IC/93/65
INTERNAL REPORT
(Limited Distribution)

International Atomic Energy Agency
and
United Nations Educational Scientific and Cultural Organization
INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS

THE RNA WORLD, AUTOMATIC SEQUENCES AND ONCOGENETICS

K. Tahir Shah
International Centre for Theoretical Physics, Trieste, Italy
and
World Laboratory, Lausanne, Switzerland.

MIRAMARE - TRIESTE

April 1993

ABSTRACT

We construct a model of the RNA world in terms of naturally evolving nucleotide sequences assuming only Crick-Watson base pairing and self-cleaving/splicing capability. These sequences have the following properties.

1) They are recognizable by an automaton (or automata). That is, to each k-sequence, there exist a k-automaton which accepts, recognizes or generates the k-sequence. These are known as automatic sequences. Fibonacci and Morse-Thue sequences are the most natural outcome of pre-biotic chemical conditions.

2) Infinite (resp. large) sequences are self-similar (resp. nearly self-similar) under certain rewrite rules and consequently give rise to fractal (resp. fractal-like) structures.

Computationally, such sequences can also be generated by their corresponding deterministic parallel re-write system, known as a DOL system. The self-similar sequences are fixed points of their respective rewrite rules.

Some of these automatic sequences have the capability that they can read or "accept" other sequences while others can detect errors and trigger error-correcting mechanisms. They can be enlarged and have block and/or palindrome structure. Linear recurring sequences such as Fibonacci sequence are simply Feed-back Shift Registers, a well known model of information processing machines.

We show that a mutation of any rewrite rule can cause a combinatorial explosion of error and relates this to oncogenetical behavior. On the other hand, a mutation of sequences that are not rewrite rules, leads to normal evolutionary change. Known experimental results support our hypothesis.

1. Introduction. From the computational viewpoint there are three rather curiously related aspects of genes - the formal linguistic, the information processing, and the self-similarity of nucleotide sequences. The linguistic aspects of the genome and the genetic code has been discussed by many authors (see, e.g., Searls, 1992). Rewrite systems, especially the L-systems were invented and applied by Lindenmayor (1987) and others (see, e.g., Prusinkiewicz and Hanan, 1989) to the problem of cell lineages. The information processing aspect is not known to be modelled despite the fact that introns seem to behave like automata that move around and cut, splice and possibly detect errors in replication. Information contained in nucleotide sequences would be useless if there are no mechanisms to read, sort and process that information. How this is done is itself information contained in other sequence. But these sequences are neither expressed nor they need to be expressed. The self-similar sequences are known to generate fractals. Graphic studies of fractals show their remarkable similarity to many biological organs and plants. The fractal nature of biological organs is curiously related to cell development and differentiation properties and the underlying structure.

Mathematicians have studied recently deterministic parallel rewrite systems in the context of non-linear dynamics and computation theories and have shown that such systems, called DOL systems, generate recurrent sets which are self-similar and give rise to fractal structures (Dekking, 1982a and 1982b). It may simply not be a coincidence, but it is highly plausible that some kind of relationship exists between the syntactical structure of the RNA and DNA, the genetic code, the cell development and differentiation cycle (or all of them) and the fractal nature of biological organisms. We argue on the basis of existing experimental evidence that such a fractal generating structure of genomes must have been inherited from early prebiotic nucleotide sequences of RNA. A further analysis suggests that they are the precursors of introns. We shall call them proto-introns. Once, such a structure and (later) a code have been devised by nature, both remained unchanged throughout evolution - it is too costly to change it all.

The prebiotic environment, most likely containing only simple chemicals, suggests that the generation of early nucleotide sequences with these properties must have been a simple process too. One can possibly consider only a very simple process because of the nature of prebiotic chemistry, that is, there were only a few chemicals present at the time having only simple combinational properties. A complex formation process like the one we know exists in the genome of multicellular organisms at present requires a complex series of mechanisms, conditions, and material resources (e.g., protein enzymes). The sequence generation and replication mechanisms must have been very simple and self-sustaining also because (accepting the fact) there were no complex proteins to act as enzymes to supplement the activity of ribozymes.

In this paper, therefore, we suggest a model in which a simple process inducing the polymerization of nucleotides gives rise to interesting sequences having the right kind of properties - automaticity (this term will be clarified later) and self-similarity. Not only do these sequences carry information and replicate, but some of them do information processing within the genome. They have aperiodic structure and their self-similarity under certain map could give rise to fractal structures at higher levels of life forms. Higher dimensional space can be filled by lower dimensional sequences in a manner like fractals. For infinite sequences such properties are well known. We show that finite-length sequences with self-cleaving and splicing capability give rise to such characteristics.

For the evolution of higher life forms it is imperative that earliest living systems must have been capable of both replication and complexification. Complexification here means evolving towards longer and longer sequences with more and more complex structure. In this way more and more functions are added. Replication in this case means reproducing the same sequence of nucleotides over and over again exactly or with low mutation rates but not necessarily using the template method and protein enzymes, i.e., sequence-to-complement and complement-to-copy method, as is the case for DNA. An example in Section 3 clarifies this comment. Information storage and replication, that is the transfer of information both for internal genetic mechanisms such

as reading a sequence, error detection during or after replication and other similar tasks, and the genetic message to reproduce the next "generation", are all difficult to explain for the early prebiotic period. There were no protein enzymes and the replication mechanism would not have reached the complexity known to exist in the contemporary eukaryotic cell. It was considered too difficult to understand how first nucleotide sequences could have replicated without a protein catalyst, which in its turn could not arise without a nucleic acid genome to encode it.

However, the discovery that ribosomal RNA intron of the protist *Tetrahymena* excises itself from the rRNA precursor without a protein catalyst, implies that the very first molecules might have been RNA replicases. That is, such molecules catalysed their own replication. This hypothetical RNA would have functioned as both the information carrier and replication enzyme. Such self-replicating introns, an RNA element that can splice itself out of an RNA molecule and then cleavage itself (i.e., it can remove and/or insert itself), is the fundamental unit that we wish to study in our model.

The first task, however, is to understand how nucleotide sequences were formed, whether they were random or followed some specific formation mechanism resulting in some specific kind of a structure. How were these structures to serve some specific purpose during the course of evolution, such as information processing? How does the behavior of these sequences change under mutation?

The next section summarizes a few relevant aspects of the RNA world. In Section 3, we shall describe a model of prebiotic RNA based on the properties of appropriate DOL systems. In Section 4, a brief mathematical discussion on DOL systems, especially on Morse-Thue and Fibonacci sequences is given. Section 5 deals with oncogenetical aspects and mutation problem. Section 6 concludes the discussion along with a few speculations.

2. The RNA World and Related Issues. It has become clear in the past decade that the RNA world - an early form of the living world - played a crucial role in the origin of higher life forms. Early oceans were ideal for the development of polymer

chains due to their suitable temperature range as well as the presence of required chemicals, such as oligonucleotides, peptides and oligosaccharides. It is speculated that most probably many viroid and virus-like chains that were capable of replication by a process analogous to crystallization were formed in the initial stages of prebiotic life (Diener, 1989). According to some theories, the evolution of self-replicating chains have been the ribose type RNA initially, which involves two phases of polymerization. In the first phase, activated nucleotides combined directly either in the solid state or in solution, to form the primary polynucleotides. In the second phase, the primary strands served as templates helping the synthesis of secondary complementary polynucleotides from activated monomers in acquous solution, governed by specific base-pairing hydrogen bonding (A-U and G-C). Orgel et al. (1987) studied the non-enzymatic prebiotic chemistry of both types of polymerization, and in particular, the template reactions. We use both phases simulataneously to generate polynucleotide templates which grow incrementally into longer and longer chains.

What are the minimum requirements for the construction of a chemical evolving system? Two informational processes are considered fundamental to the operation of an evolving system (Watson et al., 1987):

(1) The genetic information must be replicated in order to compensate for the inevitable loss of individual copies due to chemical degradation.

(2) The genetic information must be expressed as a behavoiral phenotype so that its usefulness can be assessed by natural selection.

That is, the basic requirements are that the informational polynucleotide sequence can be replicated and expressed as a behavioral phenotype in a way that is sequence dependent in addition to requirement that the ability to carry out replication and expression be part of the expressed phenotype. Both these processes rely on the use of polymers enzymes in biology. Clearly, both issues are discussed widely. The replication of polymer chains has been a most important theme of discussion and there are two aspects that need consideration. One is, of course, the widely studied enzymatic activity. The other is the information carried by a sequence and how, internally within

the sequence, information is processed. The RNA replication during the prebiotic period had occurred in the absence of presently known mechanisms (e.g., in the living cell's own RNA) using both protein (as enzymes) and template.

There are two principal models on the origin of RNA and its replication. The models in question are those proposed by Eigen and Orgel. In Eigen's model (Eigen and Winkler-Oswatitisch, 1992) of RNA, enzymes are present but no templates. On the contrary in Orgel's model (1987; 1992) only templates are used and enzymes do not take part. To support the hypothesis that RNA was an ancestral molecule of life, we should be able to explain how RNA, or for that matter, any kind of nucleotide sequence can be made without prior presence of templates or enzymes. Obviously, in this regards the first phase mentioned above should not be left to pure chance. It should come out as the most likely event to occur.

Given that there were no (proteins) enzymes in the prebiotic era, we now focus our attention to non-ezymatic replication and especially the work by Orgel and coworkers (see Orgel, 1987; 1992) for a review and bibliography), who classifies non-enzymatic replicating systems as:

1. Those systems that do not involve direct replication of a polymeric molecule, but may help to create a favorable environment for such replication, e.g., system with autocatalytic syntheses of one of the building blocks of replicating polymers.

2. Systems involving replication of informational polymers - conservation of sequence information is crucial during the replication process. Such systems are referred to as 'informational replicating systems'.

Most researchers discard non-informational part of replicating systems for various reasons such as that they cannot evolve by natural selection, or they do not store information in a stable way (Orgel, 1992). From mathematical point of view it is not obvious whether a sequence is carrying information or not, unless one can prove that there does not exist any language (in the formal sense) to which this sequence is associated - not an easy task. Commonly, the so called non-informational part is referred to that part which is not expressed. A sequence, such as an intron which is not

expressed, does not imply that it does not carry any kind of information. The intronic parts do carry information, but for other purposes. Certainly, it will not show up in any functionality of an organism. Since the intronic part is also replicated there must be some reason that nature keeps these segments of a genome.

Non-enzymatic template-directed RNA synthesis: The most attractive candidates for the role of information macromolecules are considered those compounds that have inherent template structure. Templating simplifies the task of information transfer during the replication process. The ideal candidate would be a polymer that acts as a template to direct the synthesis of additional copies of itself. *The issue is how inherent template structure were formed initially.* A template is an idealized concept. It is nothing but a sequence of purine and pyrimidine bases attached to the suger-phosphate support. In essence, the process of sequence formation that we shall describe in section 3 is the process of template formation.

The only known (Joyce, 1987) systems of this type are those where self-replication has been demonstrated in the laboratory using chemically modified RNA substrates and well chosen self-complementary template. One of the most appealing properties of RNA (and other RNA like materials) is its inherent template ability. *There is a preferential interaction between complementary nucleotides based on the specificity of Crick-Watson pairing.*

How are fundamental sequences formed? In the laboratory experiment it was shown (see for example Joyce, 1987) that the interaction of monomers in free solution results in the formation of dimers. These dimers are subsequently bound to the template and begin to elongate proceeding in the 5'->3' direction. The overall yield is determine by the relative rates of initiation and elongation. They are in their turn dependent on the stability and the concentration of both monomer and template. The rate of yield of oligomers (beyond the dimers) and the chain length of the average oligomers is proportional to the rate of elongation relative to initiation.

It was shown by Inoue and Orgel (see for example Joyce, 1987; Orgel, 1992) that *monomers are incorporated into newly synthesized oligomers if and only if their complement is present in the template.* This important result demonstrates that the specificity of Crick-Watson pairing is sufficient to provide the basis for information transfer in template directed oligomerization reactions. An interesting laboratory case (see for example, Orgel, 1987) of replicating informational polynucleotides is the following.

The sequence CCGCC has been shown to catalyse the synthesis of GGCGG from a mixture of activated G and C nucleotides. This was the first clear demonstration of information transfer in a non-enzymatic template-directed reaction system that used the 'defined-sequence' template CCGCC to direct the condenstaion of GGCGG. The yield of GGCGG was less than 20% of the templates. Although the elongation of GG to GGC was very slow, however, once GGC is formed, it is converted efficiently to GGCG and GGCGG.

There are some interesting characteristics of chemical reaction. For instance, small differences in template size and sequence can make a considerable difference in the efficiency of the template-directed reaction, e.g., CCGCC is reasonably good template, CGC is completely inert. It is not clear if there are such differences for longer chains. It was noted (by Orgel and his coworkers) that as long as monomers are joined to the growing chain by a 3' -- 5' linkage, elongation can continue, yielding oligomers whose length may even exceed the length of the template. This can be explained as due to the phenomenon of "chain sliding". The complementary strand is able to move relative to the template, exposing previously occupied template sites that can then be used to direct further elongation. The complementary strand may slide in such a way as to allow synthesis to continue up to the length of the second template.

There were some unexpected results. Products of the form GGC were most abundant, which implies that GG is converted more rapidly to GGC than to GGG. Major products upto 12-mer were of the form GGGC and GGC. However, for major product of length 12 nucleotides and longer were all of the form GGC. The crucial

question is now whether any of these RNA sequences are associated with any interesting behavioral phenotype such as autocatalytic activity, or simply whether they are used for internal information processing? In fact, we shall construct automatic sequences of the same type and discuss their properties.

Earlier we mentioned correct and faithful replication of genome as an essential condition that any life self-sustaining system has to satisfy. This includes the capability of catalyzing chemical reactions during the self-replication process. The question of catalytic behavior can be addressed in the context of self-replication in our model as follows:

1) Each of our fundamental sequences starting from C was catalyzed by its complementary sequence starting from G.

2) Each half of the palindrome in a fundamental sequence can catalyse the other half.

Another crucial problem in any theory of the origin of life is to provide a faithfull replication mechanism that is almost always error free. If it is not possible to device a perfect and error free mechanism then it must contain a mechanism to detect errors in replication and correct them. Since due to natural conditions an error-free replication is not always guarenteed, an almost impossible task, it is more likely that nature developed error-detection and error-correction mechanisms in order to make sure that the end product is the desired one. A failure of such mechanisms could lead to bizarre behavior of life forms, such as production of cancerous cells and similar behavior radically different from normal evolutionary changes which may occur due to mutations. *We stress that a consideration of these mechanisms is as important as the genetic message itself.*

Since we consider the replication mechanism along with both the error-detection and error-correction mechanisms essential for life, we emphasize that there should be two components of any genome: the *information part* and the *information processing part.* The information processing is purely internal to a genome. It is not necessary that it should be expressed. The error-detection and error-correction mechanisms are to be considered as components of the information processing and internal guidance system

of a genome. The automaticity deals with the question how a sequential structure of necleotides can achieve this. This is the issue that we try to resolve.

Since in our model, Morse-Thue sequence doubles its size as it grow, we point out a chemical mechanism by which larger genomes could have evolved. Insertion type point mutations within recognition site would abolish cleavage at that site and molecules twice the size of the previous ones would be synthesized. This leads to what has been called the structural periodicity. If so, remnants of structural periodicity might be detectable in present day genomes of great antiquity. For example: all viroids (except hop stunt viroid) exhibit structural periodicities characterized by repeat units of 12-, 60- or 80-nucleotides residues depending on the viroid species.

To end we quote Joyce (1989). He asked " Could it be that the first genetic molecule did not have template properties and replicated in some other way? We must postulate that some mechanism existed to ensure that a particularly useful sequence would facilitate the production of additional copies of itself." The most important feature of RNA is that it combines genotype and phenotype in a single molecule, so that replication of RNA enables Darwinian evolution to occur at the molecular level. This question has motivated us to propose a mathematical model described in the next section.

3. The Model. In the present model, the RNA world initially consisted of only nucleotides C, G, A and U, in the nucleotide soupe but no chains were present. All other chemicals which were necessary for chain formation were also present. This is our starting point. From this point we construct incrementally templates of larger and larger size and distinctly different sequential structure. Given that there is a chemical affinity between complementary bases (in purine and pyrimidine) C, G, and A, U, the well-known Crick-Watson base-pairing, we propose the following scenario:

Let us take a starting molecule C. To illustrate the mechanism of *incremental template formation* in a simple manner consider only C and G nucleotides. If there is a G nearby it is highly likely that it might attach itself to C, or simply, it is attracted towards C and

then concatenate itself to either left or right of C, i.e., attach itself to the sugar-phosphate base. We shall call this process left- or right-concatenation, as the case may be. Assume from now onwards that C-G pairs either attach to each other or simply are attracted towards each other. To form a sequence, bases are to be attached to the sugar-phosphate support. This is possible in two ways. The complementary nucleotide is simply attracted and attach itself to the sugar-phosphate support (from now onwards we shall call it simply the support and bases together as nucleotides), or if a nucleotide attaches itself to its complement (forming a double strand) then it has to detach itself before concatenating or unfolding. Let us now assume that it unfold itself in the manner that G is concatenated to C on the support. The word 'unfold' does not really illustrate the mechanism properly for only two nucleotides. However, once the chain is a few nucleotides long, it is easy to understand * what is unfolding and how it is different from left- or right-concatenation.

In the next step, two new nucleotides attach themselves to this chain of two elements. It can unfold itself in two distinct fashion. In the first case, it may unfold like an hairpin, while in the second case, the newly attached sequence detach itself and concatenate itself to the original sequence as shown below. Thus the two possible ways of forming sequences are:

1) Attract and concatenate

2) Attach and unfold or detach and concatenate the whole block to the left or right of the original sequence.

---

* The following diagram illustrates the point for an arbitrary sequence

```
CGGCG
                unfolds
          ---------------------------->  CGGCGCGCCG
GCCGC

CGGCG
          detaches & concatenates
          ---------------------------->  CGGCGGCCGC
GCCGC
```

That is, we have two different sequences CGGC and CGCG. Repeating this process further, we get:

CGGC ----> CGGCGCCG
GCCG

CGCG ----> CGCGCGCG
GCGC

Since the second option gives rather a trivial, periodic sequence, we do not consider it any more in the present context. There are many sequences of this type in nature. However, the first option gives a series of interesting sequences as follows:

C
CG
CGGC
CGGCGCCG
CGGCGCCGGCCGCGGC
...

A complementary set of sequences can be obtained by starting with G as follows:

G
GC
GCCG
GCCGCGGC
GCCGCGGCCGGCGCCG
....

The complementary sequence can also be obtained mathematically (section 4) by simply replacing every occurrence of C by G and of G by C. One may speculate at this point that two complementary sequences may have formed the precursor of early DNA or double strand RNA sequences by simply attaching themselves to each other. It may not be unreasonable to infer that this had occurred given that the formation of complementary sequences is equally probable.

These sequences have many interesting mathematical properties. For instance, an infinite sequence of this type replicates itself under the substitution rule C --> CG and G --> GC (i.e., replace each occurrence of a symbol by what is on the right-hand side of the rule). A mathematical interpretation is that this sequence counts the number of G's or C's mod GG (or mod CC in the complementary sequence) occurring in another sequence, namely the binary representation of Integers (Mauduit, 1992; Cobham, 1972; Salomaa, 1985). Although the sequence is aperiodic, its spectrum is highly periodic due to correlations induced by the self-similarity property. In fact, it is a well known sequence discovered by A. Thue in Number Theory and by M. Morse (Morse and Hedlund, 1987) in Symbolic Dynamics and it is now referred as the Thue-Morse sequence. In the theory of automata such a sequence is generated by a particular re-write system in the class of deterministic, context-free and parallel, referred as DOL system (Salomaa, 1985). L-systems were discovered in the cell-lineage problem initially and later computer scientists and linguists studied them formally.

For the RNA if one take for granted its self-cleaving/splicing capability, then it is possible to generate many distinct types of finite blocks of nucleotides (symbols) by using rewrite rules compatible with the chemistry. We construct an example in Section 4 to illustrate this procedure. Furthermore, in order to show the equivalence between chemical processes and rewrite systems we show how can we generate the sequences of nucleotides described above using only two rewrite rules and a single axiom as the starting symbol. To simplify our discussion we shall continue to consider only two symbols C and G instead of all four. A generalization to four or more symbols is trivial.

The above sequence generating process based on Crick-Watson base pairing, self-cleaving/splicing, and concatenating can now be shown to be equivalent to the following rewrite (or substitution) system. Consider the following rewrite rules:

1) C --> CG, and
2) G --> GC.

One may chose the starting symbol to be either C or G. To obtain the desired sequence, one must apply both rules in parallel at each step of the iteration process. For instance, begin with C as the starting symbol. Substitute CG for each occurrence of C. The first iteration gives CG, since only one rule is applicable. Repeat this process, i.e., apply now both rules (1) and (2) in parallel to each occurrence of C and G. The new sequence is CGGC. Remember that when rules are not applied in parallel (i.e., one rule at each iteration instead of both) the result will be CGG or CGC, but not CGGC. When this is the case, we shall refer to this as the sequential application of rules. Similarly, one can obtain CGGCGCCG, and longer and longer sequences. This specific rewrite system belongs to the class of DOL rewrite systems mentioned above.

It is equally possible that in nature some sequences have been synthesized which are mathematically equivalent to those obtained by sequential application of rewrite rules. In fact, they are equally probable. For instance:

G --> GC --> GCG --> GCCG
C --> CG --> CCG --> CCGC --> CCGCC
G --> GC --> GCG --> GCGG etc.

In fact, laboratory experiments which we mentioned in Section 2 clearly show the in vitro synthesis of CCGCC, GGCGG, GCCG, GGC etc.

This type of rewrite system can give rise to a large variety of nucleotide sequences. However, they are not necessarily fixed points of the substitution maps (rewrite rules) and therefore not self-similar. Thus we have two types of sequences: those which are fixed point (usually of infinite length) of the substitution map and those which are not. Sequences recognizable by finite automata are referred as *automatic sequences*. Another useful classification could be in terms of automatic and non-automatic sequences. Only some automatic sequences are also self-similar.

Information Processing Mechanisms There are many known examples in Genetics which support the idea that there are nucleotide sequences whose sole purpose is to

operate on other parts of the genome in order to achieve some specific goals such as cleaving, splicing, error detection and so on. Typical examples of such mechanisms are jumping of sequences, recoding and reading inter-cellular signals.

The genetic code dictates how nucleic acid sequence is translated into amino acid sequence. In some (a minority) mRNAs there is another set of instructions contained in the mRNA sequence that specifies an alteration in how the genetic code is applied (Gesteland et. al., 1992). There are two ways to do this.

1) Instructions alter the linear mechanism of readout;

2) The "meaning" of the code words is altered.

This phenomenon is referred by the Gesteland et al., as *re-coding* phenomenon. This is an example of information processing mechanism that has nothing to do with any function of a life form. It is purely internal to genome. For example, the E. Coli RF2 gene mRNA *programs* some 30% of the ribosomes to change to +1 reading frame after codon number 25 in order to complete the synthesis of the active protein. The signal sequence in the mRNA which does this re-coding has two components: the frame shift site (codons 25 &26) and an upstream sequence termed a stimulator (that encourage frame shift event).

Transposans or transposable elements is another example of internal information processing mechanisms. Transposons can jump around in an organism's genome altering its gene expression is fairly well-known and understood (N.J. Dibbs, 1991). Instead of making a large number of copies at many parts of a genome, nature uses transposons whenever they are needed and where ever. Certainly, this is a better solution to information processing requirements.

Introns are segments of genome that is used during translation. Nevetheless, they are replicated like the 'information carrier' part. If nature has kept them during the evolution there must be some function for which introns are useful. It is known that introns sometime contain sequences that regulate transcription. For other genes, introns act after transcription to stabilize RNA.

It is widely known that there many types of intron. Especially, class I and class II are object of great interest recently. In class I models the pre-mRNA introns are considered self-splicing and were present in the ancestral gene. The class II models negate all of the above. Many theories consider viroids as escaped introns. However, according to Diener (1989) sequence similarities between introns and viroids are only coincidental. He shares the view that viroid and viroid-like RNAs are phylogenetically older than introns. Diener suggests that plant RNAs are more plausible candidates than introns as "living fossils" of a precellular RNA world. Circular RNAs are also considered relics of precellular evolution (c. 350 nucleotide long). It is known that introns frequently divide DNA into regions that encode functional domains or subdomains in protein. Exon shuffling of recognized domains are also known now. That is, the transportation of a function from an area to another area in the gene. This supports the thesis that intron sequences are information processing units without which exonic part would not be translated or expressed properly and it would be useless. Much evidence suggest (Darnell and Doolittle, 1986) an early existence both of introns and of shuffling of introns between genes. We agree with the view that introns seemed unlikely to have entered pre-existing genes, all coded in a contiguous fashion, in the absence of an already functioning splicing mechanism to ensure their accurate removal. Therefore, introns existed in the early RNA world and conclusively they are ancient as exmplified by their constant presence through very long times in evolution.

We are now in a position to set forth the above qualitative discussion into a mathematical hypothesis. Given the description of the nucleotide chain formation we propose the following mathematical model.

1) Many nucleotide sequences are generated by DOL or IL type rewrite systems. We call them automatic sequences.

Consequently, this gives rise to:

a) Self-replicating and replicating self-similar sequences such as Morse-Thue and Fibonacci sequences.

b) Fractal nature of biological growth.

c) A small (point) mutation in the rewrite rule, i.e., sequences that act as automata and does information processing and other such operations on other sequences, causes a catastrophic change in the sequences copied. However, a point mutation in the information carrying exonic part of the genome leads to normal slow evolutionary changes. A rewrite rule represents the mechanism of creating automatic sequences.

2) Precursor of introns or proto-introns, transposons and viroids are automatic sequences.

Because of their simplicity and that viroids do not express themselves we shall identify them with what we called automatic sequences. Their sole purpose is to do information processing if information sequences are available. Thus viroid are automatic sequences and represent mechanisms for information processing for internal control and guidance. Transposons are sequences with some important mechanism that is activated by moving to different sites where it might add some novelty to the genetic patrimony. Surely this is more efficient than creating the same sequence at many places.

3) All automatic sequences do information processing tasks within RNA and genome.

What can we predict? *We expect blocks of sequences of the kind shown in earlier sections are to be found repeatedly in all species.* They represent the living fossil of early fundamental sequences. They may occur in some viroids and small viruses. Therefore, the hypothesis can possibly be tested on:

1) Viroids and evolutionary remnants of fundamental sequences. This could include, for instance, homoboxes, ALU sequences and segments of introns.

2) Suppressor genes, and especially the p53 protein oncogene.

Going a step further and extending the above arguments and evidence we may suggest that present day exons carry information/data, while introns are processors or instructions (in the sense of programs). Their function is to read other sequences, detect errors and trigger error-correction mechanism and sometime act as catalytic RNA (e.g., in *Tetrahymena thermophila*).

A remark on the suppressor mechanism is not out of place in the present context. A sequence can be made inactive by simply attaching a complementary sequence to a crucial part or to all of it. Suppression would be then to simply attach a complementory sequence to an automatic sequence. A mutation will some how break this pairing and the suppression mechanism will not function. The example of p53 gene is an important case in point. It is known that point mutations *in p53 and other suppressor genes at some specific site are enough to fail the suppression mechanism.* We shall discuss this issue in Section 5.

Mutations cause other changes as well. Since Morse-Thue sequence are perfect palindromes, early RNAs were also perfect palindromes. However, later mutations caused "BLOBS" to be formed. Blobs are those parts of hair-pin arrangements of a double strand where pairings between complementary bases break down due to mutations. In our model early RNA had no intronic or exonic parts - everything was both genotype and phenotype since the genetic code was not yet developed. Initially there was no distinction between biological functions and internal mechanisms. Later, when some nonautomatic sequence joined together with automatic sequences, then perhaps a distinction between them begins to show up. Possibly the separation took place at the time when eukaryotes and prokaryotes diverged. The existence of internal information processing sequence or automatic sequences played a key role in the evolution and development of complex life forms.

4. Sequences and Automata.

4.1. *Process Operators* We shall now show how to generate in a simple manner some "natural" binary sequences, which are known to be self-similar with respect to certain transformations. As mentioned earlier, such sequences are also known as automatic sequences because they can be generated and recognized by automata. Notice that the generation of binary sequences described in the sequel is similar to that we explained in section 1 purely on the basis of chemical properties. In some recent

mathematical works, Fibonacci, Morse-Thue and other DOL sequences are referred as "paper folding" sequences (Dekking, 1982b).

Let $\Sigma = \{ 0, 1 \}$ be the alphabet consisting to only two letters, 0 and 1. Consider them to be "complementary" (analogous to Crick-Watson complementary pair) to each other in the sense that 0 --> 1 and 1 --> 0, i.e., zero can be replaced by 1 and 1 by zero everywhere to obtain its complementry copy without changing its basic structure.

Let the symbol $\binom{s}{\underline{s}}$ denotes a process that given a subsequence s and its complement $\underline{s}$ (all symbols replaced by their respective complements), concatenate s and $\underline{s}$, where $\underline{s}$ is always to the right of s. Call it right-concatenation (as in section 1). Similarly $\binom{s}{\underline{s}}$ denotes the same process as above that $\underline{s}$ is concatented to the left of s. Call it left concatenation. The symbol $\binom{s^k}{s^{k+1}}$ denotes a process that concatenates k-th iterate to the right of (k + 1)-th iterate. This notation and its significance in the present context will become clear by the examples given in the sequel.

4.2. *The Morse-Thue Sequence* Start now with 0.

$\binom{0}{1}$ right-concatenation gives 01 (iteration 1)

Take the new sequence 01, and find its complement and right-concatenate to it. Now,

$\binom{0\ 1}{1\ 0}$ gives 0110 (iteration 2)

The new sequence is now 0110. Repeating once again the same operation

$\binom{0\ 1\ 1\ 0}{1\ 0\ 0\ 1}$ gives 01101001.

Thus using any k-th application of $\binom{s}{\underline{s}}$ for k = 0, 1, 2, 3, ..., we can generate a series of ever larger sequence, each time it doubling in its size.

k = 0: 0
k = 1: 01
k = 2: 0110
k = 3: 01101001
k = 4: 0110100110010110

and so on.

One may choose the starting symbol 1 instead of 0. The following set of sequences are obtained.

k = 0: 1
k = 1: 10
k = 2: 1001
k = 3: 10010110
k = 4: 1001011001101001

and so on.

The two set of sequences are complementary in the sense that one (of the same length) can be obtained from the other by applying the map 0 --> 1 and 1 --> 0. Furthermore, the two sets of sequences can be generated by using the following substitution or rewrite rules:

0 --> 01
1 --> 10

Starting from 0 or 1, and inserting 01 and 10, respecively for each occurrence of 0 and 1, in parallel, one can obtained the above two sets of sequences. This kind of substitution is called a fixed length substitution since for each letter, a word of fixed length is substituted. The parallel application of rules means that appropriate rewrite rule should be applied to all symbols at any iteration. Applying these rule ad infinitum one obtains the infinite word of Morse-Thue.

More formally, let $\Sigma = \{ 0, 1 \}$ be the alphabet and the mapping $R = \{R_1, R_2\}$, where R_1: 0 --> 01 and R_2: 1 --> 10. Denote $R^k(0)$ and $R^k(1)$ as the k-th application of R_1 and R_2, in parallel, to 0 and 1, respectively. Denote now the sequences:

$u_1 = R^\infty(0) = 011010011001...$
$u_2 = R^\infty(1) = 100101100110...$

for k --> ∞ .

Sequences u_1 and u_2 are called the fixed-point of the map R, because they reproduce themselves under the map R. In other words, they remained invariant. The infinite Morse-Thue sequences are also self-similar since their structure remains the same after each application of the map R despite their doubling in size. They have blocks of palindromes. In fact, any Morse-Thue finite word is a palindrome for k even.

4.3. *The Fibonacci Sequence* Fibonacci or so called rabbit sequences are known to occur in many domains of biology. For instance, some flowers are known to have petals equal to Fibonacci numbers. Mathematically, Fibonacci numbers are defined by the recursion relation

$$F_n = F_{n-1} + F_{n-2}$$

The first few numbers are 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, ... The ratio of successive number is the so called Golden ratio $g = 1 + \sqrt{5}/2$. Fibonacci sequences have some remarkable information processing and optimization properties. For example, they are used for optimal linear search and Fibonacci heaps (or priority queues) support arbitrary deletion from an n-item heap in O(log n) amortized time and all other standard operations in O(1) amortized time (Fredman and Tarjan, 1987). Fibonacci lattices are also known to occur in quasi-crystals also. Thus Fibonacci numbers and their binary representation are well studied in number theory literature. However, we shall point out

only a few interesting properties that are of relevance to RNA world nucleotide sequences.

How can we generate a Fibonacci sequence either using our process operator or rewrite rules? Consider at first the process operator $\binom{s^k}{s^{k+1}}$. Starting symbol can be 0 or 1, as it was in the above case of Morse-Thue sequence.

Starting symbol: 0

Take the complement of 0 and concatenate 0 to the right side of it, and apply now the operator $\binom{s^k}{s^{k+1}}$

$\binom{0}{1}$ gives now: 10

Take the new sequence 10 and concatenate it with the original one (the complement of 0 in this case), i.e., apply again the operation symbol $\binom{s^k}{s^{k+1}}$.

This gives: 101.

Repeating this process, one obtains the following set of sequences;

k = 0: 0	That is:	$\binom{0}{1}$	gives 10
k = 1: 10		10	gives 101
k = 2: 101		101	gives 10110
k = 3: 10110		10110	gives 10110101
k = 4: 10110101		etc.	

and so on. Its complementary sequence can be obtained, either by replacing 0s by 1s and 1s by 0s, or by starting from 1 and applying iteratively the process operator $\binom{s^k}{s^{k+1}}$.

These set of sequences can also be generated by applying the rewrite rules: 1 --> 10 and 0 --> 1, in parallel, starting either from 0 or 1. Both of these sequences generated by $\binom{s^k}{s^{k+1}}$ are the famous Fibonacci sequences. Fibonacci sequences are also self-similar, automatic and have interesting properties like Morse-Thue sequences. The operator $\binom{s^k}{s^{k+1}}$ represents what we termed the "Incremental Splicing" as explained

above. It is simple to verify that this is equivalent to iterative parallel application of rules C --> CG and G --> C. The third iterant and thereafter, each block of Fibonacci sequence has an interesting palindrome structure.

Theorem (deLucca, 1981):

For all $k \geq 3$, one has F = P d, where P is palindrome and d = CG if k even and d = GC if k is odd.

4.4. *Self-Replicationg Blocks* A historically well-known example is due to E. Post, the famous (00, 1101) tag machine problem where a sequence 011011101110100 is reproduced in six steps (for details, see Minksy (1967)). The following example will illustrate how certain fixed block sequences can be generated through repeated application of rewrite rules and cleaving.

Let Σ = { C, G, A, U } be the alphabet. Consider the following rewrite rule:

C --> CG
G --> AG
A --> AU
U --> CG.

Let the starting symbol be C. Then, the following growing chain of symbols will arise:

C --> CG --> CGAG --> CGAGAUAG --> CGAGAUAG.AUCGAUAG -->

An inspection reveals that we have a finite number of blocks, in this case 3 blocks, which continue to be reproduced ad infitum. The two blocks of the above sequence:

1) CGAGAUAG and 2) AUCGAUAG

will replicate in the following manner. Applying the rewrite rules to AUCGAUAG will give:

AUCGCGAG.AUCGAUAG,

where AUCGCGAG is a third distinct block different from (1) and (2). Now apply the rules to AUCGCGAG, obtaining the sequence AUCGCGAG.CGAGAUAG, i.e., the third and the first blocks. If we continue to apply rewrite rules, we shall go generating various combinations of the following three blocks:

(1) CGAGAUAG (2) AUCGAUAG (3) AUCGCGAG

of eight symbols each.

Consider another set of rewrite rule which is known to generate Heighway Dragon Curve (Dekking, 1982). The rules are:

A --> AU
U --> GU
G --> GC
C --> AC

and the starting symbol is A. We shall be able to generate the following four distinct blocks:

(1) AUGUGCGU (2) GCACGCGU (3) GCACAUAC (4) AUGUAUAC

The above two examples suggests that a fixed length (in the above cases, of length 2) rewrite rules generate *structural periodicity* in the form of blocks which are repeated as the chain grows. In nature, such structural periodicities are know to occur (cf. section 2). It can be proved that the process is independent of starting symbol and the same set of blocks are generated again and again. However, by changing the order of letters in Σ and rewrite rules in various combinations, we can find many other blocks.

4.5 *Other Examples of Automatic Sequences*

1. Milnor-Thurston Sequence:

There are other interesting sequences which have structure similar to Morse-Thue sequence. For instance, in Symbolic Dynamics, a sequence known as Milnor-Thurston (Milnor, 1988; Bai-lin, 1991) kneeding sequence has some interesting properties. It is of the form:

CGGC.GC.GCCG ...

Again, the dots in between the symbols are placed only for reading convenience. If we consider simply the above block of 10 symbols (nucleotides), it is clearly a variant of Morse-Thue sequence, or more precisely, its first ten symbols. The only difference is that the central subblock is GC, which is absent in the Morse-Thue case. Around the subblock GC on both sides are two 4-symbols subblocks, each complement of the other. We may fold it to get a hair-pin configuration as:

```
   C G G C
G
C
   G C C G
```

The Milnor-Thurston block can be considered as a mutation (insertion of two nucleotides GC in the central position) of a Morse-Thue block.

2. Baum-Sweet Sequence (Mauduit, 1992):

This sequence can be generated by the following four symbols and their respective rewrite rules.

Σ = { C, G, A, U } and C --> CG, G --> AG, A --> GU, U --> UU. If the starting symbol is C, then

C --> CG --> CGAG --> CGAGGUAG --> CGAGGUAGAGUUGUAG ... is a Baum-Sweet sequence.

3. Rudin-Shapiro Sequence (Dekking, 1982b):

Like Morse-Thue sequences, this sequence counts the number of 11 mod 2, in a binary representation of positive Integers (Morse-Thue sequence counts 1 mod 2). Thus the corresponding automata can be used for error detection in another sequence. Let Σ = {C, G, A, U}, then a Rudin-Shapiro sequence can be generated by rewrite rules:

C --> CA, G --> UA, A --> CG, U --> UG.

For a starting symbol C, we have:

C --> CA --> CACG --> CACGCAUA --> CACGCAUACACGUGCG ...

It can also be generated by only two symbols, i.e., Σ = { C, G } and rewrite rules:

CC --> CCCG, CG --> CCGC, GC --> GGCG, GG --> GGGC

If the starting symbols are CC, then

CC --> CCCG --> CCCGCCGC --> CCCGCCGCCCCGGGCG ...

The corresponding automata counts GG mod 2 in the CG representation of numbers.

4. Fredholm Sequence:

This sequences is generated by three symbols C, G, U, or C, G, A, or other combinations of C, G, U, and A, by rewrite rules:

C --> CG, G --> GU, U --> UU (or other similar rules). For example,

CGGCGCCCGCCCCCCCGCCC...

which represents the powers of 2 in three symbol representation, i.e., it represents the sequence 1 2 4 8 16 32 ...

4.6. *Properties of Automatic Sequences* We now return to an exact relationship between sequences and automata. The notion of rewrite systems, substitution systems, computing endomorphism, recurrent sets, symbolic dynamics, and fractal grammar are all intimately related to each other. Rewrite rules are also referred as computing endomorphisms or substitution rules in dynamical system theory. Fractals are the "phase space" of the computation and a fractal curve is a computational trajectory of the respective rewrite system (for a detailed discussion, see for example Shah, 1993). Their relationship with the theory of automata was established by Cobham (Cobham, 1972). The theorem of Cobham states that a sequence generated by p-substitution and a p-automata are equivalent. A further generalization is by Christol et al. which states:

Theorem (Christol et al, 1980):

Let Σ be a finite and non-empty alphabet.

Let $S = (s^k)$, an element of Σ^N and let p be a prime number. The following conditions are equivalent:

1) There exist $q = p^n$ such that S be an image of an algebraic element of $F_q((X))$;

2) S is the image of a sequence generated by a p-automaton.

3) S is the image of a fixed point of a p-substitution.

The equivalence of (2) and (3) is given by Cobham theorem. In other words, a sequence is p-recognizable if and only if it is generated by a p-substitution. On the other hand, condition (1) lead us to linear recurring sequence on a finite field and the coding theory. Linear recurring sequences are simply feed-back shift registers modelling computing machines. We shall not enter into this discussion. Their relationship with dynamical systems and ergodic theory is well-known (Furstenberg, 1988). These sequences are expressible in terms of formal series algebraic over the field of rational functions. The characteristic functions of sequences recognized by p-automata are always number-theoretic Mahler functions (Loxton, 1988).

The notion of a Tag Machine, another name for substitution automaton, goes back to monogenic canonical system of E. Post (Minsky, 1967). A finite p-automaton

corresponds to a uniform tag machine which substitutes for each symbol a word of length p. Some of these systems are known to be as powerful as Turing machine and they are capable of computing general recurive functions (Theorem 14.6-1, Minsky, 1967).

The Morse-Thue sequence is generated by a 2-automata. It checks the parity of one symbol, i.e., can detect errors in a code. Thus Morse-Thue sequence is a set of states of a 2-automaton that performs the parity check on binary representation of integers. Early nucleotides, therefore, have this capability of detecting errors in other sequences which is a necessary condition for a faithful replication.

Computational Complexity:

Computational complexity is an active field of computer science that deals with the study of the inherent complexity of an algorithm in terms of computational resources (time, memory, number of CPUs etc.). The number of steps needed to form a given nucleotide sequence gives information about its temporal formation and growth rate. The length of the chain computing a word gives information about the computational or generative time (Berstel and Brlek, 1987). This notion was introduced by A. A. Diwan to define the computational complexity of a sequence. A word of length n over a q-letter alphabet can be computed in $n / \log_q n$ (base q) steps. For the base q = 2, for $\Sigma = \{C, G\}$, the number of steps is $n/ \log_2 n$ (base 2). However, this computational complexity reduces for overlap-free words to which Morse-Thue sequence belongs. In a overlap-free case, each word can be computed in logarithmic length - the chain length of overlap-free words is proportional to $\log |w|$. The length of optimal chain for the Fibonacci word w_k is shown by Diwan (1991) to equal to k. The following theorem of Bousquet-Melou state that there are only 19 minimal chains for any k starting from k = 3.

Theorem (Bousquet-Mélou, 1992):

Let M(k) be the number of minimal chains computing the k-th application of the rewrite rules of Morse-Thue sequence, then:

$M(1) = 1$

$M(2) = 5$

$M(k) = 19$ for $k \geq 3$.

The proof is by construction.

These results inspire us to think that nature tend to choose those words (sequences) which are optimal with respect some important parameter (e.g., energy) and use minimal resources. We are tempted to propose, therefore, that in early RNA world those chains were preferred where the number of computational steps were minimal. The minimal complexity sequences most likely dominated the RNA world and probably they do today as well.

5. Mutations and Oncogenetics.

5.1. *Mutation of A Sequence*:

Consider a block (a sub-word) of Morse-Thue sequence

CGGCGCCG

Suppose there is a mutation in the 6th place (from the left) of the form, a deletion and an insertion. One C is deleted and a G is inserted in its place. The muted sequence is:

CGGCGGCG

Apply now the rewrite rule C --> CG and G -->GC, equivalent to chemical process that leads to a right-concatenation of the complementary sequence. One gets:

CGGCGCCGGCGCCGGC, whereas the wild-type would be
CGGCGCCGGCCGCGGC

Enlarging again the sub-block CG (of the wildtype) and GC of the mutant to

CGGC and GCCG, CGGCGCCG and GCCGCGGC etc., notice that the mutant part will continue to give larger and larger complementary sequence. Eventually, the

sequence will converge to a sequence containing blocks of GCCG and CGGC in a nested form, palindrome in a palindrome. Thus the mutation of a automatic sequence of Morse-Thue type will not have drastic effects on its block structure. However, this is not the case with a point mutation of a rewrite rule.

5.2. *Mutation of A Rewrite Rule*:

Consider, for instance:

C --> CG & G --> GC is changed to C -->CC & G --> GC

after a point mutation. This means almost a new set of rewrite rules. Now start from C and G, respectively. The following two radically different sequences are obtained:

C -->CC --> CCCC --> CCCCCCCC --> ...
G -->GC --> GCCC --> GCCCCCCC --> ...

If the mutation occurs after a certain number of replications, e.g., given CGGCGCCG and the muted version of rewrite rules, one obtains:

CCGCGCCCGCCCCCGC.

After a certain number of replications the sequence will be dominated only by Cs and an occasional G appearing. Whether we start from a single nucleotide or from a block of, Morse-Thue sequence, there is a radical change in its structure. However, if the mutation is corrected after a certain number of iteractions, the sequence will eventually return to a combination of GCCG and CGGC blocks.

5.3. *Convergence of Random Sequences*:

Since chemical Crick-Watson base pairing plus splicing is shown to be equivalent to the following rewrite rules (Section 2):

C --> CG and G --> GC,

any random sequence, after a certain number of application of these rules, will converge eventually to a sequence type that contains combinations of two fundamental complementary sequences, GCCG and CGGC. These blocks are "seeds" of fractal generating sequences.

5.4. *Oncogenetics*:

The crucial event in the development of a cancer can be traced back to alterations of the founder cell. The genetic (and not some epigenetic) origin of cancer is now established (Weinberg, 1991). Cancer genes have been discovered in the chromosomes of tumor cells - called oncogenes. It is these genes that become activated in the conversion of normal founder cell into a cancer cell.

Unconstrained growth is the most obvious trait of cancer cells. Also, they exhibit a shape that is very different from that of their normal counterparts. They fail to respect the territorial rules that confine normal cells to particular tissues. They rely to an unusual extent on aerobic mechanism: energy converting processes that do not depend on oxygen. The outer membrane of cancer cells are also different from the normal cells. Consequently, cancer cells are radically different from normal cells in all their aspects.

We have shown above that even a single point mutation of a rewrite rule, which in essence represent an automaton, leads to drastic change in the corresponding sequence generated. A simple analogy will clarify what it means. The situation is similar to when an information processing machine is slightly damaged. It may function but gives out its output in an abnormal manner and radically changed or it may not function at all. On the other hand, if the data is slightly different, then its output will be only slightly different. Therefore, mutations of automatic sequences have serious consequences as compared to mutations of information carrying sequence, e.g., exons, which do not process information.

We hypothesized earlier (cf. Sec.4) that mutation of rewrite rules, that is, that part of a sequence which represents the rewrite rule, if altered, can cause drastic output changes. We proposed in Section 4 that *introns, viroids and transposons are automatic sequences*. Therefore, in the context of oncogenes we have the following:

Mutations of introns in a genome are responsible for an erratic behavior, for instance, during cell division process.

The general understanding among geneticists (Weinberg, 1991) is that: A single, centrally acting cellular element is turned on and is then able to elicit a large number of changes in the phenotypes - all at the same time. The evidence thus points towards a simple mechanism. A case in experimental evidence is that of virus-induced tumors.

Another important factor that lend credibility to our hypothesis that if the viral genes were *lost* from the proliferating tumor cells or were *inactivated* experimentally, the cells were returned to normal states. In fact, it was found that in viral-induced tumors not only were the viral genes required to initiate the process of transformation; their continued presence and activity were necessary to maintain the tumor phenotype.

Our model suggests precisely this kind of behavior. The rewrite rules should remain muted at each application. Otherwise, if normal rewrite rules were applied after a certain number of application of mutated rules, the sequence generated will return eventually to their normal or wild type form. In other words, if the rewrite rules were corrected or muted rules were inactive, normal replication process would lead to (statistically) normal cells. The drastic change stops. This confirms what we showed mathematically that sequence deterioration is increamental. Now to be more precise, we rephrase our hypothesis on the mutation of rewrite rules as follows:

Proto-oncogenes represent the rewrite rule(s)/automata part of the gene.

To support the above statement we now describe more experimental facts. It is known that the information for being a tumor cell is transferred from one cell to another by DNA molecule - in a small single active segment. In our model this active and

small segment is an automatic sequence (representing an automaton) whose function could be, e.g., it reads other parts of gene.

If one models enzymatic mechanism(s) as rewrite rule(s) and RNA as the sequence generated by it, i.e., if automatic sequences represent not only internal mechanisms of reading, error-detection etc. proposed above, but also enzymes. Then, the mutation of rewrite rules does make a sense directly. However, there is no chemical or biological significance of the rewrite rules - not at least directly. Rewrite rules are simply a mathematical representation of chemical processes leading to formation of sequences. *Mutations of re-write rules can be interpreted as damage or change in the chemical processes or mechanism.* Mutation of mechanism can cause large error in the output, just like a defective machine can give radically different output as compared to the situation where only input data is muted (changed a little bit) but the machine processing that data is normal.

The oncogene is a slightly altered version of the normal gene which is called a proto-oncogene. The experimental evidence suggests that the genes ancestral to human proto-oncogenes must already have evolved when a common ancestor of human beings and flies lived in the Precambrian. Proto-oncogenes would not have been kept almost unchanged over such a long time unless they were and continued to be indispensable - something perhaps related with the control of cellular proliferation. This again *confirms* our hypothesis that proto-onogenes are part of some essential information processing and control mechanism.

We now give an example (Weinberg, 1983; Brandt-Rauf, 1992) of a point mutation that converts a benign proto-oncogene into an active oncogene. The critical 350 nucleotide long segment in the oncogene and the one in the corresponding proto-oncogene differ in only one base: a guanine in the proto-oncogene is replaced in oncogene by a thymine. Thus a simple one-point mutation G --> T, in the 5000-nucleotide normal human gene could convert it into an oncogene and make it cancer producing! Although, in this case the sequence is not an intronic part since the mutation occurs at GGC --> GTC, i.e., converting a Glycine to Valine. Apparently that single

substitution causes the protein to assume novel function enabling it to profoundly alter cellular metabolism. However, according to our model, it is something more than a change in the protein function. Some part of information processing or inter-cellular communication message is effected by this change. It has possibly something to do with the tumor suppressor mechanism.

p53 is considered to be a tumor suppressor gene. (Oren, 1992) A large body of evidence suggests that p53 gene may well be the most frequent target for genetic alterations in human cancer. These alterations range from complete deletion of the genes to a variety of different point mutations. As a consequence of these alterations there is the cancellation of the tumor suppressor activity of the normal p53. In many cases this is acheived through point mutations in p53 leading to pronounced conformational changes (Brandt-Rauf et al., 1992). There are now many indications that p53 may play a central role in the control of cell proliferation, cell survival and differentiation (see, e.g., Oren, 1992). Nevertheless, despite its role in such crucial processes, mice can still develop apparently without any defect in the total absence of p53 (Donehower et al., 1992]. This raises the possibility that p53 may become critical only when normal growth control is lost!

Now, we will consider something that supports our hypothesis. It is known that in most cases, the mutations involve various domains of the protein that have been *highly conserved through evolution.* This is consistent with the idea that such mutations interfere with a *basic feature* of the protein that is essential for its proper biochemical functioning. What are, therefore, the functions of p53?

A number of p53 - mediated activities have been identified experimentally. The most frequent observation was that wild type p53 can *induce a growth arrest*, which occurs primarily in the G1 phase of the cell cycle. There are two additional biological processes that may account for its tumor suppressor properties.

1) Cell deaths/elimination: the loss of p53 function may allow cells to survive illegitimately.

2) Induction of differentiation.

These findings support the notion that, at least in certain cell types, the loss of p53 may contribute to tumor progression by arresting the cells in a relatively immature, continuously self-renewing state. Wild type p53 may also be involved in the control of cell senescence, its loss thereby promoting the establishment of continuously proliferating tumors. Among its biochemical functions, studies using the replication of DNA viruses as a model system confirm the role of wt p53 in the control of DNA replication.

Some researchers speculate (Oren, 1992) that p53 binds to specific sequence elements that control the initiation of cellular DNA replications and directly represses initiation. Another possibility is that it is a transcriptional regulator. The ability of a protein to act as a promotor-specific transcriptional activator usually requires the selective binding of this protein to defined DNA elements. Indeed wt p53 was found capable of sequence-specific binding, while p53 mutants fail to do so (Kern et. al., 1991; El-Deiry et. al., 1992).

What is p53 Essential For? Cells could survive without p53, but there are many indications that p53 plays central role in the control of cell cycle progression, and perhaps other key processes such as differentiation and programmed cell death. The wild type p53 may some how be involved in sensing the DNA damage and imposing G1 growth arrest, i.e., p53 reads damages. In other words it is a proof reader!

It is likely that so long as the DNA is intact the presence of p53 may not be of great importance to the cell. However, once DNA damage is occurred, normal p53 function will be required for this damage to be eliminated. *This is in conformity with our model.* p53 is involved in information processing, i.e., proof reading and possibly activating repair mechanism in order to avoid an incorrect replication. The mutation of internal mechanisms will cause incrementally abnormal trait of cell. First, a few replications of a wild type cell will be very close to normal. The distinctiveness of cells will increase with newer and newer generations, until they are completely different from the original.

6. Conclusion and Speculations. We have shown that certain mathematical structures such as automatic sequences are a natural outcome of prebiotic Chemistry. Many properties known to mathematicians have their counterpart in genetics. This encouraged us to go further and hypothesize that viroids, proto-introns, internal guidence sequences, and transposons are all automatic sequences involved in information processing. There seems to be no other reason why nature has kept them so long. More importantly, we have shown that even a single point mutation of these sequences is the cause of a drastically abnormal behavior during the cell division process. A defect in the mechanism (i.e., automaton) has severe consequences as compared to a change in the input data. This is the main message.

Furthermore, we have shown that automatic sequences, computing machines and the finite field (of numbers) are inter-dependent. This lead us to speculate on coding aspects. We emphasized that it is natural and computationally less expensive to compensate for errors through the use of error correcting codes than to device a 100% reliable information transmission and replication system. It is simply impossible to avoid noise in any information system. Nature must have, therefore, used some method to detect and correct errors. A simple error-detection and correction technique is to make code as redundant as possible by adding extra symbols. In the present context there are many possiblties, e.g.,

1) The intronic part of the genome is a simple error-detection mechanism and is not a part of the code to produce proteins.

2) There is a built-in redundancy in the genetic code consisting of three nucleotides - only 20 amino acids are coded using 64 combinations. At least, in eight amino acids, it seems the actual code is made of only 2 nucleotides while the third ranges over all four. In other cases the third ranges over only two of them. From the theory of error-correcting codes over a finite field we know that an extra symbol can only be used for only a single error detection but no correction. However, putting together all three: redundancy in the genetic code, Morse-Thue type sequences to detect errors through

parity count, and one extra symbol, nature has deviced a fairly robust system of information transmission from one generation to another.

Acknowledgments The writing of this paper was made possible by the grant of a fellowship from the World Laboratory, Lausanne, Switzerland. The author gratefully acknowledges encouragement and kind support by Professor A. Zichichi, President of the World Laboratory.

Thanks are due to Professor Abdus Salam, the International Atomic Energy Agency and UNESCO for hospitality at the International Centre for Theoretical Physics, Trieste, and Professor J. Chela-Flores for a critical reading and many exciting discussions.

LITERATURE

Berstel, J. and Breck, S. 1987. On the length of work chains *Info. Processing Letters* **26**, 23-28.

Bousquet-Mélou, M. 1992. The number of minimal word cahins computing the Thue-Morse Word" *Information Processing Letters* **44**, 57-64.

Brandt-Rauf, P.W. et al. 1992. Conformational Effects of Selected Cancer-Related Amino Acid Substitutions in the p53 Protein *Jour. Biomolecular Structure and Dynamics* **10**, 253-264.

Cech, Thomas R. 1986. RNA as an Enzyme *Scientific American* **255**(6) 76-84.

Cech, Thomas T. 1987. The Chemistry of Self-Splicing RNA and RNA Enzymes *Science* **236**, 1532-1539.

Christol, G. et al. 1980. Suites Algebriques, Automates et Substitutions *Bull. Soc. Math.* France **108**, 401-419.

Cobham, A. 1972. Uniform Tag Sequences *Math. Systems Theory* 164-1991.

Darnell, J.E. and Doolittle, W.F. 1986. Speculation on the Early Course of Evolution *Proc. Natl. Acad. Sci.* (USA) **83**, 1271-1275.

deLuca, A. 1981. A Combinational property of the Fibonacci Words *Info. Processing Letters* **12**, 193-195.

Dekking, F.M. 1982a. Recurrent Sets *Advances in Math.* **44**, 78-104.

Dekking et al. 1982b. *Mathematical Intelligencer* **4**, 130-138, 173-181 and 190-195.

Dibb, N.J. 1991. Proto-splice Site Model of Intron Origin *J. Theor. Biol.* **151**, 405-416.

Diener, T.O. 1989. Circular RNAs: Relics of Precellular Evolution? *Pure Math. Acad. Sci.* (USA) **86**, 9370-9374.

Diener, T.O. 1991. Subviral Pathogens of Plants: Viroids and Viroidlike Satellite RNAs *FASEB J.* **5**, 2808-2813.

Diwan, A.A. 1991. Optimal Word Chains for the Fibonacci Words in *National Seminar on Theoretical Computer Science*, IMSC Report 115, Madras, India, edited by P.S. Thiagarajan, pp.26-33.

Donehower, L.A. et al. 1992. Mice Deficient for p53 are Developmentally Normal but Susceptible to Spontaneous Tumors *Nature* **356**, 215-221.

Eigen, M. and Winkler-Oswatitsch, R. 1992 Steps towards Life. Oxford Univ. Press.

El-Deiry et al. 1992. Definition of a Consensus Binding Site for p53 *Nature Genet.* **1**, 45-49.

Fredman M. and Tarjan, R.E. 1987. Fibonacci Heaps and their uses in Improved Network Optimization Algorithm *Jour. ACM* 34, 596-615.

Furstenberg, H. 1988. *Recurrence in Ergodic Theory and Combinational Number Theory* Princeton Univ. Press, Princeton (NJ), USA.

Gesteland, R.F. et al. 1992. Recoding: Reprogammed Genetic Decoding *Science* **257**, 1640-1641.

Hao, Bai-lin 1991. Symbolic Dynamics and Characteristic of Complexity *Physics* **D51**, 161-176.

Joyce, G.F. 1987. Nonenzymatic Template-Directed Synthesis of Informational Macromolecules. Cold Spring Harbor Symp. VOL.LII, 41-51.

Joyce, G.F. 1989. RNA Evolution and the Origins of Life *Nature* **338,** 217-224.

Jürgensen, H. and Lindenmayer, A. 1987. Inference algorithms for developmental systems with cell lineages *Bulletin of Math. Biology* **49**, 93-122.

Kern, S.E. et al. 1991. Identification of p53 as a sequence-specific DNA Binding Protein *Science* **252**, 1708-1711.

Lidl, R. and Niederreiter H.,1986. *Introduction to Finite Fields and Their Application* Cambridge Univ. Press, Cambridge (Chap. 6).

Lothaire M. 1983. Combinatories on Words Addison-Wesley Reading (MA).

Loxton J.H. 1988. Automata and Transcedence in *New Advances in Transcedence Theory* edited by A. Baker, Cambridge Univ. Press, pp.215-228.

Mauduit, C. 1992. Propriét´s arithmériques des substitutions in Seminaire de Théorie des Nombres, Paris, 1989-90,S. David, editor, Birhauser, Berlin.

Milnor J. and Thurston, W. 1988. Lecture notes in Math. No.1342, p.465.

Minsky, M.L. 1967. *Computation: Finite and Infinite Machines* Prentice-Hall (NJ) USA.

Morse, M. and Hedlund, G.A.,1987. *Symbolic Dynamics* Collected Papers of Marston Morse, World Scientific.

Oren, M. 1992. p53 the ultimate tumor suppressor gene? *FASEB J.* **6,** 3169-3176.

Orgel, L.E. 1987. Evolution of the Genetic Apparatus: A Review. Cold Spring Harbor Symp. Vol.LII, 9-16.

Orgel, Leslie E. 1992. Molecular Replication *Nature* **358**, 203-209.

Prusinkiewicz P. and Hanan, J. 1989. Lindenmayer Systems, Fractals, and Plants. Lecture Notes in Biomathematics No.79, Springer-Verlag, Berlin.

Salomaa, A. 1985. *Computation and Automata* Cambridge Univ. Press.

Scott, E.K. et al. 1992. Oncogenetic Forms of p53 Inhibit p53-Regulated Gene Expression *Science* **256**, 827-830.

Searles, D.B. 1992. The Linguistics of DNA *American Scientists* **80**, 579-591.

Shah, K.T. 1993. *Automata, Neural Networks and Parallel Machines* Chapter 6, World Scientific, Singapore (in press).

Watson, J.D. et al. 1987. *Molecular Biology of the Gene* Vol.II, The Benjamin/ Cummings Publ. Co., Menlo Park, CA, USA.

Weinberg, R.A. 1983. A Molecular Basis of Cancer *Scientific American* 249, No.5, 102-116.

Weinberg, R.A. 1991. Tumor Suppressor Genes *Science* 254 1138-1146.