YOUVAN

John N. Augustine, Mail Stop 921-107
Unsolicited Proposal Manager U.S.
Department of Energy National Energy
Technology Laboratory 626 Cochrans
Mill Road P.O. Box 10940 Pittsburgh, PA
15236-0940 Email:
DOEUSP@NETL.DOE.GOV

February 5, 2015

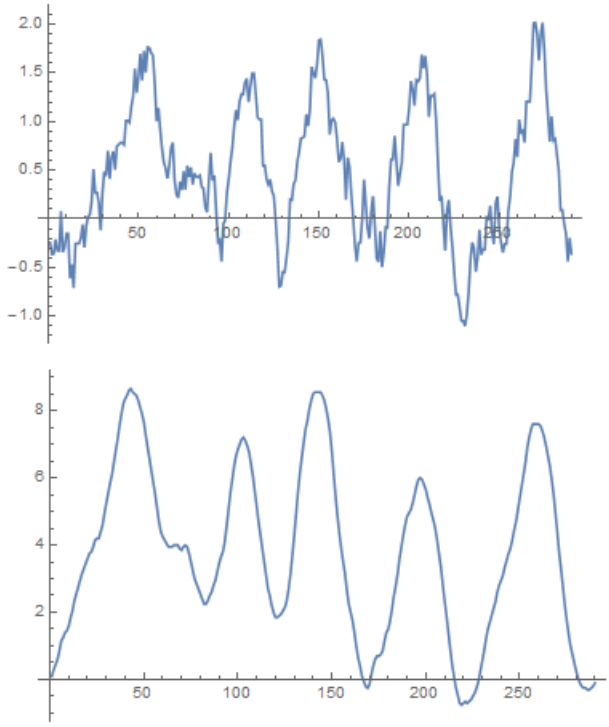
DOEUSP:

Youvan Inc. herewith submits a grant proposal, "The Structure of the Genetic Code / SAGC". SAGC is an acronym for the *Search for an Alternative Genetic Code*. We will be looking in new environmental metagenomics libraries for evidence of the use of a non-standard genetic code. This is a very high risk endeavor. If we fail, other goals will be accomplished: 1) The genetic code is likely to be universal; therefore, protein sequences deduced from giga-base sequencing are likely to be correct. 2) Our methods use *Mathematica* for all programming. We find this platform highly advantageous for research level molecular biology combined with research level mathematics for use in genomics. Libraries of software functions are being developed. The overall structure of the code (the exact placement of codons) is being decrypted.

There are 14 months of programming and writing in this proposal. It will lead to a virtual book with source code. One should look at our work as partially academic – not necessarily "all for-profit". If there is a commercial effort, it will be to find better algorithms, write better code, and run on high-end desktop computers, thus taking the load off supercomputers. If we went to market today, we would use an Intel i7 (4 nodes) with 32 GB of RAM, and 1 TB of solid state drive. The computer would be preloaded with *Mathematica* and our code. Total cost would be about $7,000. As my work progresses, I would be very pleased to deliver a software-loaded computer to a DOE lab – as I have done previously under a NASA grant. On certain problems involving SVD, I am already confident that I can xx - redacted- xx computer with a desktop computer at the "**sweet spot**" for consumers – as described above. That involves the "YI" pseudoinverse, described in this proposal.

Here is an example of such an approach. For genomic-sized hydropathy plots, I was faced with supercomputing. Instead, I wrote a new algorithm based on H = T – A, assigning +1 to T in the second position of the codon, -1 to A, with all other 10 positions / nucleotides set to zero. A moving average of

20 codons was taken, *without translation,* to calculate hydropathy. The figure (below) shows the M subunit from the photosynthetic reaction center, known to have 5 transmembrane alpha helices. The upper plot is from the standard Kyte and Doolittle method based on 20 amino acid parameters. The lower plot is our H=T-A method, using only half the second position of the codon, or 1/6th of the total DNA sequence.



Kyte & Doolittle Hydropathy Plot

20 Amino acid parameters

Up to 20 more positional weights

Discovered **1983**

Youvan H = T – A Plot

2 Parameters using 1/6th of the DNA

No translation into protein

Discovered **2015**

There are several other notable results in this proposal. We found the genetic code to be a matrix of tuples and discovered that it is a rare type of over-determined matrix with an *exact solution*, thus obviating the need for Singular Value Decomposition. SVD scales as a cubic in computation and as a quadratic in memory. This type of scaling is usually considered unusable, but there has been no alternative method. Again, this falls into a paradigm of using research level mathematics and molecular biology with better algorithms and better code to reduce computational load. Such computers are much more affordable for *both research and teaching labs*.

I have an excellent relationship with Wolfram Research, makers of *Mathematica*. Their sales now total in the millions of units, some being site licenses. xx - redacted- xx The theme in *Mathematica* is that “everything is a list”. What could be more harmonious with genomics? With the recent release of Version 10 came parallelization, cloud computation, and further development of functions. Most of known mathematics is covered. Their virtual book, too big to be printed, is how I learned discrete mathematics and programming after I was crippled by a violent crime.

Please do not discount this proposal because of the style of writing. There is a huge deficit in writing and publication on this subject. I am already at a stage where a 300 – 500 page virtual book needs to be written. Therefore, this proposal had to fill in for 25 years of missing work. That goes back to the time when we first ran SVD and created a 12 parameter hydropathy plot directly from the DNA sequence at MIT. That paper is attached.

xx - redacted- xx

We have substantial cost sharing in this proposal. I owe a lot to DOE and NIH for my education and lifetime grant support. Now, I try to reenter the field and become productive again. My limitations are only physical. I can't travel and I do most of my work from bed. However, we have prepared a new facility where I can work comfortably and provide typical office space for employees xx - redacted- xx

Sincerely,

Doug Youvan

Douglas C. Youvan
Principal Investigator and President
Youvan Inc.

Unsolicited Proposal Cover Sheet

Proposal Receipt Date _____ USP# _____

DOE Amount Requested: $155,000 per year

TIN: 47-2694183

Future Proposed Cost Share ~ 40%

Plus 14 months of B&P and IR&D time, completed ~ $180,000

Total Project Value ~ $250,000 per year

Congressional District: KS 02

Organization Name: Youvan Inc.

101 S. Cayuga

Frontenac, KS 66763

Contact Information: doug@youvan.com 620-875-0108

Douglas C. Youvan, Principal Investigator

Proposal Title: “The Structure of the Genetic Code / SAGC”

Proposed Duration: 36 months

Organization Type: 8

Support Type: 4

Socio-Economics Type: 2 and 3

This proposal does contain proprietary information.

This proposal may be subjected to external review.

No other organizations are considering this proposal.

. Doug Youvan

_____

Douglas C. Youvan, President

Youvan Inc.

DUNS: 07-967-5766

SAM Active: 079675766 / 7AXH3

Budget and Financials

Youvan Inc. was formed on January 6, 2015 as a Kansas Subchapter S Corporation. We have attempted to reduce the cost of this grant to DOE by approximately 40% through Youvan Inc.'s contributions of funds and services:

- xx - redacted- xx
- Doug Youvan will contribute 25% of his time at no charge.
- **No overhead**, whatsoever, will be charged to DOE with the exception of employer FICA and employee health insurance. Facility and facility charges, equipment, office supplies, upkeep, computers, software licenses, bid and proposal time, insurance, property taxes, consulting, etc. will be paid by Youvan Inc.

We request:

Direct:

75% of Youvan's salary @ $180,000 per year	$ 135,000

Indirect:

Employer FICA	$ 7,000
Health Insurance	$ 13,000
Total per annum	$ 155,000
Total (3 years)	$ 465,000

Facilities

Because of Dr. Youvan's physical disabilities, a 2,000 square foot building was built at a cost of $300,000 that is attached to his home. The facility, its computers, office equipment, utilities, grounds, and upkeep are contributed for work on this grant at no cost to the DOE. Youvan's home is securely separated from this facility such that it is appropriate for employees (in progress). Youvan is basically house-bound, and he can move from his office to his home through a secure door - all on one level. The same property has three external buildings that can be used for expansion. All of this is accessible within a two acre plot directly off the main street of the small town (population 3,000) of Frontenac, KS. Pittsburg State University (PSU), a regional university with 6,000 students, is six miles away. Youvan received his B.S. in Biology from PSU in 1974, so it is an excellent source of work/study students.

Publications

Immediately after the 2003 injuries, I began to retrain in discrete mathematics via *Mathematica*.

I am missing publications, but not the work, from about 2001 to 2015. Self-published virtual e-books appear to be the only resolution to this problem. One book is already finished: "Pseudocolor in Pure and Applied Mathematics".

The work is broader than the title would suggest. The entire book needs to be brought up to the current version of Mathematica (10).

My current biography is on Wikipedia. A full biography pre-year 2000 is available on www.youvan.com .

The Structure of the Genetic Code / SAGC

Abstract

Studies of the pure and applied mathematics of the genetic code are an ongoing process, wherein the code is treated as a mathematical entity and attention is paid to advances in molecular biology. Having solved the genetic code as an over-determined matrix, we will investigate other sparse matrices and variations on a method that shortcuts the use of Singular Value Decomposition (SVD) – but gives exactly the same solution. In applied mathematics, we will develop an algorithm to survey the ever-increasing DNA sequence database for protein genes that appear to be encoded by an alternative genetic code. This effort is the **S**earch for an **A**lternative **G**enetic **C**ode (**SAGC**). If an alternative genetic code (AGC) were to exist within the environmental metagenome, we anticipate the most likely major alteration in the code would be a switch of codons involving A/T with G/C at the second position of the codon, thus maintaining the overall high correlation of the structure of the code with protein hydropathy. We believe that even if negative results are obtained, we will have at least insured that the proteome is being correctly translated. Negative results also imply the genetic code is universal, and this greatly simplifies genomics. Perseverance in continuing to run this algorithm on all new metagenome data is essential far into the future, as new niches of unculturable microbes continue to be sequenced (Chaffron *et al*., 2010; http://www.cm2bl.org/samples.html). There are diverse research initiatives and even commercial companies that are now actively engaged in these studies (http://www.broadinstitute.org/scientific-community/science/projects/microbiome-projects; http://metagenombio.com/) Additionally, identification of Complementary Proteins (CPs), found from *E. coli* to humans, with simultaneous translation of one sequence of DNA into two functional proteins (-1 and +1 frames) gives us an opportunity to look at an ingenious coding strategy. CPs also provide constraints for a better understanding of the structure of the genetic code. If the genetic code is just a “frozen accident”, we should realize that it is a very special frozen accident. Perhaps a better view is that the code froze after a high degree of optimization, with the last step locked by the code’s ability to encode CPs with reversed and inverted protein hydropathy. Other applied work includes the development of an evolution program that treats the genetic code as an ‘intelligent look up table’ (iLUT) that can optimize its layout by mutation and recombination within an evolving pool of fit codes.

Introduction

If a Molecular Biology "luminary" had said that the structure of cytochrome c was a "frozen accident" and people believed him, what would it have done to our understanding of protein structure and function? This was certainly easier to believe with the genetic code, where one change would have been lethal, and there was no apparent reason for any of the layout except for wobble (Alkatib *et al*., 2012; Demeshkina *et al*., 2014). Anyone attempting to explain the structure of the code might seem to be trying to find evidence for Intelligent Design. This remains unacceptable. I look at the structure of the code for purposes of discovery. With gigabases of sequence from unculturable metagenomic microbes (Lombardot *et al*., 2006), we can't do experiments on individual organisms. And in the face of this mountain of data, one tantalizing question arises: Is the genetic code universal? The task of answering this question falls to theorists to see whether consistent hypotheses can be developed (Jestin & Kempf, 2009; Koonin & Novozhilov, 2009; Szostak, 2015). (It should be noted here that by using the term "alternative code", I am not generally referring to organisms wherein there are only a few codon reassignments; see, e.g., McCutcheon *et al*., 2009; http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi .)

With tens of billions of bases sequenced, it is hard to understand why the genetic code is consistently under-studied. After the discovery of inverted hydropathy in Complementary (or "Antisense") Proteins, I would have expected more protein chemists to seize on this new window into the structure of the code (see, *e.g*., Chou *et al*., 1996; McGuire & Holmes, 2005). This is one of the reasons that a book on *The Structure of the Genetic Code* is very much needed. As this proposal shall indicate, I am not finished with the formulation. On a five year time schedule, I am compelled to finish this work and write at an understandable level for both mathematicians and molecular biologists.

You will see in this proposal that hydropathy plots (of the photosynthetic reaction center proteins) and the underlying Kyte and Doolittle hydropathy values (Kyte and Doolittle, 1982) have come into play several times in the past 30 years, leading me through a process of better understanding the genetic code. It turns out that hydropathy appears to be the major organizing factor in the genetic code (Figure 1). Serendipitously, while starting to write this proposal, I found the photosynthetic reaction center L and M subunits have full length ORFs in the -1 frame (antisense). These putative proteins have reversed and inverted hydropathy plots relative to the known reaction center subunits. The possible significance of this finding will be discussed below.

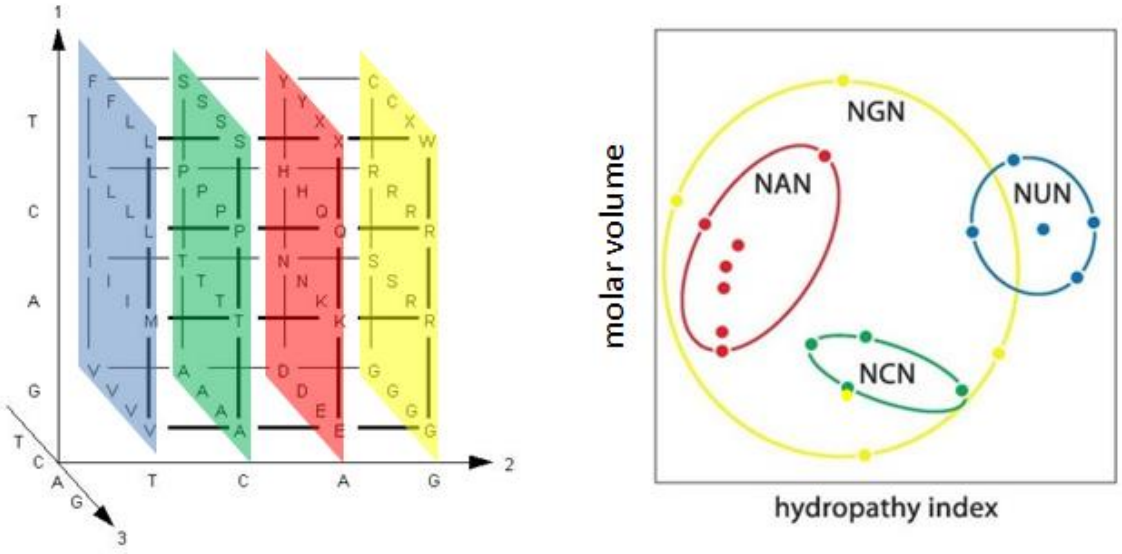


Figure 1. The genetic code (left) and the various codons mapped into amino acid hydropathy vs. molar volume space (right). Whether coincidental or not, patterns in the genetic code are useful. As early as 1986, we exploited patterns in the code to synthesize combinatorial mutagenesis cassettes that targeted particular groups of amino acid residues. For example, mixed synthesis of an oligonucleotide carrying NTN gives five hydrophobic residues with DNA complexity reduced from 64 to 16. That in turn was generalized to "target sets" of amino acids constructed from phylogenies or prior mutagenesis experiments to generate a higher throughput of mutants. The computer program we developed to aid in this targeting (CyberDope) found its best application in therapeutic peptides. As part of our genomics effort, CyberDope has already been rewritten in Mathematica as CyberDopant. All of this comes into play again, as researchers continue to discover Complementary Proteins synthesized from both the +1 and -1 frames on opposing DNA strands. The main determinants of the structure of the protein coded as a reverse complement in the -1 frame are the hydropathy difference (NAN versus NTN) and the combined values for *molar volume x hydropathy* space (NCN versus NGN). These patterns also give us an idea of how codon assignments could feasibly be switched if an alternative genetic code were to be found.

Reading Frames and Hydropathy Inversion

Inversion of hydropathy is a direct consequence of the intrinsic symmetry of the genetic code, and as I reviewed the literature, it became clear that a significant percentage of proteins, found in organisms from *E. coli* to man, have these so-called Complementary Protein pairs. As you might imagine, if DNA codes for two proteins simultaneously from the +1 and -1 frames, the constraints placed on the genetic code would help a theorist see more clearly into the nature of the code. See Figures 2-5 for further illustration.

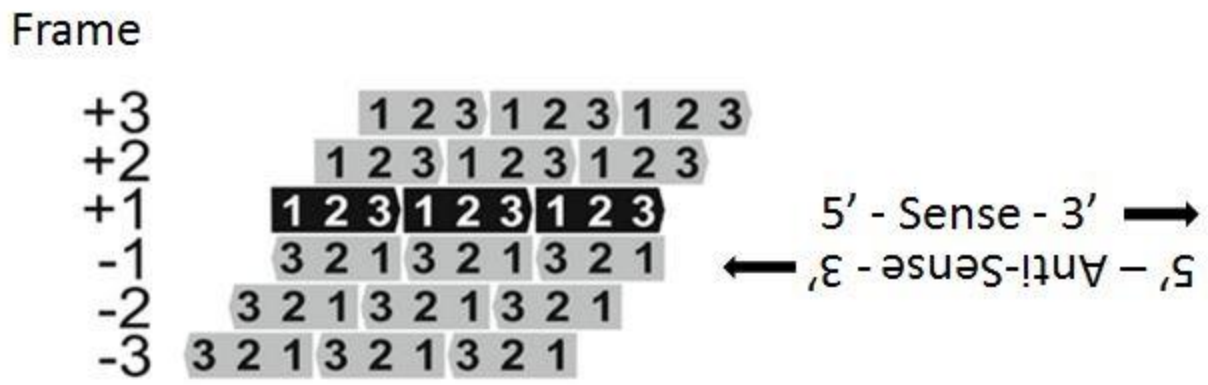


Figure 2. The effect of reading frame on coding patterns and hydropathy inversion. (Modified from www.alterorf.cl) Six reading frames are shown, with the nucleotides grouped into three codon triplets per row. The notation to the right side is intended to show what the terms 'reversed' and 'inverted' mean for the sense (black) and anti-sense hydropathy plots of 'Complementary Proteins'. The effect of sequence reversal in the numerical plot is exactly the same as in the text example. Note that frames +1 and -1 are not staggered with respect to each other; however, there is a reversal in the direction that the double helix is read. There are a number of publications that discuss the frequency, identity, and function of anti-sense RNAs, and that identify some proteins known to be expressed in this way. (See, *e.g.*, http://www.alterorf.cl/Publications/Publications.htm)

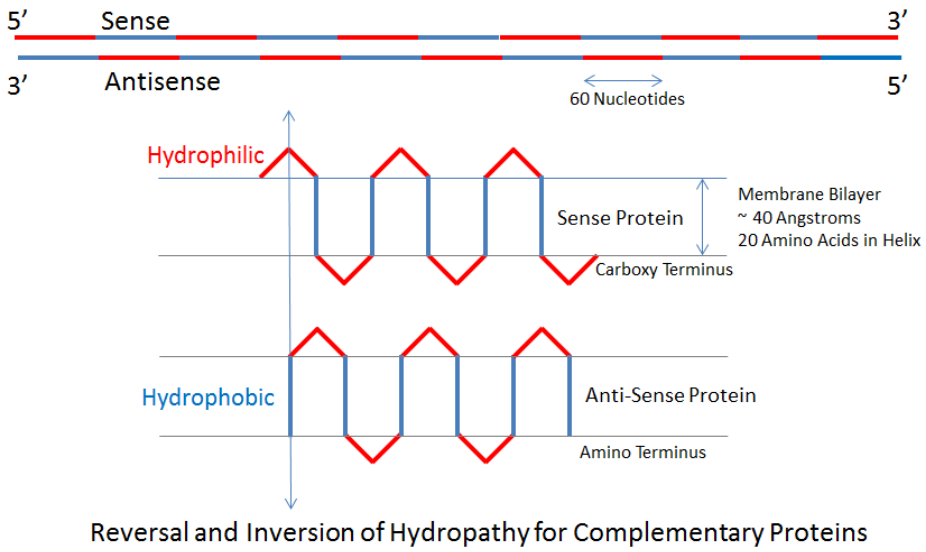


Figure 3. Reversal and inversion of hydropathy for complementary proteins. Red color is used here to indicate the DNA sequence that encodes hydrophilic residues, as well as the protein sequence that contains the hydrophilic residues amino acid residues. The same principle is used in blue for hydrophobic regions. The long vertical line with arrows indicates that the alignment of the sense and anti-sense proteins is 180 degrees out of phase. In a hydropathy plot, this causes inversion (see below). Also note the reversal in sequence between sense and anti-sense that is simply due to the antiparallel encoding of genes and proteins. For a globular protein, these same transformations would result in the

hydrophobic interior and the hydrophilic exterior exchanging polarity and turning inside out in the antisense protein (not shown).

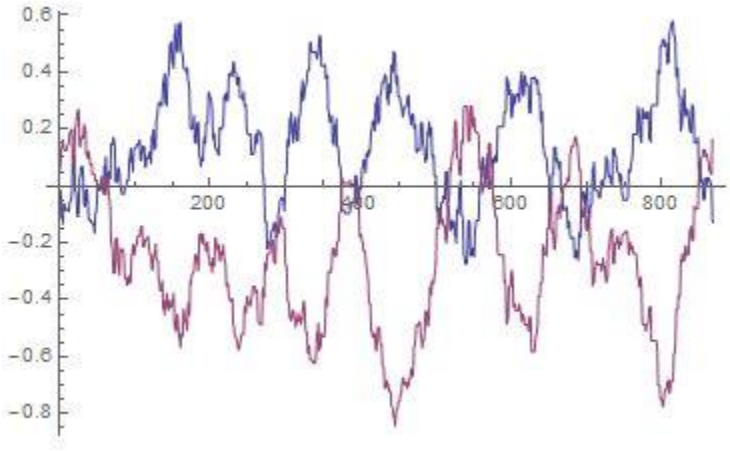


Figure 4. ***Rhodopseudomonas capsulatus*** **(original name) reaction center M subunit hydropathy plot.** Deduced polypeptide sequence for the +1 (sense) frame is blue and the -1 (antisense) frame is magenta. Left-to-right reading of the blue line (sense, +1) is the canonical amino- to carboxy-terminal direction for this type of plot, with the axis labeled in residue number. However, the magenta line (antisense, -1) is reversed and reads in the carboxy to amino direction. This is the 'reversal' aspect of a Complementary Protein pair. The 'inversion' of hydropathy is a reflection through the x-axis. A separate plot (not shown) from my program shows that stop codons are generated in all of the other four reading frames and that there is nothing remarkable about their hydropathy plots.

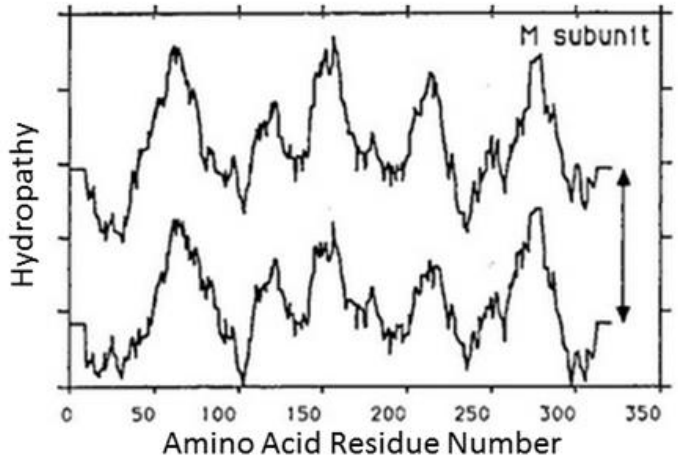


Figure 5. Amino acid- versus nucleotide-determined hydropathy plots for a membrane protein. In our 1990 publication (Yang *et al*., 1990; attached as an appendix), we showed that the hydropathy plot of a protein could be determined directly from its DNA sequence, without translation. In addition, reducing the 20 Kyte and Doolittle values to 12 nucleotide-determined hydropathy (NDH) values (without loss)

shows that the genetic code is organized for hydropathy. Displayed here are the running averages of the hydropathy values for the M-subunit of the photosynthetic reaction center, known by X-ray crystallography to have five transmembrane alpha helices. The upper trace is a standard K&D plot using the values for the actual amino acids. The lower plot instead uses the 12 NDH values, displaced for clarity.

In Appendix 1, I have included perhaps the most important medical research paper (abstract) to date on Complementary Proteins. Pendergraft *et al*. (2004) have a detailed mechanism and experimental evidence for a CP being involved in auto-immunity. They are not the only group to find such a relationship. It would certainly be interesting to know whether the expression of these proteins could be triggered by mutagens or radiation.

Nucleotide-Determined Hydropathy Values

Nucleotide-determined hydropathy can be represented by a 64 x 12 matrix, as shown in Figure 6 (below).

1 0 0 0 1 0 0 0 1 0 0 0	k
1 0 0 0 1 0 0 0 0 1 0 0	n
1 0 0 0 1 0 0 0 0 0 1 0	k
1 0 0 0 1 0 0 0 0 0 0 1	n
1 0 0 0 0 1 0 0 1 0 0 0	t
1 0 0 0 0 1 0 0 0 1 0 0	t
1 0 0 0 0 1 0 0 0 0 1 0	t
1 0 0 0 0 1 0 0 0 0 0 1	t
1 0 0 0 0 0 1 0 1 0 0 0	r
1 0 0 0 0 0 1 0 0 1 0 0	s
1 0 0 0 0 0 1 0 0 0 1 0	r
1 0 0 0 0 0 1 0 0 0 0 1	s
1 0 0 0 0 0 0 1 1 0 0 0	i
1 0 0 0 0 0 0 1 0 1 0 0	i
1 0 0 0 0 0 0 1 0 0 1 0	m
1 0 0 0 0 0 0 1 0 0 0 1	i
0 1 0 0 1 0 0 0 1 0 0 0	q
0 1 0 0 1 0 0 0 0 1 0 0	h
0 1 0 0 1 0 0 0 0 0 1 0	q
0 1 0 0 1 0 0 0 0 0 0 1	h
0 1 0 0 0 1 0 0 1 0 0 0	p
0 1 0 0 0 1 0 0 0 1 0 0	p
0 1 0 0 0 1 0 0 0 0 1 0	p
0 1 0 0 0 1 0 0 0 0 0 1	p
0 1 0 0 0 0 1 0 1 0 0 0	r
0 1 0 0 0 0 1 0 0 1 0 0	r
0 1 0 0 0 0 1 0 0 0 1 0	r
0 1 0 0 0 0 1 0 0 0 0 1	r
0 1 0 0 0 0 0 1 1 0 0 0	l
0 1 0 0 0 0 0 1 0 1 0 0	l
0 1 0 0 0 0 0 1 0 0 1 0	l
0 1 0 0 0 0 0 1 0 0 0 1	l
0 0 1 0 1 0 0 0 1 0 0 0	e
0 0 1 0 1 0 0 0 0 1 0 0	d
0 0 1 0 1 0 0 0 0 0 1 0	e
0 0 1 0 1 0 0 0 0 0 0 1	d
0 0 1 0 0 1 0 0 1 0 0 0	a
0 0 1 0 0 1 0 0 0 1 0 0	a
0 0 1 0 0 1 0 0 0 0 1 0	a
0 0 1 0 0 1 0 0 0 0 0 1	a
0 0 1 0 0 0 1 0 1 0 0 0	g
0 0 1 0 0 0 1 0 0 1 0 0	g
0 0 1 0 0 0 1 0 0 0 1 0	g
0 0 1 0 0 0 1 0 0 0 0 1	g
0 0 1 0 0 0 0 1 1 0 0 0	v
0 0 1 0 0 0 0 1 0 1 0 0	v
0 0 1 0 0 0 0 1 0 0 1 0	v
0 0 1 0 0 0 0 1 0 0 0 1	v
0 0 0 1 1 0 0 0 1 0 0 0	x
0 0 0 1 1 0 0 0 0 1 0 0	y
0 0 0 1 1 0 0 0 0 0 1 0	x
0 0 0 1 1 0 0 0 0 0 0 1	y
0 0 0 1 0 1 0 0 1 0 0 0	s
0 0 0 1 0 1 0 0 0 1 0 0	s
0 0 0 1 0 1 0 0 0 0 1 0	s
0 0 0 1 0 1 0 0 0 0 0 1	s
0 0 0 1 0 0 1 0 1 0 0 0	x
0 0 0 1 0 0 1 0 0 1 0 0	c
0 0 0 1 0 0 1 0 0 0 1 0	w
0 0 0 1 0 0 1 0 0 0 0 1	c
0 0 0 1 0 0 0 1 1 0 0 0	l
0 0 0 1 0 0 0 1 0 1 0 0	f
0 0 0 1 0 0 0 1 0 0 1 0	l
0 0 0 1 0 0 0 1 0 0 0 1	f

Figure 6. Nucleotide-determined hydropathy matrix. A triplet codon with an alphabet of four bases requires 12 binary digits for full separation of variables (3 positional times four alphabetic). The genetic code then requires 64 rows of these twelve binary entries yielding a 64 x 12 matrix. I believe this is the only logical representation of the genetic code, but other than in my 1990 publication, it seems not to be used. The n x m dimensions for this matrix are 64 x 12, which is determined by only two factors: the alphabet size, *a* = 4, and the word length, w =3, such that it is (*a*^*w*) x (*a* x *w*). The most succinct binarization of the genetic code (left) uses 12 columns (3 positions x 4 nucleotides) and 64 rows (codons) to properly separate the variables (X1 to X12). These 'X values' are the column headings set left to right in a format of first position (A, C, G, T), second position, and third position of the codon. The matrix on the right is a 'properties matrix', and in this case it shows the hydropathy values of Kyte and Doolittle for the R-group of the amino acids, except for the three stop codons, which are padded with a mean value of zero.

Table 1. SVD Determined Values*

Hydropathy (C = 0.92)

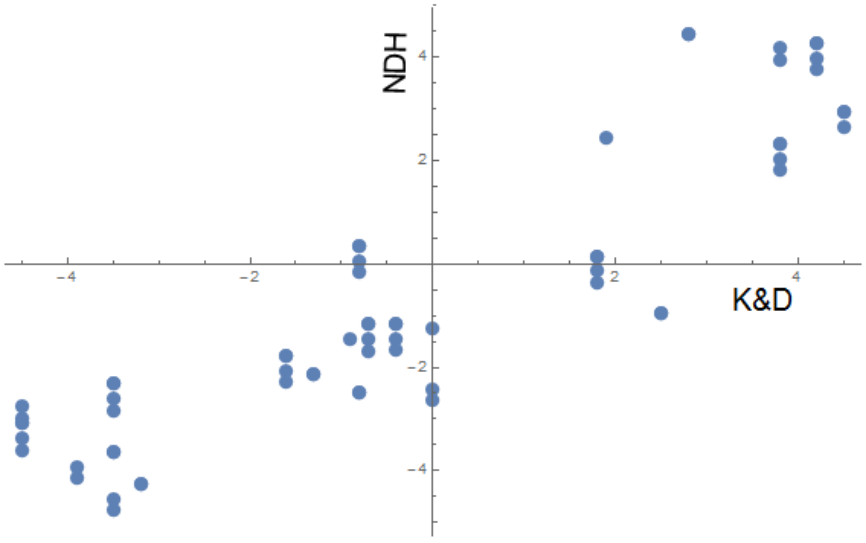
	A	G	C	T
Codon Position 1	-0.57	0.76	-1.18	0.52
Codon Position 2	-2.92	-1.48	-0.09	4.02
Codon Position 3	-0.35	-0.50	0.19	0.19

Molar Volume Å^3 (C = 0.78)

	A	G	C	T
Codon Position 1	48.4	12.7	62.2	58.1
Codon Position 2	58.6	37.3	13.4	72.0
Codon Position 3	48.3	53.6	39.7	39.7

* SVD determined nucleotide values for hydropathy **[h]** and molar volume **[v]** using 61 non-stop codons. These vectors were solved using the subroutine "svdcmp" found in reference 14. All 61 nucleotide-determined hydropathy or molar volume values can be generated from the twelve elements of **[h]** or **[v]**, respectively. For example, the nucleotide-determined hydropathy value for the cysteine codon TGC (binary format: 000101000010) is 0.52 + (-1.48) + 0.19 = -0.77. A nucleotide-determined value for an amino acid possessing several different codons is determined by calculating a weighted average (corresponding to the degree of degeneracy ≤ 6). The correlation coefficient (C) is equal to unity for an exact linear correlation between the nucleotide-determined values and the corresponding physico-chemical values for the amino acid residues (plots not shown).

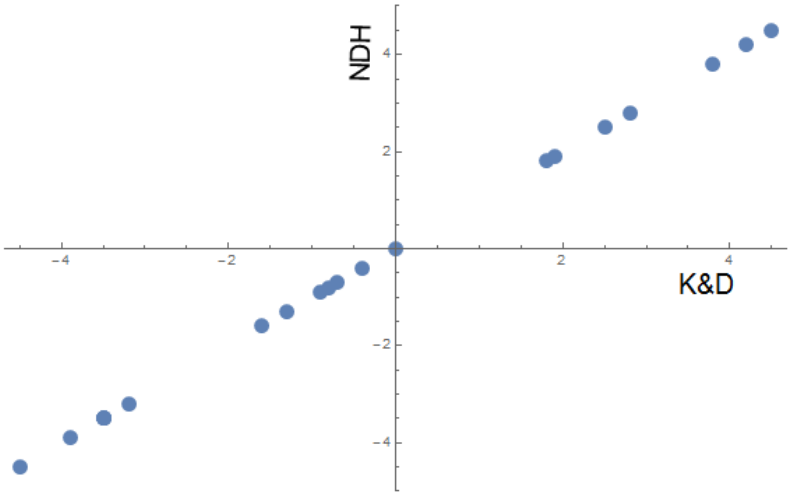
Figure 7. SVD-determined values for amino acid hydropathy and molar volume as applied to codons. The strong correlation between hydropathy and the second codon position (from our 1990 SVD publication, Yang *et al*., 1990). Mary M. Yang was the programmer, but the original code (written in MFC C++) is no longer available. I wrote a Mathematica program *de novo* and obtained exactly the same results. Most importantly, A and T in the second position have Nucleotide-Determined Hydropathy (NDH) values of -2.92 (hydrophilic) and +4.02 (hydrophobic) in K&D units. Correlation with molar volume was poor, but I have since found an excellent correlation for 2nd position G and C with the product of hydropathy times molar volume, essentially an area as swept out in Figure 1 (right side). The small print gives instructions on how to re-create an amino acid's hydropathy using these 12 values. The fact that the 20 K&D values could be reduced to 12 NDH values is evidence in itself that the code has a hydropathy structure that is additive at the nucleotide level. The full paper is Appendix 3.




```
{intercept → -0.528, slope → 0.816}

c = {0.903}
```

Figure 8. Example of the output of a Mathematica program designed to calculate the hydropathy values of all 61 codon (minus 3 stop codons) using the 12 NDH values. Slight differences in the correlation coefficient arise from the fact that the stop codons were included and padded with zeros.




```
{intercept → -0.528724, slope → 0.816882}

{{0.903815}}
```

Figure 9. Effect of eliminating redundancy. A second plot, from the newer Mathematica program, in which the NDH values used to produce the 61 codon values are now averaged into 20 amino acid values as redundancy is eliminated. Those value are in turn compared with the original K&D values. In other studies, not shown, I found that using partially randomized codes from a pool of 21! possible codes (including no change in the redundancy pattern, but all amino acid assignments randomized) showed that this level of correlation placed our wild-type genetic code in the 99.975 percentile for correlation with hydropathy.

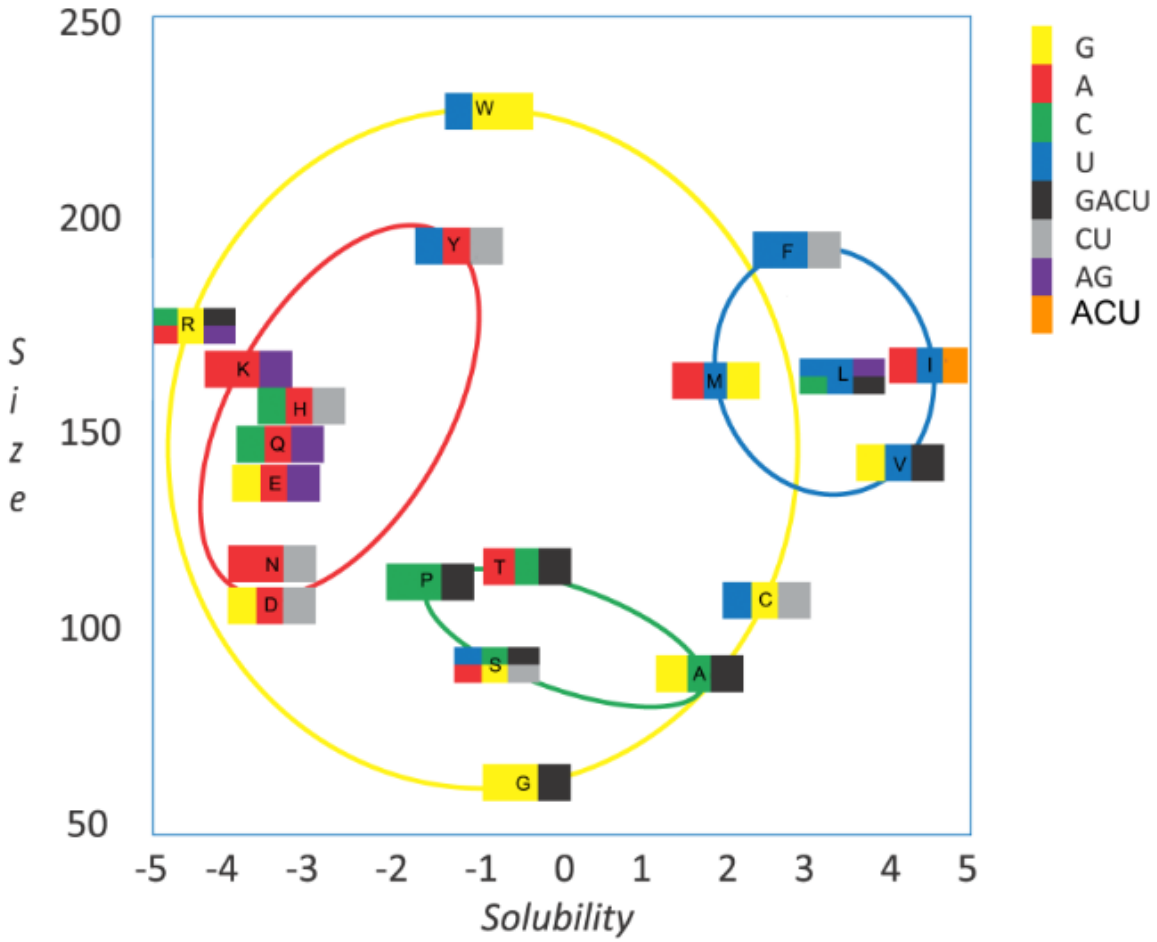


Figure 10. Amino acid properties mapped onto various codons. From: http://upload.wikimedia.org/wikipedia/commons/e/e6/Elliptical_Genetic_Code.JPG . Redrawn in May, 2008 from Figure 4 in http://www.complexity.org.au/ci/vol01/fullen01/html/ which attributes the graph to Yang *et al*., 1990. The condensed version of this figure is now the fourth figure in Wikipedia's Genetic Code article which I inserted a few years ago. It gets about 600,000 hits per year. Size is given in cubic angstroms, and solubility is in the K&D scale.

SVD and a Short-Cut Solution, YI

Something truly remarkable and unexplainable happened during these theoretical analyses. The most likely encoding of the genetic code by a mathematician would almost certainly be a 64 x 12 matrix, wherein each row is marked by three "ones", indicative of the codon's sequence. *However, we found this sparse matrix has an exact solution.* The scrolling of the 1's within this matrix can be formed using Tuples; however, what is essential is that all words (codons) are present using all possible letters (nucleotides). Thus, it appears that a "Tuple Matrix" is special among sparse matrices and matrices in general. They have an exact pseudoinverse. Once the pseudoinverse is obtained, then the properties of a 64 x 1 matrix (*e.g.*, hydropathy) can be mapped to the 12 nucleotide determined variables.

There are two ways to create the 64 x 12 matrix, given my Mathematica code shown below. This is followed by the definition of Tuples which gives the formula for the canonical output that causes the 1's in this sparse matrix to 'scroll' in an orderly manner. The variables "alpha" and "word" are used in the text as "*a"* and "*w"*.

1) alpha=4;word=3;
dict=Partition[Flatten[Tuples[IdentityMatrix[alpha],word]],(alpha*word)];

2) nucgc=Tuples[{a,c,g,t},3];
a={0,0,0,1};c={0,0,1,0};g={0,1,0,0};t={1,0,0,0};
nucgc=Flatten[nucgc]; nucgc=Partition[nucgc,12];

with the function Tuples defined from the Mathematica virtual book as:

> Tuples[list, n] generates a list of all possible n-tuples of elements from list. The order of elements in [*list*, *n*] is based on the order of elements in *list*, so that [{a_1,…,a_k},*n*] gives {{a_1,a_1,…,a_1},{a_1,a_1,…,a_2},…,{a_k,a_k,…,a_k}}.

I have now proceeded through two different levels of mathematics to transfer the values of a properties vector (64 x 1) to 12 values (denoted x1 – x12) that recast a property of the amino acid residues onto nucleotide-determined values. The first values obtained were nucleotide-determined hydropathy (NDH) values which captured the propensity of the genetic code to be consistent with hydropathy. About 50% of the Kyte and Doolittle hydropathy values are captured by only two NDH values: x5 and x8, second position A and T, respectively, that can be used to reconstruct the actual amino acid residue hydropathy parameters with a high correlation coefficient.

The first level of math was Singular Value Decomposition (SVD), which was done in about 1988. (For a later example of how SVD can also be applied to analyzing microarray data, see Alter *et al*., 2000.) This is proper and highly accurate, but computer intensive because of the time complexity (Holmes *et al*., 2007), and too abstract to convey to most audiences. A few years ago, I noticed a pattern in various test SVD matrices and deduced a new pseudoinverse, "YI", which is equal to:

$$D^{\mathrm{T}} a^{-(w-1)} - \frac{\frac{w-1}{a^w}}{w}$$

First one takes the transpose of the 64 x 21 matrix "*D"* (also designated 'dict' in some of the Mathematica code). This can be done with no scaling by simply redefining rows and columns. The transpose simply rotates the matrix 90 degrees clockwise. Note that there is a multiplicative factor next to the transpose. Then another factor is subtracted. For the 64 x 21 matrix, these numbers are 1/16 and -1/96 respectively. That is true of all scroll matrices with *a*=4 and *w*=3, regardless of the data.

Also known as the "Youvan Inverse", this equation operates on a special class of matrices, defined as Scroll Matrices, D, where *w* is symbolic of word length and *a* is the alphabet size. In a discrete mathematics definition, I compute the scroll matrix "D" using Mathematica syntax:

D = Partition[Flatten[Tuples[Reverse[IdentityMatrix[a]],w]],(a*w)]

I have shown the canonical Singular Value Decomposition (SVD)-derived pseudoinverse of matrix D, "PI", scales very poorly in computation time as *a* and *w* increase in size. Again, using Mathematica syntax, with equivalence denoted "==" , the Mathematica Solve function will return "True" for:

YI == PI, for all values of *a* and *w* > 0

At the largest values of *a* and *w* that I could practically compute, *i.e*., (5, 5), I showed that the PI solution requires 59,000 seconds of CPU time (on an Intel i5 computer), whereas the YI solution is found in less than 62 milliseconds, approximately one million times faster.

There is no rigorous solution as to why the genetic code would just happen to have a binary representation as a sparse matrix with an exact solution. Due to the 2nd position A/T identification with hydropathy, it is of interest to find the best property vector that fits the remaining base pair, G/C. The 2nd position G covers a very wide range of hydropathy/ molar volume values, while C codes for a tightly focused group of residues with moderate hydropathy and small size. I will try to rearrange my exact solution to the pseudoinverse to solve for a representation of what an ideal 2nd position G/C property vector would look like. It will also be of interest to see if we can solve for equations related to the sharing of coding capacity between complementary proteins.

Rules and Alternative Codes

A cursory analysis of the genetic code will already show some simple rules. These rules are taken to be correlations, not causative. Causation can be implied if the correlation has a corresponding molecular mechanism.

To begin with, in no case is a stop codon opposed by a reverse complement in the code that belongs to an amino acid with a redundancy of 1. If this were not the case, such amino acids could not be used in a complementary pair, or else they would generate stop codons in their partner. We have barely begun to look at the constraints placed on the code by CPs.

The entire problem seems to fall too deeply into mathematics and molecular biology to have caught the attention of someone who can work in both fields. So, I am very happy to do this work and produce a book by the same title as this proposal. I need to be able to explain CPs to a mathematician and to explain sparse matrices and special solutions to a molecular biologist by adding background material in both fields to the book. I feel that is a very good use of DOE's funding for this project as genomics grows and well trained researchers are needed.

Please note how redundancy plays a second, important role between an amino acid and the reverse complements. In general, a residue with a degeneracy of *d* can have *d* different residues as reverse complements. In switching the 1st and 3rd positions, we can bring a 3rd position N (all coding for the same amino acid) into four different first positions (A, C, G, T), all of which code for different amino acids. Thus the redundancy aspect of the wobble in the 3rd position becomes a primary identifier of amino acids when transferred to the 1st position of the codon. This provides for a more relaxed rule on splitting the coding capacity between the two strands.

At this time, it is very important to annotate phylogenetic protein alignments to distinguish proteins that come from an organism with some verification of the code versus those that are deduced from organisms in which the code has never been verified by protein sequencing. The environmental metagenomic data falls in the latter category by default. Once incorrectly translated proteins enter these phylogenies, it makes it easier for other incorrect sequences to align (http://www.ebi.ac.uk/Tools/emboss/) or appear to be normal variation. This compromises the proteome and a very large body of work. We can't easily get direct protein sequence data from an unculturable organism, so our only alternative at this time is to put some good algorithms in place that will detect translation by an alternative code.

A major rearrangement of the code that does not affect methionine or the stop codons will be the hardest to detect. Every lab generating data from environmental metagenomics should run more sophisticated algorithms to detect for such changes in the code. The most likely algorithm to detect such changes will probably use hydropathy and detect that a complementary protein does not have reversed and inverted hydropathy. Next most likely would be an algorithm that sees that some amino acids are aligning with the phylogeny at a high percentile while another group of amino acids is poor at alignment. With some expert logic, if the alignment for all amino acids is improved, one has an argument that the code is likely changed. We should at least flag such a protein sequence as having been deduced with no direct protein sequencing in the organism and that it has tripped an AGC algorithm. For now, that protein should be provisionally excluded from phylogenetic trees and listings for reasons stated above. While this might look like quality control, in reality the scientists behind these detection algorithms are conducting original research into the possible existence of alternative codes. Once the procedure is established, the job should pass to bioinformatics experts. These analysts will take precautions to ensure that deduced polypeptide sequences with or without any supportive protein sequence data are identified as such. Confirmatory protein data can come from as little as one example protein per organism.

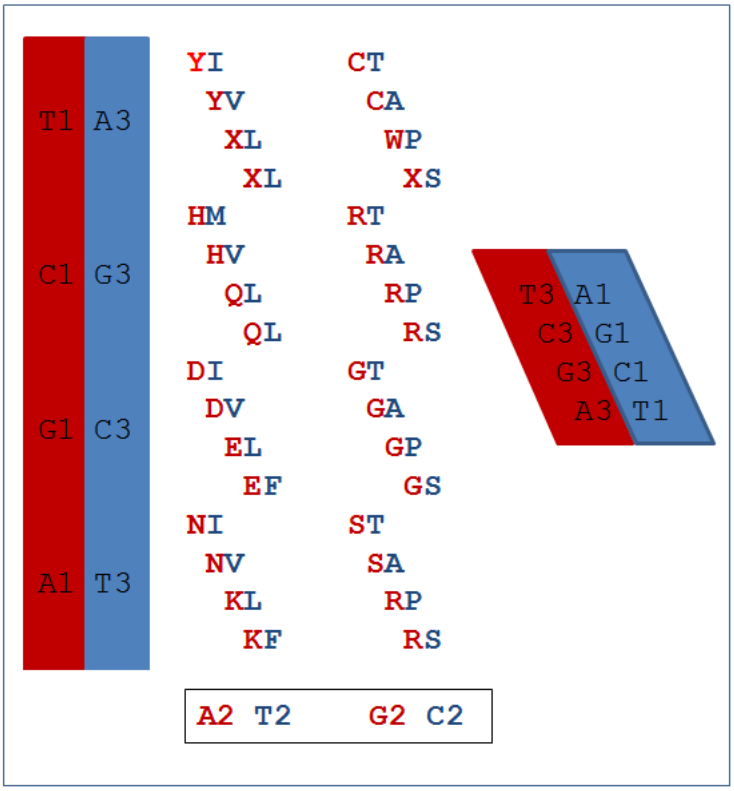


Figure 11. Relationship between 'sense' codons and 'antisense' codons in a hypothetical complementary protein pair. It is important to understand that the sequence of the antisense protein in a CP pair is entirely determined by how codons and their anticodons are grouped. One could see this as a second translation of the code, first by the sense codons and then by the 'reverse complement' codons. The key for this code might seem scrambled, but it is not: Positions are given in pairs consisting of reverse complements. For example, the complementary pair of amino acids Y and I (upper left) are read as T1A2T3 and A3T2A1 or TAT and ATA, which are reverse complements. If Y is in the sense strand (using this codon) we can immediately identify that I is in the antisense strand. Physicochemical correlations can also be observed. If one follows the two vertical and staggered columns above A2 and T2, it is obvious that these code for hydrophilic and hydrophobic residues, respectively.

I have extensive original research on the genetic code that has not been published. Properly, it is time to write a book by the same title as this grant proposal. This comes at a time when the genetic code, except for minor changes, is beginning to look universal. Sequencing of environment metagenomic samples that show no signs of alternative codes will help to solidify the idea that the code is indeed universal. This is also the correct time to be very careful about deducing protein sequence data from primary DNA structure without any confirmation of actual protein sequencing in unculturable organisms. One aspect of this proposal is to design algorithms that would check these deduced sequences for signs of an alternative code.

Another aspect of this proposal presents evidence that our genetic code is truly remarkable and has received far too little attention. No matter how the code arose, it carries features that yield high correlation coefficients with the two most important factors in protein structure and function, i.e., the hydropathy and molar volume of the amino acid R-groups. Some of these features have been recently revealed in the physical chemistry of Complementary Proteins – a situation that arises when proteins are coded in unison from the same sequence of DNA. This protein pair uses a full length antisense ORF

alongside with the sense strand encoded protein. Here we see part of the structure of the code: The antisense protein has a reversed and inverted hydropathy profile. This is due to nucleotide determined hydropathy values at the second position of the codon. "T" is hydrophobic and "A" is hydrophilic. In addition, a second position "C" versus "G" encodes residues that have either low or high dispersion in hydropathy-molar volume plots.

Work Plan

Our research is aimed at uncovering organization in the genetic code that is pertinent to protein structure and function and evolution. To the extent that we use some advanced mathematical and computational methods, we are also interested in expanding the use of algorithms to study the organization of the genetic code and to possibly search within metagenomic DNA sequence libraries for major alterations in the seemingly universal genetic code– in particular for the unculturable microbes exhibiting syntrophy. From a mathematical perspective, we also plan to continue to study and exploit a recent advance we made in the mathematics of rectangular matrices filled with Tuples that are normally solved by a Penrose-Moore algorithm better known as Singular Value Decomposition or SVD. We found that Tuple matrices have an exact solution, bypassing the need for computationally intensive SVD. So far, we have shown a quantitative relationship with amino acid R-group hydropathy and the second position of the codon. Work in progress should also show a second position correlation with the dispersion of R-group molar volume. Both of these features use rules that cause reversal and inversion of the physiciochemical properties of Complementary Proteins, i.e., proteins translated using the antisense strand. We also plan to develop an optimization program that will use features of Complementary Proteins as an example. Simultaneously, we will attempt to find global maxima and global minima in complex problems using the paradigm of a structured look up table (LUT) patterned after the genetic code. This is especially advantageous given that the underlying over-determined matrix is the only one we know that can be solved rapidly and exactly for a pseudoinverse. This inverse can then be taken through a dot product with a properties column vector to recast phenotypic vectors on a genotypic string analogous to DNA. Our company's general theme is to combine advanced biological phenomena with advanced mathematics and computation using the biological system as a model for algorithm development. Max-min problems modeled after the expression of sense-antisense proteins using a structured LUT is especially interesting and novel.

'Subset homology' can be defined as an effect that will potentially be observed in phylogenetic alignments. What we look for is a difference in the percent identity between the deduced amino acid sequence of a newly sequenced gene and the consensus sequence. If we see that one subset of amino acids has a much higher percent identity than another set of amino acids, then we are potentially seeing a protein gene that is encoded by an alternative genetic code. By inspecting the subset that has poor identity, we deduce a putative, new code. If we are correct, then the percentage identity between this new sequence and the consensus sequence should increase substantially.

For example, in a hypothetical altered genetic code, if we switch the encoding of the hydrophilic residues (NAN) with the hydrophobic residues (NTN), the homology between this improperly translated deduced polypeptide and the consensus sequence will be lost. By inspection of this obvious switch, one guesses the new code and sees that the NTN and NAN groups of amino acids increase to a percent identity originally observed in the unaltered NGN and NCN subsets of amino acids. Of course, the switching of codons within the altered genetic code does not have to be this systematic, and a computer program with a fitting algorithm is then appropriate and will be written. The number of altered codes observed in a diverse metagenomic library could range from having only one altered code to seeing a different code for

each new sequence (other than the wild-type code). The wild-type code is expected to have a very high frequency, possibly greater than 99%.

The bacteriorhodopsin data of Venter from the Sargasso Sea has 1800 new bacteriorhodopsin sequences. One wonders how the percent identity cut-off threshold was established to qualify a protein for this list. By examining sequences in the database that appear to have a very low percent identity (below his cut-off), we have the possibility of using the subset homology algorithm to increase the percent identity of one or more of these sequences by an amount that would cause it to move above the cut-off. This dataset has very rich mining opportunities, and as more libraries such as this one are sequenced, we need to run the subset homology algorithm with perseverance. A single positive sequence that passes statistical criteria is of great significance. It will also yield a partial deduction of a new altered code that can be run on all libraries with the idea that it is the most frequent altered code. If true, we would expect unclassified sequences to suddenly become part of known families of proteins with a significant degree of identity.

Venter's entire dataset of the Sargasso Sea study (Venter *et al*., 2004) has been deposited: http://www.ncbi.nlm.nih.gov/books/NBK6855/ as WGS study AACY000000000A described as a "total of 41 different samples were taken from a variety of aquatic habitats collected over 8,000 km. 7.7 million sequencing reads were obtained from size-fractionated samples, yielding 6.4 million contiguous sequences, totaling 5.9 Gbp of nonredundant sequence. These were further processed into about 3 million assemblies (scaffolds)."

Rich Roberts came to an identical and simultaneous solution to this problem. Our emails actually crossed:

"The approach you suggest of looking for aligned protein sequences where there are recurrent patterns of unexpected amino acids interspersed in obvious homologs is the way I had been thinking about after your first email. You wouldn't need to do this on everything, but rather check just a few key proteins that are generally well conserved, such as those involved in intermediary metabolism, energy utilization, ribosomal function, transcription, replication etc. You could easily do a mockup of this using existing entries in GenBank where we know (or at least think we know) what genetic code is being used and then comparing examples of things like Mycoplasmas where we know the code is unusual. It would be easy to write an algorithm to then go looking at say Craig Venter's data set or some of the more recent metagenomic sets. However, the latter may be misleading since I suspect that most metagenomes are composed primarily of strains that exchange DNA, which would imply a common code. The most likely set of organism with unusual codes would be some of the extremophiles or groups of isolated organisms such as coral symbionts or other specialized marine environments."

Another prediction based on these observations suggests that an alternative code, necessarily an odd number of positions) could increase to five positions and drop alphabet to 2 nucleotides. Thus, 2^5 =32; 32>20. Only one major rule could be supported by the center (third) position, and that would be a division of hydropathy. However, the entire set of amino acid residues would go into only two groups – blurring the distinction between extreme and average hydropathy values. This analysis is entirely numerical, and no consideration is made for underlying biochemistry for synthesis pathways

Algorithms can process strings and add information to them. For example, the canonical ordering of Tuples generates a progressive pattern. Now the question is whether the code can add order to strings while they are undergoing mutation. An equilibrium could be produced or the code could continue to add information. This mechanism utilizes the redundancy of the code – it is not a straightforward LUT in the conventional sense. 64 different inputs cause 21 outputs. This is unusual.

It would be interesting to look at the hypothetical complement of many known proteins using protein structural prediction programs. This would be followed by the use of our expert logic to see whether there is any indication of what might be a functional protein. Such a protein could be a vestige of a time when the protein under study actually had a CP, but has now diverged, and the genes are no longer arranged in this manner. One clue in this process might be to consider whether the activities of the CPs could be related in metabolism. The method will be no better than the investigator's knowledge of protein structure and function, therefore someone is needed who is very good at this subject and able to visualize many different protein structures as possible motifs for the CP.

The genetic code is identified as an 'intelligent look up table '(iLUT) organized to carry rules related to amino acid residue hydropathy and molar volume. The code is structured to resolve random mutational events (point mutants and crossovers) in favor of enhanced protein function. The code's basic structure allows it to encode half of the amino acids by hydropathy and the other half by the area of dispersion in a metric of hydropathy multiplied by molar volume. These rules are setup to support the evolution of two proteins at once, one in the sense strand and the other antisense. The discovery that the code's redundancy provides coding capacity for rules should generalize to other optimization problems.

The following example reasoning demonstrates how cryptography can inform molecular biology and vice versa: (1) Assume that the only thing known about the genetic code is that it is triplet. The exact coding is unknown. (2) Also assume you can sequence protein but not DNA. (3) Further assume that you can identify a CP pair as using the same DNA but being transcribed in an anti-parallel manner. After sequencing this protein pair, you notice something: hydrophobic amino acids in one protein align with hydrophilic residues in the other strand. Whereas reading the sense protein tells us nothing about the code, comparing these two sequences does. A good cryptographer should be able to deduce that a rule exists at the second position for hydropathy using complementary base pairs. In other words, nothing can be deduced about the "first encoding" of the sense protein, but as soon as we see its CP, we also see part of the code. That is one reason why studying the structure of the genetic code is important: We learn the restrictions on CP pairs. CP's are not uncommon. One might then ask, "Which came first? The structural features of the code, or the need for certain physicochemical relationships between the CP pair (e.g., hydropathy)?" Furthermore, this cryptographer might also see that when certain amino acids within a small area of hydropathy x molar volume space are used, then the CP has a certain group of residues that sweep out a larger area of HMV space. The same questions arises: Which came first? Did this effect in hydropathy x molar volume space exist before the code froze, or is this just a consequence of an inadvertent structure to the code? With further study of the code and with a second example from an alternative code, perhaps we can answer this question.

One should look at the genetic code in an analogous manner to how we look at protein structure and function. We look for relationships and motifs in structure rather than discuss the statistics of the sequence. We look at the genetic code as a Single Event (so far) where statistics do not apply. If we find a relationship that involves only two codons, we take it "as is". We weigh evidence for a particular hypothesis rather than discard such a relationship because it so improbable. No one would do such a thing with a protein structure. There we look at minute details and incorporate them into a model. That is also the proper way to evaluate the genetic code.

We have some bioinformatics work to do. All sequence data that we will analyze will be collected and stored locally. Solid state drives are used in our new computers. These drives provide virtual memory for computation (see, *e.g.*, http://www.sandisk.com/enterprise/ulltradimm-ssd/). One of the more difficult tasks is categorizing entries into the Proteome as to whether there has been any confirmation of the wild-type genetic code in those organisms. Another difficult task is to ascertain whether any of the CPs are

known to be expressed through experimental studies. For unculturable microbes, we will initially assume there is neither conformation of a wild-type code nor observation of any CP pairs being expressed.

The genetic code is treated as a Look Up Table (LUT). We don't consider the origin of any apparent structure in the code. In fact, research on the code may have been impeded by the idea that anyone looking at the structure of the code is somehow an advocate of Intelligent Design (ID). Structural aspects of the code have now been observed in Complementary Proteins, where both sense and antisense mRNA are translated in to protein. By asking how one DNA sequence can code for two proteins, we see that the genetic code is entirely responsible for structural aspects (e.g., inverted hydropathy) of both proteins. In terms of alternative genetic codes, these relationships are sufficiently well known that we are able to predict other codes that would conserve the ability to encode two proteins at once.

Researchers apparently did not realize it is possible to develop a hypothesis regarding the code and then test it. In this particular case (future prediction), statistics is relevant. I am testing my model with statistics, not the origin of the code. It is not until we find a molecular biology mechanism that fits this hypothesis and numerical experiment that we would start thinking about correlation and causality. Statistically infrequent correlations might be convincing, but there is nothing about the scientific method that says an outcome must be improbable.

Consider a hypothetical situation wherein we could determine that all the DNA's coding capacity went into the sense protein. The only information relayed to the anti-sense protein is through the reverse complementation setup by the genetic code. If the antisense protein were functional, it means the information to make it so came from the genetic code. It seems unusual that reversing a protein's sequence, inverting its hydropathic residues (2nd position A/T), and changing the dispersion of the hydropathy x molar volume profile (2^{nd} position C/G) could lead to functionality. This needs to be investigated. As a theorist, I feel that it is important for me to find indicators of the percent coding going into each CP pair. Perhaps codon usage could be used. If frequent codons are used in sense and infrequent codons are used in antisense, then one might conclude the coding capacity is entirely in sense. The control would be from proteins that don't have a CP. This obviously needs more evaluation.

In order to make the process more productive, one might develop a list of niches and a strategy for dividing a given amount of sequencing among niches to maximize the probability of finding an alternative code. One of the first questions is whether to run pilot rDNA assessment for molecular cladistics to measure diversity and find organisms close to the base of the phylogenetic tree. One of the most important resources in this project is the viewpoint of many different scientists, some of whom have experience in exploration for new genes for industrial processes (enzymes). Others will have important information on how these communities form and interact with an aim of finding the most primitive of the primordial bacteria that survive to this time. This is a thought experiment that could become a reality. The price tag is high, so we want to maximize the chances of finding alternative codes by bringing together our best thinkers from diverse backgrounds. Ironically, failure to find an alternative genetic code after such a search might be preferable to some. It simplifies molecular biology down to a universal genetic code and makes it more likely what we deduce from DNA properly enters the proteome without much probability of error. However, from our standpoint, the algorithms to detect alternative code will continue to run as long as we are performing metagenomic sequencing.

DOE Mission

The DOE is heavily invested in gigabase large-scale DNA sequencing projects, dating back to the early beginnings of genome projects in 1987 (http://www.genome.gov/12011239). As sequencing moves towards unusual environmental niches, the probability of finding alternative genetic codes increases. This proposal is proactive in detecting alternative genetic codes before such data corrupts the Proteome. We are also proactive in methods to potentially decipher a new code based solely on DNA sequence, as it is highly improbable that biochemical methodology can be applied to these unculturable bacteria. If the genetic code is universal, our methods will be important for empirically proving it. A second aspect of our proposal that is key to DOE missions is to further our understanding of Complementary Proteins (CPs), where one DNA sequence codes for two proteins. From *E. coli* to man, a significant percentage of antisense mRNA is transcribed into protein. Roles for CPs range from those just detected for the photosynthetic reaction center as well as a role in autoimmune diseases in humans. The structural relationship between the CP protein pair is determined entirely by the genetic code. By treating the genetic code as a mathematical identity, we continue to discover structural relationships involving hydropathy and molar volume, physical chemical factors directly imbedded in the genetic code. These studies are facilitated by the use and extension of higher mathematics and computation – areas in which the DOE has an interest in seeing American scientists being trained and excelling in their research. Beyond bioinformatics, we work with both advanced molecular biology concepts and mathematical algorithm development with application to genomics in a manner requiring brilliant theorists in a think-tank environment. To stay at a technological advantage, the US must take a leading role in bringing mathematicians and molecular biologists together. In still another area appropriate for the DOE, we must understand the metagenome of various environmental niches. Studying these metagenomes provides the ultimate technique for reporting on the health of the environment. Moreover, at the same time, actual discovery of new genes could lead to cleaner industrial processes by providing novel enzymes. However, as this proposal suggests, the first thing we must do is determine whether the genetic code has alternatives, or whether it is universal, in order to maintain the accuracy of the rapidly growing proteome database Finally, for purposes of teaching, developing our next generation of scientists, and facilitating research in genomics, a book by the same title as this proposal will be written to demonstrate the utility of interactions between mathematicians and molecular biologists. The Sargasso Sea data that we will be using, and which came from Venter's lab, was made possible in part by support from the Department of Energy Genomes to Life program and the Office of Science U.S. Department of Energy grant no. DE-FG02-02ER63453.

Computation

To approach this highly complex problem, one must consider whether to implement parallel versus distributed processing. I actually prefer the latter because it is more similar to how I develop. The only additional capability needed is communication between an *ad hoc* group of workstations with about equal capabilities among users with a *Mathematica* licenses. A master program would simply divide the work load among input files to these piered workstations and then collect and combine the output files. The programs running on the workstations would be identical, so I am familiar with how they behave. With a recursive process running millions of times, it is best for me to have at least 16 GB of RAM per CPU.

With parallel processing I see configurations with far less RAM per CPU. Just as important as hardware configuration is how a very high language such as Mathematica can properly interface with new parallel boards and use these boards efficiently. I really don't see a general solution for that problem, and I would prefer to decide for myself how the task is run on the parallel boards. To do this, I would have to learn how to run a Xeon-based workstation and learn the appropriate low level language for parallel processing. This is the type of job that requires a division of a company, not a sole investigator. So, for now, the plan is to do distributed processing on piered workstations (across the internet) all running Mathematica. I can do that with the Wolfram Research developers if I contribute a powerful workstation to the pier.

Something that is not well known is that Mathematica programs (or computationally intensive parts of programs) can be compiled in Mathematica. Wolfram suggests that the function Compile is best used on expressions that involve considerable arithmetic calculation. Their functional calls (e.g., Sin) have already been optimized. They state a factor of 20 can be achieved in compilation of complicated arithmetic statements – such as I use. While I have discussed parallel and distributed computing, it will be my goal to write optimized code that will run on a single, high-end workstation. I find that researchers involved in large scale computation always emphasize: "write better code".

I don't want this to sound like boasting, but I am certain my YI shortcut of SVD is correct. Given a large enough matrix, YI could reduce billions of years of computation on DOE's Titan to only a few seconds on a Raspberry Pi. I think that should be known by others within the DOE community that might be working on computationally-intensive problems.

Perhaps the strongest constraints on the genetic code occur in the case of CP's for which both proteins are functional and every codon for one protein is the reverse complement of the other protein – with an additional reversal of the carboxy – amino termini of the proteins. The genetic code could facilitate this double encoding if reverse complement codons encoded amino acids in a manner 'logical' for protein structure and function. For some unknown reason one constraint is clear: The second position complement uses an A/T rule to invert the hydropathy of one protein relative to the other. We already see a few more 'rules', whether coincidental or not:

1. No amino acids are lost from the sequence of either protein, because neither of the single degeneracy M or W residues have a reverse complement of a stop codon.
2. This is also true of the nine cases for double-degeneracy amino acids. None are opposed by two stop codons.
3. The wobble positions of one protein becomes the first position of the other. The former differentiates amino acids based on a GA versus CT division, whereas all four bases determine the amino acid in the first position of the codon.
4. Hairpins in mRNA generate a reverse complement of codons within the same strand, symmetrically arranged at growing distances from the hairpin's loop.
5. Codons using G/C at the 2nd position have a 2-fold increased degeneracy over codons using A/T.
6. Second position G versus C shows an extreme difference in the encoded residue hydropathy x molar volume area. G sweeps out extremes and a central serine residue whereas C is tightly clustered (see Figs. 1 and 10).
7. Overall, redundancy is essential for patterns to develop in the code. Because of the need for one position to remain unchanged (#2), only an odd number of positions (*e.g.*, 3) will support a concise rule for a property to transfer from the sense to antisense strand.

These differences taken together and combined with protein structure / function parameters yield an extremely complex and sophisticated system to study. These are existing and factual structural parameters of the code which do not depend on origin, statistics or the degree of optimization.

We should consider any and all molecular biology paradigms that might place constraints on the code. This includes RNA splicing and combinatorial diversity of the immune system. As a general paradigm, combining advanced concepts from molecular biology with mathematics is a route to both understanding the structure of the code and developing new algorithms. For example, beginning with our early work on recursive ensemble mutagenesis, it is clear that an evolution program should be written that makes full use of findings from molecular biology. For example, what we see in CPs in the cell can be transferred to an evolution program that simultaneously minimizes and maximizes the solution to a benchmark problems such as the Traveling Salesman Problem (TSP).

Methods for analyzing a database of homologous proteins can roughly be divided into two approaches: 1) heuristic, and 2) systematic. A heuristic approach can make major rearrangements of the code based on what we already know. A systematic approach makes incremental changes to the code that progress to a recognizable set of changes that can be grouped. Heuristic switches in the code could include 1) switching NAN with NTN codons, or 2) switching NCN with NGN codons, or 3) switching NNA/G with NNC/T codons. These indicate rearrangements based on hydropathy, HMV dispersion, or wobble, respectively. In one version of the systematic approach, we subject the entire phylogeny alignment of a particular protein to N^2 – N = 64^2 – 64 possible 'mutations' of the genetic code where every codon is individually changed to every other codon. For both the heuristic and systematic approach, we are looking for an increase in the percent identity between a phylogenetic consensus sequence of a particular protein and a test of every occurrence of the protein for increased identity. The heuristic approach can make major gains in one step whereas the systematic approach requires 64 recursive steps.

As an example of the systematic approach, consider Venter's Sargasso Sea data of 1,800 bacteriorhodopsin proteins. A complete analysis requires a 3-D matrix 64 x 64 x 1800 (~ 7.4 million numbers) wherein each element indicates the gain or loss in identity between that particular sequence and the consensus sequence when the corresponding codons are changed from one to the other (on the x and y axis). A positive result in this matrix is indicated by a plane that would show perhaps a 50% increase in sequence identity by changing about half of the codons. In fact, the changes might be so numerous that realignment of the sequences is necessary as the codons are changed. At the end of this process, we will want to see a ranking of the proteins showing the highest percentage gain. In addition, some automation of the types of changes causing increase in identity might be attempted.

We assume these sequences have already been checked for stop codons and for key amino acid residues in a particular hydropathy cycle. However, if we move into the bulk of uncategorized sequences for proteins below Venter's identity cutoff, hydropathy and stop codons will be checked. Our best candidate might be in this uncategorized data if the change in the code is extreme. If possible, we will want to capture proteins as low as 30% identity wherein a major improvement in the code could add 40% to the score, with 30% divergence. If these proteins are indeed primitive, then the final percentage divergence with the consensus sequence could be high. Initially, a very good protein homology detection program will be required, and I plan to write that from scratch in Mathematica. I've previously worked with the Wolfram staff to minimize seek time (see: http://en.wikipedia.org/wiki/Hard_disk_drive_performance_characteristics) into very large matrices using techniques that are not described in the support literature. I've also worked with very large 3D parametric plots.

Some of the most interesting proteins to study with respect to the structure of the genetic code happen to be membrane proteins involved in energy transduction (photosynthetic reaction center and bacteriorhodopsin). The RC has CPs, but bR does not. The large bR database from Venter's metagenomic sequencing of the Sargasso Sea DNA would seem to be our best chance (at present) to find alternative codes among sequences they have left uncategorized.

For our Evolution Program, we will start with a single DNA sequence and the protein for which it codes. Beginning with random genetic codes, we will score successive iterations of program for improvement of the code for correct translation of the DNA into protein. We will begin with an ensemble of random genetic codes. Much like GA's, we anticipate improved motifs in the code will propagate and recombine with each other. The chance of finding the actual genetic code would seem to be infinitesimal (one in 1.5 x 10^84) and there is nothing to say that there is anything unique about the wild type code. However, if we use the same set of evolving codes to translate a DNA sequence and its reverse complement (separately), we might expect acceleration in the rate at which these codes progress to ones with favorable rules. Those rules would be analogous to the 2nd position A/T rule for hydropathy and the G/C rule for dispersion in H-MV space.

To preserve motifs in these developing codes, we will attempt to recombine them as 3-D structures rather than simple strings. Out of the tic-tac-toe box, volumes ranging from 1/16th to half the cube can be traded between two different cubes. What usually survives in a genetic algorithm is fit strings – analogous to proteins. In this case, it is fit genetic codes that survive. The GA-based protein engineering method known as Recursive Ensemble Mutagenesis (REM) which employed carefully randomized combinatorial DNA cassettes (Arkin & Youvan, 1992; Delagrave *et al*., 1993), also had an unusual mechanism of survival involving genetic codes. They were randomized, but we did look at the canonical rate of (protein) string survival. The strings survived by being reverse translated into DNA by a program analogous to CyberDope. CyberDope collected mutagenesis data for each site in the protein ensemble, and then it calculated new 'doped' sequences (compatible with a DNA synthesizer) to start the next cycle. In the case of REM, we actually reduced this to experimental practice. That is why there was a requirement for the DNA sequence that was being "fed back" to be compatible with a DNA synthesizer. I already have most of CyberDope rewritten, hereafter under the name of program's new name, CyberDopant.

I have begun to research general problems in science and engineering where the technique of Singular Value Decomposition (SVD) is applicable and that are impossible to compute using today's fastest computers. This is due to the extremely poor scaling of SVD with matrix size. If one were to increase the size of a matrix by 10^3, SVD computation time will increase by a factor of 10^9 and memory by a factor of 10^6, whereas the YI method would only increase CPU and memory by 10^3. The literature shows attempts to use SVD in image processing and compression. One rapidly gets into problems involving factorization, Hilbert space, and Eigenvalues. In one of its most basic uses, linear combination of atomic orbits results in molecular orbitals. I dare to say that the literature is deficient on such topics because researchers abort plans to use SVD as soon as they see the highly punitive scaling. I will also search Principal Components Analysis (PCA) because this is often the forerunner to SVD analysis. The only additional requirement for YI over SVD is that the dataset have a complete listing of 'words' for a given 'alphabet'. Representation of data as Tuples is convenient but not necessary. Each tuple will become a row in the matrix that is a^w by a*w in size. Given my background, my most relevant interest is hyperspectral imaging, where PCA has already been used.

Concepts for Further Development

I have a list of mostly undeveloped concepts that range from algorithm development to commercialization, all pertinent to work on the structure of the genetic code. Many of these ideas will require additional staff, and their prioritization is extremely important to the success of future research and the company as well. Full expansion of some of these concepts will be useful for a book. However, as a theorist it is difficult to draw the line between completing the next concept (possibly a computer program) and taking the time to describe something that might be easily finished. A rapidly advancing field makes this decision even more difficult. The situation is very analogous to trying to write molecular biology textbooks in the 1970 and 1980's as new techniques and findings were being published weekly. Now, 15 years after the sequencing of the genome, it seems like we still do not have time to pause and reflect. Here is where I stand:

1. Instead of working with high complexity, we need to find new algorithms to bypass these problems entirely. Such major steps cannot be expected on any kind of timetable. I could view my YI bypass of SVD as a once-in-a-lifetime finding. The conclusion is that more people thinking along these lines are essential.
2. While parallel computers are an intriguing topic, at this time I believe it is better to work with bundled workstations. This could change in just a few years. I will wait until I see the equivalent of Artificial Intelligence (AI) assigning memory. This is CS problem that is too crucial to the entire field to be addressed by a single theorist or small company. Hardware and intermediate programming languages are required – neither of which we design or write.
3. Much information might be contained in the 'discard pile' of large metagenomics projects. That could happen simply by analyzing stop codons before protein homology is studied. With the expense of sequencing, it is important to go through and take a second crop of these data. In the search for an alternative genetic code, it is almost likely the data (if any) will have been left unprocessed. The amount of data to be processed is immense and requires new algorithms.
4. Bioinformatics is a supporting field; however, the first iteration of a new algorithmic process, one is working at a research level in both mathematics and molecular biology. More people need to be trained in this approach before it is reduced to systematized bioinformatics.
5. In terms of a commercial effort that will facilitate work in this field, I can envisage a "Metagenomic Computer" that facilitates the development of algorithms by graduate students and postdocs. Using genomics programs written in Mathematica, they could be only weeks away from writing a break-through algorithm rather than spending their time on computer architecture, tracking down an assortment of software, and finding all the databases. For metagenomics at least, I am close to being able to setup such a system for a minimal amount of money.
6. If we formalize the *Search for an Alternative Genetic Code* (SAGC), input is needed from an expert group of researchers to determine how to distribute sequencing efforts between a number of sites and how to apportion the number of base pairs per site. The success of the project has much to do with experience in gene mining and the dynamics of particular niche communities. Even if we don't find an alternative code, this project begins to build a bio-geo picture of the Earth's metagenome. This map could be as rich as one for minerals and fuels: It shows us where to look for particular types of biodiversity and it begins the process of mapping our microbial ecosystem.

References

Alkatib, S., Scharff, L.B., Rogalski, M., Fleischmann, T.T., Matthes, A., Seeger, S., Schöttler,M.A., Ruf, S., & Bock, R. (2012) The contributions of wobbling and superwobbling to the reading of the genetic code. *PLoS Genet*. 8:e1003076.

Alter,O., Brown, P.O., & Botstein, D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci*. 97:10101–10106.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*. 215:403-410. http://blast.ncbi.nlm.nih.gov/Blast.cgi

Arkin, A.P. & Youvan, D.C. (1992) An algorithm for protein engineering: Simulations of recursive ensemble mutagenesis. *Proc. Natl. Acad. Sci. U.S.A*. 89:7811-7815.

Chaffron, S., Rehrauer, H., Pernthaler, J., & von Mering, C. (2010) A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res*. 2010. 20:947-959.

Chou, K.C., Zhang, C.T., Elrod, D.W. (1996) Do antisense proteins exist? *J. Protein Chem.* 15:59-61.

Delagrave, S., Goldman, E.R. and Youvan, D.C. (1993) Recursive ensemble mutagenesis. *Protein Eng*. 6:327-331.

Demeshkina, N., Jenner, L., Westhof, E., Yusupov, M., & Yusupova, G. (2012) A new understanding of the decoding principle on the ribosome. *Nature* 484:256-259.

Dokmanic, I., Kolundzija, M., & Vetterli, M. (2013) Beyond Moore-Penrose: Spare pseudoinverse. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6526-6530. http://infoscience.epfl.ch/record/182698/files/DokmanicKV13.pdf

Holmes, M.P., Gray, A.G., & Isbell, C.L. (2007) Fast SVD for Large-Scale Matrices. In: *NIPS*2007 Workshop on Efficient Machine Learning: Overcoming Computational Bottlenecks in Machine Learning.* http://sysrun.haifa.il.ibm.com/hrl/bigml/files/Holmes.pdf

Jestin, J.L. & Kempf, A. (2009) Optimization Models and the Structure of the Genetic Code. *J. Mol. Evol.* 69:452-457.

Koonin, E.V. & Novozhilov, A.S. (2009) Origin and evolution of the genetic code: the universal enigma. *IUBMB Life* 61:99-11.

Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol*. 157:105-132.

Lombardot, T., Kottman, R., Pfeffer, H., Richter, M., Teeling, H., Quast, C., & Glockner, F.O. (2006) Megx.net—database resources for marine ecological genomics. *Nucleic Acids Res*. 34: D390–D393.

McCutcheon, J.P., McDonald, B.R., & Moran, N.A. (2009) Origin of an Alternative Genetic Code in the Extremely Small and GC–Rich Genome of a Bacterial Symbiont. *PloS. Genet.* 5: e1000565.

McGuire, K.L. & Holmes, D.S. (2005) Role of complementary proteins in autoimmunity: an old idea re-emerges with new twists. *Trends Immunol*. 26:367-72.

Pendergraft III, W.F. *et al*. (2004) Autoimmunity is triggered by cPR-3(105–201), a protein complementary to human autoantigen proteinase-3. *Nat Med*. 10:72-79.

Szostak, J. (2015) The Origin of Life on Earth. II. Protocell Membranes & III. Non-Enzymatic Copying of Nucleic Acid Templates. Online Webinar at: http://www.ibiology.org/ibioseminars/evolution-ecology/jack-szostak-part-2.html .

Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M., Solovyev, V.V., Rubin, E.M., Rokhsar, D.S., & Banfield, J.F. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428:37-43.

Venter, J.C., *et al*. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304:66-74.

Yang, M.M., Coleman, W.J. & Youvan, D.C. (1990) Genetic coding algorithms for engineering membrane proteins. In: *Reaction Centers of Photosynthetic Bacteria* (Michel-Beyerle, M., *ed*.) Springer-Verlag, Berlin, pp. 209-218.

xx - redacted- xx

xx - redacted- xx