NONLINEAR INFORMATION-KEEPING IN BULK MATERIALS

Davis, J.[1,2‡], Kaiser, T.[3], Shultz, T. [3], Van Vugt, B. [4], Church, G. M.[1‡]

[1] Department of Genetics, Blavatnik Institute, Harvard Medical School

[2] Department of Biology, Massachusetts Institute of Technology

[3] Blink AG, Jena, Germany

[4] Department of Humanities, University of Amsterdam

[‡] Corresponding authors. jdavis@genetics.med.harvard.edu (J.D.); gchurch@genetics.med.harvard.edu (G.M.C.)

Many common materials are made up of diverse constituent substances. Examples of bulk materials such as sand and concrete, mixtures of fragrances that make up aromas of flowers and perfumes, and even solutions of minerals in drinking water fall into this category. These materials can hold information by precisely varying amounts of their respective components. Consecutive data can be stored non-consecutively, and then reacquired in proper sequence by distinguishing incremental chemical or physical properties of materials in these mixtures. So long as such non-linearly encoded material is a homogenous mixture, a small aliquot can hold information identical to information contained in a larger body of the same material.

INTRODUCTION

Letters typed in one character at a time to make up a line of text and the consecutive sequence of nucleotide bases that make up synthetic DNA molecules are both examples of linear encoding. An encoding method is considered “nonlinear” when data resides in random mixtures rather than as information rendered in step-by-step sequence. Units of nonlinearly encoded data are not physically connected to each other and so, are not arranged into adjacent sets. Nevertheless, practical data storage allows for recovery of familiar forms of digital content.

Non-linearly encoded materials may include unique markers or, “barcodes” (1). Addition of such markers to initial data allows random mixtures of information-keeping units to be reconfigured into familiar linear formats as a part of the process of data retrieval. Since corresponding sets of unique markers must first be created, then added to, and later removed from final assemblies, methods of marker-dependent recovery of input data call for increases in both the volume of encoded material and decoding complexity.

Microscopic physical or chemical “taggants” are sometimes added to bulk materials to distinguish authentic products from counterfeits, provide traceable lot numbers and company names, detect cross-contamination or dilution of proprietary products in agriculture, pharmaceuticals, plastics, inks, and explosives. Like markers described above, industrial taggants are additive. Where input data is large enough to require multiple taggants, each variety of data-holding taggant would also have to contain positional data to guide assembly of decoded information. Here again, data consigned to storage in the form of encoded taggants would increase total volume of encoded material together with corresponding increases in complexity of data retrieval.

Linear data formats can be efficiently entered into and recovered from non-linear storage media without the need for added markers or taggants. Instead, queueing of units holding stored data can be accomplished by sorting random mixtures according to properties of materials that are naturally incremental, and do not increase volume of information storage.

Unique ingredients of multi-component materials (such as sand) can be initially encoded as numbers indicating percentages of the whole or, in numbers of grams-per-kilogram or, in unit numbers obtained by other consistent forms of measurement. Since linear data formats are organized in chronological order (first-to-last), these numbers are then decoded incrementally according to some secondary property. For any given material, there are many different characteristics to choose from, including optical, mechanical, electrical and thermal properies. In this example, increments of molecular weight serve as a “clock” where numbers indicating quantities of the lightest substance are gathered first, then the next heaviest, then the next, and so on. A random mixture of information-keeping units can thereby be reconfigured into a correct linear sequence.

Volume of information that can be entered into such non-linear data storage is limited by the number of different substances included in a particular storage medium. Techniques for encoding of bulk materials may never achieve information density (as in bits-per-given-volume) to rival storage capacity of DNA molecules, but more than 100 million chemical compounds that make up our world comprise a large selection of encodable substances (2). Many thousands of additive natural and synthetic materials can be included in “sand” and other bulk materials.

DISCUSSION

More than 350 genes have been identified that contribute to the aroma of roses (3-5) and data could be encoded by regulating the expression of these genes. Unfortunately, since the rose is a woody plant, the process of rose genetic modification would take several years at best, require a dedicated laboratory, and the full-time attention of one or more highly skilled professionals. These considerations, taken together with requisite facilities and protocols regarding disposition of healthy recombinant plants, would mean that a project to encode information in the aroma of a rose would be both expensive and time-consuming.

Another approach to aroma-encoding could be realized by modifying components of perfume (see: Appendix 1). This would require formulation of a fragrance with many different compounds, but manufacturers are known to draw on more than 3000 ingredients to make perfume (6). In fact, more than 800 ingredients can be involved in the development of a formula for a single fragrance (6-8). Furthermore, many ingredients of popular perfumes are uniform synthetics made up of single molecules rather than complex mixtures that make up most natural fragrances. Single-molecule synthetics would be easier to compound in a non-linear encoding scheme based on incremental material properties.

A variety of analytical techniques could be used to recover digital information from encoded aromas. These include chemical sensor arrays (9-10). Likewise, instruments of mass spectrometry, gas chromatography-mass spectroscopy (GC-MS), and liquid chromatography–mass spectrometry (LC–MS) might also be applied to detect and quantify unique ingredients of aromas (11-12). Less sophisticated techniques might also be used to

determine which ingredients comprise a particular fragrance and the relative quantities in which they appear. Liquid chromatography utilizes wicking materials to separate individual substances present in liquids and determine their relative abundances. Moreover, infrared and UV imaging could be used with this kind of analysis to reveal otherwise invisible substances.

A still less problematic proof-of-concept for nonlinear information-keeping in bulk materials has been the opportunity to demonstrate encoding of ordinary drinking water.

METHODS

In May of 2022, the water authority in the German city of Jena provided lists of minerals and ranges of appearance in the output of several different wells that make up their municipal water supply (see: Appendix 2). Water was subsequently encoded by varying amounts of soluble minerals that normally appear in drinking water and in concentrations considered safe to drink.

Once mineral salts go into aqueous solution, their crystalline structures dissociate into sets of positively charged *cations* and negatively charged *anions.* In this case, 5 soluble minerals were dissolved into an aqueous solution with ultrapure water so that unit quantities of their component elements measured in mg/liter H_2O exactly correspond with numerical ASCII code for alphabetical letters in the word, “truth” (See: Appendix 3). The order of letter assignments (t-r-u-t-h) correspond in turn, with incremental molecular weight of ions in solution.

First, a concentrated solution was created in 100ml containing:

1.2710g *NaCl*
4.0071g $MgCl_2\ 6H_2O$
0.9533g *KCl*
0.3668g $CaCl_2\ 2H_2O$
4.0303g $Mg(NO_3)_2$

So that 1ml of the concentrated solution added to 1 liter extra pure water yields:

Na	5mg/l	= 101
Mg	9mg/l	= 1001
Cl	21mg/l	= 10101
K	5mg/l	= 101
Ca	1mg/l	= 01

Since 5-bit ASCII codes are expected, zeros are added as suffixes to smaller numbers to convert them into 5-bit numbers (see Appendix 4) :

Na	5mg/l	= 10100	= T
Mg	9mg/l	= 10010	= R
Cl	21mg/l	= 10101	= U
K	5mg/l	= 10100	= T
Ca	1mg/l	= 01000	= H

CONCLUSION

In 1946, Jorge Luis Borges published a one-paragraph short story about a map of the world that coincided with itself, concluding that such a map was useless to the science of geography*. Now it seems that a map of the world that coincides with the world might not turn out to be so utterly useless after all.

An unspoken aspiration of information-keeping in DNA is the notion of a greater biological archive, one in which the whole terrestrial biome becomes an incontrovertible, perpetual record of human culture and civilization. Granted, many obstacles would have to be overcome, not in the least of these are ethical protocols that concern disposition of recombinant materials outside of the laboratory, but that is the dream.

Techniques for mass-encoding of non-biological bulk materials suggest that parts of the Earth itself might also be configured to hold a legacy of human archives. Since encoding substances can be added to bulk materials in relatively miniscule amounts, mineral content of deserts and glaciers could be modified with secondary substances dropped from the air, much like dispersal of flame-retardant mixtures released from aircraft over large swaths of landscape. These ideas may have some implications for CETI (Communication with Extraterrestrial Intelligence) too. Surfaces of biota-free environments such as comets and asteroids might be intentionally modified to serve as message keeping bodies for dispersal into interplanetary and interstellar environments.

Codes are systems of signals, symbols, textures, colors, shapes, and forms coupled with procedures for meaningfully combining them. Information science, on the other hand, has become the study of structures intentionally removed from meaning and purpose. In this sense, it makes no difference whether a binary digit represents the toss of a coin, or the fate of the universe. Yet, words and images are created not only to be transmitted and received. “Information” and “meaning” are both abstract nouns. Every character of text, every number and pixel, is a cipher, an estimation. Nothing is simple. Nothing is just zero or one, black or white. Everything has texture, nuance, the measure of what is true and real and what is not. Code is language, mathematics, and art. It is what we use to hold ideas, the products of reason. Capacity to create and understand meaning is perhaps the single most defining aspect of human nature, the power of imagination to transform and illuminate human lives.

Encoding “truth” into water with 41 mg/l of harmless minerals embodies a higher truth. The tragedy of history is that it is a record more of loss than content. Despite more than 100,000 years of accumulated knowledge, we have forgotten, destroyed, or abandoned much more than humanity can ever remember. Time and again, malice and indifference have caused the brightest flowers of culture and civilization to fade and pass away. With this project, the authors imagine a new kind of library that can help to defeat the tragedy of history, and where “truth” is the first word saved.