KATIE WHITEFIELD

# The Perils of Data Transfer: Error Detection and Correction in Genetics and Computing

This article arose as a consequence of a conversation started between my mother and myself regarding the *Above the Threshold* installation for the Art and Genetics Exhibition. In attempts to explain what she was doing and some of the theory behind it, we started a dialogue about the similarities of DNA as a code and my experiences as a computer programmer working with code.

The implications of erroneous data transfer or copying in both genetics and computing are potentially disastrous. Mistakes in DNA replication can be fatal or lead to disease whereas corrupted data can range from complete data loss to data being unknowingly overwritten. The importance of identifying errors as they are transferred from one location to another is therefore paramount. Computer scientists have developed techniques to detect and correct errors in data transmission, which seem to mimic the commonly occurring methodologies that nature employs.[1]

Data needs to be transferred for a plethora of reasons. The Internet and all associated networking equipment rely on it; it is the means through which computers can “talk” to other computers or networks. Data can even be transmitted internally within a device, such as when random access memory (RAM) sends data to a processor. In nature, DNA needs to replicate to allow cell replication, not only to allow generational reproduction but also survival. The replication process is happening constantly in every cell of a human body simply so it continues to live. DNA also needs to transfer data so proteins can be produced, which are necessary for the structure and function in living tissue and organs.

Different strategies have emerged from nature to allow for accurate replication and transcription of DNA. The code for DNA is contained within the sequence of four nucleotide bases - thymine, cytosine, adenine, and guanine, commonly referenced as T, C, A, and G. In both replication of DNA, and transcription of DNA to mRNA, one strand directly correlates to another. Translation of mRNA to form a string of amino acids occurs using triplet coding and allows for some redundancy in the third base of each codon. In computer programming, instead of four bases, a unit of information is expressed as either a 1 or a 0, known as a bit. Aggregated series of bits are sent and received in various operations in computing systems. Computer scientists construct rules to try to account for errors during data transmission.

In computing, one of the simplest methods of error detection is the inclusion of a parity bit. A parity bit is an additional bit that is added to a sequence of bits used to detect whether an error has occurred. An even or odd parity scheme is agreed. Given seven binary bits, an eighth bit is added that makes the sum of all the 1's in the sequence even or odd. For example, using an even parity bit, a sequence of 0100110 results in an additional 1 as the eighth bit, making a sum of an even number of 1's: four. This data is then transmitted and checked at the other end. To detect an error, the number of 1's are compared to the parity bit; if the number is still even, then the result is correct. However, if the sequence comes through as say 00001101, then the number of 1's sum to three, despite the parity bit indicating that the sum should be even. Hence, this discrepancy between how many 1's there are and how many the parity bit state, indicate an error.

Parity bit checking was widely used for its simplicity and ease of implementation, but is by no means a foolproof means of error detection. If either two, four, or six bits flip (or even the parity bit itself) in the data transmission, then the parity bit would read as correct even though there were multiple errors in the transmission.

Repetition is another technique programmers use that involves sending the same bits multiple times, and if the grouped bits are not all the same, an error is detected. If the sequence 10110001 was to be sent, then this sequence would be sent as 111 000 111 111 000 000 000 111. A binary triplet read as 101 would mean an error has occurred. The majority of binary bits in the triplet that are the same could then correct the error so the triplet 101 would assume to translate to 1. This technique means the amount of data necessary to transfer might be trebled and is certainly not a scalable solution. Repetition may be able to detect the odd bit that is erroneous but longer sequences of wrong bits, called burst errors, would not be flagged.

Richard W. Hamming developed a much more efficient form of error correction in the 1940s.[2] Hamming codes use parity bits that indicate whether a small group of bits are correct. Due to the overlaps of which parity bits correspond to which data bits (calculated from an algorithm), a bit flip can not only be detected, but also corrected. They are highly effective at detecting one off errors but are not robust against burst errors. Hamming codes still have practical applications today, where an extended version of it is commonly used in memory storage.

DNA is a molecule comprised of two intertwining sugar phosphate backbones, the one strand bonded to the other through nitrogenous base pairs. When DNA is replicated, its famous double helix structure is unravelled into two strands. Because each of the four bases only couples with another specific base, each strand on its own contains enough information to recreate its corresponding strand. During replication, the enzyme DNA polymerase can thus move along each strand and add the correct base. The replication process does get approximately one in one billion nucleotides incorrect.[3] DNA polymerase also engages a proofreading phase, where it checks the base pairs are matched correctly. It is similar to a Hamming code being able to determine and flip an incorrect bit if it is reading off the template strand but also shares the same problem in that a 'burst error' along the strand would not be able to be corrected by the DNA polymerase.

During protein synthesis, after a DNA strand has been unzipped and transcribed into messenger RNA, the mRNA is rendered into a string of amino acids by a matching process called translation. Three bases from the mRNA strand are read together and correspond to a specific amino acid.

There are 64 possible combinations of codons, yet they only match to 20 amino acids and one stop signature. This method of translation uses the same ploy of redundancy that data transmission does in order to get the desired result. The aforementioned replication technique in computing, where the triplet binaries 101, 110, 011, and 111 would all map to a 1, is similar to mRNA translation. For example: the amino acid glycine is coded by the triplet codons GGT, GGC, GGA, and GGG. Thus an error in the last base would still result in the same amino acid, and is known as the "wobble effect."[4]

When an error fails to be detected in either genetics or data transmission, then in both instances the error will persist and be unable to be observed by future detection systems. Using a simplistic example: 10011100 is transferred (where the last bit is a parity bit using an even scheme) and received as 11111100. No error is detected from this transmission and if this data is then stored on the receiving system every time it is read or transferred it will contain an error. When an error happens in DNA replication it can be even more disastrous and lead to deleterious effects or be fatal. Error correction systems then clearly need to be extremely reliable and consistent.

The timeline of the development of error detection and correction in data transmission came clearly before our understanding of how DNA replicates itself and creates proteins. However, the techniques developed for error correction in computer and DNA data transmission, whilst clearly different in implementation details, share a rather striking similarity in approach. The addition of redundancy of information, as well as a proofreading phase in which errors can be amended, are core methods of error correction in both data transmission systems and genetics. As humans we endeavour to develop the best strategies for problems, such as error rates in transmission, but our solutions seem to unknowingly lean towards those that the natural world already applies.

**Katie Whitefield** graduated with a BSc in 2013 and continued on to gain a Post Graduate Diploma in Computer Science from Otago University in 2014. Currently she is a programmer and mobile app developer for the BBC in London.

[1] Thomas Thompson, *Error-Correcting Codes Through Sphere Packings to Simple Groups* (Washington: The Mathematical Association of America, 1983), vii. Subsequent mathematical methods of encoding messages to ensure correctness when transmitted over noisy channels led to discoveries of extremely efficient lattice packings of equal-radius balls, especially in 24-dimensional space. In turn, this highly symmetric lattice, with each point neighbouring exactly 196,560 other points, suggested the possible presence of new simple groups as groups of symmetries. Indeed, new groups were found and are now part of the "Enormous Theorem".

[2] Ibid.

[3] Lawrence Loeb, et al., "Multiple Mutations and Cancer," *Proceedings of the National Academy of Sciences* Vol. 100 (1974): 776-781, http://www.pnas.org/content/100/3/776.full.

[4] Francis Crick, "Codon-Anticodon Pairing: The Wobble Hypothesis." *Journal of Molecular Biology* Vol.19 (2) (1996) 548-555, http://www.sciencedirect.com/science/article/pii/S0022283666800220.