Optimized and Cost Considering Huffman Code For Biological Data Transmission

Abstract

Keywords:

1. Introduction

In the recent years, application of battery-powered portable devices, e.g. laptop computers and mobile phones has increased significantly. Proper representation of digital data and their transmission efficiency has become a primary concern for digital community because it affects the performance, reliability, and the cost of computation in both portable and non-portable devices. CMOS technologies were developed in order to reduce the power consumption both in data processing and transmission. In order to increase transmission speed and reduce transmission cost, parallel data transmission methods are widely used. However, parallel transmission is limited to short distance communications, e.g. locally connected devices, internal buses. Ruling out the possible availability of parallel transmission links over long distance, we are left with its serial alternative only. If we attempt to transfer big files, e.g. DNA sequences, over a serial transmission link then it would take a significant amount of time. However, we can not overlook this problem because at present parallel processing is increasingly used to increase throughput and in parallel processing architecture, processing units are usually distributed in different physical locations and task sharing is a must in such architecture.

Data encoding techniques came into action to improve the data transmission efficiency over serial communication medium by compressing data before transmitting. Efficiency can be measured in terms of incurred cost, required storage space, consumed power, time spent and likewise. Data must be encoded to meet the purposes like: unambiguous retrieval of information, efficient storage, efficient transmission and etc. Let a message consist of

sequences of characters taken from an alphabet Σ, where $\alpha_1, \alpha_2, \alpha_3 \ldots, \alpha_r$ are the elements that represent the characters in the source Σ. The length of α_i represents its cost or transmission time, i.e., $c(\alpha_i) = length(\alpha_i)$. A codeword w_i is a string of characters in Σ, i.e., $w_i \in \Sigma^+$. If a codeword is $w_i = \alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{in}$, then the length or cost of the codeword is the sum of the lengths of its constituent elements:

$$\text{cost}(w_i) = \sum_{j=1}^{n} c(\alpha_{ij}) \tag{1}$$

If all the elements of a codeword has unit cost or length then the cost of the codeword is equivalent to the length of the codeword. However, it is not necessary for the elements in the codeword to have equal length or cost. For example, in Morse Code all the ASCII characters are encoded as sequence of dots ($\cdot$) and dashes($-$) where a dash is three times longer than a dot in duration [1]. However, the Morse code scheme suffers from the prefix problem [2]. Ignoring the prefix problem, Morse Code results in a tremendous savings of bits over ASCII representation. Using Morse Code, we can treat the binary bits differently; 0 as a dot and 1 as a dash. Even if we consider the voltage level to represent the binary digits then they are still different. Table 1 shows the logic level to represent binary digits in CMOS and TTL technologies.

Table 1: Example of binary logic level

Technology	0	1	Notes
CMOS	0 V to $\frac{V_{DD}}{2}$	$\frac{V_{DD}}{2}$ to V_{DD}	V_{DD}= supply voltage
TTL	0 V to 0.8 V	2 V to V_{CC}	V_{CC} is 4.75 V to 5.25 V

As the unequal letter cost problem is not new therefore it has been addressed by different researchers. The more general case where the costs of the letters as well as the probabilities of the words are arbitrarily specified was treated in [3]. A number of other researchers have focused on uniform sources and developed algorithm for the unequal letter costs encoding [4, 5, 6, 7, 8]. Let $p_1, p_2, \ldots, p_n$ be the probabilities with which the source symbols occur in a message and the codewords representing the source symbols are $w_1, w_2, \ldots, w_n$ then the cost of the code W is:

$$C(W) = \sum_{i=1}^{n} cost(w_i) . p_i \tag{2}$$

The aim of producing an optimal code with unequal letter cost is to find a codeword W that consists of n prefix code letters each with minimum cost c_i that produces the overall minimum cost $C(W)$, given that costs $0 < c_1 \leq c_2 \leq c_2 \ldots \leq c_n$, and probabilities $p_1 \geq p_2 \geq \ldots \geq p_n > 0$.

Huffman code is an efficient data compression scheme that takes into account the probabilities at which different quantization levels are likely to occur and results in fewer data bits on the average. It is widely used to compress biological data, however, all the techniques use its classical form where bits are treated equally. Out of many variation of the Huffman code where cost of bits are treated unequally, the most recent version is dercribed in [9].

The rest of the paper is organised as follows: Section 2 presents the background study of the issues in biological data processing and the Huffman code. The proposed approach is described in Section 3. Experimental results and discussion are presented in Section 4 . Finally, concluding remarks are presented in Section 5.

2. Background

2.1. Issues on Biological Data Transmission

The size of biological data including DNA sequences increase with an ever expanding rate and will be bigger and bigger in the future. These Biological data are stored in biology database, the exponential growth of these database become a big problem to all biological data processing methods [10]. Different operation will be applied to these data such as, searching [11],e-mail attachment [12], alignment [13], and transmission on distributed computing [14]. Interestingly, biological data compression can play a key role in all biological data processing.

A recent deluge of interest in the development of new tools for biological data processing, these all algorithms needs an efficient methods for data compression. The main objective of data compression methods is minimizing the number of bits in the data representation. In [15] authors propose a new general data structure and data encoding approach for the efficient storage of genomic data. This method encode only the differences between a genome

sequence and a reference sequence, the method use different encoding scheme from fixed codes such as Golomb, Elias codes and variable codes such as Huffman codes. Other methods based on same idea to encode only the difference between reference sequence and the target one, Authors in [12] uses Huffman code for encoding difference between sequence to sent it as an email attachment, but these methods suffer that they must sent the reference sequence for at least one time for each species.
Wang and Zhang [16] proposed a new scheme for referential compression of genomes based on the chromosome level. The Algorithm aim to search for longest common subsequence between matching parts and the differences encoded using Huffman coding.
All previous studies focus only on the differences and the relation between continuation of the sequence, and without improvement of the encoding scheme.

2.2. Huffman Codes

In computer science and information theory, Huffman code is an entropy encoding algorithm used for lossless data compression. It takes into account the probabilities at which different symbols are likely to occur and results into fewer data bits on the average. For any given set of symbols and associated occurrence probabilities, there is an optimal encoding rule that minimises the number of bits needed to represent the source. Encoding symbols in predefined fixed length code, does not attain an optimum performance, because every character consumes equal number of bits irrespective to their degree of contribution to the whole message. Huffman code tackles this by generating variable length codes, given a probability usage frequency for a set of symbols. It generates prefix-code to facilitate unambiguous retrieval of information. A scheme of prefix code assigns codes to letters in Σ to form codeword w_i such that none of them is a prefix to another. For example, the codes $\{1, 01, 001, 0001\}$ and $\{000, 001, 011, 111\}$ are prefix-free, whereas the code $\{1, 01, 100\}$ is not, because 1 is a prefix in 100.

Applications of Huffman code are pervasive throughout computer science. The algorithm to completely perform Huffman encoding and decoding is explained by [17]. It can be used effectively where there is a need for a compact code to represent a long series of a relatively small number of distinct bytes. For example, Table 1 shows 8 different ASCII characters, their frequencies, ASCII codes and the codewords generated for those symbols using Huffman code. It is seen from the table that the codeword to represent each character

is compressed and the most frequent character gets the shortest code. In this example, the compression ratio obtained by Huffman code is 64.16%.

Table 2: Example of application of Huffman Code to compress ASCII characters

Symbols	Frequency	ASCII Code	Codewords using Huffman Code
A	50	01000001	00
B	35	01000010	101
C	42	01000011	110
D	22	01000100	1001
E	65	01000101	01
F	25	01000110	1111
G	9	01000111	1000
H	23	01001000	1110

There are many other variants of Huffman codes that compress source data to reduce data size and/or transmission cost. For example, Mannan and Kaykobad introduced block technique in Huffman coding which overcomes the limitation of reading whole message prior to encoding[18]. In classical Huffman coding scheme, the letter costs are considered as equal. The unequal letter cost versions of Huffman codes scheme are proposed in [19, 20, 21, 22]. In the unequal letter cost version of the classical Huffman code, letters of the alphabet are considered as unequal. Recently, in [9] a method is proposed to show the effects of unequal bits cost on classical Huffman code. The idea of this method is to assign the most frequent symbol the minimum cost and the least frequent symbol the maximum cost code, whereas classical Huffman code assigns most frequent symbol the minimum length and the least frequent symbol the maximum length code.

3. Approach

3.1. Proposed Scheme

A genome is a stretch of DNA (or RNA) that codes for a polypeptide (protein), that is a set of amino acids bound together in specific order. Each

genomic sequence consist of nucleotide bound together, which are interpreted by the cellular machinery in groups of three, called triplets. This the main raison to divide the whole sequence in a set of triplet and give a code to each triplet. The first step in the optimised cost considering algorithm is cutting the sequence in triplet, then compute the frequency of each triplet in the whole sequence. This table of frequencies are used by the cost considering Huffman code to generate low cost code to each triplet (frequency). Finally these codes with frequencies are used by the optimised cost considering algorithm to generate the optimal allocation with a given penalty.

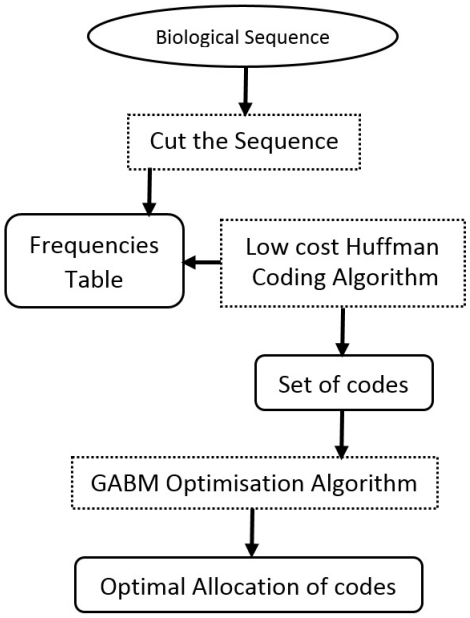

Figure 1: The proposed Scheme

3.2. Power Efficient Huffman code

3.3. Optimisation of the Codes

3.3.1. Problem formulation

The problem of finding the best allocation of codes to each symbol can be modelled as an Assignment Problems with Constraint, the problem is formulated as follows :

Algorithm 1 Cost-considering / Unequal bit cost Coding

Require: Distinct symbols contained in the message to be encoded and their frequencies

Ensure: Non-uniform / variable letter cost i.e, Cost-considering balanced tree

1: **for** each distinct symbol i **do**
2: $Enqueue\,(max\_Q, frequency\ [\ i\ ])$
3: **end for**
4: create a root node
5: $cost\ [\ root\ ] \leftarrow 0$
6: $Enqueue\,(min\_Q, cost\ [\ root\ ])$
7: Define costs of the left and right child of the binary tree
8: **repeat**
9: $cost\_of\_parent\_node \leftarrow Dequeue\,(min\_Q)$
10: create $left$ and $right$ child for this node
11: $cost\ [\ left\_child\ ] \leftarrow cost\_of\_parent\_node + left\_child\_cost$
12: $Enqueue\,(min\_Q, cost\ [\ left\_child\ ])$
13: $cost\ [\ right\_child\ ] \leftarrow cost\_of\_parent\_node + right\_child\_cost$
14: $Enqueue\,(min\_Q, cost\ [\ right\_child\ ])$
15: Mark parent node as explored
16: **until** $2\,(n-1)$ nodes are created
17: **while** $min\_Q \neq \emptyset$ **do**
18: $leaf\_node \leftarrow Dequeue\,(min\_Q)$
19: $frequency[leaf\_node] \leftarrow Dequeue\,(max\_Q)$
20: **end while**
21: **for** each parent node j **do**
22: $frequency\ [\ j\ ] \leftarrow frequency\ [\ left\_child\ ] + frequency\ [\ right\_child\ ]$
23: **end for**
24: **repeat**
25: **if** conflict between nodes **then**
26: resolve conflict by swapping conflicted nodes
27: calculate and reassign cost of all affected nodes
28: calculate and reassign frequency of all affected nodes
29: **end if**
30: **until** all conflicts are resolved

Definition: Given a set of codes $C = \{C_1, C_2...C_n\}$, and a set of frequencies $C = \{Q_1, Q_2...Q_n\}$. For each code we have the length of the code $|C_i|$ (number of bits) and the cost of the code S_{C_i} (cost of ones and zeros), the objective is to assign to each frequency a code in order to minimize the total number of bits, while respecting the initial assignment total cost S_t with a given penalty λ. This penalty coefficient represent the allowed Marge can be sacrificed for cost to optimize number of bits
The Objective Function is :

$$Minimise \sum(|C_i| \times Q_j) \tag{3}$$

while :

$$\sum(|S_{C_i}| \times Q_j) \leqslant (\lambda + 1)S_t \tag{4}$$

3.3.2. Basic Genetic Algorithm

Genetic Algorithm (GA) is a bio-inspired meta-heuristics algorithm developed by []. GA is a stochastic optimization algorithm imitate the natural evolution process of genomes. GA started by generate a population of random feasible solutions, the optimization process of GA is as follow, and we select two solution among the population, by one of the well-known selection techniques. This two selected solution will be considered as two parents, we generate two other new solutions from the two selected solution (Sons), this new solutions can be mutate according to a given mutation probability. The quality of each solution is computed with the fitness function which control the evolution of the GA population by the deletion of the worst solution and insertion of the good solutions among parents and sons. This processes is repeated until the stopped criteria is achieved which can be the number of generation or if the population is stabilized.

3.3.3. GA for Bits minimisation

The main objective of the GA optimisation algorithm for bits minimisation (GaBm) problem is to assign to each frequency a specific code. The GaBm population is generated randomly from the different codes, and the affectation of codes to different frequencies given by the low cost considering algorithm to initial population to ensure that the final solution is better or at least equal to the solution given by the low cost considering Huffman code algorithm (step 1). The optimisation process of the genetic algorithms start with the selection of two solution randomly from the population (step

3). After that the operations of genetic algorithms are apply for the initial population optimization (see figure 3). Firstly the crossover operation are applied to these two selected solution (considered as parents) to generate two new solutions (considered as sons)(step 4). These two children may contain conflict like finding a duplicated code allocated to two different frequencies in the solution, so a regulation step is done to ensure the correctness of the solution (see figure 3). Secondly these two new solutions are mutated according to a predefined probability (step 5), to ensure a good diversification on the space solutions. Each new generated solution must satisfy the cost constraint, these children must be valid with satisfying the cost constraint. The next step is to add these two new solutions (children) to the population (step 7) (see figure 2). Finally the new population are ranked by fitness (step 8), and the worst solution are deleted until the initial size of the population are achieved (step 9). the whole process are repeated until the max number of operation is achieved (step 10).

Algorithm 2 GA for bits minimisation

1: Population initialization (P). Max number of generation not achieved
2: Select two solutions S_1, S_2 form P.
3: Crossover S_1, S_2 to generate S_{11}, S_{21} (Children).
4: Mutate S_{11}, S_{21}.
5: Validate children with cost constraint (equation 4).
6: Add children to population
7: Rank the population by fitness
8: Remove worst candidates until population limit
9: Return to 2;
10: Display the best solution from the population P;

4. Results And Discussion

The approach has been evaluated with different real biological data (genomes), these genomes from The National Center for Biotechnology Information (NCBI) available on (http://www.ncbi.nlm.nih.gov) []. In table () the different data set are described with the size of each of them and the references on the biological data bank.

The table 3 present the results founded by the basic Huffman code, cost considering algorithm and optimised cost considering algorithm. The result

Table 3: Dataset description

Data sets	**Name**	**Size (bp)**	**Reference**
Genome 1	Mycobacterium smegmatis	6,988,302	CP009496
Genome 2	Amycolatopsis benzoatilytica	8,704,271	NZ_KB912942
Genome 3	Mycobacterium rhodesiae NBB3	6,415,739	CP003169
Genome 4	Streptomyces bottropensis ATCC 25435	8,955,726	NZ_KB911581
Genome 5	Mycobacterium smegmatis str. MC2 155	6,988,269	CP009494
Genome 6	Mycobacterium smegmatis MKD8	7,092,137	NZ_KI421511
Genome 7	Bradyrhizobium WSM471	7,784,016	NZ_CM001442
Genome 8	Amycolatopsis thermoflava N1165	8,677,910	NZ_CM001442
Genome 9	Bacillus thuringiensis Bt407	6,026,843	NZ_CM000747
Genome 10	Bacillus thuringiensis serovar thuringiensis	6,323,123	NZ_CM000748
Genome 11	Pseudomonas aeruginosa 9BR	6,801,503	NZ_AFXI01000001
Genome 12	Bacillus thuringiensis serovar berliner ATCC	6,260,142	NZ_CM000753
Genome 13	Bacillus thuringiensis serovar pakistani	6,037,513	NZ_CM000750
Genome 14	Pseudomonas aeruginosa LES400	6,591,121	CP006982

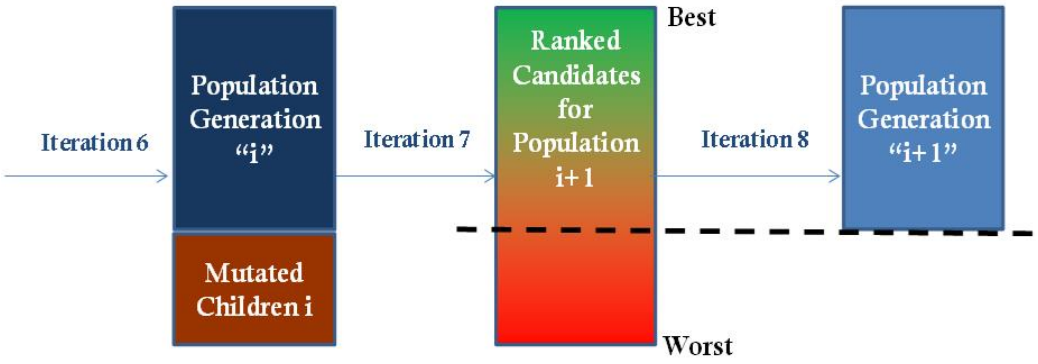

Figure 2: Population Update for genetic algorithm

show that the number of bits of Huffman algorithm is the minimum number among the other algorithm but the cost is very high. The cost considering algorithm improve the quality of the generated codes in terms of cost but the number of bits. The optimised cost considering algorithm try to find the best allocation of codes to frequencies while the cost constraint respected. The table 4 present the best founded number of bits with different penalty, for each genome we find the max number of useful penalty, after this value, increasing the penalty are in-useful (See figure 3), the number of bits achieve the minimum number but the cost stop decreasing (point 1 in figure 3) and this number of bits stabilized while the cost still increasing until it stabilized also (point 2 figure 3), after this point the cost and number of bits are stabilized.

5. Conclusion

In this paper we have proposed a new approach for efficient data compression using Huffman code and optimised strategy. The new approach is divided into phases, firstly a cost considering Huffman algorithm are proposed which reduce the cost of the generated codes, these codes are secondary passed by the optimisation algorithm to reduce the global number of bits using a cost penalty.
The proposed approach is tested with biological genomic sequence and a performance comparison is made with the standard Huffman code and the cost considering without optimisation. Simulation results showed that the

Table 4: GABM for comparison for cost and number of bits of different approaches with different datasets

Data sets	Huffman Algorithm		CCA		OCCA	
	Cost	**Bits**	**Cost**	**Bits**	**Cost**	**Bits**
Genome 1	76787151	37256819	67272097	41270715	67272097	40459089
Genome 2	100425402	48778740	88334137	54602397	88334137	53281409
Genome 3	75940155	36860555	66629579	40873035	66629579	40128089
Genome 4	103552729	50047265	90758205	56109303	90758205	54281713
Genome 5	82234926	39772838	71876260	44199916	71876260	43338060
Genome 6	83454842	40370894	72945595	44805759	72945595	43945321
Genome 7	92539488	44977876	81176987	49688369	81176987	49052869
Genome 8	99613856	48183688	87033015	53778627	87033015	52820213
Genome 9	71876739	34874617	62637908	38314532	62637908	37684296
Genome 10	75324432	36576034	65658000	40226924	65658000	39558890
Genome 11	80766360	39092450	70462778	43020460	70462778	42260476
Genome 12	74560825	36179579	65001120	39772574	65001120	39077012
Genome 13	71562941	34737083	62335103	38118291	62335103	37629173
Genome 14	78261299	37843793	68212426	41745748	68212426	40979588

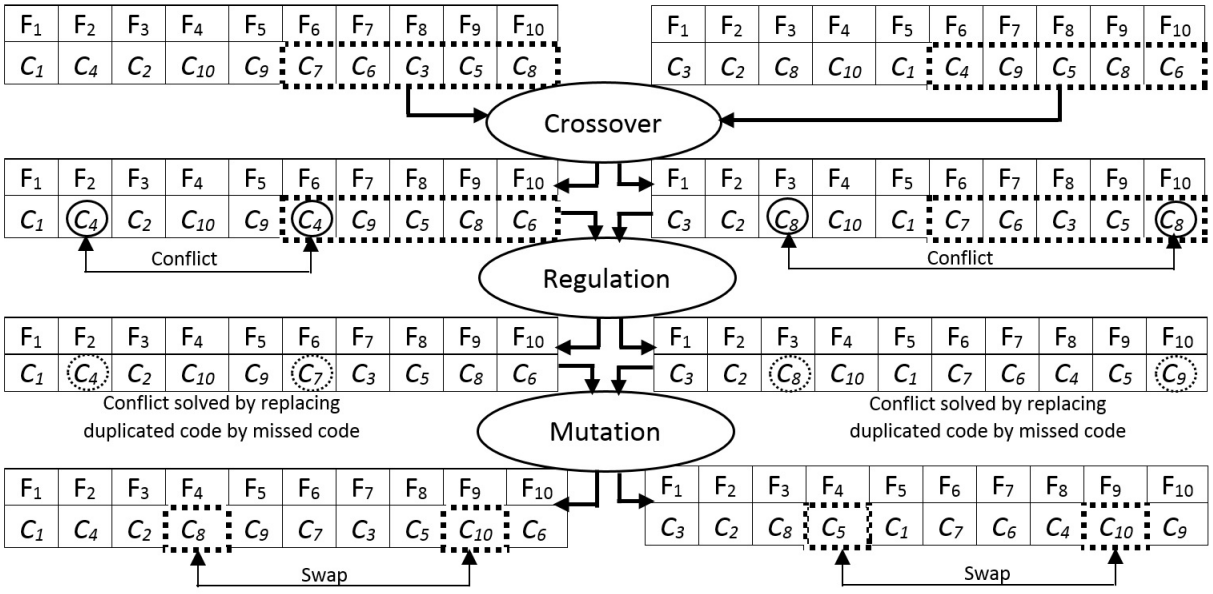

Figure 3: Operations of genetic algorithm

proposed approach is more robust and efficient compared to other competing algorithms because its penalty based optimisation strategy to search the best allocation of codes to different frequencies.

References

[1] W. A. Redmond, International morse code, Microsoft Encarta 2009 [DVD] (1964) 275–278.

[2] P. D. Grunwald, P. M. B. Vitany, Kolmogorov complexity and information theory, Journal of Logic, Language and Information 12 (2003) 497–529.

[3] R. Karp, Minimum-redundancy coding for the discrete noiseless channel, IRE Transactions on Information Theory 7 (1) (1961) 27–38.

[4] E. N. Gilbert, Coding with digits of unequal costs, IEEE Transactions on Information Theory 41.

Table 5: influence of penalty on bit minimization

Data sets	**Cost**	**Bits**	$\lambda(\%)$
Mycobacterium smegmatis	69956907	37933325	4
Amycolatopsis benzoatilytica	92682206	49593802	5
Mycobacterium rhodesiae	69292943	37581057	4
Streptomyces bottropensis	96107532	50292416	6
Mycobacterium smegmatis. MC2	74730060	40325704	4
Mycobacterium smegmatis MKD8	76528930	40723850	5
Bradyrhizobium WSM471	84396016	46056472	4
Amycolatopsis thermoflava	92205957	48298731	6
Bacillus thuringiensis serovar thuringiensis	65138929	35283557	4
Bacillus thuringiensis Bt407	68258009	37234983	4
Pseudomonas aeruginosa 9BR	72568094	40166110	3
Bacillus thuringiensis serovar berliner	66949136	36822940	3
Bacillus thuringiensis serovar pakistani	65431387	34963921	5
Pseudomonas aeruginosa LES400	71603920	38158306	5

[5] R. M. Krause, Channels which transmit letters of unequal duration, Information Control 5 (1962) 13–24.

[6] B. Varn, Optimal variable length codes -Arbitrary symbol cost and equal code word probability, Information Control (19) (1971) 289–301.

[7] D. Altenkamp, K. Mehlhorn, Codes: Unequal probabilities, unequal letter costs, Journal of the Association for Computing Machinery 27 (3) (1980) 412–427.

[8] Y. Perl, M. R. Garey, S. Even, Efficient generation of optimal prefix code: Equiprobable words using unequal cost letters, Journal of the ACM (JACM) 22 (2) (1975) 202–214.

[9] S. Kabir, T. Azad, A. S. M. A. Alam, M. Kaykobad, Effects of unequal bit costs on classical huffman codes, in: 17th International Conference on Computer and Information Technology, 2014, pp. 96–101.

[10] P. F. T. G. L. H. W. H. D. P. H. R. K. M. S. S. S. P. S. T. O. W. .

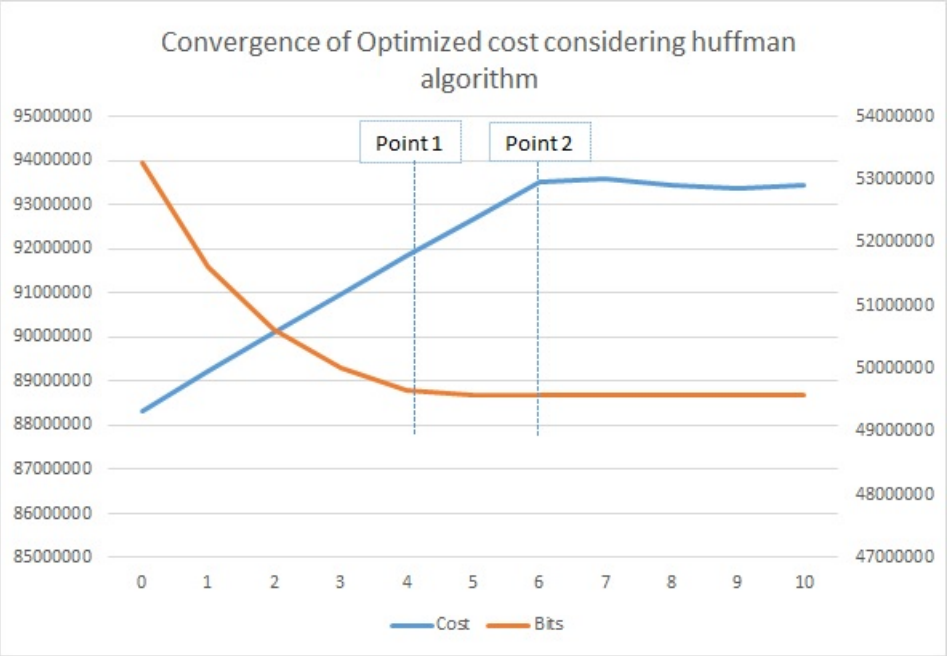

Figure 4: Convergence of Optimized cost considering Huffman algorithm

S. Y. R. Doug Howe, Maria Costanzo, Big data: The future of biocuration, Nature 455 (2008) 47–50.

[11] F. Valentin, S. Squizzato, M. Goujon, H. McWilliam, J. Paern, R. Lopez, Fast and efficient searching of biological data resourcesusing eb-eye, Briefings in bioinformatics 11 (4) (2010) 375–384.

[12] S. Christley, Y. Lu, C. Li, X. Xie, Human genomes as email attachments, Bioinformatics 25 (2) (2009) 274–275. doi:10.1093/bioinformatics/btn582.

[13] R. Chenna, H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, J. D. Thompson, Multiple sequence alignment with the clustal series of programs, Nucleic acids research 31 (13) (2003) 3497–3500.

[14] T.-H. Chang, S.-L. Wu, W.-J. Wang, J.-T. Horng, C.-W. Chang, A novel approach for discovering condition-specific correlations of gene expres-

sions within biological pathways by using cloud computing technology, BioMed research international 2014.

[15] M. C. Brandon, D. C. Wallace, P. Baldi, Data structures and compression algorithms for genomic sequence data, Bioinformatics 25 (14) (2009) 1731–1738.

[16] C. Wang, D. Zhang, A novel compression tool for efficient storage of genome resequencing data, Nucleic acids research 39 (7) (2011) e45–e45.

[17] J. Amsterdam, Data compression with huffman coding, BYTE 11 (5) (1986) 98–108.

[18] M. A. Mannan, M. Kaykobad, Block huffman coding, Computers and Mathematics with Applications.

[19] M. Golin, N. Young, Prefix codes: Equiprobable words, unequal letter costs, SIAM JOURNAL ON COMPUTING 25 (6) (1996) 1281–1292.

[20] M. J. Golin, C. Kenyon, N. E. Young, Huffman coding with unequal letter costs, in: ACM Symposium on Theory of Computing, 2002, pp. 785–791.

[21] P. Bradford, M. Golin, L. Larmore, W. Rytter, Optimal prefix-free codes for unequal letter costs: Dynamic programming with the Monge property, JOURNAL OF ALGORITHMS 42 (2) (2002) 277–303.

[22] M. J. Golin, C. Mathieu, N. E. Young, Huffman Coding with Letter Costs: A Linear-Time Approximation Scheme, SIAM JOURNAL ON COMPUTING 41 (3) (2012) 684–713.