PERIODICO di MINERALOGIA Volume 88, No. 6, 2019

Epigenomics: Unifying Genomics and Proteomics for an Efficient Functional Genomic Approach of the Current Genetic Analysis

NATALIA CUCU

Institute of Genetics, Department of Transgenesis, University of Bucharest, 1-3 Aleea Portocalilor, Bucharest, Romania

Abstract

One of the major tasks of the current molecular genetics is the understanding of the mechanisms involved the control of the complex eukaryote gene expression, for practical purposes of solving medical, pharmaceutical, agricultural, industrial, environmental and recently arisen ethical problems. Despite the numerous advances in this domain, especially those based on the individual gene and protein sequencing, the initial goals of it to explain life in its complexity could not have been completed: the factors orchestrating the gene activity from the genotype till the phenotype, during the development, ontogeny and even evolution of organisms are still incompletely defined. An additional epigenetic information to the already established domains of genomics and proteomics offers at present the opportunity to unify these ones, into a newly emerged domain: epigenomics. This one is the only able to explain the key early lacking element concerning the impact of the environment on the genes and proteins function: the control of the spatio-temporal gene expression. By preserving the genetic information or the primary nucleotide sequence, the chromatin modification by DNA methylation is now the focus of the new domain for targeting the potential solutions in medical prevention/cure and any economic problems linked with the environmental factors-such as adjustment and biodiversity.

Keywords: DNA methylation, chromatin modeling factors, gene expression, genetic and epigenetic information, genomics, proteomics and epigenomics, functional genomics, high-throughput analysis, large-scale analysis, preventive medicine

Introduction

The advances of the recently finished world-wide Human Genome Project and of other currently running projects on functional and comparative genomics have lead already to the explosion of knowledge about our genes and have made numerous connections with many social and economic domains, especially with medical and agricultural ones. We are convinced now that our genetic make-up is deeply involved in our lives. In spite of a huge amount of data has given to biologists the opportunity to explain numerous life events at the cell structures, functions and even at their genetic determinism level, yet an identified part of vital processes, particularly linked with the genome-environmental communications, has still remained undefined. As a Nobel Prize laureate, Jacob, has underlined, “we have started to understand the cell, but not the tissue and organ”. It is still a matter of investigation how cells

of an organism, programmed by the same genotype, can differentiate in a specific tissue at the right moment and in the right environmental conditions, as conforming to an additional spatio-temporal order.

This limitation has been imposed by so far approaching the *genome variation* only at the *nucleotide level*, thus considering the unique interactions between *the phenotype* as the final gene expression and its *corresponding primary DNA structure, or sequence*. Studies on the genome variations have been therefore carried out especially based on random single gene *mutants*, lacking of a certain detectable phenotype. In part, this has been efficiently catalyzed by the recently finished program of sequencing and mapping the human genome. It had required for modern rapid, precise and accurate techniques that facilitated the acquisition of raw sequence data and thus permitted subsequently complex genetic analyses. In this way, the *technological advances in automation and bioinformatics* contributed to the emergence of a new discipline of biology, called *genomics*. It can be broadly defined as *the generation and analysis of information about genes and genomes*.

Undoubtedly the completion of the Human Genome Project was possible only because of the improvement in such approach and the sequencing technologies. So, the rapidly sequencing of individual genes by automated capillary electrophoresis using fluorescent nucleotides has allowed researchers to screen for its primary structure the entire 14,8 billion base-pair human genome over just nine months (VENTER et al 2001; MARTIN and NELSON, 2001). The benefit of such discoveries is already obvious in the domain of medicine, allowing the actual developing of new domains like *molecular diagnosis* (identification of new clinically phenotypes and their associated genes) and *pharmacogenetics* (chemically targeting certain genes and DNA sequences or certain wrong acting metabolic pathways).

Besides all the mentioned advances the *genomic sequence has serious limitations*: it does not specify the protein-protein interactions or the specific protein location in the cell under various conditions. It has been also proven, that transcript abundance levels do not always correlate with the protein abundance level; moreover, one cannot specify from the genomic sequence that a gene may be translated into protein or rather may function as an RNA. On the other hand, recent discoveries have shown that one does not have to look at an entire gene coding sequence in order to detect certain disease linked mutations: a recent discovery is the fact that often diseases, in spite of the fact that they have a genetic component, they are not determined by a single gene, but rather are influenced by a *multigene system activity*. Meanwhile, complex diseases, like arteriosclerosis, cannot be predicted only by genetic testing: the phenotypes do not always match the risk predicted by the study of a certain gene. Identification of SNPs (single nucleotide polymorphism) representing the right variants in the considered gene has often proved to be a better approach for such purposes. Such polymorphism, as molecular markers, does not affect the gene regulation or the function of encoded proteins, but can be considered pinpoint for the location of those specific genes responsible for the disease phenotype (BROPHY and JARVIK, 2000; JARVIK, 2001).

Such new features may also indicate the fact that the function of our genotype is determined not only by *the sum of our component genes* (the former, classical definition of the genotype). It indeed indicated that our cultural, social and physical environments and histories have more of an impact on shaping who we are and how we function as complex eukaryotes than our genetic make–up (WILLIAMS, 2001). The requirement for the *correlation between the genomes, the cell type in which such genomes reside and the environment where they express their phenotype* arose in this context.

This review is aiming at a state of the art presentation of the already worldwide emerging domain of epigenomics: as the only one able to offer the opportunity to study the above mentioned genome-environment relations through a new epigenetic information, based on the DNA methylation markers. Such genome modification, localized not only in the gene coding sequence, but rather in the so-far called "junk", mostly repetitive DNA, proved to have dramatic effects on the spatio-temporal gene expression. Therefore, it has been generally agreed that such DNA modification may represent new additional information regarding the gene function, which enables the communication between the genome and the environment.

The knowledge of its basic concepts would help the implementation in our country of such an approach in order to figure certain important solutions for solving medical, agricultural or environmental problems by proper biotechnological and analytical methods.

Functional Genomics represents the proper domain for studying and deciphering the eukaryote normal and pathological development in its complexity

The need for study genes in their biochemical, cell, organism and environmental context imposed the developing of a new domain for characterizing the basic correlations between the structure and the function of genes: the so called "*functional genomics"*.

The cellular functions are not carried out by *nucleic acids*, but by *proteins*; moreover, the function of these molecules, through complex protein-protein interactions, may be directly or indirectly correlated with the final phenotype behavior at the complex level of the eukaryotic organism. The term *„proteomics"*, derived from „proteome", which means the complete set of proteins expressed by a given genome has been therefore designed to define a new, parallel domain. It is dealing with the identification and characterization of the protein function and amino acid sequences (MARTIN and NELSON, 2001). Interestingly, a closer look at the protein make up of a cell has revealed that protein modifications may not be apparent from the nucleotide or amino-acid sequences, but may rather be modeled by other informational levels. Therefore, from a medical point of view, in pharmaceutical interventions and diagnostic tests, the advances in this field have often given successful solutions for targeting proteins and not genes. The new approach has already been possible by the development of accurate techniques of two-dimensional gel electrophoresis (2DE) and recently, a more rapid and large scale approach by mass spectrometry (MS), yeast two – hybrid systems and protein arrays.

Functional genomics is divided conceptually in two distinct approaches. The *gene driven approaches* use *genomic information* for identifying, cloning, expressing and characterizing genes at the molecular level. The *phenotype driven approaches* analyze phenotypes from random mutation screens or naturally occurring variants (mouse mutants, human disease) to identify and clone the gene(s) responsible for a certain phenotype. Such a large approach domain of the actual functional genomics has imposed different and complementary *experimental models* for the elucidation of gene function at its final levels of expression. So, h*igh throughput analysis of gene products* (transcripts, proteins) and *biological systems* (cell, tissue or organism) using *automated procedures* allowed classically performed experiments for single genes or single proteins (e.g. generation of mutants, analysis of transcript and protein) to be extended to the so called "*large – scale*" level (e.g. through numerous intergenic or protein-protein interaction). The new genome and proteome wide

analysis represents actually the *modern systematic* effort of understanding the function of genes and gene products into the biochemical, cellular and organism context (YASPO, 2001).

However the newly acquired systematic large-scale data from the more global view of the genome and proteome permitted to decipher the function of single genes, they have limitations regarding the entire genome approach and moreover, at the organism level. It contributed to deepen the already existing *gap* between the *accumulation of sequence* information and the *understanding the entire orchestration of normal and pathological development in eukaryotes*. One cause of this limitation, as it has been already above mentioned, resides in the fact that characterizing the function of a gene is not straightforward since it depends on the *molecular context within a given cell type* and in a *particular cellular microenvironment.* Last but not the least is in this context the *intergenic interaction within a given genome*, which does not change the DNA sequence, but rather its further expression in the given cell type and the given cellular environment.

Such information - additional to that gathered from the nucleotide or amino acid sequencing - represents the recently defined *epigenetic information.* It is dealing with the activity of those molecules which are involved in the activity of the *different levels of the gene expression*: transcriptional, posttranscriptional, translational and post translational ones, provided that the primary correlation between the nucleotide sequence and the corresponding coded amino-acid sequence remains unchanged. Therefore, epigenetics does not operate basically through mutagenesis but rather by more versatile DNA modifications, which permit the specific direct interaction with precise transcriptional tissues specific factors and with certain indirect interactions with the environmental ones.

None of the mentioned *functional genomics* domains, genomics and proteomics, although each of them is fruitful in itself for providing important information, could answer the most important questions about the genetic shaping of life and its vital processes in a given environment. Many advances in the two-mentioned areas have successfully demonstrated the involvement of diverse *epigenetic phenomena*. Most prominent among these are recent studies on gametic and embryonic *imprinting*, genetic hierarchies in *embryonic development* and *cytodifferentiation*, *senescence* and perhaps most recently, on epigenetic mechanisms *of gene activation/inactivation in cancer* (RAZIN, CEDAR, 1994; RAZIN, KAFRI, 1994; RAZIN, SHEMER, 1995; RIGGS, 1995; BARLOW, 1993; BARTOLOMEI and TILGHMAN, 1997; LAIRD and JAENISCH, 1996, 1999; JONES, 1999; SZYF, 2001). Epigenetics is an approach that views these and other such complex phenotypes from the genomic level *down*, rather than from the genetic level *up* by providing powerful *in*- sights into the functional interrelationships of genes in development of organisms (CHRISTMAN et al, 1977; EHRLICH and GAMA-SOSA, 1982; MEYER, 1995; MATZKE and MATZKE, 1995ab; MARTIENSSEN, 1996, JANOUSEK et al, 1996), in health and disease, respectively (BECK et al, 1999).

The promises of the recent (1999) established Epigenomics consortium, as a continuation of the Genome Sequencing one, is therefore to use new, specific large scale analysis by conventional high-throughput sequencing of DNA modifications other than mutations, like DNA methylation. The specific methods including PCR – sequencing methods applied to bisulphite-treated DNA, DNA mass spectrometry, methylation sensitive - PCR and DNA restriction may now provide an organism's „epigenotype" as a complement and additional information for the known, almost sequenced "genotype" (BECK, OLEK and WALTER, 1999).

Specialists are now agreed that "epigenomics" allows a better understanding of the whole - genome scope of the phenomena we are observing as biologists. Epigenesis is after Strohmann (BECK et al, 1999) one system that provides explanations for the complex functional attributes of cells and organisms, a set of "informational systems and operating rules" that *complemen*t genes and genomes, by one part and proteins, by the other part. High throughput analysis of epigenomics is the only one able to unify and complete the functional genomics data into a more *global,* comprehensive approach, that is conducted in a systematic fashion. By such approaches, genomics, together with proteomics and completed by epigenomics has already begun to produce the results through the identification and characterization of individual genes and, recently, the pattern of gene expression in a normal and pathologic way, particularly of neoplastic cells. Not only is it now possible to characterize an individual for his health status, but one has the tools for performing the risk assessments and predicting the tendency of his own genome towards certain interactions with the environmental conditions or with certain chemotherapeutic strategies.

DNA (methylated) minor bases as the major molecular marks of the epigenetic information

Epigenomics performs its analysis on genetic variation of organisms through the new basic definition of gene: this is not only characterized by the „fixed" mutation into the genome, or more precisely, the genotype, which makes a clear correlation between the nucleotide and amino acid sequences, but its definition is completed by the so-called "space-temporal" determinism of its expression. The new term is designed to define factors which represent that information regarding the expression of the gene at the right moment of its development and at the same time in the right cell type or tissue.

This *additional information to* the **genetic** one is referred to as the **epigenetic information**. It is *heritable* like the genetic information, however it is *not static*, but more versatile and *dynamic,* as it represents the permanent link with the changing environment which is continuously shaping the individual behavior of organisms. In the meantime it leaves the primary DNA structure (the genetic information of nucleotide sequence which codes for a specific amino acid sequence and function of a protein) unchanged, but instead it characterizes the *apparently minor changes in the already established DNA sequence (such as the methylation* of its major bases). However, the so-called "minor" bases proved to have dramatic effects upon the above-mentioned spatio-temporal expression of genes. Such DNA modifications are **not mutations** but represent major **genome molecular marks** for the regulation of replication and transcription of genes and moreover, at the tissue and organism level, for the control of the correct, normal cytodifferentiation and development (RAZIN, RIGGS, 1980; RAZIN, 1984; Szyf, 1991)

As it has already been underlined, epigenesis is referred to as those activities of the informational macromolecules, nucleic acids and proteins, performed at the various stages of the gene expression machinery: transcriptional, posttranscriptional, translational and posttranslational. The changes in the pattern of these activities are imposed by both endogenous and exogenous determinants: on one side, the given **location of specific genome marks** which are defining a specific modification tendency towards the effect of various external agents and, on the other side, the context of specific environmental factors whose impact upon the cell is transduced up to the DNA level. Such DNA-environment communication is conducted through the above-mentioned steps of the epigenetic information, which permit the *tissue specific gene expression in a given context of external*

and endogenous factors (MONK, 1995)**.** Understanding the factors and the mechanisms acting at each level of epigenesis is now possible only by means of the information gathered from functional genomic large-scale analysis. The great benefit from such a global epigenetic approach, based on specific, precise genome and proteome elements, is the knowledge of the mechanisms underlying the cytodifferenciations and development of complex eukaryotic organisms.

The best known epigenetic signals residing into a genome are the methylation sites, which are represented by the above-mentioned minor bases. Among these, *5-methyl-cytosine* (5mC) is by far the most important cue of marking the genome for gene expression regulation (EHRLICH and WANG, 1981; BARTOLOMEI and TILGHMAN, 1997; SIEGFRIED and CEDAR, 1997; WALSH and BESTOR, 1999). This apparently minor base is derived from the major one, cytosine, by the very complex action of a special enzyme in epigenetics: *DNA methyltransferase*.

Despite the fact that 5mC had been detected long (about 50 years) ago into the eukaryote DNA (CHARGAFF, 1953; VANYUSHIN et al, 1970), the DNA methylation processes have been so far considered the "black box" of the gene expression domain in eukaryote organisms (SELKER, 1990), the controversial discussions linked to it still representing an important body of concern in current scientific articles. In the history of the DNA methylation domain many opinions on the biological role of the minor base 5mC into the eukaryote genome agreed that this one is involved in important cell processes, such as: the defence reaction, similar to the prokaryote restriction/modification system, against mobile elements- transposons and the viral/prokaryote pathogens, the marking of the heterochromatin silenced regions and through this idea, the imprinting and the X-chromosome suppression, similar to the ancestral DNA involvement into the *Chlamydomonas* maternal inheritance processes (SAGER et al, 1984).

The main arguments of those who doubtedly tackled this domain were linked to the so far elusive detection of the 5mC in the eukaryote experimental models of *Drosophila*, *Caernohabditis elegans* and yeasts. Indeed, in the past, numerous experiments have defined *Drosophila* as a standard organism without DNA methylation, which led to certain conclusions regarding the restriction of such minor DNA modification processes to the reduction of the so-called "transcriptional noise" in the eukaryote genomes. According to such opinions, the mechanisms by which the decrease in transcription of certain unnecessary genes may involve, in the absence of DNA methylation, only the chromatin modeling factors, like those acting in the very compacted *Drosophila* genome.

Yet, the recent discovery of the DNA methylation processes in the very early stages of *Drosophila* embryogenesis (LYKO, 2001) has suggestively indicated that the domain has no more obstacles for its development. The demonstration of an active and quite different developmental regulation of the DNA methylation processes in the fly, in the early stages of its embryogenesis, opened the possibilities of tackling the other mentioned model genomes by new technical and conceptual approaches of DNA modification processes and moreover to get more insights into their roles in gene expression with the eukaryote model *Drosophila*.

The 5mC and the DNMTase- the key elements in the epigenetic information

The biochemistry of the methylation reaction

At present it is well established that in addition to the four bases comprising the genetic information, a modified base, 5mC, that plays an important role in the epigenome may arrive in the DNA double helix during embryogenesis and later, during the cytodifferentiation

and specific adaptation to various environmental context. It is assumed that such DNA marking processes differentiate the gene expression pattern in different type of cells, in different developmental moments of live. Genes are composed of *subsequences of the DNA identical in each cell type* of the entire complex eukaryote organism. In *different cell types*, unique sets of genes are switched on, giving those cells a *unique and functional identity.* The action of 5mC marks in DNA is actually referred to maintaining as the two different features: the identity of the DNA subsets of genes throughout the organism and the establishment of a functional identity in different cell types, in specific environmental conditions. The action of such marks is therefore linked with *tissue specific gene expression*, which in turn is known to be regulated through the interaction of the different *chromatin conformations in a specific gene and the transcriptional machinery*. Noteworthy is the actual opinion regarding the causality of these processes: DNA is not the cause, but the effect of the action of other specific transcriptional factors, which are linked to the chromatin modeling. 5mC is now considered a *"molecular lock"* of the active/inactive chromatin conformations, which comprise the coordinated action of other remodeling elements (like the specific chromatin modifying proteins: histone acethylase/deacethylase or HAC/HDAC and the specific methylated DNA binding proteins or MBDPs) (EHRLICH and GAMA-SOSA, 1982; CAIAFA, 1995; NG et al, 1999; NG and BIRD, 1999; HENDRICH and BIRD, 1998; SUDARSANAM and WINSTON, 2000; AHRINGER, 2000) .

DNA cytosine methylation is a chemical modification of the cytosine residues into certain DNA sequences. The reaction occurs through the substitution of the 5C hydrogen from the cytosine pyrimidine ring by the single carbon methyl group (m). This chemical group is delivered by a specific donor, common for all types of organisms, either pro- or eukaryote ones, S-adenosyl methyonine (SAM or AdoMet). The reaction is catalyzed by an enzyme DNA cytosine methyltransferase (DNMTase) (EC 2.1.1.37) (Figure1a).

a.

b.

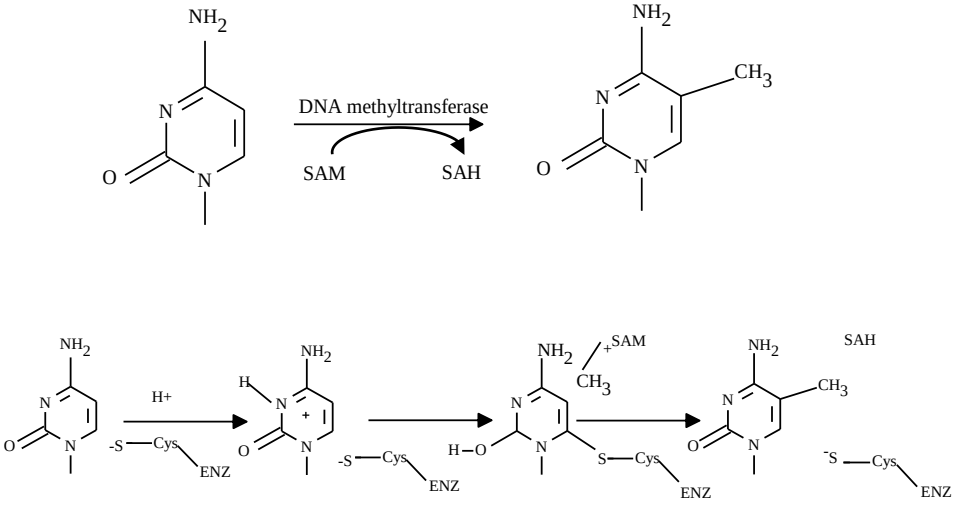

Figure 1a, b. (a)The general equation of the biochemical reaction of the cytosine residue methylation; (b)The mechanism of the SAM - methyl group transfer on C5 - cytosine residue.

The modification of cytosine residues in the DNA genome does not occur at random but it is coordinated with certain other nuclear processes. This base is unique, as it is not

inserted by the DNA polymerase into the nascent DNA string, from the bulk of free major bases trinucleotides. It is inherited in the nascent, unmethylated DNA sequence, by the postreplicative action of the enzyme DNMTase. Therefore, after the faithful replication of the DNA sequence, the nascent double helix DNA, formed from different, parental-methylated and new-unmethylated DNA chains, is subsequently modified. Specific cytosine residues of the sequence, corresponding to the distribution of the methylated cytosine residues on the parental DNA chain are targeted by the DNMTase, thus maintaining the methylation pattern from the initial replicated genome. Besides this passive, mitotic, process of methylation (which implies the common known DNMTase 1 or the so-called ”maintenance” methylase and a parental DNA template), there are two other active, replication independent ones, which do not require a DNA template: (i) one that methylates the cytosine residues in de novo sites, the so-called “de novo” methylase and (ii) another one, detected as a demethylase , involved into a methylated cytosine excision repair process (VANYUSHIN, 1984; OKANO et al, 1999; BHATTACHARAYA et al, 1999; RAMCHANDANI et al, 1999) .

The enzyme structure

As a true biochemical reaction, the cytosine methylation is dependent on the action of all its components: (i) the DNA substrate, represented by the cytosine residue embedded into certain secondary double stranded DNA structure, (ii) the SAM methyl group donor and (iii) the enzyme DNMTase. Indeed, certain DNA sequences, SAM concentration and the structure of the protein DNMTase strongly influence the biochemical reaction.

The enzyme reaction involves the following steps specifically controlled by the enzyme functional domains: (i) the recognition of the substrate by the enzyme, (ii) the activation by bonding of the donor, (iii) the bonding of the C6 of the cytosine pyrimidine ring through certain amino-acid components of a specific functional domain of the enzyme and the consequent rearrangements of the electronic pattern inside the ring, (iv) the methyl group cleavage from the SAM donor and its transfer to the activated C5 of the cytosine residue and finally, (v) the release and stabilization of the modified, methylated cytosine and the resulting S-adenosine homocysteine (SAH). The representation of such interactions is outlined in Fig.1b, reproduced after SMITH et al (1994) with the courtesy of the organizers of DNA methylation society web site.

It is obvious that the reaction catalyzed by the enzyme implies a specific orchestra of interactions between the *specific protein functional domains* and the *components of the reaction*. Indeed, the interesting structure of the DNMTase enables it to perform the so-called *multifaceted activity* on **DNA substrate** and, surprisingly, in certain cell-cycle important **protein-protein** interactions (SZYF, 2001). In Figure 2 the distribution of its specific functional domains and specification of their roles in the enzyme activity are represented (LEONHARDT et al, 1992; SZYF, 2001).

Understanding the methylation roles as molecular markers in eukaryote genome structure, chromatin, may be facilitated by a survey of the enzyme reaction features, comprising the substrate and reaction type specificity and the influence of the donor concentration or Km.

The substrate specificity of the enzyme reaction

It has been mentioned earlier that the true substrate for DNMTase is represented by the cytosine residue in a double helix DNA. This means that the activity of the enzyme depend on its ability to pass through the entire complex packaging of the chromatin in order to access the cytosine residues embedded in superstructures. The conformational structures of

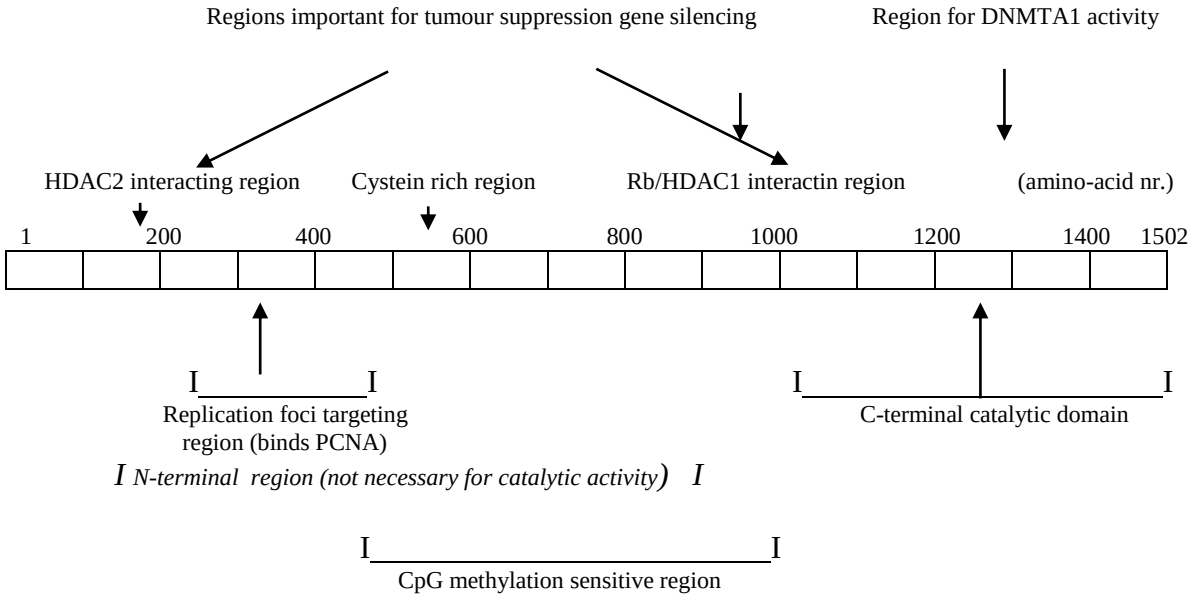

Figure 2. The protein structure of DNMTase; the distribution and roles of its functional domains (PCNA- proliferating cell nuclear antigen; HDAC-histone deacetylase; Rb- retinoblastoma protein tumor suppressor).

chromatin however are of much help for the enzyme activation and accessibility. This is confirmed by a consistent body of literature data which agreed that not all the cytosine residues are targeted by the enzyme for modification in their C5 position. Abnormal superstructures specifically activate DNMTase, at least like those in the so-called "hairpin" conformations, as such structures alter *the rigid parameters of the B-DNA double helix* and determine the *flipping out the cytosine residues* for performing on them the chemistry. It is obvious now that such altered structures are found in the preferential sequences, like those repetitive purine-pyrimidine repetitions (ROBERTS, 1955; Figure 3 and Figure 4).

Repetitive CpG sequences are therefore intensely claimed to represent the preferred substrate for the enzyme accessibility and action on the C residues. Mammalian and plant genome have been studied for such sequences, the results indicating indeed an intense methylation of the cytosine in the repetitive CpG sequences, which assures actually *symmetrical disposal* of the methyaltion spots in a double helix DNA. Recent findings indicated however that C in other, *asymmetrical repetitive sequences* might be methylated in mammals and plants. These are represented by trinucleotide, such as CAG repetitions, particularly in mammals and CpNpG repetitive sequences particularly in plants.

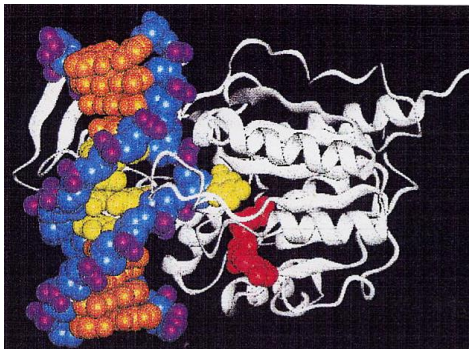

Figure 3. The van der Waals representation of the three component interaction involved in the biochemical reaction of DNA methylation : double helix DNA and its Cytosine residue flipped out of it (yellow), SAM (red) and enzyme (white) (Reproduced after the image offered by the courtesy of the organizers of the DNA methylation society website).

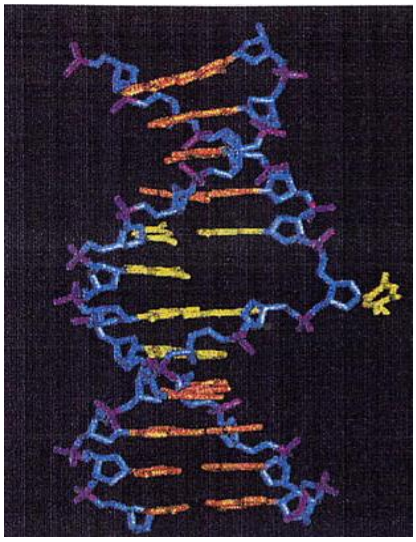

Figure 4. The perspective representation of the DNA structure showing the flipping out of its cytosine residue due to the constrains imposed by the CpG repetitive sequence (CHEN et al, 1991: CHENG, 1995).

Another important substrate linked feature of the enzyme reaction is that not all the C residues in the above-described sequences are methylated: there occurs a specific preferred attack by the enzyme of the so-called "critical" cytosine residues in such primary structures. The critical methylation sites are established during early embryogenesis and then transmitted during the subsequent stages of the development, assuring the cytodifferentiation, or the specific (clonal) transmission or inheritance of the established distributions of the methylation spots.

The earlier mentioned distribution of the 5mC through the repetitive DNA sequences represents the so-called "methylation pattern". This is a specific tissue and determines a specific regulation of gene expression in specific cell types, which provides an alternative explanation of the cytodifferentiation and tissue specific gene expression. Numerous literature data indicate a very complex activity of the eukaryote DNMTases in order to choose the right cytosine residues for methylation and the specific methylation pattern establishment. The initial data in the domain of DNA methylation through the eukaryote embryogenesis indicated the great importance of such chromatin marking processes for the subsequent correct transmission of the established pattern of methylation for the entire complex organism. It is actually dependent on the contributions of the involved gametes chromatin, marked in their turn by specific methylation patterns. The altered DNA methylation process during the embryogenesis proved to have dramatic effects upon the life and development of the individual (LI and coworkers, 1992).

It is therefore obvious why nature has focused on such processes for their accurate coordination through the multifaceted activity of a central protein in these vital scenes: DNA methyltransferase. Such complex activity implies on the one hand, the precise formation of the individual methylation pattern in its early embryogenesis and then, on the other hand, the faithful maintenance of a specific methylation pattern in the different cell lines forming a specialized tissue. Therefore, different DNMTase activities for the coordination of the two facets are required: one that targets cytosine residues on an unmethylated double helix (dh) DNA (the so-called "de novo" DNMTase) and one that reproduces the DNA methylation pattern from a template, hemimethylated dh DNA (the so-called "maintenance" DNMTase). Recently, there has been detected a demethylation DNMTase activity, similar to the DNA repair processes, in chicken and mouse (SZYF, 2001).

A very important feature of the DNMTase is that linked to another facet of its activity: the *mutagenic* one. Certain conditions of the biochemical reaction, as mentioned earlier, may not induce the *methyl group transfer* from the SAM donor, to the C5 of the cytosine pyrimidine ring, but rather trigger the *deamination* of either cytosine, or even already methylated cytosine residues. Such reaction specificity depends on the SAM pool: the deamination is favorised by the decreased SAM concentrations, as the results of the studies performed on cell cultures deprived of the donor has been shown (LAIRD, 1998). It is obviously that the impact on the nucleotide sequence results in this condition in transition mutations: C-U or mC-T. While the first one is easily detected and repaired by the DNA glycosylase eukaryote system, the second one has been intensively detected in tumorigenesis and explains the numerous cases of cell transformation (SHEN et al, 1992; EHRLICH et al, 1990; FREDERICO et al, 1993).

Therefore, numerous body of data are currently aiming at finding specific molecular DNA methylation marks in tumorigenesis, particularly in the study of the interaction of the environmental known and potential carcinogens and of the chemotherapeutic strategies with the individual genomes, particularly high risk or cancer prone. Such approaches might be very helpful further for the so called preventive medicine.

Having in mind that certain methylation forms of DNA may be derived in different moments of the DNA processing into the nucleus, such different enzyme activities can be explained by coupling the DNA methylation processes with the replicative and transcriptional functions of chromatin.

DNA replication

De novo and demethylase activities are independent of the DNA replication and have specificity for an unmethylated substrate, whereas the maintenance activity can be performed only on a hemimethylated, nascent dh DNA from the replicative process. The demethylase activity is an *active* one, instead, the *de novo* and *maintenance* activities entail a *passive* process, which depends on the specific *inhibition* of the enzyme.

The active, demethylase, and the passive, *de novo* DNMTase act during the early embryogenesis stage, for the establishment of the individual, unique, global methylation pattern, which does not need a methylation template. The maintenance enzyme is active during the subsequent steps of cytodifferentiation for the faithful clonal inheritance of the tissue specific DNA methylation patterns, and therefore depends on the mitotic, replicative processes.

One can withdraw the biological role of the DNMTase in the mentioned vital processes: the coordination of the two very accurate activities, the formation and replication of the methylation pattern. An alteration in such coordinated activities explains actually the derivation of the normal developmental paces towards the pathologic, particularly, transformed ones. Therefore, the detection of the methylation patterns and watching their faithful replication in specific tissues was assumed to be a good approach for identifying the marked genomic regions in disease molecular diagnostics. Moreover, the modulation of the enzyme's functional domains in order to assure a specific enzyme activity from the complexity of its multifaceted is also recently emerged approach in pharmacogenetics.

DNA transcription

One of the earliest indications that the pattern of methylation plays an important role in controlling gene expression was the remarkable observation of DNA methylation level inverse relationship with the transcriptional status of a gene. Furthermore, an interesting correlation between the distribution of methyl groups and the distribution of inactive chromatin has been proved. The hypothesis that the role of DNA methylation involves the

marking of inactive genes was supported by numerous data indicating that 5'regions of the active genes are hypomethylated, while those of the inactive ones are hypermethylated (SIEGFRIED et al, 1999).

The present concepts are agreed on the two, direct and indirect, mechanisms used by DNMTase to suppress the gene activity. First, methylation of a gene CpG site in a recognition sequence by the transcriptional factors may block the required interaction of the promoter with the RNA polymerase machinery. The discovery that interference with the transcription factors may not be determined only by the CpG sites, which are not present in the recognition regions of all genes, determined the elucidation of an alternative mechanism of the gene suppression by DNMTase. This last one implies the involvement of the earlier mentioned chromatin conformations in active or inactive forms (SZYF, 1991, 2000, 2001).

Chromatin is an important and dynamic regulator of transcription. Cellular signals can activate/inactivate certain genes by modeling the packaging of certain genes that are involved in structures that silence transcription. Misregulation of chromatin structure can cause incorrect gene activation or improper gene silencing. DNA methylation is one major factor that recruit chromatin-modifying complexes that silence transcription.

The methylation process of the cytosine residues in the genomic DNA is an apparent minor modification but having a tremendous potential for major modifications into the gene expression. One of the central tasks of the DNA methylation domain has been the elucidation of the causality with the silencing processes conducted by the chromatin modeling. It has recently been established that DNA **methylation is not the cause, but rather the consequence of the action of such chromatin remodeling, silencing factors**. The signals of silencing a wrong acting or a foreign gene are therefore the first to act on the part of the chromatin–environment signal transduction pathways. Such signals determine the activation of DNA methyl transferase, basically by specific, abnormal superstructures of the DNA double helix, like those hairpin formed ones. The activated DNMTase recruit the remodeling factors, acting as the so-called **„molecular lock"** of the suppressed conformations already established in the chromatin.

The conformational changes may be induced in chromatin by specific modifying factors for the counterpart of the DNA in the nucleo-protein complexes: the histones. Thus, the acetylation/deacetylation of these basic components of chromatin determine specific conformations in promoter or coding gene regions, which, in turn, interact with the DNMTase machinery. So, it has been proved that methylation of a region around a transcription regulatory site recruits proteins that bind methylated DNA regions (methylated DNA binding proteins or MBDP). These interact and activate the histone deacetylases (HDAC) that contribute on their part to the formation of inactive chromatin structures around the gene. Also, recently, two MBDPs have been detected whose activities may be linked also to a demethylase and, respectively, a thymidine glycosylase (SZYF, 2001) (Figure 2).

As the formation and stabilization of different active/inactive regions in chromatin need to be accurately perfomed, one can assume a very specific action of DNMTase and its interactions with the chromatin remodeling factors in this context. The important vital processes, like imprinting in gametes and embryogenesis are also linked to such perfect distribution of the active/inactive regions in chromatin, which have a dramatic effect upon the choice for a normal or pathologic development of the entire organism. The approaches of methylation footprints on the chromatin will yield valuable information on genome stability and overall chromatin packaging phenomena in certain environmental conditions.

Considering the facts that DNA methylation pattern in somatic cells is not static, but rather changing with the physiological signals, which are in turn determined by the

developmental stage, one can predict the complexity of the action of the above described DNMTase. A recent established idea is that the proper inheritance of the replication pattern of methylation is an outcome of the combined action of different DNMTs and the demethylase activities that are guided by the chromatin structure. It has proven that the above mentioned demethylase (repair) activity is determined by the state of acetylation or modification of the histone components of the chromatin. Both methylation pattern of DNA and the acetylation pattern of histones are inherited through the replication process. According to this hypothesis, the error in transmission of such patterns may arise if the epigenomic information embodied in the chromatin structure is changed. Knowledge of the mechanisms involved in such changes may be therefore be exploited for solving numerous medical or agricultural problems linked with the genetic stability of genes.

Conclusions

The summary presentation of the epigenomic information, essentially fundamented on DNA methylation processes was aiming at a better understanding of the preferential approach of this domain in the actual molecular genetic analyses, for a better scaling up of efficient strategies in functional genomics. These are only basic parts of the emerging concepts regarding the DNA methylation process involvement in deciphering epigenomics. The understanding of such fundamental ideas has lead however to the actual consideration of the methylation of chromatin in explaining the vital processes and suggesting new solution for solving important problems of pathologic, particularly neoplastic development. Noteworthy, numerous aspects commented in this review have been derived particularly from the human genome domain, as the main part of the foreign literature is abunding in such data and the major concern of the biotechnological and ethical problems is much concerned of the human health and rights. Nevertheless, the experimental models developed for DNA methylation study are based on animal and especially plant systems, the last ones being more attractive from an ethical point of view and mainly due to its extraordinary flexibility in terms of their tolerance to high variations in DNA methylation levels. Moreover, the molecular mechanisms of DNA methylation involvement in eukaryote development may be relevant to be deciphered in any of these models, as numerous factors implied in such complex processes have been conserved through evolution.

Aknowledgements: This is a review based on the updated Ph.D thesis dissertation presented in 2000. I thank for helpful advises Prof. Dr. Lucian Gavrila, the coordinator of my Ph.D. training program.

References

J. AHRINGER, *Trends in Genetics*, **16**(8), 351-355 (2000).

D. E. AYER, *Trends Cell Biol*, **9**, 193-198 (1999).

T. BESTOR, J. TYCKO, *Nat Genet*, **12**, 363-67 (1996).

M. S. BARTOLOMEI, S. M. TILGHMAN, *Annu Rev Genet*, **31**: 493-525 (1997).

B. P. BARLOW, *Science*, **260**, 309-10 (1998).

S. K. BATTACHARAYA, S. RAMCHANDANI, N. CERVONI, M. SZYF, *Nature*, **397**, 579-83 (1999).

S. BECK, A. OLEK, J. WALTER, *Nature Biotechnology*, **17**, 1144 (1999).

A. P. BIRD, *Trends in Genetics*, **11**, 94-100 (1995).

A. P. BIRD, A. P. WOLFFE, *Cell*, **99**, 451- 454 (1999).

V. H. BROPHY, G. P. JARVIK, R. J. RICHTER, L. S. ROZEK, G. D. SCHELLENBERG, C. E. FURLONG, *Pharmacogenetics*, **10**(5), 453-60 (2000).

P. CAIAFA, A. REALE, *Gene*, **157**(1-2), 247-51 (1995).

B. R. CAIRNS, *Trends in Cell Biology*, **11**(11), S15-S21 (2001).

E. CHARGAFF, C. F. CRAMPTON, R. LIPSCHITZ, *Nature*, **172**, 289-292 (1953).

I. CHEN, A. M. MAC MILLAN, W. CHANG, Y. K. EZAZ NIKPA, W. S. LANE, G. L. VERDINE, *Biochemistry*, **30**, 11018-11025 (1991).

X. CHENG, *Annu Rev Biophys Biomolec Structure*, **24**, 293-98 (1995).

J. M. CRAIG, W. A. BICKMORE, *Nature Genetics*, **7**, 376-381 (1994).

M. EHRLICH, R. Y. H. WANG, *Science*, **212**, 1350-1357 (1981).

M. EHRLICH, M. A. GAMA-SOSA, *Nucl Acids Res*, **10**, 2709-2721 (1982).

M. EHRLICH, X. Y. ZHANG, N. M. INAMADAR, *Mut Res*, **238**, 277-286 (1990).

M. ESTELLER, P. G. CORN, S. B. BAYLIN, J. G. HERMAN, *Cancer Res*, **61**, 3225-3229 (2001).

M. FAGARD, H. VAUCHERET, *Annu Rev Plant Physionl Plant Mol Biol*, **51**, 167-194 (2000).

E. J. FINNEGAN, K. A. KOVAC, *Plant Mol Biol*, **43**, 189-201 (2000).

L. A. FREDERICO, T. A. KUNKEL, B. R. SHAW, *Biochemistry*, **32**, 6253-60 (1993).

F. FUCKS, *Nat Genet*, **24**, 88-91 (2000).

B. HENDRICH, A. BIRD, *Mol Cell Biol*, **18**, 6538-47 (1998).

S. E. JACOBSEN, *Curr Biol*, **9**, 617-619 (1999).

B. JANOUSEK, J. SIROKY, B. VYSKOT, *Mol Gen Genet*, **250**(4), 483-90 (1996).

G. P. JARVIK, E. WIJSMAN, *ASHG **1***, OCT 13 (2001).

P. JONES, *Trends in Genetics*, **15**, 34-37 (1999).

P. W. LAIRD, R. JAENISCH, *Annu Rev Genet*, **30,** 44-64 (1998).

P. W. LAIRD, R. JAENISCH, *Human Molec Genet*, **3**, 1487-95 (1999).

H. LEONHARDT, A. W. PAGE, H. U. WEIER, T. H. BESTOR, *Cell*, **71**, 865-73, (1992).

E. LI, *Cell*, **69**, 915- 926 (1992).

F. LYKO, *Trends in Genetics*, **17**(4), 169-172 (2001).

R. MARTIENSSEN, *Curr Biol*, **6**, 810-813 (1996).

R. A. MARTIENSSEN, E. J. RICHARDS, *Current Opinion in Genetics and Development*, **5**, 234-242 (1995).

D. B. MARTIN, P. S. NELSON, *Trends in Cell Biol*, **11** (11), S60-S65 (2001).

M. A. MATZKE, A. J. P. MATZKE, *Trends in Gen*, **11**(1), 1-3 (1995).

M. A. MATZKE, A. J. P. MATZKE, *Plant Physiol*, **107**, 679-685 (1995).

P. MEYER, *Euphytica*, **85**, 359-366 (1995).

P. MEYER, I. NIEDENHOF, M. TEN LOHUIS, *EMBO J*, **13**, 2084- 2088 (1994).

M. MONK, *Dev Genet*, **17**(3), 188-97 (1995).

Y. NABU, T. KAKUTANI, J. PASZKOWSKI, *Curr Opin Gen Dev*, **11**, 247-257 (2000).

H. H. NG, Y. ZHANG, B. HENDRICH, C. A. JOHNSON, B. M. TURNER, H. ENDJUMENT-BROMAGE, P. TEMPST, D. REINBERG, A. BIRD, *Nat Genet*, **23**, 58-61 (1999).

H. H. NG, A. BIRD, *Curr Opin Genet Dev*, **9**, 158-63 (1999).

M. OKANO, D. W. BELL, D. HABER, E. LI, *Cell*, **99**, 247-57 (1999).

J. PASZKOWSKI, S. A. WHITMAN, *Current Opinion in Plant Biology*, **4**, 123-129 (2001).

S. PRADHAN, URWIN NAR, G. I. JENKINS, R. L. P. ADAMS, *Biochem J*, **341**, 473-476 (1999).

S. RAMCHANDANI, S. K. BHATTACHARAYA, CERVONI, M. SZYF, *Proc Natl Acad Sci USA*, **96**, 6107-6112 (1999).

A. RAZIN, A. D. RIGGS, *Science*, **210**, 604-610 (1980).

A. RAZIN, *in DNA Methylation: Biochemistry and Biological Significance*, A.Razin, H.Cedar, A. D. Riggs eds., Springer Verlag, p.127, 1984.

A. RAZIN, H. CEDAR, *Cell*, **77**, 473-476 (1994).

A. RAZIN, T. KAFRI, *Prog Nucl Acids Res Mol Biol*, **48**, 53-81 (1994).

A. RAZIN, R. SHEMER, *Hum Mol Genet*, **4**, 1751-5 (1995).

A. D. RIGGS, *Cytogen Cell Genet*, **14**, 9-25 (1995).

R. J. ROBERTS, *Cell*, **82**, 9-12 (1995).

K. D. ROBERTSON, A. P. WOLLFE, *Nature Rev*, **1**, 11-19 (2000).

M. R. ROUNTREE, K. E. BACHMAN, S. B. BAYLIN, *Nature Genetics*, **3**, 269-2777 (2000).

R. SAGER, H, SANO, C. T. GRABOWY, *Curr Top Microb Immunol*, **108**, 157-173 (1984).

E. U. SELKER, *TIBS*, **3**, 103-107 (1990).

J. C. SHEN, W. M. RIDEOUT, P. A. JONES, *Cell*, **71**(7), 1073-1080 (1992).

Z. SIEGFIRED, S. EDEN, M. MENDELSOHN, X. FENG, B. Z. TSUBERI, H. CEDAR, *Nat Genet*, **22**, 203-6 (1999).

S. SMITH, L. NIU, D. J. BAKER, *Proc Natl Acad Sci USA*, **94**(6), 2162 (1994).

P. SUDARSANAM, F. WINSTON, *Trends in Genetics*, **16**(8), 345-351 (2000).

M. SZYF, *Ann NY Acad Sci*, **910**, 156-174 (2000).

M. SZYF, *Biochem Cell Biol*, **69,** 764-767 (1991).

M. SZYF, *Frontiers in Bioscience*, **6** (1), 599-609 (2001).

M. SZYF, *TIBS*, **15**, 233-238 (1994).

B. F. VANYIUSHIN, S. G. TKACHEVA, A. N. BELOZERSKY, *Nature* , **225**, 948-949 (1970).

M. VAIRAPANDI, N. J. DUKER, *Nucl Acids Res*, **21**, 5323-27 (1993).

J. C. VENTER, *Science*, **291**, 1304-1351 (2001).

J. T. WACHSMAN, *Mutation Res*, **375**, 1-8 (1997).

C. P.WALSH, T. H. BESTOR, *Genes Dev*, **13**, 26-34 (1999).

S. A.WEITZMAN, S. CEDAR, *Mutation Research*, **386**, 141-152 (1997).

S. WILLIAMS, *Sources (UNESCO)*, **132**, 3 (2001).

M. G. YEBRA, A. S. BHAGWAT, *Biochem*, **34**, 14752-57 (1995).

M. L.YASPO, *Trends in Molecular Medicine*, **7**(11), 494-500 (2001).

F. YU, *Nucleic Acid Res*, **28**, 2201- 2206 (2000).