CMLS, Cell. Mol. Life Sci. 60 (2003) 1775–1178
1420-682X/03/091175-4
DOI 10.1007/s00018-003-23147-z

CMLS Cellular and Molecular Life Sciences

Visions & Reflections

Stochastic gene expression during cell differentiation: order from disorder?

A. Paldi

Institut Jacques Monod, INSERM E0367, and École Pratique des Hautes Études, 2, place Jussieu, 75005 Paris (France), e-mail: paldi@ijm.jussieu.fr

Received 17 April 2003; received after revision 15 May 2003; accepted 19 May 2003

Understanding cell differentiation in multicellular organisms remains one of the central questions of biology. According to the prevailing view of contemporary developmental biology, differentiation of multicellular organisms is based on precisely regulated communication between cells via diffusible molecules or direct cell-to-cell contacts. These signals are the basis of embryonic induction, where one cell instructs others to adopt a particular developmental fate. A molecular signal is supposed to initiate differentiation by specifically activating one or several regulatory genes, which, in turn, also specifically activate other downstream regulator and/or effector genes. The activation of such a regulatory cascade leads to the expression of a new set of genes that progressively guides the cell toward commitment into a lineage and ultimately to its differentiated state. This process of hierarchically ordered and sequential expression of genes is usually referred to as a 'genetic program'. The role of the regulatory genes is, of course, crucial, since they code for the 'instructions' of the program in the form of transcription factors that are able to bind nucleotide sequence motifs in the regulatory regions of the target genes and initiate the process of transcription by recruiting the other components of the transcription machinery.

The *elementary event* of the whole process is the activation of a previously inactive individual gene. How does the activation occur? The current view is based on the idea that the default state of a gene is repressed and it is targeted for activation by specific factors [1]. On the other hand, assembly of the DNA into chromatin makes the interaction of the transcription complex proteins with the promoter of an inactive gene impossible because the stable chromatin structure prevents access of the proteins to the target sequence. This point is crucial to explain gene activation and has been explicitly identified in the literature [2]. To be transcribed, the chromatin must first be remodeled to make the DNA accessible to the proteins of the transcriptional machinery. How might they recognize the promoter of the gene to be activated? According to current thinking '… one of the key roles of gene-specific transcriptional activators is to target promoter regions for the unfolding or remodelling of chromatin structure' [2]. However, what is true for the components of the general transcription initiation complex must also be true for specific transcription factors. For a sequence-specific transcription factor protein to bind to its target sequence, the chromatin around the recognition site must be remodeled, but the chromatin remodeling cannot be targeted specifically to a particular sequence without first recognizing it. Thus, the current thinking leads to a 'catch 22' situation frequently encountered by deterministic models. The process of gene activation in differentiating cells is usually studied using methods of molecular biology (e. g. Northern blots, RT-PCR or microarrays), which require large cell populations (from hundreds up to several millions of cells). Under these conditions, a gradual change in the transcript level following stimulation is usually interpreted as regulation of transcription. However, such observations only reflect the average steady-state level of the gene transcript in the whole population, not the actual transcription status of a single gene. Nevertheless, the conclusions are systematically drawn at the level of a single gene copy in an individual cell. The resulting models that aim to explain the regulation of genes are typically presented as simple cartoons built up of a solid line representing the gene with different regulatory

or coding sequence motifs as boxes on it, and of circles and squares indicating the different protein molecules. The possible interactions between these players are qualitatively shown by arrows, the quantitative characteristics are usually ignored.

A conceptually different model of how a previously inactive gene is activated has emerged from recent single-cell studies. In vivo fluorescent microscopy studies have revealed an intrinsically dynamic behavior of nuclear proteins and other molecules [3]. Nuclear components are constantly challenged by random Brownian motion of small molecules, which results in the energy-independent high mobility of all micro- and macromolecules in the nucleus. This constant 'vibration' is critical for their function. The rapidity of the movement is slowed down by repeated transient interactions of the molecules with each other or by the less mobile structures and, thus, depends on the frequency and the strength of these interactions. The overall result is a constant flux of all nuclear components between different compartments. The morphologically observable nuclear compartments are, in fact, highly dynamic and represent the steady-state equilibrium between the on and off rate of the components. Of course, the functional status of the proteins and other molecular components determines the composition and the morphology of the compartments that can be considered as self-organizing structures. Smaller-size structures such as chromatin are also formed by a steady-state in and out flux of their various components. Chromatin proteins dissociate from their binding sites on the DNA or on other proteins, diffuse and bind again if they encounter other free binding sites. As a result, individual protein molecules constantly roam the nuclear space. The average retention time of a molecule on its binding site is approximately several seconds for high-mobility group (HMG) proteins, several minutes for H1 histones, more than 15 min for heterochromatin protein1 (HP1), and up to several hours for core histones. Therefore, even a stable structure such as heterochromatin is, in fact, highly dynamic and differs from the less stable euchromatin only by the lower exchange rate of its constituents.

It is this energy-independent, highly dynamic and continuous exchange of protein building blocks that makes the structural remodeling of chromatin around a gene promoter possible without any specific targeting mechanism. When a site becomes exposed by the spontaneous dissociation of a nucleosome or chromatin-binding proteins, transcription factors can use the opportunity to gain access to DNA and bind to it. Such opportunities arise frequently in the euchromatin. Although less frequent, they also occur constantly even in heterochromatic regions. Since the time window of opportunity created by the dissociation of a chromatin-forming protein is short and many different proteins can bind to the same site, the competition is rough and the success rate depends on availability, i.e. the concentration of a given component [4, 5]. The concentration varies widely. The total number of protein molecules in a cell may be lower than the number of potential binding sites [6]! The higher the concentration (availability) of a factor, the higher the probability that it can use the opportunity to bind to its target site. The bound activator can, within its residence time, recruit other available factors from the environment. If no partner is recruited, the factor dissociates. Although cooperative interactions may facilitate the process, all successive steps of transcription complex assembly follow the same stochastic logic.

High mobility by diffusion is a simple and economic way for any molecule to reach any site in the nucleus and find its target within a very short time without any specific targeting or signal recognition. This is a random process. More important, the combination of high mobility and high exchange rate of different molecules forming the chromatin and the functional enzymatic complexes acting on it enables the initiation of transcription anywhere in the genome. The assembly of an active transcription initiation complex is an intrinsically stochastic process making the *transcription of a single gene a probabilistic event*. Transcriptional activation of any previously silenced gene is possible and depends on the dissociation-association rate of the chromatin molecules on the one hand and the level of transcription-activating factors on the other.

The probabilistic model of gene expression is fully supported by all single-cell studies of gene expression [for a review see ref. 7]. The main conclusion from these studies is that the level of gene expression in a tissue or cell population can be described in terms of probability of transcription of the gene in individual cells. Genes with a high probability of transcription are frequently transcribed in an individual cell and are detected as 'highly' expressed in a cell population because they are simultaneously expressed in many cells. The situation is the opposite for 'low'-expressing genes. The average messenger level is low in a cell population because the number of cells transcribing the gene at the same time is low and not because of a low level of transcription in all cells.

What is the relevance of these observations to cell differentiation? In general, the entire process of differentiation is accompanied by changes in the expression probability of many genes in each cell and ends with a state where most of the cells express the genes typical for the differentiated phenotype.

On the basis of our understanding of the probabilistic nature of gene activation, a new model for cell differentiation can be proposed that is devoid of the circular reasoning of 'which came first'. The gradual change in the expression level of a gene in a differentiating cell population reflects the increase in the probability of its expression in the individual cells. Differentiation is usually in-

duced by a substantial change in the immediate microenvironment of the cells that may include e.g. alteration of the concentration or nature of nutrients, altered cell-to-cell contacts, physical deformation or changes in oxygen concentration. Importantly, the cell itself constantly contributes to the change of its own environment, e.g. by secreting metabolites and extracellular matrix substances or by interacting with other cells. The consequence of these changes is that the physiological state of the cell is no longer adapted to its environment. To optimize cell function under the new conditions, expression of new physiologic functions is required. For example, activation of new metabolic pathways may be necessary to metabolize new substrates. Synthesis of new surface molecules may be required to improve cell-to-cell communication or to induce e.g. structural changes or migration. From this point of view, induction of differentiation is reminiscent of a stress, and differentiation is a response to it. Although the cells of a multicellular organism are not usually considered as autonomous units, but simply as the instruction-executing part of the whole organism, this logic is widely accepted for the explanation of life phase transitions of simple organisms such as slime molds or unicellular organisms. Without a doubt, chemical signaling plays a key role in the detection of environmental changes by a cell, but, as I argued above, targeted activation of specific genes is unlikely. The best strategy for 'finding' new genes to be expressed is a 'random walk.' A non-specific increase in the basal level of gene expression due to acceleration in mobility of chromatin constituents provides an opportunity for transcription for all genes, even for those previously silenced. During the initial phase following induction, each cell in the population will display individual patterns of gene expression that result from the random combinations of the 'on' or 'off' states of many genes. If the expression of a new gene contributes to the better adaptation of the cell to the new requirements, its transcription is stabilized. If a new gene product has no effect, its expression ceases. In this way, using a random strategy, each cell is able to find the combination of genes whose expression is necessary for the better adaptation to the new environment. When examined at the level of the entire cell population, a gradual increase in the expression level of these genes is detected in parallel with the process of differentiation, giving the impression of a tightly regulated change.

How can a cell first increase the basal level of gene expression non-specifically then stabilize a new combination of expression profiles? The key role in both phases belongs to epigenetic chromatin modifications. Since the probability that a DNA sequence will be expressed depends on the exchange rate of the molecular components of the chromatin around it, any cellular process which increases or decreases the exchange rate will influence gene transcription. Indeed, 'epigenetic' covalent modifications such as methylation, acetylation, phosphorylation, poly-ADP-ribosylation of histones and other chromatin components, and also CpG dinucleotide methylation in the DNA all act on the stability of the chromatin structure through altering the strength of interactions between the different components [8]. Both transcriptionally active and inactive chromatin regions have typical 'epigenetic signatures.' Some combinations of these modifications favor the stability of the chromatin structure typically found in heterochromatin (i.e. low component exchange rate), while other combinations promote the high mobility of the nucleosomes usually found in euchromatin. Therefore, the difference between heterochromatin and euchromatin is not static but dynamic, and resides in the difference in the exchange rate of their components. In other terms, the frequency by which transcription is initiated on a promoter depends largely on the epigenetic modifications of the regulatory regions. This property is illustrated by the example of imprinted genes. These genes are characterized by different combinations of epigenetic modifications on the two parental alleles within the same cell [9]. Single-cell studies of imprinted-gene transcription have shown that transcription is initiated randomly on both alleles but with a different frequency [10]. Epigenetic modifications are usually reversible and both the forward and reverse reactions are catalyzed by enzymes which are specific only for the substrates, and not for the DNA sequence. These reactions are often mechanically linked and form a network with feedback and feed-forward loops. The network of epigenetic modification reactions incorporates all the properties of a network, such as multistability and robustness, that provide the means for epigenetic memory and inheritance [8, 11].

As outlined above, transcriptionally active and inactive chromatin regions both have a typical 'epigenetic signature' representing by and large two stable equilibrium states of the network of various modifying reactions acting on chromatin proteins. Interestingly, the substrates used for epigenetic modifications are all key molecules in the basic energy metabolism of the living cell [12]. The substrate for acetylation of proteins is acetyl-CoA. The universal methyl donor for DNA and protein methylation is S-adenosyl-methionine, which is synthesized from the essential amino acid methionine and ATP. ATP is essential for the activity of chromatin-remodeling complexes, which maintain the open chromatin structure of actively transcribed genes. NAD^+ is the substrate for poly-ADP-ribosylation, an essential covalent modification of many nuclear proteins. The intracellular concentration of these molecules is a reliable indicator of the metabolic state of the cell. Since the rate of enzymatic reactions depends primarily on the substrate concentration, epigenetic modifications of the chromatin are highly likely to be strongly influenced by the actual energetic state of the cell. For ex-

ample, starving cells have a high concentration of acetyl-CoA (derived from the degradation of lipid reserves) and a high concentration of NAD^+, but a lower than normal ATP concentration and a lack of methionine. As a result, the rate of epigenetic modifications that promote chromatin mobility will increase and the rate of stabilizing modification will decrease in a DNA sequence-independent fashion. This shift in the equilibrium of epigenetic modifications will aid the emergence of new gene expression patterns. If metabolic homeostasis of the cell is re-established, the 'epigenetic equilibrium' will also shift in favor of the stabilizing reactions making the activation of new genes less likely. Activity of the ATP-dependent chromatin remodeling enzyme complexes will contribute to the maintenance of the transcriptional activity of the already active genes.

Without doubt, the real situation is much more complex than the above-depicted simple model. Nevertheless, the metabolic link between epigenetic modifications and gene expression regulation seems evident. The model is supported by the fundamentally stochastic logic underlying transcription of individual genes and by our present knowledge of epigenetic mechanisms. It allows developmental biologists to consider cell differentiation without invoking a predetermined specific mechanism. It is supported by experimental observations at the level of single cells [see for example refs 13–15] and is in accordance with the apparently regulated changes of gene expression observed at the level of whole cell populations.

The inherent logic of the model is that of a simple selection-adaptation process. The first step of differentiation can be characterized by an increase in intercellular variation, and then the most adapted variants are selected. Interestingly, a theoretical model of cell differentiation based on Darwinian selection was first proposed 20 years ago [16, 17] and the stochastic nature of individual gene expression has been known for more than 10 years [18]. The validity of the variation/selection model is well accepted for some particular cases, such as the differentiation of antibody-producing lymphocytes [19]. In the light of recent progress in our understanding of mechanisms underlying gene expression and cell differentiation, stochastic processes may play a more important role in the development of individual living organisms than has been previously suspected. Erwin Schrödinger in his landmark work, *What is Life* suggested that the apparently tightly regulated macroscopic order of life is based on the microscopic order of its constituents. Half a century after the publication of this work, the deterministic view still dominates biology. However, we are starting to realize that 'we have been Newtonians for the past several decades in our thinking about gene action. It is time to become Darwinian' [20]. Is it possible that in biology also, just as in the physical world, macroscopic order is based on the stochastic disorder of its elementary constituents?

Acknowledgements. I thank my colleagues T. Heims, T. Imamura, J.-J. Kupiec, S. Saint-Juste and P. Sonigo for the helpful discussions and critical reading of the manuscript.

1 Struhl K (1999) Fundamentally different logic of gene regulation in eukaryotes and procaryotes. Cell **98:** 1–4
2 Fry C and Peterson C (2001) Chromatin remodeling enzymes: who's on first? Curr. Biol. **11:** R185–R197
3 Mistelli T (2001) Protein dynamics: implications for nuclear architecture and gene expression. Science **291:** 843–847
4 Henikoff S (1996) Dosage-dependent modification of position-effect variegation in Drosophila. Bioessays **18:** 401–409
5 Ahmad K and Henikoff S (2001) Modulation of transcription factor counteracts heterochromatic gene silecing. Cell **104:** 839–847
6 Nan X, Campoy F and Bird A (1997) MeCP2 is a transcriptional repressor with abundant binding sites in genomic chromatin. Cell **88:** 471–481
7 Hume D (2000) Probability in transcriptional regulation and its implication for leukocyte differentiation and inducible gene expression. Blood **96:** 2323–2328
8 Ahmad K and Henikoff S (2002) Epigenetic consequences of nucleosome dynamics. Cell **11:** 281–284
9 Paldi A (2003) Genomic imprinting: could the chromatin structure be the driving force? Curr. Topics Dev. Biol. **53:** 115–138
10 Jouvenot Y, Poirier F, Jami J and Paldi A (1999) Biallelic transcription of Igf2 and H19 in individual cells suggests a post-transcriptional contribution to genomic imprinting. Current Biology **9:** 1199–1202
11 Schreiber S and Bernstein B (2002) Signaling network model of chromatin. Cell **111:** 771–778
12 Lehninger A (1975) Biochemistry, 2nd ed., Worth, New York
13 Hu M, Krause D, Greaves M, Sharkis S, Dexter M, Heyworth C et al. (1997) Multilineage gene expression precedes commitment in the hemopoietic system. Genes Dev. **11:** 774–785
14 Levsky J, Shenoy S, Pezo R and Singer R (2002) Single-cell gene expression profiling. Science **297:** 836–840
15 Heams T and Kupiec J (2003) Modified 3′-end amplification PCR for gene expression analysis in single cells. BioTechniques **34:** 712–716
16 Kupiec J (1997) A darwinian theory for the origin of cellular differentiation. Mol. Gen. Genet. **255:** 201–208
17 Kupiec J (1983) A probabilist theory for cell differentiation, embryonic mortality and DNA C-value paradox. Specul. Sci. Technol. **6:** 471–478
18 Ko M (1992) Induction mechanism of a single gene molecule:. Bioessays **14:** 341–346
19 Lederberg J (1988) Ontogeny of the clonal selection theory of antibody formation. Ann. N. Y. Acad. Sci. **546:** 175–187
20 Greenspan R (2001) The flexible genome. Nat. Rev. Genet. **2:** 383–387