Comparison of Protein Corpuses

Wedajo Diribi and Kumudha Raimond

***Abstract*—This paper presents a comparison of two protein corpuses. The protein corpus is a data set of four files for evaluating the performance of protein compression algorithms. Although past studies reported compression rates of protein sequences with varying degrees of success, there are wrongly stated claims and confusing results in some standard publications arising from inappropriate comparison of the data sets. To emphasize the difference and similarity of the data sets, the content of the files in the two protein corpuses are compared with respect to the size in bytes and repetitions of amino acids. In addition, comparison is made based on difficulty of compressing the files in the corpuses. The results indicate that the two protein corpuses possess different regularities. Besides, nine general purpose compression algorithms outperform the results reported by biological compressors on one of the corpus and comparable results on the other corpus.**

***Index Terms*—Protein corpus, compression rate, protein compression, biological compressors, general purpose compressor.**

I. Introduction

Life scientists face tough challenges in that large amounts of scientific data need to be analyzed and queried [1]. Dealing with large amounts of scientific data is closely tied to the advances of compression algorithms. The algorithms are required to have good performance evaluated on standard data set. Canterbury text corpus [2] and protein corpus [3] are two examples of standard data sets used to compare the performance of text and protein compression algorithms respectively. The focus of this paper is given to protein data sets.

The ultimate goal is to compare two protein data sets that have been used in most international publications on protein compression. The first data set, which is named in this paper as Protein Corpus 1 (PC1), was used in [4], [5] and [6]. The second data set, named as Protein Corpus 2 (PC2), was used in [7] and [8]. PC1 and PC2 contain a set of four files. These files are Haemophilus Influenzae (HI), Homo Sapiens (HS), Methanococcus Jannaschii (MJ), Saccharomyces Cerevisiae (SC).

To accomplish the goal, the numbers of amino acids in each file are identified and the contents of each corpus are analyzed to assess similarity and difference of PC1 and PC2. Then general purpose compressors are applied to the two data sets so as to compare the results with the rates reported on biological compressors. In the process the extent of compressibility of the data sets are observed. The compression rates obtained by applying general and biological compressors on both data sets are also compared with Shannon entropy base line.

Shannon source coding theory [9] sets fundamental limits on the performance of data compression algorithms. The theory, in itself, does not specify exactly how to design and implement these algorithms. It does, however, provide some hints and guidelines on how to achieve optimal performance. The size of the alphabet for protein consists of 20 symbols (A C D E F G H I K L M N P Q R S T V W Y). Thus the base line of protein entropy rate is $\log_2 20 = 4.32192$ bits per character (bpc). According to Shannon, the best possible lossless compression rate is the entropy rate.

The paper is organized as follows. The compression algorithms are discussed in section II. The results are presented in section III and some conclusions are drawn in section IV.

II. Algorithms

The difficulty of compressing PC1 and PC2 are tested by applying ten general purpose compressors to the data sets. The results are compared with compression rates reported by seven biological compression algorithms. To simplify the document the details of the algorithms are not discussed. For readers interested to refer further, supportive references and websites are indicated.

A. General Purpose Compressors

Unless specified in brackets as (setting: option), the results for general purpose algorithms are based on their default settings. The algorithms are listed as follows: block reduce compressor (bred3) [11, 12], Burrows-Wheeler compressor version 0.99 (bwc v0.99) [13], Nanozip alpha version 0.08 (setting: Nz_cm) [14], ash07 [15], Advanced Block sorting Compressor version 2.4 (ABC v2.4) [16, 17], context tree weighting version 0.1 (CTW v0.1) [18, 19], FreeArc 0.666 [20], ZZIP v0.36c [21], WinRK 3.1 (setting: best symmetric) [22], WinRAR 3.51 [23].

B. Biological Compressors

The biological compressors are listed as follows: CP [4], PPM [4], LZ-CTW [5], XM [6], ProtComp [7], Adjeroh et al. [8] and Model 3 of Benedetto et al. [10]. The compression rates of ProtComp, Model 3 of Benedetto et al., LZ-CTW, XM, CP and PPM on PC1, and Adjeroh et al. and ProtComp on PC2 are obtained from respective publications and author's websites.

III. Experimental Results

The experiment is performed on a workstation with Intel Pentium Dual CPU T3400 2.16GHZ 2.17GHz and 2.00GB of RAM. The operating system is Microsoft Windows Vista

Manuscript received April 1, 2012; revised May 5, 2012.

W. Diribi is with Department of Electrical and Computer Engineering, Addis Ababa Institute of Technology, Addis Ababa University, Addis Ababa, Ethiopia (e-mail: wedlovmam@gmail.com).

K. Raimond is with the Department of Computer Science and Engineering, Karunya University, Coimbatore, India (e-mail: kumudharaimond@yahoo.co.in).

Home Premium version 6.0.6001 Service Pack 1 Build 6001. The compression results are calculated from the size of real encoded files. The compression rates are reported in bpc.

Table I compares the number of amino acids, the content of the sequences and consecutive maximum repetitions of amino acids in the sequences of PC1 and PC2. Note that the naming of the four protein sequences (HI, HS, MJ and SC) are the same for PC1 and PC2. The protein sequences of PC1 and PC2 have exactly the same number of amino acids (509519, 3295751, 448779 and 2900352 for HI, HS, MJ and SC respectively). This is the cause of ambiguity. The four files in PC1 and PC2 with the same name and equal size may leads to wrong conclusion of thinking as if the files are identical. The size of the file in bytes is equal to number of amino acid. The content of the files and repetitions of amino acids in the files are, however, different. For instance, the first ten amino acids for HI in PC1 and HI in PC2 are MAIKIGINGF and RHMAQTSLWF respectively.

Similarly the other combinations also reflect different ordering of amino acids. Maximum repetition refers to the maximum number of consecutive amino acids in the sequence, for instance, AAA…, CCCCC…, DDDD. The maximum repetitions of amino acids seen in the four files of PC1 are more than that of PC2. It can be concluded from table I that the statistical distribution and repetitions of amino acids in the four sequences of PC1 and PC2 are different.

Table II shows the compression rates in bpc for HI, HS, MJ and SC in PC1. From the table II results, three biological compressors (XM, ProtComp, Model 3 of Benedetto et al.) and two general purpose compressors (Nz_cm, WinRK 3.1) offer excellent compression rates in the range of 3.99 to 3.94 bpc (note that small values imply better performance). XM and ProtComp are the two best compressors on the four protein sequences in PC1. Besides fourteen compressors are able to compress the files in PC1 better than the base line of protein entropy ($\log_2 20$ = 4.32192 bpc). This implies that protein sequences are indeed compressible with better compression rates. PPM and WinRAR 3.51 fail to compress the four files in PC1 to less than the entropy base line.

As can be seen from table III, nine general purpose compressors are performing much better than biological compressors on the four files of PC2. All general purpose compressors and two biological compressors are able to compress the files in PC2 better than the base line of protein entropy. The performances of biological compressors (ProtComp and Adjeroh et al.) are bad compared to the general purpose algorithms on PC2.

The results of table II and III show that PC1 is difficult to compress than PC2. Note that rows labeled colored in table II and III show the results of biological compressors.

The time performance in table IV is performed to emphasize the difference in distribution of amino acids in PC1 and PC2. Although the size in bytes and number of amino acids in the two corpuses are the same, the time performance of the algorithms in the respective files of PC1 and PC2 are different. A significant difference is observed in compression times of ABC v2.4 and ZZIP v0.36c. The compression times in seconds of ABC v2.4 on HI, HS, MJ and SC are respectively 0.581, 4.353, 0.534 and 3.762 for PC1 and 81.858, 53.506, 23.419 and 280.237 for PC2. Similarly compression times of ZZIP v0.36c on HI, HS, MJ and SC are respectively 0.68, 3.52, 0.6 and 3.15 for PC1 and 353.60, 209.88, 97.23 and 435.68 for PC2. This significant difference in compression time could be an indication that the two corpuses posses different statistical distribution of amino acids.

IV. Conclusion

According to the tabular results, there is an important difference in compressibility rates, compression time and statistical distribution of amino acids in the files of the two protein corpuses.

The results obtained in table II and III reflect the fact that PC1 is more difficult to compress than PC2 regardless of contrary claim of past work.

On the files in PC1 the biological compressors perform marginally better than general purpose compressors (table II) but on the files in PC2 almost all general purpose algorithms outperform the biological compressors (table III).

TABLE I: Number of Amino Acids, the Contents of the Sequences and Maximum Repetitions of Amino Acids in the Files of Pc1 And Pc2

Protein Sequences	Number of Amino Acids		The First Ten Amino Acids		The Last Ten Amino Acids		Maximum Repetitions	
	PC1	PC2	PC1	PC2	PC1	PC2	PC1	PC2
HI	509,519	509,519	MAIKIGINGF	RHMAQTSLWF	AMLIQQLLAK	ACFAREPDEW	7	5
HS	3,295,751	3,295,751	NMALLVGLLV	TFPFSDPDKY	IIYIHLRKRE	EPNACTTVLM	42	8
MJ	448,779	448,779	MSYFSLTEFA	FMLVLVFSAG	LLEMCKRIGK	GRTVFFPELL	9	6
SC	2,900,352	2,900,352	MSITNGTSRS	GWKMGVELWD	CRDSSREVGE	SECRADGPEL	45	6

The ranking of the algorithms are also different on the two corpuses. For instance, ash07 is the best general purpose compressor on the files of PC2 (table III) but it is ranked seventh on the files of PC1 (table II). Similarly ProtComp is the second best performing biological compressor on the files of PC1 (table II) but it is ranked eleventh on files of PC2 (table III).

The compression rates of general purpose compressors on PC2 are in the range of 2.81157 to 2.0902 (table III). This range is too far from the base line of protein entropy.

TABLE II: COMPARISON OF PROTEIN COMPRESSION IN BITS PER CHARACTER ON THE 4 FILES OF PC1 (SMALLER VALUES IMPLY BETTER PERFORMANCE)

Compressors	HI	HS	MJ	SC	Average
XM	**4.1022**	**3.7860**	**4.0002**	**3.8850**	**3.9434**
ProtComp	**4.108**	**3.824**	**4.008**	**3.938**	**3.9695**
Nz_cm	4.12078	3.84216	4.02200	3.91476	3.97492
WinRK 3.1	4.14197	3.82424	4.03710	3.90711	3.97760
Benedetto et al. (Model 3)	**4.10**	**3.93**	**4.00**	**3.95**	**3.99**
LZ-CTW	**4.1177**	**4.0055**	**4.0279**	**3.9514**	**4.0256**
ash07	4.16956	3.96390	4.06671	3.99098	4.04778
ABC v2.4	4.16298	4.08523	4.06377	4.10101	4.10324
CTW v0.1	4.14734	4.09287	4.05146	4.13928	4.10773
CP	**4.143**	**4.112**	**4.051**	**4.146**	**4.113**
FreeArc 0.666	4.26833	4.06500	4.21436	4.15299	4.17517
ZZIP v0.36c	4.29182	4.21310	4.23543	4.25602	4.24909
bwc	4.29344	4.21481	4.23884	4.26256	4.25241
bred3	4.31187	4.39292	4.25294	4.30404	4.31544
$\log_2 20$	4.32192	4.32192	4.32192	4.32192	4.32192
WinRAR 3.51	4.66218	4.35157	4.56802	4.40331	4.49627
PPM	**4.881**	**4.639**	**4.734**	**4.854**	**4.777**

TABLE III: COMPARISON OF PROTEIN COMPRESSION IN BITS PER CHARACTER ON THE 4 FILES OF PC2 (SMALLER VALUES IMPLY BETTER PERFORMANCE)

Compressors	HI	HS	MJ	SC	Average
ash07	1.20703	2.92425	1.83117	2.39832	2.09019
Nz_cm	1.20932	3.01506	1.83700	2.45147	2.12821
FreeArc 0.666	1.22243	3.55218	1.86106	2.43171	2.26684
WinRK 3.1	1.32225	3.33753	2.02148	2.70031	2.34539
WinRAR 3.51	1.39786	3.50806	2.13207	2.83875	2.46918
CTW v0.1	1.47170	3.34080	2.17581	3.00346	2.49794
ABC v2.4	1.67101	3.22614	2.43298	2.97286	2.57574
ZZIP v0.36c	1.66257	3.46579	2.29757	3.17956	2.65137
bwc	1.70140	3.60008	2.41977	3.52505	2.81157
Adjeroh et al.	2.55	3.41	2.27	3.11	2.84
ProtComp	2.34	3.91	2.87	3.44	3.14
bred3	4.22984	4.11436	4.16828	4.17338	4.17146
log220	4.32192	4.32192	4.32192	4.32192	4.32192

TABLE IV: COMPARISON OF TIME IN SECONDS TAKEN TO COMPRESS THE FILES OF PC1 AND PC2 BY GENERAL PURPOSE ALGORITHMS

Compressors	PC1				PC2			
	HI	HS	MJ	SC	HI	HS	MJ	SC
Nz_cm	2.97	18.24	2.67	15.84	2.86	19.35	2.47	16.19
ABC v2.4	0.581	4.353	0.534	3.762	81.858	53.506	23.419	280.237
ash07	6.661	43.867	5.850	39.687	6.428	41.871	5.631	39.172
CTW	10.5	89.1	9.6	91.2	7.3	81.0	7.2	73.1
ZZIP v0.36c	0.68	3.52	0.6	3.15	353.60	209.88	97.23	435.68

As a result, PC2 is not enough to evaluate protein compressibility. However, the compression rates on PC1 (table II) are at the vicinity of the baseline entropy. This implies that PC1 is preferable to PC2 to evaluate protein compressibility.

From the results of table II and III, majority of the compressors compress PC1 and PC2 at a compression rate better than the protein baseline entropy. This implies that protein is indeed compressible as concluded in earlier works [5, 6, 7].

It can generally be concluded that PC1 is better than PC2 to test protein compressibility. We, however, recommend testing protein sequence compressibility on both data sets for better result.

REFERENCES

[1] H. Tian, R. Sunderraman, I. Weber, H. Wang, and H. Yang, "A protein structure data and analysis system," *Proc. of the IEEE Engineering in Medicine and Biology 27th Annual Conference*, pp. 2847 – 2850, Shanghai, China, Sept. 2005.

[2] ftp.cpsc.ucalgary.ca:/pub/text.compression.corpus/text.compression.corpus.tar.Z.

[3] http://www.data-compression.info/Corpora/ProteinCorpus/index.html

[4] C. G. Nevill-Manning and I. H. Witten, "Protein is incompressible," *Proc. of the Data Compression Confrence* (DCC'99) , pp. 257–266, Snowbird, Utah, USA, March 1999.

[5] T. Matsumoto, K. Sadakane, and H. Imai, "Biological sequence compression algorithms," *Genome Informatics*, pp. 43 – 52, 2000.

[6] M. D. Cao, T. I. Dix, L. Allison, and C. Mears, "A simple statistical algorithm for biological sequence compression," *Proc. of Data Compression Conference* (DCC'07), pp. 43 – 52, Snowbird, Utah, USA, March 2007.

[7] A. Hategan and I. Tabus, "Protein is compressible," *Proc. of the 6th Nordic Signal Procecessing Symp*. (NORSIG'04), pp. 192–195, Espoo, Finland, June 2004.

[8] D. Adjeroh and F. Nan, "On compressibility of protein sequences," *Proc. of Data Compression Conference* (DCC'06), pp. 422–434, Snowbird, Utah, USA, March 2006.

[9] C. E. SHANNON, "A Mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379 – 423 and 623 – 656, Oct. 1948.

[10] D. Benedetto, E. Caglioti, and C. Chica, "Compressing proteomes: the relevance of medium range correlations," *EURASIP Journal on Bioinformatics and System Biology*, July 2007, doi:10.1155/2007/60723.

[11] M. Burrows and D. J. Wheeler, "A block-sorting lossless data compression algorithm," *Research Report* 124, Digital System Research Center, Palo Alto, California, 1994.

[12] D. Wheeler, "An implementation of block coding," *Computer Laboratory*, Cambridge University, England, October 1995.

[13] ftp://ftp.stack.nl/pub/users/willem/

[14] http://www.nanozip.net

[15] http://ctxmodel.net/files/ASH/

[16] http://www.data-compression.info/ABC/index.htm

[17] Juergen Abel, "Improvements to the Burrows-Wheeler compression algorithm: After BWT stages," University Duisburg-Essen, 2003

[18] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context-tree weighting method: basic properties," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 653–664, 1995.

[19] http://www.ele.tue.nl/ctw/manual.html

[20] http://www.freearc.org/Download.aspx

[21] http://www.debin.net/zzip/

[22] http://www.softpedia.com/get/Compression-tools/WinRK.shtml

[23] http://www.rarlab.com/

Wedajo Diribi Feyissa received his post graduate degree in Computer Engineering from Addis Ababa University, Addis Ababa, Ethiopia in 2011; and BSc degree in electrical engineering from Arba Minch University, Arba Minch, Ethiopia in 2007. He worked as lecturer in electrical engineering department of Arba Minch University

At present, he is pursuing second level specialization in wireless systems and related technologies at Polytechnic University of Turin, Torino, Italy. He became a member of International Association of Computer Science and Information Technology (IACSIT). He is a reviewer of International Journal of Computer and Electrical Engineering (IJCEE).

Dr. Kumudha Raimond received her B.E from Madras University and M.E from Government College of Technology, Coimbatore and Doctoral degree from Indian Institute of Technology, Madras, India. She is having thirteen years of teaching experience along with three years of industrial experience in General Electric, India.

She was working in Electrical and Computer Engineering Department, Addis Ababa University, Addis Ababa, Ethiopia. At present, she is working in Karunya University, India. Her research interests are intelligent systems, adhoc protocols, wireless sensor networks, image processing, compression, watermarking and biometric, biomedical and bioinformatics applications. She is a member of International Association of Computer Science and Information Technology (IACSIT) and Machine Intelligence Research Lab.