

A Block Coding Method that Leads to Significantly Lower Entropy Values for the Proteins and Coding Sections of *Haemophilus influenzae*

G. Sampath

Department of Computer Science, The College of New Jersey, Ewing, NJ 08628

sampath@tcnj.edu

Abstract

A simple statistical block code in combination with the LZW-based compression utilities *gzip* and *compress* has been found to increase by a significant amount the level of compression possible for the proteins encoded in *Haemophilus influenzae*, the first fully sequenced genome. The method yields an entropy value of 3.665 bits per symbol (bps), which is 0.657 bps below the maximum of 4.322 bps and an improvement of 0.452 bps over the best known to date of 4.118 bps using Matsumoto, Sadakane, and Imai's *lza-CTW* algorithm. Calculations based on a compact inverse genetic code show that the genome has a maximum entropy of 1.757 bps for the coding regions, with a possibly lower actual entropy. These results hint at the existence of hitherto unexplored redundancies that do not show up in Markov models and are indicative of more internal structure than suspected in both the protein and the genome.

1. Introduction

The rapid increase in available biological sequence data in several publicly accessible databanks [14-19] and the use of advanced distributed computing methods [6] have led to the use of increasingly complex computational methods designed to extract biological, chemical, and physical correlates from those sequences. On the one hand, the growing volume of sequence data available makes it possible to search for deep structural properties in the sequences and relate them to biological function at several levels, from the molecular to the cellular, in ways that are more promising and statistically more significant than was possible earlier with smaller data sets. Thus new algorithms have been designed explicitly for this purpose or existing ones modified to work with biological data [10], and special data structures designed to work efficiently with the massive amounts of data available [13]. On the other hand, there is a growing interest in reducing the amounts

of storage required to cope with the exponential rise in the sizes of the databanks. This has led to the development or adaptation of a wide range of data compression techniques [11, 12], new and old, tailored to the task of compressing genome and protein sequences. Currently, however, the results obtained from the application of these methods appear only to reinforce the long-standing belief that biological sequences (which appear to be random [3] even though they have of necessity hidden regularities [7]) are not compressible in any significant way. Such a belief, which is reflected in the title of a recent communication [9], stems in part from the fact that the vast majority of attempts to compress biological sequences have resulted in entropies that are not very far from those predicted by information theory [1]. Hidden patterns in most sequences studied seem to be difficult to extract even with the most painstakingly designed compression schemes [2, 4-9]. In spite of this negative result, interest in biological sequence compression has not slackened because any gains, however small, are likely a reflection of constraints in internal structure. Such structure, which may not otherwise be evident in a sequence, may help understand the physical and chemical properties of the macromolecules coded by it.

The purpose of the present communication is to report on a significant amount of compression achieved with the proteins coded by *Haemophilus influenzae* (*Hi*) (the first genome to be fully sequenced [14]). An appreciable amount of compression has also been obtained with the genome itself, although it is not as large as with the proteins. One possible consequence of this result is that it holds out hope for similar successes with other sequences, perhaps through the use of the methods outlined here or with new approaches. Thus, further investigation can help determine if the compression achieved with *Hi* is to be regarded as a rare occurrence (which would make *Hi* unique, itself a subject worth studying) or significant decreases in entropy are possible with other proteins and genomes.

In Section 2, a brief introduction is given to the data compression problem as it affects biological sequences

and its relation to entropy measures. In Section 3, a block code is developed for the proteins coded by H_i based on their statistical distribution. Section 4 describes a procedure to compress the protein sequences and presents results comparing its performance with that of other methods as reported in the literature. Section 5 presents an efficient way of mapping residues (that is, amino acids) in the protein sequence to the codons (triplets) in the parent DNA that leads to a non-trivial decrease in entropy of the latter. Section 6 offers some conclusions and directions for future work.

2. Biological sequence compression, information theory, and the Shannon limit

Almost all known biological functions can be traced to the sequence of nucleotides or bases known as DNA that is found in the cells of every organism. The base sequence in the DNA is transcribed one-to-one in the cell to another sequence of nucleotides known as RNA (whose bases are identical to the bases of DNA with one small difference, noted below). The latter in the presence of a number of catalysts is translated into a set of proteins, which are sequences of amino acids that contribute to a host of cellular and higher-level functions of the organism. The 'genetic code' used in the translation maps a triplet of bases or 'codon' to an amino acid. The code is degenerate with 61 of the 64 possible codons mapping to 20 amino acids (the remaining three coding for terminators). The entire DNA sequence for an organism is called a 'genome', it consists of segments of coding regions and non-coding regions (referred to as 'introns'). A coding region that codes for a single protein is called a 'gene'. In most species, the non-coding regions are far greater in size than the coding regions. For example, the human genome consists of about 3 million bases arranged in 23 pairs of 'chromosomes' each of which contains a number of related genes. Only about 10% of the genome is coding, the other 90% is non-coding with most of it known or suspected to be involved in the chemistry of transcription and translation.

At each of the three levels of information encoding in a cell, the macromolecules involved (DNA, RNA, and protein) can be viewed as sequences of symbols (where a symbol corresponds to a base) in an alphabet, and the transformation steps as string operations. In information theory [1], a sequence is a string on an alphabet whose information content can be related to its entropy. The entropy E of a sequence is closely related to the alphabet on which the sequence is based and is defined as

$$E = - \sum_{i=1}^N p_i \log_2 p_i \quad \text{bits per symbol (bps)}$$

where p_i is the probability (or normalized frequency) of occurrence in the sequence of the i -th symbol in an alphabet of N symbols. When the probabilities are all equal and equal to $1/N$, the entropy is simply $\log_2 N$ bps. This is commonly referred to as the *Shannon entropy* of the alphabet. DNA's alphabet is the set {T, C, A, G} (corresponding to the bases thymine, cytosine, adenine, and guanine), while RNA's is {U, C, A, G} (corresponding to the bases uracil, cytosine, adenine, and guanine). The protein alphabet is a set of 20 amino acids {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}, where the letters in the set are the 1-letter codes for alanine, cysteine, aspartic acid, glutamic acid, phenylalanine, glycine, histidine, isoleucine, lysine, methionine, leucine, asparagine, proline, glutamine, arginine, serine, threonine, valine, tryptophan, and tyrosine, with corresponding three-letter codes ALA, CYS, ASP, GLU, PHE, GLY, HIS, ILE, LYS, LEU, MET, ASN, PRO, GLN, ARG, SER, THR, VAL, TRP, and TYR. The Shannon entropy values for the three alphabets are respectively 2, 2, and 4.322 bps.

The entropy as defined above is a first order measure that measures redundancy at the single character level, the limit of which is the equiprobable or Shannon entropy. There may also be higher-order dependencies among the symbols that can lead to greater redundancy in a sequence, this can be measured using a more general definition of entropy that is based on viewing a sequence as a discrete-time stochastic process [7]. Such a generalized entropy measure provides a quantitative measure of redundancy in biological sequences that can sometimes be correlated to physical properties such as, in the case of RNA and protein, the shape that the macromolecule attains in space. When the sequence is random, it tends to have symbols that occur with the same frequency, and hence leads to an entropy close to the Shannon value. From the information theoretic point of view, the farther away (towards 0) the measured entropy is from the Shannon value, the greater the redundancy and hence the lower the information content.

Biological sequences can also be approached from the point of view of their storage requirements: the presence of redundancies indicates the possibility of using less storage. Thus the entropy of a sequence can also be viewed as the number of bits required to store a symbol, this number being computed as an average over the whole sequence. The lower the entropy, the less the storage required. One can then consider ways of recoding the sequence to yield a compressed form, whose length in turn is a measure of the redundancy in the sequence. While biological sequences appear to be random [3] the fact that only a few thousands of proteins have been found in cells suggests that DNA is a 'purposeful' molecule [7]. Thus not every combination of bases is a valid code sequence, which implies hidden

regularities that can lead to a viable protein and therefore a lower entropy value. Likewise not every amino acid sequence can be a protein because of physical constraints on the shape that a protein can attain in space. All of these considerations have led to concerted attempts at compressing biological sequences using a wide range of available algorithms. However, such attempts have generally had limited success. The measured entropies of the sequences that result have either been not far from the Shannon limit or, worse, in many cases higher, with the original sequence expanding in size. For example, the third chromosome in yeast has about 300000 bases and is very resistant to compression [4], a large number of proteins show a similar resistance [4]. This resistance of both DNA and protein to compression has led to the use of more advanced pattern recognition methods and complex Markov models [2, 4-9] that examine a sequence for patterns and dependencies between a symbol and the symbols preceding it in a variable width window. When applied to the proteins encoded by a set of fully sequenced genomes the resulting decreases in entropy are on the order of 0.2 to 0.3 bps [7-9]. With the genome *Haemophilus influenzae* (*Hi*), the lowest entropy obtained so far has been 4.118 bps using the *lza-CTW* algorithm of Matsumoto, Sadakane, and Imai [8], which is 0.204 bps below the Shannon limit, the next best being 4.143 bps using Nevill-Manning and Witten's *cp* algorithm [9]. In comparison, in the work reported here an entropy level of 3.665 bps (which is 0.657 bps below the maximum) has been obtained for the proteins of *Hi*. The details are given in the following sections.

3. A block code for the proteins coded by *Haemophilus influenzae*

Compression of the proteins coded by the genome sequence of *Haemophilus influenzae* is based on a procedure that codes each residue with a 2-part code.

3.1 Partitioning the amino acids

First a frequency distribution of occurrence of the 20 amino acids in the proteins is obtained and sorted in increasing order. Let this sorted distribution be $\mathbf{F} = [f_1, \dots, f_{20}]$ and the corresponding amino acid array $\mathbf{A} = [a_1, \dots, a_{20}]$. \mathbf{A} is divided into N groups $\mathbf{G} = [\mathbf{G}_1, \dots, \mathbf{G}_N]$ where $\mathbf{G}_i = [i_1, \dots, i_k]$ and $\mathbf{G}_i \leq \mathbf{G}_{i+1}$. Here \leq is a total ordering defined on the groups: the frequency of occurrence of every residue in \mathbf{G}_i is less than or equal to that of every residue in \mathbf{G}_{i+1} , $1 \leq i \leq N-1$. A residue in the protein is then coded with a number g , $1 \leq g \leq N$, corresponding to the group \mathbf{G}_g that it belongs to.

3.2 Coding the elements in each partition

Next the same residue is assigned a number p that codes the position of the corresponding amino acid in the group. If $|\mathbf{G}_g|$ is 1, this number is omitted as there is no need to distinguish the single amino acid from another in the group.

Thus every one of the residues in the protein has a code pair (g, p) . (p may be absent, as mentioned above). A protein with L residues can then be coded as a sequence of such pairs. Alternately, it can be represented by two strings: a group membership string $\mathbf{g} = g_1 g_2 \dots g_L$ and a position string $\mathbf{p} = p_1 p_2 \dots p_L$, where L' can be $< L$ because, as noted above, an amino acid may be the only one in its group in the partition.

4. Compressing the proteins coded by the *Haemophilus influenzae* genome

It is shown next that compression can be achieved by applying the well-known utilities *compress* and *gzip* separately to the \mathbf{g} and \mathbf{p} strings. The value g_i in the i -th pair can be stored in $\lceil \log_2 N \rceil$ bits. The value p_k ($k = 1, \dots, L'$) requires s_k bits, where

$$0 \leq s_k \leq g_{\max} = \max \{ \lceil \log_2 |\mathbf{G}_1| \rceil, \lceil \log_2 |\mathbf{G}_2| \rceil, \dots, \lceil \log_2 |\mathbf{G}_N| \rceil \}.$$

The storage requirement (in bits) for the protein is then

$$L \lceil \log_2 N \rceil + \sum_{k=1}^{L'} s_k$$

The partitioning of the amino acid array \mathbf{A} as described in Section 3.1 is done to reflect the frequencies of occurrence of the amino acids in the protein. This is effectively a Huffman-like code in which the more frequently occurring amino acids are in smaller groups, which require fewer bits to encode membership in the group. To avoid wasting bits in coding g , it is of advantage to choose N as a power of 2. For example, the 20 amino acids can be divided into 8 groups of i_1, \dots, i_8 amino acids, with the frequencies of occurrence decreasing to the right in the list, or 4 groups i_1, \dots, i_4 . Then the g in each residue's code pair (g, p) is coded with 3 or 2 bits respectively. The sizes of the groups in a partition can be chosen similarly, but choosing a power of 2 for each $|\mathbf{G}_i|$ once again tends to give the best results. For example, the partition could be $\{1, 1, 2, 2, 2, 2, 4, 4, 4\}$ in the first case, or $\{4, 4, 4, 8\}$ in the second. Experiments with the protein sequence of *Hi* led to the first of these choices as the one giving the best results. The decoding is straightforward: the group for the i -th residue is the i -th symbol in the \mathbf{g} file, and its position in the group is either not required (if the group has only

one member) or the value corresponding to the next symbol in the **p** file.

The individual strings **g** and **p** are now compressed using a number of available utilities. Table 1 shows results obtained on a Sun Ultra-5 workstation with a 500 MHz processor and 512 Mbytes running Solaris. Data for *Hi* were taken from the Protein Corpus [18]). The block coding time was negligible and the compression time very small, neither was recorded. In the latter case, it ranged from near zero with *compress* to a few seconds with *gzip* (switch -9). The *gzip* version used was 1.2.4 (dated 18 August 1993). The version number for *compress* is not available (*compress* is not stable as it changes with the OS), the version used was the one in effect in April 2003.

Since the alphabet for each of the strings is limited to symbols used to represent small integers one would expect their behavior with compression algorithms to be similar to that shown by DNA (with its alphabet size of 4). However the outcome is different from the expectation. Because the decoding is to be done after decompression, the **g** and **p** files can be compressed separately and by algorithms that are different. As seen in Table 2 below, this mixing and matching leads to an unexpected drop in the total size of the compressed data when **p** is compressed with *gzip* and **g** with *compress*. (A small amount of storage, less than 50 bytes, that is used for 11 gene region separators in the source, is not included.)

Table 3 (adapted from [8]) gives an idea of how significant the change in entropy from the Shannon value is in comparison with values reported for other methods. (Entropy values above the Shannon value are italicized, that for the current method is in bold.)

5. Using compressed proteins to compress DNA

Attempts to compress DNA sequences have generally been unsuccessful. Although there have been instances [7, 8] in which entropies as low as 1.1048 have been found, these lower values are usually obtained with sequences that contain long non-coding segments (which are either introns or so-called 'junk' DNA). For coding segments, values in the range 1.84 through 1.95 are more typical [7]. Based on empirical data available, it is generally the case that the coding sections of a genome are very resistant to compression, while the intron/'junk' segments appear to be more susceptible. This is often attributed to the small alphabet and the tendency of coding DNA sequences to be uniformly distributed on the alphabet. (It is widely believed that this uniform distribution is a result of DNA having evolved over a long period of time, a consequence, perhaps, of the third 'law' of thermodynamics.) But here too *Hi* runs counter to the general behavior of DNA. The following paragraphs show how the coding sections of *Hi* can be compressed to 1.757 bps.

File	File size (bytes)
Protein file (see [18])	509519
g file	509508
p file	453560
g compressed with <i>gzip</i> (switch -9) A	205192
p compressed with <i>gzip</i> (switch -9) B	45485
g compressed with <i>compress</i> C	187836
p compressed with <i>compress</i> D	117301

Table 1. Results of compressing the block coded files g and p

Shannon entropy for amino acids (bps)	4.321928
Shannon storage for protein file (bytes)	275140
Storage required by combining B and C from Table 1 (bytes)	233321
Entropy resulting (bps)	3.665030

Table 2. Entropy resulting from combining different compressed versions of g and p

Compressor	Entropy (bps)
(Shannon)	4.32198
<i>Compress</i>	4.7702
<i>Bzip2</i>	4.324
<i>Gzip -9</i>	4.6712
Arith	4.1557
lz-ari (1M)	4.1270
Normal PPMD+	4.862
Adapted PPMD+	4.151
CTW20(8)	4.1381
CTW20(16)	4.1378
lz-CTW(8)	4.1185
lza-CTW(8)	4.1177
CP(1)	4.149
CP(2)	4.146
CP(3)	4.143
Current method	3.665

Table 3. Comparison of results with earlier work (see [8] for details)

As mentioned in Section 2, the genetic code is degenerate, with more than one codon coding for an amino acid. The following table lists the number of codons for each amino acid, it also contains information on the reverse code needed to go from amino acid to codon as discussed in the next paragraph.

If the coding segments of a genome are translated to amino acid sequences, they can be recovered only if the codon coding for a residue is known. This requires storing a coded form of the source codon for each residue, which requires from 0 to 3 bits, leading to a third reverse code file **r** if the genome is to be compressed (see Table 4). If the storage required for the compressed sequence protein ($= |g| + |p|$, obtained from the translation of genome to protein sequence) plus the storage for **r** is less than the Shannon storage value of the genome sequence, then compression of the genome is considered to occur. Consideration must also be given to allocation of storage to the gene separators in the DNA so that the gene boundaries can be recovered during decompression. There are 1709 genes with 85% of them having a coding density of c. 1070 base pairs per gene [14]. With no compression and at 4 bytes per gene (2 for the length, 2 for the offset within the genome) this generates a fourth file **b** with 6836 bytes. (This is in

contrast with the Protein Corpus [18], where the objective of its designers seems only to be decreasing the entropy. The corpus stores the residues of *Hi* in the form of 12 strings corresponding to the 12 gene regions in the genome with a separator between two successive regions, rather than as individual proteins. Apparently this is based on the assumption that the gene boundaries are accessible through the parent DNA, which is considered to be available separately.) Table 5 shows the results from applying this method to the protein sequences from *hi*, with **r** and **b** stored uncompressed.

As noted in [7] if the fact that only 61 codons code for amino acids (the other three coding for terminators) is taken into account, then the Shannon entropy of DNA is $\log_2 61 / 3 = 1.977$ bps. Using this in the computation above, the entropy for the compressed translated genome is lower than the Shannon value by 0.22 bps, which, in the current state of DNA compression, is still an appreciable reduction. A slight improvement may occur if the reverse code for Isoleucine (I) is stored as a prefix code. The reverse code file **r** and the gene boundary file **b** also may be compressible, results are incomplete at this time. (At the very least, the gene boundaries do not need 4 bytes, 3 should suffice since most length and offset values are less than 2048.)

No of codons (X)	Amino acids	No of amino acids	Reverse code size = $\lceil \log_2 X \text{ bits} \rceil$
1	M, W	2	0
2	F, Y, C, H, Q, N, L, D, E	9	1
3	I	1	2
4	V, P, T, A, G	5	2
6	L, S, R	3	3

Table 4. Reverse code requirements for the genetic code

Compressed storage for <i>Hi</i> proteins	A	233321 bytes
Reverse code for 509519 residues	B	96144 bytes
Gene boundary information (file b)	C	6836 bytes
Total storage required for genome	D	336301 bytes
Shannon storage required for 3×509519 bases (assuming 2 bps for DNA)	E	382889 bytes
Entropy of compressed translated genome with reverse codes and gene boundaries	F	1.757 bps
Reduction in entropy level	G	0.243 bps

Table 5. Results of compressing the coding sections of *hi*

At the present time, the method does not seem to be a general one, but studies are ongoing, with results inconclusive at this time. The method when applied in its original form to the proteins coded by the three other genomes in the Protein Corpus [18] (being the two fully sequenced ones, *Methanococcus Jannischii* and *Saccharomyces Cerevisiae*, and the third partially sequenced one, *Homo Sapiens*) yielded negative results (the sequences expanded). Modified versions are under development with tests underway on these genomes as well as a range of other proteins in various protein databases [14-17, 19]. Firmer conclusions about their compressibility or otherwise should be possible after results are available.

6. Conclusions and directions

The following are some conclusions that can be drawn from the above study:

- 1) It may not be reasonable to infer from the failure or success of a compression method working on one or more sequences that it will fail or succeed with others [9]. Attempts to devise general compression schemes are often frustrated because an algorithm that works with a genome or protein may not work with another (and vice versa). As seen above, it is possible that existing methods in combination or suitably modified (or perhaps in conjunction with others to be devised) might succeed in lowering the entropy of one or more protein sequences or their parent genes, or both, which may have been resistant to known compression schemes. One likely cause for the failure of compression schemes to lower the entropy by significant amounts when applied to sequences across the board is the wide diversity present at the molecular level. A protein or genome (or a small set of related ones) may therefore have to be matched with a compression scheme specific to it. This further suggests classifying biological sequences based on their compressors, a possibility that is currently being investigated.
- 2) If tests on a wider range of data show that the method described here works only on *Hi* but not on other

sequences, it would perhaps make *Hi* unique, which then raises the question what makes it so.

3) Based on the method described above a sufficient level of compressibility in protein would imply compressibility of the parent DNA as long as the **r** and **g** files are not too big (comparatively speaking).

4) The fact that *Hi* has significantly lower entropy than the maximum (for both the proteins and the set of its encoding genes) implies a level of redundancy that does not show up in Markov models. This in turn hints at the existence of more internal structure than suspected at both the protein and the genome level and points to the need to look for its causes.

Acknowledgement. The manuscript was rewritten based on revisions suggested by the referees.

7. References

- [1] Ash, R. B. *Information theory*. Interscience Publishers, New York, 1965.
- [2] Chen, X., Kwong, S., and Li, M. "A compression algorithm for DNA sequences and its applications to genome comparison." *Genomic Informatics* **10**, 52-61, 1999.
- [3] Curnow, R. N. and Kirkwood, T. B. L. "Statistical analysis of deoxyribonucleic acid sequence data - a review." *Journal of Royal Statistical Society (Series A)* **152**, 199-220, 1989.
- [4] Grumbach, S. and Tahi, F. "A new challenge for compression algorithms: genetic sequences." *Information Processing and Management* **30**, 875-866, 1994.
- [5] Lanctot, J. K., Li, M., and Yang, E. "Estimating DNA sequence entropy." *Proceedings 11th Annual ACM-SIAM Symposium on Discrete Algorithms*, 409-418, 2000.
- [6] Larson, S. M., Snow, C. D., Shirts, M., and Pande, V. S. "Folding@Home and Genome@Home: Using distributed computing to tackle previously intractable problems in computational biology." In: *Computational Genomics*, R. Grant (ed.), Horizon Press, 2002.

- [7] Loewenstern, D. and Yianilos, P. N. "Significantly lower entropy estimates for natural DNA sequences." *Journal of Computational Biology*, **6** (1), 1999.
- [8] Matsumoto, T., Sadakane, K., and Imai, H. "Biological Sequence Compression Algorithms." *Genomic Informatics* **11**, 43-52, 2000.
- [9] Nevill-Manning, C. and Witten, I. H. "Protein is incompressible." *Proceedings of IEEE Data Compression Conference*, 257-266, 1999.
- [10] Pevzner, P. A. *Computational Molecular Biology: An Algorithmic Approach*. MIT Press, Cambridge (Mass.), 2000.
- [11] Sayood, K. *Introduction to Data Compression*. Morgan Kaufman, San Francisco, 1996.
- [12] Solomon, D. *Data Compression: The Complete Reference*. Springer-Verlag, New York, 1997.
- [13] J. S. Vitter. "External memory algorithms and data structures: dealing with massive data." *ACM Computing Surveys* **33**(2), 209-271, June 2001.

Sources on the Web

- [14] Center for Biological Sequence Analysis (CBS). <http://www.cbs.dtu.dk/index.html>, <http://www.cbs.dtu.dk/services/GenomeAtlas/Bacteria/Haemophilus/influenzae/Rd/>.
- [15] GenomeNet. <http://www.genome.ad.jp>.
- [16] National Center for Biotechnology Information (NCBI). <http://www.ncbi.nlm.nih.gov>.
- [17] Swiss-Prot. <http://us.expasy.org/srs5>.
- [18] Protein Corpus. <http://www.Data-Compression.info>.
- [19] TIGR Database (TDB). <http://www.tigr.org/tdb>.