

Please note that this manuscript has yet to be accepted by *Complexity International*

# A hybrid model for DNA: mathematical analysis

Huen Y.K.

*Cah Research Centre, P.O.Box 1003,  
Kent-Ridge Post Office, Singapore 911101*

Email: [webbooks@singnet.com.sg](mailto:webbooks@singnet.com.sg)

## Abstract

In a previous online paper in Complexity International, the author has developed the general theory of k-nary strings of which DNA as 4-nary strings are members of this family. Therefore properties of k-nary strings must be applicable to DNA when viewed as strings. But DNA must also satisfy biochemical properties from its biomolecular model. From the theory of k-nary strings, the theory of genomic instability was developed which enables one to compute the level of instability of a genome using a defined unit of measurement called Compressibility Ratio of Comprat. As Comprat is a wholistic unit of measurement, this suggests the emergence of a hybrid model for DNA incorporating both reductionistic and wholistic properties. Mathematically speaking, for such a hybrid model to work properly, reductionistic and wholistic proofs must be nonintersecting. The hybrid model could signal the emergence of a new type of complex system not covered by any existing definition in Complexity Theory.

## 1. Background

In a previous online paper, the author has established mathematical properties of DNA using sequence algebra or SA [Ref.3]. The theorems developed are perfectly general being applicable to any k-nary string including DNA as 4-nary strings. Another online paper followed closely on the heels of this paper in which the author has proved mathematically that genomes are intrinsically unstable and that the role of “junk DNA” or introns is to reduce this instability [Ref.5]. Genomic instability is measured by a unit called Compressibility Ratio or Comprat where absolute equilibrium is associated with unity value of Comprat. Comprat can also be interpreted as the distance displaced from equilibrium of a genome and that the degree of genomic instability is proportional to this value. Since the equilibrium state is based on the Comprat value of the natural number sequence, no DNA-sequence can reach this state via the process of genomic stabilisation. This implies that life is permanently teetering at the brink of chaos or disequilibrium and that the state of equilibrium is to be avoided as it will spell

stagnation and extinction [Ref.5]. The theory of genomic instability could provide explanations to some puzzling questions from molecular biology posed in this paper.

The theory of sequence algebra or SA, the theory of first order compressibility of strings, and the theory of genomic instability were employed in mathematical analysis. SA has been under development since 1994. Readers needing more information on SA are referred to previously published papers in the Reference [Ref. 7 to 13].

## 2. A hybrid model for DNA

Traditionally, properties of DNA-sequences are explored using biochemical methods which is reductionistic. This is because the whole theoretical structure is built upon the four nt-bases like brick work. On the other hand, the genomic instability model of DNA is developed without any detailed knowledge of DNA-codings and is therefore wholistic. In this paper, it is demonstrated that the inclusion of wholistic analysis in DNA will not cause any anomaly or paradox provided reductionistic and wholistic analysis are nonintersecting. Since this is a mathematical paper, all proofs are presented as theorems.

Theorems 1 to 14 are either under review or published and are listed in the Appendix without proofs [Ref.3]. Theorem 15 to 19 are withheld as these are not relevant to the present topic [Ref.4]. Theorem 20 to 21 have been proved in a late paper [Ref.5]. New theorems added in this paper will start from 22 onward.

**Theorem 22: Two axiomatic theories, one reductionistic and the other wholistic, cannot intersect either exactly or partially.**

**Proof:** A reductionistic axiomatic theory is founded on reductionistic axioms whilst a wholistic axiomatic theory is founded on wholistic axioms. If perfect intersection exists, then a wholistic axiom can be proved mathematically from one or more reductionistic axioms or vice versa. But axioms are definitions which cannot be proved. Thus there can be no exact intersection between two such theories. Even if there is partial intersection, it still means the above paradox still persists. Therefore a reductionistic theory and a wholistic theory cannot intersect. Q.E.D.

**Corollary 1 to theorem 22: Unification of a hybrid theory into a single theory is impossible.**

**Proof:** According to theorem 22, the two theories cannot intersect. Unification cannot be attained with two nonintersecting subtheories because this process requires the integration of two theories into a single theory where intersections are unavoidable. Q.E.D.

**Theorem 23: The theory of genomic instability is wholistic.**

**Proof:** The Compressibility Ratio or Comprat value which measures the degree of genomic instability does not depend on any detailed knowledge of DNA-codings. Therefore Comprat must be a wholistic unit of measurement. Therefore the theory of genomic instability is wholistic. Q.E.D.

**Theorem 24: The larger the size of a genome, the closer it approaches equilibrium.**

**Proof:** In general the higher the evolutionary ranking of a species, the larger is its genome size although there are exceptions. This means that higher ranked species are closer to equilibrium and therefore to early extinction. For a general proof, we need to prove using a general genome generated by a random number generator. This should be valid since Comprat is wholistic as its evaluation does not depend on detailed DNA-codings. Here is the proof:

```

general_genome:sum((random(4)+1)/z^i,i,1,k);
natural_number_sequence:sum(1/z^i,i,1,k);
In the above two formulae, only k is a variable, so that for
the computations of string lengths, these formulae are modified as
follows:
general_genome:(string_length("sum((random(4)+1)/z^i,i,1,)" +string_length(k));
natural_numseq:(string_length("sum(1/z^i,i,1,)" +string_length(k));
Therefore:
Comprat =
general_genome:(string_length("sum((random(4)+1)/z^i,i,1,)" +string_length(k));
-----
natural_numseq:(string_length("sum(1/z^i,i,1,)" +string_length(k));
To compute over a range of k values, we use the following
Masya 2.2 programline.
for k from 1 thru 1000000000 step 1000000 do print(genome_size(k)
*float(( string_length("sum((random(4)+1)/z^k,k,1,)" +string_length
(k))/( string_length("sum(1/z^k,k,1,)" +string_length(k))));

```

*Table 1. Comprat vs Genome size*

```

-----
1.75 genome_size(1)
1.66667 genome_size(101)
.....
1.66667 genome_size(901)
1.63158 genome_size(1001)
.....
1.63158 genome_size(9901)
1.6 genome_size(10001)
1.6 genome_size(99901)
1.57143 genome_size(100001)
.....
1.52174 genome_size(10000001)
.....
1.52174 genome_size(90000001)
1.5 genome_size(100000001)
.....
1.5 genome_size(990000001)
1.48 genome_size(1000000001)
.....
1.48 genome_size(9990000001)
-----

```

It can be seen from the computed values in Table 1 that Comprat does decrease with increasing size of genome. Q.E.D.

Note that in the hybrid model, the wholistic component is the theory of genomic instability. In the next section, the author would attempt to find plausible explanations to some puzzling questions in molecular biology using the theory of genomic instability.

### 3. Some puzzles from molecular biology

#### **Q1: How does one explain the wide variations of genome sizes within each taxonomic grouping?**

The wide variations of genome sizes within each taxonomic grouping or phylum can be explained in terms of genomic instability. It has been demonstrated numerically that the Compressibility Ratio or Comprat does not decrease smoothly but are interrupted by alternating peaks and troughs with increasing lengths of “junk DNA” or introns [Ref.5]. Both genes and “junkDNA” are growing over time in evolution. Growth in lengths of “junkDNA” may have a historical background ranging over millions of years but the lengths associated with each species is purely fortuitous as the length could settle into any local stable valleys in the genomic stabilisation process. This explains why there are such wide variations of genome

size. Therefore one should expect to find within each phylum, species with a low percentage of junkDNA. One could assume that originally all species had low percentage of junkDNA but over a long period of time junkDNA grew in lengths which affected some species more than others.

**Q2: Is there any validity in the relation between genome size and evolutionary progress?**

The validity of the relation has already been proved in a previous paper [Ref.4] but for copyright reason, since this paper has not yet been put online, reference to it has been put on hold. Nevertheless, we can prove this based on the answer from Q1 where it is proved that genome size do vary due to genomic instability. But within every taxonomic grouping, there are species which have low percentage of junkDNA and these are the ones of most interest since if one plots minimum genome size from each phylum against evolutionary progress, the curve is quite smooth showing that there is validity between minimum genome size within each phylum and evolutionary progress.

**Q3: Could general properties of DNA be predicted without detailed knowledge of DNA-codings?**

For general properties, the answer is No. Only specific properties related to algorithmic information can be predicted and these are wholistic properties which do not depend on the detailed knowledge of DNA-codings. Amongst these, the most useful is the theory of genomic instability. This is the theory which provides plausible explanations to the puzzles raised in this section.

**Q4: From theorem 24, it has been proved that the Comprat value decreases with increasing size of genome (see proof in Theorem 24). Does this mean that species in higher taxonomic groupings are closer to extinction than lowly ranked species?**

Theorem 24 asserts its mathematical truth. However this topic has not been studied in depth within the theory of genomic instability. But there are scattered evidences pointing to its validity. For example, we know that prokaryotes are immortal but these have very short genomes without the benefit of genomic stabilisation by junkDNAs or introns. On the other extreme, we know that mammals have large genome size which are closer to equilibrium than prokaryotes. Man with its large sized genome is most probably closer to extinction than some lowly worms or fungi.

**Q5: Are prokaryotes genomically more unstable than eukaryotes?**

The answer is Yes because prokaryotes have smaller genome sizes which do not benefit from genomic stabilisation. In general, prokaryotes will have higher Comprat values than eukaryotes. Therefore they are further removed from the equilibrium state and are therefore more unstable. They will mutate more rapidly than eukaryotic species further up the evolutionary scale.

**Q6: How does prokaryotes achieve genomic stabilisation without “junkDNA”?**

Right now the answer is not completely clear. We guess that a genome existing as a closed ring will depend on genomic vibrational modes in the ring to achieve stability. The mechanism may parallel that of a closed superstring which also exhibits vibrational modes. For example, if a prokaryotic genome has a size of 6000 nt-bases, and if there is a periodic substring within it with a periodic interval of  $6000/k$  where  $k$  divides the numerator exactly, then a vibrational mode could be established. Such a vibrational mode is a periodic standing wave which is expected to contribute to a lower value of Comprat. This could contribute to genomic stabilisation of prokaryotes.

**Q7: We know that there are regulatory functions within “junkDNA” which have nothing to do with genomic stabilisation. Won’t these affect the Comprat values?**

In the human genome, 97% are labelled as “junkDNA” which may also contain short stretches of DNA sequences with regulatory functions other than genomic stabilisation. But these are short stretches compared to the wholistic length of “junkDNA” and will have no effect on the value of Comprat. So junkDNA is quite versatile being able to tolerate the presence of other regulatory functions whilst performing its main role in genomic stabilisation. For example, even the junk repeats left by virus insertions could contribute to genomic stabilisation. But one must concede that these subsidiary regulatory functions are local and cannot be predicted by a wholistic theory such as the theory of genomic instability.

**Q8: DNA is being called a complex system. Does the hybrid model fit any existing definition in Complexity Theory?**

If the definition of Complexity Theory according to Lucas is accepted, then the answer is No! [Ref.1]. Complexity Theory is still an evolving “science” without a set of definitive analytical tools and methods of quantification. The development of SA and the general properties of k-nary strings should be considered innovations in Complexity Theory. From this point of view, DNA is a novel complex system which does not fit existing definition in Complexity Theory. Since Complexity Theory is evolving, its definitions cannot be static. The hybrid model would require Complexity Theory to accept both reductionistic and wholistic analysis. This may go against the grain of some complexity scientists who believe that the paradigm should be purely wholistic. Thus a hybrid model may be a compromise solution in Complexity Theory.

**Q9: What justification is there that the theory of genomic instability is a wholistic theory?**

The Compressibility Ratio or Comprat value which measures the degree of genomic instability does not depend on any detailed knowledge of DNA-codings. Therefore Comprat must be a wholistic unit of measurement. Therefore the theory of genomic instability is wholistic. Clearly weighing a book to predict its printed contents is an impossibility. But the length of a string is a measure of its algorithmic information and not absolute information. There lies the difference.

**Q10: Could genomic instability provide a plausible explanation on the Cambrian Explosion?**

So far, we think the theory of genomic instability could provide the most plausible explanation on this event. The Cambrian Explosion should be viewed as the result of a major reverse mutation cluster. In this theory, genomic instability is measured by Compressibility Ratio or Comprat. We start with the Comprat equation of the genome of a species as follows:

$$\text{Comprat of Genome} = \text{Comprat of (Genes + junkDNAs)}$$

Suppose on the RHS the gene content was suddenly bumped upward by some unknown catastrophes such as an intensive radiation shower which killed off most species but some species still remained viable with much increased gene contents. These meant that some species had received a step jump in gene contents and became highly unstable. To stabilise, two options were available. In the natural process, the slow path would have been taken which required a long period to increase the lengths of junkDNAs or introns. But because of high instability, the faster path was to shed some genes to lower the Comprat values. The shedding of genes could be a random process resulting in the proliferations of highly varied life forms most of which went extinct but some remained viable. This explains why over a short time in the Cambrian Explosion, so many phyla suddenly appeared [Ref.6]. Since mutations occurred

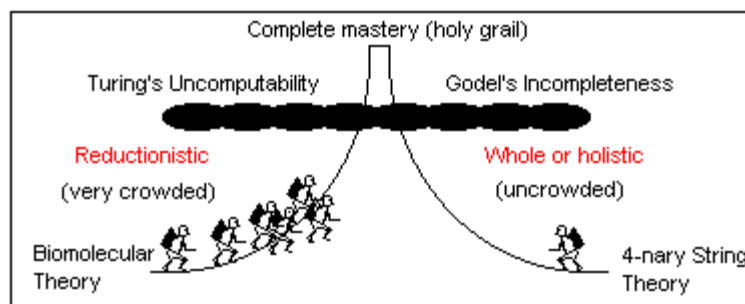
in clusters as reverse mutations, there were no connections between the ancestors and the progenies. In fact the ancestors could appear to be more modern than the progenies. Bird-Dinosaur controversy could be caused by such a phenomenon. If reverse mutation is valid, then Bird could be the ancestor of Dinosaurs.

**Q11: Is it true that the above questions cannot find explanations from traditional sources?**

This is true because all these questions require wholistic explanations which could be provided by the theory of genomic instability. If one scans the DNA-codings like one reading a book, one will not find any entries there. This is because the analogy between DNA and any human language is most probably wrong. These messages are wholistic and can only be interpreted from the genome as a functional whole. Biological databases are neutral which explain the surfeit of controversies posed by various conflicting theories. But plausible explanations are provided by the theory of genomic instability.

**Q12: Will the hybrid model be the theory of everything in molecular biology?**

No such a claim has been made. This answer is best provided graphically in fig.1:



*Fig.1*

By adopting the hybrid model, the exploration space is immediately doubled. If the peak in Fig.1 is Mount Everest, then molecular scientists have been working from the Southern Slope (Indian Subcontinent) paying no attention to the opportunities offered by the Northern Slope (Tibetan Plateau). There are plenty to do even if no attempt is mounted to scale the peak. After more than 2000 years, number theorists are still not within sight of the peak even though number sequences are so much simpler than DNA-sequences. How much more difficult will it be in our attempt to master the complexity of a system such as DNA?

## 4. Conclusions

In the hybrid model theory developed in this paper, it is found that whilst biomolecular theory is reductionistic, the theory of genomic instability is wholistic. It is proved that for a hybrid model to work, the two subtheories must not intersect. It is also proved mathematically that the hybrid model theory cannot be unified into a single theory. There is also the dichotomy of difference in philosophy between biochemists and pure mathematicians where the former are experimentalists whilst the latter are abstractionists. Twelve puzzles in biomolecular theories have found plausible explanations based on the theory of genomic instability. The hybrid model may signal the emergence of a new paradigm in complex system model building.

## Appendix

### *List of previously published theorems*

Theorem no. 1 to 11 was published in Ref.3 and Theorem no.12 to 14 can be found in Ref.4 (recently submitted to Online Bioinformatic). Theorem no. 15 to 19 are also from Ref.4 are withheld as these are found irrelevant to the present topic.

Theorem 1: The Relcomp of an infinitely long string is uncomputable.

Theorem 2: A 2-1tuple is the smallest tuple which can be compressed by the most concise formula.

Theorem 3: All finite k-nary strings of lengths greater than k characters are always compressible.

Theorem 4: The incompressibility of a k-nary string is conserved in upward migration but lost in downward migration.

Theorem 5: The probability of Randomness appearing in a k-nary string of length k is given by  $p_k = k!/k^k$ .

Theorem 6: The condition of Relcomp = 0 can never occur in a finite k-nary string.

Theorem 7: Randomness is associated with Relcomp = 1 and it has maximum Information Content.

Theorem 8: The string length of the most concise formula is almost linearly proportional to the logarithmic function of the string length of its expanded Taylor-Laurent sequence.

Theorem 9: The Absolute Information Content (or AIC) of a quasi-periodic k-nary string is given by the logarithmic function of the sum of the Absolute Information Content of the k resolved strings where each represents on symbol.

Theorem 10: Any k-nary string can be resolved into the sum of k most concise formulae each representing a single symbol.

Theorem 11: The longer the DNA-sequence, the closer the agreement in Relcomp between the natural number sequence and the DNA-sequence.

Theorem 12: All DNA-sequences have degrees of instability.

Theorem 13: DNA-sequences attempt to reduce instability by increasing the lengths of junk DNA's but the process cannot be completed.

Theorem 14: If a genome with AIC value of AIC1 is increased to AIC2 where  $k = AIC2/AIC1$ , then the string length of the genome can be predicted by AIC Hypothesis.

Theorem 15 to 19: display withheld.

Theorems 20 and 21 are published in Ref.5:

Theorem 20: Introns or junk DNA's carry finite amount of Algorithmic Information.

Theorem 21: Introns or junk DNA's help to stabilize unstable genomes by the mechanism of negative feedback but the control actions cannot be completed.

## References

- (1) Chris Lucas: Quantifying Complexity Theory.
- (2) Huen YK: Brief comments on the Theory of First Order Compressibility of Strings. Submitted to Complexity International (under review).
- (3) Huen YK: Mathematical Properties of DNA:Algebraic Sequence Analysis, Volume 1: 42-59, 2001. Online Journal of Bioinformatics. Submitted for review to Online Journal of Bioinformatics.
- (4) Huen Y.K.: AIC Hypothesis on C-value Paradox, Submitted for review to Online Journal of Bioinformatics.
- (5) Huen YK: Brief comments on junk DNA: is it really junk? Submitted to Complexity International (under review).
- (6) CAMBRIAN EXPLOSION/ORIGIN OF THE PHYLA.
- (7) Huen YK A matrix map for prime and non-prime numbers, Int J Math. Educ. Sci. Technol 6: 913-920, 1994.
- (8) Huen YK The twin prime problem revisited, Int J Math. Educ. Sci. Technol 6: 825-834, 1997.
- (9) Huen YK Is pie periodic? Int J Math. Educ. Sci. Technol 1, 19-26, 1998.
- (10) Huen YK Order analysis in sequence algebra, Int J Math. Educ. Sci. Technol 2: 259-269, 1999.
- (11) Huen YK Explicit formulations versus recursive formulations – studies based on the Fibonacci sequence and the factorial function, Int J Math. Educ. Sci. Technol 6: 788-795, 2000
- (12) Huen YK An Introduction To Sequence Algebra, ISBN 981-04-0866-8, The Library of Congress, U.S.A., Certificate TX 4-252-656, 1999 <http://www.cahresearch.com/>
- (13) Huen YK Advanced topics in sequence algebra, ISBN 981-04-1641-5, 1999. <http://www.cahresearch.com/>