med. Syst.* **2015**, *39*, 161. [CrossRef] [PubMed]
22. Cho, G.Y.; Lee, G.Y.; Lee, T.R. Efficient Real-Time Lossless EMG Data Transmission to Monitor Pre-Term Delivery in a Medical Information System. *Appl. Sci.* **2017**, *7*. [CrossRef]
23. Peng, Z.; Wang, G.; Jiang, H.; Meng, S. Research and improvement of ECG compression algorithm based on EZW. *Comput. Methods Programs Biomed.* **2017**, *145*, 157–166. [CrossRef] [PubMed]
24. Zhang, G.; Wu, T.; Wan, Z.; Song, Z.; Yu, M.; Wang, D.; Li, L.; Chen, F.; Xu, X. A method to differentiate between ventricular fibrillation and asystole during chest compressions using artifact-corrupted ECG alone. *Comput. Methods Programs Biomed.* **2017**, *141*, 111–117. [CrossRef] [PubMed]
25. Hach, F.; Numanagic, I.; Alkan, C.; Sahinalp, S.C. SCALCE: Boosting sequence compression algorithms using locally consistent encoding. *Bioinformatics* **2012**, *28*, 3051–3057. [CrossRef] [PubMed]
26. Ziv, J.; Lempel, A. Compression of Individual Sequences via Variable-Rate Coding. *IEEE Trans. Inf. Theory* **1978**, *24*, 530–536. [CrossRef]
27. Ziv, J.; Lempel, A. Universal Algorithm for Sequential Data Compression. *IEEE Trans. Inf. Theory* **1977**, *23*, 337–343. [CrossRef]
28. Roguski, L.; Deorowicz, S. DSRC 2–Industry-oriented compression of FASTQ files. *Bioinformatics* **2014**, *30*, 2213–2215. [CrossRef] [PubMed]
29. Stanford, K.I.; Middelbeek, R.J.; Goodyear, L.J. Exercise Effects on White Adipose Tissue: Beiging and Metabolic Adaptations. *Diabetes* **2015**, *64*, 2361–2368. [CrossRef] [PubMed]
30. Petrovic, N.; Walden, T.B.; Shabalina, I.G.; Timmons, J.A.; Cannon, B.; Nedergaard, J. Chronic peroxisome proliferator-activated receptor gamma (PPARgamma) activation of epididymally derived white adipocyte cultures reveals a population of thermogenically competent, UCP1-containing adipocytes molecularly distinct from classic brown adipocyte. *J. Biol. Chem.* **2010**, *285*, 7153–7164. [CrossRef] [PubMed]
31. Huang, Z.A.; Wen, Z.; Deng, Q.; Chu, Y.; Sun, Y.; Zhu, Z. LW-FQZip 2: A parallelized reference-based compression of FASTQ files. *BMC Bioinf.* **2017**, *18*. [CrossRef] [PubMed]
32. Fritz, M.H.Y.; Leinonen, R.; Cochrane, G.; Birney, E. Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res.* **2011**, *21*, 734–740. [CrossRef] [PubMed]
33. Nicolae, M.; Pathak, S.; Rajasekaran, S. LFQC: A lossless compression algorithm for FASTQ files. *Bioinformatics* **2015**, *31*, 3276–3281. [CrossRef] [PubMed]
34. Al-Okaily, A.; Almarri, B.; Al Yami, S.; Huang, C.H. Toward a Better Compression for DNA Sequences Using Huffman Encoding. *J. Comput. Biol.* **2017**, *24*, 280–288. [CrossRef] [PubMed]
35. The Gzip Homepage. Available online: https://www.gzip.org/ (accessed on 23 August 2018).
36. Pinho, A.J.; Pratas, D. MFCompress: A compression tool for FASTA and multi-FASTA data. *Bioinformatics* **2014**, *30*, 117–118. [CrossRef] [PubMed]
37. Mohammed, M.H.; Dutta, A.; Bost, T.; Chadaram, S. DELIMINATE—A fast and efficient method for loss-less compression of genomic sequences. *Bioinformatics* **2012**, *28*, 2527–2529. [CrossRef] [PubMed]
38. Yang, C.; Chen, X.; Xiang, Y. Blockchain-based publicly verifiable data deletion scheme for cloud storage. *J. Netw. Comput. Appl.* **2018**, *103*, 185–193. [CrossRef]
39. Goni, A.; Burgos, A.; Dranca, L.; Rodriguez, J.; Illarramendi, A.; Bermudez, J. Architecture, cost-model and customization of real-time monitoring systems based on mobile biological sensor data-streams. *Comput. Methods Programs Biomed.* **2009**, *96*, 141–157. [CrossRef] [PubMed]
40. Leinonen, R.; Sugawara, H.; Shumway, M. The Sequence Read Archive. *Nucleic Acids Res.* **2011**, *39*, D19–D21. [CrossRef] [PubMed]
41. The Top 10 Causes of Death, Fact Sheets. Available online: http://www.who.int/mediacentre/factsheets/fs310/en/index1.html (accessed on 23 August 2018).
42. Ischemic Cardiomyopathy: Symptoms, Causes, and Treatment. Available online: https://www.healthline.com/health/ischemic-cardiomyopathy (accessed on 23 August 2018).

43. Friedrichs, I.; Bingold, T.; Keppler, O.T.; Pullmann, B.; Reinheimer, C.; Berger, A. Detection of herpesvirus EBV DNA in the lower respiratory tract of ICU patients: A marker of infection of the lower respiratory tract? *Med. Microbiol. Immunol.* **2013**, *202*, 431–436. [CrossRef] [PubMed]
44. Shoemaker, S.A.; Scoggin, C.H. DNA molecular biology in the diagnosis of pulmonary disease. *Clin. Chest Med.* **1987**, *8*, 161–171. [PubMed]
45. Taneja, T.K.; Sharma, S.K. Markers of small cell lung cancer. *World J Surg. Oncol.* **2004**, *2*, 10. [CrossRef] [PubMed]