

BUDAPEST UNIVERSITY OF TECHNOLOGY AND ECONOMICS
FACULTY OF MECHANICAL ENGINEERING
GÉZA PATTANTYÚS-ÁBRAHÁM DOCTORAL SCHOOL

**ENGINEERING APPLICATION OF BIOLOGICAL MODELS;
DEVELOPMENT OF ENTROPY BASED ADAPTIVE
IMAGE COMPRESSION METHODS**

PhD Thesis Booklet

Sándor Piros
electrical engineer

Supervisor: Péter Korondi PhD, DSc.

Mechatronics, Optics and Engineering Informatics Department

Budapest, 2014

1 Introduction

The main feature of interdisciplinary researches is, that the area under investigation is examined by the methods of a variety of fields together – for example biology, mathematics, engineering and information technology – and at the end each discipline yields discipline specific results.

Long before the beginning of my PhD studies I have started to study different areas of biology, to obtain a deeper knowledge mainly in the fields of molecular biology, cell biology and developmental biology. I try to comprehend biological processes by an engineering approach, to explain certain phenomena, on the other hand to learn ideas through close observation of nature, to be utilized in the field of engineering. Acquired new knowledge, particularly in the field of modeling and simulation of dynamic systems and experience gained in MATLAB programming was helpful in this. The examination of the mysteries of nature is a sort of "reverse engineering".

The target objective was, to propose hypotheses about the possible structure how the information is stored and compressed in the chromosomes, which refers to the three-dimensional structure of an organism. Human body consists of more than 10^{14} cells, they belong to one of the (about) 100-200 different cell types. Our genome is 3.3×10^9 base pairs long only (four different DNA bases exist in nature). In the course of my research work I dealt with: how this 'assembly' information (body plan) is programmed and how an organism develops through a series of divisions based on that?

To achieve this goal, I have studied biology from the existing literature, especially which were related to chromosome structure and topology. It is likely that this type of information is incorporated not only in the genome, i.e. the DNA sequence, but also as how the DNA chain is bound to the chromosome scaffold proteins, or how the DNA chain's methylation pattern changes in the course of cell division cycles.

In this particular case, we can expect not only to find an efficient method for compressing a 2, 3 or multidimensional data set, but it can be useful for the biology itself too, serving with explanations of certain phenomena. By creating an expressive model, we might be able to explain anomalies, like certain type of cancerous alterations during cell divisions.

Every multicellular organism develops from a single cell, the zygote. During cell division, a cell always produces two descendent daughter cells. There are various descriptions in literature review of biological experiments, which suggest that there is causal relationship between the replicated DNA strands and the types of the individual daughter cells. Based on that, it was

possible to introduce a cell-numbering system.

By studying the process of cell division and cell differentiation the conclusion came that this process is in fact looks like an inverse discrete wavelet transformation.

This transformation, derived from nature was worked out for two cases:

- for the case of logical variables,
- for the case of discrete variables.

We managed to develop compression methods for these 2, 3 or multi-dimensional transformed data sets, e.g. for images, 3-D motion pictures or for vector fields.

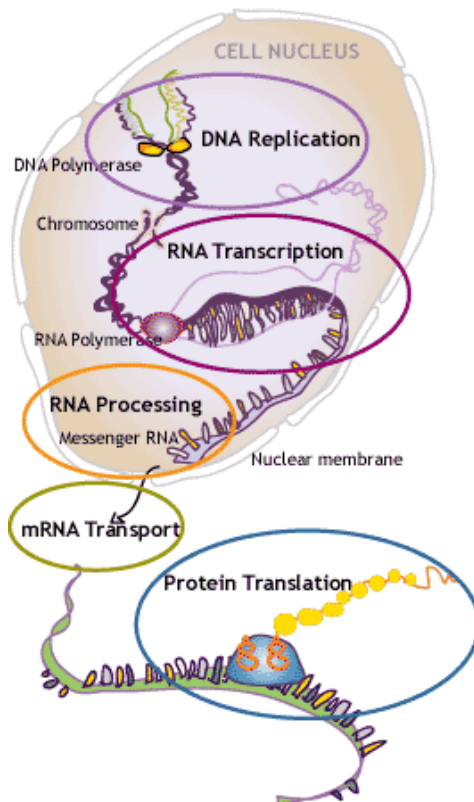


Figure 1 Central dogma of molecular biology:
"DNA makes RNA and RNA makes protein" (nobelprize.org)

2 Literature Background

In the field of biology we can find information systems at multiple levels. The organisms are grouped in various ways, for example like unicellular and multicellular. Multicellular organisms must possess a bi-level information system, or data set, first of all individual cell-type specific information about its cell components and, secondly, for the set-up information of all the cells and tissues that make up the whole organism.

Human body contains at least 10^{14} cells, all of them are originated from only one by consecutive cell divisions. After every cell division cycle the number of cells is doubled. The number of cells after the n^{th} cycle is 2^n and the number of all previous stages or the number of cell divisions is $2^n - 1$. The number of cycles to produce this quantity (10^{14}) of cells should be more than $\log_2(10^{14}) = 47$, or even more taking into account of apoptosis (programmed cell death) or other cell losses. What is the amount of information to be inherited if it were not compressed in the genome? The number of different cell types is in the range of 100-200. Similarly to the genetic code, DNA base triplets are coding for the possible 20 amino acids ($4^3 = 64$), on this analogy it needs base quadruplets (at least 4 nucleotide bases long words) to code for the different cell types ($4^4 = 256$). To describe the type of each and every cell of the body, this information alone requires $10^{14} \times 4 \text{ bases} = 4 \times 10^{14}$ nucleotide base pairs from our genome. Contrarily the length of our genome is 3.3×10^9 base pairs only, so this kind of information, related to body plan, seems to be highly compressed.

Cell lineage patterns follow perfect binary tree structure; DNA strand replications also follow binary tree structure. Recent experiments hint, that there could be high correlation between them. Considering this fact and the nature of cell divisions (symmetric or asymmetric), embryonic development could be described as a kind of wavelet transformation.

Since the DNA was discovered, this molecule is known to be the material of inheritance (Figure 1). These DNA molecules are contained in the cells mostly in the form of chromosomes. The principal role of DNA is the long-term storage of genetic information. The unit of DNA that carries this genetic information is called a gene. Each gene serves as a template on how to build a protein molecule. Proteins are among the most essential macromolecules, which perform vital tasks for the cell functions and serve as building blocks in tissues. The orientation of the genetic information defines the protein composition and their functions for each cell. Proteins are assembled from 20 different amino acids; usually a few hundred of them are connected into a chain by peptide bonds. The amino acid sequence of the protein is

coded by the messenger RNA (mRNA) and the “factory” where the “assembling” takes place is the ribosome. The ribosome is a conglomerate of RNA molecules and proteins. The template of these RNA molecules is stored inside the chromosome in the form of double helix DNA chains. This should contain the genetic information used in the development of living organisms too.

Coding sequences formulate only about 2-3 percent of the genome, the remaining 98% is the so called junk DNA. Some of this junk DNA has known function, the investigation of the remaining portion for function and role is the present and future challenge of molecular biology.

3 Computational Methods

3.1 Second Generation Wavelet Transformation of a 1D Signal

Traditional signal processing transformations, like Fourier transformation and even first generation wavelet transformations too, work well for infinite or periodic signals. Images and large part of the datasets are bonded in size and usually are not periodic, but they have a favorable property, that neighboring elements are correlated to each other. Figure 2 illustrates the second generation wavelet transformation of a one-dimensional (1D) signal.

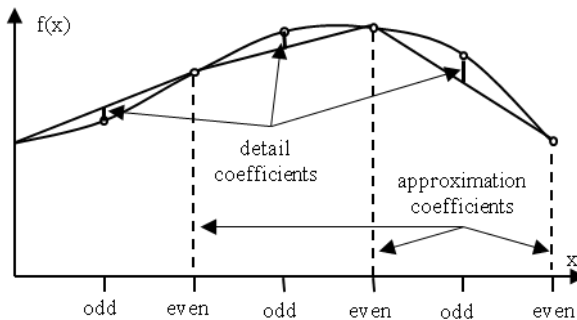


Figure 2 Illustration of the second generation wavelet transformation

General wavelet transformation comprises mainly two steps, filtering (to compare the signal to the kernel) followed by a downsampling process. In the case of second generation wavelet transformation there are three steps. First step is splitting the data into two parts, for example to even and odd, predict odds using even part, thus generating the detail coefficients and finally update even part using detail coefficients. Inverse transformation has similar steps in opposite order. The simplest use of this wavelet transformation is

for 1D signal.

Filtering:

In a 1D system we have an input signal x , and the transformed signal is y , the running variable is m , then the linear constant coefficient difference transformation of the signal looks like:

$$\sum_{(k) \in \mathcal{R}_a} a_k y(m-k) = \sum_{(k) \in \mathcal{R}_b} b_k x(m-k)$$

The region of the filters are \mathcal{R}_a and \mathcal{R}_b , a_k and b_k are real numbers. By rearranging this difference equation we get, that if $a_0 \neq 0$ and $(0) \in \mathcal{R}_a$ the transformed signal would be:

$$y(m) = - \sum_{(k) \in \mathcal{R}_a} a_k' y(m-k) + \sum_{(k) \in \mathcal{R}_b} b_k' x(m-k)$$

for example let the direction of recursion be left to right.

3.2 Two-dimensional (2D) System

How to carry out wavelet transformation in higher dimensional systems? In a general 2D spatial system the two variables are m and n , the linear constant coefficient difference equations could be written as:

$$\sum_{(k,l) \in \mathcal{R}_a} a_{k,l} y(m-k, n-l) = \sum_{(k,l) \in \mathcal{R}_b} b_{k,l} x(m-k, n-l)$$

Whereas x is the input image, y is the transformed image, $b_{k,l}$ are the feed forward coefficients, $a_{k,l}$ are the feedback coefficients. The solution for y :

$$y(m, n) = \sum_{(k,l) \in \mathcal{R}_a} a'_{k,l} y(m-k, n-l) + \sum_{(k,l) \in \mathcal{R}_b} b'_{k,l} x(m-k, n-l)$$

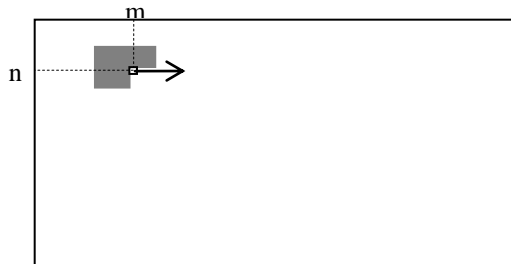


Figure 3 Filtering progression in a 2 Dimensional system

The progression could be e.g. horizontally left to right, vertically up to

down by each row, Figure 3 gives an example.

Wavelet transformations and the progression of cell division and differentiation chain, all are hierarchical transformations. Figure 4 shows example for the transformation of a square shape image into fragments of similar shapes, where L is the low frequency component, H is the high frequency component of the signal.

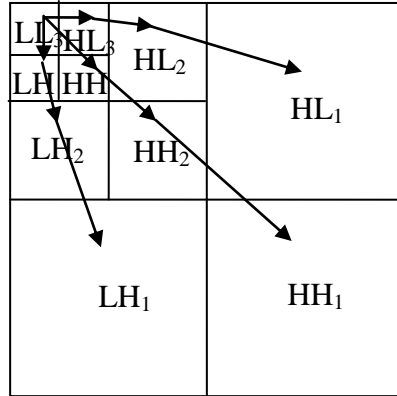


Figure 4 An example of a hierarchical image transformation in 2D system

Compression methods are very helpful to reduce data volume for archiving or transmitting images, data sets or packages, to reduce memory space, reduce required transmission time duration. The extent of compression depends on the rate of correlation between the elements of the dataset to be compressed. There are significant differences between compression methods. The higher the correlation (similarity) between elements of a neighborhood, the higher is the potential compression ratio of the compressed signal compared to the original signal.

H is the entropy of gray scale image, p is the probability of a given intensity value in the image, if $b=2$, then we get the unit of entropy in bits. For example in Figure 5 a and b both are 16 grades gray scale images, each grade with equal probability, so the two images have the same entropy value, in our example this value is equal to 4. To achieve high compression ratio, entropy should be low. If entropy really relates to the measure of disorder, then entropy values should differ between case a and b images. Common sense dictate, that image a could be compressed more efficiently than image b , although both of them contains exactly the same number of each pixel values.

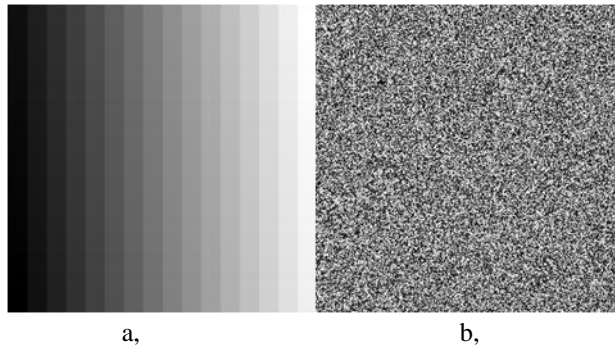


Figure 5 similar entropy value 16 grades gray scale images, each grade with equal probability (John Loomis)

In general image entropy is defined as:

$$H(x) = - \sum_{i=1}^n p(x_i) \log_b(p(x_i))$$

Compression methods are lossless or lossy. This bio inspired compression method is fully lossless, but could be modified to increase compression ratio further by modifying it to a lossy compression method. The requisites of this compression method are: the signal should be discrete and its region should be bounded.

3.3 *Matlab Modeling and Simulation*

All the image processing, modeling and simulation work mentioned in this thesis were performed by Matlab. Matlab is a matrix based numerical computational program with graphical interface to visualize the results. There are several toolboxes available to help the user to accomplish different tasks. Image processing toolbox was used to acquire images for further manipulations – the presented image transformation and compression jobs and visualization of tRNA movement. All computations and simulations for the publications in this theme were done by Matlab.

Matlab strength however lies in matrix calculations, direct matrix algebraic operations are much faster with it, compared to cycles. Filtering shown in Figure 3 for example suggests, the structure of the transformation program could be arranged as a double cycle along x and y axes. Due to the hierarchical structure of processes developed here, they can be solved by direct matrix operations, avoiding organizing cycles, which increases the magnitude of the processing speed.

Of course there are other available alternatives too, like Scilab or LabVIEW for instance. Scilab is very similar to Matlab, main distinction, that Scilab is freely available, but unfortunately there is no any reliable image processing toolbox on hand for it.

LabVIEW is a graphical programming language. All the above mentioned assignments could be solved by LabVIEW as well. (This is the future plan, image transformation method is already worked out by LabVIEW.)

4 Thesis Presentation

4.1 Thesis I

Every multicellular organism develops from a single cell, the zygote. During cell division, a cell always produces two descendent daughter cells. Descriptions of biological experiments found in literature suggest, that a causal association exists between the replicated DNA strands and the types of the descendent cells. Based on it a cell-numbering system was introduced.

What is the benefit to establish a cell numbering system? First of all we could be able to name and identify each and every individual cell of the body of a given organism, on the other hand to find out the structure or method of body plan coding.

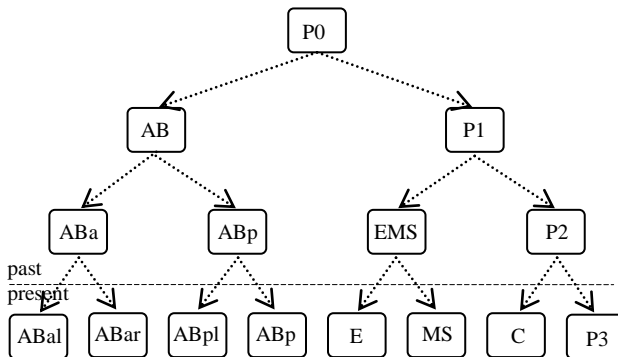


Figure 6 Cell lineage of *C. elegans*. P0 was the zygote cell, yielding 8 descendent cells after the third cell division cycles

Cell divisions follow a binary tree structure. The history to draw cell lineages is dating back to the discovery of microscopy, when cells and tissues were first discovered. The most studied organism is a worm *Caenorhabditis elegans* with 959 somatic cells.

Figure 6 shows the cell lineage of *C. elegans* after the third cell division

cycle, the present number of cells in this stage is 8 and the total number of previously existed cells is 7. Cell divisions follow a well choreographed and documented pattern.

DNA molecules are double helixes. How to tell apart each other if they are all identical? There is a relative DNA strand naming method to call one strand Watson strand and the complementary as Crick strand. It would be better to generate a consensus in naming the individual strands. DNA replication starts at the origin of replication initiation regions. They contain a unique DNA sequence pattern. One strand is rich in thymine (T) so consequently the complement strand is.

An autonomously replicating sequence (ARS) consensus strand:
 $5' \sim \text{ATTTATGTTT} \sim 3'$

so the corresponding complementary strand should contain the sequence:
 $3' \sim \text{TAAATCAAAA} \sim 5'$

This way we can call the strand containing the first ARS sequence as **T** strand and the adenine (A) rich complementary strand as strand **A**.

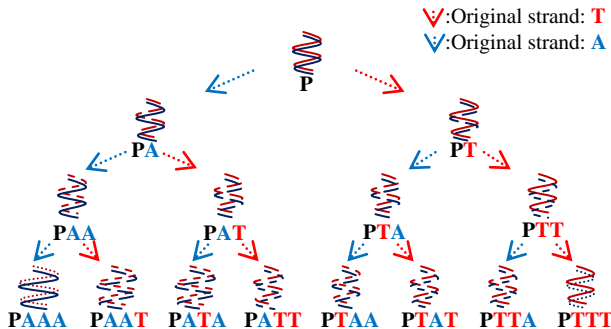


Figure 7 DNA strand identification

Development of a multicellular organism originates from the zygote and cell divisions follow a binary tree structure. For the sake of simplicity take an example of an organism, with a single chromosome only (like the Australian jack jumper ant, *Myrmecia pilosula* their genome is squeezed into a single pair of chromosomes, moreover males are haploid, so they have a single chromosome only). Meselson, M. and Stahl, F.W. (1958) proved, that during cell division DNA strand of each chromosome duplicates semi-conservatively. We can draw a binary tree, like the one shown in Figure 7, to follow each and every old and newly synthesized strand. Cell **P**, the ancestor contains the pair of originator DNA strands and both of its descendants carry one old and one new strands. Letter **A** or **T** indicates which one of the two is the original strand.

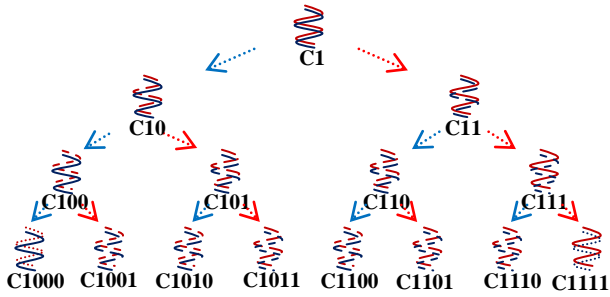


Figure 8 Replacing **A** and **T** letters by numeral 0 and 1

To create a cell numbering system, letters are replaced by numbers, the last digit of the ID number is set to 0 when strand **A** was the original strand and number 1 in the case of letter **T** (Figure 8). For example cell number **C1000** is a cell in the third generation (after the third cell division cycle), containing the original **A** strand of the zygote (**C1**), cell **C1111** of the 3rd generation holds the original **T** strand, and so on.

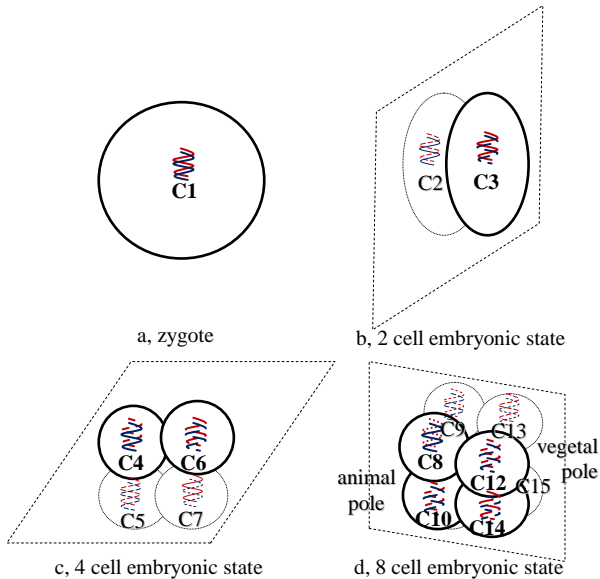


Figure 9 Embryonic development by consecutive cell divisions: different orientation planes are marked

There is an ongoing debate about immortal strand hypothesis, but there is no argument about the existence of high degree correlation level between

cell type and DNA identity of the descendant's chromosomes. If we can say, that there exists correlation between cell lineage and replicated chromosomes \rightarrow replicated DNA strands, we can rightly make logical connection between these two binary tree structures (Figure 6 and Figure 7).

Figure 9 combines this double entity by illustrating a developing embryo and identifying its cells by these corresponding cell numbers. Of course it is a fictitious numbering now, since there is no any experiment carried out yet for correct identification, however there are experiments pointing out its existence. Cell numbers are shown here as decimal numbers (ex. $C1111_2$ binary cell number is the same as $C15_{10}$ decimal cell number).

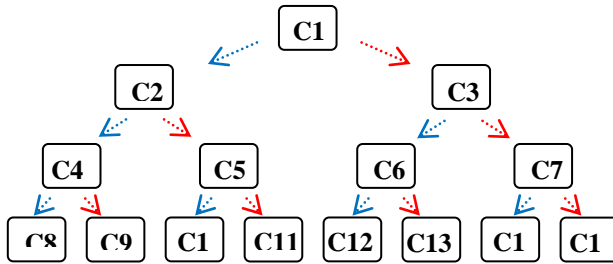


Figure 10 Cell division binary tree

Looking at this cell division tree (Figure 7 and Figure 10), the most intriguing thing we can notice is that we could establish two major cell division directions. Following the direction of the original **T** strands or following **A**-s. Practically it is possible to hide two different creature's genome in the form of one double-helix DNA strand.

	$y_1=0; y_0=0$	$y_1=0; y_0=1$	$y_1=1; y_0=0$	$y_1=1; y_0=1$
$x_1=0;$ $x_0=0$	$C_{16}=a_{00,00}$	$C_{18}=a_{00,01}$	$C_{24}=a_{00,10}$	$C_{26}=a_{00,11}$
$x_1=0;$ $x_0=1$	$C_{17}=a_{01,00}$	$C_{19}=a_{01,01}$	$C_{25}=a_{01,10}$	$C_{27}=a_{01,11}$
$x_1=1;$ $x_0=0$	$C_{20}=a_{10,00}$	$C_{22}=a_{10,01}$	$C_{28}=a_{10,10}$	$C_{30}=a_{10,11}$
$x_1=1;$ $x_0=1$	$C_{21}=a_{11,00}$	$C_{23}=a_{11,01}$	$C_{29}=a_{11,10}$	$C_{31}=a_{11,11}$

Figure 11 Pixel numbers and the corresponding binary tree identification numbers

In our discussion for simplicity and for the sake of better understanding we will represent this sample organism as a 2D object (Figure 11), which elements are the daughter cells. After each cycle the number of elements is

doubled, one time in x direction, next to the y direction. When we look at this resulting arrangement, it resembles to a two dimensional image with 2^n pixels after the n^{th} cycle (in our example this image has 4×4 pixels after the 4^{th} cycle).

Let's look at this image by other way around, an image is given of a size of $M \times M$ pixels, where $M=2^m$. Following the procedure opposite direction that we have seen in nature, halving the number of elements by each cycle, we create an image transformation algorithm. Our aim is, to find out to which known algorithm it resembles? Because in each cycle we halve the number of pixels, it means, that we downscale the image by a factor of 2. Discrete wavelet transformation (DWT) does the same (Figure 12), decompose the image with a high pass and a low pass filter. $h(n)$ high-pass filter provides the detail coefficients. $g(n)$ filter provides the approximation coefficients. After downsampling the image high frequency coefficients we get the first level coefficients and downsampling the low frequency coefficients we could get the further (second and so on) level coefficients.

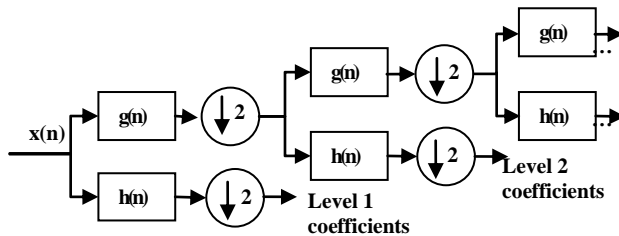


Figure 12 General DWT representation using cascading filter bank

Summary

By proposing a cell numbering system we could be able to name and identify each and every individual descendant of a predecessor chromosome (a zygote, practically the first cell after sexual cell division or crossing over). We can assume that cell division (asexual) is well choreographed, so we can suppose that exactly the same half of the DNA double-chain should remain always in the old cell and always the other goes to the new cell (immortal strand hypothesis). Linking cell lineages and this numbering system could help our understanding about how body plan information could be stored in our genome. Each cell belongs to one of the few hundred cell types. The resulting new cells occupy an available new location, position in the 3 dimensional space, it resembles a kind of image. Considering the pattern of cell differentiations after divisions, this process best could be interpreted as a kind of wavelet transformation. Describing cell divisions and differentiations

brings us closer to the understanding, how body plan information is compressed, what is the possible structure of this information. Based in this, author has developed an algorithm to transform and compress datasets for storage and transmission purposes.

4.2 Thesis 2

Classification of Cell Divisions.

There are two types of cell divisions: symmetric and asymmetric cell divisions, but in either case at least one of the daughter cells should look like as their predecessor.

Symmetric cell divisions

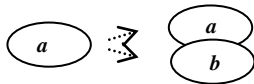
After completion of a mitotic cell division cycle of a eukaryotic cell, it produces two identical daughter cells.

Type “*a*” cell yields two “*a*” type cells.



Asymmetric cell divisions

Contrary to symmetric cell division the resulting daughter cells are of different types. One would be the same like their precursor and the other one would be different type.



In case of **intrinsic asymmetry** type “*a*” cell would develop and divide into one “*a*” and one different “*b*” type immediately. In the case of **extrinsic** (induced) **asymmetry**, daughter cells are similar directly after cell division but later their environment could induce differentiation, one of them could become different by interaction or by signaling with its neighbors.

Let us see the first cell division cycle, regardless that it is symmetric or asymmetric cell division, anyhow one of the progenies would be similar to the original, therefore $C2$ is similar to $C1$, $C3$ might be the same or might be dissimilar, depending its fate coded in this organism’s body plan. $C3=C1+D1$

In case of symmetric cell division $C3=C1$, so $D1=0$ or in case of asymmetric cell division $C3 \neq C1$ and this makes $D1 \neq 0$. And so on... in the moment illustrated in Figure 13 the actual cells are $C8 \dots C15$, $C1 \dots C7$ cells

were the previously existed stages. In this example $C8=C4=C2$ $C10=C5$, $C12=C6=C3$ and $C14$ is definitely similar to $C7$ ($C14=C7$). Values (cell type) of $C9$, $C11$, $C13$ and $C15$ cells depend on $D4$, $D5$, $D6$ and $D7$.

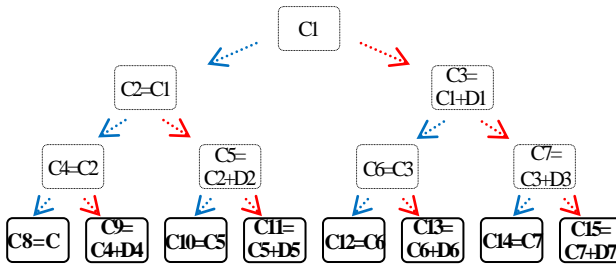


Figure 13 Cell differentiation

When we look at the cell division-cell differentiation biological process from the output end, this “reverse cell division” could be analyzed and considered as an image transformation. Figure 14 shows what is the main difference between this reverse biological progression and the general discrete wavelet transformation. Because after any type of cell division anyhow one of the daughter cells should be identical to their common predecessor, it means that in the reverse consideration the low pass filtering degenerates to a simple downsampling. The corresponding wavelet representation instead of the general case shown in Figure 12, would become like the one in Figure 14.

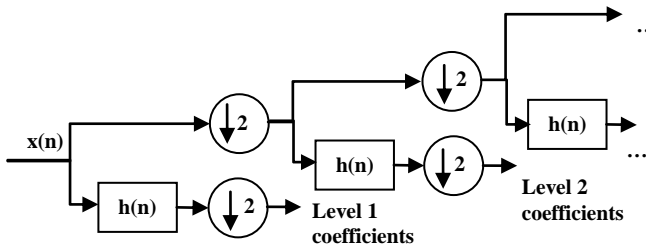


Figure 14 “Reverse cell division” representation with cascading filter bank

Image transformation of Thesis 2 will result a binary tree structure and the detail coefficients will be the element of this tree. Regardless what was the dimensionality of the original signal, image or dataset, the result would be a binary tree shape detail coefficients “heap” and a single remaining approximation coefficient (Figure 17). If the number of elements in the original

dataset was: 2^N , then the number of steps to establish the complete binary tree is N , the number of detail coefficients are $2^N - 1$ (extreme case: $N=1 \rightarrow$ gives one detail coefficient and one approximation coefficient). In case the number of elements is not equal to power of 2, then better to insert padding elements.

A number of the detail coefficients i.e. several elements of the binary tree are 0 or close to 0 values if the pixels are closely correlated to their neighbors. When some part of an image is undisturbed or the pattern is predictable, then the corresponding branch of the detail coefficient binary tree contains only 0 values.

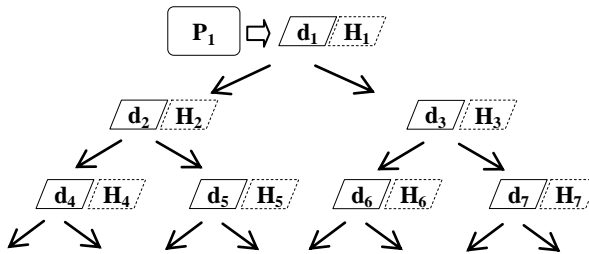


Figure 15 P_1 is the first approximation element and d_1 to d_{2^N-1} are the detail coefficients of the transformed image. “ H ” variable suggest the entropy level in the given branch of the binary tree

If we are certain that a branch of the tree contains only 0 values, then we can omit that branch of the tree. Let us introduce an internal variable H , this shows us whether the signal in the branch is disturbed or undisturbed. H is a kind of entropy value; let the entropy be $H=0$ when the branch of the tree contains only $d=0$ values and $H \geq 1$ when any of the detail coefficients is higher than 0 in that branch. Figure 15 depict these entropy values besides the detail coefficients of the binary tree.

How to generate these entropy-like coefficients? Its value should be equal to 0, when all of the detail coefficients of that branch of the binary tree were zero and moreover it should have similar value like the detail coefficients from which it is generated.

$$H_n = \frac{d_n + \frac{H_{2n} + H_{2n+1}}{2}}{2}$$

$$D_n = d_n - \frac{H_{2n} + H_{2n+1}}{2}$$

$$E_n = H_{2n+1} - H_{2n}$$

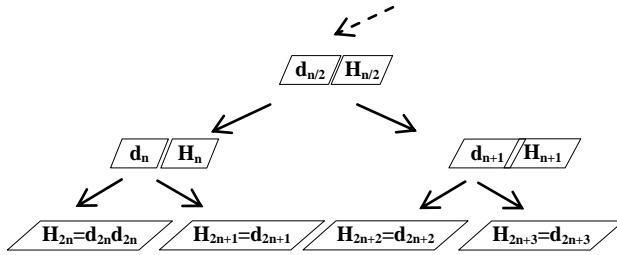


Figure 16 The end of a branch of the “ H ” entropy level binary tree

There are two types of extreme cases: the top of the tree, d_1 and H_1 and the other end, at the lowest level, at the very beginning of the compression.

At the lowest level let the H entropy value be equal to the ‘ d ’ detail coefficients and H_1 is at the top of the tree:

$$H_1 = \frac{d_1 + \frac{H_2 + H_3}{2}}{2}$$

$$D_1 = d_1 - \frac{H_2 + H_3}{2}$$

$$E_1 = H_3 - H_2$$

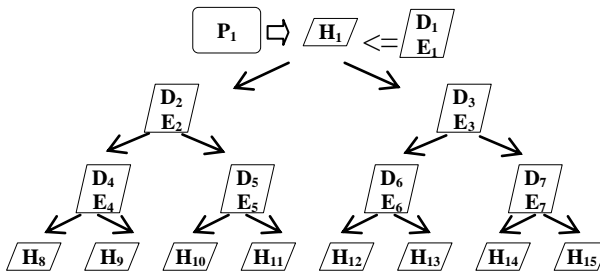


Figure 17 D and E coefficients of the compressed image

The H variable hint the entropy level in the given branch of the binary tree, see Figure 16.

Summary

This thesis is about an algorithm, to transform and compress discrete dataset for storage and transmission purposes. The transformation is based on the same algorithm how a multicellular organism develops from a single cell by consecutive cell divisions and differentiations. This algorithm could provide an economical way for handling a correlated or highly correlated discrete or quantized analog signal, image or any multidimensional data set.

The means of this algorithm is a special binary tree structure, inspired by developmental biology, the way, how a multicellular organism develops from a single cell. This observed natural process is an inverse transformation so on this base is possible to prepare a corresponding wavelet transformation method.

There are different methods to compress digital signals, datasets or images, each of them has some special characteristics. The procedure introduced here is entropy based adaptive compression method, its main feature is the hierarchical structure and entropy based compression. Wavelet transform of a dataset could be arranged to a binary tree like structure, which is the starting point of this compression. Degree of complexity of an image varies, different parts have different information content. Detailed regions require more thorough description than relaxed, predictable parts. To accomplish this task a new type of variable is introduced, resembling the entropy of the particular area of the image. The higher is the value of this variable indicating entropy, the higher is the importance of the area's detail coefficients.

4.3 Thesis 3

Any multi-cellular living organism develops from a single cell through consecutive cell divisions. After cell division one of the daughter cells retains the properties of the initial cell, the other might be the same or different. Any cell of an organism contains the same genetic information. The type of a cell is determined by its gene expression profile, some of the genes are enabled for transcription and the others are repressed, through epigenetic modifications, like DNA methylation and demethylation. In Figure 18 we represent the status of a particular gene. In the first generation we have only one cell, the status of this gene is G_1 (either logical 0 or 1). In the second generation one of the daughter cells retains this status,

$$G_{10}=G_1$$

the other sister cell's similar gene could have the same status or the opposite.

$$G_{11}=G_1 \text{ or } G_{11}=\sim G_1$$

*(\sim logical NOT)

How to describe the differentiation of this particular gene of the daughter cell? d represents whether the expression of the gene is the same or altered after cell division.

$d=0$ means same status, $d=1$ stands for altered status.

$$d_1= G_{11} \oplus G_1$$

*(\oplus logical operation XOR)

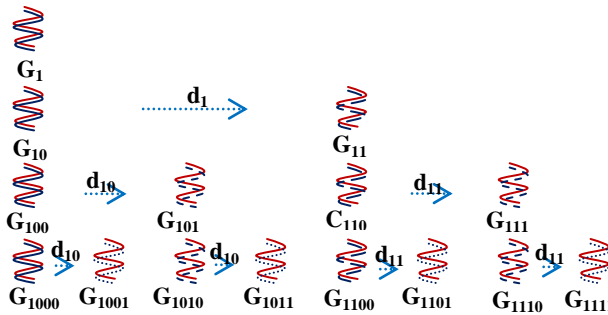


Figure 18 Cell division and differentiation

Development of an organism is well predictable; these factors are stored somewhere in the genetic information. Since always one cell divides into two new cells this process looks like a binary tree. Also these d values could be represented by a binary tree (shown in Figure 19). Indexing is in binary system. d is a logical variable, its value either 0 or 1.

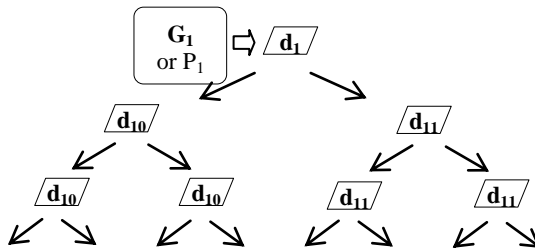


Figure 19 Binary tree shape representation of cell differentiation (G_1 is the initial state of G gene or P_1 is a pixel of an image)

The number of cells of a multi cellular organism is order of magnitude greater than the number of base pairs of the DNA chains in the nucleus. (There are 4 different nucleotides; the sequence of these nucleotides in the DNA chain of the chromosomes is called genetic information.) Moreover the number of genes in the genome is in the order of tens of thousands. That means, the genetic information concerning cell differentiation must be highly compressed.

In Figure 11 matrix elements could be the status of a gene in a biological system or it could be the value of a pixel. By consecutive divisions we reach to the fully developed organism. If it is about image processing, this is actual-

ly an inverse image transformation.

When we have a set of logical variables and the elements of this dataset are correlated to each others, then we are able to predict any element of it higher than 50% accuracy. To predict the value of an element usually we compare it to its direct neighbor or neighbors. The principal of wavelet transformation and some other transformation's too, is comparison between elements. The result of this kind of transformations is usually a perfect binary tree. The higher the predictability of the elements, the higher is the percentage of zeros than the ones in the transformed image. The aim of any image transformation to binary images is, to reduce the number of 1-s compared to zeros. This could increase the efficiency of any compression method. In Thesis 2 it was shown, that cell division and differentiation resembles to a kind of wavelet transformation, better to say an inverse transformation. This thesis gives an idea about how binary data set transformation and compression is used by nature, how this method could be utilized by us for image or dataset compression.

In Figure 19 G_I could be the status of a gene in a biological system or it could be the value of a pixel. By consecutive divisions we reach to the fully developed organism. If it is an image processing algorithm, this is actually an inverse transformation. When some part of an image is undisturbed or the pattern is predictable, then the corresponding branch of the detail coefficient binary tree contains only 0 values.

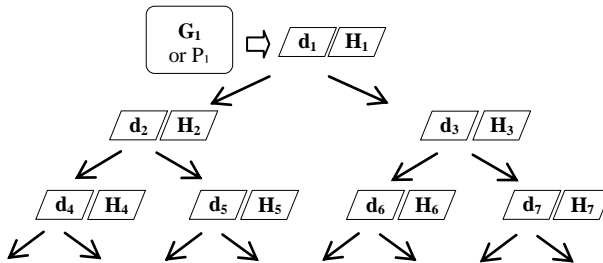


Figure 20 “H” variable suggest the entropy level in the given branch of the binary tree

If we are certain that a branch of this tree contains only logical 0 values, then we can omit that branch of the tree. Let us introduce an internal logical variable H , this show whether a branch is disturbed or undisturbed (Figure 20).

$H=0$

when the entropy of that branch is 0 and

$H=1$

when the entropy is higher than 0.

The result of the bio-inspired wavelet transformation is a binary tree, the newly introduced internal variable entropy coefficients are also structured into a similar size tree shape (Figure 21).

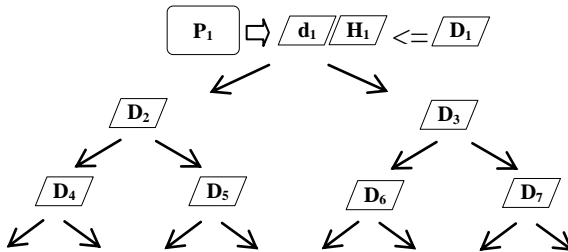


Figure 21 “d” detail coefficient and “H” entropy coefficient is replaced by “D” multi-valued variable

Instead of d detail coefficients we have to introduce a different coefficient, let’s call it D . D values obviously should contain information about the relevant value of the given d coefficient and should update the value of the internal entropy variable H .

Multi-Valued Logical Variable

D variable has to have more than two values, not only “true” or “false”. Besides coding for the detail coefficient d , it should give additional information for updating H entropy internal variable too. D is a kind of fuzzy variable.

Summary

Any multi-cellular living organism develops from a single cell, through consecutive cell divisions. After cell division one of the daughter cells retains the properties of the initial cell, the other might be the same or different. Any cell of an organism contains the same genetic information. The type of a cell is determined by its gene expression profile, some of the genes are enabled for transcription and the others are repressed, through epigenetic modifications. This thesis explains an algorithm, which was developed to transform logical dataset for storage and transmission purposes. This algorithm could provide an economical way for handling a correlated or highly correlated logical data set. The means of this algorithm is a special tree structure to transform a multidimensional dataset into a binary tree structure, following a natural algorithm. There are some potentially possible approaches shown,

how to compress this kind of bitmap. The main advantage of this lossless method is that the resolution of the reconstructed image is doubling by each step.

4.4 Thesis 4

The aim was to develop a model of tRNA molecular movement in bacterial cytoplasm and run simulations according to different tRNA concentrations and velocity conditions. The main criterion required in protein synthesis is the availability of the necessary amino acid in the vicinity of the ribosome in due time. Therefore, we examine the spatial movement or placement of aa-tRNA molecules in the cytoplasm—viewed from the perspective of that particular aa-tRNA (charged with a specific amino acid).

A 3-dimensional model and Matlab simulation program was developed, where particles are considered to follow the rules of Brownian motion. The speed of differently sized particles can be approximated using literature data from Project CyberCell *E.coli* Statistics. The average rate of amino acid assembly is around 20 ms per base, so if our simulation results show a rate of assembly that fits into this time frame, it is possible to conclude that the selection process is purely statistical or random.

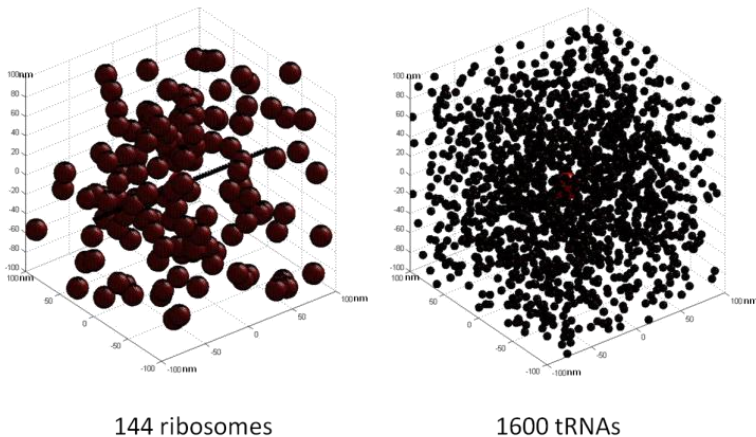


Figure 22. Simulation space with mRNA, 144 ribosome, 1600 tRNAs

Simulation results show, that a prospective consecutive cognate tRNA has no realistic chance of reaching a ribosome at reasonable rate in normal circumstances. The simulation gives one or more than one tRNA to hit the target only, when the number of cognate tRNAs or the time interval is exag-

gerated. Simulation space and particles are shown in Figure 22 and Figure 23.

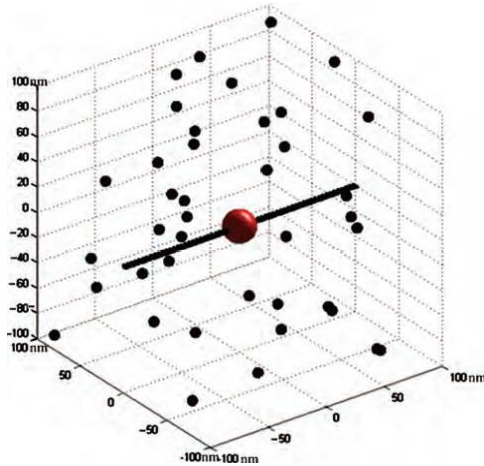


Figure 23 Simulation of cognate tRNAs movement.

Since the presented model proved that it is virtually impossible for the tRNA to reach the A site of the ribosome by random motion, we have tried to find possible explanations for this biological phenomenon. Three different scenarios are suggested, summarized in Table 1.

Summary

Transfer RNAs (tRNAs) can recognize a specific amino acid from a possible pool of 20. They are able to transport these protein building-blocks to the ribosome, the site where amino acids assemble into protein chains. Accurate and rapid selection of tRNAs by the ribosome is critical for cell survival. The aim of this thesis is to develop a preliminary and simple model of tRNA molecular movement in the bacterial (*Escherichia coli*—*E. coli*) cytoplasm. Spatial movement/placement of aminoacyl-tRNAs (aa-tRNA or charged tRNA) were examined in the cytoplasm—viewed from the perspective of that particular aa-tRNA. To achieve this goal, a kinetic model of the interaction between messenger RNA, ribosome, and RNA molecules is developed. The purpose of the simulations is to examine the conditions necessary for the tRNA to deliver a particular amino acid to the ribosome within a biological timeframe. Simulation results show that it is unlikely that tRNAs are able to reach the “A site” of the ribosome by random movement.

Furthermore three potentially probable mechanisms were proposed to explain tRNA pre-selection (to distinguish it from initial selection, it is re-

ferred in this presentation as “pre-selection”) in prokaryotes.

Table 1. Comparison chart of possible explanations

Hypothesis	Pros	Cons	Opinion
1. The ribosome stores tRNAs and preselects them.	-	Timeframe is not sufficient.	Very unlikely.
2. Signaling exists between the ribosome and the cognate tRNA.	Logical explanation. No contradiction against it.	There is no observation on the existence of signaling or specific force between the tRNA and the ribosome. It cannot explain the wobbling effect.	Possible. Should be proved or disproved by conducting biological experiments or observations.
3. The tRNAs reach the ribosome in a preselected manner.	Logical explanation. RNA-RNA interactions are known phenomena. It can explain the wobbling effect too.	There is not any observation.	Possible. Should be proved or disproved by conducting biological experiments or observations. New modeling and simulation is needed.

5 Summary of the Thesis

5.1 *Thesis 1*

As far a multicellular organism is considered a three-dimensional object, just like the progression of cell division and differentiation can be described by wavelet transformation. To support this:

- 1) set up a cell numbering system to support the applicability of wavelet transformation,
- 2) proved the existence of that wavelet transformation, which can be represented by the same binary tree, like cell division progression,
- 3) showed that similarly to the spatial orientation of cell division, downsampling required for wavelet transformation may have different spatial orientations.

Related publications: [5], [7], [10], [11], [13].

5.2 *Thesis 2*

In the course of formatting the thesis about a biologically inspired wavelet transformation and compression method for multi-dimensional image, discrete data set:

- 2a. Exemplified analogy between cell division binary tree on chromosome-level and image or discrete database wavelet transformation in conformity with Thesis 1. The transformed dataset is suitable for the subsequent compression.
- 2b. Defined an entropy like term resembling the orderliness of an image. Applying this, a hierarchical discrete dataset or image compression method was developed.

Related publications: [3], [4], [8], [10].

5.3 *Thesis 3*

The status of a gene is actually considered as a logical variable, its changes in the course of cell divisions can be represented as a binary tree, likewise a logical binary data set, bitmap or a monochromatic image too, therefore:

- 3a. An algorithm was developed, which can describe the gene-level cell division and differentiation, and can transform a multi-dimensional bitmap into binary tree structure. This logical variable transformation shows structural similarity to the discrete wavelet transformation described in Thesis 2.
- 3b. An entropy concept was created for this binary tree transformation that is capable to compress logical data sets, binary images. On this basis a hierarchical entropy based image compression method was worked out to compress bitmaps or binary images using Boolean algebra and multi-valued logic variables.

Related publications: [6], [9], [10], [12].

5.4 *Thesis 4*

A preliminary kinetic model was set up to imitate the physical movement of transfer RNA (tRNA) in the cytoplasm of *Escherichia coli* bacteria. The obtained simulation results verify the hypothesis that tRNA molecules are not able to approach the ribosome by random Brownian motion at a rate which would be sufficient to maintain the speed of protein synthesis.

Related publications: [1], [2], [14].

6 Application Possibilities

An example for Thesis 3 is shown in Figure 25, (Figure 24 is the original image) subplot 15 is the first downsampled level (256x128 pixels) and backwards 14...0 are the approximation coefficients (simple downsampling). Figure 26 shows the corresponding detail coefficients, most of the detail coefficients have 0 values (black).

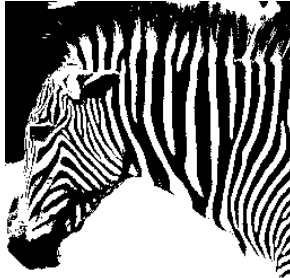


Figure 24 Image to be compressed; size 256x256 pixels

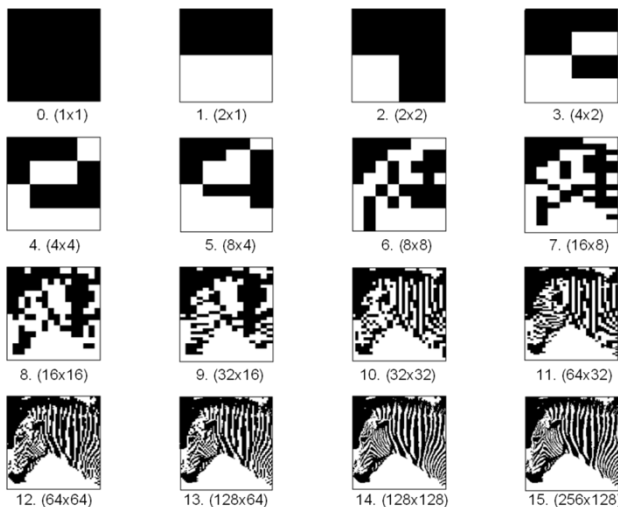


Figure 25 Image transformation and inverse transformation, to revert back the original image (logical variables)

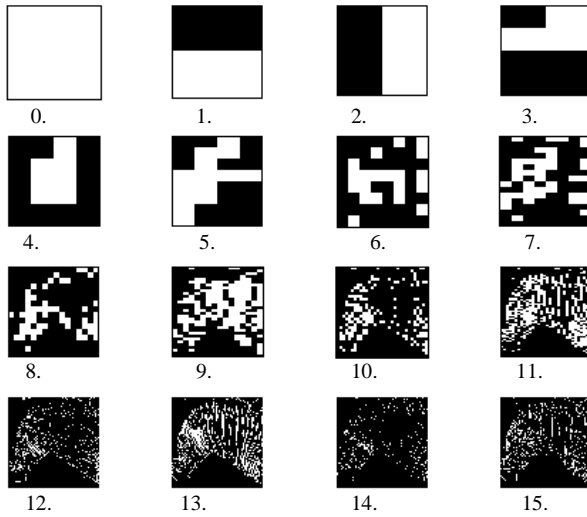


Figure 26 Detail coefficients of the transformed image

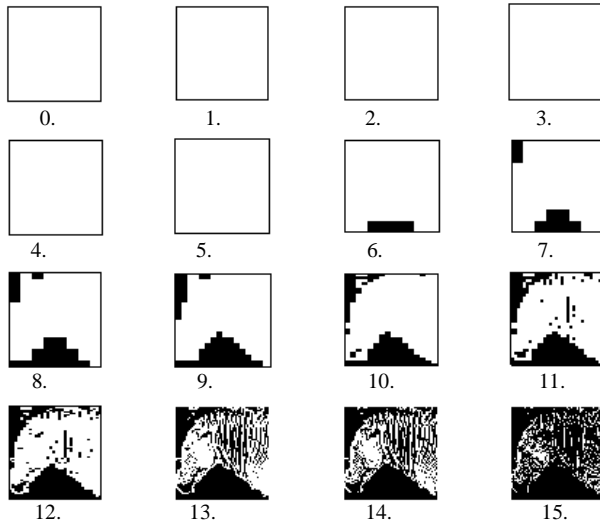


Figure 27 H entropy values of the image

In case of the previous example image, Figure 27 shows the entropy values in all 16 levels calculated by compression method explained in Thesis 3.

Figure 28 shows an example to image transformation according Thesis 2.

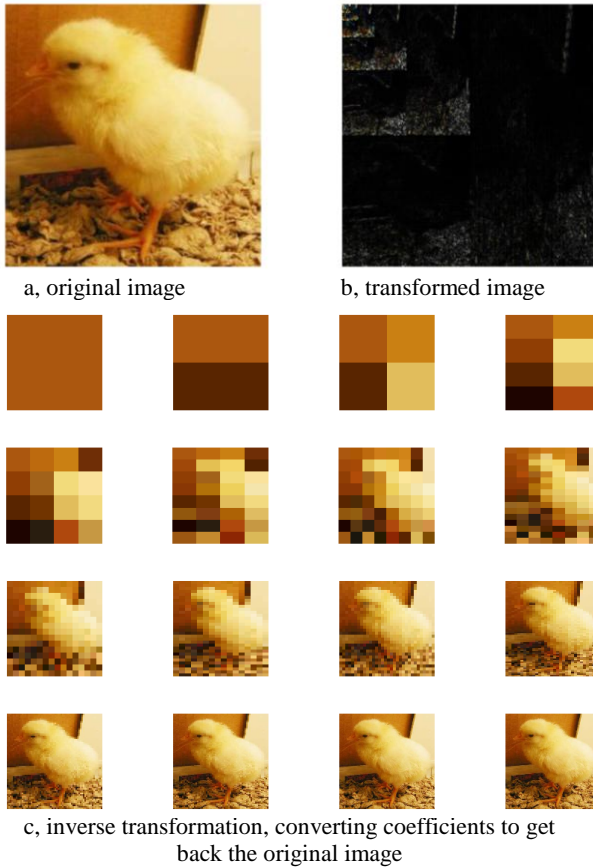


Figure 28 Image transformation

The most common and very simple compression method is the run-length encoding. These cell division inspired transformations coupled with these compression method grant a way by which we are getting closer and closer to the original image through a series of higher and higher resolution pictures.

Pros and cons of this method: this method provides similar compression ratios than other methods, but it takes us step by step closer to the reconstructed image, in each step we get a realistic image only in lower resolution. One useful utilization could be to code for object's location in intelligent space in robotics.

7 Publications Related to the Thesis

7.1 Article in International Journal of IF

- [1] Piros S., Husseini G.A. *Preliminary Modeling of Transfer RNA Kinetics in the Cytoplasm of Escherichia coli Bacteria*, *Advanced Science Letters* 3:(1) pp. 28-36. (2010) Independent citations: 2, IF: 1.253
- [2] Piros S., Husseini G.A. *Possible Physical Mechanisms of tRNA Pre-Selection in the Cytoplasm of Escherichia coli Bacteria*, *Advanced Science Letters* 3:(1) pp. 37-42. (2010) Independent citations: 1, IF: 1.253

7.2 Article in International Journal

- [3] Piros S., Korondi P. *Practical Implementation of Second Generation Wavelet Transformation on 2D Images by Transformation Matrixes*; *Recent Innovations in Mechatronics*, (accepted)
- [4] Piros S., Korondi P., *Entropy Based Adaptive Image Compression Procedure* (submitted)

7.3 Article in International Conference Proceedings

- [5] Sandor J Piros, *Tracing the Immortal Strand: Cell Numbering and Lineage Representation as Wavelet Transformation*, *Proceedings of the IASTED International Conference: Computational Bioscience (CompBio 2011)*. Cambridge: Acta Press, 2011. pp. 414-419. (ISBN:978-0-88986-889-2)
- [6] Piros Sandor J, Korondi Peter, *Compression Method for Binary Tree Like Bitmaps*, *Eighth International Symposium on Mechatronics and its Applications, ISMA'12: American University of Sharjah (AUS) & Emirates Aluminum (EMAL)*. New York: IEEE Press, 2012. pp. 1-5. Paper 6215197. (ISBN:978-1-4673-0860-1)

7.4 Article in Hungarian International Conference Proceedings

- [7] Piros S, Korondi P, *Developmental Biology from Informatics Point of View*, *11th IEEE International Symposium on Computational Intelligence and Informatics CINTI 2010: Proceedings*. Budapest: IEEE Hungary Section, 2010. pp. 225-228 (ISBN:978-1-4244-9279-4)
- [8] Piros SJ, Korondi P, *Biologically Inspired Informatics; Algorithm for Discrete Data and Signal Processing*, *Proceedings IEEE/ASME International Conference on Advanced Intelligent Mechatronics*. Piscataway: IEEE, 2011. pp. 972-977. (ISBN:978-145770838-1)

- [9] Piros S. J., Korondi P, *Biologically Inspired Informatics; Algorithm for Logical Data Processing*, 2nd International Conference on Cognitive Infocommunications: CogInfoCom 2011. New York: IEEE Computer Society Press, 2011. pp. 1-4. (ISBN:9781457718069)
- [10] Sandor J Piros, Peter Korondi, *Who said that pixels should be squares?* Proceedings of CERiS'13 - Workshop on Cognitive and Eto-Robotics in iSpace. Budapest: BME, 2013. pp. 40-45. (ISBN:978-963-313-086-5)

7.5 Article in Hungarian Conference Proceedings

- [11] Piros Sándor, Korondi Péter, *Informatika a Biológiában, Biológia az Informatikában*, Informatika a felsőoktatásban 2011 konferencia. Debrecen: Debreceni Egyetem Informatikai Kar, 2011. pp. 262-268. (ISBN:978-963-473-461-1)

7.6 Conference Participation

- [12] Sándor Piros, *Biologically Inspired Algorithm for Logical Data Processing*, Proceedings of CERiS'13 - Workshop on Cognitive and Eto-Robotics in iSpace. Budapest: BME, 2013. pp. 46-49. (ISBN:978-963-313-086-5)
- [13] Piros Sándor, Korondi Péter, *Sejtvonal és sejtdifferenciálódás reprezentálása wavelet transzformációval*, Kutatói hálózatok kapcsolatépítése - Debrecen az élettudományi kutatásokért konferencia. Place and date of conference: Debrecen, Hungary, 29/05/2014 *poster*
- [14] Sandor Piros, Peter Korondi, *Modeling of Transfer RNA Kinetics in the Cytoplasm of Escherichia coli Bacteria*, I. Innovation in Science - Doctoral Student Conference 2014. 207 p. Szeged, Hungary, 02/05/2014-03/05/2014. (Doktoranduszok Országos Szövetsége, Biológiai és Kémiai Tudományok Osztálya) Szeged: Magyar Kémikusok Egyesülete, 2014. pp. 108-019. (ISBN:978-963-9970-52-6)

7.7 Other Publications

Article in International Journal

- [15] Piros S., *Autonomous Cleaning Robot For Intelligent Building*, Analele Universitatii Din Oradea Fasciola Management Si Inginerie Tehnologica / Annals Of The University Of Oradea Fascicle Of Management And Technological Engineering X (Xx):(3) Pp. 2.35-2.41. (2011)