The 6th International Conference for Informatics and Information Technology (CIIT 2008)

FILE SYSTEM ORGANIZATION IN MINIMAL BIOLOGICAL SYSTEMS

Nevena Ackovska
Institute of Informatics, FSNM
"St. Cyril and Methodius" University
Skopje, Macedonia

Stevo Bozinovski
Department of Mathematics
and Computer Science,
South Carolina State University
Orangeburg, SC, USA

Gjorgji Jovancevski
University American College
Skopje, Macedonia

ABSTRACT

This paper elaborates on different organizations of the files in the genetic systems. These organizations of the file system are also detected in the human created file systems. They include different groupings of genes related by similarity, tasks, segmentation and packing. The other part of this paper elaborates on another important concept, the concept of genomes: set of all genes of specie. There are some valuable lessons that could be learned by observing how the nature handles its file system.

I. INTRODUCTION

The genetic material is the substance that contains the information specifying the inherited characteristics of an organism. It wasn't until 20th century that people realized that the DNA is the inheritance substance. Before that people believed that the proteins are inheritance carriers.

DNA is a double helix of two intertwined polynucleotides. The polynucleotides are usually referred to as strands of the DNA.

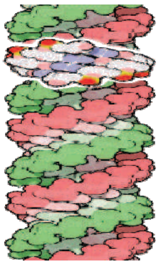


Figure 1: DNA molecule

Today it is a common knowledge that the genetic material is carried out by segments on DNA named genes. Genes encode different characteristics of an organism named phenotypes. They give attributes to an organism, and they have values.

If a gene gives attribute "eye color" then its values (alleles) would be "black", "brown", "blue", and "green". However, the genes as segments on DNA do not encode a phenotype attribute directly. All they do is encoding proteins and RNA's. Therefore, we will use the definition that: Genes are segments of DNA that encode either RNA or proteins. Also, one could encounter slightly different definitions [1, 2].

Genes are found of both strands of DNA [1]. However, they are read only in one direction, so called 5' → 3' direction (Fig. 2). Sometimes genes for both sides of the DNA overlap.

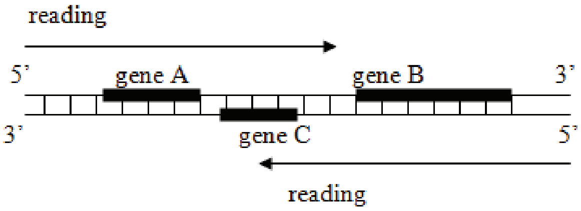


Figure 2: Reading direction of DNA

II. GROUPS OF GENES

We find interesting that there are several types of gene groups found in DNA. Here we will consider the concepts of: similar groupings, related tasks groupings, segmented genes and chromosomes. We will discuss them in the sequel.

A. *Groups of similar genes*

The groups of similar genes are divided in two major subgroups: Clusters of same genes and Clusters of similar genes

1) Same gene Clusters

These clusters represent groups of same particular gene that spreads through some region of DNA. Example of a same gene cluster is the cluster of 5S rRNA gene, that is contains many copies of the same gene [3]. This particular cluster is needed for fast parallel production of rRNAs (Figure 3). In humans there is a cluster containing 2000 copies of that gene.

It seems that the concept of redundancy and clustering of same genes adds to the speed of production of some vital resources of the cell [4]. And indeed, in the moment when a need for many ribosomes is sensed, many copies of this particular file must be produced in parallel, because in the living cell, no waiting for vital resources is allowed. Even worse, cell doesn't store material for future need; it assembles the needed material in the moment of the need.

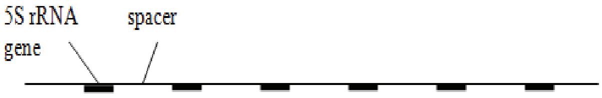


Figure 3: Cluster of 5S rRNA gene

2) *Similar gene Clusters*

Some clusters contain similar genes but not exactly the same. They represent a gene family.

One example is the gene family for the globin polypeptides in vertebrates. The combination of those globin polypeptides with a haem peptide gives hemoglobin, which is very important protein which is the oxygen transporter in vertebrates' bloodstream.

Haemoglobin is complex molecule, containing four globin protein chains and a haem protein. Two of the proteins are *alpha chains*, two are *beta chains*. So a haemoglobin protein can be represented as (2α + 2β + hem).

The haemoglobin gene clusters in humans are located in chromosomes 11 (β-shaped polypeptides) and 16 (α-shaped polypetides), as represented in the Fig.4. In various stages of human development various genes are accessible. This particular gene is a member of another very important gene group – genes responding to developmental events.

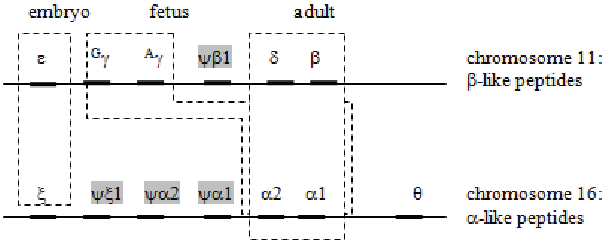


Figure 4: The globin family genes

As shown in the Fig. 4. the embryo phase activates genes ξ and ε so the haemoglobin is like (ξ, ξ, ε, ε). The fetus phase activates genes (α2, α1, $^G\gamma$, $^A\gamma$). In the adult phase the genes (α2, α1, δ, β) are active.

The shaded genes in the hemoglobin cluster (Fig.4) are not functional, and are called *pseudogenes*. Probably they loosed their function during the evolution.

The evolution explores alternatives, by duplicating a gene and then waiting a mutation in a gene. Whichever of the genes prevail, it will be selected by evolution.

B. *Clusters of genes related by a common task - Operons*

Operon is a specific cluster of genes, related by a common control function. It contains several genes that are released simultaneously to perform a specific task. Existence of operons was postulated by Jacob and Monod [5] and was subsequently confirmed by experimental evidence.

Operon is a control structure that controls expression of a number of genes simultaneously.

An operon based control system releases several genes and produces several proteins simultaneously. The proteins, for example enzymes, deal with the cell event that triggered the activation of the operon [6]. This is represented in the Figure. 5.

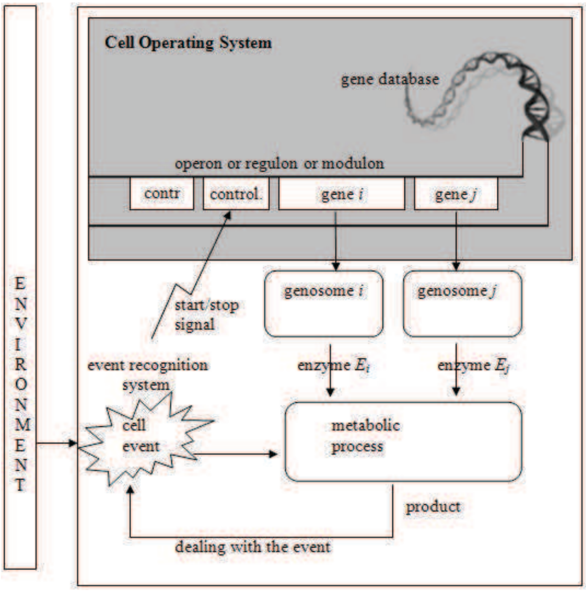


Figure 5: Dealing with an event by a cell control circuit

The genosome in Fig. 5 denotes transcription-translation machinery that, given a gene, produces a protein or RNA.

In order to illustrate how the operon structure functions, let us consider a simple example of milk consumption. Once the milk enters our intestines, resident E. Coli bacteria receive signals of the presence of the sugar lactose. That is the material that should be processed into glucose and galactose. Figure 6 shows the processing of the lactose input [7].

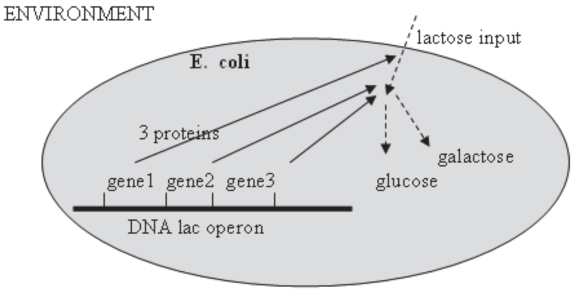


Figure 6: Processing of an external event: lactose enters a cell

Once a lactose molecule is sensed by the E. Coli signal proteins, they start processing a file containing 3 programs (genes). The enzyme RNA polymerase will transcribe three mRNA source codes. They will be compiled into three proteins: β-galactosidase, lactose permease, and β-galactoside transacetylase. They will work as three robots and will disintegrate lactose into galactose and glucose. The process is highly regulated and real-time. The lactose related robots exist only when the lactose is present, when they reach level of up to 5000 molecules in only few minutes. When lactose is not present any longer, the lactose regulating robots are not assembled. Furthermore, all those lactose processing robots will be disintegrated, and their components will be recycled into other robots or other needed cell components.

There are several levels of hierarchical control systems over the operon level. Several operons can be controlled by a regulatory system which is denoted as regulon, and a set of several operons and regulons can be controlled by a control system denoted as modulon [8, 9]. The number of hierarchical levels of the cell control system is unknown, even for the prokaryotes [9].

C. *Segmented genes*

Some genes are segmented. In between two segments of such a gene, there is a segment that could be spacer, or could contain another segment of another gene. The gene segments are named **exons**, and the strings between exons are named **introns**. Exons are found in eukaria, archea, viruses, but not in bacteria.

One prominent example of a segmented gene is the gene of a cystic fibrosis protein (Figure 7). It is a transmembrane regulator.

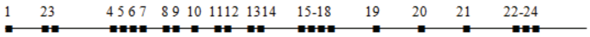


Figure 7: Gene of cystic fibrosis protein

As shown in Fig. 7, the gene of the cystic fibrosis protein contains 24 short exons. They are distributed over a DNA string of length 250,000 bp. Only 4% of that length is the actual gene. The average length of each exon is 227 bp. Introns could be much bigger, 2 bp to 35 Kbp

Exons are part of the modular software, and are used to produce variety of proteins from the same memory structures. Figure 8 shows this concept.

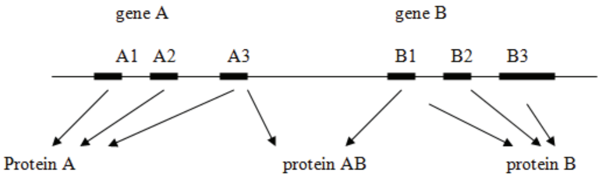


Figure 8: Reusable software – exons

The reusability of the exons is used for example in the process of building antibodies in the living creatures.

D. *Chromosomes*

The genes are organized in chromosomes. Each chromosome is one string of packed DNA. They are segments of DNA clearly visible when DNA replicates. The number of chromosomes varies in different organisms: each organism has specific number of chromosomes. Table 1 shows different number of chromosomes in some living beings.

Table 1: Numbers of chromosomes in different beings

organism	#chromosomes	#genes	genes/chromosome
E. coli	1	2800	2800
yeast	16	8750	550
human	23	~25000	~1100

In a human genome there are 23 pairs of chromosomes (Fig. 9). While each of the first 22 chromosomes contains a copy of itself, the X and Y chromosomes have no copies. Figure 9 represents a set of male human chromosomes. A human female would have two X chromosomes and no Y chromosome.

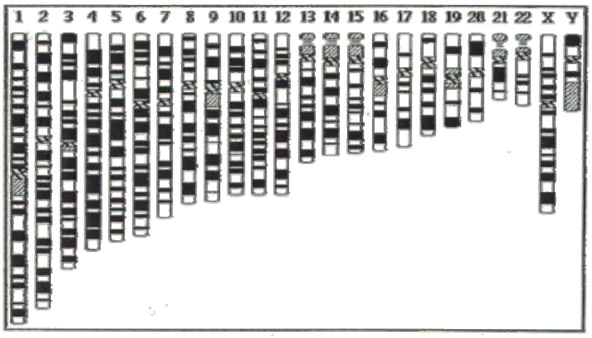


Figure 9: Set of human chromosomes – male

III. GENOMES

It is important to study the content of a DNA in sense of distribution of the genes and other sequences along a DNA.

A genome is a set of all genes of a species, organized into chromosomes. It actually presents an organism genotype that produces a particular organism phenotype.

Prokaryotes usually contain a single chromosome in their genome.

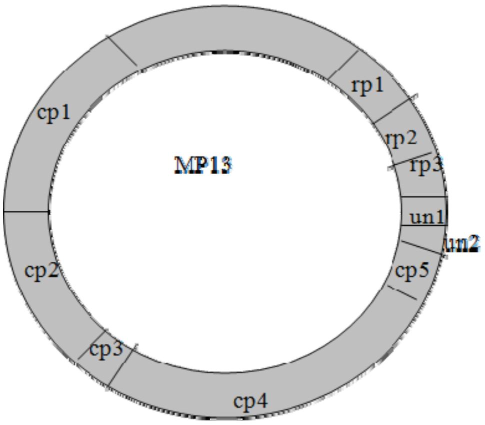


Figure 10: Phage genome M13

The simplest genomes are those of bacterial viruses, bacteriophages or simply, phages. Having small genomes, phages contain small number of genes. Figure 10 presents the M13 phage genome. It is on a circular DNA. This phage has: 5 genes for capsid formation, 3 genes for self-replication, and 2 unknown function genes. It is very interesting that the genome of such a simple organism has genes with unknown function.

The next figure represents the genome of another phage - the ϕX174 genome. It is interesting because it shows overlapping genes in the same genome.

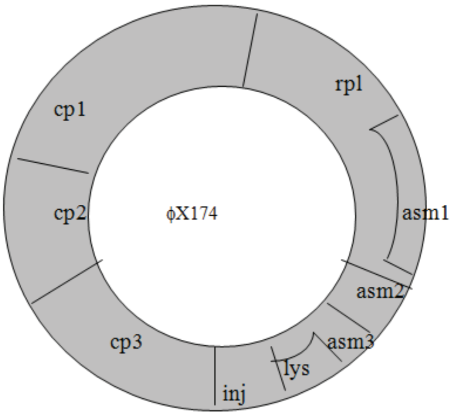
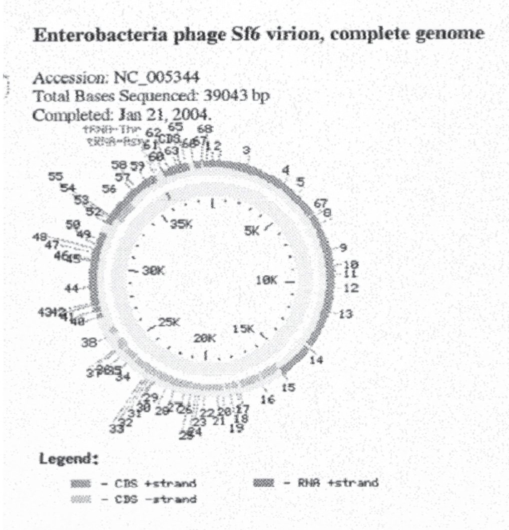


Figure 10: Overlapping genes in a genome

Today the search for genome data is made easier by the numerous web sites that offer genome search, and genome pattern matching. For example, Fig. 11 represents the genome of Sf6 virion given on National Center for Biotechnology Information (NCBI) [10].



Enterobacteria phage Sf6 virion, complete genome

Accession: NC_005344
Total Bases Sequenced: 39043 bp
Completed: Jan 21, 2004.

Figure 11: Sf6 virion representation by NCBI

There are several genome database portals where one can find structured and functional information about genomes of different species. Some of them are Kyoto Encyclopedia of Genes and Genomes (KEGG) [11], NCBI, Human Genome Database (GDB) [12] etc.

IV. Conclusions

This paper presents an overview of the file system organization in the minimal biological systems. The file system is divided in files that are being processed as a whole in the genetic processes. Those groups are divided in 4 subgroups: similar groupings, related tasks groupings, segmented genes and chromosomes.

As we could observe, in bio-systems everything is packed and protected. The packing enables protection and coding.

The same gene clusters introduce the concept of redundancy. This is very vital factor since redundancy ads to the speed of production of vital resources for the cell.

On the other hand, segmentation of one file in several pieces enables re-usability of those pieces for construction of different end products.

Another very interesting concept is the concept of genome. A genome is a set of all genes of a species. It actually presents an organism genotype that produces a particular organism phenotype. Some genomes are fully discovered. Other species still have genes that have unknown function. Genomes could be observed in various genome database portals.

References

[1] T. A. Brown, *Genetics, A molecular Approach*, Chapman & Hall, 1992

[2] S. Bolsover, J. Hyams, S. Jones, E. Shephard, H. White, *From Genes to Cells*. Willey-Liss, 1997

[3] T.A Brown, *Genomes*, 2-nd Ed., Willey-Liss, 2002

[4] N. Ackovska, S. Bozinovski, G. Jovancevski, "A New Frontier for Real – Time systems – Lessons from Molecular Biology", *Proc. IEEE SoutheastCon 2007*, pp. 224-228

[5] Monod J., Pardee A., Jacob F. "The genetic control of cytoplasmic expression of 'inducibility' in the synthesis of b-galactosidase by Escherichia coli", *Journal of Molecular Biology* 1, pp. 165-178, 1959

[6] S. Bozinovski, G. Jovancevski, N. Bozinovska "DNA as a real time, database operating system". Proc SCI 2001, Orlando, 2001, pp. 65-70

[7] N. Ackovska, S. Bozinovski, Gj. Jovancevski "Real-Time Systems – Biologically Inspired Future", *Journal of Computers*, Vol. 3, No. 3, pp. 56-63, Academy Publisher, 2008

[8] A. Engel "Beyond CIM: Bionic manufacturing systems in Japan" *IEEE Expert*, 79-81, August 1990

[9] J.Lengeler "Metabolic networks: a signal-oriented approach to cellular models." *Biological chemistry* 381, 2000

[10] http://www.ncbi.nlm.nih.gov/

[11] http://www.genome.jp/kegg/

[12] http://www.gdb.org/