## DNA Does Sudoku

# The amazing story of how cells revise and safeguard their own DNA - with a Sudoku-like genetic code matrix 

By Perry Marshall

Imagine that someone gives you a mystery novel with an entire page ripped out.
And let's suppose someone else comes up with a computer program that reconstructs the missing page, by assembling sentences and paragraphs lifted from other places in the book.

Imagine that this computer program does such a masterful job of repair that most people can't tell the page was ever missing.

DNA does that.


In the 1940's, the eminent scientist Barbara McClintock damaged parts of the DNA in corn maize. To her amazement, the plants could sometimes reconstruct the damaged section. They did so by copying other parts of the DNA strand, then pasting them into the damaged area.

This discovery was so radical at the time, most of her colleagues doubted her. (40 years later she won the Nobel Prize for this work.)

And we still wonder: How does a tiny cell possibly know how to do.... that???
A French HIV researcher and computer scientist has now found at least part of the answer. Hint: The instructions in DNA are not only linguistic, they're beautifully mathematical. There is a mathematical matrix that ensures the integrity of the genome. It's remarkably similar to Sudoku.

Computers use something called a "checksum" to detect data errors. It turns out DNA uses checksums too. But DNA's checksum is not only able to detect missing data; sometimes it can even calculate what's missing. Here's how it works.

In English, the letter E appears $12.7 \%$ of the time. The letter $Z$ appears $0.7 \%$ of the time. The other 24 letters fall somewhere in between. So it's possible to detect data errors in English just by counting letters. (You can also use this to tell the difference between English and French, because the letter percentages in French are different.) This is a key concept in communication theory. It's called an ergodic pattern - the statistical signature of a language.

In DNA, some letters also appear a lot more often, like E in English, and some much less often. But... unlike English, how often each letters appears in DNA is controlled by an exact Sudoku formula that is hidden within the genetic code table.

When cells replicate, they count the total number of letters in the DNA strand of the daughter cell. If the letter counts don't match certain exact ratios, the cell knows an error has been made. So it abandons the operation and kills the new cell. Failure of this checksum mechanism causes birth defects and cancer.

Dr. Jean-Claude Perez started counting letters in DNA. He discovered these ratios are highly mathematical and based on "Phi", the Golden Ratio 1.618. This is a special irrational number, sort of like Pi. Mathematicians and philosophers have written about this number for thousands of years.

Leonardo Divinci's famous "Vitruvian man" celebrates the Golden Ratio and the proportions of the human body.

Perez' discovery was published in the scientific journal Interdisciplinary Sciences / Computational Life Sciences in September 2010. His paper is called "Codon populations in single-stranded whole human genome DNA Are fractal and fine-tuned by the Golden Ratio 1.618." Before I explain how it works, allow me to explain just a little bit about the genetic code.


DNA has four symbols, T, C, A and G. These symbols are grouped into letters made from combinations of 3 symbols, called triplets. There are $4 \times 4 \times 4=64$ possible combinations.

So the genetic alphabet has 64 letters. The 64 letters are used to write the instructions that make amino acids and proteins.

I have no idea how he did it, but Perez somehow figured out that if he arranged the letters in DNA according to a 3-dimensional T-C-G-A table, an interesting pattern appeared when he counted the letters.

He divided the table in half as you see below. He took single stranded DNA of the human genome, which has 1 billion triplets. Single stranded means that even though DNA has two complimentary bases, one on each side of the helix i.e.

TCTTAAG
AGAATTC
He is only considering the half of the strand that we label
TCTTAAG

He counted the population of each triplet in the DNA and put the total in each slot. I'll show you the results in a minute, but let me pause and explain Sudoku first.

## Sudoku and DNA

Have you ever played Sudoku? It's a crossword puzzle with numbers instead of letters. Many newspapers publish a Sudoku puzzle each day, right next to the crossword puzzle.

In Sudoku, you start with a partially filled out grid like this:

| 5 | 3 |  |  | 7 |  |  |  |  |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 6 |  |  | 1 | 9 | 5 |  |  |  |
|  | 9 | 8 |  |  |  |  | 6 |  |
| 8 |  |  |  | 6 |  |  |  | 3 |
| 4 |  |  | 8 |  | 3 |  |  | 1 |
| 7 |  |  |  | 2 |  |  |  | 6 |
|  | 6 |  |  |  |  | 2 | 8 |  |
|  |  |  | 4 | 1 | 9 |  |  | 5 |
|  |  |  |  | 8 |  |  | 7 | 9 |

The objective is to fill a $9 \times 9$ grid with digits so that each column, each row, and each of the nine $3 \times 3$ boxes that compose the grid contain all of the digits from 1 to 9 . When you're done, the completed puzzle looks like this:

| 5 | 3 | 4 | 6 | 7 | 8 | 9 | 1 | 2 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 6 | 7 | 2 | 1 | 9 | 5 | 3 | 4 | 8 |
| 1 | 9 | 8 | 3 | 4 | 2 | 5 | 6 | 7 |
| 8 | 5 | 9 | 7 | 6 | 1 | 4 | 2 | 3 |
| 4 | 2 | 6 | 8 | 5 | 3 | 7 | 9 | 1 |
| 7 | 1 | 3 | 9 | 2 | 4 | 8 | 5 | 6 |
| 9 | 6 | 1 | 5 | 3 | 7 | 2 | 8 | 4 |
| 2 | 8 | 7 | 4 | 1 | 9 | 6 | 3 | 5 |
| 3 | 4 | 5 | 2 | 8 | 6 | 1 | 7 | 9 |

Some people play 3D Sudoku puzzles. A Sudoku version of the Rubik's Cube is called the Sudoku Cube.


What does this have to do with DNA? Cells use a 3-D Sudoku-like matrix to check for errors in DNA replication.

In Sudoku, each number has to appear once in each row, column and $3 x 3$ box. DNA does a different kind of Sudoku. If we make a Sudoku matrix where each box contains the total count a specific letter in single-stranded DNA, this matrix has very special properties. In DNA's matrix, each of its 64 letters (triplets, or codons) has to appear a precise percentage of the time.

This makes for a fascinating accounting problem if you arrange the 64 letters into a 3D, $4 \times 4 \times 4$ table:


Each dimension of this cube represents one of the three symbols in a triplet. The vertical axis is the $1^{\text {st }}$ symbol, the left-to-right axis is the $2^{\text {nd }}$ symbol, and the front-to-back axis is the $3^{\text {rd }}$ symbol. This 3D table becomes a very convenient way to arrange the genetic code.

Dr. Perez counted the total number of TTT's in single-stranded DNA. You can put the total in the "front and center" bucket above, where each axis is "T." You can count the number of ATT's and put the total in the bucket just below that one, and so on, until each bucket has a total count in it.

You then divide the table in various ways and look for symmetries, according to the scheme below.

TC G A


Perez split the table in half like this:


When he added up the letters, the ratio of total white letters to gray letters was $1: 1$ - if he divided the table down the middle. And.... it didn't matter whether he divided the table in half across the X , or Y , or Z-axis, the exact ratio of $1: 1$ still held.

This turned out to not just be roughly true. It was exactly true, to better than one part in one thousand, 1.000:1.000.

However, if he colored the middle 2 sections white and the edges grey, the ratio of white letters to gray letters is 0.690983 . ('lll explain the significance of this number in a minute.) And again, it didn't matter whether he divided along the X axis, or Y axis, or Z axis. The ratio was the same.

So for three ways of dividing the table (half vs. half), the ratio of white letters to gray letters is 1.000:1.000. And for the other three ways of dividing it (middle section vs. edges), the ratio is 0.690983 .

## What's so important about $\mathbf{0 . 6 9 0 9 8 3}$ ?

0.690983 is (3-Phi)/2 where Phi is the Golden Ratio. The Golden Ratio is the famous number 1.618 which is found in the pyramids, seed patterns in sunflowers, leaf patterns in plants and all throughout nature. (If you don't know about the Golden Ratio, it's a natural ratio that has fascinated mathematicians for 2500 years. Look it up on Wikipedia.)

When Perez combined these 6 symmetries together, each of the 64 cubes in the matrix converted to an exact value. It made 32 pairs of triplets where each member of the pair occurs the exact same number of times as its twin. You get a set of mathematical stairs with 32 exact steps.

Isn't this fascinatingly similar to Sudoku? In Sudoku, each cell and column has to have each number from 0 through 9 . In DNA, each symmetry in the table has to be balanced within $0.1 \%$. Otherwise a copying error has occurred.

Perez found that the ratios of these 32 steps was exactly the same as the length of the sides of a famous pattern in fractal geometry, the "Dragon Folding Curve":


DRAGON paperfolding fractal curve and human genome:
POPULATIONS of each of the 64 codons are controlled by POSITIONS of the same codons in genetic code map
$\underset{\text { folding faces }}{ }$ ODD $\quad$ EVIding faces

In other words, the frequency each letter in DNA is governed by a pattern in fractal geometry.

Dr. Perez discovered some other interesting facts:

- Similar patterns with variations on these same rules are seen across a range of 20 different species, from the AIDS virus to bacteria, primates and humans.
- Each character in DNA occurs a precise number of times; each has a twin. TTT and AAA are twins and appear the most often; they're the DNA equivalent of the letter $E$.
- This pattern creates a stair step of 32 frequencies, a specific frequency for each pair.
- The number of triplets that begin with a $T$ is precisely the same as the number of triplets that begin with A (to within $0.1 \%$ ).
- The number of triplets that begin with a C is precisely the same as the number of triplets that begin with $G$, to within $0.1 \%$.
- The genetic code table is fractal - the same pattern repeats itself at every level. The micro scale controls conversion of triplets to amino acids and is in every biology book. The macro scale, newly discovered by Dr. Perez, checks the integrity of the entire organism. It detects copying errors.
- Perez is still discovering additional patterns within the pattern.

I am only giving you the tip of the iceberg. There are other rules and layers of detail that I'm omitting for simplicity. Perez presses forward with his research; more papers are in the works.

You can read Dr. Perez' paper at www.ncbi.nlm.nih.gov/pubmed/20658335. Perez doesn't describe this as Sudoku; I made that connection myself, and mapped Perez' 2D tables to the 3D charts you see above. I also helped him translate this paper from French into English before it was published.
You can access it free at http://www.scribd.com/doc/53748414/2010-Perez-Codon-Populations-in-Single-Stranded-Whole-Human-Genome-DNA-Are-Fractal-and-Fine-Tuned-by-the-Golden-Ratio.

OK, so what does all this mean?

- Neo-Darwinism always insisted that evolution occurs through random copying errors of DNA, filtered by natural selection. Perez' work shows copying errors cannot be the source of evolutionary progress, because if that were true, eventually all the letters would be equally probable. It would be statistically impossible for random copying errors to produce this kind of consistent ergodic pattern.
- This proves that useful evolutionary mutations are not random. Instead, they are controlled by a precise Evolutionary Matrix to within $0.1 \%$.
- When organisms re-arrange and repair their genomes through transposition, these rules are still kept.
- When organisms exchange DNA with each other through Horizontal Gene Transfer, the end result still follows the pattern.
- When organisms multiply their genomes through polyploidy, the end result still obeys the pattern. The above statements are ostensibly true because genomes obey these rules both before and after transposition, Horizontal Gene Transfer and genome multiplying events.
- DNA is able to re-create destroyed data. It appears to do this, in part, by computing checksums in reverse - almost like calculating the missing contents of a page ripped out of a novel.

No man-made language has this kind of precise mathematical structure. DNA is a tightly woven, highly efficient language that follows extremely specific rules. Its alphabet, grammar and overall structure are ordered by a beautiful set of mathematical functions.

A few more interesting patterns....
The most common pair of letters (TTT and AAA) appears exactly $1 / 13$ th as often as all the letters combined - consistently, in the genomes of humans and chimpanzees.

If you put the 32 most common triplets in Group 1 and the 32 least common triplets in Group 2, the ratio of letters in Group1:Group2 is exactly 2:1. And since triplet counts occur in symmetrical pairs (TTT-AAA, TAT-ATA, etc), you can group them into four groups of 16.

When you put those four triplet populations on a graph, you get a peace symbol:


I'm not at all suggesting that the peace symbol has been "hiding" in DNA all along, waiting to tell us to be peaceful. But do look at the symmetry. There is a completely abstract symmetrical relationship between four groups of 16 letters in the genetic alphabet, consistently seen in a vast array of genomes from HIV to mice to humans.

Does this precise set of rules and symmetries appear random or accidental to you?
This is part of why it is possible for DNA to be a code that is self-repairing, self-correcting, self-re-writing and self-evolving. It's a clue to understanding how your cells repair their DNA as they deal with sunburns, toxins and aging. It reveals a level of engineering and sophistication that human engineers could only dream of. Best of all, it's elegant.

## This is a Significant Discovery

I wrote the book "Industrial Ethernet," published in 2002 by the International Society of Automation, now in its $2^{\text {nd }}$ edition. As a communications engineer who knows how 1 's and 0 's make the Internet go 'round, I believe Dr. Perez' work points the way to many more $21^{\text {st }}$ century discoveries in biology - and computer science.

Why? Because no man-made code is anywhere near this elegant. It clearly demonstrates that an abstract mathematical model precedes the genetic code and controls its activity at a very high level. (As a communication engineer, Chargaff's rules greatly arouse my curiosity. I believe they deserve much more attention than they've received.)

In physics, String Theory is an attempt to formulate a mathematical model that attempts to unify quantum mechanics and general relativity. It models the universe as flowing from a multi-dimensional mathematical matrix.

You could think of the Evolution Sudoku Matrix as a similar way of viewing the cell and the genome. A sort of "String Theory" of biology. A set of abstract mathematical rules that govern the behavior of the code, and the paths of evolution. Controversial? Certainly. But a path that warrants further study.

Further discoveries may have far-reaching implications for cancer research, immunology, HIV research and evolutionary theory. This is not merely an interesting footnote; I believe is a key to cracking the genetic code and fully understanding our genomes.

Cancer has sometimes been described as "evolution run amok." (See Cancer models, genomic instability and somatic cellular Darwinian evolution by Mark P. Little). When cells divide, the error-checking mechanism is supposed to kill the daughter cell when there is an error; this is called "apoptosis." Failure for this to happen is what causes cancer.

Dr. Perez has noted interesting distortions of this matrix in cancer cells. Cancer is what happens when the cell's will to survive overrides the rules of this matrix. I suspect solutions for cancer research are hidden in this matrix.

## Dr. Perez has also found that some chromosomes in human DNA do not follow this rule, and these chromosomes are much more susceptible to being hijacked by the HIV virus.

Productive research in medicine and computer science will come from intensive study of mathematical patterns. This is just the tip of the iceberg. There is so much more here to discover!

## A Golden Ratio Evolutionary Hypothesis

I believe Perez' discovery is only the first of many. Further research will uncover more layers of order. Specifically, I suspect that this matrix also governs the local linguistic and grammatical rules of the genetic language. It defines what combinations of genes and chromosomes "make sense," as opposed to combinations that are genetic gibberish.
I predict we'll find another Sudoku matrix within this matrix. That smaller-scale patterns dictate the structure of genes and groups of genes.
The Evolution Matrix enables organisms to make highly educated guesses when they mutate in response to changes in their environment. When bacteria exchange information through horizontal gene transfer, the matrix dictates which combinations are allowable and which ones are not. Thus the matrix creates tracks for evolution to run on.

Perez compares the letters in DNA to a deck of 52 playing cards: As you take cards off the deck, they appear to be random, but you know there are four queens and four kings, four aces; each kind of card comes in spades, clubs, jacks and hearts; and that if you know what's been taken out of the deck, you know what's left in the deck.
Just like a genius poker player in Vegas, DNA maintains the integrity of its 'card deck' by carefully counting every card.

## The Evolution Matrix:

## Rails for Evolution to Run On

Among the many books l've bought along my journey of discovery was "The Plausibility of Life" by Kirschner and Gerhart. It encouraged me in my idea that evolution follows certain strict paths. They show how with the enormous diversity of life, many patterns occur again and again.
"There seems to be a conserved developmental module common to all appendages of all insects. Then, in the individual appendages in the lines of insects, the process was modified and supplemented, to give great variety."
"Looking at these changes, one has the distinct impression that the modifications involve rearrangement, local shortening or lengthening, or truncation, but are not anything basically new."
They zoom out and look at the grand sweeping patterns in the evolutionary tree:
"If we follow the path from the bacterium-like ancestor toward humans, we find repeated episodes of great innovation. New genes and proteins arose in each episode. Afterward, the components and processes settled into prolonged conservation. The existence of "deep conservation" is a surprise. To some biologists it is a contradiction of their expectations about the organism's capacity to generate random phenotypic variation from random mutation. To some, it borders on paradox when held against the rampant diversification of anatomy and physiology in the evolutionary history of animals.
"Widespread conservation must reflect some limit on the organism's freedom to generate viable variation in all directions under the impact of mutation. If phenotypic variation is really smoothly continuous in all directions, then all components should vary and nothing should be conserved for very long.
"Darwin's supposition that change is pervasive has to be replaced with the view that in the history of life some things change and others do not, and that change occurs in spurts and then becomes fixed and subsumed in all subsequent organisms. Superimposed on top is ongoing anatomical and physiological innovation.
"To go still deeper, why did the history of life on earth progress in this way instead of in the way Darwin surmised, where every aspect of the organism would be subject to change?
"Core processes that transformed the organism seem to have been invented only a few times, in episodes, on the way to multicellular animals."

In their book they describe a variety of organisms and features, showing that underneath the thousands of variations is a basic chassis that remains the same. They call this a "theory of facilitated variation." They observe that:

1) Variation is not random
2) Core processes are conserved across millions of years and species
3) DNA information processing and protein synthesis are identical across all organisms; Hox genes play similar roles everywhere, and limbs form the same way in all land vertebrates.
4) Most evolution since the Cambrian explosion has not been new processes, nor has it been wholesale alterations to the core. Rather it's been re-combinations of details within the same core process.
5) Core processes are easily re-combined to generate new features and creatures. "In response to an input, one or a few outputs are selected among those [possible] states and retained, often by stabilization."


Human


Dog


Bird


Whale

Limbs of different animals are similar because they all use the same mathematical "toolbox" - the evolutionary Sudoku matrix - to generate their forms.

Kirschner and Gerhart conclude:
"Rather than staggering like a drunken sailor, evolution marches along a myriad of paved pathways, changing direction without instruction, but taking large, forceful steps and avoiding many lethal obstacles."
They describe the texture of evolution very accurately. At the same time they still seem to be grasping for an underlying theory or guiding mechanism that describes the nuts and bolts.

Can you see that the Evolution Matrix could be a plausible structure for controlling these changes that they describe? It's a set of rails for evolution to run on - the "paved pathways" - that take large, forceful steps and avoid lethal obstacles.

## Evolution is not random.

## And it's not even slightly accidental.

It's driven by Transposition, Horizontal Gene Transfer, Epigenetics and Genome Doubling.

- Transposition, when cells re-arrange genes and large portions of chromosomes to adapt to changes in the environment
- Horizontal Gene Transfer, cells exchanging segments of DNA to instantly gain new features;
- Epigenetics, shaping heredity without altering the DNA sequence;
- Genome Doubling - DNA doubling to expand "hard drive space" and make room for novel features.

Look these up in Wikipedia if you're not familiar. They're the systematic mechanisms of evolution. Recommended book: "Evolution: A View from the $21^{\text {st }}$ Century" by James A. Shapiro. It describes these mechanisms in great detail.
Since organisms ranging from viruses to humans follow the matrix, then ostensibly all these mechanisms are likewise subject to the matrix. All are modular, all operate within definite "rails." When bacteria exchange genetic material, the newly adapted genome still obeys these rules.

Kirschner and Gerhart's description of evolution reminds me of a Swiss Army Knife, equipped with various blades that perform tasks as necessary.
Dear reader, I propose to you that the as-yet undiscovered properties of the Evolution Matrix hold the key to new breakthroughs. Not just in evolutionary theory but for immediately practical modern technologies and medicine.

Speak to me. Discuss this article and the Golden Ratio Matrix by posting a blog comment at http://www.cosmicfingerprints.com/mathematics-of-dna/

## More Articles, Audios and Videos:

Intelligent Bacteria: Cells are incredibly smart:
http://www.cosmicfingerprints.com/intelligent-bacteria/
"If you can read this, I can prove God exists"
http://www.cosmicfingerprints.com/prove-god-exists/
Evolution: The Untold Story:
http://www.cosmicfingerprints.com/evolution-untold-story/

## Additional References:

"Decoding non-coding DNA Codes: Human Genome Meta-Chromosomes Architecture" by Jean-Claude Perez http://www.scribd.com/doc/57828784/jcperezBeijing032011.

If you're able to read French, get Perez' book "Codex Biogenesis," available on Amazon. His English website is http://golden-ratio-in-dna.blogspot.com.

Chargaff's Rules: Wikipedia http://en.wikipedia.org/wiki/Chargaff\'s_rules

