Joe Davis

Romance, Supercodes, and the Milky Way DNA

An Artistic Principle of Transanimation

This paper will explain in some detail how high resolution digital images may be precisely coded into molecules of synthetic DNA.[1] Various elements of molecular biology, mathematics, and information science are relevant to the topic, yet these important technical and scientific aspects surround a strong poetic theme.

Artists of the Golden Age fanatically pursued mimetic reproduction of the natural world, especially of the human body itself, for which the Greeks sought nothing less than "perfect knowledge". From idealized proportions of the human figure, they derived the classical foundations of music, architecture, and even of science and mathematics. The strong artistic tradition surrounding this "search for self" has in many historic examples included the search for some special power over elusive qualities of vitality and function that distinguish life and death.

The quest for "secrets of life" that preoccupies literature and the history of art is now of course ever more intensely pursued in laboratories of so-called "life sciences" worldwide. Here, at least insofar as certain biomolecules are concerned, the age-old dream of "bringing-to-life" inanimate matter is suddenly no longer the stuff of magic, myth, legend, or for that matter, of divine intervention. Even so, perhaps the most dramatic and sweeping attempts to bring inanimate matter to life are not really to be found in either art or molecular biology. Rather, they are embedded in recent scientific attempts to communicate with extraterrestrials. With an alchemy of rockets, plaques on space probes, powerful radar transmitters, and binary messages beamed into space, science attempts to animate the entire cosmos.

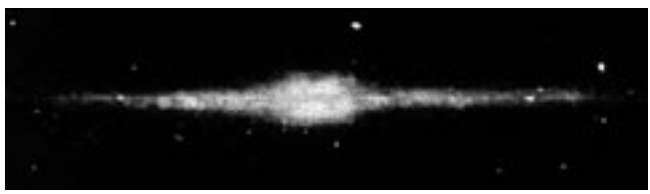
In 1986, I combined mathematical strategies previously used to compose messages for extraterrestrial intelligence with standard techniques of DNA synthesis to create my first synthetic DNA molecule. This molecule, *Microvenus,* was created as a work of art in collaboration with Dana Boyd and Jon Beckwith at Harvard Medical School and Hatch Echols at the University of California Berkeley. It contained graphic rastar information for an ancient Germanic rune used to represent "life"[2] and by simulacrum, an image of external female genitalia heretofore censored from graphic representation in serious scientific messages for extraterrestrial intelligence. Synthetic DNA created for a subsequent project, *Riddle of Life*[3], realized the molecular implications of model-based communications which were originally exchanged between Nobel laureates Max Delbrück and George W. Beadle in 1958. *Riddle of Life* DNA, created in 1993 in collaboration with Burkhardt Wittig's laboratory at the Free University of Berlin, is coded with Max Delbrück's English language text, "I am the riddle of life; know me and you will know yourself." Both *Microvenus* and *Riddle of Life* molecules have since been inserted into the DNA of virus-like bacterial vectors called plasmids, and subsequently cloned into laboratory strains of *E coli* bacteria. With the examples of text and simple line graphics in the form of DNA, these earlier works suggest that DNA may eventually find some special usefulness as a data storage medium for conventional computer databases. Encoding strategies used to create *Microvenus* and *Riddle of Life* molecules would be inefficient for that purpose, however, and were not intended to be directly compatible with conventions for the organization and operation of computer databases. The plan to develop both a computer-friendly and a biochemically practical encoding strategy for the conversion of an ordi-

nary computer file into a DNA sequence is consistent with the scientific or technical ideal of a such a biological database. The reasons I have decided to capture a picture of the Milky Way in this form are however, hopelessly romantic.

Many years ago, a friend showed me a sketchbook containing illustrations for a children's story about a child who could find no happiness until she found a mouse who had a map of the whole world in its ear. Although it is of no technical or scientific relevance, I was also inspired by the fact that the taxonomical name of the familiar flower otherwise called "forget-me-not," is Myosotis which is from the Greek meaning, "mouse's ear".

A Gene-sized Picture of the Milky Way

The first high-resolution picture data to be coded into a sequence of DNA bases[4] is a map of the Milky Way galaxy. This image has been coded into a 3867-mer DNA molecule (a molecule with 3867 bases). A synthetic DNA molecule having 3867 bases is large.[5] In fact, a 3867-mer will be among the largest synthetic DNA molecules ever made. It will be comparable in size to many genes known to appear in nature, and larger than some plasmids.[6] Unlike the genes of organized life-forms, this one will not be translated into significant proteins or enzymes by the various elements of cellular machinery. Instead, this will be a molecule specifically intended for translation solely by technological means. Data used to create the Milky Way DNA map were originally collected in space with instruments on board NASA's Cosmic Background Explorer (COBE) satellite.[7] Before results of the COBE experiments became available in the early 1990's, intragalactic dust clouds obscured astronomers' view of large parts of our own galaxy. Interferometric infrared sensors on the COBE spacecraft produced the first high resolution maps of the entire galaxy, including unprecedented images of the galactic core.[8] The portfolio of these findings may comprise the most important advances in cartography since the contributions of Gerhardus Mercator in the sixteenth century.



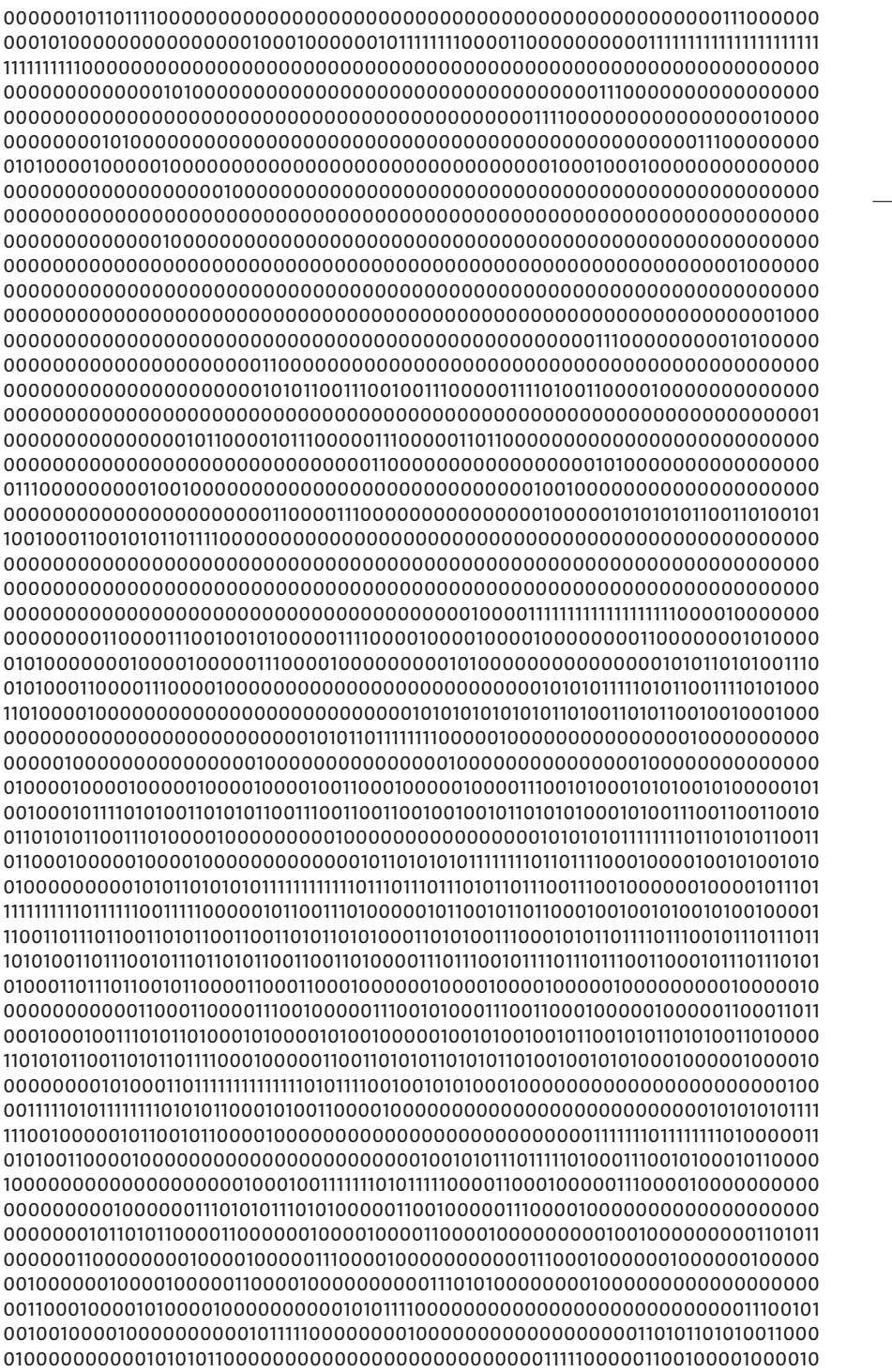
COBE image of the Milky Way galaxy

The COBE image of the Milky Way has had its primary existence as "on" and "off" states, first, in the semiconductors of detectors in space, and then again in the solid state memory of computers that the original data were transmitted to. The graphic image above is only *one* of the ways in which this information can be expressed.

Computer Codes

"0's" and "1's" that correspond to the COBE Infrared image of the Milky Way—or for that matter, the binary identity of *any* computer file—can be easily obtained with one of various "editors" commonly included in software packages that come with computers at point-of-purchase. These ubiquitous computer desktop tools allow for quick and easy interconversion of binary "picture data" into several standard forms. Common computer picture formats such as JPEG, GIF, TIF, etc., may be viewed either as text files composed of alphanumeric characters or as picture files composed of video image units called "pixels". I used a picture editor written for Macintosh called "ResEdit" to obtain binary data corresponding to the Milky Way image.

Binary equivalent of COBE Milky Way image



0001111111101010101000000000000000000001001010010011011000100000100001111111101 0
1010100000000000000000010110101010111010001100100000111011111101011011000000010
0000001011100111001100010100001010010001111111101010100001000000001000001101011
0101010010010000101001001111111010101001000000000000000000001111011100100110000 1
0000000001111110100100000000000000000000000000110001010010010000100000000
0011111010000001000000000000000000000001100010100100100001000000000000011101011
0000000000000000000000000000110001100010110000100000000000011101001000000 0
0000000000000000000000100001000010100000000000000000001111101000000000000000000
0000000000000000100001000000110000100000000000011101001000000000000000000000000
0000000000100001000011000000000000000000011101110000000000000000000000000000000
0010000100000011100001000000000000000000000001110000000000000000000000001010000
0000001000000000000000000000000000000000011100000000010101110000000000000011110
0000000000000000111000000000000000000010000000000000000000100000000000000000011000
0000000000001100000000000000000100000000000000000000000000000000000110100000000
0000110000000000000000100000000000000000101000000000000000001010000000000000001011111
0000000001011111000000000000001010001111111111111110000000000000010000000010000010
1000101110101011010010110001101101011010101000110100101101101011001011010101000100
0000110000101101110011001000010000000110000100000000000000000000000101000000000
0000000001100000000000000111000010010010101100110100101100100011001010110111100 1
0000001100100011001010110001101101111011011010111000001110010011001010111001101110
0110110111101110010000000000000000000001010000000000000000000100000000000000000000
0000010000100001101001000000110000101110010011001010010000001101110011001010 1
1001010110010001100101011001000010000000111010001101111001000000111001101100101 0
1100101001000000111010001101000011010010111001100100000011100000110100101100011
011101000111010101110010011001010000110100000000000000000000000000011111111

Because it would be a maddening task to carry out even modest computer operations using only two binary digits, human programmers address binary computer memory with informational superstructures that compile binary characters into the 16 hexadecimal, or base-16 numbers, "0" through "F". In order to save time, I also used a hexadecimal, or *hexdex* equivalent of the Milky Way image to code the COBE Milky Way image into DNA. The same ResEdit picture editor that yielded binary data also provided a hexdex equivalent of the Milky Way image.

Hexadecimal equivalent of COBE Milky Way image:

02DE000000000000700500011 02FF0C00FFFFFFFF000000000000000000000500000000700000
0000000001E0001000A00000000000070050820000000022000000001000000000000000000
0000000000000000100000000000000000000000000000040000000000000000000000000
000000000400000000000000700500000030000000000000000005672707A6100000000000000
000000010001617070 6C000000000000000300005000070008000000048000000000001870001
05566964556F00000000000000000000000000000000000000000000000000000000010FFF
FE10001872507842100C050501083840140005 6A728C384000000ABEB3D4684000000AAAB4
D648800000056FF0400040004000400040004000421042131043945 4A0445EA6ACE66496A
8A73326ACE84010000AAFF6ACD882100005AAFF6F1094A4015AAFFF7775B9C810BBFFBF3E03
A0B2D89294873766B335A8D4E2B7B97775372ED6668772F773177546ECB0C6204210401040031
8720E51CC4106362275A28520949656A686ACD6F1066AD5A4A88210051BFFF5E4A88000010F
AFF562984000000AAFE41658400000 0FEFF41984000004AEFA3958400004 4FEBE188384000
040EAEA0C838400000 0B58604218401200D606010838007102040408418400EA0100000C42
8400AF0000001CA48400BE01000035A98400AB0000003E0C8421FF5500004A4D8821FF5500
005AAE8C83BF5B01017398A148FFAA1010655490A4FEA900003DC98401FA40000018A48400F
A04000018A48400EB00000018C58400E900000008428000FA00000008418400E90000000
4218000EE0000001083840000380000A0040000000700AE00780007000100010003000 30
0040000000D000C00100050005000BE00BE0028FFFE001010517569636B54696D65AA20616
E6420610000280003000E12566964656F2064665636F6D70726573736F7200002800080000
210D20617265206E65656465642074 6F207365652074686973207069637 4757265 0D000000FF

DNA Numbers

Given a mathematical interpretation of DNA, the interconversion of digital information and DNA sequences is a straightforward mathematical operation. DNA molecules are variable modular assemblies that have at least conceptual parity with the mathematical structure of computer memory. The most obvious difference is that DNA "memory" is recorded with four integers, rather than with two binary numbers.

The four DNA "numbers" are the four movable parts of DNA molecules called "DNA bases" : *cytosine, thymine, adenine,* and *guanine,* or "C", "T", "A", and "G". In double-stranded DNA molecules, these bases assemble in pairs that form the central rungs of the ladder-like structure of DNA. "C" always pairs with "G." and "G" with "C"; likewise "A" always pairs with "T," and "T" with "A." Nature uses DNA molecules with variable sequences of bases to hold information in a way that is analogous to the way in which computer memory is stored on hard disks, magnetic tape, CDs, and semiconductor "chips."

Interpolation of the "on/off" quanta of semiconductor states depends on the idea that "off" is *less* than "on"; "Off" states are construed to be "0" and "on" states to be "1." The notion of quantity can also be used to logically increment the DNA bases. Fortunately, none of the four bases are exactly the same size. Thus, each can be assigned an incremental value that corresponds to its relative molecular weight:

molecular weights of DNA bases

molecule	molecular weight	incremental value
cytosine	111.10	0
thymine	126.11	1
adenine	135.13	2
guanine	151.13	3

Given the above increments any numerical database can be translated into a DNA sequence. A key for the inter-conversion of DNA, binary, and hexadecimal numbers is provided below.

DNA-binary-hexdex number key

C = 00; T = 01; A = 10; G = 11

CC = 0000 = (0)
CT = 0001 = (1)
CA = 0010 = (2)
CG = 0011 = (3)
TC = 0100 = (4)
TT = 0101 = (5)
TA = 0110 = (6)
TG = 0111 = (7)
AC = 1000 = (8)
AT = 1001 = (9)
AA = 1010 = (A)
AG = 1011 = (B)
GC = 1100 = (C)
GT = 1101 = (D)
GA = 1110 = (E)
GG = 1111 = (F)

If *noms de plume* (such as "X" and "Z") are usd to represent either the "A" and "C" hexadecimal characters, or the "A" and "C" DNA bases, then translation from one into the other can easily be carried out on the desktop of a small personal computer with the *change,* or *replace* function of a text editor like Microsoft Word. The Milky Way image can therefore be directly translated into the following 2936-mer DNA code, but there is a problem.

2936-mer Milky Way DNA (primary strand):

CCCAGTGACCCCCCCCCCCCCCCCCCCCCCTGCCCCTTCCCCCCCTCTCCCAGGGGCCGCCCC
CGGGGGGGGGGGGGGGCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCTTCCCCCCCCCC
CCCCCTGCCCCCCCCCCCCCCCCCCCCCCCCCCCCTGACCCCCCCTCCCCCCAACCCCCCCCCCCC
CCCCCCCCTGCCCCTTCCACCACCCCCCCCCCCCCCACACACCCCCCCCCCCCCCCTCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCTCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCTCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCTCCCCCCCCCCCCCCCCCCCCCCCTGCC
CCTTCCCCCCCCCCCCCGCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCTTTATGCATGCCTGAAT
ACTCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCTCCCCCCCTTACTTGCCTGCCTAGCCC
CCCCCCCCCCCCCCCCCCCCCCCCCGCCCCCCCCTTCCCCCCCCTGCCCCTCACCCCCCCCCCCCCT
CACCCCCCCCCCCCCCCCCCCTACTGCCCCCCCTCCTTTTTATAATTATCTATTTAGGCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCTCCGGGGGGGGGGACTCCCCCCCTACTGCATTCCT
GACTCCACTCCCCGCCCTTCCTTCCCTCCACCGACTCCCCTTCCCCCCCTTTAAATGCAACGCC
GACTCCCCCCCCCCCCCAAAGGAAGCGGTTCTAACTCCCCCCCCCCCCCAAAAAAAGTCGTTATCA
CACCCCCCCCCCCCCTTTAGGGGCCTCCCCCCCTCCCCCCCTCCCCCCCTCCCCCCCTCCCCCCCTCCC
CCCCTCCACTCCTCCACTCGCTCCTCCGATTCTTTCAACCAATCTTGAAATAAAGCGATATATCAT
TAAAACAATGCGCGCATAAAGCGAACTCCCCTCCCCCCCCAAAAGGGGTAAAGCGTACACCACTCC
CCCCTTAAAAGGGGTAGGCTCCATTCAATCCCCTTTAAAAGGGGGGTGTGTGTTAGATGCAC
CTCCAGAGGGGGAGGGCGGACCAGCGAACCAGCAGTACATCAATTCACTGCGTGTATAAGCGCGT
TAAACGTTCGACAAGTGAGATTGTGTGTTCGTGCAGAGTTATATAACTGTGCAGGTGTGCGCTTGT
GTTTCTAGAGCAGCCGCTACACCTCCACTCCTCCCCTCCTCCCCCCGCTACTGCACCGATTCT
GCGCTCCTCCTACGTACACATGTTAACAACTTCACCATTCATTATTTAAATAACTAAAGCGTTAGGCTC
CTATAAAGTTTAATCAAACACCACTCCCCTTCTAGGGGGGGTTGATCAAACACCCCCCCCCCTCCG
GAAGGGGTTTACAATACTCCCCCCCCCCCCCCAAAAGGGATCCTTATTACTCCCCCCCCCCCC
CGGGAGGGGTCCTAAATACTCCCCCCCCCCCCTCAAGAGGAACGATTCTTACTCCCCCCCC
CTCTCGGGAAGGACTACACCGACTCCCCCCCCCTCCCGAAAGAAACCGCACCGACTCCCCCCCCCC
CCCAGTTACTACCTCCACTACTCCCCTCACCCCGTTACCTACCCTCCACCGACTCCCCCTGCTCCCAC
CTCCCTCCCACTCCTACTCCCCCGAAACCCTCCCCCCCCCCCGCTCCAACTCCCCCAAGGCCCCCCCCC
CCCCTGCAATCACTCCCCCAGGACCCTCCCCCCCCCCGTTAAATACTCCCCCAAAGCCCCCCCCCCCC
CGGACCGCACTCCACTGGGGTTTTCCCCCCCCCTCAATCGTACACCACTGGGGTTTTCCCCCCCCT
TAAAAGAACGCACCGAGGGTTAGCCCTCCCTTGCGATACAACTTCACGGGGAAAAACTCCCTCC
TAAGTTTCATCCAATCGGGAAAATCCCCCCCCCCGGTGCATACTCCCCTGGAATCCCCCCCCCCCCTA
CAATCACTCCCCCGGAACCTCCCCCCCCCCTACAATCACTCCCCCGAAGCCCCCCCCCCCCCTACGCT
TACTCCCCCGAATCCCCCCCCCCCCCCACTCCAACCCCCCCGGAACCCCCCCCCCCCCCACTCCTACTC
CCCCGAATCCCCCCCCCCCCCCTCCACTACCCCCCCGAGACCCCCCCCCCCCCCTCCACCGACTCCCCC
CCCCCGACCCCCCCCCAACCCCTCCCCCCCCCCCCCCCCTGCCCCAAGACCCCTGACCCCCCCCTGCCCC
CCCTCCCCCCCTCCCCCCCGCCCCCCCGCCCCCCTCCCCCCCCCCCCCCCCGTCCCCCCGCCCCCCTCCC
CCCTTCCCCCCTTCCCCCCAGGACCCCAGGACCCCCAACGGGGGGGACCCCCTCCCTCCTTCTTGTT
TAATTACGTAAGTTTCTAATTAGTTATTAAAACACCTACTTAGATATCCACCTACTCCCCCCCCCAACCC
CCCCCGCCCCCCGACTCATTTATAATTATCTATTTAGGCACCTATCTATTTACGTAGGTAGTTGCCTG
CATATTTGCGTGCGTAGGTGCACCCCCCCCCAACCCCCCCACCCCCCCCCCACTCCGTCACCTACTTG
CATATTCACCTAGATATTTATTTATCTATTTATCCACCTGTCTAGGCACCTGCGTATTTATTCACCTGTC
TAACTAATTGCGCACCTGCCTAATTACGTGTCTGTTTGCATATTCCGTCCCCCCCCCCCCGGGG

The problem with this sequence has to do with the blackness of space. Black color fields are translated as repeats of "0000" and "0" respectively in binary and hexdex, and as repeating sequences of the DNA "CC". Because of the predominance of a black color field in the COBE image of the Milky Way, there are long runs of poly"C" in the

corresponding DNA sequence. While any shift in the shape of a single-stranded DNA molecule[9] caused by the electromechanical torsion of an individual base would be very difficult to measure, the accumulated torsional effects of many identical bases can influence the formation of loops and coils which might not otherwise form in a more heterogeneous molecule. A DNA molecule is normally very flexible when suspended in an aqueous solution. In this flexible form, DNA has subtle structural characteristics that are recognized by various other molecules that interact with it. Enzymes that function normally in association with heterogeneous DNA will tend to "slip" or "skip" when acting on long poly-C repeats. Furthermore, standard techniques for *sequencing* or "reading back" DNA are unreliable for sequencing long poly-C repeats. In this case also, there is a "skipping" problem that has to do with how DNA molecules pass through chemical gels in a process called "gel electrophoresis" that is an essential part of standard sequencing techniques[10]. As it is, the 2936-mer Milky Way DNA sequence would be extremely difficult to synthesize and clone into the reproductive machinery of living cells, and equally difficult to sequence, or read back, with existing technology. Ironically, long repeats of poly-C are known to exist in nature, but only in the so-called "junk DNA." Junk DNA is not acted on by the processes of "transcription" and "translation"[11] which are involved in the operational dynamics of functioning DNA.[12] Like synthetic poly-C DNA, poly-C junk DNA molecules cannot be easily sequenced and so with few exceptions do not reside in existing genome databases. It is therefore impractical to search for homologies of the 2936-mer Milky Way DNA in the archives of genome research.

DNA Supercodes

Numerical data cannot be directly translated into usable DNA sequences after all. Practical, DNA "memory" will have to contain data without being biochemically problematic. In order to create such a working biological information repository for the Milky Way picture data, or for any other generic, extrabiological database, corresponding DNA sequences can be recoded into second-generation sequences that are 1.) "biochemically friendly", 2.) contain first-generation extrabiological data with high fidelity, and 3.) do not dramatically increase the size of the original database.

To that end, a secondary encryption strategy, or "DNA supercode" is presented in this paper that allow for the translation of a given DNA sequence into a series of second generation or supercoded sequences that can, by various intervening operations, be precisely converted back into the original sequence. Supercoded Milky Way sequences retain both the original COBE picture data and the logical system of quantitative increments used to translate it into DNA in the first place.

I originally experimented with several different data-handling strategies that were based variously on character rotation and doublet- and triplet-encryption of the original sequence. Inevitably however, symmetries imparted by each successive scheme would in turn impart a biochemically significant symmetry into the supercoded DNA sequence and furthermore, some of these schemes would dramatically increase the size of the molecule. Supercodes that simply factor first-generation data seem to be unreliable means for the creation of extrabiological databases with unexpanded, biochemically suitable DNA molecules. Although each of the sequences examined had aspects that could be conveniently synthesized and assembled, only an arbitrary patchwork of these supercodes could be used to create a practical Milky Way molecule. A "road map" of the same arbitrary patchwork of codes would be required to decode original data. In theory, a special "road map" or decoding primer could be included with each molecular database. This scenario does not seem to be a realistic one, however, because it allows for unrestricted and time-consuming complexity. It seems likely that

a decoding requirement for supplementary "roadmaps" would significantly expand the original volume of data. In my opinion, randomly composited assortments of factor-based supercodes do not represent practical means to build biological information repositories for conventional databases.

On the other hand, since generic data is itself not "uniform," varying structural aspects (including undesirable ones) of its translation into DNA can be expected to occur randomly. DNA supercode strategies described above are based on continuous and regular modifications of original data. These can render undesirable structural elements more topologically complex, but the inherent symmetry of uniform adjustments can produce structural problems where none previously existed. A more perfect supercode would be an asymmetrical one capable of variable encryption that can be tailored to solve different kinds of problems. Certain natural operations of the genetic code function in just this way.

Degenerate Supercode

Nature must conserve specific genes in the milieu of evolutionary change. The influences of natural selection and the genetic machinery of sexual reproduction keep the context of natural DNA sequences in constant flux. Nature somehow manages to reproduce very specific proteins with DNA code that is always being actively rewritten. To accomplish this, the sequence of bases in a given gene can be "restated" in many different ways. Each of these alternative sequences can be translated into cell products, or proteins that are identical to the product of the original sequence. Nature's ability to carry out such restatements is based on what is called the degeneracy of the genetic code. To explain this quality of degeneracy, I will take a moment to re-examine the basic processes by which information is stored in DNA, transcribed into RNA, and translated into protein.

Transcription and Translation

Long molecules of DNA are ultimately translated into all of the substances that make up living things. In a process biologists call "transcription," a copy, or template of one side of the DNA duplex is written into a different kind of nucleic acid called RNA (ribonucleic acid). With the aid of an enzyme called "RNA polymerase", information from the original DNA molecule is copied into a variety of RNA molecule called messenger RNA, or mRNA. mRNA[13] molecules are structurally identical to DNA molecules with two principal exceptions: 1.) The ribose sugars in RNA have an extra oxygen molecule attached to them, and 2.) the DNA base thymine ("T") is transcribed into the smallest mRNA base, uracil ("U").

mRNA molecules are intermediary agents in the process by which the original DNA code is translated into protein. In cells, the information stored in mRNA molecules is processed by hour glass-shaped structures called ribosomes. These attach themselves to mRNA molecules and "read them out," three bases at a time. Then, with the help of yet another RNA molecule called transfer RNA, or tRNA, a new template is made. This third-generation template is not written into a nucleic acid like DNA or RNA. Instead, information from the original DNA molecule—and the RNA intermediary—is made into protein. This new template is one that, like mRNA, corresponds to the sequence of bases in the original DNA molecule. As a general rule, the ribosome adds one amino acid for every three mRNA bases. For each three-base segment, or triplet codon, in the original DNA molecule, a corresponding codon is found in the mRNA template which is translated into an amino acid by the action of a ribosome and tRNA.

There are only 20 amino acids in almost all living things.[14,15] Nature uses the same 20

amino acids to build structures as diverse as tomatoes and human beings. Amino acids combine to become peptides, and these in turn combine to form proteins. Nearly everything in the natural biological world is made from, or by interaction with, protein. This final template is much larger than either DNA or mRNA. It is this copy that ultimately becomes the living organism itself.

It is easy to think about these varied operations as if they were operations on a factory floor : Genomic DNA comprises the original "drawings". The mRNA molecules are "blueprints" distributed to workplaces. Ribosomes and tRNA molecules are the cellular "assembly lines and factory workers" that carry out construction of long chains of amino acids that correspond to the original factory drawings. These are, in summary, the biological operations of transcription and translation.

DNA is mapped into an mRNA copy which is acted on by cellular operations that translate one of 20 amino acids from each nucleic acid triplet. A set of 64 triplets can be made from four individual bases. Each of these triplets directs the production of one of 20 amino acids. This association of 64 nucleic acid triplets and 20 amino acids—often represented in the form of a rectangular chart—is called the genetic code.

The Genetic Code :
nucleic acid triplets and amino acids

first place	URACIL	CYTOSINE	ADENINE	GUANINE	third place
URACIL	UUU-PHE	UCU-SER	UAU-TYR	UGU-CYS	URACIL
	UUC-PHE	UCC-SER	UAC-TYR	UGC-CYS	CYTOSINE
	UUA-LEU	UCA-SER	UAA-STP	UGA-STP	ADENINE
	UUG-LEU	UCG-SER	UAG-STP	UGG-TRP	GUANINE
CYTOSINE	CUU-LEU	CCU-PRO	CAU-HIS	CGU-ARG	URACIL
	CUC-LEU	CCC-PRO	CAC-HIS	CGC-ARG	CYTOSINE
	CUA-LEU	CCA-PRO	CAA-GLN	CGA-ARG	ADENINE
	CUG-LEU	CCG-PRO	CAG-GLN	CGG-ARG	GUANINE
ADENINE	AUU-ILEU	ACU-THR	AAU-ASN	AGU-SER	URACIL
	AUC-ILEU	ACC-THR	AAC-ASN	AGC-SER	CYTOSINE
	AUA-ILEU	ACA-THR	AAA-ASN	AGA-ARG	ADENINE
	AUG-MET	ACG-THR	AAG-LYS	AGG-ARG	GUANINE
GUANINE	CUU-VAL	GCU-ALA	GAU-ASP	GUU-GLY	URACIL
	GUC-VAL	GCC-ALA	GAC-ASP	GGC-GLY	CYTOSINE
	GUA-VAL	GCA-ALA	GAA-GLU	GGA-GLY	ADENINE
	GUG-VAL	GCG-ALA	GAG-GLU	GGG-GLY	GUANINE

AMINO ACIDS :

LEU (Leucine)	GLY (Glycine)	SER (Serine)	ALA (Alanine)
GLU (Glutamic Acid)	PRO (Proline)	VAL (Valine)	THR (Threonine)
LYS (Lysine)	ARG (Arginine)	ASP (Aspartic Acid)	GLN (Glutamine)
ILEU (Isoleucine)	ASP (Asparagine)	PHE (Phenylalanine)	TYR (Tyrosine)
CYS (Cysteine)	HIS (Histidine)	MET (Methionine)	TRP (Tryptophan)
STP ("Stop")			

Although there are 64 places in the code, with rare exceptions, only 20 amino acids, and *stop,* are coded for in nature. The 64 places in the genetic code contain three-letter "words" for only 20 distinct "meanings" (more if the three *stops* are counted as "meaning"). There are 44 more codons in the genetic code than necessary to code for these 20 genetic 'meanings' so that in most cases several different codons may be used to code for a particular amino acid. This synonymy allows for considerable flexibility in the composition of DNA codes to direct the construction of particular proteins. In fact, owing to this flexibility, even small proteins can be described with astronomical numbers of alternate DNA sequences. This is the quality that biologists call the "degeneracy"[16] of the genetic code. If for some reason any part of a given DNA sequence becomes troublesome, there are many other ways that particular part can be rewritten to solve the problem without altering the protein that is to be translated.

Note that the degeneracy of the genetic code can always be used to conserve the identity of particular translation products, but not the precise sequence identity of DNA molecules that previously coded for them.

If on the other hand, DNA triplets are interpreted as codons for *numbers* rather than as codons for amino acids, then the degeneracy of the genetic code can be exploited to create supercoded DNA sequences from which an original DNA sequence can be precisely recovered. I will now describe how such a system can be used to supercode synthetic DNA molecules—such as the Milky Way DNA—for the construction of biochemically practical extrabiological databases.

So that this supercode operates with a variability similar to that of the genetic code, 64 triplets are used to represent 20 numbers in the same way that nature uses 64 triplets to represent 20 amino acids. Thus, this kind of supercode operates with the mathematical base-20 so that each triplet is used to signify a base-20 number from "0" to "J" (decimal numbers 0 through 19). Again, in order to mimic the natural degeneracy of the genetic code, these 20 numbers are mapped to the 64-place code according to the natural distribution of amino acids, with four exceptions : the triplet "CCC" is set aside to represent "C"; "UUU" ("TTT") represents "T"; "AAA" is "A"; and "GGG" is "G".

In conventional written language, the frequency of appearance of particular alphabetical characters in any body of text is governed by the specific lexical and colloquial characteristics of the language in question. The letter "e" for instance, is the most frequently used letter in the English language whereas this is not necessarily the case for any other human language. Likewise, amino acids are translated from DNA triplets according to certain species-specific frequencies of use. That is, certain DNA triplets are translated into amino acids more or less frequently depending on which species of organism the DNA comes from. Here, I decided to use the frequency of amino acids normally translated in the cells of *Homo sapiens* in order to increment triplets representing numbers "0" through "J".

In *Homo sapiens,* the approximate order of frequency of translation of amino acids and the base-20 numbers that can be logically attributed to them (quantities corresponding to frequency of appearance) are as follows (source: Dr. Jeff Spitzner)[17]:

Translation Frequency	Amino Acid	Base-20 Number
(1)	Lucine	0
(2)	Glycine	1
(3)	Serine	2
(4)	Alanine	3
(5)	Glutamic acid	4
(6)	Proline	5

(7)	Valine	6
(8)	Threonine	7
(9)	Lysine	8
(10)	Arginine	9
(11)	Aspartic acid	A
(12)	Glutamine	B
(13)	Isoleucine	C
(14)	Asparagine	D
(15)	Phenylalanine	E
(16)	Tyrosine	F
(17)	Cysteine	G
(18)	Histidine	H
(19)	Methionine	I
(20)	Tryptophan	J

These number/codon assignments might seem counter-intuitive because the largest numbers are assigned to the least frequently used codons. In general, computer databases do not contain long runs of identical values that would in turn be coded for with large number values. The result is that a degenerate supercode would more frequently represent data with smaller number values than with larger ones. Thus, since number assignments are based on frequency of use, the codons that would be more frequently used are assigned to lower number values which would tend to be used more frequently in ordinary computer code.
In the following key, base-20 number equivalents and the codons for "C", "T", "A", and "G" (hollow-body text) are expressed in a format normally used to represent the genetic code.

Key for a base-20 degenerate supercode

first place	URACIL	CYTOSINE	ADENINE	GUANINE	third place
URACIL	UUU-PHE-𝕋	UCU-SER-2	UAU-TYR-F	UGU-CYS-G	URACIL
	UUC-PHE-E	UCC-SER-2	UAC-TYR-F	UGC-CYS-G	CYTOSINE
	UUA-LEU-0	UCA-SER-2	UAA-STP-*	UGA-STP-**	ADENINE
	UUG-LEU-0	UCG-SER-2	UAG-STP-X	UGG-TRP-J	GUANINE
CYTOSINE	CUU-LEU-0	CCU-PRO-5	CAU-HIS-H	CGU-ARG-9	URACIL
	CUC-LEU-0	CCC-PRO-ℂ	CAC-HIS-H	CGC-ARG-9	CYTOSINE
	CUA-LEU-0	CCA-PRO-5	CAA-GLN-B	CGA-ARG-9	ADENINE
	CUG-LEU-0	CCG-PRO-5	CAG-GLN-B	CGG-ARG-9	GUANINE
ADENINE	AUU-ILEU-C	ACU-THR-7	AAU-ASN-D	AGU-SER-2	URACIL
	AUC-ILEU-C	ACC-THR-7	AAC-ASN-D	AGC-SER-2	CYTOSINE
	AUA-ILEU-C	ACA-THR-7	AAA-ASN-𝔸	AGA-ARG-9	ADENINE
	AUG-MET-I	ACG-THR-7	AAG-LYS-8	AGG-ARG-9	GUANINE
GUANINE	CUU-VAL-6	GCU-ALA-3	GAU-ASP-A	GUU-GLY-1	URACIL
	GUC-VAL-6	GCC-ALA-3	GAC-ASP-A	GGC-GLY-1	CYTOSINE
	GUA-VAL-6	GCA-ALA-3	GAA-GLU-4	GGA-GLY-1	ADENINE
	GUG-VAL-6	GCG-ALA-3	GAG-GLU-4	GGG-GLY-𝔾	GUANINE

A key for the conversion of base-20 numbers to decimal (base-10) numbers is provided below:

Base-20 to decimal key

0 – 0	20 – 10	40 – 20	60 – 30	80 – 40	100 – 50
1 – 1	21 – 11	41 – 21	61 – 31	81 – 41	101 – 51
2 – 2	22 – 12	42 – 22	62 – 32	82 – 42	102 – 52
3 – 3	23 – 13	43 – 23	63 – 33	83 – 43	103 – 53
4 – 4	24 – 14	44 – 24	64 – 34	84 – 44	104 – 54
5 – 5	25 – 15	45 – 25	65 – 35	85 – 45	105 – 55
6 – 6	26 – 16	46 – 26	66 – 36	86 – 46	106 – 56
7 – 7	27 – 17	47 – 27	67 – 37	87 – 47	107 – 57
8 – 8	28 – 18	48 – 28	68 – 38	88 – 48	108 – 58
9 – 9	29 – 19	49 – 29	69 – 39	89 – 49	109 – 59
10 – A	30 – 1A	50 – 2A	70 – 3A	90 – 4A	100 – 5A
11 – B	31 – 1B	51 – 2B	71 – 3B	91 – 4B	101 – 5B
12 – C	32 – 1C	52 – 2C	72 – 3C	92 – 4C	102 – 5C
13 – D	33 – 1D	53 – 2D	73 – 3D	93 – 4D	103 – 5D
14 – E	34 – 1E	54 – 2E	74 – 3E	94 – 4E	104 – 5E
15 – F	35 – 1F	55 – 2F	75 – 3F	95 – 4F	105 – 5F
16 – G	36 – 1G	56 – 2G	76 – 3G	96 – 4G	106 – 5G
17 – H	37 – 1H	57 – 2H	77 – 3H	97 – 4H	107 – 5H
18 – I	38 – 1I	58 – 2I	78 – 3I	98 – 4I	108 – 5I
19 – J	39 – 1J	59 – 2J	79 – 3J	99 – 4J	109 – 5J

To operate this supercode, the three "stop" codons, "UAA" (TAA), "UGA" (TGA), and "UAG" (TAG) are used—as they are in nature—to *stop,* or terminate the process of translation, but also in this case, to *start* the data "reading frame" of a molecule as needed. Because the most efficient use of supercode would allow for selective editing of a given sequence, I made an arbitrary decision to use two of the "stop" codons to turn on and off the supercode "editor" such that:

UAA or TAA (*) = Factor the following sequence + delete this codon

and,

UGA or TGA (**) = Unedited sequence follows + delete this codon

The third "stop" codon is used to delete an inserted sequence:

UAG or TAG (X) = delete the following sequence + delete this codon

In an example of a statement using the first codon (TAA) , the sequence, "AATCCCCCC-CCCCCCCCCCCCCCCCCCCCCCCCC" can be supercoded as "TAATCTAAATTTGGTCAACCC." Note that "TAATCTAAATTTGGTCAACCC" is only one of many sequences that can be generated by the base-20 degenerate supercode to precisely describe the sequence "AATC-CCCCCCCCCCCCCCCCCCCCCCCCCCCCCC"

Where: TAATCTAAATTTGGTCAACCC

= TAA (factor the following sequence + delete TAA)
+ TCT (2) AAA (adenine)
+ TTT (thymine)
+ GGT (1) CAA (B) CCC (cytosines) [base-20]

= (delete TAA) + 2 adenine + 1 thymine + 1B cytosine

= (delete TAA) + 2 adenine + 1 thymine + 31 cytosine [base-10]

= AATCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC [DNA]

In an example of a statement using the second codon (TGA), the sequence, "CGCTAGCTGCGATA" could be supercoded as "TGACGCTAGCTGCGATA." Here, only one correct supercode statement can be generated : "CGCTAGCTGCGATA = TGACGCTAGCTGCGATA" where :

TGACGCTAGCTGCGATA
= TGA (unedited sequence follows + delete TGA)
+ CGCTAGCTGCGATA
= CGCTAGCTGCGATA [DNA]

The third stop codon, "UAG" (TAG), is devoted to supercode not directly involved in coding for the Milky Way picture data. Its role will be detailed later in this paper.
Because "stop" codons in both the original sequence *and* in the supercode cannot be conveniently treated simultaneously, all of the "stop" codons in the original 2936-mer sequence were translated into supercoded sequences that do not contain the original "stops" (examples follow):

TAA = TAATTTTCCAAA
TGA = TAATTTGGGAAA
TAG = TAATTTGGCAAAGGG

Another option would be to use both "TGA" and "TAA" supercode statements such that:

TAA = TGATATAAAAA
TGA = TAATTTTGAGA
TAG = TAATTTAAATGAG

After supercoding all "stops" in the original sequence, any "stop" codons then directly appearing in the edited sequence are supercode "instruction statements" that will ultimately be deleted in the process of decoding. The same decoding process will recover the original set of "stop" codons and their correct positions in the 2936-mer Milky Way DNA sequence
Because it is degenerate, trillions of different supercode sequences could be composed that would all decode into exactly the same Milky Way picture data. Many of these would mimic the activity of naturally occurring DNA in living cells and would not normally be distinguishable from natural DNA. Thus, the base-20 degenerate supercode can be used to create computer databases in DNA molecules that can be manipulated and sequenced with tools currently available to molecular biology.
In addition to its role as a data handler, the base-20 degenerate supercode can be used to customize a sequence in order to minimize *in vivo* translation of supercoded databases into unwanted peptides or proteins and furthermore, to facilitate the assembly of the numerous products of DNA synthesis (oligonucleotides) into fully double-stranded, gene-sized molecules such as the Milky Way DNA. First, there is the matter of random translation:

In vivo Translation of Supercoded Databases

DNA sequences that have evolved for the purpose of directing the production of protein are highly specialized. By comparison, sequences generated by the conversion of ordinary computer files into DNA would have very little or no biological activity and the chances that any pathological agents could be biologically expressed (by transcription and translation) from such databases are extremely low. Since there is a possibility that all generic, extrabiological databases may be eventually coded into DNA sequences, and since the exact conditions leading to the *in vivo* translation of these

sequences into pathological agents cannot be anticipated absolutely, several aspects of the supercode have been exploited to limit *in vivo* translation of supercoded DNA:

"Stop" Codons

As previously noted, all supercode statements begin with "stop" codons as "instruction" or qualifying "statements" and these are distributed throughout the supercoded DNA sequence.

Moreover, unlike triplet-based "reading frames" of natural genetic instructions, the "reading frames" of supercode statements are not always constrained to be written down in sets of three characters. Because stop-initiated supercode instruction statements can frequently slip from one natural reading frame to another, stop codons can be distributed throughout multiple reading frames of supercoded sequences.

"Start" Codons

Just as specific stop codons are used to signal the end of a DNA sequence that is to be translated into protein, the beginning of a translated sequence also contains a special signal, the one to initiate translation. The "ATG" triplet (codon for methionine)[18] usually assumes the role of this special "start" codon that signals the beginning of a sequence that is to be translated into protein. To further reduce the possibilities for *in vivo* translation of random, the supercode can be used to eliminate "start" codons from all 6 reading frames of a given sequence. As a demonstration of this capability, virtually all "ATG" codons have been eliminated from the supercoded Milky Way DNA.[19]

Supercode Assists for Gene Assembly

Owing to text-length limitations set for contributions to this publications I have omitted details regarding my specific plans for assembly of the Milky Way DNA molecule from constituent parts. Note also that for the same reason I have omitted details of the structure and automated synthesis of DNA.

Various schemes for assembly of large synthetic DNA molecules with hundreds or thousands of base pairs require the use of large numbers of individually synthesized oligonucleotides. In the case of the Milky Way DNA, construction of the entire sequence involves the assembly of at least 45 individual fragments. One of the operations of DNA supercode has been reserved to assist in the repair of any errors that might occur in the process of creating such large-scale assemblies. As previously noted, two of the three stop codons ("TAA" and "TGA") are used as "instruction statements" to direct the encryption (supercoding) or non-encryption of a particular sequence. The third "stop" codon, "TAG" is dedicated to precede supercode statements that are to be intentionally deleted in the process of decoding. In the following supercoded Milky Way DNA sequence the "TAG" codon has been used to insert unique recognition sites (bold typeface below) for restriction enzymes at positions flanking each of the approximately 100-mer fragments that comprise the complete sequence. The supercode was also used to "re-state" (remove) duplicate restriction enzyme recognition sites in order to maximize the number of unique ones. These unique recognition sites allow for the convenient separation of any one the constituent fragments with the use of just one or two restriction enzymes. The cut out error-containing fragment can then be separated from the remaining error-free sequences by a method (gel electrophoresis) that separates DNA molecules according to size. Thus, if an error is found in any part of the assembled sequence, only one error-corrected 100-mer oligonucleotide fragment need be resynthesized. Re-assembly of the corrected "gene" would then involve the assembly of two larger pre-assembled fragments and one

error-corrected fragment of approximately 100 base-pairs in length. This supercoded Milky Way DNA sequence consists of 348 discreet supercode statements which all begin with either "TAG", "TAA", or "TGA" and is flanked by two 18-mer "arms" which are included to facilitate assembly into bacterial vectors .

3867 Milky Way Supercode DNA

T**GGATCC**CCC**GAAGAC**CCTGACCCAGTGTAAAAAGGATCACCCTTTTGAGCCCCTTCCCCCCCTCT
CCCAGGGGCTGACTGAGCCCCCGTA**GACGTC**TAATATGGGTAGTTAAGGCTGTCCCTCCTTTTGTC
CCTTTGGGGGCAGACCCTTTGGGAAAACACCCTTTGTGCCCAGCAAATA**GACCGGT**TAAGGCTCAC
CCTTTTGAGCCCCTTCCACCTAAAAATATCCCTGAACACTAAAAATAATATCCCTTTGCCGCACCCTTT
TA**GGGCCC**TAATCGTGTCCCTTTGCAACACCCTTTGGTCCACCCTTTTGAGCCCCTTTA**GTGCAC**TA
AAACCCCGGGG**GTATCC**CCTGATTTATGC**ATGCTAG**GTGACTGTAAAAATAGTGAATACTTAAGGTT
GGCCCTTTACGCCCTGATTACTTGCCTGCCTATAAGGGGGCAAGCCCGGGTAG**CCTAGG**TAAAAGC
CCTGATTTAAAAGCCCTTTTGAGCCCCTCATAAAATCCCTGATCATAAGGACTGCCCTGATACTGTAA
ACTCCCT**AGATCT**TGATCCTTTTTATTAAAGCAAATTTT**GATATC**TATTTATAATCTGGGTAACCAAGA
CCCTGATCCTAAAGAGGGTGAACTCCCCCCTAG**TCTAGA**TAGTGACTACTGTGACATTCCTGTAAAA
ATGACTCCACTCCCCGCCCTTCCTTCCCTCCACCGACTCCCCTTCCCCCCCTTTAG**TCATGA**TAGGG
TAATTTAGCAAATAGT**GAATGC**TAGGGTGAAACGCCGACTTAAAATCCCTGAAAAGGAAGCGGTTCT
TAAAGCAAACCCTAG**CCGCGG**TGATTAAAATCCCACGAAAGGGTGATCGTTATCACATAAAATCCCG
CATTTTGAAGGGGCCTTAAACGCCCTTTACACCCTTTTA**GTCGAC**TAAACTCCCTTTACGCCCTTTAC
CCCCTTTACGCCCTTTTGACCACTCCTCCACTCGCTCCTCCGATTCTTTCAACCAATCTTGT**AGCTTA**
GCTAAAAATAGTGAAATTAAGCGAAAGGGTGACGATATATCATTT**AGTACT**TAAGCGAAATAGTGAA**C**
AATTGA**GCGCGC**TGAATTAG**CGGCCG**TAAGCCAAAGGGTGACGAACTCCCCTTAAAAGCCCTGAAA
AAGGGGTTAAGCGAAATAGTGAGCGTACACCACTCCCCCCTTTA**GGTCTC**TAAGCGAAATAGTGAAG
GGGTATAATCTGGGTGACTCCATTCAATCCCCTTTTAAGAGAAAGTCGGGTTTTGAGTGTGTTTA**GGC**
GCCTAAAAAGGGTGAATTAAGGGTGACACCTCCAGAGGGGGAGGGCGGATAG**GGATC**TGACCAGC
GAACCAGCAGTACATCAATT**CACTGCGTG**TGATATTAAAGCAAAGGGTG**ACGCGT**TTAAGCCAAACC
CTGAGTTCGACAAGTTAAGGGAAAGGGTGAATTGTGTGTTC**GTGCAGA**GTTATATATAAAAATAGTG
ACTGTTGAGCAGGTGTGCGCTTGTGTTTCTTAAAAATAGTGAGA**GCAGC**CGCTACACCTCCACTCCT
CCCCTCCTCCCCCCGCTACTTAAGGGTGACACCGATTCTGCGCTCCTCC**TACGTA**C**ACATGT**TATAG
TAAAAATAGTGACAACTTCACCATTCATTATTTTAAGCGAAATAGTGATATAGTGAACTATAAAGCAAA
TAGTGAGCGTTTAG**CCAAGG**TAAAAATAGTGAGGCTCCTATATAAAGCAAATAGTGAGTTTATAAAAA
TAGTGATCAAACACCACTCCCCTTGATCTTAAAAATAG**TCCGGA**TGAGGGGGGGTTGTAAAAATAGT
GATCAAACTAAAAAGATCCCTGATCCGTGAGAAGGGGTTTACAATACTTAAAATCCCTAG**CCTCAGC**T
GAAAAAGGGTGAATCCTTGATATTACTTAAAATCCCTGAGGGAGGGGTCCTATAAAGCAAATAGTGAT
ACTTAACAACCCTAGT**AGGCCGGCC**TGATCAAGAGGAACGATTCTTACTTAACGACCCTTTTGACTC
GGGAAGGACTACACCGACTTAACGACCCTGATCCCGAAAGAAATAG**GCTAGC**TGACCGCACCGACT
TAAAATCCCTGAAGTTACTACCTCCACTACTCCCCTCACCCCGTTACCTACCCTCCACCGACTCCCC
CTTAG**AAGCTT**T**GAGCTC**CCACCTCCCTCCCACTCCTACTCCCCCGAAACCCTTAAGATCCCGGGTG
ACTCCAACTCCCCCAATAATCAGGGTAG**CGGTCCG**TAAAATCCCTTTTGAGCAATCACTCCCCCAGG
ATAAGCTCCCTTTCGACCCTGAGTTATAATCGAAATAGTGATACTCCCCCAAAGTAG**ATCGAT**TAAAA
TCCCTGAGGACCGCACTCC**ACTGGG**GTTTTTAAAAGCCCTGATCAATCGTACACC**ACTGGG**GTTTTT
AAAAGCCCTAG**CTCTTC**TGATTATAAAAATAGTGAAAGAACGCACCGAGGGTTTAAAAATAGTGAGCC
CTCCCTTGCGATACAACTTCACGGGGAAAACTCCCTAG**CTCGTG**TGATCCTATAG**CATATG**TAAAAAT
AGTGAGTTTCATTAAAGTCCCTGAAATCGGGAAAATTAAAGACCCTGAGGTGCATACTCCTA**GACTA**
GGTCTGACCTGGAATTAAATACCCTGATACAATCACTCCCCCGGAACCTTAAGATCCCTGATACAATC
TGAACTCCCCCGAAGTAG**CCATAGAATGG**TAAAATCCCTGATACGCTTACTCCTAATAGTGACCCGA
ATTAATTCCCCTGAACTCCAACCCCCCCGGAATAATTCCCCTA**GTGGCCA**TGAACTCCTACTCCCCC
GAATTAATTCCCCTGATCCACTACCCCCCCGAGATAAAATCCCTGATCCACCGACTTAAGATCCCTAG
CTCGAGTGAGATAAAGACCCTGAAACCCCTTAATATCCCTAATTTTGAGCCCCAAGACCCCTGTAAAA
AACGCCCTGATTGAGTAAACGCCCT**AGACTAGT**TAATTTACGCCCTTTACGCCC**GGGAC**ACCCGGG
TAAGTCCCCTTTTACCCCTGAGTTAAGTTCCCGGGGTTCCCTAATTTGTTCCCTAG**CAATTG**TAATCG
TTTGTACCCTCGTTTGTACCCTGAAGGACCCCAGGACCCCCATAAAAACCCACAGGGTGAACCCCCT
CCCTCCTTTAG**ACGCGT**TGACTTGTTTATAAAAATAGTGATTATAACCCTGAGTATAAAAATAGTGAGT
TTGATCTATAAAAATCGTTTTAAAAATAG**AATATT**TGAGTTATTTAAGAGAAATAGTGACACCTACTGA
TTTAAAAATAGTGAGATTGAATCCTGAACCTACTTAAAGACCCTGAAATAG**CACGTG**TAAAAGCCCGG

GGTTCCCGGGTGAACTCATTTATATAAAAATAGTGATTATCTATTTTAAAAATAG**CCATGG**TAGGGTA
GTGAGGCACCTATCTATTTACGTTAAAAATAGTGAGGTTAAAAATAGTGAGTTGC**CTGCA**TATTTGCG
TGCGTTAAAAATAG**TGCGCA**TAGTGAGGTGCATAACGACCCTGAAATAAACACCCAAAGATCCCT
GAACTCCGTCACCTACTTGCATT**GAATTC**ACCTTAAAAATAG**ACCTGGT**TGAGATTGAATTTATTTATCTA
TTTATCCACCTGTCTTAAAAATAGTGAGGCACCTGATGCGTATTTATTTGACACCTGTCTTAG**ACCGG**
TTAAAGCAAATAGTGACTATAAAAATAGTGATTGCTAAGGGTGACACCTAG**CGTACG**TGATGCCTATA
AAAATAGTGATTACTGAGTGTCTGTTTAG**TCGCGA**TAGGTGATTGAGCATATTCCGTTAAATACCCGA
AGGGTAGG**GTCTTC**GGG**CTGCA**GG

The supercode has also been used to install unique recognition sites that flank that part of Milky Way DNA that contains the Milky Way picture data because additional sequences are included at each terminus that are complementary to cloning vectors. These terminating sequences contain special "palindromic" restriction sites so that only one enzyme *(BpuA I)* can later be used to excise selectively only the sequence containing picture data from DNA of the cloning vector(s) into which it has been assembled.

Recovery of Visual Images

The COBE map of the Milky Way galaxy has now been compiled into a sequence of DNA bases. A DNA supercode has been employed to adjust this sequence so it will have structural and biochemical parity with the molecular apparatus of living cells. A corresponding set of oligonucleotides can now be efficiently synthesized and assembled into a DNA molecule with 3867 bases. Conventional techniques can then be used to insert this molecule into any of various cell libraries, or "biological carriers." Likewise, existing tools and techniques can be used to recover the original sequence
A variety of DNA sequencing strategies is available. The most advanced automated DNA sequencing machines in current use deliver DNA sequence information directly into computer memory. Whatever method is used, Milky Way DNA sequence information that has been recovered from a biological repository can be re-entered into computer memory in text form, and then rapidly converted back into the original image with the same (or similar) "desktop" tools originally used to create the DNA sequence from picture data. First, supercode protocols are reversed and desktop *replace* functions used to revert the supercoded 3867-mer DNA sequence into the first-generation 2936-mer. The desktop *replace* function and DNA-to-hexadecimal key can then be used to convert the first-generation DNA sequence into numerical hexadecimal data. At this point, the COBE Milky Way image can be reconstituted in 10 steps on a Macintosh desktop with the ResEdit picture editor :

1. Copy hexadecimal data as text.
2. Invoke (click mouse button on icon) ResEdit. application.
3. Invoke "CREATE NEW FILE."
4. Invoke "CREATE NEW RESOURCE."
5. Select "PICT" resource.
6. Invoke empty "PICT" file in "(FILENAME)" window
7. Invoke "OPEN USING HEX EDITOR" from active file in "PICTs FROM (FILENAME)" window.
8. Select [highlight] all data [zeros] in "PICT ID = (#)" window.
9. Paste Milky Way hexadecimal data text file
10. Close "PICT ID = (#)" [hex editor] window.

The COBE Milky Way digital video image should now appear automatically in the "PICTs FROM (FILENAME)" window. "OPEN RESOURCE EDITOR" yields a full-size image.

The Mouse's Ear

Living organisms are known to express discreet "biological periods" that accurately correspond to local planetary cycles.[20] These periods describe interrelationships of the sun-moon-Earth system so accurately that if an estimate can be made for the mass of only one of these bodies, the masses of the other two, and the distances between all three can be calculated. Newton's fundamental equation *F=ma* (Force = mass x acceleration) which describes the motion of all objects can be reconfigured to describe the motions of bodies in planetary models. Force becomes the universal force of gravity, called the gravitational constant (K).[21] Mass becomes the mass of interacting planetary bodies such as Earth-sun or Earth-moon ($M_1 + M_2$). Acceleration of an object in curvilinear (orbital) motion equals its angular velocity multiplied by the radius of curvature, or the distance between objects described (R) divided by time, or orbital period (p). Newton's law of planetary motion is written as:

$K (M_1 + M_2) = R^3 / p^2$

This law describes circular rather than elliptical orbits, but the amateur "bio-astronomer" will find this equation adequate to determine rough estimates of mass and distance where biological periods would be included as p. To find the mass of the Earth and Earth-sun distance for instance, one would first introduce an estimate for solar mass as M_1 (mass of most observable stars can be estimated by various means), and the 365-day annual period as p. Once terrestrial mass has been determined, a similar equation could be written for the Earth-moon planetary model.

For the purpose of this discussion, it is sufficient to point out that mice and other living organisms already inherently possess subtle "maps" of the local cosmos, and that an artificial gene containing astronomical information may be to some extent, redundant. At present, several methods for the creation of recombinant, or "transgenic" mice are known to biologists. One method involves the use of specially weakened retroviruses as vectors. Ordinary viruses "take over" the genetic machinery of infected cells for the purpose of creating new viruses. Retroviruses take this covert action a step further and actually insert their genes into the genomic DNA of cells they have infected. One of these, the Moloni virus, has been genetically engineered to have no pathological properties while retaining the ability to infect—and permanently integrate its genes into—host cells. Conventional techniques are used to "cut-and-splice" foreign DNA into the DNA of the Moloni virus. Biologists now routinely use the Moloni, and other retroviruses to insert experimental genes into the genomic DNA of laboratory mice.

Another method, called "oocyte injection," involves the micromechanical and biological manipulation of mature egg cells, which are subsequently fertilized and surgically implanted in the uterus of a surrogate female. The first step in this method calls for the removal of mature egg cells (oocytes) from the ovary (in this case, from the ovary of a mouse). At about 1 millimeter in size, mouse oocyte cells are large enough to be seen with the naked eye. Then, with the aid of a microscope, pure ("foreign," or synthetic) DNA is directly injected into oocytes using very small glass tubes called micropipettes. Once foreign DNA has been injected into an oocyte, it is somehow permanently integrated into one of the cell's chromosomes [the exact details of this process of integration are still not completely understood]. The artificially manipulated oocyte is fertilized *in vitro,* and then surgically implanted into the uterus of a surrogate female mouse. From this point on, the transgenic embryo develops normally. The offspring of the surrogate mouse are screened for the presence of the new gene, and a pure strain of mice that carry the gene is produced with traditional techniques of animal husbandry.

Notes

1 Deoxyribonucleic acid (DNA) that is identical in structure and function to DNA that occurs naturally in biological organisms can be created artificially by chemical synthesis.

2 "Microvenus", *Art Journa.* Spring, 1995

3 "'Genetic Art' builds cryptic bridge between two cultures," *Nature.* No.378. p229. 1995

4 The "DNA bases" are explained on p. 221 pp of this paper.

5 The largest synthetic DNA molecule I have found to date is one that Midland Certified Reagent Co. Molecular Biology Group (3112 Cuthbert Ave., Midland, TX 79701) has constructed, a synthetic gene with 7000 DNA bases for a Boston-area biotechnology firm (unpublished).

6 Plasmids are autonomous, virus-like entities that themselves contain whole collections of genes.

7 COBE results also included important findings not discussed in this paper, including a map of the cosmic microwave background that has profoundly influenced scientific theories about cosmology and the primordial ("big bang") event.

8 The COBE image coded into the Milky Way DNA is a "never before seen" near infra-red image of the Milky Way. The image was compiled from a combination of data gathered with COBE's Diffuse Infrared Background Experiment (DIRBE), one of three separate COBE scientific experiments. Data for the image was gathered with DIRBE's liquid helium-cooled detectors at intervals within the first six months in orbit and released in April 1990. It shows the Milky Way from an edge on perspective with the galactic north pole at top, south pole at bottom and galactic center at the center. The image was collected from vantage points within our own solar system which lies close to the galactic plane. The picture combines images obtained at several near infrared wavelengths. The dominant source of light at these wavelengths is from stars within our own galaxy. Even though our solar system is part of the Milky Way, the view looks distant because most of the light comes from the population of stars that are closer to the galactic center than our own sun. No image of the Milky Way galactic disc and spiral arms has ever been made because vantage points needed for the collection of such data are many thousands of light years distant.
The COBE spacecraft was launched on November 18, 1989 on board the last NASA-owned Delta rocket from Vandenberg Air Force Station, CA. COBE was specifically designed to study radiation from the "Big Bang." (Source: NASA COBE gif comments)

9 Double-stranded DNA molecules are synthesized one strand at a time.

10 For reasons that are not completely understood, poly-Cs are "skipped over" in conventional acrylamide DNA sequencing gels.

11 Biological processes of *transcription* and *translation* will be more fully described later in this paper

12 The fact that "junk" DNA does not undergo continuous chemical manipulation involved in protein translation seems to be associated with the fact that junk DNA is "conserved," that is, it is highly unlikely to undergo directed editing, or mutations of the kind that that ordinary DNA is subjected to.

13 In the biological process of transcription, an enzyme called RNA polymerase is attached to double-stranded DNA molecules forming a complex called a "transcription bubble". The RNA polymerase forms a single strand of mRNA that is mapped to only one strand of the "parent" DNA. That is, a single-stranded mRNA molecule is created that is a template made from only the 5'-to-3' side of a parent DNA molecule

14 Lehninger, 1975

15 In addition to the standard 20 amino acids, several others of relatively rare occurrence have been isolated in some specialized types of proteins. These include hydroxyproline, hydroxylysine, desmosine, and isodesmosine. Several very unusual methylated amino acids have been found in certain muscle proteins including methylhistidine, methyllysine, and trimethyllysine. All of these are derivatives of some standard amino acid. Over 150 other amino acids are known to occur biologically either individually or in combined form, but never in proteins.

16 Mathematicians also use this term to describe problems that have a variety of solutions. Instead of a single correct solution that is finite point or number, a *degenerate* solution is a "plateau" of correct points or numbers.

17 Note that "stop" codons are the least frequently translated DNA triplets, and fall into 21st place, behind the 20 amino acids.

18 In rare cases, several other codons may also be used to signal the start of protein translation in certain organisms. The codons CTG and GTG may sometimes be used as "start" codons.
19 Only one "ATG" codon in the Milky Way DNA sequence was not supercoded because it lies in a unique recognition site for a restriction enzyme. This single remaining ATG codon is flanked by nearby "stop" codons so that only a few amino acids can be translated.
20 Biological analogs have been found for the 365-day annual period that describes the Earth's orbit around the sun ; the 27.3 day "monthly" period that corresponds to the period of the moon's orbit around the earth; the 24-hour diurnal cycle (period of the Earth's rotation on its own axis) ; and the "circadian rhythms" which are actually seasonal periods of light and darkness that vary according to the tilt of the Earth's axis. Although some organisms manifest these periods more profoundly than others, it is probably safe to assume that astronomical periods are reflected in the "biological clocks" of all living organisms.
21 6.670×10^{-8} dyne cm^2 / gm^2

References

Joe Davis. "Microvenus," *Art Journal*. Spring, 1995

Steve Nadis . "'Genetic Art' builds cryptic bridge between two cultures," *Nature.* 378 229. 1995

Albert L. Lehninger. "Biochemistry," (second addition), Worth Publishers, Inc., New York, NY, 1975

Roland Brousseau, Wing Sung, Ray Wu, and Saran A. Narang. "Synthetic Gene Assembly, Cloning and Expression," *Synthesis and Applications of DNA and RNA* (Ed. by Saran A. Narang.) Academic Press, Inc; Harcourt Brace Jovanovich, Publishers, Orlando, Florida, 1987

K.L. Agarwal, H. Buchi, M. H. Caruthers, N. Gupta, H. G. Khorana, K. Kleppe, A. Kumar, E. Ohtsuka, U. L. RajBhandary, J. H. van de Sande, V. Sgaramella, H. Weber, and T. Yamada, *Nature.* No. 227. pp.27–34. 1970

H. G. Khorana, K. L. Agarwal, P. Besmer, H. Buchi, M. H. Caruthers, P. J. Cashion,

M. Fridkin, E. Jay, K. Kleppe, R. Kleppe, A. Kumar, P. C. Loewen, R. C. Miller, K. Minamoto, A. Panet, U. L. RajBhandary, B. Ramamoorthy, T. Sekiya, T. Takeya, and J. H. van de Sande, *Journal of Biological Chemistry.* No. 251. pp. 565–570. 1976

Alberto Di Donato, Mena de Nigris, Nello Russo, Sebatiano Di Biase, and Guiseppe D'Alessio, "A Method for Synthesizing Genes and cDNAs by the Polymerase Chain Reaction," *Analytical Biochemistry.* No. 212. pp. 291–293. 1993