# A Functional Model of Cell Genome

Alessandro Fontana[1] and Walter Steven Fraccaro[2]

[1]IEEE member, Milan (Italy)
[2]University of Milano-Bicocca, Milan (Italy)
alessandro.fontana@ieee.org

**Abstract**

This paper is concerned with a model of the cell genome called Artificial Genome, that tries to model some aspects of the cell cycle, in particular those related to gene expression, cell differentiation and cell growth. The functioning of the model during interphase and mitosis is explained in detail through an example, that shows how the four functional categories of the Artificial Genome (Functions, Code, Data and Buffer) interact to determine the phenotype. The capacity of the model of generating phenotypical patterns, represented as 2-dimensional shapes, is explored through a simulation, that evolves in 9 cycles a cell to become a small face made up of 132 cells. Finally some parlallels between the Artificial Genome and the natural one are discussed.

## Introduction: the Genome

The discovery of the structure and replication mechanism of the DNA, performed by J. Watson and F. Crick in 1953, is undoubtedly one of the most important achievements of the 20th century. Another major milestone, achieved in the 21st century thanks to the simultaneous efforts of the international Human Genome Consortium and of the private company Celera Genomics, has been the sequencing of the human genome. Other species's genomes (including Drosophila Melanogaster, C. Elegans, mouse, etc.) are also currently available.

Having at disposal the whole base sequence, the biologists are now faced with the challenge of interpreting the meaning and functioning of the sequence, that is understanding how the code is utilized to produce what is called the "phenotype".

Some known facts about the genome follow. The genome is made up of chromosomes (23 pairs in human), that are very long molecules of deoxyribonucleic acid (DNA), composed of basic elements called Nucleotides, whose distinctive parts are four sub-elements called Bases: Adenine, Cytosine, Guanine, Thymine. Triplets of these bases (codons) are *transcribed* into RNA and then *translated* into amino-acids, the building-blocks of proteins, which in the end determine the cell's (chemical) behaviour and function.

Some *unknown* facts about the genome follow. One of the biggest surprises coming from the sequencing of the DNA of man and other species is that the genome, which can be very large (3 GB in human), seems almost empty, meaning that the overwhelming majority of it is made up of non-coding sequences (like for example "TTAGGG", which is repeated several times), never transcribed into RNA and never translated into proteins. The coding part represents only 2% of the whole base sequence; the remaining 98% of the sequence, that apparently has no function, has been labeled "junk DNA". However, recent experiments have given evidence that "junk DNA" may be involved in regulating the activity of the coding part.

According to the replication mechanism proposed by Watson and Crick, when a cell duplicates each of the two daughter cells inherits one of the two single-stranded helices of the mother cell's DNA, on which the DNA-polymerase enzyme rebuilds the complementary chain, thus yielding a perfect double-stranded copy of the parental DNA: as a result all the cells of an organism have the same DNA. A major question that arises is then how can different cells (e.g. a nervous cell and a heart cell) of the organism behave differently, to perform different functions. This is achieved by selectively activate in each cell only a subset of genes by means of several biochemical mechanisms (e.g. the genes that don't have to be activated are "silenced" through the methylation of the corresponding genome segments, which inhibits their transcription). However, the question how a cell knows which parts of the DNA have to be silenced and which not, is still unanswered.

A strictly related question is how can the zygote, through a series of duplications, grow into a fully developed organism (embryogenesis). Also this process seems to imply a series of selective activations and de-activations of different groups of genes, whose control mechanism is still largely unknown.

It is the aim of this paper to present a functional model of the cell genome that, without going into biochemical details, tries to model some key aspects of cell dynamics, like gene expression, differentiation and morphogenesis. The remainder of the paper is organized as follows. In the following three sections the model of the genome is introduced and its functioning during interphase and mitosis is explained in detail through an example. In the subsequent two sections a simulation is shown that demonstrates the potential of the model and some

parallels between the artificial genome and the natural one are outlined. The final section draws the conclusions.

## Artificial Genome

The information content of the Artificial Genome is divided in three categories:

- Functions
- Code
- Data

The Functions hold information embedded in intrinsic mechanisms, not expressed explicitly (that is as a sequence of quaternary digits 0–1–2–3 called bases) and cannot be modified during the organism's life: they can be compared to a computer hardware.

The Code holds information that is expressed explicitly but again is not (normally) modified during the organism's life: it can be compared to the computer software that contains instructions (code). Modifications occurring to the Code are called *Mutations*.

The Data contain information that is expressed explicitly and is normally modified during the organism's life: they can be compared to the computer software that contains data.

There is also a fourth category, called Buffer, which contains information that is used temporarily during the cell cycle and then discarded. The Buffer does not need to be inherited or duplicated and therefore is not part of the Artificial Genome.

We note that, according to the said definitions, the Artificial Genome is not the same for all the cells of the organism, that will have the same Functions and Code but, in general, different Data (and Buffer).

The overall information content of the Artificial Genome is made up of the following units:

| Unit Name | Symbol | Category |
|---|---|---|
| Gene Activator | GAF | Funct. |
| Gene Controller | GCF | Funct. |
| Gene Transcriptor | GTF | Funct. |
| Cell Duplicator | CDF | Funct. |
| Gene Activation Code | GAC | Code |
| Gene Control Code | GCC | Code |
| Gene Transcription Code | GTC | Code |
| Duplication Plan Code | DPC | Code |
| Cell Type | CT | Data |
| Development Stage | DS | Data |
| Gene Number | GNB | Buffer |
| Exon Address | EXA | Buffer |
| RNA | RNA | Buffer |

As already said, the last three units are not strictly part of the Artificial Genome.

We will now analyze in detail all these units and we will explain how they interact to determine the cell cycle events, but first let us introduce the concept of Artificial Phenotype.

The Artificial Phenotype, that is the phenotype of the Artificial Genome, is represented as a drawing on a bi-dimensional grid, where each grid cell represents an artificial cell and the colors represent the result of the cell's gene expression. The picture below gives an example of Artificial Phenotype.

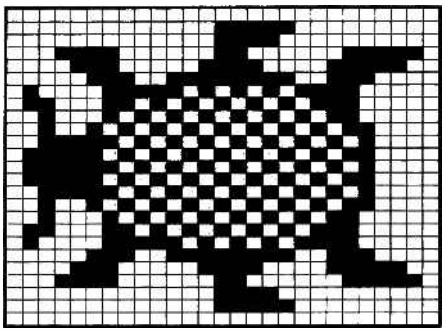

Figure 1: Example of Artificial Phenotype.

Let us now get back to the cell cycle, which is divided in two phases: interphase and mitosis.

## Cell Cycle: Interphase

During the interphase the genes are activated and expressed and the phenotype is modified accordingly. Let us assume that at the beginning of the interphase the organism has the following phenotype (two cells expressing no genes –color white–, numbers inside represent Cell Types):

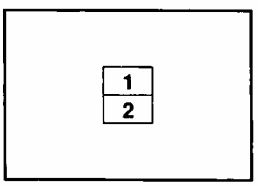


Figure 2: Phenotype at the beginning of interphase.

In the interphase, three functions are executed in sequence (Inputs > Outputs indicated in the parentheses)

| | | |
|---|---|---|
| Gene Activator | (GAC,CT,DS | > Gene #) |
| Gene Controller | (GCC,Gene # | > Exon Addr.) |
| Gene Transcriptor | (GTC,Exon Addr. | > RNA) |

The whole process is triggered by the couple of variables (CT,DS) and ends with a modification of the phenotype.

## Gene Activation

The function Gene Activator reads the couple (CT,DS) and GAC, and writes Gene Number. Let us have a closer look at these units.

The data-type number Development Stage (DS) is simply a counter that is incremented at each mitosis (with the exception of the "idle" case, explained later). It is assumed, for simplicity's sake, that all cell cycle events are synchronous and have no duration.

Beside the Development Stage, another data-type number is assigned to each cell, representing its "type": the Cell Type (CT). The same type can be assigned to more than one cell, so types can be interpreted as a set of categories into which the whole of the cells is divided, conceptually corresponding to tissues or organs. Cell Types can vary with time, allowing to model aspects related to the development.

As far as GAC is concerned, it is basically a matrix whose rows represent the different Cell Types of the organism (ranging from 0 –zygote– to n) and whose columns represent the Development Stages (ranging from 0 to m). In the following picture an example of GAC & DPC combined is reported.

For each (CT,DS) couple, four digits are given, two for GAC (grey background, used for interphase) and two for DPC (white background, used for mitosis). The GAC indicates the gene(s) activated for each (CT,DS) couple (only one, for simplicity).

| | DEVELOPMENT STAGES | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | | | | 1 | | | | 2 | | | | 3 | | | |
| | GAC | | DPC | | GAC | | DPC | | GAC | | DPC | | GAC | | DPC | |
| CELL TYPES 0 | 1 | 0 | 1 | 0 | | | | | | | | | | | | |
| 1 | | | | | 1 | 0 | 1 | 3 | | | | | | | | |
| 2 | | | | | 1 | 1 | 0 | 2 | | | | | | | | |
| 3 | | | | | | | | | 1 | 0 | 1 | 2 | | | | |
| 4 | | | | | | | | | 1 | 1 | 1 | 2 | | | | |
| 5 | | | | | | | | | 1 | 2 | 1 | 2 | | | | |
| 6 | | | | | | | | | 1 | 3 | 1 | 2 | | | | |
| 7 | | | | | | | | | | | | | 1 | 0 | 1 | 3 |
| 8 | | | | | | | | | | | | | 1 | 0 | 1 | 3 |
| 9 | | | | | | | | | | | | | 1 | 2 | 1 | 3 |
| 10 | | | | | | | | | | | | | 1 | 2 | 1 | 3 |
| 11 | | | | | | | | | | | | | 1 | 3 | 1 | 3 |
| 12 | | | | | | | | | | | | | 1 | 3 | 1 | 3 |
| 13 | | | | | | | | | | | | | 3 | 0 | 1 | 3 |
| 14 | | | | | | | | | | | | | 1 | 2 | 1 | 3 |
| 15 | | | | | | | | | | | | | | | | |
| | INTERPH. | | MITOSIS | | INTERPH. | | MITOSIS | | INTERPH. | | MITOSIS | | INTERPH. | | MITOSIS | |

Figure 3: GAC (Gene Activation Code) and DPC (Duplication Plan Code).

The function Gene Activator gets the couple of values (CT,DS) and reads the digits written in the GAC table at these coordinates. If we assume DS = CT = 1 the read digits are [1 0]. This array, written into Gene Number, indicates the gene to be activated.

## Gene Control

The function Gene Controller reads Gene Number and GCC and writes Exon Address. Let us have a look at these units. A picture of GCC is reported below.

| GENE # | | EXON ADDRESS #1 | | | | EXON ADDRESS #2 | | | | EXON ADDRESS #3 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 2 | 0 | 0 | 0 | 0 | 1 | 0 | 3 | 0 | 1 | 0 | 3 | 3 |
| 0 | 3 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 1 | 0 | 1 | 1 | 1 | 1 | 1 | 2 | 0 | 1 | 1 | 2 | 1 | 3 |
| 1 | 1 | 1 | 1 | 3 | 1 | 1 | 1 | 2 | 3 | 1 | 2 | 0 | 3 |
| 1 | 2 | 1 | 1 | 2 | 0 | 0 | 1 | 3 | 2 | 1 | 3 | 3 | 1 |
| 1 | 3 | 0 | 3 | 0 | 2 | 1 | 3 | 1 | 2 | 1 | 3 | 1 | 3 |
| 2 | 0 | 1 | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 1 | 3 | 3 | 3 |
| 2 | 1 | 0 | 0 | 0 | 3 | 0 | 0 | 3 | 3 | 0 | 1 | 3 | 3 |
| 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 2 | 2 |
| 2 | 3 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 2 | 1 | 0 | 0 | 0 |
| 3 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 2 | 2 | 1 | 0 | 1 | 0 |
| 3 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |

Figure 4: GCC (Gene Control Code).

The GCC contains the code necessary to sort out the exons, that is the portions of the coding sequence that will be transcribed. The table has 16 rows, corresponding to the total number of genes (in this example). Each row is divided into n blocks (three in our case, grey cell indicating block start), each of which represents the address of an exon. The array returned by the function Gene Activator ([1 0]) is looked up in the two left columns of the GCC table. The following digits of the matching row are then used to build up the exon addresses. For each block, the first digit (1) tells whether the exon is to be transcribed or not, that is whether the following three digits are to be read or discarded. In our case the addresses of the three exons are [1 1 1], [2 0 1], [2 1 3]. This array is written into Exon Address.

## Gene Transcription

The function Gene Transcriptor reads Exon Address and GTC and writes RNA, putting together the exons. A picture of GTC is reported below.

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 2 | 0 | 0 | 1 | 2 | 3 | 0 | 1 | 2 | 3 | 0 | 1 | 2 | 0 |
| 0 | 1 | 2 | 3 | 3 | 0 | 3 | 3 | 1 | 1 | 2 | 1 | 0 | 2 | 2 | 3 |
| 0 | 2 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 3 | 0 | 1 | 2 | 3 |
| 0 | 1 | 2 | 3 | 0 | 1 | 3 | 1 | 0 | 1 | 2 | 3 | 0 | 2 | 2 | 2 |

Figure 5: GTC (Gene Transcription Code).

GTC represents the "coding" portion of the genome, the one that is actually transcribed into RNA. The exon addresses returned by Gene Controller tell which cells

must be read in the GTC table to build-up the RNA sequence. The first digit of an exon address indicates the row of GTC, the other two digits the position in the row. In our example, the read sequence is [0 2 0]: this value is written into RNA. Finally, RNA is used to modify the phenotype, thus completing the expression of the gene. The mapping between the RNA values and the relevant phenotypical effects is reported in the Color Table below.

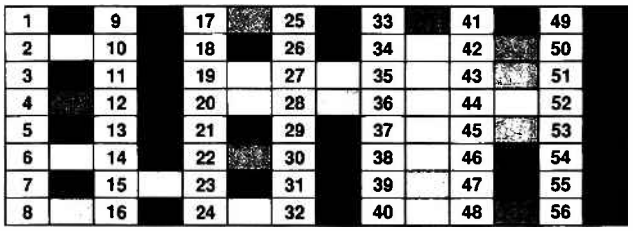


Figure 6: The Color Table maps RNA values to the corresponding phenotypical effects, represented as colors.

In our example, the corresponding base 10 number of the RNA [0 2 0] (16), is looked up in the Color Table and the relevant Cell Type (1) is colored with the corresponding color (dark grey). The phenotype becomes:

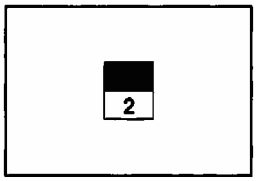


Figure 7: Phenotype at the end of interphase.

This procedure is carried out for all the cells. When all the cells have been processed, the interphase is over. The work-flow of the interphase is summarized in the following diagram.

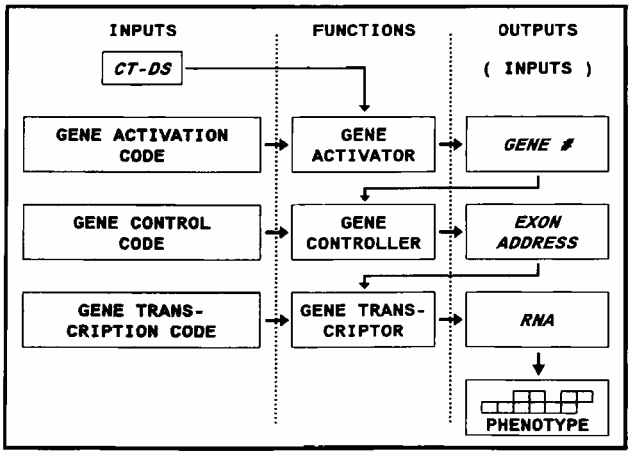


Figure 8: Interphase work-flow (Functions are indicated, Code is highlighted in grey, Data and Buffer in italic).

## Cell Cycle: Mitosis

During mitosis the cell duplicates, giving origin to two daughter cells; this operation is executed by the function Cell Duplicator, which reads the couple (CT,DS) and DPC and modifies the phenotype.

The Duplication Plan Code contains the scheme for the duplication of the whole organism, that tells each Cell Type in each Development Stage if and how it must duplicate. This information is contained in two digits: the mitosis type and the displacement type.

The mitotis type has four possible values: Grow, Differentiation, Aging and Idle. If the value is "Differentiation", the two daughter cells are assigned two new Cell Types, starting from the first "free" (not in use by any cell) Cell Type. If the value is "Grow" the two daughter cells inherit the Cell Type of the mother cell. In both cases DS is incremented by one for the two daughter cells. If the value is "Aging" the DS is incremented for the cell but no duplication occurs. If the value is "Idle" nothing happens. The displacement type has also four possible values: down, up, right, left. It indicates the position of the second daughter cell, while the first remains at the same position of the mother cell.

A summary of mitosis and displacement types and the relevant coding is reported in the following table.

| Mitosis Type | Coding | Displacement Type | Coding |
|---|---|---|---|
| Grow | 0 | 2. daught. cell down | 0 |
| Differentiation | 1 | 2. daught. cell up | 1 |
| Maintain (aging) | 2 | 2. daught. cell right | 2 |
| Idle cycle | 3 | 2. daught. cell left | 3 |

A picture of GAC & DPC combined is reported in Figure 3. DPC is made up by the digits associated to the mitosis period (white background), representing the mitosis type and the displacement type.

In our example, at the coordinates (CT = 1, DS = 1), we have [1 3]. The first digit (1) indicates that during this mitosis a differentiation is performed; the second digit (3) indicates that the displacement type is "2. daughter cell left". As a result, two new Cell Types, 3 and 4, are created. At the coordinates (CT = 2, DS = 1) we have [0 2], meaning that no new Cell Type is created and the second daughter cell is placed to the right. The final phenotype is the following (no genes are expressed during mitosis: cells are all white):

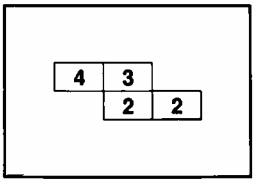


Figure 9: Phenotype at the end of mitosis.

## Simulation

A simulation has been carried-out, where the following code (here reported in a compacted form)

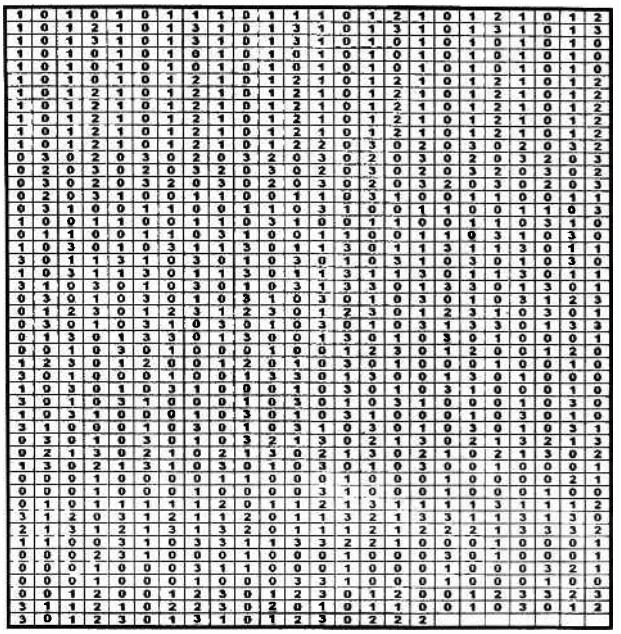

Figure 10: The simulation code. The 1. block represents GAC-DPC (grey), the 2. GCC (white), the 3. GTC (grey).

was used to develop the zygote to become in 9 cycles a small face made up of 132 cells.

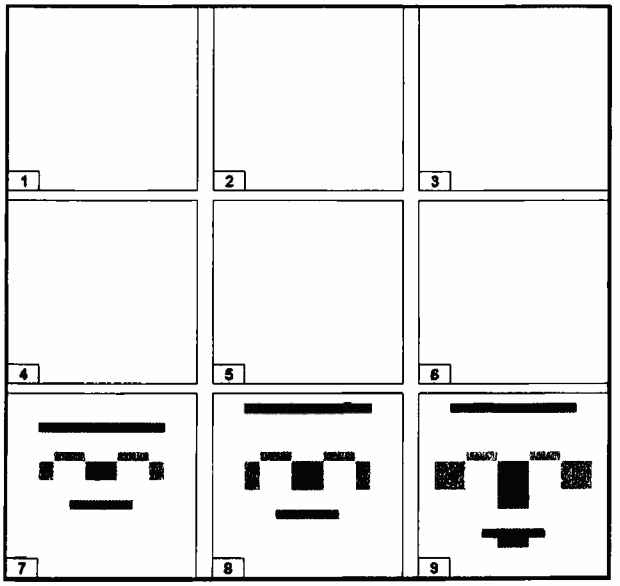


Figure 11: The simulation phenotype.

Some statistics relevant to the simulation:

- Total genome size: 1145 bases
- GAC and DPC %: 75%
- GCC %: 20%
- GTC ("coding" portion) %: 5%
- Number of cells: 132

In this simulation the vast majority of the sequence is represented by the Duplication Plan Code, the Gene Activation Code and the Gene Control Code, while the "coding" part, represented by the Gene Transcription Code, accounts for just 5% of the total.

## Discussion: What the Proposed Model Can Tell on the Nature of the Genome

Hereafter some observations on the proposed model.

In the proposed model the cascade of events that occur during interphase and mitosis is triggered by the values taken by CT and DS, that tell the cell who it is and what it has to do during the different phases of the cell cycle. The (CT,DS) couple represents therefore the primary source of information for both differentiation and morphogenesis.

The gene expression mechanism may look complicated (as the natural one, that makes us of exons, introns, three types of RNA, does), but it has the advantage of allowing a more flexible means of modulating the genes. Without the addressing mechanism provided by GCC, modifying the genes would necessarily require changing some digits in the coding sequence GTC. With such mechanism this can also be achieved by changing the exon addresses, thus modifying the combination of the picked exons. This gives more possibilities to a (either artificial or natural) genetic algorithm to explore and exploit new regions of the search space. Furthermore, the addressing mechanism offers an explanation for the phenomenon of gene overlapping, by which a DNA portion can participate to more genes.

The first part of the simulation (until DS no. 7) has been constructed using only the mitosis type "differentiation", generating a genotypical variety without a corresponding phenotypical variety (all the cells expressed the same gene - color light grey). The choice of this development scheme was motivated by the necessity to "mark" the different parts of the organism in order to be able, at a later stage of development, to selectively trigger the expression of different genes in those parts. We define the "cell diversity ratio" as the number of cell types divided by the total number of cells:

$$CDR = \text{number of cells types} / \text{number of cells}$$

This parameter can be taken as a measure of the "diversity" of the organism: if it equals unity, it means

that the organism's cells are all (genotypically) different; if it is less than 1, it means that more cells share the same cell type. CDR is equal to 1 in the first DS's of the simulation, then progressively decreases as more type "grow" mitoses are been used. The similarity of the face at DS = 7 with a child's face and the face at DS = 9 with a more adult face could be a clue that something similar occurs also in nature.

The proposed framework can be used to model genetic diseases. An interesting artificial genetic disease is induced by mutations in the (CT,DS) couple, that make the cell "jump" onto another point of the GAC-DPC matrix, causing the re-activation of embryonal development mechanisms and the expression of the relevant genes. The presence in (some) colon-cancer cell clones of the Carcinoembryonic Antigen (CEA - normally produced by embryo cells) could be a clue that something similar occurs also in nature.

In the model the "non-coding" part of the genome is made up mostly by the GAC-DPC matrix. The dimension of such matrix can be increased either by increasing the number of cell types (a measure of the "complexity" of the organism) or by increasing the number of development stages (a measure of life duration). This could explain why certain plants have very large genomes, even larger than humans: they live longer.

The proposed coding scheme can of course be optimized through compression techniques, e.g. storing only the differences between one DS and the next one: this would allow the storage of more information with the same genome size.

The described model can be enriched by introducing a description of inter-cell dynamics, that is modeling the chemical signals that cells exchange to self-organize as a population (most of the models present in literature are actually concerned primarily on inter-cell dynamics).

We said that, when a mitosis of type "differentiation" is performed, the first "free" cell types (i.e. not used by any cell) are assigned to the daughter cells, which seems to imply a kind of central control that has no biological grounding. An alternative, biologically more plausible method consists in representing cell types with binary sequences; the cell types of the daughter cells are then generated by appending at the end of the mother cell's sequence a "0" for the first daughter cell and a "1" for the second. With such a method, the cell type holds the cell's "duplication history" starting from the zygote onwards.

Not necessarily the proposed model implies a violation of the dogma by which all cells have the same DNA, if the natural counterpart of the (CT,DS) couple, which holds the part of the Artificial Genome that varies between cells of the same organism, is realized through epigenetic mechanisms.

## Conclusion

A model of the genome has been proposed, that tries to reproduce some aspects of the cell cycle, with particular reference to gene expression, cell differentiation and morphogenesis. The model is a rather high-level one, nevertheless it seems to hold the potential to model real biological phenomena, even though there are for sure cell biology aspects it cannot explain: further efforts are needed to customize the model for a better adherence to these aspects. A final remark: our work follows the philosophy of the artificial life community, to try to reproduce life as it could be, with the hope it leads us to a better understanding of life as it is.

## References

De Garis, H. 1999. Artificial Embriology and Cellular Differentiation. Evolutionary Design by Computers: 281-295.

Eggenberger, P. 1997. Evolving Morphologies of Simulated 3d Organisms Based on Differential Gene Expression. In Proc. of the 4. European Conference on Artificial Life. Cambridge, MA: MIT Press.

Furusawa, C., and Kaneko, K. 2000. Complex Organization in Multicellularity as a Necessity in Evolution. Artificial Life 6(4): 265-281.

Gibbs, W. W. 2003. The Unseen Genome: Beyond DNA. Scientific American, 2003-12.

Hood, L., and Galas, D. 2003. The Digital Code of DNA. Nature. 2003 Jan 23; 421(6921): 444-448.

Kitano, H. 1995. A Simple Model of Neurogenesis and Cell Differentiation Based on Evolutionary Large-Scale Chaos. Artificial Life, 2(1):79-99.

Moore, M. J. 1996. When the Junk isn't Junk. Nature 379: 402-403.

Olovnikov, A. M. 1996. The Molecular Mechanism of Morphogenesis: a Theory of Location DNA. Biochemistry, Vol. 61.

Prodan, L., Tempesti, G., Mange, D., and Stauffer, A. 2002. Embryonics: Electronic Stem Cells. In Proc. of the 8. International Conference on Artificial Life, 101-105. Sidney, Australia: MIT Press.

Reil, T. 1999. Dynamics of Gene Expression in an Artificial Genome – Implications for Biological and Artificial Ontogeny. In Proc. of the 5. European Conference on Artificial Life, 457-466. New York, NY: Springer Verlag.