Data Compression – An Overview and Trends in Genomics

Introduction
The magical touch of compression can be felt in many modern computer and communication technologies. As from email communications to the sharing of pictures through WhatsApp, in lightning speed is possible only because of compression. It is always time consuming to move large files as such over internet network and that even demands for a higher bandwidth. The best practice is to shrink the files by throwing away the redundant data. For instance as shown in Fig. 1, taking into advantage of human visual system, it is possible to reduce the size with varying levels of degradation of quality acceptable for different purposes.

Even in our day-today life knowingly or unknowingly we use compression. The arrangement of things in the best possible manner with the available space is also an example for compression. So, from a layman point of view, compression is the process of discovering structures that exist in the data [1]. Data compression is widespread in number systems, natural languages and even in mathematical notations. It plays a very important role in communications technology, especially the digital multimedia. Many portable practical developments like mobile computing, digital & satellite TV, computer systems such as memory structures & disks employ data compression. But then, centuries before the developments of technology, poets and painters used the principle of compression. The imaginative skill of artist helped to concise an elaborate message on a piece of paper with minimum words or images, which may actually require pages and pages of explanations. For example, a small extract of the well-known poem Elegy written in a country churchyard by Thomas gray is shown below in Fig. 2.

The poet describes that many blessed beauty in the nature is unseen since we limit ourselves to understand it. There are many intrinsic meaning be perceived in these lines. In general, inorder to understand the paintings or poetry, one should know the principle behind the artistic creation which is even true for the data compression technology. In short we can define data compression as the process of transforming a data from one representation to another so that it takes less storage space or less transmission time. Most of the real world data have inherent redundancy in the form of structural similarity or some hidden patterns. Exploiting these redundancies will help us to represent data in less number of bits. Hence, in this context, data compression may be viewed as an art of representing information in a compact form [2]. Even if the technology is improving for better mass storage system; the regular increase in data always urges a need for compression techniques.

A look back in history reveals the envisaging concept laid out by Claude E. Shannon in his famous 1948 landmark paper “A Mathematical Theory of Communication” helped to frame the concept of data compression which is inevitable in many fields of communication [23]. Annals shows the strong inspiration of Harley’s paper in proposing the mysterious concept of information. The term information has nothing to do with its meaning in common parlance. The intuitive ideas shed out by Shannon helped to relate surprise and information [3] which will be explained in detail in the subsequent sections. Shannon proposed even the limit at which a message can be transmitted from one end to another through channel without loss of information.

The abstract concept of information proposed by Shannon forms the foundation of all technological advancements, in the field of data storage and transmission systems.

Fig. 1: An example of Image compression for different picture resolution 409 KB, 37KB and 25KB (Photo: Lakshmy Gopalaswamy by Hareesh N)

Fig. 2: Data Compression Analogy with Poetry (Source: http://en.wikipedia.org/wiki/Elegy__Written_in_a_Country_Churchyard)

Fig. 3: Claude E. Shannon – Father of Information Theory (Source: http://www.nndb.com/people/934/000023865/)
Compression

Any source of information can be translated into an efficient representation using compression techniques for better storage and transmission. Compression may be lossy or lossless. If the compressed file can be reproduced exactly similar to the input file, then the scheme is called lossless compression. Text compression is an example for lossless compression. On the other hand, if the reconstructed file is not exactly as input file, then the scheme is lossy. Video compression is an example for lossy compression. Any compression algorithm consists of two stages as shown in Fig. 4. A source model, which describes the redundancy of given message followed by the selection of an optimal encoding technique for a much precise and smaller representation of the message[1].

Example 1:

\[
\text{T}_\text{h}_\text{e}_\text{s}_\text{s}_\text{,}_\text{n}_\text{t}_\text{i}_\text{.}_\text{M}_\text{s}_\text{s}_\text{i}_\text{.}_\text{g}_\text{.}_\text{C}_\text{l}_\text{a}_\text{.}_\text{d}_\text{e}_\text{.}_\text{S}_\text{h}_\text{.}_\text{n}_\text{.}_\text{n}_\text{.}_\text{r}_\text{n}_\text{.}_\text{a}_\text{d}_\text{.}_\text{T}_\text{h}_\text{.}_\text{M}_\text{.}_\text{k}_\text{.}_\text{.}_\text{n}_\text{g}_\text{.}_\text{f}_\text{.}_\text{n}_\text{f}_\text{o}_\text{r}_\text{m}_\text{.}_\text{l}_\text{.}_\text{i}_\text{n}_\text{.}_\text{T}_\text{h}_\text{.}\text{.}_\text{r}_\text{y}
\]

Even though a few letters are missing, still we will be able to read the text as “The essential missing Claude Shannon and the making of information theory”. As explained by Shannon, “any one speaking a language possesses an enormous knowledge of the statistics of the language. Familiarity with the words and grammar enables to fill in missing or incorrect letters[4]”.

This in fact form the fundamental of any compression algorithm.

Mathematically, information is inversely proportional to probability (\(p\)).

\[
I = - \log (1/N) = -\log P, \quad \text{…………………….(1)}
\]

In the subsequent sections, the commonly used compression term Information and Entropy is explained with a few analogies which is further followed by the current compression trends in the field of genomics.

Information

Information is a common term that we encounter in our daily life. The term has broadly been used in many different areas with many intuitive meanings, which generally create confusion. As per Shannon, the semantic aspect of communication is irrelevant[1]. All forms of messages like text, images, audio or video can be transmitted in two states like “yes” or “no”. Information may be then defined as minimum number of yes or no questions to determine the state like “on(1) or off(0)”. Any system, which is defined by two states has their fundamental atom as bits. Hence bit is the unit of information.

In the 1948 land mark paper Shannon quoted “If the number of messages in the set is finite then this number or any monotonic function of this number may be regarded as information”[2]. Thus, as proposed by Shannon, Information may be mathematically expressed as

\[
I = \sum_{i=1}^{k} \log p_i, \quad \text{…………………….(2)}
\]

The famous simple prediction game example[5] will certainly help one to understand the intuitive meaning of information. To state with, imagine any random number from 1 to 100. One can predict the number by asking logical yes or no questions. For example, one can reduce the search space by directly asking whether the number is less than 50. Now the search space reduced to one half, which actually increase our confidence to predict the number. Further, one can ask is it less than 25, so that again the search space is reduced. More logical questions like is it prime number or is it odd number, help us to predict the exact number. How can we connect this prediction game to information theory concepts? In the example the total possible combination of numbers is 100. Information is the logarithm of all possible combinations. Hence \(\log(100) = 6.6\), so nearly 7 questions are required to correctly guess the exact number. On the other hand, it is possible for us to say nearly 7 bit of information is present in the event.

As another example, consider a sequence from the source of 4 alphabets \(\{A, T, G & C\}\).
AATGGCACCT
Let p(A), p(T), p(G) & p(C) be probability of occurrence of A, T, G & C respectively.
\[ p(A) = \frac{2}{10} = 0.2 \]
\[ p(T) = \frac{3}{10} = 0.3 \]
\[ p(G) = \frac{2}{10} = 0.2 \]
\[ p(C) = \frac{3}{10} = 0.3 \]

The information content of A, T, G, C can be computed as \( I(A) = \log_2(A) = 2.32 \) bits, \( I(T) = \log_2(T) = 1.74 \) bits, \( I(G) = \log_2(G) = 1.74 \) bits and \( I(C) = \log_2(C) = 1.74 \) bits. Almost all symbols have equal probability, hence uncertainty is more. Each symbol have around 2 bit of self information and the total information content is
\[ I = \sum p_i \log_2 p_i = 8.17 \text{bits} \]

As another example consider the tossing of biased coin with \( P(H) = 1/5 \) and \( P(T) = 4/5 \), then \( I(H) = -\log_2(1/5) = 2.32 \) bits and \( I(T) = -\log_2(4/5) = 0.1 \) bits. The occurrence of tail is more, which means the event is almost certain, hence the self-information is low. The occurrence of head is low, hence self-information is high.

**Entropy**
Entropy is a measure of uncertainty or lack of information. It denotes “the amount of surprise created on us”\(^{10}\). For instance, the following news “School locks up UKG student in dog house” certainly create more surprise than “School locks up a dog in dog house”. Since the first incident is rare to happen, the number of bits required to encode is more compared to the second news.

In general, Entropy is the average amount of information produced from the event. It intuitively provides the number of bits per symbol actually required storing data. Thus entropy provides a bound for lossless compression. Mathematically, Entropy (H) is weighted average of the probability \( (p_i) \) of occurrence of all possible events.
\[ H = \sum p_i \log_2 (p_i) \]

As mentioned earlier, the significance of Information entropy is that it tells us the minimum number of bits required to encode the message digitally. This would mean that if one measures the entropy of a message, he can know if there is a scope for compression of that message. The number of bits required to represent English text, if all letters and space are considered to have the same probability, is \( \log_2(27) = 4.75 \) bits. But the underlying structure of English language clearly states that, the probability of occurrence of all letters in a message is not uniform. Based on standard estimation of probability of occurrence of English alphabet
\[ \sum p_i \log_2 p_i = p(a) \log_2 (p(a)) + p(b) \log_2 (p(b)) + \ldots + p(z) \log_2 (p(z)) = 11.34 \text{bits} \]

As per Shannon, “If the language is translated into binary digits(“0” or “1”) in the most efficient way, the entropy H is the average number of binary digits required per letter of the original language.”\(^{32}\). This shows that English text can be ideally represented using 4 bits based on the probability of its occurrence of each alphabet.

Consider the coin tossing experiment with the following outcomes HTHTTTTHHT

**Compression Trends in Genomics**
Compression waves have alleviated bottlenecks in many different fields ranging from internet service to the multimedia industries, its healing touch can even be felt in the field of Genomics. Compression is one technology that helps to shrink data there by storing in the same disk in a more effective approach. The New biology pulls out a new form of biological data- DNA (De-oxy ribonucleic acid) that helps to reveal the mysteries of life. DNA is equivalent to a text file with four alphabets \( \{A, T, G, C\} \), which forms the genetic code that runs our life.

DNA is responsible for the unique traits which is passed on to offspring through both parents and this macromolecule determine the variation in gene accountable for the look of hair or eye. For example, the Malayalam actor Indrajith due to his inherit traits has a strong resemblance to his parents Mallika and Sukumaran as shown in Fig. 8. It is even interesting to highlight the fact that not only the curly hair or the long nose but also the day to day activities of every cell is being controlled by the secret code engraved deep inside the nucleus of cell.

The human body system maintains a symphony with one hundred, million, million cells \((100,000,000,000,000)\). The symphony is controlled through the code

<table>
<thead>
<tr>
<th>Case</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>4/10</td>
</tr>
<tr>
<td>T</td>
<td>6/10</td>
</tr>
</tbody>
</table>

\[ H = \sum p_i \log_2 (p_i) = [0.4 * \log_2 (0.4) + 0.6 * \log_2 (0.6)] = 0.97 \text{bits / symbol} \]

Let us again consider the example: AATGGCACCT

The entropy H is calculated as
\[ H = -[0.3 * \log_2 (0.3) + 0.2 * \log_2 (0.2) + 0.2 * \log_2 (0.2) + 0.3 * \log_2 (0.3)] = 1.81 \text{bits / symbol} \]

The probability of occurrence of all the 4 symbols in a sequence is almost same. These messages may be represented ideally using two bits. The statistical nature of language helps to “reduce the entropy” by selecting a proper model. This knowledge in turn helps to store the message in more efficient manner. With the fundamental idea of compression we cordially invite our readers focus into the compression trends in Genomics.
reside inside the nucleus of cell which one inherit through hereditary. Human nucleus contains 23 pairs of chromosomes. Each chromosome contains a twisted ladder shaped DNA (Deoxyribonucleic acid) molecules. Two strands of DNA are known as coding strand and template strand. Each of them is complement of the other. And these two strands are connected by Hydrogen bonds. In DNA strands, Adenine always combines with its complement Thymine and Guanine always combines with its complement Cytosine. Human genome is made of 3 billion genetic letters. Ever since the complete draft of the first human genome in 2003, the biologist are marveled by many insightful surprise. It took nearly 13 years to publish the first draft of human genome for $1 billion. A decade later, with high throughput sequencing technology genomic data is growing and the cost is dramatically decreasing. The price of sequencing have gone down to $5000 which is further expected to drop down. Currently the genomic data is accumulated in Petabyte scales. Storage requirement for a petabyte may result in stacking of DVDs for nearly 2 miles tall. The huge accumulation of data is surpassing all hardware requirements for storing the data.

As shown below, sequencing technologies hyped a long way than Moore’s law.

Even in the midst of data horror, greater understanding of individual genome is of great interest by physicians and scientist. Early intervention of genetic risk, disease prediction and treatment were made possible with genetic understanding. Moreover the prescribed rate of drug dosage for each individual is revolutionizing the personalized medicine industry too. In the future, DNA sequences need to be kept in hand –held like the credit/debit card for medical decision making. Considering the demand for processing, analyzing, transmitting and storing the huge data, DNA Data compression seems a viable choice to manage the flood of data.

The new big voluminous data, addresses many computational challenges.

As mentioned earlier, the language of DNA has 4 alphabet. Hence, the Shannon’s information entropy is close to 2 bits per base and this forms the upper limit to encode the bases which is close to a naïve encoding of 2 bits per base. Understanding the nature of DNA data and exploiting its repeat properties help to frame an expert source model capable to compress the DNA sequences. As the data explosion continues to prevail we expect that novel compression algorithms has to flourish for effective DNA data Management.

References

Fig. 7: A sample DNA sequence

Fig. 8: An example for DNA traits- Resemblence of Malayalam film actor Indrajith with his parents (Source: http://en.wikipedia.org/wiki)

Fig. 9: Sequencing Cost trend 2002-2014 (Source: http://www.nature.com/news/technology-the-1-000-genome-1.14901)

Fig. 10: Personal Genome Card (Photo by Hareesh N.)

Biji C.L. is currently working towards her Ph.D. from University of Kerala. She is interested in communicating science through popular science magazines and has earlier contributed to CSI communications.

Manu K. Madhu is an M. Tech. student of School of Computer Sciences, M G University, Kottayam. Apart from the academic life, he is a passionate poem writer and he enjoys cooking.