Gene and Genon Concept: Coding versus Regulation -- A Conceptual and Information-Theoretic Analysis of Genetic Storage and Expression in the Light of Modern Molecular Biology

Klaus Scherrer
Juergen Jost

SFI WORKING PAPER: 2007-08-018

SANTA FE INSTITUTE

GENE AND GENON CONCEPT : CODING VERSUS REGULATION

A CONCEPTUAL AND INFORMATION - THEORETIC ANALYSIS

OF GENETIC STORAGE AND EXPRESSION

IN THE LIGHT OF MODERN MOLECULAR BIOLOGY

Essay

Klaus Scherrer[1] (Paris) and Jürgen Jost[2] (Leipzig)

1. Institut Jacques Monod, CNRS and Univ. Paris 7, Paris, France
2. Max Planck Institute for Mathematics in the Sciences, MPI MIS, Leipzig, Germany

Correspondence to : Klaus Scherrer, Institut Jacques Monod, CNRS and Univ. Paris 7, 2, place Jussieu, F-75251 Paris-Cedex 5, France. Tel./ Fax : +33 1 4707 5231, E-mail scherrer@ijm.jussieu.fr

Correspondence to : Jürgen Jost, Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany. Tel.: +49-341-9959-550, Fax: +49-341-9959-555, E-mail: jost@mis.mpg.de

Submitted to "Theory in Biosciences"

Table of Contents

Abstract

We analyse here the definition of the *gene* in order to distinguish, on the basis of modern insight in molecular biology, what the gene is coding for, namely a specific polypeptide, and how its expression is realized and controlled. Before the coding role of the DNA was discovered, a gene was identified with a specific phenotypic trait, from Mendel through Morgan up to Benzer. Subsequently, however, molecular biologists ventured to define a gene at the level of the DNA sequence in terms of coding. As is becoming ever more evident, the relations between information stored at DNA level and functional products are very intricate, and the regulatory aspects are as important and essential as the information coding for products. This approach led, thus, to a conceptual hybrid that confused coding, regulation and functional aspects. In this essay, we develop a definition of the gene that once again starts from the functional aspect. A cellular function can be represented by a polypeptide or an RNA. In the case of the polypeptide, its biochemical identity is determined by the mRNA prior to translation, and that is where we locate the gene. The steps from specific, but possibly separated sequence fragments at DNA level to that final mRNA then can be analysed in terms of regulation. For that purpose, we coin the new term "genon". In that manner, we can clearly separate product and regulative information while keeping the fundamental relation between coding and function without the need to introduce a conceptual hybrid. In mRNA, the program regulating the expression of a gene is superimposed onto and added to the coding sequence in cis – we call it the *genon*. The complementary external control of a given mRNA by trans-acting factors is incorporated in its *transgenon*. A consequence of this definition is that, in eukaryotes, the gene is, in most cases, not yet present at DNA level. Rather, it is assembled by RNA processing, including differential splicing, from various pieces, as steered by the genon. It emerges finally as an uninterrupted nucleic acid sequence at mRNA level just prior to translation, in faithful correspondence with the amino acid sequence to be produced as a polypeptide. After translation, the genon has fulfilled its role and expires. The distinction between the protein coding information as materialised in the final polypeptide and the processing information represented by the genon allows us to set up a new information theoretic scheme. The standard sequence information determined by the genetic code expresses the relation between coding sequence and product. Backward analysis asks from which coding region in the DNA a given polypeptide originates. The (more interesting) forward analysis asks in how many polypeptides of how many different types a given DNA segment is expressed. This concerns the control of the expression process for which we have introduced the genon concept. Thus, the information theoretic analysis can capture the complementary aspects of coding and regulation, of gene and genon.

(1) Introduction

The concept of the gene was introduced before the onset of molecular biology, in the wake of the work of Mendel (Mendel, 1866) and Morgan (Morgan et al., 1915) and their successors in the early 20th century. At that time, it meant a basic unit of heritable phenotypic properties (Johannsen, 1909) (cited in (Roll-Hansen, 1989)). Molecular biology identified the structures underlying these properties, that is, the molecules coding for or carrying out specific functions. The phenomenal success of modern biochemistry and molecular biology, however, rather than clarifying the relationship between inheritance, coding and function, eventually led to confusion about the basic concept, the gene. Most investigators today tend to identify a gene with a certain, more or less contiguous stretch of DNA that codes for some specific functions. When looking at the biochemical details, however, this practice becomes rather contorted, with all kind of exceptions and twists, and is, as we shall argue in this paper, problematic not only on practical, but also on conceptual grounds.

Originally, before the molecular carriers of function were understood and the coding aspect came to the foreground, a gene had been conceived as a simultaneous unit of inheritance, mutation and function. The principles of mutation are easy to understand at the biochemical level. The basic type of mutation is the exchange of a single nucleotide in the DNA. A single nucleotide, however, is too small to count as a unit of function. Such mutations may affect one or several functions and play a role in cis/trans tests; but in most cases, they are neutral at the functional level. Other genetic mutations consist of deletions, insertions, transpositions, inversions, or duplications. Again, they involve a certain piece of DNA, whose location and size, however, need not respect any unit of function or regulation, and therefore, they are not necessarily related to a specific phenotypic trait. For these reasons, the gene cannot be characterised as a unit of mutation. Moving the discussion to units of inheritance, first of all, there is the fact that, under asexual reproduction, entire genomes are reproduced (faithfully, unless mutations occur), and thus, here the basic unit would be too large. Under sexual recombination, it seems that units of recombination tend to respect functional boundaries, but this is not inherent in its basic mechanism, but rather represents a secondary adaptation. Furthermore, modern molecular biology essentially focuses on biological function and not on inheritance. Therefore, in our gene concept, we shall concentrate our analysis on the functional aspects. Of course, we realize that in the perspective of evolutionary biology, different conceptual emphasis has lead to utilizations of the term “gene” that are different from ours.

Looking at some, apparently rather authoritative example in the wake of the genome sequencing project (Snyder and Gerstein, 2003) defined a gene as ”a complete chromosomal segment responsible for making a functional product” and then discusses *five* criteria for identifying genes in the DNA sequence of a genome, open reading frame, sequence features (like codon bias), sequence conservation, evidence for transcription, and gene inactivation (the possibility for mutating or inactivating the product by direct gene disruption or RNA interference). These criteria are rather heterogeneous, and such a gene concept could have at best a heuristic value (c.f. also (Griffiths and Stotz, 2006)). More recently, the ENCODE project (ENCODE Project Consortium, 2007) shifted the emphasis from the DNA sequence to the collection of transcripts, and besides the discovery of many transcripts of unknown or at least non-protein-coding function, the intricacies of the regulation process came into focus again. This also led to a redefinition of the gene in (Gerstein, 2007)[1] as “a union of genomic sequences encoding a coherent set of potentially overlapping functional products”. In other words, it is realized that there is no one-to-one relation between a coding sequence at DNA level and a functional product. In this situation, the quoted definition then abandons both the gene as a coding and as a functional unit, and entirely suppresses the regulatory effects that mediate between those two aspects. This may be acceptable for some purposes; however, it is the aim of the present paper to argue for strict conceptual definitions. Before starting that enterprise, let us go further back in history, in order to better appreciate the difficulties involved in clarifying the gene term.

In the beginning of modern molecular biology, genetically identified functions could be related first to polypeptides and then to DNA (Hershey and Chase, 1955). Seymour Benzer (Benzer, 1959; Benzer, 1961; Benzer and Champe, 1961) then introduced the concept of the *cistron* (contiguous genomic elements acting in cis, essentially the protein coding sequence), a concept to be extended by Jacob and Monod (Jacob and Monod, 1961). This related the gene to an un-interrupted piece of DNA, able to complement a function in a cis/trans test. On that basis, the identification function = gene = polypeptide = continuous piece of DNA = cistron appeared plausible (Fig. 1).

[1] This paper appeared only after the review (Scherrer and Jost, 2007) of our gene concept had appeared and the present paper had been submitted.

While this was an important step, it turned out to be too simple. The reason is that while some polypeptides, like pancreatic RNase, assume a function by themselves, in most cases a genetically determined function is based on a higher order complex of polypeptides. Furthermore, these polypeptides typically may interact with low Mr compounds as Heme, vitamins, metal ions, etc. This means that several polypeptides or genes have to co-operate to secure a function. At that point, Jacob and Monod coined the notion of an “operon” constituted by several, possibly cooperating genes. Other problems then emerged with the discovery of regulatory genes. As an example, let us consider the lac repressor gene.

The *lac* function obviously has phenotypic effects. In fact, it is based on operator action involving the repressor; but the lac repressor gene is not part of the cistrons controlled by the operator. Indeed, the gene coding for the *lac*-repressor protein, which has to attach to the DNA sequence of the operator, placed *in cis* upstream to the genes in the operon, is encoded far away (Fig. 2). So, in what sense can we speak of all these distinct elements as of one gene?

In molecular biology of eukaryots, the situation is still more intricate and complicated as some researchers soon found. In bacteria, transcription and translation are tightly linked in a single physical complex. In eukaryotes, in contrast, the DNA is stored in the nucleus, which is the site of transcription, whereas the polyribosomes, where translation takes place, are located in the cytoplasm and thus removed from the DNA. The mRNA becomes autonomous, thus, and new types of controls become possible at that level. An untranslated region (5'-side UTR) preceding the coding sequence in the mRNA is needed to avoid a functional overload of the initial bases of the mRNA string. For both, chemical and steric reasons, the initial bases of the mRNA string cannot at the same time recognise and interact with the ribosome and bear the initiation triplet. However, there is also a 3'-side UTR at the end of the mRNA chain that, in the case of some genes (e.g., the Prion mRNA), can include more nucleotides than the coding sequence itself. These untranslated regions, being contiguous and in cis, on both sides of the coding sequence, clearly constitute problems for the original concept of the gene. Even worse for that concept, one and the same coding sequence can have different 3'-side UTRs, depending on cell type or expression timing of a given gene. The question then is whether the expression of one function at different times or in different cells (e.g. the myosin light chains (Kelly et al., 1995)) should count as a single gene.

Obviously, we can go on with problems and difficulties: In particular, mRNA form ribonucleoprotein complexes (mRNPs) in eukaryotic cells. More precisely, specific proteins recognise and attach to specific sequence motifs along the mRNA chain. This happens not only in the UTRs, but inside the coding sequence itself, for instance as shown in the case of globin mRNAs (Dubochet et al., 1973),. This indicated the existence of protein binding sites that are superimposed onto the coding sequence, as can be seen in EM pictures of mRNPs (insert in Fig.4), possibly with a specific code (Auweter et al., 2006) of protein-RNA interaction. It is a basic experimental fact that (ribosome-)"free" mRNPs, as found *in vivo* outside the translation machinery of the polyribosomes, are not translatable *in vitro*, unless most of the RNP proteins are removed (Civelli et al., 1980). Thus, it seems that these proteins assume some kind of repressor function. In passing, we note that RNPs are also capable of forming higher order complexes. These are assembled by interaction with other proteins or cellular structures as, for instance, at the level of the nuclear matrix (Ioudinkova et al., 2005; Lawrence et al., 1989; Razin et al., 2004) or the cytoskeleton (c.f. review in (Lawrence and Singer, 1991; Scherrer and Bey, 1994).

Even more devastating for the original gene concept is the existence in eukaryots of giant precursor RNA and its gradual processing (Scherrer and Darnell, 1962; Scherrer et al., 1963; Scherrer et al., 1966) (Georgiev et al., 1963) (review in (Scherrer, 2003)). Pre-mRNA "splicing" shows that the coding sequence is in most cases fragmented at the genomic level. In other words, only fragments in place of entire genes are stored in the DNA (Berget et al., 1977; Chow et al., 1977). From the point of view of the original genetic definition of the gene, and of the cistron concept, this means that the gene has to be *created* from its parts encoded in the DNA before it can be expressed. The phenomenon of differential splicing, implying that the same stretch of DNA can contain the information for different genetically identifiable functions, definitely suggests to conceptually and terminologically distinguish the gene as a function from its genomic counterpart in form of DNA.

Under these circumstances, how are we to deal with this situation where no single term is adequate to capture all types of information involved in the expression of a single genetic function? Clearly, we need to distinguish and isolate the essential units of the process of gene expression, from both the mechanistic and logical point of view. Necessarily, this process will require new concepts and terms; we shall boldly enter this path. In the end, we shall not only find ourselves equipped with precise definitions for gene expression in terms of Molecular Biology, but we shall also be able to devise and apply mathematical algorithms that can analyse gene storage and expression in terms of information processing. A short version of the proposal to be presented here has been published recently (Scherrer and Jost, 2007).

(2) Coding versus Control: the Genon Concept

Genetic function is carried out by proteins composed of folded polypeptides. Their amino acid sequences are read off in the process of translation from the coding sequence contained in the mRNA. The mRNA coding sequence is the elementary counterpart of the biological function, and therefore constitutes the natural starting point for a gene definition wishing to capture biological function. This leads to Benzer's original definition of the gene in terms of molecular biology, meaning the uninterrupted nucleic acid stretch that, as already mentioned above, was called "cistron" (Benzer, PNAS 1961) within the Jacob-Monod model (Jacob and Monod, 1961) of the operon (Fig. 2). Since translation is faithful (although coding is redundant due to the degeneracy of the genetic code), this mRNA sequence constitutes the equivalent of the polypeptide chain as the underling unit of genetic function and analysis. The important point here is that this uninterrupted nucleic acid stretch emerges only at the level of the mRNA. In particular in eukaryotes, it is typically not yet present at the DNA level as an uninterrupted sequence, but fragmented into exons. This implies that, at DNA level, the gene cannot yet be directly identified. Therefore, instead of looking forward from the DNA to the final mRNA prior to translation, we rather should look backward and understand how such a gene is assembled from pieces in the genome prior to its expression. At this point, however, in addition to the coding sequence itself, we have to take into account the existence of a program for the formation of the mRNA and its expression in time and space; this aspect also needs to be conceptualized.

More specifically, to implement this program, we have both, cis-acting receptors in the transcript and trans-acting factors in the milieu. The cis-acting receptors form sequence-motifs contained in the same strand of DNA or RNA as the fragments of the coding sequence. On the other hand, the trans-acting factors act on the signals placed in cis. Combined, they form the program that, in a sequence of many different steps, generates the gene within a given cellular space and at a specific time (Figure 3).

Let us list some of the many steps involved in gene expression, roughly in their temporal order (in fact, it is important to take a comprehensive view here): chromatin modification and activation, transcription and formation of pre-mRNPs, processing (including splicing) and transport of the pre-mRNP, formation and export of the mRNP to the cytoplasm, activation (or, perhaps more accurately, de-repression) of mRNA and, finally, translation. The cis-program of this process is specific for each individual gene, i.e. mRNA or polypeptide to be formed, although the same signals, in different combinations, can be utilized for the expression of different genes. We have coined the term "*Genon*" (contraction of "*Gene" and "operon* ") for the cis-acting program associated to a specific gene at mRNA level, as contained in the original nucleic acid sequence of DNA and pre-mRNA (Scherrer and Jost, 2007). The ensemble of trans-acting factors bearing on a given genon as contained in an mRNA will be called its "*transgenon*" (Figure 4).

Actually, the terminology needs to proliferate a little at this point. To express that a polycistronic pre-mRNA and/or a Full Domain Transcript (FDT) can control in cis one or several coding sequences, we propose the term “pre-genon” for the program contained in those structures. Prior to transcription, at the DNA level, the “proto-genon” includes in addition the signals for transcription activation (Figure 5).

A "poly pre-genon" then controls more than one gene in case of a polycistronic pre-mRNA (several (fragmented) coding sequences in a row) or contains the fragments of several genes to be created by differential splicing. A "mono pre-genon" occurs when a single gene is contained in a genomic domain, or at later steps of processing of a polycistronic or polygenic pre-mRNA. This mono pre-genon accompanies the pre-mRNA of an individual gene. When an mRNA is produced by alternative splicing, the remaining elements of its pre-genon form its genon. The distinct genon in the mRNA eventually formed includes all cis-acting signals, whether superimposed onto the coding sequence or contained in the 5'- and 3'-side UTRs. We also propose the term "holo-genon" for the sum of all (proto-)genons at the level of the entire genome.

In view of the distinction between cis and trans acting elements, the concept of the genon is meant to capture the cis-program. From this perspective, the effects of the trans-acting factors are indirect and relegated to the transgenon. Of course, we may change the perspective and consider each trans-acting factor of protein nature also as the result of a gene and its own genon.

It is a consequence of our concept that there are at least as many genes and genons as distinct open reading frames (ORFs) encoded in the genome. This means that in the human genome there

would exist about 500 000 genes, each controlled by its own genon and producing a specific polypeptide (Scherrer and Jost, 2007); this number may be carried to one million gene products if RNA genes and regulatory RNAs, including RNAi, are taken into consideration. According to current estimates, these genes and genons would arise from about 30 000 genomic domains (Pennisi, 2003; Venter and (et al), 2001) and produce at least as many Full Domain Transcripts (FDTs) and/or pre-mRNAs (Scherrer and Jost, 2007), and pre-genons of poly- or mono-genon type. Such a DNA domain might encode only a single gene, but more typically several genes, in the form of juxtaposed polycistrons or via differential splicing, as the products of one or several transcriptional units present within a domain. Thus, from our point of view, the about 30 000 genomic domains would encode at least 500 000 genes, each responsible for a specific polypeptide or other functional product.

The genon and its precursors act at transcriptional and post-transcriptional levels and expire with mRNA translation and its eventual degradation once they have fulfilled their function. Therefore, the translation step is the natural cut-off point for our analysis. A complete picture should include all aspects of the control of gene products, of their types as well as of their numbers, that is, RNA and protein degradation as well as biogenesis and the interplay and coordination of biosynthesis and degradation. However, we shall not treat here the post-translational programs governing gene expression, nor the catabolic side of protein homeostasis.

(3) Gene Expression and Regulation

To prepare the subsequent discussion of the genon concept, we now shall discuss the types of information involved in gene expression and the various types of gene products. According to our conceptual strategy, gene expression is governed by the coding sequence and the genon. The genon contains on one side the program in cis carried by the mRNA during the process, and on the other the program in trans, constituted by the transgenon representing the factors controlling and regulating the process between transcription and translation.

The products of gene expression can be of protein or RNA nature; these products may carry out some structural or enzymatic function, or may control gene expression in a mechanistic or regulative manner. This suggests a two-fold distinction, between protein and RNA genes (**P**-genes vs.**R**-genes for short), and between structural and controlling genes (**s**-genes vs. **c**-genes). As these distinctions are independent, we thus have sP-genes and sR-genes as well as cP-genes and cR-genes.

It has to be kept in mind, however, that some types of gene products may act simultaneously in several of these categories, for instance as sP and cR genes (e.g. the SRA protein gene involved, as an RNA, in differential splicing (Hube et al., 2006)).

(3.1) Protein genes

By definition, "protein-gene" implies that the corresponding gene function is carried out by a protein, constituted by one or several polypeptides.

(3.1.1) The protein-gene is the equivalent of the triplet-based coding sequence in the mRNA

As outlined above, the coding sequence is the mRNA equivalent of the gene, being defined by genetic analysis carried out at the level of the *phenotype*. The outcome of this analysis constitutes the *genotype* as the ensemble of defined inherited functions. Such physiological functions are based on the expression of an ensemble of unit functions. The unit function, subject to mutation, is carried by the polypeptide in its nascent [2]/ form (see Fig. 1). The actual function is exerted in general by a quaternary protein complex, which may integrate several identical and/of different proteins, possibly modified chemically, as well as by low Mr co-factors of organic or inorganic chemical nature.

The unit of a coding sequence is the *triplet* of nucleotides which, according to the genetic code, directs during translation of an mRNA the choice of a given anticodon carried by a given tRNA. Due to the degeneracy of the code, incorporation of an identical amino acid (aa) may be directed by different triplets. Within our argument, an essential feature based on this fact is that, according to the triplet chosen for a given amino acid, a different nucleotide sequence is formed at the level of the mRNA. In

[2] / "Nascent": we use this term in its strict logical meaning of "at birth", or the final product when released from the site of formation.

consequence, a different secondary structure of the nucleic acid arises which may be "recognised", for instance by proteins or interfering RNAs interacting with the RNA.

On the other hand, given amino acids and hence triplets are not equivalent within the polypeptide chain. Indeed, the same type of amino acid may assume different "functions" at the level of the secondary protein structure, in terms of hydrogen-bonding or ionic interaction, once the polypeptide chain is folded in the 3D space. This type of function may hence be projected back onto the corresponding genomic sequence. To single out a given triplet, its position within a coding sequence and/or exon should be labelled. One may hence conceive a notation for the position of a given triplet within a coding sequence. A possible and efficient description, compatible with alignment schemes in bioinformatics, is the following: Chromosome /genomic domain /maximal open reading frame /exon /triplet position within exon. (One should note, however, that there is as yet no generally agreed and universally employed convention in bioinformatics for describing the position of a triplet in the genome of a species.)

Accordingly, a given triplet (formally or as a physico-chemical entity) can be followed from the genomic DNA to a collection of amino acids within polypeptides. This is an essential feature when handling the triplet and its information content by a mathematical approach.

(3.1.2) Structural protein genes (sP-genes)

By definition, structural protein genes contribute to cellular structure and function either directly or via enzymatic activities. They may constitute the building blocks of the nuclear and plasmatic membranes, the endoplasmic reticulum, the nuclear matrix and the cytoskeleton. As enzymes they govern the intermediary metabolism as well as protein, RNA or lipid biosynthesis and degradation. There are the proteins acting as the mechanistic and enzymatic carriers of the system of protein biosynthesis, which do not discriminate among specific types of DNA, pre-mRNA or mRNA. Among the latter are, e.g., the RNA polymerases, the non gene-specific splicing factors, the non-specific transport factors as "exportin" or NLS (nuclear localisation signal) (Rodriguez et al., 2004) binding to RNA sequences, the translation initiation and elongation factors, the poly(A)- (Grossi de Sa et al., 1988) and CAP-binding proteins (Furuichi and Shatkin, 2000) (see review in (Shatkin and Manley, 2000)).

(3.1.3) Regulatory protein genes (cP-genes)

Regulatory protein genes control gene expression from chromatin activation to transcription and translation; they may function as repressors and activators of transcription, or act at post-transcriptional levels by interaction with pre-mRNA and mRNA. Four sets of such regulatory proteins can be distinguished: (1) the of non-histone type chromatin proteins, as the transcription factors (TFs; see e.g. (Latchman, 1990; Martin, 1991) as well as the histone- and DNA-modulating factors which control local remodelling of chromatin, allowing or not accessibility of the transcription machinery to DNA (Felsenfeld, 1992; Felsenfeld and Groudine, 2003) (for a review see (Gasser, 2002; Kouzarides, 2007)). (2) The nuclear pre-mRNA binding proteins which interact, in specific sets, with given types of pre-mRNA, in statu nascendi (Daneholt, 2001; Dreyfuss, 1986; Dreyfuss et al., 2002; Maundrell and Scherrer, 1979) as well as at the level of the nuclear matrix (De Conto et al., 2000; Maundrell et al., 1981; Razin et al., 2004). There are several hundreds of relatively acid proteins bound by hydrophobic bonds, and relatively fewer (some dozens) basic ones binding possibly by ionic interaction (the "histone-type" of pre-mRNP proteins) (Maundrell and Scherrer, 1979). (3) The cytoplasmic proteins binding in variable sets the non-translated mRNAs (Civelli et al., 1980; Spohr et al., 1970; Vincent et al., 1977; Vincent et al., 1981); these proteins bind in general by hydrophobic interaction, they act positively in guiding cytodistribution of mRNA (Arcangeletti et al., 2000; De Conto et al., 1999; Maundrell et al., 1979) , and negatively as cytoplasmic repressors (Civelli et al., 1980; Maundrell et al., 1979; Vincent et al., 1983). (4) The prosome particles (Martins de Sa et al., 1986; Schmid et al., 1984) (De Conto et al., 2000; De Conto et al., 1999; Ioudinkova et al., 2005) (review in (Scherrer and Bey, 1994), a population of protein complexes built of 2 x 14 subunits in variable composition, which bind on one side to chromatin, pre-mRNA and cytoplasmic repressed mRNA, and on the other to the nuclear matrix and the cytoskeleton (c.f. Fig 6 and 7). (Most interestingly, these same 20S particles act as the core of the 26S proteasomes, the main catabolic system able, in conjunction with the ubiquitin system, to degrade selectively specific proteins (Coux et al., 1996)).

The proteins binding nucleic acids at DNA and RNA levels, the non-histone chromatin proteins, the pre-mRNP and cytoplasmic mRNP proteins, all three constitute distinct populations of proteins including several hundreds and, possibly, up to several thousand members in animal cells. Since they seem to act on specific genomic domains, and on RNPs including specific types of mRNA, it follows

that they must act in a pleiotropic manner, constituting *sets* of proteins singling out, in combinations, specific (pre-)genons.

Among these cP-gene products two types should be distinguished: (1) those which act on specific individual genons regulating, hence, the expression of specific genes, and (2) those which control the expression of whole sets of genes or gene families. Among the latter are, for instance, some types of transcription and translation factors.

(3.2) RNA genes

By definition, "RNA-gene" implies that the corresponding gene function is directly carried out by an RNA, in association or not with proteins.

(3.2.1) Structural RNA genes

The most important member of this class of RNA is the ribosomal RNA (rRNA) which serves as the scaffold of ribosomal subunits by organising the sequential alignment of ribosomal proteins (Scheer and Hock, 1999; Tschochner and Hurt, 2003). The rRNA has, in addition, ribozyme functions (Steitz and Moore, 2003). The metabolic precursor of the rRNAs found in the the small (16S and 18S rRNA, resp. in prokaryots and eukaryots) and large (23S and 28S rRNA) ribosomal subunits is the nascent pre-rRNA (45 S in eukaryots: (Scherrer et al., 1963; Spohr et al., 1976)). It is of the same nature aligning, in addition to the proteins ending up in the final ribosome, proteins which*, in eukaryotic cells,* never leave the nucleolus (Scheer and Benavente, 1990; Tschochner and Hurt, 2003). They have a structural role in ribosome biosynthesis and the nucleolar dynamic architecture; pre-ribosomes form the fibrillar centre of the nucleolus, whereas the final ribosomal subunits constitute its granular zone 5 (Figure 6D).

(3.2.2) Regulatory RNA genes

Among the RNAs intervening in control of gene expression we have to distinguish those which handle many types of (pre-)mRNA without individual selection, in contrast to those which selectively recognise and control, in a sequence-specific manner, *individual* types of (pre)mRNA. The latter allow strict recognition and control of individual gene expression, whereas the former RNA may discriminate among *classes* of, but not of individual mRNAs.

(3.2.2.1) Non-discriminating RNA regulators

The most straightforward example of such RNA is the tRNA class which select individual triplets in the coding sequence. Availability of specific types of tRNA corresponding to types of (degenerate) triplets, or the many chemically modified tRNA types may influence and coordinate the expression of classes of mRNA.

Similar limited regulatory function is exerted by the U-type RNAs involved in splicing (Valadkhan, 2005; Will and Luhrmann, 2005). The snoRNAs (small nucleolar RNA) guide modifications (including ribose O-methylation) of (pre-)rRNA and, possibly, even mRNA; they are tissue-specific in higher eukaryotes (Dennis and Omer, 2005) but are present already in *Archaea*. A particular feature of snoRNAs in higher eukaryots is that they are often encoded in introns of pre-mRNA (Filipowicz and Pogacic, 2002).

(3.2.2.2) Discriminating RNA regulators : siRNA and miRNA

The advent of RNA interference (RNAi) marked the unexpected discovery of sequence-specific mRNA silencing by natural antisense transcripts (Sontheimer, 2005; Tang, 2005). This type of post-transcriptional regulation may occur also at pre-mRNA level in the nucleus (Matzke and Birchler, 2005).

The basic mechanism of RNAi is the synthesis, by an RNA-dependant RNA polymerase and an RNA replicase, of double stranded RNA copies of target RNAs, in particular of mRNA. From such RNA double-strands, several hundred bp long, short 21-25nt long fragments are cut out within the RISC RNA-protein complex. Two classes of interfering RNA are reported, the small interfering RNA (siRNA) and the micro RNA (miRNA) which form distinct siRISC and miRISC complexes (for a recent

review see (Sontheimer and Carthew, 2005)). SiRNAs induce destruction of the target mRNA after sequence-specific hybridisation whereas miRNAs silence temporally the target mRNA.

(4) Genomic information not directly related to gene expression

The object of this chapter is to point out that large parts of the genome relate to other mechanisms than gene expression *per se.* Some relate to replication, genetic transmission and meiotic recombination, and others to the static and dynamic organisation of chromatin; the latter may bear, eventually, on gene expression. Indeed, one of the most striking conceptual developments in recent years was the gradual introduction of the notion of space in genome organisation and gene expression, in addition to the classical concepts of regulation in time and according to physiological change.

The nucleic acids carrying the genome and the gene expression machinery must assume at least two basic functions : (1) contain the information relating to the genes and, (2) serve as the physical support for this information. Function (1) is all evident within the definition of gene and genon developed above, whereas the implications of function (2) are less clear.

First, the support of genetic information has to obey the necessities of various quite distinct functions as, e.g., (i) long term storage of genetic information, (ii) its transmission from generation to generation and (iii) the intra-cellular mechanisms of gene expression including selective transport of the transcripts to the sites of translation (see Figure 7), and post-transcriptional regulation in adaptation to physiological conditions. The chemically quite inert DNA seems well suited for safeguard and transmission of information, whereas the more reactive and flexible RNAs are, seen their chemical and physical properties, better suited to adapt to the necessities of gene expression.

It is often forgotten that both, DNA and RNA, act *a priori* as the mechanical support of genetic information and have to adapt to stringent rules deriving from their own physico-chemical properties. Concerning information storage and regulation at DNA level, an important factor coming into play is, for instance, the quite high physical rigidity of the DNA double strand which does not allow free and random movements, in particular in the conditions of high viscosity in the cellular nuclei. There are limits to folding up of hetero- and euchromatin and, e.g., to rapid "flip-flop" movements of DNA loops assumed to operate according to some popular models (de Laat and Grosveld, 2003). Furthermore, relating to what may be called "chromosome mechanics", coming into play in mitotic replication and meiotic recombination, as well as sister-chromatid exchange, these rules sometimes may supersede the information content relating to the genes *per se*.

As suggested above, the genomic DNA may have an architectural function organising both, overall nuclear as well as local chromatin organisation. Cavalier-Smith pointed out already that there is a correlation between DNA organisation and chromosome architecture influencing both, nuclear size and linear chromosome organisation (Cavalier-Smith, 1978); speaking of “nucleoskelettal DNA” (S-DNA), as opposed to “genic DNA” (G-DNA), implicitly he proposed a relation of DNA and nuclear matrix. Recently, Képès proposed in his solenoid model that there is a correlation between transcription factor and promotor attachment sites and the higher order chromatin organisation; moreover, this organisation is suggested to be transcription-pattern dependent (Kepes and Vaillant, 2003; Képès, 2003). This points to interdependence of 3D genome and transcription organisation, i.e. the static and dynamic nuclear architecture, as discussed below within the Unified Matrix Hypothesis (c.f. (4.1) and (Scherrer, 1989)) Actually it can be assumed that many so-called transcription (initiation ?) factors are proteins of the nuclear matrix (c.f. below 5.5.2), and that promoters (Auboeuf et al., 2007; Auboeuf et al., 2005) may carry out some their functions at the level of the RNA-dependant nuclear matrix (Ioudinkova et al., 2005; Razin et al., 2004).

Surprisingly neglected by actual Molecular Biology is the fact that DNA and RNA have to operate in a 3-dimensional space; and passive "crystallisation" or interaction of macromolecules can not possibly explain all of genomic and cellular 3D organisation. (DNA "knows" that there is iron and light in the world, but seems to have "forgotten" that its environment is a 3D space !). When genes are being expressed, their reconstitution from RNA fragments in course of splicing, as well as the physical transport of mRNA from sites of transcription to those of expression, have to be organised in the 3D space and necessitates a precise dynamic architecture in space and time. Within these mechanisms, relating to semi-static and dynamic nuclear architecture, the positions of exons and the sites of RNA-protein interactions within the transcripts obey certain rules, which must be compatible with the selectivity of RNA processing and its implementation in the 3D space. Furthermore, since it became obvious that the genome is distributed in specific, experimentally identifiable sectors of the nuclear space, assigning specific positions to chromosomes and genomic domains, the organisation of the DNA *itself* in 3D must be taken into consideration. Figure 8 outlines the conceptual consistency of

organisation in space, common to DNA, RNA and proteins; the basis is the "architectural" necessity to place sites of action and interaction in precise 3D positions relative to each other. The mutual interdependence of the exonic fragments of genetic information and the biophysical properties of its physical support lead, inevitably, to the notion of additional genomic information necessary to rule these processes.

That the nuclear DNA might carry information other than that related to the genetic code could be inferred for a long time on the basis of data pointing to its possible role in cellular structure. The C-value paradox (Cavalier-Smith, 1978; Commoner, 1964) showed a correlation of cellular and nuclear size (the prime architectural feature !) with DNA content. Later, comparing amphibian erythrocytes in species with a DNA content varying up to 100 times, it was found that these differences bear on repetitive DNA; interestingly, in these species the complexity of the *transcribed* genome remains comparable (Rosbash et al., 1974). Furthermore, most of such repetitive DNA was found to be AT-rich, with little or no coding sequences.

That DNA may have a structural role independent of its gene content is also demonstrated by the phenomenon of the "petit" mutants in yeast (Bernardi, 2005). Petit mutants have non-functional vestiges of mitochondria, which contain, however, normal-sized mitochondrial DNA. It was found that in such mutants the mitochondrial genes were progressively lost and, surprisingly, replaced by stretches of almost pure A+T (Bernardi, 2005). There seems to exist, thus, a mechanism subject to selective pressure, which maintains the length of mitochondrial DNA constant independent of the gene content. A similar case may exist in the kinetoplast of Trypanosomes, where the DNA of the organelle is largely composed of gene-less A+T-rich stretches (Shapiro and Englund, 1995).

In chromosomes also, there are DNA segments which relate to structure rather than gene content. The genome is subdivided into genomic domains. The definition of genomic domains may be based either on the organisation of DNA, chromatin and/or chromosomes; or on functional considerations, such as units of replication or
transcription. As pointed out in the "Cascade Regulation Hypothesis" (CRH; see below and Figure 10), conceived in the sixties (Scherrer and Marcaud, 1968) and laid out in final form in 1980 (Scherrer, 1980), the most straightforward illustration of genomic domains are the bands in the polytene chromosomes observed in some insects as *diptera* (Fig. 6C). Their salivary glands contain *bona fide* interphase cells, which actively express many genes and predominantly those at the basis of silk secretion. By order of magnitude, in drosophila there are as many cytogenetically observable polytene chromosome bands as units of meiotic recombination

(Judd et al., 1972) (NCBI Map Viewer, 2006); there is hence coincidence of physical and genetic units of function. From these bands spring up, upon developmental or experimental activation, the so-called "RNA puffs" (Fig. 6C), signs of transcriptional activity visible in the optical microscope (Grossbach, 1974) . A band may produce a single or several pre-mRNAs but corresponds, obviously, to a unit of transcriptional regulation. In some types of insects, the family of *Sciaridae*, the phenomenon of "DNA-puffs" occurs, where DNA has to be replicated locally, as a prerequisite for transcriptional activation (Glover et al., 1982) . In this case, the unit of transcriptional control corresponds, to units of replication as well (Fig 6C and (Lara, 1987).

There is, thus, good reason to consider the interbands of polytene chromosomes as borders of genomic domains. All the more since some molecular biological and biophysical facts point to the same interpretation. Interband DNA has some qualities of insulators, as defined by molecular genetics (Gaszner and Felsenfeld, 2006) and are, e.g. in the case of the drosophila gene *Gipsy*, visible in the cell nuclei after cytochemical staining (Gasser, 2002; Gaszner and Felsenfeld, 2006). Finally - and most interestingly - the interbands correspond to sites of Z-DNA formation (Nordheim et al., 1986).

The higher order organisation of DNA into genomic domains is embedded into the super-organisation of chromatin and chromosomes, which divide the genome into individual segments. Phenotypically very similar animals of closely related species may have vastly different numbers of chromosomes. Indeed, the fusion of the 46 telomeric chromosomes of *Mus Musculus* into the 23 metacentric chromosomes of *Mus Posciavino* (Capanna et al., 1976) will still produce a mouse, albeit of a different size. And the 6 chromosomes of *Muntjacus Muntjak* or the 46 of *Muntjak Reevesi* will be able to condition an almost identical phenotype (cf. Lima de Faria, 1980); they maintain, however, a similar pattern of R- and G-bands (c.f. review in (Sumner, 1982)). At this level of organisation, other types of genomic information is encoded which bears only indirectly on gene expression. We shall discuss here the 3D organisation of DNA and some phenomena, which might be singled out as "chromosome mechanics".

(4.1) The 3D DNA organisation according to the Unified Matrix Hypothesis

The Unified Matrix Hypothesis (UMH) was an early attempt to give a logical interpretation to the - apparently - surplus DNA, lightly qualified as "junk" (Ohno, 1972) (c.f. discussion in (Scherrer, 1989)). Starting from the C-value paradox showing linear correlation between DNA content and relative size of cells (Cavalier-Smith, 1978), the proposition was made that a major part of the 95 % of DNA not coding for proteins might have, essentially, an architectural function.

A straightforward illustration of this proposition was the phenomenon of *ectopic pairing* (Barr and Ellison, 1976; Cohen Jr, 1976; Kaufman et al., 1948); (Ananiev et al., 1981) of polytene chromosomes observed in the salivary glands of drosophila and other systems of "giant" chromosomes (the latter are the result of DNA replication without disjunction of the daughter DNA strands which remain physically aligned up to 10.000 times). Ectopic pairing consists in physical connections by cables of, apparently, nucleo-protein nature, linking distant sites within and in between chromosomes (Fig. 9 A) These connections run typically from interband to interband and in between telomeres. They have been mapped in details (Fig. 9 B) providing genetically significant patterns (Kaufman et al., 1948). Of particular importance to the emerging matrix concept was the fact that several such ectopic cables suspend the nucleolus in a particular position relative to the chromosomes (see Fig 9A and (Ananiev et al., 1981). They must, hence, include the DNA of the nucleolar organiser sequences. The nucleolus was known for some time already to occupy specific positions in the nucleus of non-transformed cells differentiating normally (see Fig. 9A). The idea arose, thus, that ectopic pairing might reveal a basic mechanism implemented in any normal interphase cell, having normal chromosomes based on double-stranded, non-amplified DNA.

On this basis the proposition was made within the UMH (Figure 9 C, D) that, quite in general, the nuclear DNA was organised in a 3D network, where proximal and distal chromosome sites were connected by bi-functional MARs (matrix attachment regions) keeping chromosome domains and sites of transcription in specific spatial positions (Fig. 9 C, D). At those positions, transcripts are formed, processed and exported to the nuclear periphery. A straightforward example of this process is the nucleolus where pre-rRNA is processed (see review in (Tschochner and Hurt, 2003)) and from where subribosomal rRNA is exported, as a component of the ribosomal subunits (see also Fig. 6D).

The main conceptual implication was that *shear DNA length amounts to genetic information*, independent of its sequence. This proposition of the UMH allowed to logically interpret several features hitherto difficult to understand, as e.g. the phenomenon of the "Chromosome Field" (Lima de Faria, 1979; Lima de Faria, 1983; Lima-de-Faria, 1980) showing the topological maintenance in evolution of groups of genes within the chromosome organisation, as shown in Figure 9 (E, F), and allowed propositions to explain, for instance, the specificity of sites of chromosome crossing-over in some types of leukemic cells.

This is not the place to further develop this theory; suffice to say that in recent years more and more relevant data could be placed within the originally loose frame of the UMH. The recent reports about "kissing chromosomes", showing that distant chromosomal sites must be linked physically, to allow the expression of specific genes within "3D gene regulation", is a most eloquent illustration of its basic concept (Kioussis, 2005; Spilianakis et al., 2005). In the meantime more and more data accumulated which point to a quite strict organisation of the genome and gene expression in the nuclear space (Bolzer et al., 2005; Cremer and Cremer, 2001; Cremer et al., 2000; Stadler et al., 2004). Genes seem to reside in specific places and mRNA is brought to cytoplasmic sites of, sometimes functional significance as, e.g., when muscle-specific mRNAs (resp. RNPs) are transported to the intra-cellular sarcomeric plates of myotubes in order to be translated locally, there where the proteins shall be assembled (Foucrier et al., 1999; Foucrier et al., 2001; Fulton and Alftine, 1997).

Here we need just to point out that there exist basic functions of DNA that are only indirectly related to gene expression. The UMH indicates disjunction of the actual genome size, which varies vastly within the C-value correlation, in particular in its repetitive elements, from gene expression. As pointed out above, in the same group of species with vastly varying DNA content, the sequence complexity of the expressed genome may remain almost constant (Rosbash et al., 1974). However, the static and dynamic DNA architecture seem to play vital functions, which are maintained in evolution, independent of DNA and gene content.

Although the overall architectural function of DNA seems dissociated from the specific mechanisms of protein biosynthesis, an architectural function in gene expression of the transcripts as well became more and more evident. The observations of an RNA-dependant nuclear matrix (De Conto et al., 2000; Maundrell et al., 1981; Nickerson, 2001; Penman et al., 1982) carried by the primary transcripts and their processing products (Ioudinkova et al., 2005) shows, that the genon-related program

encoded in pre-mRNA and mRNA must also satisfy an architectural function, as originally suggested by the UMH (Scherrer, 1989). We need to distinguish, however, this type of dynamic architectural function from the basic one, carried essentially and directly by the DNA, which is implemented prior to onset of transcription; it remains static in a given type of differentiated cell.

One may propose that the DNA defines the overall nuclear architecture *per se* and, in particular, the euchromatic part of chromatin which is unfolded and DNase-sensitive. The directly DNA-dependant 3D network is more "static" than the dynamic RNA-dependent architecture. It is liable to modification, however, in the process of cell differentiation, when the relative parts of hetero- and euchromatin are modified. The concept of "Quantal Mitosis" (see below in chapter 5.2.1.2) proposed by Holtzer and Weintraub (Holtzer et al., 1975; Holtzer et al., 1972) was based on the fact that, in course of differentiation, there are special types of cell divisions when further differentiation is blocked, at precise stages of differentiation, by substitution of thymidine (T) by bromo-desoxyuridine (BudR) which is without any effect later on. BudR substitution reduces the dissociation constant of DNA-binding proteins, as observed already for the lac-repressor (Wick and Matthews, 1991).

On the other hand, there is the transcript-dependant, dynamic nuclear architecture as a result of RNA transcription, processing and transport. It is encoded in the (pre-)mRNA and its (pre-)genons. However, in both cases - the non-transcribed as well as the transcribed genome - the architectural function turns one-dimensional DNA and RNA into 3D structures, into which the coding parts are inserted. This conceptual deduction seems liable to explain to some extent the 95% of "surplus" DNA in a logical manner.

(4.2) Meiotic recombination, synaptonemal complex and chromosome mechanics

Another type of genetic information fixed by evolution into the genome without being directly involved in gene expression may be related to mechanisms termed, possibly, *Chromosome Mechanics.* This term relates again to the fact that the nuclear DNA not only carries several types of information, but is at the same time the mechanistic carrier of the information contained. Whereas molecules like DNA or RNA are carriers of information and of genon-related signals and provide, thus, information for the process of gene expression, the nuclear DNA in addition provides the structural organisation for the interaction of such biomolecules. Thus, here, in contrast to the typical fluid situation elsewhere in the cell where molecules have to find each other on the basis of mutual affinities, we see a spatial structure that enables specific interactions and prevents others. This is a type of information to be distinguished from the coding and regulatory information.

Applied to the genon concept, this means that in the nucleic acid backbone, within the cis-program of the holo-genon, coding, functional, and structural aspects are intertwined whereas in the transgenon the regulatory or controlling features dominate.

Thus, merely mechanistic criteria of the information carriers and their higher order complexes must be respected as solidity, flexibility and folding characteristics, adapted chemical stability ("*DNA is granite and RNA butter*"), viscosity, etc. In some phases of physiological life, these physical and chemical criteria have to prime over the information contained in the signals carried by individual biomolecules.

A particularly interesting illustration of such phenomena is meiotic recombination and sister chromatid exchange which imply the formation of the synaptonemal complex as the physical basis of meiotic crossing over (Colaiacovo, 2006; Kleckner, 2006). There, the two DNA strands with their gene fragments in the derived chromatin structure have to align point by point, down to the individual exon, in order to allow precise breakage of the DNA strands and their ligation to the opposite ones. If this condition is not satisfied, as is often the case in interspecies crosses, meiotic recombination cannot proceed and the DNA is dissolved. Of course, in most species other barriers have evolved which preclude interspecies mating prior to the molecular interactions outlined. However, chromosome mismatch represents the ultimate molecular mechanism at the basis of the species barrier, as clearly visible in the case of crosses of horse and donkey (c.f. Fig. 8 in (Scherrer, 1989) and (Chandley et al., 1974)) resulting in mule and hinny; those creatures - though going strong - are incapable to reproduce. This example is particularly speaking since, surprisingly, fertile crossing-over in species having vastly different DNA content - and cell size - is possible in some cases (Bennett, 1982), best illustrated for some plant species, the *Secale*, e.g. (which, thanks to this phenomenon bearing on the size of seeds, are at the basis of the "green revolution" in world nutrition). There, the chromosome alleles of the parent species match to align, but their surplus DNA folds out from the strictly aligned axis of the synaptonemal complex, in opposite loops of very different size (according to a proposition of (Rees et al., 1982)). This process is a particularly striking example of "chromosome mechanics; it

implies the existence of an independent mechanism which lays down signals for meiotic alignment which seems to be largely independent of all other genomic information.

(5) Development of the Genon concept

(5.1) The genon acting in cis is carried by sequence motifs in the mRNA

As defined above, the genon represents a regulatory program superimposed and attached to a given coding sequence. It is materialized in cis by the ensemble of signals within mRNA secondary structure that control the expression of the contained coding sequence. These signals are either present in the coding sequence or in the 5'- and 3'-side UTR of the mRNA sequence; the mRNA sequence carrying a given program is, therefore, longer than the coding sequence which it contains. In this manner, a specific cis-genon is defined for every gene (Figure 4).

The implementation of the genon-program in cis is carried out in trans by nucleic-acid binding proteins (NABPs) on the one side, and by interfering small RNAs (siRNAs, miRNAs) on the other; all together, these factors constitute the transgenon, the program in trans.

(5.2) Proto- and Pre-genon as well as the final genon placed in cis relate to the Cascade of Regulation

We will restrict here discussion to the *cellular* regulation cascade (outlined in Figure 10) including the steps of gene expression leading from the zygotic genome to the final polypeptide. Logically the holo-cascade of regulation (c.f. Scherrer 1980) may start with the creation of the individual genome of an organism from the gene pool of the species. At the other end of the cascade, we have post-translational events allowing for eventual functional expression of a gene.

The cis-acting program of the individual genon is encoded in the proto-genon of a genomic DNA domain; after transcription, it is carried further by the pre-genon within a Full Domain Transcript (FDT) and pre-mRNA, which may include a single or several genes. An individual pre-genon is represented, hence, at DNA level as well, and is expressed relative to other pre-genons by activation of individual genomic domains, individual transcriptional units, or by differential splicing of a pre-mRNA or poly-pre-genon, according to programs of cell differentiation or physiological adaptation. After transcription, the pre-mRNA or FDT will be processed according to its pre-genon and its complement of factors within the holo-transgenons (see below) of a given cell.

Within the Cascade of Regulation, specific gene expression in a given eukaryotic cell may be subdivided in (at least) the following steps:

(1) Organisation of the DNA in the 3D-space and formation of the DNA-dependant matrix (Step 1 in Fig.10).
  (1.1) Organisation of Chromatin into chromosomal territories.
  (1.2) Formation of differentiation-specific local chromatin networks and the DNA-derived nuclear matrix.
(2) Activation of Chromatin Domains for eventual transcription of individual transcriptional units contained in a domain (step 2 in Fig.10).
(3) The primary transcripts (step 3 in Fig.10).
  (3.1) Synthesis of the FDT or individual primary pre-mRNA.
  (3.2) Association of nuclear RNA-binding proteins to pre-mRNA forming the pre-mRNPs.
  (3.3) Formation of the RNA-derived nuclear matrix by integration of the pre-mRNPs.
(4) Processing of pre-mRNPs (step 4 in Fig.10).
(5) Differential splicing and formation of the pre-mRNP including exons of a single coding sequence (step 5 in Fig.10).
(6) Final processing of pre-mRNPs (step 6 in Fig.10).
(7) Import of mRNA into the cytoplasm (step 7 in Fig.10).

(8) Formation of cytoplasmic inactive (ribosome-)free mRNP with concomitant replacement of the majority of nuclear (pre-)RNP-type proteins by cytoplasmic ones (step 8 in Fig.10).
(9) Activation of mRNA and polyribosome formation (step 9 in Fig.10).
(9.1) Replacement of mRNP proteins by translation factors, forming the translated mRNPs.
(9.2) Formation of polyribosomes by association of 40 S and 60 S (native) ribosomal subunits forming functional ribosomes.
(10) Translation of the coding sequence in mRNA (step 10 in Fig.10).
(11) Formation of the nascent primary polypeptide and secondary protein structure (the genon has expired).

In addition, at several steps of biochemical information processing RNA interference (RNAi) takes place, in the nucleus as well as in the cytoplasm, by physical elimination or temporary masking of mRNA sequences by siRNAs or miRNAs. Another important but not clearly localised mechanism of information processing is RNA editing, by which a coding sequence in an already present (pre-)mRNA can be modified (review in (Koslowsky, 2004)).

Mechanisms of expression and regulation within the cascade operate mainly by association of regulatory proteins and of interfering RNAs, and by the action of the enzymes involved in the transcription and processing machinery, including control of RNA editing. The physical support of the carriers of information are the nuclear matrix and the cytoskeleton, as well as the endoplasmic reticulum for proteins to be exported.

The known biochemical steps of DNA and RNA activation, of RNA processing and transport occur within the "Cascade of Regulation" which stepwise reduces the information content of the genome to that of a single gene, ultimately. It shall be pointed out, however, that in terms of information processing, information is gained during this process, to the extent that uncertainty about the eventual selection of a given triplet in the DNA, to be expressed within a polypeptide, is gradually reduced. The potential information of the genome thus becomes effective.

The content in genomic information is currently evaluated in terms of what in the biological literature has been called "sequence complexity", that is the length of non-repetitive DNA or RNA (Britten and Kohne, 1968; Hough et al., 1975). By now, the latter is well documented by published sequences of various species, in particular, the human genome with approximately 3.25 $x10^9$ base pairs (Lander et al., 2001.; Pennisi, 2003; Venter and (et al), 2001), of which 90 % may be unique sequence DNA. In contrast, a gene, e.g. that of the human alpha-globin, includes about 600 nt only; this represents a fraction of and, thus, a selection in regulation of 1 in 10^6 or 10^7. However, such degree of selection is beyond the possibility of a direct process based on biochemical or biophysical mechanisms, essentially for reasons of chemical thermodynamics, enforcing thus multi-step selection as the rule.

More precisely, there are 3 main reasons for stepwise regulation of gene expression:

(1) Noise: As pointed out in the Cascade Regulation Hypothesis (CRH), published first in 1968 (Scherrer and Marcaud, 1968) and in more final form in 1980 (Scherrer, 1980), such a degree of selection (10^6 to10^7 in eukaryots) is only conceivable within a series of sequential selection steps. Indeed, in our physical and chemical world, in any direct selection step, no better resolution than about 1 in 10^3 is possible. Signal to noise ratios within the rules of physical and chemical thermodynamics are the limiting factors; indeed, the degree of association of any controlling factor is limited by its dissociation constant. However, in E. coli with a few thousand genes to be controlled, accordingly, direct selection of genes or operons at genome level seems possible.

(2) Effort: There exist different search strategies that, in principle, could be employed for the selection. If one performs the selection in a single step, one needs to screen all the available elements to find the right one. The selection effort is then proportional to the number of items to be scanned, that is in case of the human genome, of the order 10^6 or 10^7. As explained above, this effort is far too large to be biologically realistic. The other extreme is search by binary alternatives. Here, in the first step, the set of items to be searched is divided into two classes of equal size, and one selects one of those. In the next step, that class is again divided into two classes and the process is repeated until after log *N* (the binary logarithm of the number *N* of elements) steps, the desired element is found. In the present case of *N* of the order 10^6 or 10^7, this amounts to about 20 to 23 steps, where in each step only a choice between 2 alternatives has to be made, so that the total search effort is about 45 which is rather small, and in fact the best one can achieve. However, the number of steps involved is two large to be biologically plausible. Thus, a compromise between the two extremes seems to be the biologically best solution: Instead of scanning only 2 alternatives in each step, one scans a larger number, for instance 10, that is, one divides the set to be searched not into 2, but into 10 subsets. The

number of steps required in our example then reduces to 6 or 7, with an effort of 10 in each step. Thus, the total effort is 60 or 70 which is sufficiently small, and this is achieved in a small number of steps. (see also chapter (IV.D.) in (Scherrer, 1980)).

Within gene expression, selection effort means mainly the number of regulative factors needed within the transgenon. To keep their number in the genome low - for obvious reasons - cP-regulators have to operate in *sets of combinations* at the different selection steps. Furthermore, protein is expensive to cell metabolism whereas RNA is relatively cheap. This may be one of the reasons why anti-sense RNA, possibly emerging already in merely chemical, complex pre-biotic systems, has been maintained in evolution (Rich, 1961) . The cP-genes are more sophisticated; indeed, proteins can process input into output, that is, perform different tasks depending on the input they receive from low Mr molecules or allosteric effectors, in contrast to anti-sense RNAs, which by themselves essentially operate in an on/off mode.

(3) Reaction speed: Gene expression is a long and complex process. When physiological adaptation must be rapid, the necessary information may not possibly be called from the genome : gene information must be stored close to the place of action, in the extreme case in form of pre-proteins as, e.g., trypsinogen, turning into a functional enzyme upon a simple biochemical signal. We have introduced the term "peripheral memories" (see Fig. 3) for the epigenetic storage of genomic information closer to the sites of expression (Scherrer, 1980). These may take the form of pre-mRNPs (including fragments only of genes) or silent, repressed mRNP complexes. In most cells, partially processed transcripts turn slowly over in the nucleus, and cytoplasmic mRNAs shuttle between the expressed and silent states, turning over individually, independent of active translation (Spohr and Scherrer, 1972). Most important, large fractions of the genome are stored as RNA in the metaphase cells as well as in the oocytes, to allow epigenetic transfer of information in between generations of cells and organisms; in the latter case, this information allows *de novo* gene expression and regulation after fertilisation of the egg.

The expressed part of the genome can be measured by modern micro-array techniques, which give numbers of genes represented in a given cell isolate, and from which the non-repetitive sequence length, in terms of (known) RNA-sequence, might be calculated. However, such data are at present not available in a comprehensive manner, in relation to the biochemical steps of the gene expression cascade. We have therefore to rely on the published data of sequence length ("sequence complexity") measured by re-association kinetics in hybridisation assays which are expressed as Cot- (for DNA (Britten and Kohne, 1968; Pearson, 2006)) or Rot- (for RNA (Birnstiel et al., 1972)) values (c.f. also (Imaizumi-Scherrer et al., 1982)).

Early hybridisation data indicated that 10 – 20 % of the nuclear DNA is transcribed in most species, even in highly specialised cells as the red blood cell, where 90% of the protein output is globin (Imaizumi-Scherrer et al., 1982). Those data represented - by necessity - the more stable, partially or fully processed RNAs; indeed, the primary transcripts are highly unstable. Very large transcripts start to be processed and spliced even prior to transcription termination, as can be observed directly by EM of transcription complexes (Osheim et al., 1985). These old data are, thus, compatible with the more recent notion that up to 60 % of eukaryotic genomes might be transcribed eventually, at one time or another, in a cell of an organism (Fantom Consortium and Riken Genome Groups, 2005). The latter figure points once more to the only recently adopted basic fact that transcriptional regulation represents a minor part only of regulation of gene expression, as pointed out in the Cascade Regulation concept. The idea of pre-eminence of transcriptional controls is a vestige from prokaryotic models; in contrast, most regulation in eukaryotes is post-transcriptional. The main reason for this is that there are many more tasks to accomplish in the cell for transcripts than carry protein coding information. Furthermore, as pointed out already, regulation close to peripheral gene expression sites is more rapid and, hence, more efficient than calling up a gene from the genome.

In the following, we will discuss the individual steps of the cascade of regulation in view of the genon concept.

5.2.1 Organisation of the DNA in the 3D-space (Step 1 in Fig.10)

The first step of the regulation cascade involves the selection of the chromatin fraction to be eventually activated in a given cell. The zygotic genome is being subdivided into stem cell lines according to the mechanisms of (lineage) determination (review in (Tiedemann et al., 2001)), which are at the root of cellular differentiation. Almost all of the non-repetitive part of the genome is transcribed at one time or another in an organism; and in any case during diplotene stage of oogenesis when lampbrush chromosomes are formed (see Fig. 6 B). Along with this first reduction of genomic information to that potentially to be expressed in a given tissue or cell goes the process of

heterochromatin formation. This implies the permanent or semi-permanent silencing of part of chromatin, which is biophysically put aside, and the organisation of the remaining euchromatin into a 3D network where every genomic domains finds its assigned place in the nucleus. This process operating at genome level involves not only the proto-genons of individual genomic domains but as well the DNA intercalated in between them. We may thus have to take in consideration *the holo-protogenon*.

(5.2.1.1) Organisation of Chromatin into differentiation-specific chromosomal territories

Originally based on spurious observations (Lawrence et al., 1989), the notion that nuclei of differentiated cells are subdivided into chromosome territories, seems actually quite well established (cf (Cremer et al., 2000; Stadler et al., 2004) and recent review in (Albiez et al., 2006) (Lanctot, 2007)). In positive correlation with this concept is the fact that also the condensed metaphase chromosomes occupy established places in the metaphase plates, as know for some time, e.g. for *secale* species (Bennett, 1982). More recent data illustrate the same fact since, most interestingly, when the actin-skeleton is mechanically extracted from living metaphase cells (by an antibody-coated hook attaching the cellular matrix just under the plasma membrane), the metaphase chromosomes are extracted in a precise sequential order (Maniotis et al., 1997). Furthermore, maternal and paternal chromosomes in mice remain linked in separate sets throughout the first 3 cell divisions after fertilisation (Odartschenko and Keneklis, 1973). These data point to the existence of a basic mechanism that keeps chromosomes in assigned places relative to a physical (nuclear) matrix, in metaphase as well as in interphase cells.

As suggested in the Unified Matrix Hypothesis (Scherrer, 1989) discussed above (cf. 4.1), the DNA may form a 3D-network spanning the nucleus; in turn, it might prime position information by the DNA-induced nuclear matrix via MAR binding proteins. This organisation might be perpetuated throughout cell division when the nuclear membrane is dissolved, chromosomes condensed and, hence, DNA largely removed from the matrix; a network essentially constituted of proteins then spans the entire cell (Fig. 9D). Beyond the data outlined 20 years ago in the UMH, to our knowledge, little new facts pointing to such a mechanism are at present known. Nevertheless, a model suggesting that, upon unfolding, the metaphase chromosomes are re-inserted into pre-established territories is, apparently, not in contradiction with actual knowledge.

The first selection step within the cascade of regulation leading to the expression of specific genes is, thus, the organisation of chromosome territories (Cremer and Cremer, 2001).

(5.2.1.2) Formation of differentiation-specific *local* chromatin networks and the DNA-derived nuclear matrix

The next selection step leading to the eventual expression of a specific gene, concerns the organisation of a chromosome territory into repressed or activated domains, the latter to be placed into specific expression-relevant positions within the nuclear architecture (Lawrence et al., 1989). This process depends largely on the holo-protogenon including, in addition to genomic domains, the DNA placed in between.

In course of differentiation, chromatin is remodelled. This is illustrated by mutual conversion of hetero- and euchromatin, as observed originally by light microscopy; chromatin modification is actually subject to intensive studies (Grewal and Jia, 2007; Horn and Peterson, 2006; Kaeser and Emerson, 2006). A particular but little known phenomenon in relation to chromatin modification is "quantal mitosis", as defined by (Holtzer et al., 1975; Holtzer et al., 1972). The basis of the latter concept are observations on erythro- and myoblasts in early differentiation. If in DNA thymidine (T) is largely replaced by (BudR), hematopoiesis is fully blocked at specific steps of early embryogenesis, whereas in later steps of terminal differentiation, gene expression is fully normal, in spite of full substitution of (T) by (BudR). The latter is known to block induction of the lac gene in E. coli (Fried and Crothers, 1981; Wick and Matthews, 1991), since the dissociation constant of individual repressors is altered; by analogy, it might also prohibit chromatin remodelling in differentiating cells prohibiting exchange of chromatin proteins attached to the DNA, as suggested in the UMH (Scherrer, 1989).

After normal mitosis the chromosomal DNA will be unfolded into the holo-transgenon pre-existing prior to cell division; alternatively, the composition of transgenon-related factors might have been changed and contain new or modified MAR-binding proteins and other remodelling factors. In the latter case, this process might imply dissociation and re-association of trans-acting factors from the protogenon in cis, remodelling, hence, the chromatin locally. Expression of novel cP-genes in late G1 phase, time-programmed or induced by external factors, might control this process. Chromosome territories would be grossly maintained but within a territory, the structure of euchromatic chromatin altered.

In addition, the relation of neighbouring territories may be modified as well. Indeed, as EM observations have revealed, chromosome territories are interlaced (Nickerson, 2001; Nickerson et al., 1995) and, according to the phenomenon of "chromosome kissing" recently reported (Kioussis, 2005), distant DNA loops of different chromosomes have to interact to allow for some differentiation-specific gene expression ("3D-gene regulation"; cf. (Spilianakis et al., 2005)).

Once established as outlined above, within chromosome territories individual genomic domains will form local areas of euchromatin, where specific genes are localised and eventually will be transcribed. The organisation of such local domains will be influenced mainly by the condensation status of chromatin, modulated by histone modification. The DNA of heterochromatic areas being repressed they will, hence, not participate in the 3D organisation of the *local* chromatin network.

For the immunoglobulin gene domains, within this step of the cascade, prior to folding into a 3D network and eventual activation for transcription, the DNA itself is modified. Genomic regions containing the fragments of immunoglobulin genes are re-associated, by joining of the CJ and V regions; the latter, in addition, are sequence-modulated in function of particular antigens operating within a particular immune response (Lennon and Perry, 1990; Tonegawa, 1983). This implies permanent elimination of DNA regions and hence, part of the proto-genon of a genomic domain, under the influence of a particular holo-transgenon. Furthermore, this mechanism clearly illustrates the importation of information from the exo-system, along the concept of the "Exo-cascade" (see below and Figure 11).

The holo-transgenon of a given differentiated cell will provide non-histone and MAR-binding proteins and, accordingly, chromatin will fold into DNA loops. Furthermore, bi-functional MAR-binding proteins might crosslink such loops and structure the local 3D organisation of the DNA network, within and between chromosome territories, prefiguring the sites where transcription and processing factories will spring up.

The activated part of chromatin carrying the gene fragments is potentially DNase sensitive. In contrast, the nuclear matrix DNA protected by MAR-type proteins is highly resistant. About 1-2 % of nuclear DNA is DNase resistant, it includes repetitive DNA and is in general AT-rich. In this AT-rich fraction the sites binding MAR-protein are inserted; the latter sites are quite often GC-rich, but may be AT-rich as well. In polytene chromosomes, the MARs are inserted into the interband DNA separating individual gene domains, which represent the units of transcription (Fig. 6 B). In some biological systems, they correspond also to units of replication, in view of the local DNA amplification observed in some *Sciaridae* species (Lara, 1987; Santelli et al., 2004). Furthermore, interband DNA has some qualities of insulators, as defined by molecular genetics (Gaszner and Felsenfeld, 2006) and are, e.g. in the case of the drosophila gene *Gipsy*, visible in the cell nuclei after cytochemical staining (Gasser, 2002; Gaszner and Felsenfeld, 2006). As already mentioned, the interbands include systematically sites of Z-DNA formation (Nordheim et al., 1986); this fact is most interesting since it points again to the possibility of modification of the local chromatin structure s, possibly at the origin - or a consequence - of activation or inactivation of a genomic domain under the influence of the nuclear transgenon.

The fragments of DNase-resistant matrix DNA, the MARs, are clearly integrated into apparently organised filamentous networks, observed in the nuclei after exhaustive DNase treatment and extraction by high ionic strength buffers (see Fig. 7 c,d). In intact DNA, the MARs might hence constitute the organisational principle of the protein part of the nuclear matrix network, aligning sequential protein assemblies. Being placed at specific sites in the genomic DNA they may indirectly organise the 3D DNA-matrix backbone, as suggested in the UMH (Scherrer, 1989).

(5.2.2) Activation of Chromatin Domains for eventual transcription (Step 2 in Fig.10)

In this step, facultative heterochromatin may be transformed into euchromatin; but not all euchromatic DNA is by necessity transcribed, eventually. The classical criterion for chromatin liable to be activated is its DNase sensitivity (Razin et al., 1985; Stalder et al., 1980; Travers, 1999). DNase sensitive DNA segments can be actively transcribed once transcriptional repressors are eliminated, possible activators are present, and the RNA polymerase machinery is put in place. Typically, such domains remain DNase-sensitive after arrest of transcription (Groudine and Weintraub, 1981; Weintraub and Groudine, 1976). In terms of the genon concept, a selection among chromatin domains and individual transcriptional units eventually present in such domains, operates within their corresponding proto-genons in cis. Concomitantly, the holo-transgenon of a given differentiated cell is constituted whose factors may interact with the local DNA, upon local nucleosome decondensation.

Present consensus assumes the intervention of transcription factors (TFs) and promoters which might render the DNA liable for transcription. Recently, however, data supporting other types of interpretation appeared. Transcription factors, for instance GATA-protein binding sites, are spread all along a genomic domain of, e.g., the human or chicken globin domain (Cantor and Orkin, 2002;

Shimizu and Yamamoto, 2005); this is not logically compatible with a function in *initiation* of transcription but rather in support of elongation of the transcription products and facilitation of their progress to the processing machinery. Another possibility is that TFs might be part of the domain-specific nuclear matrix, which contributes to the liberation of the transcripts from the DNA and initiation of the transport system. Particularly clear are the recent data showing that the so called "promoters" of transcription may exert their action at the RNA processing level and not in initiation of transcription exclusively (Auboeuf et al., 2007; Auboeuf et al., 2005), as was believed for some time (see below) .

At this step of the regulation cascade, the holo-transgenon has to provide for the regulatory proteins which interact with the DNA at specific sites provided by the proto-genon in cis; and for the enzymes which locally modify the histones, by acetylation and methylation, altering the local chromatin in particular at the sites of transcription initiation, which are often situated at the level of the co-called locus control regions (LCRs) (Anguita et al., 2001; Flint et al., 2001; Tuan, 1989). A particularly interesting phenomenon is the attachment to specific DNA sites of proteins which are, possibly, later on transferred to and carried along by the transcripts. An early example of this process is the large T-antigen of polyoma virus and SV-40, which binds to the origin of DNA replication, but also to viral and cellular transcripts (Darlix et al., 1984). Interestingly, some MAR binding proteins were identified as previously sequenced pre-mRNP (or Hn-RNP) type proteins (von Kries et al., 1994). These types of proteins have a higher affinity to DNA than to RNA; their selection and sequential arrangement may, thus, take place at genomic level, and their transfer, from DNA to RNA, occur in course of transcription (c.f. below Chap. 5.2.3.2).

(5.2.3) The primary transcripts (step 3 in Fig.10)

In this regulative step, the information of the proto-genon of a genomic domain is reduced to that of the pre-genon, which is carried along by the RNA. The primary transcripts, which may include fragments of several genes, carry the cis-information for alternative processing of the transcript into one or several mRNAs, and their transport to the nuclear membrane in time and space. Under the control of the corresponding transgenon picked up by the RNA in formation, the primary pre-mRNP is formed. The latter, in terms of mass contains 3 times more protein than RNA. In situ hybridisation with probes for specific genes shows that partially processed pre-mRNA may accumulate first at the nucleolar periphery prior to moving to specific processing centres (Iarovaia et al., 2001), from where the mRNA is exported to the nuclear periphery along apparently specific tracks (see Fig. 7; and (De Conto et al., 1999; Iarovaia et al., 2001)).

These processes are controlled by the factors constituting the holo-transgenon of a given nucleus : presence or absence of specific TFs and of factors involved later-on in gene-specific processing (splicing) and transport, decides the fate of a given transcript in time and space.

(5.2.3.1) Synthesis of the FDT or primary pre-mRNA.

Formation of the primary transcripts starts with the local opening of the DNA double helix, at or in the vicinity of the LCR (locus control region) of a domain (Anguita et al., 2001; Flint et al., 2001; Tuan, 1989) under the influence of factors allowing eventual attachment of the complex of the RNA polymerases 1, 2 or 3. Concentrating here on pre-mRNA and polymerase 2, this process starts with the sequential attachment of transcription initiation factors (recent review in (Chen and Rajewsky, 2007)), among them the ubiquitous TATA binding protein. According to the signals in the proto-genon of such a domain, more specific factors will be picked up - if available - in the holo-transgenon of a given nucleus.

The formation of a primary transcript or FDT is in itself a multi-step process. Early experiments with the drug DRB have shown that a checkpoint exists after less than 1000 nt where the polymerase may stall, and fall off eventually (Egyhazi et al., 1999). Thus, some control may take place during the RNA elongation phase, most likely dependant on the presence or activation status of TFs .

(5.2.3.2) Association of nuclear RNA-binding proteins to pre-mRNA forming the pre-mRNPs.

As soon as the RNA is made it is covered by proteins to form the pre-mRNPs in *statu nascendi*. According to the genon concept, this happens under the direction of the cis-acting pre-genon, in function of the factors available in the particular holo-transgenon present in a given nucleus.

Basically, four types of factors may be distinguished that bind to pre-mRNP (for more details c.f. chapter 5.3.1) :

(1) Factors carried over from the DNA. According to recent data, these include promotor binding factors (Auboeuf et al., 2007; Auboeuf et al., 2005)as well as MAR-binding proteins (Darlix et al., 1984; Von Kries et al., 1991).
(2) The "classical" pre-mRNP (also called HnRNP) proteins of relatively basic charge (pI), the "histones" of the pre-mRNP; there are less than 50 components known (Dreyfuss et al., 2002; Maundrell and Scherrer, 1979).
(3) The acidic pre-mRNP proteins (Maundrell et al., 1979; Maundrell and Scherrer, 1979); proteomic analysis may allow to identify up to 500 components. Most of the factors of the processing and splicing machinery (Choi et al., 1986; Kim and Dreyfuss, 2001) as well as the transport (Kindler et al., 2005) and export factors (Rodriguez et al., 2004) are of this type.
(4) The ambivalent prosomes (Schmid et al., 1984)(review in (Scherrer and Bey, 1994)), the population of mRNP-binding particles (Mr 720.000) of variable subunit composition (which act also downstream of gene expression in proteolysis, as the core of the 26S proteasomes (Arrigo et al., 1988; Coux et al., 1996)}(Collins and Tansey, 2006)). Prosomes are part of (pre-)mRNPs (Martins de Sa et al., 1986), nuclear matrix (De Conto et al., 2000; Ioudinkova et al., 2005) and the cytoskeleton (Arcangeletti et al., 2000; Ioudinkova et al., 2005) (see Fig. 7).

Present in the nuclear sap or in specific compartments (e.g., the so-called "speckles" (Handwerger and Gall, 2006) constituting a pool of splicing factors), these factors represent the holo-transgenon from which an individual transgenon is picked up by a particular pre-genon. The composition at RNA level of the pre-mRNP factors is in constant modification since, during RNA processing, parts of the pre-genon are discarded, having fulfilled their function.

(5.2.3.3) Formation of the RNA-derived nuclear matrix by integration of the pre-mRNPs.

Concomitant to, or subsequent to formation of the pre-mRNP, the pre-mRNA "in statu nascendi" with the inherent pre-genon, is integrated into the nuclear matrix (see Fig. 7c,d; and c.f. (Ioudinkova et al., 2005; Nickerson, 2001)) . Of particular interest, in this context, is the fact that actin was recognised recently as a component and co-factor of all three types of eukaryotic polymerases (review in (Grummt, 2006; Haeusler and Engelke, 2006; Mayer and Grummt, 2006; Sims et al., 2004)), Since the nuclear matrix seems to be constituted by actin up to 30%, this fact points to the possibility that RNP formation and matrix integration may be simultaneous processes.

Interestingly, in the adult chicken the genomic region of the productively expressed adult globin genes alpha major and minor are relatively resistant to DNase, but not the embryonic gene pi which is transcribed abortively; in the transcriptionally silent final erythrocyte, the full globin domain was found to be DNase sensitive (Groudine and Weintraub, 1981; Weintraub and Groudine, 1976). This might be interpreted in the sense of a close association to the DNA-derived nuclear matrix with the transcripts in *statu nascendi,* liable to be transcribed and expressed.

Considering pre-genon and the corresponding holo-transgenons, their interplay will decide about the stabilisation, temporary storage or expression of the (fragmented) genes in the pre-mRNA. This process gradually reduces within the cascade of regulation the information content in terms of genes and genons present.

(5.2.4) Processing and differential splicing of pre-mRNPs. (step 4 in Fig.10)

As pointed out above, RNA processing and transport can be interrupted at several metabolic steps, resulting in the constitution of "peripheral memories" (see Fig. 3), in form of partially or fully processed pre-mRNA. Gene expression may, thus, be interrupted before the genes are constituted physically. Upon a specific signal, the transiently stored pre-mRNA may be processed and transported further. In steady state, such partially processed RNA turns slowly over in the nucleus, constituting the major fraction of nuclear RNA and a large part of the nuclear mass.

Processing of pre-mRNA represents the major regulative process in gene expression. Indeed, transcribed gene fragments are either temporarily stored in the nucleus or degraded, or else selected for productive splicing and transport of mRNA to the nuclear membrane. Indeed, during this process, 90% of transcribed sequence is eliminated either transiently or permanently (Kiss, 2006; Scherrer, 2003; Soller, 2006; Spohr et al., 1974). Accordingly, the pre-genon is also reshaped and reduced in information content down to the individual gene-specific genons, under the control of the factors of the holo-transgenon in a given nucleus.

Overall RNA processing occurs in steps. Early data showed that there are discrete steps in terms of size of the transcripts, RNA turnover times and sequence complexity. E.g. in avian erythroblasts,

RNA of very high Mr (among them globin RNA of up to 33Kb) with half-lives (t/2) of 20 min., intermediate size RNA with t/2 of 3 hours, and smaller nuclear RNA of up to 12 hours could be identified as 3 classes of decreasing RNA complexity (Imaizumi-Scherrer et al., 1982; Spohr et al., 1974; Spohr et al., 1976).

In positive correlation with these old findings on global RNA processing, recent *in situ* hybridisation data indicate that primary globin transcripts occupy diffuse, not clearly defined sites in the nucleoplasm, that a large part of the transcripts accumulate around the nucleoli when RNA processing is interrupted, whereas productively processed and exported globin (pre-)mRNA form two distinct processing centres (PCs). The highly unstable primary transcripts would hence end up in the PCs, where intermediary products of globin pre-RNA processing accumulate and transport to the cytoplasm starts (see Fig. 7 b). Between transcription and accumulation in the PCs, the presence of a site of transient peri-nucleolar residence is likely, although not observable when processing is normal and hence rapid; it is an old notion that the nucleolus plays a role in pre-mRNA processing (Deak et al., 1972; Hernandez-Verdun, 2006; Kiss, 2006; Maxwell and Fournie, 1995; Warocquier and Scherrer, 1969).

A most important feature of RNA processing concerns the nuclear matrix. As outlined above (chapter 5.2.3.3), the primary transcripts constitute the backbone of the RNA-derived nuclear matrix (see Fig. 7 c,d; and c.f. (loudinkova et al., 2005; Nickerson, 2001; Razin et al., 2004)). This mechanism ensures that every segment of RNA, with its associated protein complexes and enzymatic processing factors governing differential gene expression and site-specific transport, is placed in a precise position in space. This feature represents the ultimate justification for the, apparently, excessive size of nuclear transcripts (c.f. discussion in (Scherrer, 2003). During RNA processing, this 3D-organisation is continuously remodelled, when parts of pre-mRNA and, hence, pre-genons are gradually eliminated. The consecutive sites of residence of specific gene transcripts reflect this process, as well as the fact that every RNA fragment, ending up eventually in a gene, has to be handled individually according to time, physiological state and the dynamic architecture of the nucleus.

(5.2.5) Formation of the final pre-mRNP including exons of a single coding sequence. (step 5 in Fig.10)

During processing, eventually a pre-mRNA containing the exons of a single gene is formed containing, hence, a unique pre-genon. The most decisive mechanism operating at this step is differential splicing (Blencowe, 2006; Cuperlovic-Culf et al., 2006; Missler and Sudhof, 1998), and the differential choice of polyadenylation sites (Edwalds-Gilbert et al., 1997), as well as the involvement of untranslated regions in processing (Hughes, 2006). This implies either splicing, resulting in the differential composition of a final pre-mRNA with a unique set of exons or else, the separation of individual genes present as rows of consecutive exons in an FDT. The latter may be observed, e.g., for the globin genes (Broders and Scherrer, 1987; Broders et al., 1990) which form relatively stable and hence observable final pre-mRNAs (Therwath and Scherrer, 1982). In this process, intergenic transcripts are eliminated and, hence, part of the pre-genon.

The system controlling this process is once more the holo-transgenon of a given nucleus, which is modified according to cellular differentiation, during embryogenesis as well as in terminal differentiation. Concerned are the ubiquitous or partly selective factors and enzymes involved in splicing, among them the U-type small RNAs, resp. their RNP complexes. Less well known are the factors which govern the putative gene-specific splicing.

(5.2.6) Final processing of pre-mRNPs. (step 6 in Fig.10)

In parallel with pre-mRNP processing, part of the RNA is degraded. There is elimination of introns and intergenic RNA as the basic mechanism of processing. However, there is also elimination of part of the exonic and other functional RNA as a selective process under the control of the transgenon; RNA interference may also play a role at this level (Matzke and Birchler, 2005).

The final pre-mRNA is transformed into mRNA with its unique genon, ready to be exported to the cytoplasm. Accordingly, factors constituting a specific transgenon are by now associated with the mRNA. Final processing may be concomitant with export; e.g., the last intron of globin pre-mRNA is eliminated just prior to export . (In the general case, the nucleus does not contain mRNAs, and the cytoplasm no pre-mRNA). Though it is not clear at present if final processing entails by necessity export of the mRNA to the cytoplasm, nevertheless, a final selection step at this level has to be taken into consideration.

From the nuclear processing centres (PCs), mRNA is exported to distinct sites in the cytoplasm prior to being dispersed (see Fig. 7b). Many types of mRNA are then ubiquitously dispersed, as globin

mRNA , whereas others end up in specific cytoplasmic sites as, e.g., desmin mRNA in the sarcomeric discs of muscle cells (Fulton and Alftine, 1997). This selective transport through, possibly specific nuclear pores (Blobel, 1985), is guided by cis- and transgenon and involves the nuclear matrix, the cytoskeleton and cofactors of mRNPs, as for instance the prosomes.

(5.2.7) Import of mRNA into the cytoplasm (step 7 in Fig.10)

It is possible, although actually not established, that import of mRNA operates in a gene-specific manner. At this crucial step of the cascade - in view of the threshold of the nuclear membrane - qualitative selection and hence reduction of gene- and genon-specific information might operate.

The machinery of mRNA import is concentrated in the nuclear pore complex (Maco et al., 2006; Rout and Blobel, 1993). Already the first EM images of Hans Rees of RNA squeezing through the nuclear membrane showed, suggestively, a huge plug of RNA on the nuclear side being fragmented in the cytoplasmic compartment. In the meantime, the process of nuclear export has been analysed in many details (see review in (Cole and Scarcelli, 2006; Fahrenkrog and Aebi, 2003; Kutay and Guttinger, 2005; Rodriguez et al., 2004)). However, we do still not know about the biochemical composition, on either side of the nuclear membrane. of the mRNPs to be transferred. Theoretically, a transfer mRNP was postulated but never biochemically identified, in relation to the rather well characterised nuclear pre-mRNPs (Maundrell and Scherrer, 1979), and the repressed and translated mRNP complexes in the cytoplasm (Civelli et al., 1980; Maundrell et al., 1979; Vincent et al., 1980). More recent investigations show the implication of specific but rather ubiquitous factors involved in nuclear import and export; these operate at the level of the nuclear pores and seem to be in general non-discriminating for specific mRNA (Rodriguez et al., 2004). The biochemical nature of the mRNP complexes subject to these shuttling factors is unknown.

(5.2.8) Formation of cytoplasmic inactive (ribosome-)"free" mRNP (step 8 in Fig.10).

The final mRNAs entering the cytoplasm carry their unique genons which are exposed to the cytoplasmic holo-transgenon, allowing them to pick up sets of factors corresponding to their individual genons, resp. transgenons. This process results in an almost total exchange of mRNA associated proteins relative to those of the nuclear pre-mRNPs. A notable exception are the already mentioned factors involved in mRNA exportation which shuttle between both compartments. Furthermore, the prosomes are found on both, nuclear pre-mRNPs and cytoplasmic silent mRNPs .

The holo-transgenon as defined by proteomic analysis of silent mRNP complexes includes several hundred proteins, in their majority of rather acidic pI. The composition of factors in a given cellular compartment is in constant change in function of physiological adaptation, controlled by internal agents as well as by factors from the environment. The proteins directly attached to silent RNAs act as genuine cytoplasmic repressors (Civelli et al., 1980; Vincent et al., 1983; Vincent et al., 1981).

The advent of RNA interference has given a new dimension to cytoplasmic repression (Jackson and Standart, 2007; Sontheimer, 2005; Sontheimer and Carthew, 2005): siRNAs destroy mRNA in a gene-specific manner whereas the miRNA mask the mRNA transiently, in a manner similar to the RNP proteins.

(5.2.9) Activation of mRNA and polyribosome formation (step 9 in Fig.10)

Within the genon concept, mRNA activation is controlled by the factors available within the holo-transgenon of a given cytoplasm. There are - in competition - the selective repressive factors of the silent mRNP on the one side, and on the other the rather ubiquitous translation initiation and elongation factors associated to the translated mRNA. The existence of a third class of putative factors might be postulated on theoretical grounds; those selecting individual mRNAs to change their repressed or active status.

Many facts indicate that translation *per se* is a compulsory, constitutive mechanism. Translation factors are ubiquitous and present in relatively high concentrations in the cell sap, whereas the proteins associated to the repressed mRNP, as well as prosome subunits, are only found within the complexes and not in free form (Maundrell et al., 1979; Vincent et al., 1981). Therefore, cytoplasmic regulation might be essentially negative and depend on the biosynthesis, assembly and activation of repressing factors in a local holo-transgenon.

(5.2.9.1) Activation (de-repression) of mRNA by exchange of repressing mRNP proteins for translation factors, forming the translated mRNPs.

The (ribosome-)free cytoplasmic mRNPs outside the polyribosomes are translationally repressed, *in vivo* and *in vitro*. To render them *in vitro* translatable, the associated factors have to be almost completely removed . This implies that the factors of the transgenon associated to the repressed mRNA must eventually fall off, being either inactivated by chemical modification or allosteric effectors, or else removed by digestive enzymes as, possibly, the proteasomes (Baugh, 2004; Coux et al., 1996).

As already outlined, activation of mRNA is a reversible process; mRNAs may shuttle between the active and repressed states. If e.g. globin mRNA is translated to about 90% in erythroblasts, another abundant mRNAs in the same cells, the mRNA for the PABP (Poly(A) binding protein), is found to be 90 % repressed in terminal differentiation. Other mRNAs, as e.g. that for lipoxygenase, staying at constant level throughout reticulocyte maturation, is translated only during a short period, prior to being terminally repressed (Thiele et al., 1979). In nerve cells, mRNA is transported along specific axons; it is considered crucial that translation remains repressed after arrival at the destination site (e.g., a postsynaptic micro-domain) until an appropriate activation signal is received (for a recent review c.f. (Eberwine et al., 2002; Twiss and van Minnen, 2006)). Cytoplasmic repression is, thus, a crucial step of control of selective gene expression. In avian erythroblasts, by RNA complexity measurements, the presence in the cytoplasm of about 2000 different mRNAs was found, whereas only about 200 were actively translated, among them the globin mRNAs accounting for 90% of the protein output (Imaizumi-Scherrer et al., 1982).

If for repressed mRNP the presence of mRNA-binding proteins (Vincent et al., 1983; Vincent et al., 1981), and prosomes (Scherrer and Bey, 1994) in gene-specific sets was shown, as well as - in contrast - the ubiquitous nature of translation factors on translated mRNA (Civelli et al., 1980; Maundrell et al., 1979), the detailed mechanisms of selection of specific mRNA to be activated are unknown, as yet.

It seems possible that the proteasome system might play a role in mRNA activation, by liberating the 5'-side UTR for interaction with initiation factors (Maundrell et al., 1979; Olink-Coux et al., 1992). Since specific types of prosomes (also called 20S proteasomes), core of the 26S proteasomes, are associated with particular silent mRNA, it is tempting to speculate that such specific prosomes might be integrated into the 26S proteasome, the *in vivo* proteolytically active complex, to cleave repressive factors on an already selected mRNA (Baugh, 2004).

(5.2.9.2) Reversible formation of polyribosomes by association of the ribosomal subunits assembled into functional ribosomes.

Once the mRNA available for translation, first the 30S and later on the 50S ribosomal subunits associate to form ribosomes and the functional translation machinery. Since in steady state, translation factors do generally not discriminate specific mRNAs, there might be little intervention of the particular genon at this level.

However, when physiological conditions change, polyribosomes disintegrate. This is most spectacular in heat shock conditions. When Hela cells are brought up to 42°C, all polyribosomes fall apart within 20 minutes (Warocquier and Scherrer, 1969); within the same process the cytoskeleton and part of the prosomes disintegrate as well (Olink-Coux et al., 1992). If high temperature is maintained, polyribosomes reform partially within another 30 minutes; but protein output has changed, qualitatively and quantitatively.

Activation and inactivation of individual mRNAs must obey more subtle mechanisms. They may operate under the impact of changes of factors within the holo-transgenon of a given cell. Interestingly, the prosome-proteasome system may play a role in this process, eliminating selectively translation initiation factors from specific mRNA in translation (Baugh, 2004).

A major role is played by RNA interference in transient or final repression of specific mRNAs, either directly or indirectly. SiRNA and miRNA might block mRNA upon import or when mRNA segments become transiently accessible during translation. RNA interference is actually subject to most active investigation and no general conclusions seem possible as yet (see chapter 3.2.2.2).

The prosomes might participate in RNA interference as well; prosomes isolated biochemically or by immuno-precipitation contain up to 10% of small RNAs which have the capacity to block protein biosynthesis *in vitro* (Civelli et al., 1980) . Suggestive EM pictures were published at an early time showing polyribosomes and, interestingly, prosome-like particles associated wherever the mRNP chain emerges in between ribosomes (Fig. 12 and 14 in (Spohr et al., 1970); c.f. review in (Scherrer and Bey, 1994)), and ribosome-free mRNA, distinct of repressed mRNP, with only prosome-like particles attached were observed occasionally (Granboulan and Scherrer, 1969, unpubl. obs.). However, it is likely that other types of controlling factors participate in mRNA activation and inactivation by formation of the respective mRNPs.

(5.2.10) Translation of the coding sequence in mRNA. (step 10 in Fig.10)

Once translation has started, little regulatory intervention occurs in steady state that might involve genon and transgenon. Translation initiation is more temperature-sensitive than elongation; in less than optimal physiological conditions, ribosomes run off (Chezzi et al., 1971). Generalising this principle, one may speculate that in such conditions mRNA might be exposed to the transgenon for eventual regulatory interventions, according to the activity state of competing translation or repressing factors.

During translation, the rules of the genetic code and the translation machinery prevail by selection of triplets and assembly of the polypeptides. In steady state, the genon is hence put to rest as far as the coding sequence is concerned. In contrast, the 5'-side and 3'-side UTRs are likely to play a role by interacting with regulating proteins and interfering RNAs. Interestingly, polyribosomes have a tendency to form circles (e.g. (Christensen et al., 1987); as found at an early time by George Palade observing the so-called "rosettes" in the electron microscope - prior to the discovery of polyribosomes (Warner and al, 1963).

(5.2.11) Formation of the nascent primary polypeptide and higher order protein structure, gene function and protein homeostasis (step11 in Fig.10)

Once the polypeptide has formed the genon, by definition, expires and the factors of the protein world modulate the nascent polypeptide to assume secondary, tertiary and quaternary structure, which, eventually, will assume the genetic function based on one or a set of cooperating genes. These post-translational processes of gene expression and its control have to be most complex; we will here not enter these matters. However, it seems important to point out, that gene expression must obey homeostasis of protein biosynthesis and degradation. Mechanisms coordinating protein biosynthesis and catabolism must exist, by necessity.

The main operator in clearing misfolded, or otherwise defect polypeptides is the Ubiquitin-proteasome system (Coux et al., 1996); the proteasome core, the prosome or 20 proteasome participates as the key operator. Possibly, this 20S particle may shuttle between the mRNPs and the 26S proteasomes. Nature may, hence, have made the economy of still another system in charge of coordinating the biosynthetic and catabolic pathways (Scherrer and Bey, 1994). The prosome / proteasomes system in itself represents a complex machinery of differential action, due to the compositional variability of the basic core prosome particle.

Forming a molecular cylinder, the prosome has the capacity to interact bi-functionally at either end, as can be directly observed in stress fibres of the cytoskeleton (Arcangeletti et al., 2000; Arcangeletti et al., 1997). It may hence associate with mRNPs and simultaneously recognise cellular structures, as the nuclear matrix and the cytoskeleton of actin and intermediate filament nature (Arcangeletti et al., 2000; Arcangeletti et al., 1997; De Conto et al., 2000) (review in (Scherrer and Bey, 1994). When the prosome core eventually integrates the 26S proteasome complex, target protein recognition is delegated to the 19S modulator complexes associating at both ends, which opens the proteolytic chamber, and shields the prosome surface from external interactions. The ubiquitin system identifying proteins for degradation (recent review in e.g. (Ciechanover, 2006)), upstream of the 26S proteasomes, as well as the 19S modulator complex which, in an ATP-dependant manner unfolds, gates and actively feeds doomed polypeptides into the proteolytic chamber, may secure gene-specific catabolism.

(5.3) The transgenon, the trans-acting program carried by the factors acting onto a given (proto/pre-)genon placed in cis

For this discussion we exclude all mechanisms directly related to constitutive and basic protein biosynthesis within the frame of the genetic code as, e.g., the ribosomes and the basic tRNA machinery.

The cis-genon as outlined above is materialised by the ensemble of factor-binding sites within an individual mRNA. These sites are recognised by protein or RNA factors supplied by the program in trans. The factors selected by a single genon constitute its specific *transgenon*, which is available - or not - within the *holo-transgenon* of a given cell, nucleus or cytoplasm (Fig. 4).

By definition, the *holo-transgenon* corresponds to the *hologenon* of an organism or a single cell; concerned are all factors, might it be protein or RNA, able to respond to the cis-program encoded in DNA or RNA and related to a gene to be expressed. We need to distinguish here between the hologenon of an organism and that of specific cells in that organism, because of the presence of

differences in their genomes. Not only carry the cells of the immune system particular adapted genomes, but also other differentiated cells may incorporate genetic modifications like transpositions in their DNA. In addition, epigenetic effects as well create differences between cells affecting the expression control exercised by the genon and its precursors.

Regulation of transcription, and hence of programs of differentiation and physiological change, is in part under the influence of cell-external signals (see the "Exo-cascade" formulated in Figure 11). Genomic systems have been generated and gene expression evolves in function of the ecosystem. Controls from the environment dominate also regulation of cellular gene expression, although some constitutive cell-internal expression programs are carried out. Signals from the outside touch off synthesis of, e.g. transcription factors, influence their activity status or trigger their shift from the cytoplasm to their targets in the chromatin to be activated (Wu et al., 2006; Zhu and McKeon, 2000). They may, hence, also largely control the generation of all the factors which influence the fate of the transcripts on the gene expression pathway.

The genon is embedded in the pool of trans-acting factors recognised by the receptor oligomotifs in cis. The presence of these factors is hence crucial for the execution of the expressions program encoded into the genon. In addition to being passively picked-up by the oligomotifs in cis, these factors have a discriminative regulatory function as well. Their presence or absence controls the implementation of the cis-program; they may, furthermore, be in active or inactive state. Since proteins, like logical gates (as utilized in computers or electronic chips), are capable of integrating many types of input, small MW agents may influence directly or as allosteric effectors the factor-receptor interactions.

The transgenon, carried by cP-genes and cR-genes, is built up by the normal mechanisms of gene expression and regulation, leading to the synthesis of DNA- and RNA-binding proteins, the synthesis of siRNA and miRNA within the frame of RNA interference and of all other types of cR-genes which might affect differential regulation of gene expression.

(5.3.1) Nucleic acid-binding proteins as carriers of the transgenon

Proteins cover all types of RNA in the cell. In case of mRNA and pre-mRNA, it was shown at an early time by electron microscopy that proteins are aligned all along the RNA molecules (Dubochet et al., 1973), protecting specific sequences from degradation by RNase (see Fig. 4 and (Goldenberg et al., 1979)). By mass, messenger ribonucleoproteins (mRNPs) include 3 times more protein than RNA. One of the roles for such proteins is to protect RNA from degradation by different types of RNases, which are natively active and abundant in the cellular sap; naked RNA is hence rapidly destroyed.

RNP-type proteins bind in a RNA-sequence dependant manner. The poly(A)-binding proteins (PABPs), attached to the 3'-side tail (length : 50-200 A residues) of the mRNA protect about 12-20 A-residues at a time (Baer and Kornberg, 1983) ; larger RNA-binding proteins may cover up to 50 nt, as is the case, e.g., for the viral large T antigen of SV-40 and Polyoma virus (Darlix et al., 1984). The RNP-type proteins include amino acid sequence motifs, recognising the sequence motifs in DNA and RNA. The mRNA, hence, includes sequentially such protein-binding oligomotifs. Actually, only in rare cases the oligo-nucleotide sequence is known that binds a given protein. In addition to the PABP one might mention, e.g., the IRE-BP (Iron Response Element binding protein), a protein binding a motif in the 5'-side or 3'-side UTR of the mRNA, respectively for Ferritin and Transferrin (Thomson et al., 1999). There is not even a conventional term to name such motifs in RNA and DNA; we hence introduced the term "*oligomotif*" (Scherrer and Jost, 2007). For the recognition motifs in proteins, one might use the term "*aa-motifs* ", reserving that of oligomotif for DNA and RNA.

Early proteomic studies of 20 years ago allowed people to estimate that there are several hundred (up to 1000) acidic, non-histone proteins attached to DNA, and as many to pre-mRNA and FDTs (Maundrell and Scherrer, 1979). In the cytoplasm of a given cell, there may exist 500 - 1000 different proteins in the repressed mRNPs; a specific mRNA binds a specific combination of such proteins (c.f. review in (Scherrer and Bey, 1994)). When the ribosomes are split off the mRNA in polyribosomes *in vitro*, the translated mRNA is found also in mRNP form, which includes in mass 3 times more protein than RNA. The proteins characteristic for the translated stage are of much fewer - a few dozen - types and seem to be ubiquitously bound to different mRNAs, recognising in particular the 5'- and 3'-side UTRs; a translation-specific PABP binds the Poly(A) tail (Edmonds, 2002; Grossi de Sa et al., 1988; Shatkin and Manley, 2000).

These observations indicate that there must be a "code" governing the interaction of a limited number of NABPs in chromatin and mRNPs which, in general, are specifically DNA- or RNA-binding proteins. Relatively new data have confirmed, however, the old observation that the same protein may bind both, DNA and RNA, as outlined above. This was originally observed for the large T-antigen of SV 40 and polyoma virus (Darlix et al., 1984; Khandjian et al., 1980) and more recently confirmed for a series of DNA-binding MAR proteins (von Kries et al., 1994)), by sequence identified as - already

known - pre-mRNA binding proteins (pre-mRNP or HnRNP proteins); this may represent an interesting exception rather than the rule. The existence of different rules for binding of proteins to DNA and RNA must be assumed. Within this discussion we may use the terms of ***DNP- and RNP-code*** (Auweter et al., 2006) for the system governing protein-NA interactions. Specific binding may be by sequence-motifs of about 15nt minimal length for which we have coined above the neologism of "oligomotif". An oligomotif would thus interact with an aa-motif in a protein; this interaction can occur either directly, or imply a mechanism of "induced fit". In terms of information processing, in this exchange, the holo-transgenon generating the binding factors would represent the sender, and the oligomotif in DNA and RNA the receiver of these signals, acting according to the rules of chemical thermodynamics.

In addition to the signals encoded in the oligomotifs of the primary RNA sequence, there are post-transcriptional modifications of the transcripts (review in (Shatkin and Manley, 2000)) which may be recognised by the factors in trans. There is internal methylation of mRNA (Perry and Kelley, 1975; Perry and Scherrer, 1975) as well as "Capping" (review in (Banerjee, 1980)), of the 5'-side of (pre-)mRNA (consisting in the 5'-5' addition of GTPs including differential methylation), and there is polyadenylation of the 3'-side (Munroe and Jacobson, 1990). $Poly(A)_{50\text{-}200}$ is recognised by the PABPs of several types which are different in case of the nuclear pre-mRNA, the cytoplasmic repressed and, eventually, the translated mRNA (Mangus et al., 2003). The existence of a poly(A) tail and the corresponding PABP is a factor essential for translation (Gorgoni and Gray, 2004; Grossi de Sa et al., 1988). During processing, the transcripts may be cleaved and the site of scission recapped and polyadenylated; primary transcripts may extend far beyond the aauaaa polyadenylation site (??). Indeed, there are enzymatic systems known to add monophosphates and the 5'-5'-triphospates on the 5'-side (Barbosa and Moss, 1978; Venkatesan et al., 1980), as well as poly(A) on the 3'-side of the processing products (Shatkin and Manley, 2000). By definition, all these processes are post-transcriptional; thus, the enzymes carrying out these modifications are to be accounted for as factors involved in the generation of the trans-program, implemented by the holo-transgenon of a cell.

The transgenons carried by cP-genes are constituted by the normal mechanisms of gene expression and regulation by protein biosynthesis.

(5.3.2) RNA interference and the transgenon

The second mechanism - recently discovered - of transient or final repression of specific mRNAs is RNA interference. Si- and miRNAs might block mRNA upon import to the cytoplasm, or during translation when mRNA segments become accessible as pointed out above. The phenomenon of RNA interference is at present most actively investigated and no general conclusions seem possible as yet. Actually, little could be said with any chance of precision, beyond the general considerations outlined above (see chapter 3.2.2.2).

It is however evident that RNA interference represents at the same time a highly gene-specific system of control, liable to recognise precise RNA targets. It is hence at the same time more efficient but also less sophisticated than the regulatory protein factors. Indeed, the latter are capable to integrate controls to a much higher extent. The si- and miRNAs may represent primitive slots operating in an on/off mode. But this system as well has to be managed upstream by protein factors, not only enzymatic system involved in its generation, but also mRNP proteins. Being single as well as double stranded, interfering RNA is a target for any type of (pre-)mRNA binding protein as well.

RNA interference is likely to have evolved prior to RNA-binding proteins, possible already in pre-biotic systems. RNA hybridisation is the most basic process of RNA stabilisation and neutralisation. Later, chemically more sophisticated systems of RNA-protein recognition and mutual stabilisation may have evolved, much before the tRNA based protein-coding revolution happened, opening the gate to life and evolution.

(6) Mathematical Analysis of Genetic Information and Gene Expression

(6.1) General considerations

The proposition of the Genon concept is not only thought to redefine the gene in unambiguous terms and allow better comprehension of gene expression and regulation; the ultimate goal is to provide a scheme clear enough to allow us the application of mathematical methods in analysis of

genes and genomes. Here again we have to separate the definition of the gene *per se* from the programs that guide their expression in time and space.

The restriction of the definition of a gene to the coding sequence, constituted by the assembly of coding triplets, considerably facilitates the development of algorithms in view of mathematical analysis; as we will see below, the approach to be taken seems quite straightforward. It is evident, however, that the gene as a function represents more than the coding sequence and its equivalent in terms of the nascent polypeptide. Chemical modifications and the formation of secondary, tertiary and quaternary protein structure are not exclusively encoded in the primary amino-acid sequence; external factors as well govern the assembly of the structures underlying the functions expressed within the phenotype. Therefore, additional programs must exist which control this process; some programs may entirely or largely be encoded in a given genome but in addition, factors from the ecosystem seem to play a major role in the final gene function.

In line with our general conceptual decision of taking translation as the cut-off point, we here only take into account pre-translational processes and restrict gene expression to the formation of the primary structure. Nevertheless, the analysis to be presented can in principle also be extended to post-translational events.

In addition to the gene *per se* just mentioned, our information theoretical analysis will be concerned with the program of gene expression, that is, with the genon. Again, it is natural to begin with the program in cis, i.e., the ensemble of genon-related signals encoded in DNA and RNA. After that, we turn to the holo-transgenons, i.e., the ensemble of factors from trans, either provided by the genome or the environment of cell and organism, that interact with the cis-program.

The scope of this task can be seen by an overview of the types of decision-making programs that will come into play, following the mechanisms of gene expression exposed above.

The first program in cis to be considered is the final genon itself, as encoded in the mRNA, which carries the information governing the expression of its gene; this analysis concerns essentially pre-translational controls, as transport of mRNA, its cytoplasmic activation or repression, and the effects of co-translational factors. Logically, this analysis has to be extended to the pre-genon carried by the primary transcript of a genomic domain and will, hence, concern RNA processing (including splicing), post-transcriptional repression and storage in the nucleus, as well as eventual activation and transport to the nuclear membrane. Within the nuclear cis-program, in addition to the pre-genon as defined above, we have also to take into account the proto-genon, including additional cis-signals encoded in the DNA which serve chromatin activation and onset of transcription. Of course, the pre-genon is included in the proto-genon; but the latter includes signals that operate at DNA level exclusively. Among those are the sites where transcription factors attach, operating at the level of genuine promotors and, possibly, some types of enhancers.

Formulating algorithms of control one has to take into account, furthermore, the fact that some decisions are made at DNA level which are born out at pre-mRNA level only; indeed, some proteins binding to specific DNA sequences are carried over to the pre-mRNA. Most often not taken into consideration, this is an important basic mechanism, which makes possible the sequence-related assembly of proteins with high affinity for DNA, and hence binding specificity, which, once assembled, are carried over to the RNA in *statu nascendi*. In addition, it has to be pointed out that, at the level of analysis and comprehension actually possible, it is not clear in many cases whether so-called promotor and enhancer effects really bear on transcription (most "transcription" tests are actually based on translation products, as CAT or luciferase), or rather on the stabilisation of transcripts and the efficiency of their expression at the level of translation.

The analysis of the trans-program is obviously more difficult than that of the cis-program since the trans-program includes a rather heterogeneous set of factors that interact with the cis-genon encoded in the mRNA. This makes it necessary to utilize a classification according to the different types of factors.

The first step of that classification distinguishes factors produced by the genome itself from factors provided by the environment. The genome produces DNA and RNA-binding proteins as well as the small RNAs involved in RNA processing and the recently discovered RNA interference (RNAi). External factors provided by the environment include mineral ions, chemical compounds not produced internally (as some vitamins), diverse sources of energy, light (as a source for photosynthesis or as a signal for circadian rhythms), gravity (providing for example a gradient for spatial diffusion according to weight), etc. In between these two types of factors are the ones produced by other cells in a multicellular organism, like hormones, cytokines, and other secondary cell messengers. Here, for simplicity, we shall concentrate on genome-dependent trans-factors.

On one hand, we have to take into account those factors that physically interact with the cis-genon, and on the other hand, we have those that modulate the action of DNA- and RNA-binding factors in an indirect manner. Examples include protein-protein interaction and ionic or allosteric modulation, as well as interaction by cell-external factors like cytokines.

In order to appreciate the logic of the formal analysis, it is advantageous to start with biological simplifications, and approximate biologically realistic scenarios only gradually. In this regard, one might hence start with a single genon in a given mRNA that has available all possible trans-factors occurring within the holo-transgenon of a given genome. The genon in the mRNA then only needs to select the appropriate trans-factors corresponding to its specific transgenon. The situation becomes more specific when the mRNA carrying the genon in cis is immersed into the cytoplasm of a given specialized cell. It then encounters only about 500-1000 RNA-binding protein factors with which the perhaps 20-50 signals in the specific cis-genon interact. Instead of varying the specificity at the trans-side, we may also turn to the cis-side and consider, instead of a single genon in cis, an entire hologenon that is exposed to all the factors in the trans-program of a cell, as in the case of sperm DNA entering the ooplasm of an egg.

(6.2) The questions

The general question to be asked in terms of information theory concerns the information content, at the various and subsequent levels of gene storage and expression, of a gene as a product as well as the result of the expression program that led to its eventual realisation. Standard analysis is concerned with the amount of information about the biochemical identity of a polypeptide contained in its coding sequence. That, however, takes such a polypeptide out of its cellular context. First of all, a polypeptide is not simply read off from a coding sequence stored somewhere in the DNA, but, as we have amply explained, it is the result of an intricate regulation process leading to the coding sequence at mRNA level prior to translation. This involves contributions from regulatory elements in cis as well as from factors provided by trans, and this should also be conceptualized as a sequence of information processing steps, and as such, it should be made amenable to an information theoretical analysis. Secondly, what is biologically relevant is not simply the biochemical identity of a protein, but - in addition to its spatial shape which, however, is not addressed here - its relation to other proteins, in terms of numbers and differences between types. Taking these issues into consideration, our analysis will deal with the reduction of uncertainty about the outcome of gene expression, following the steps of the expression program just outlined, and where this uncertainty is quantified in terms of different possible polypeptides within some biologically relevant class. However, the genome is not the exclusive source of information guiding this program; as outlined in chapter (5.3.3) and Figure 11, there is continuous influx of information from the Exo-system, surrounding a given cell as well as an organism ("ecosystem"). Therefore we have to develop our analysis from the genome to the product as well as from the periphery of organism and cell to the genome.

As we will see, different formalism will apply to the "forward" and the "backward" analysis in terms of input from the genome, or from the exo-system, the latter bearing essentially on the holo-transgenons (excluding input in the frame of evolution). For any such analysis, it is essential to specify what one assumes as known and what one wants to know.

A clear-cut illustration of this problem is the number of different polypeptides imaginable within the rules of the genetic code: there are 64 triplets (minus 2 - the start and stop codons) coding for 20 amino acids. Assuming average length of a polypeptide of, say, 500 amino acids, the number of all combinatorial possibilities is astronomically large, much beyond any range that evolution could have possibly explored. There are essentially two ways out of this impasse: to assume rules of possible sequence correlations or else, to put into the game the proteome as derivable from the sequence analysis of genomes published. Practically, these approaches have their limits since, in both cases, our knowledge is approximate, at best. Therefore we will have to resort to experimentally founded assumptions to carry out this analysis. Concerning the human genome, e.g., we may hence assume the existence of about 500.000 different polypeptides to be potentially expressed, and up to one million gene products altogether, counting sR and cR genes and taking into consideration RNAi.

Entering our analysis, for a polypeptide actually expressed in a cell we can ask about the sequence at DNA or RNA level that is coding for it; this is the classical application of information theory to molecular biology. It deals with the selection of a given gene and leads to the issue of the degeneracy of the genetic code. Another aspect of this question is the localization of that coding sequence in the DNA.

Our main interest here, however, concerns the opposite direction, that is, going forward from a (piece of) coding sequence in the DNA to the polypeptides (or other functional products) that it will eventually get expressed in. For such a coding sequence at DNA level, we already know the amino acid that each triplet is coding for. Looking only at this sequence we do not know, however, whether, and if so, when, where, and in which quantity that sequence is expressed in the cell under

consideration. Thus, there is some uncertainty here, and we shall be concerned with quantifying that uncertainty.

In order to perform this quantification according to the rules of information theory, we need to specify the options available. Thus, we need to list those polypeptides in which our sequence could possibly be expressed. (Of course, in a particular situation at hand, it may not get expressed at all; this is one of the possible options.) It is now important to realize that there are some choices to be made here; we have to agree about what prior knowledge we already wish to admit. If we do not wish to admit any prior knowledge, we need to consider all combinatorially possible amino acid sequences (up to some specified length). As already pointed out, this is a very large number. We may therefore wish to impose some restrictions, in order to reduce the number of options and to include only more realistic ones according to the given cellular condition. We could restrict ourselves to consider only the amino acid sequences that are biochemically possible in the sense that they can give rise to well folded proteins, or to polypeptides that have been identified in some cell and are listed in some data base. We could even assume more prior knowledge, namely that we consider only those polypeptides that occur in the proteome of the organism in question. Or, finally, we could restrict our considerations to the ensemble of polypeptides present in the cell at the time of investigation. In any case, whichever of those ensembles we choose, the uncertainty then consists in identifying which member of the ensemble in question is realized by the expressed coding sequence, and also in which quantity. If there were no regulatory mechanisms like alternative splicing, silencing, or other decisions on the expression pathway, the expressed product itself would be completely specified by the (fragmented) coding sequence at DNA level. Still, however, the number of expressed copies is not yet determined. Repression mechanisms at various stages of the expression pathway could result in no expression at all, whereas repeated transcription/translation or other multiplicative steps could result in multiple products. Finally, mechanisms like alternative splicing even make it impossible to predict the biochemical identity of the expressed product from the coding sequence alone.

In the sequel, we shall set up the information theoretic scheme to quantify these uncertainties and to assess the relative contributions of the coding part, the gene, and the regulatory part, the genon, in resolving these uncertainties. Thus, the total information, needed to specify the types and numbers of functional products produced from a giving coding region at DNA level, is a sum that will be decomposed in the parts attributed to the gene and the genon. Numerical estimates (presented elsewhere in detail) will show that the by far larger part is the one coming from the coding sequence, whereas the contribution of the genon is rather small. As genon and transgenon are rather complex, involving many binding sites in cis and binding factors from trans, this indicates that the genon is doing more than just providing this little amount of additional information to resolve some ambiguities about the products derived from a coding region. Among the contributions of the genon not quantified here is the regulation in space and time, that is, the contribution of information about when and where in the cell some gene is to be expressed, in addition to type and amount of product. Another aspect is the stabilisation of expression in a fluctuating milieu with unpredictable external perturbations. According to Ashby’s law of requisite variety, entropy, that is, information, is needed to compensate for those fluctuations and perturbations, and that information then will be not visible in the final product; but it must be provided by genon and transgenon.

(6.3) Information theory and molecular biology

(6.3.1) The concept of information

For the purpose of applying information theory to gene expression, we should first discuss the concept of information itself. Our starting point will be the theory of Shannon. In that theory, a sender composes a message from the elements of a code agreed upon with the receiver. The receiver knows the probabilities p_i with which the individual code words i appear, but apart from that, he does not know the content of the message before receiving it. Thus, before receiving the message, his uncertainty about the actual content of the message to be received is given by the entropy

$$I = -\sum_i p_i \log p_i \tag{1}$$

where we take the binary logarithm (that is, $\log 2 = 1$).[1] The standard convention $0 \log 0 = 0$ expresses the fact that no information is gained from events that cannot occur. Also, when some $p_i = 1$ (and consequently all the other $p_j = 0$), we use $1 \log 1 = 0$, meaning that when we are already certain about an event, we do not gain information either.
The formula (1) leads to equating the information content of the received message with the uncertainty present before receiving it, and that uncertainty is quantified by the entropy (1). In that sense, the information received is a reduction of uncertainty. Uncertainty, as expressed by the entropy (1), is converted into information. This is similar to a conversion of potential energy into kinetic energy. The reduction of one is the gain of the other, and the quantities are the same, and are measured in the same units. Thus, for our purposes, entropy is equivalent to potential information.

(6.3.2) Ensemble and sequence entropy

In our applications to molecular biology, we shall be concerned with sequences (of nucleotides or amino acids). For such a sequence, we want to know its composition, that is, we want to know which element (nucleotide or amino acid, resp.) occurs at each position. This is the information we are after. For formalizing this, there exist two alternative approaches, and in this section, we want to discuss those. One approach consists in simply taking the set of all possible sequences under the given circumstances as an ensemble and then quantify how much information is needed to specify a particular sequence within this ensemble. The other approach looks at the individual positions in the sequence in turn and quantifies how much information is needed to specify which nucleotide or amino acid occurs at that particular position. When we do this for each position and take correlations between the various positions into account, we can again quantify the information needed to determine the composition of our sequence. We shall now describe these two approaches in more formal terms. Suppose that we are given an ensemble of N items of M different types x with relative frequencies or probabilities $p(x)$.[2] The information about the size of the ensemble is given by $\log_2 N$. Since this is so simple we shall mostly suppress it in the sequel. The **ensemble entropy** is then given by

$$I = -\sum_x p(x) \log p(x) \tag{2}$$

Without further knowledge, all the relative frequencies $p(x)$ should be assumed equal, according to Jaynes' principle of maximal ignorance, and

$$I = \log M. \tag{3}$$

This is the maximal possible value of I, given the number of types. Refinements through additional knowledge then decrease the entropy; examples include

[1] The negative sign in front of the sum arises here to make the whole expression positive, because the p_i take values between 0 and 1, and therefore, their logarithms are negative. Equivalently, we may write $I = \sum_i p_i \log \frac{1}{p_i}$, that is, absorb the minus sign inside the logarithm.

[2] "Relative" here expresses the normalization $\sum_x p(x) = 1$.

- observations of relative frequencies, restriction of the ensemble, or
- encoding of regularities, or
- physical considerations, where we have some kind of an energy function, in the terminology of statistical physics a Hamiltonian H that leads to the Boltzmann-Gibbs distribution

$$p(x) = \frac{1}{Z} e^{-\beta H(x)}. \tag{4}$$

(Here, the factor Z, the so-called partition function, serves to achieve the normalization $\sum_x p(x) = 1$, and β is a factor that regulates how strongly differences in the value $H(x)$ of the Hamiltonian translate into differences in probability.)

In molecular biology, we are not working with arbitrary ensembles, but often with ensembles of sequences, and for such ensembles, there is an alternative approach to entropy. Let S be a sequence of length n of "symbols" a drawn from an "alphabet" A of size $|A|$, occurring with relative frequencies p_a. Each position in the sequence then has entropy $I_{pos} = -\sum_a p_a \log p_a$. Without further knowledge about sequence regularities, S has entropy

$$I_S = nI_{pos} = -n \sum_a p_a \log p_a. \tag{5}$$

Here, without further knowledge, all the p_a are equal ($= 1/|A|$) so that

$$-\sum_a p_a \log p_a = \log |A|, \tag{6}$$

and

$$I_S = n \log |A|. \tag{7}$$

Since there are $M := |A|^n$ different such sequences, this is the same as the ensemble entropy $\log M$ above, cf. (3). Again, refinements through additional knowledge decrease entropy; examples include

- unequal distribution of the p_a, in which case (5) becomes smaller than (7),
  or
- sequence correlations leading to the consideration of block entropies

$$-\sum_\nu p_\nu \log p_\nu, \ \nu = \text{block of length } l. \tag{8}$$

The block entropies become smaller than the entropy (5) when the probability of occurrence of a symbol at a particular position also depends on the symbols in its vicinity. In other words, sequence entropy can get decreased when the symbol probabilities are context dependent.
One should note, however, that the computation of block entropies is numerically feasible only for relative small values of the block length l. This is not quite as bad as expected because by the Shannon-MacMillan-Breiman theorem, the effective number of blocks is

$$2^{l\ I_{pos}} \tag{9}$$

instead of the larger number $|A|^l$ of all possible blocks. Also, iterative computation in terms of increasing block length allows for exploiting regularities efficiently. Below, we shall briefly consider this both for nucleotide and amino acid sequences.

Ensemble and sequence entropy represent two different ways of computing the same quantity, and they should therefore yield the same value. Estimates for these quantity, however, can be different, because they will employ different aspects. Thus, whereas in the case of uniform probabilities, the values (3) and (7) coincide, in other cases the estimates for the sequence entropy can yield much larger values than the ensemble entropy. The reason is that it is difficult to capture all the regularities present in an ensemble through sequence correlations, as long range correlations are not easy to track and numerically expensive to include.

(6.3.3) Applications of information theory to molecular biology

The application of information theory to molecular biology has been controversial. To clarify the issue, the following point might be useful. Usually, information theory is applied to messages. A message contains information when before receiving it one does not know the sequence of symbols in the message, that is, once the message is known that previous uncertainty is reduced. Shannon's information measure quantifies that reduction of uncertainty, that is, the difference in knowledge before and after receiving the message. This suggests that, likewise, a stretch of DNA contains information about polypeptides or phenotypic properties because knowing that DNA sequence allows one to deduce the composition of those polypeptides or those phenotypic properties. Of course, because of the intervention of other factors, the knowledge of the DNA does not lead to complete knowledge of the relevant polypeptides or phenotypes. The point is, however, that knowing the DNA reduces the uncertainty about those polypeptides or phenotypes, and this then leads to a quantification of the information contained in the DNA. The remaining uncertainty then is assigned to other factors, and the corresponding information can then also be quantified.

The point we are emphasizing here in order to avoid misconceptions about genetic information (see e.g. [4] for a recent discussion) is that for quantifying information one needs to specify first about what there is uncertainty. Uncertainty about sequence identity is different from information about types and numbers of polypeptide chains in a cell, and consequently, the information content is different as well. Below, we shall treat those different situations in turn. More precisely, we shall look at sequence, product, and process information. In each case, the information measure will be different.

In information theory, the message from the sender to the receiver has to pass through a channel, and the latter may not faithfully transmit everything emitted by the sender. The channel may introduce noise, that is, random distortions or modifications of the message. Also, there may be systematic effects decreasing the information content of the message. Different messages may be received as the same message. This is called redundancy. Redundancy can have the positive effect of error tolerance, in the context of a triplet coding for an amino acid meaning that certain mutations do not affect the amino acid coded for. Indeed, for the receiver, it does not matter which of those different messages have been chosen by the sender as long as the received message remains the same. Thus, the sender can make some errors as long as they do not change the message for the receiver.

In particular, the genetic code is redundant in the sense that the genome as the sender emits nucleotide triplets while the proteome as the receiver obtains amino acids, and several triplets of different chemical composition lead to the same amino acid.[3]

The application of information theory to molecular biology, however, should go beyond the relationship between individual nucleotide triplets and amino acids. A nucleotide and an amino acid not only have their specific chemical identity, but they are also parts of sequences, the DNA sequence, or a polypeptide chain constituting (part of) a protein. As such, in addition to their chemical composition, they are characterized by their position within that specific sequence. Moreover, the relationship between such a triplet in a specific position and the amino acids coded for by that triplet is not a relationship between individual physical objects, insofar as in a given cell, the triplet is usually expressed several times, and in different polypeptides. Each amino acid produced from the triplet can be considered as a physical instantiation of this particular triplet, and of no other triplet. There are many chemically identical triplets in the DNA, but the given amino acid as a concrete physical object is derived from precisely one such triplet.

Considering it that way, however, falls short of understanding the expression process, and if that were all that information theory can contribute, its usefulness would be rather limited. While in principle we can follow a specific expression pathway and trace the origin of a given amino acid back to a single triplet at its location in the DNA, the formation of that amino acid requires additional ingredients along the expression pathway. Some ingredients come from the cis DNA region containing that triplet. For instance the nucleotide sequence encoded in a promoter region is also needed, and enhancer and repressor sequences affect the expression.

[3] When looking at finer details of the regulation process, however, that redundancy dissolves. For example, the splicing process depends on certain recognition sites in exonic regions for the formation of certain RNPs, and here, triplets that translate into the same amino acid can be functionally different. Also, even at the translation stage, the frequency of translation depends on the presence of the appropriate tRNAs, and the more frequent triplets might also have more tRNA partners and are therefore also more frequently translated. Thus, different frequencies of triplets coding for the same amino acid can make a functional difference in the cell.

Factors in trans, which are specific for the intra- and extracellular environment, also guide the expression. The point in time within the processing sequence also affects the outcome. Thus, the relationship between specific individual chemical units is superseded by processing information that does not implement itself physically in the final product. So, on one hand, when tracing the process back in time, we have a relationship between individual chemical substances determined by their locations within specific sequences, while on the other hand, when going forward in time, we have the combination of cis and trans ingredients determining in which and in how many numbers of polypeptides a given triplet is expressed.

(6.4) Product information

6.4.1 Information in cis

6.4.1.1 The coding sequence

We have four different nucleotides, A, C, G, and T, of which DNA sequences are composed. When each of them occurs with relative frequency p_i (i =A,C,G,T), each position contributes an information of

$$I_{nuc} = -\sum_{i=A,C,G,T} p_i \log p_i \text{ bits} \tag{10}$$

In particular, when they are equidistributed, that is all $p_i = 1/4$, this information is 2 (bits). When all positions in a sequence of length N are independent, the sequence information then is $I_{seq} = N\ I_{nuc}$. Sequence correlations, however, will decrease that information. To make this precise, we need an ensemble of sequences s, and we consider subsequences of length l [4] in this ensemble and the block entropies $-\sum_\nu p_\nu \log p_\nu$ (summing over all such subsequences ν of length l, denoting their relative frequencies in our ensemble by p_ν) and let the length l become sufficiently large to capture all such sequence correlations. (In principle, such an analysis is even meaningful when our ensemble consists of a single sequence only, as long as l remains small compared with the sequence length; in any case, the maximal value of l for which the block entropies can be computed in practice is rather small, about $l = 12$)[5]

We now consider triplets of nucleotides as such triplets are the subsequences coding for amino acids. In particular, when all 64 triplets (including the 2 start/stop codons) are equally frequent (and hence, also the nucleotides are equidistributed), each such triplet contains an amount of $I_{tri} = 3I_{nuc} = 6$ bits of information.

There are 20 amino acids out of which polypeptides can be composed; we denote the relative frequency of an amino acid referred to by the index α by p_α. An amino acid thus on average requires for its specification an information of

$$I_{aa} = -\sum_\alpha p_\alpha \log p_\alpha \text{ bits.} \tag{11}$$

When all these frequencies are equal, $I_{aa} = \log 20$.

Due to the degeneracy of the genetic code which leads to redundant coding for amino acids, the information needed to specify an amino acid is smaller than the one contained in a triplet ($\log 20 < \log 64 = 6$).

Again, sequence correlations will decrease that information. In an ensemble of polypeptide chains, we consider subsequences of length l and the block entropies $-\sum_\sigma p_\sigma \log p_\sigma$ for subsequences σ of length l with relative frequencies p_σ and let the length l become sufficiently large to capture as many sequence correlations as feasible. Again, the maximal value of l for which the block entropies can be effectively computed is rather small.[6]

When f denotes the relation between triplets and amino acids, that is, $f(r) = \alpha$ when the triplet r codes for the amino acid α, and if we put $p(r|\alpha) = \frac{p(r)}{\sum_{\rho: f(\rho)=\alpha} p(\rho)}$ (conditional probability for a triplet r given the amino acid

[4] Here, biochemically, one should think of oligonucleotides; for example, $l = 2$ means pairs of nucleotides, $l = 3$ triplets, and so on.

[5] For $l = 12$, for instance, it seems that one needs to count the frequencies of 4^{12} different subsequences. As mentioned above, however, typically already for smaller values of l, not all 4^l possibilities are realized, and one can use such findings in an iterative manner to reduce the number of possibilities that one has to check for larger values of l.

[6] In fact, in the investigations of B.L.Hao and his group, it was found (personal communication) that going beyond $l = 5$ (pentapeptides) or 6 (hexapeptides) yields very little additional information and in practice rather obscures patterns.

α it codes for) when $f(r) = \alpha$, the mutual information between the collections of individual triplets and amino acids is given by

$$I_{tri,aa} = I_{tri} - (-\sum_{\alpha} p_{\alpha}(\sum_{r:f(r)=\alpha} p(r|\alpha) \log p(r|\alpha))). \quad (12)$$

The second term on the right hand side of this equation is the average of a function of the amino acids, where these amino acids are weighted with their relative frequencies. That function is the uncertainty for a given amino acid about the coding triplet. This term thus is the conditional entropy, that is, information, for the triplets given an amino acid, and it quantifies the redundancy of the genetic code. By symmetry of the mutual information (an elementary mathematical result, see [3]), $I_{tri,aa} = I_{aa,tri}$, the information gained about an amino acid from knowing a triplet. Since a triplet specifies a single type of amino acid, this expression in turn simply equals I_{aa}, the information contained in an amino acid.
Again, since there are sequence correlations, the average information needed to specify a polypeptide consisting of n amino acids is smaller than n I_{aa}. Thus, also the mutual information between nucleotide and polypeptide sequences will be different from n $I_{tri,aa}$.
We should point out that here we have computed the mutual information between the chemical compositions of amino acids and polypeptides on one hand and triplets or nucleic acid sequences on the other hand. It is a different question to infer the location of such a triplet in the DNA sequence given the chemical identity of an amino acid or polypeptide.
So far, we have presented the standard application of information theory to molecular biology. This, however, is of rather limited use, and we shall now proceed in a different direction, more in line with the general aims of this paper.

(6.4.1.2) Positional information within the coding region

When we consider the formation of an amino acid or a polypeptide, not only the chemical identity of the coding triplets is relevant. There are many chemically identical triplets in the DNA, but only one out of those is the origin of a specific given amino acid in a peptide . That triplet can be characterized and distinguished from others by its position in the nucleotide sequence constituting the DNA. This leads us to the information needed to determine that position. The position can for example be described by a coding region or maximal ORF, the sequential number of an exon within that ORF, and the position inside that exon, in an analogous manner as one localizes a word in a book by specifying a chapter, a page within that chapter, and a position on that page. Thus, when considering an individual amino acid, one can quantify the positional information about the location of the triplet coding for it in the DNA. Of course, this information cannot be derived from the triplet in isolation. Not only is the chemical identity of that triplet ambiguous because of the redundancy of the genetic code as analyzed above, but there are also many chemically identical triplets within the DNA. Therefore, the corresponding entropy, that means uncertainty, is rather high.
If we use context information, however, the situation changes. The context information comes from the polypeptide chain the amino acid is contained in. Typically, when we know such a polypeptide chain we can uniquely and unambigously identify the coding regions and exons where it is derived from, and inside such an exon, we can then also identify at DNA level the triplet from which our amino acid is expressed. (There exist exceptions to this due to the phenomenon of gene duplication, that is identical genomic regions located at different positions in the DNA coding for the same functional product.) We should point out, however, that we are assuming here that the DNA sequence as such is known and the only uncertainty is about the location of some triplet within that sequence. We have already discussed above (6.4.1.1), how to quantify the sequence information of the DNA. That information then is assumed to be known in either case considered here, that is, both, when we only know the amino acid in question and want to determine the position of the triplet in the DNA from which it is derived or when, in addition, we have the knowledge about the polypeptide chain containing that amino acid at our disposal. Thus, the alternative is between specifying the position of a triplet in the DNA sequence simply by counting, as described above, or using context information, that is, identifying the polypeptide chain containing the amino acid.[7] That latter information will be discussed in 6.4.2 below. It depends on what class

[7] assuming, for simplicity, that then the coding region in the DNA is uniquely determined; in any case, even though that need not strictly hold, the number of possible coding regions for a given polypeptide chain is rather small, and therefore, there is little remaining uncertainty

of polypeptide chains the analysis is based. In other words, we need a list of relevant proteins. The information contained in that list, as usual, depends on what we assume as known, for example certain biochemical rules that exclude some amino acid combinations, the species to which the organism in question belongs, or a specific cell type.

(6.4.1.3) The ensemble of products derived from a coding DNA region

We now leave behind the standard application of information theory to molecular biology and come to an important issue. In a living cell, from one single coding region or ORF, often a large number of polypeptides is produced, and those may be of different types, because of differential splicing and other regulation processes. Thus, we should not consider the relationship between a single DNA region and a single polypeptide, but rather the one between such a single DNA region and an ensemble of polypeptides. It is here that, in regulation, the program that we are calling the genon enters and provides specific information about the final product from our coding region that is not contained in that coding region itself; the information theoretic analysis should separate these respective contributions. Also, the contributions from the cis and trans programs interact here, and they should then be quantified in information theoretic terms.

For evaluating the information provided by the genon, we need to consider the ensemble of polypeptide chains produced under specified conditions, for example those expressed in a given cell or those that can be expressed by the genome in question. Here, when we speak of an ensemble, we always mean a collection of physical objects. These objects may belong to different types, but typically, types are represented by several of such objects, that is, not all of the objects represent different types. Thus, our ensemble consisting of individual physical objects is characterized by the types x to which these objects belong and their relative frequencies p_x with which they occur among these objects (plus an integer for the absolute size of the ensemble, as the p_x are defined as relative frequencies and not as absolute ones). Thus, we consider the ensemble of polypeptide chains produced under some specified conditions, and for each type x of polypeptide represented in the ensemble, we let p_x denote its relative frequency. We compare these relative frequencies p_x of the different types x with the relative frequencies q_x with which they can be derived from the DNA region containing the exons of the coding sequence under consideration. Of course, most q_x will be 0 because any coding region can be expressed only in a small fraction of the polypeptides present in the cell or derivable from the genome. We consider now the difference

$$I_{cis} = -\sum_x (p_x \log p_x - q_x \log q_x). \tag{13}$$

Here, the first term is the uncertainty about a polypeptide when we do not know the coding DNA sequence, whereas the second term, which is negative, that is, subtracted from the first one, quantifies the remaining uncertainty when we already know that coding DNA sequence.
In view of the preceding, we expect that I_{cis} is quite large. In that sense, our coding region encodes a lot of information about functional products. We emphasize once more that this quantity depends on the ensemble (x, p_x). As explained in the introduction, information is measured as a reduction of uncertainty, and therefore, we need to specify first about what there is uncertainty. In principle, we could consider all biochemically possible polypeptides x, even though it might be difficult to assign probabilities p_x to them. That is the situation where we don't admit any information about the genome or cell in question. We could also be more specific and admit some of the latter information. In that situation, our initial uncertainty is smaller because we already have some knowledge about which polypeptide chains could possibly occur. Therefore, the knowledge of the coding region tells us less that we did not already know than in the previous situation. This issue will again be taken up in 6.4.2.
There is one remark of fundamental importance here: While the term $-\sum_x p_x \log p_x$ in (13) is rather arbitrary because it depends on the choice of the ensemble of possible x, like all combinatorially possible, all chemically possible polypeptides, or all polypeptides occuring in a given organism or cell (an issue to be returned to below), the other term, $-\sum_x q_x \log q_x$ is not arbitrary at all, because it is derived from the frequencies of the products derived from our coding region under given circumstances. It is this latter term that is important for us and to which we shall turn in the next section.

In any case, the reader should note that compared to the beginning where we have discussed the coding information contained in a sequence, we have now completely shifted the perspective. In (13), the contribution of a cis coding region is now a residual term that is obtained by subtracting from an ensemble entropy the contribution of the regulation by genon (and transgenon).

(6.4.2) Information provided by the genon in an ensemble of functional products derived from a coding region in the DNA

We have quantified the types and numbers of polypeptides derived from a given coding region (genomic domain) by the second term in (13), that is,

$$-\sum_x q_x \log q_x. \tag{14}$$

This information cannot be found in the coding region, but is rather provided by the (proto-, pre-)genon (and the transgenon, a distinction to be addressed below). We now wish to analyze that genon contribution for the transition from the coding region to the gene in terms of information theory. In order to simplify the presentation, we start with a triplet of nucleotides in the DNA and follow its expression path. Along this path, regulation by other factors will determine the fate of the transcripts, i.e., its products, and we shall understand that as an information contribution.

Within the total protein content of a cell, we consider the ensemble of amino acids derived from the given triplet in the DNA. Whereas the chemical structure of these amino acids is the same, their number, that is, the number of copies derived from the same triplet, may vary. In addition, due to differential regulatory effects on the expression pathway, for instance differential splicing, these amino acids may find themselves in structurally different polypeptide chains. The corresponding types we identify by the index x. The information content of this ensemble of polypeptide chains now depends on what we are ready to assume as given.

Before listing some possibilities for quantifying that information content, we recall a general observation from our above discussion of the entropy: When we have a collection of physical objects, we can either list them as such, or we can seek regularities, for example identify types represented by several individual objects, to achieve a more compact representation. In the sequel, we shall begin with the naive list and then proceed to a representation in terms of types x and their relative frequencies p_x.

1. Explicit description of all physically present polypeptides in the cell containing an amino acid derived from the triplet under consideration. When no further regularities are taken into account, this becomes $n_0 I_{aa}$ where n_0 is the combined length of all these chains.[8] Of course, this is only a coarse, and not very helpful, upper estimate of the necessary information. For example, when the ensemble contains several copies, say m, of one particular polypeptide of length n, then the corresponding information can already be described by at most $\log m + nI_{aa}$ bits instead of the typically much larger number of mnI_{aa} bits. Also, as already discussed above, we can exploit sequence regularities for the individual polypeptide chains to get below nI_{aa} bits for such a chain.

2. The preceding used the class of all possible types of polypeptides. This class, however, is too large to distinguish between the different information contributions. For determining the contribution of the protogenon, we should use the class of all polypeptide chains that can be produced from the same coding sequence in the DNA, under a set of specified trans conditions. Likewise, at the level of the pre-mRNA, the possibilities are already more reduced, and the selection between them is now governed by the pregenon. At the level of the mRNA, it is then the genon that is responsible for selective gene expression. Since the same type of information theoretic analysis can be applied at each level, we shall now discuss the protogenon. The pregenon and the genon then can be handled analogously, by simply replacing the different coding regions in the DNA eventually contributing to the final product by those present in the unprocessed pre-mRNA or the unique one in the mRNA.
   So, we return to the ensemble of products that can be derived from a given coding region in the DNA.

[8] This is made more precise in the Shannon-MacMillan-Breiman theorem which tells us that the effective number of different sequences that need to be considered is $2^{n_0 I_{aa}}$ which unless the 20 different amino acids are equidistributed is smaller than 20^{n_0}. The other sequences occur with negligibly small probability. When we take sequence regularities into account, this effective number gets smaller.

Each type x of polypeptide chains present in that ensemble has a relative frequency q_x, and the average information gained by observing a specific such polypeptide chain (pc) then is

$$I_{pc} = -\sum_x q_x \log q_x. \tag{15}$$

If only one type of polypeptide chain is produced, this information vanishes. In that case, the product resulting from our triplet is already completely determined by cis, ignoring at this point the contributions of the program in trans. In order to take account of the important biological fact of non-expression, in particular by repression, the setting should be refined by also allowing for the possibility that no polypeptide at all is produced. Formally, this is handled by also including the empty polypeptide in the collection of types x, and assigning the appropriate probability to non-expression. Thus, whenever our triplet is contained in an exon and can be expressed, the ensemble has at least two members, one corresponding to suppression of expression, the other(s) to successful expression. Thus, I_{pc} is non-zero, except for the cases where the triplet either is never expressed or where it is always expressed in the same polypeptide. Since in (15), we are considering only relative frequencies, this expression does not yet capture the full information of the ensemble because it does not reflect its size, that is, the total number ν of polypeptide chains present in our ensemble. Therefore, we should refine (15) as

$$I^0_{pc} = -\sum_x q_x \log q_x + \log \nu. \tag{16}$$

There are three essential steps involved in going from our triplet to the polypeptide ensemble with entropy given by (15) or (16). The first step is the transcription which is a multiplying step in the sense that it determines how often the triplet is transcribed. (Of course, for the pregenon or the genon, this step is no longer relevant.) Out of the genomic region containing our triplet, a certain number of transcripts is produced. All those transcripts have the same composition, and thus, here only a factor, but no diversity is produced. The next step is the regulation taking place on the expression pathway. Here, no multiplication takes place as the final mRNA is formed from those transcripts (except for the indirect effect that certain RNAs might be processed faster than others originating from the same coding sequence in the DNA). The regulation here can be enhancing as well as repressing. Details have been discussed in the preceding chapters. The key point is that here, at the end of the regulation process, the diversity of the final ensemble is determined. The final step, translation, again yields a multiplicative factor as the number of times a given mRNA is translated, but no further diversity because the final mRNA already completely determines the composition of the polypeptide.

We can perform the same type of analysis for larger cis regions than triplets, for example for DNA domains containing fragments of coding sequences or ORFs. The information measures will differ when we have overlapping ORFs, that is, when one triplet belongs to several ORFs as in the case of alternative splicing or other forms of differential processing. In that case, the uncertainty about the products derived from the triplet needs to take the uncertainties about the final products about all those ORFs into account.
In particular, we can then compare the information provided by different DNA domains and thereby specify the information content of the protogenon. Let us consider a sequence s of nucleotides in the DNA, for example again a coding triplet, to start with the smallest unit relevant for the present purpose. As explained above, we have the uncertainty about the ensemble of polypeptides containing a piece of a polypeptide chain, like an amino acid in the case of a triplet, derived from s, expressed through the conditional entropy defined as

$$H(x|s) := -\sum_x q_x \log q_x. \tag{17}$$

Now, when we know a longer sequence S containing the original s, then we can compute the corresponding quantity $H(x|S)$ (where the x as before stand for those polypeptide chains that contain amino acids derived from s) which now is smaller because the additional information of $S\backslash s$ (the rest of S when s is taken away) now makes more specific predictions of the polypeptide chains possible. The important quantity expressing how much the fate of the products derived from s is constrained by the surrounding region S then is

$$H(x|s) - H(x|S). \tag{18}$$

Given s, and thus the ensemble determined by (x, q_x), we can then let the surrounding region S vary and detect from (18) the amount of allo-determination of the products derived from s.
We are now in a position to assess the information content of the protogenon. Here, we take as s (fragments of) the coding sequence for some gene, while S is a larger DNA region containing regulatory elements or other protein binding sites that are not part of the coding sequence, introns etc. The expression $H(x|s) - H(x|S)$ from (18) then quantifies the information contribution of the part of the genon captured by S.
When we wish to analyze a specific transgenon and its information contribution, then, instead of adding some further cis elements to the original sequence s, we now take some factors from trans. Then the analysis proceeds as developed above for the cis genon.
In any case, when s again is our coding region, the uncertainty

$$H(x|s) = -\sum_x q(x) \log q(x) \tag{19}$$

is precisely the amount of information about the products derived from s that comes from outside s, that is from the genon and its precursors, from the transgenon, and from external factors. By varying S, we can then quantify the various individual contributions.

Before proceeding, let us briefly make the following remark: Whereas here we have considered the situation for **P**-genes, the case of **R**-genes can be handled by the same type of analysis. Futhermore, let us recall that our analysis dealt only with the assembly of sequences during processing and differential splicing, leaving aside all other regulative interventions controlling gene expression in space and time.

(6.4.3) Sequence information of the genon

There is a different, but somewhat coarser, method of estimating the information provided by the (proto-, pre-)genon. To see this, we consider a polypeptide and look again at the case of the protogenon; we shall ask about all the DNA sites that contributed to its formation, that is, both, the coding triplets and the ones from regulatory regions that guide the process leading to that polypeptide. In 6.4.1.1, we have already studied the sequence informations for the polypeptide and the corresponding coding region in the DNA. By the same method, we can then also evaluate the sequence information of non-coding regulatory sites, both in cis and in trans, i.e., either present in the cisgenon or provided by factors from the transgenon. The former include stop codons, enhancer, promoter, repressor sites, introns that play a role in the expression pathway as binding sites for certain proteins, and the like. The relevant part of the holo-transgenon derives from the coding regions for transcription factors and all other proteins regulating or interfering with the expression pathway.
There is a problem with this approach, however. The reason is that many of the regulatory elements, while being specific to a certain degree, need not only affect the polypeptide under consideration, but also interact with the regulation of other polypeptides. Therefore, the simple sum over the sequence entropy of all contributing sites seems to overestimate their specific information content. Putting it another way, when we consider two different polypeptides, we are not allowed to simply add the corresponding sequence entropies because some of the factors may contribute to both of them, leading to an overestimate for the information needed for the two polypeptides. Nevertheless, this approach might be useful in deriving some upper bounds for the information needed to produce a polypeptide.

(6.5) Process information provided by the genon

In this section, we want to investigate the information theoretic aspects of the genon, accompanying the potential gene on the expression pathway, from a different point of view. For that purpose, we shall analyze the relative contribution of cis signals and trans factors to the information needed to express a specific gene. The basic situation is that the cis region provides certain control signals, like enhancers at the DNA stage or binding

sites for proteins forming RNP complexes at the RNA stage, whereas those binding factors then constitute the transgenon.

(6.5.1) The genon in cis

The contribution of the cis region with its combination of binding oligomotifs consists in a preselection of the possible binding elements at the particular site under consideration, out of all the proteins in the cell that can bind to DNA or RNA. We first consider one particular site s in cis, and assume for the moment that precisely one protein can bind at that site. Let p_y denote the relative frequency of the RNA or DNA binding protein y in the cell, and let q_y be the relative affinity of y at the site under consideration. For most y, q_y will be 0, because binding requires a special affinity to the site in question. We consider thus the quantitiy

$$I_s := -\sum_y p_y \log p_y + \sum_y q_y \log q_y) \tag{20}$$

where the first term represents the uncertainty about a protein in the cell without knowing the binding site, whereas the second term, which again is negative, that is, subtracted from the first one, represents the remaining uncertainty when we know the binding site, that is, when we only consider those proteins that could possibly bind at that site and their binding affinities. In view of the preceding, I_s is expected to be rather large, and this expression quantifies the specificity of the site.

The basic case from which to start thinking about the genon is where the whole expression pathway is solely controlled by cis, in the sense that all necessary factors are provided by the program represented by the transgenon, and any specificity is entirely due to selection by cis of binding factors. In that case, all $q_y = 0$ or 1, and (20) becomes

$$I_s = -\sum_y p_y \log p_y. \tag{21}$$

We have thus considered the most elementary situation. When it comes to processes of differential expression, for example in response to external signals, the ensemble of trans factors becomes variable, and therefore, we need to assign non-trivial q_y to some factors, and so we are back to (20).

Still, this needs to be expanded in two directions. First, a cis region can, and typically does, contain more than one protein binding site. When the binding properties of these sites are independent of each other, we can simply sum the expression given in (20) over all those sites, to get the process information content of that cis region. Such an independence holds when one only considers linearly RNA binding polypeptides.

In other situations, we need to modify this expression by taking correlations into account as in the previous sections. Second, there is an important combinatorial aspect because at one site, usually not a single polypeptide is binding, but some combination of such polypeptides that then biochemically form a quartenary protein complex. Furthermore, some other proteins facilitate or inhibit the binding of certain other ones. Therefore, instead of single proteins, we need to consider protein combinations, as in a language where instead of individual phonems or letters, one rather takes morphems or words as basic elements. The principle expressed in (20) still applies when one substitutes protein combinations for isolated proteins.

In summary, the process information content of a cis region is quantified by a reduction of possibilities. Therefore, it cannot be computed directly from the nucleotides forming the region, but rather depends on the proteome in the cell. This may seem paradoxical, namely that we cannot compute the information contribution of a DNA region by looking at the nucleotides, but rather need to compare the number of possible binding proteins with the smaller number of those actually capable of binding to that particular region. Of course, it is determined by the latter's nucleotides which proteins can bind there and which ones can't, but in order to do the computation we need to know which trans factors are the candidates.

(6.5.2) The contribution of the transgenon

Conversely, the information contribution coming from trans simply consists in the selection of those factors that actually bind to a given (proto/pre)genon, out of those possibilities allowed by the structure of the signals in the DNA domain as composed by its nucleotides. Thus, here the difference is between those that can possibly bind, given the concrete nucleotides, and those that are actually provided by the holo-transgenon of the given cell. Returning to (20), the uncertainty left after evaluating the information provided by cis is the term $-\sum_y q_y \log q_y$ incorporating the affinities (or the corresponding expression taking into account all the binding sites of a given cis region and their combinatorics, that is, the combination rules for the binding of several different trans factors at neighboring or otherwise related sites). The cis region allows certain proteins to bind, but it does not completely specify which ones will actually be bound. That selection is the important trans contribution that constitutes the regulation process. Therefore, when a particular protein has bound to a particular site, that site has gained an information $-\sum_y q_y \log q_y$. According to our information theoretical scheme, that information is not assigned to the cis site, but rather considered as provided by the transgenon. In the terminology of information theory, cis here is considered as the receiver for a message sent by trans, and that message consists in the specification of the binding protein.

(6.6) Conclusion

The crucial entropy (14)

$$-\sum_x q_x \log q_x \tag{22}$$

expressing the information contributed by the genon to a given product is typically quite small because the number of different products that can be derived from a given ORF or transcript is rather limited (detailed numerical examples will be presented in a subsequent paper). On one hand, it is much smaller than the term $-\sum_x p_x \log p_x$ in (13), making I_{cis} large. On the other hand, it is also much smaller than either the sequence information or the process information of the genon. Thus, it seems that a considerable loss of information is occurring between what is present in the genon and what is remaining in the product. We have already discussed another loss of information, quantified in (15), from the information contained in a triplet to the one expressed in an amino acid. That loss of information comes from the redundancy of the genetic code. The standard explanation of this phenomenon is that pairs of nucleotides can specify at most $4\times4 = 16$ amino acids, so that one needs triplets which then could specify $4 \times 4 \times 4 = 64$ amino acids, whereas only 20 are needed. In other words, the coding scheme here necessitates that more alternatives are potentially available than actually required. That redundancy can then be positively utilized for a certain error tolerance. For example, the coding of several amino acids is not sensitive to some mutations of the third position in the triplet. Also, portions of the coding sequence can form oligomotifs for the binding of proteins, and here different triplets while encoding the same amino acid could bind different proteins. Thus, the redundancy of the genetic code can be positively utilized for regulatory purposes.

The case of the genon seems different. First of all, in our computations of information, we have ignored an important aspect of the contribution of the genon. The genon not only decides what is produced among the alternatives provided by the coding sequence at DNA level, and in which quantities, but also at which place in the cell and at what time, within development and differentiation and the cell cycle, it is produced. In principle, one could also quantify this in information theoretic terms. For that, one would need to identify the spatial and temporal scale at which significant differences within the cell and its life occur.

Another explanation for the apparent information loss concerning the genon can be offered in terms of Ashby's law of requisite variety ([1], p.202ff). That law is concerned with control or regulation in the presence of external perturbations. The aim of that control then is a reduction of variety, in order to keep the system as close as possible to the goal state. In other words, in spite of perturbations with high variety that could affect the system's internal state, the system should be kept in a state of low variety. Thus, control should prevent the transmission of variety from the environment into the system. Hence, control seeks to reduce variety, in contrast to information transmission that aims at conserving variety. Active control then has to compensate each disturbance by a suitable counteraction. In particular, it needs to react differently to different

perturbations. Therefore, at least as many different counteractions are required as there are disturbances, and the internal variety of the control must be at least as great as the external variety of disturbances to be compensated. This then might also provide an explanation for the difference between the large sequence and process information contained in and provided by genon and transgenon and the small entropy contributed to the ensemble of products. Genon and transgenon achieve robust regulation in a setting of many external influences and perturbations, and the large sequence and process information might be required in order to maintain concentrations of vital polypeptides and proteins in a manner that is adapted to the state of the cell's environment, but stable against disturbances. The difference between the sequence or process entropy of the genon on one hand and the product entropy it contributes then expresses the amount of control and regulation of gene expression achieved by the genon. More precisely, this yields an upper estimate, as we do not know whether the efficiency of the genon is optimal. General evolutionary considerations might suggest, however, that the control and regulation is not too far from being optimal.

References

[1] W.R.Ashby, An introduction to cybernetics, Chapman and Hall, London, 1956

[2] M.Barbieri, The organic codes, Cambridge Univ.Press, 2003

[3] Th.Cover, J.Thomas, Elements of information theory, Wiley-Interscience, 1991

[4] U.Stegmann, Der Begriff der genetischen Information, in: U.Krohs, G.Toepfer (eds.), Philosophie der Biologie, Suhrkamp, Frankfurt/M., 2005, pp.212-230

(7) Concluding Remarks

In this paper, we have developed a definition of the gene that conceptually separates the gene as a product, from the genetic information relating to the regulation of gene expression, the latter being defined within the genon concept (Scherrer and Jost, 2007). In particular, we give up the notion of the correspondence of the gene as a functional unit and as a DNA locus. Classically, in the work of Mendel, Morgan and including Benzer, the gene had been considered as an inheritable function and basis of genetic analysis. In the sixties, knowledge about its physical basis in terms of DNA led to a picture where gene and DNA locus were equated. Such a picture, however, is simplistic because it ignores the basic fact that many steps of gene regulation are necessary to transform a genomic sequence into a collection of functional products. Crucial information necessary for this regulation process is also stored in the DNA, but obeys a different code. The gene-product is determined by the genetic code and the mechanisms of protein biosynthesis whereas regulation generally is subject to sequence-related macromolecular interaction, producing higher order complexes of DNA and RNA, involving formation of RNA-protein complexes or hybrids with regulating RNAs. Thus, both the codes and the biochemical mechanisms behind translation into a product and regulating transcription and expression, while interrelated, are clearly distinct. Therefore, a conceptually clear and practically useful gene concept needs to distinguish these two types of information, product versus process information, gene versus genon.

This emphasis distinguishes our approach both from DNA sequence based definitions in the wake of the human genome sequencing project that lost the functional aspect out of sight, and from more recent definitions that are motivated by the ENCODE project (ENCODE Project Consortium, 2007) that aims at a systematic description and classification of transcripts and lead to a conceptual hybrid between coding and functional aspects and attempt to omit regulation entirely from the gene concept (Gerstein et al., 2007; Gingeras, 2007).

To put it differently: Since there are two distinct aspects involved in the production of a collection of polypeptides from coding fragments in the DNA, namely translation of triplets into amino acids, and regulation of the assembly of those sequences of triplets from the initiation of transcription to the final mRNA prior to translation, two distinct concepts are needed. One is the gene that then becomes freed from all ballast and can again assume a pure role of a functional unit, and the other is the genon that guides and controls the assembly of the gene through the steps of the expression process.

Let us try to put our conceptual framework into perspective. Our information theoretical analysis is entirely sequential, as it is motivated by the principle of the cascade of regulation, and it integrates well a substantial body of biochemical knowledge and theoretical concepts accumulated about genome organization and gene expression (c.f. (Scherrer, 1980; Scherrer, 1989; Scherrer, 2003)).It does not, however, take the complex network of interactions between the expressions of different genes into account. This still needs to be addressed within the concepts and methods of Systems Biology; the high throughput data currently - or soon - available will be needed here.

In any case, it seems that a conceptual and information theoretical discussion of the gene has its natural point of termination at the stage just prior to translation when the coding information is read off from the mRNA, a limit adopted within this essay. After the sequence identity of a polypeptide has been determined, physical and biochemical processes take over to determine the shape in 3D of proteins as well as their spatial localization and co-localization within the cell. This then constitutes the basis of the metabolic functioning of the cell. It will be a fundamental task for the future to integrate the information-theoretic analysis developed here, which finds its natural place in the transcriptome, with a geometric approach concerning both, the proteome as well as the transcriptome.

(8) Glossary and Abbreviations

(in italics : terms in glossary)

Biological Terms

Aa-motif: short amino acid sequence interacting with a nucleic acid *oligomotif*

Alternative *splicing*: in course of *splicing*, exons can be combined in different ways so that in the subsequent steps of the expression process different functional products *(genes)* can be created from the same pre-mRNA

Cistron: contiguous genomic element acting in cis to secure a function

Controlling gene (c-gene): gene controlling the expression of other genes

Cascade Regulation: theoretical model of eukaryotic gene regulation proposing stepwise reduction of the genomic information potential in course of RNA processing and transport

Ectopic pairing : network of filaments (some known to contain DNA since running in and out of the *nucleolus*) which run in between telomeres, and link specific interbands of the four *polytene chromosomes* of Drosophila, e.g., forming a genetically fixed 3D-network which keeps every genomic fragment in a precise position in space; conceptual basis of the *Unified Matrix Hypothesis.*

EM: electron microscope

Exon: fragment of a coding sequence in the DNA placed between *introns*

FDT: full domain transcript, RNA resulting from the transcription of an entire genomic domain in the DNA; generally but not necessarily identical to *pre-mRNA* or *pre-rRNA.*

Gene: here defined as the uninterrupted nucleic acid stretch of the coding sequence in the mRNA that corresponds to a polypeptide or another functional product; thus, in eukaryotes typically not yet present at DNA level, but assembled from gene fragments (exons) in course of RNA *processing*

Genomic domain: DNA domain containing fragments of one or several genes coordinated by cis controls separated, possibly, by insulators (Gaszner and Felsenfeld, 2006), often unit of transcription and, in some cases, of replication; visible as chromatin loops in lampbrush chromosomes of birds and amphibia, and in polytene chromosomes of diptera as heterochromatic bands, representing structural units of chromosome organization and meiotic recombination, of transcription and - e.g. in Sciaridae - of replication.

Genon (contraction of gene and operon): program controlling the expression of a gene, superimposed onto and added to the coding sequence in cis, i.e.: cis-acting program associated with a specific gene at mRNA level, materialised by factor binding sites (*oligomotifs*) in an mRNA sequence, therefore encoded already in the DNA in the same strand as the coding sequence, which is fragmented into *exons* (see text for details)

Holo-genon: ensemble of all *(proto-)genons* at the level of the entire genome

Holo-transgenon: ensemble of all factors that can respond to the cis-program encoded in DNA or RNA and related to genes to be expressed

Intron: non-coding stretch of DNA placed between exons in the genomic DNA (synonymous to intervening sequence)

MAR: matrix attachment region where a DNA sequence is linked to the nuclear matrix and, hence, protected to DNase digestion

mRNA: messenger ribonucleic acid, carrying the coding sequence of a gene as well as specific signals guiding its expression (the *genon*)

NABP: nucleic acid binding protein

Nucleolus: nuclear body where the (highly amplified) ribosomal DNA is located and the ribosomal subunits synthesised

Oligomotif: oligonucleotide sequence, recognized by specific amino-acid motifs (*aa-motifs*) in nucleic acid binding proteins or by mi- or siRNAs in *RNA interference*

Operon: a sequence of *cistrons* linked in cis and transcribed into a single mRNA, representing a program of gene expression in prokaryots constituted by several, possibly co-operating genes

Peripheral (genetic) memory : genetic information temporally stored outside the genomic DNA in form of *(pre-)mRNA* and pre-proteins, allowing for delayed gene expression (e.g. maternal *mRNA* in oocytes, or proenzymes as trypsinogen)

Polytene chromosomes: giant chromosomes formed by chromatids staying synapsed together in multiple rounds of replication, where the visible band/interband pattern corresponds to units of meiotic recombination and transcription and, thus, to *genomic domains*

Post-transcriptional regulation: regulative interventions after transcription at the level of pre-mRNA and mRNA, according to the corresponding (pre-)genons; to be distinguished from *translational regulation*

Pre-genon: precursor of *genon* at *pre-mRNA* or *full domain transcript* level; the program in a transcript controlling the formation of *mRNA* and its expression

Pre-mRNA: primary transcript that is converted into *mRNA* by *processing*, including *splicing*

Pre-rRNA: primary transcript that is converted into ribosomal RNA by *processing*

Processing of RNA: mechanism of cleavage of transcripts (*pre-mRNA, pre-rRNA, FDT, etc.*) and excision and ligation of the fragments of genes which are conserved and functionally expressed (*exons*), whereas the intergenic and intervening sequences (*introns*) are destroyed

Protein gene (p-gene): polypeptide and its coding sequence, equivalent of triplet-based coding sequence in *mRNA*

Proto-genon: signals at DNA level that control - via *(pre-)mRNA* - expression of one or several genes;

includes the *pre-genon* as well as signals for chromatin modification and local activation of transcription

RNA interference: mechanism of transient or final repression of specific *(pre-)mRNAs* through specific interfering RNAs (siRNA or miRNA)

RNA-gene (r-gene): gene coding for a functional RNA

RNP: ribonucleoprotein complex, i.e, complex of RNA and proteins (selective binding of proteins to mRNA is essential for regulation of the gene expression process)

rRNA: ribosomal RNA backbone, aligning the ribosomal proteins to form the 30S (18S rRNA) and 50S (28S rRNA) ribosomal subunits; has, furthermore, ribozyme function

Splicing: particular type of RNA *processing* by internal excision of the non-coding introns from the transcripts, creating mRNAs by assembly of *exons* that contain pieces of coding sequences, possibly in several specific combinations *(alternative splicing)*

Structural gene (s-gene): gene contributing to cellular structure, either directly or via enzymatic activities

Transgenon: ensemble of trans-acting factors selected by a specific *genon* in an mRNA, acting on the signals placed in cis

Translational regulation; regulation at the level of the polyribosomes during translation of mRNA

Unified Matrix Hypothesis (UMH): postulates that a large part of the non-coding DNA has an architectural function, providing a frame for the selective interaction of specific regions in the DNA, within or between chromosomes, as seen in *ectopic pairing*

UTR: untranslated region preceding or following the coding sequence in mRNA

Mathematical Terms

Conditional probability: probability of an event or a message contingent upon the occurrence of another event or message

Ensemble entropy: uncertainty about a specific element to be chosen from an ensemble of elements with known probabilities

Entropy: uncertainty about the content of a message prior to its reception, on the basis of known probabilities for the various possible messages (see formula in text); expected information to be gained from receiving a message

Sequence entropy: uncertainty about a specific sequence composed from symbols with known probabilities and correlations

Acknowledgements

We thank our colleagues who contributed by discussion over the years to evolution of the ideas presented here, and in particular Manfred Eigen and the participants of the Klosters Winter Seminar (1997-2007). The first author thanks the Max Planck Institute for Mathematics in the Sciences (Leipzig) for its hospitality and best working conditions. The excellent secretarial help of Antje Vandenberg is gratefully acknowledged. This work was supported by the French CNRS, the Universities Paris 6 and 7, and by bioMérieux SA.

(9) References

Albiez, H., Cremer, M., Tiberi, C., Vecchio, L., Schermelleh, L., Dittrich, S., Kupper, K., Joffe, B., Thormeyer, T., von Hase, J. et al. (2006). Rise, fall and resurrection of chromosome territories: a historical perspective. Part II. Fall and resurrection of chromosome territories during the 1950s to 1980s. Part III. Chromosome territories and the functional nuclear architecture: experiments and models from the 1990s to the present. *Eur J Histochem* **50**, 223-72. Review.

Ananiev, E., Barsky, V., Ilyin, Y. and Churikov, N. (1981). Localization of nucleoli in Drosophila melanogaster polytene chromosomes. *Chromosoma* **81**, 619-28.

Anguita, E., Johnson, C. A., Wood, W. G., Turner, B. M. and Higgs, D. R. (2001). Identification of a conserved erythroid specific domain of histone acetylation across the alpha-globin gene cluster. *Proc. Natl. Acad. Sci. USA* **98**, 12114-12119.

Arcangeletti, C., De Conto, F., Sütterlin, R., Pinardi, F., Missorini, S., Géraud, G., Aebi, U., Chezzi, C. and Scherrer, K. (2000). Specific Types of Prosomes Distribute Differentially between Intermediate and Actin Filaments in Epithelial, Fibroblastic and Muscle Cells. *Europ. J. Cell Biol.* **79**, 423-437.

Arcangeletti, C., Sütterlin, R., Aebi, U., De Conto, F., Missorini, S., Chezzi, C. and Scherrer, K. (1997). Visualization of prosomes (MCP-proteasomes), intermediate filament and actin networks by "instantaneous fixation" preserving the cytoskeleton. *J. Struct. Biol.* **119**, 35-58.

Arrigo, A. P., Tanaka, K., Goldberg, A. L. and Welch, W. J. (1988). Identity of the 19S "prosome" particle with the large multifunctional protease complex of mammalian cells (the proteasome). *Nature* **331**, 192-194.

Auboeuf, D., Batsche, E., Dutertre, M., Muchardt, C. and O'Malley, B. (2007). Coregulators: transducing signal from transcription to alternative splicing. *Trends Endocrinol Metab* **18**, 122-9. Epub 2007 Feb 21.

Auboeuf, D., Dowhan, D., Dutertre, M., Martin, N., Berget, S. and O'Malley, B. (2005). A subset of nuclear receptor coregulators act as coupling proteins during synthesis and maturation of RNA transcripts. *Mol Cell Biol* **25**, 5307-16.

Auweter, S., Oberstrass, F. and Allain, F. (2006). Sequence-specific binding of single-stranded RNA: is there a code for recognition. *Nucleic Acids Res. 1994 Apr 11;22(7):1215-20* **34**, 4943-59. Epub 2006 Sep 18.

Baer, B. and Kornberg, R. (1983). The protein responsible for the repeating structure of cytoplasmic poly(A)-ribonucleoprotein. *J Cell Biol* **96**, 717-21.

Banerjee, A. (1980). 5'-terminal cap structure in eucaryotic messenger ribonucleic acids. *Microbiol Rev* **44**, 175-205. Review.

Barbosa, E. and Moss, B. (1978). mRNA(nucleoside-2'-)-methyltransferase from vaccinia virus. Characteristics and substrate specificity. *J Biol Chem* **253**, 7698-702.

Barr, H. and Ellison, J. (1976). Ectopic pairing of chromosome regions containing chemically similar DNA. *Chromosoma* **55**, 349-57.

Baugh, J., EV. (2004). 20S Proteasome Differentially Alters Translation of Different mRNAs via the Cleavage of eIF4F and eIF3. *Molecular Cell* **16**, 575-586.

Bennett, M. (1982). Nucleotypic basis of the spatial ordering of chromosomes in eukaryotes and the implications of the order for genome evolution and phenotypic variations. In *In Genome Evolution*, (ed. G. Dover, and Flavell, RB), pp. pp. 239-261. London: Academic Press.

Benzer. (1959). On the Topology of the Genetic Fine Structure. *Proc. Natl. Acad. Sci. USA* **45**, 1607-1620.

Benzer, S. (1961). On the Topography of the Genetic Fine Structure. *Proc. Natl. Acad. Sci. USA* **47**, 403-426.

Benzer, S. and Champe, S. (1961). Ambivalent rII Mutants of Phage T4. *Proc. Natl. Acad. Sci. USA* **47**, 1025-1038.

Berget, S., Moore, C. and Sharp, P. (1977). Spliced segments at the 5'terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci. USA* **74**, 3171-3175.

Bernardi, G. (2005). Lessons from a small, dispensable genome: the mitochondrial genome of yeast. *Gene* **354**, 189-200.

Birnstiel, M., BH, S. and IF, P. (1972). Kinetic complexity of RNA molecules. *J Mol Biol.* **63**, 21-39.

Blencowe, B. (2006). Alternative splicing: new insights from global analyses. *Cell* **126**, 37-47. Review

Blobel, G. (1985). Gene Gating: A Hypothesis. *PNAS* **82**, 8527-8529.

Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Muller, S., Eils, R., Cremer, C., Speicher, M. et al. (2005). Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol.* **3**, e157. Epub 2005 Apr 26.

Britten, R. and Kohne, D. (1968). Repeated sequences in DNA. *Science* **161**, 529-54.

Broders, F. and Scherrer, K. (1987). Transcription of the α-globin gene domain in normal and AEV-transformed chicken erythroblasts: mapping of giant globin-specific RNA including embryonic and adult gene. *Mol. Gen. Genet.* **209**, 210-220.

Broders, F., Zahraoui, A. and Scherrer, K. (1990). The chicken α-globin gene domain is transcribed into a 17-kilobase polycistronic RNA. *Proc. Natl. Acad. Sci. USA* **87**, 503-507.

Cantor, A. and Orkin, S. (2002). Transcriptional regulation of erythropoiesis: an affair involving multiple partners. *Oncogene* **21**, 3368-76. Review.

Capanna, E., Gropp, A., Winking, H., Noack, G. and Civitelli, M.-V. (1976). Robertsonian metacentrics in the mouse. *Chromosoma* **58**, 341-353.

Cavalier-Smith, T. (1978). Nuclear volume control by nucleoskeletal DNA, selection for cell volume
and cell growth rate, and the solution of the DNA c-value paradox. *Cell Sci* **34**, 247-278.

Chandley, A., Jones, R., Dott, H., Allen, W. and Short, R. (1974). Meiosis in interspecific equine hybrids. I. The male mule (Equus asinus X E. caballus) and hinny (E. caballus X E. asinus). *Cytogenet Cell Genet* **13**, 330-41.

Chen, K. and Rajewsky, N. (2007). The evolution of gene regulation by transcription factors and microRNAs - Review. *Nat Rev Genet* **8**, 93-103.

Chezzi, C., Grosclaude, J. and Scherrer, K. (1971). *Influence of the temperature on the repartition of ribosomal material from Hela cells*. In *1° Congresso Nazionale di Virologia*, (ed. A. e. Sanna), pp. 1-8. Parma - Salsomaggiore.

Choi, Y. D., Grabowski, P. J., Sharp, P. A. and Dreyfuss, G. (1986). Heterogeneous nuclear ribonucleoproteins: role in RNA splicing. *Science* **231**, 1534-1539.

Chow, L., Gelinas, R., Brocker, T. and Roberts, R. (1977). An Amazing Sequence Arrangeement at the 5' Ends of Adenovirus 2 Messenger RNA. *Cell* **12**, 1-8.

Christensen, A., Kahn, L. and Bourne, C. (1987). Circular polysomes predominate on the rough endoplasmic reticulum of somatotropes and mammotropes in the rat anterior pituitary. *American Journal of Anatomy* **178**, 1 - 10.

Ciechanover, A. (2006). Intracellular protein degradation: from a vague idea thru the lysosome and the ubiquitin-proteasome system and onto human diseases and drug targeting
. *Exp Biol Med (Maywood)* **231**, 1197-211. Review.

Civelli, O., Vincent, A., Maundrell, K., Buri, J. F. and Scherrer, K. (1980). The Translational Repression of Globin mRNA in Free Cytoplasmic Ribonucleoprotein Complexes. *Eur J Biochem.* **107**, 577-585.

Cohen Jr, M. (1976). Ectopic pairing and evolution of 5S ribosomal RNA genes in the chromosomes of Drosophila funebris. *Chromosoma* **55**, 349-57.

Colaiacovo, M. (2006). The many facets of SC function during C. elegans meiosis. *Chromosoma* **115**, 195-211. Epub 2006 Mar 23. Review.

Cole, C. and Scarcelli, J. (2006). Transport of messenger RNA from the nucleus to the cytoplasm. *Curr Opin Cell Biol* **18**, 299-306. Epub 2006 May 8. Review.

Collins, G. and Tansey, W. (2006). The proteasome: a utility tool for transcription? *Current Opinion in Genetics & Development* **16**, 197-202.

Commoner, B. (1964). Roles of desoxyribonucleic acid in inheritance. *Nature* **202**, 960-968.

Coux, O., Tanaka, K. and Goldberg, A. (1996). Structure and functions of the 20S and 26S proteasomes. *Annual Review of Biochememistry* **65**, 801-847.

Cremer, T. and Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet* **2**, 292-301. Review.

Cremer, T., Kreth, G., Koester, H., Fink, R., Heintzmann, R., Cremer, M., Solovei, I., Zink, D. and Cremer, C. (2000). Chromosome territories, interchromatin domain compartment, and nuclear matrix: an integrated viewof the functional nuclear architecture. - Review. *Crit Rev Eukaryot Gene Expr.* **10**, 179-212.

Cuperlovic-Culf, M., Belacel, N., Culf, A. and Ouellette, R. (2006). Microarray analysis of alternative splicing. *OMICS* **10**, 344-57. Review.

Daneholt, B. (2001). Assembly and transport of a premessenger RNP particle. *Proc Natl Acad Sci U S A* **98**, 7012-7.

Darlix, J., Khandjian, E. and Weil, R. (1984). Nature and origin of the RNA associated with simian virus 40 large tumor antigen. *PNAS* **81**, 5425-9.

De Conto, F., Pilotti, E., Razin, S. V., Ferraglia, F., Géraud, G., Arcangeletti, C. and Scherrer, K. (2000). In Mouse Myoblasts the Nuclear Prosomes are Associated with the Nuclear Matrix and Accumulate Preferentially in the Peri-Nucleolar Areas. *J. Cell Sci.* **113**, 2399-2407.

De Conto, F., Razin, S., Geraud, G., Arcangeletti, C. and Scherrer, K. (1999). In the Nucleus and Cytoplasm of Chicken Erythroleukemic Cells, Prosomes Containing the p23K Subunit Are Found in Centers of Globin (Pre-)mRNA Processing and Accumulation. *Exp. Cell Res.* **250**, 569-575.

de Laat, W. and Grosveld, F. (2003). Spatial organization of gene expression: the active chromatin hub. *Chromosome Res* **11**, 447-59. Review.

Deak, I., Sidebottom, E. and Harris, H. (1972). Further experiments on the role of nucleolus in the expression of structural genes. *J. Cell. Sci.* **11**, 379-390.

Dennis, P. and Omer, A. (2005). Small non-coding RNAs in Archaea. Review. *Curr Opin Microbiol* **8**, 685-94.

Dreyfuss, G. (1986). Structure and function of nuclear and cytoplasmic ribonucleoprotein particles. *Annual Review of Cell Biology* **2**, 459-498.

Dreyfuss, G., Kim, V. and Kataoka, N. (2002). Messenger-RNA-binding proteins and the messages they carry. Review. *Nat Rev Mol Cell Biol* **3**, 195-205.

Dubochet, J., Morel, C., Lebleu, B. and Herzberg, M. (1973). Structure of Globin mRNA and mRNA-Protein Particles: Use of Dark-Field Electron Microscopy. *Eur J Biochem.* **36**, 465-472.

Eberwine, J., Belt, B., Kacharmina, J. and Miyashiro, K. (2002). Analysis of subcellularly localized mRNAs using in situ hybridization, mRNA amplification, and expression profiling. *Neurochem Res* **27**, 1065-77. Review.

Edmonds, M. (2002). A history of poly A sequences: from formation to factors to function. *Prog Nucleic Acid Res Mol Biol* **71**, 285–389.

Edwalds-Gilbert, G., Veraldi, K. and Milcarek, C. (1997). Alternative poly(A) site selection in complex transcription units: means to an end. *Nucleic Acids Res* **25**, 2547-61. Review.

Egyhazi, E., Ossoinak, A., Filhol-Cochet, O., Cochet, C. and Pigon, A. (1999). The binding of the alpha subunit of protein kinase CK2 and RAP74 subunit of TFIIF to protein-coding genes in living cells is DRB sensitive. *Mol Cell Biochem* **191**, 149-59.

ENCODE Project Consortium. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 779-96| doi:10.1038/nature05874.

Fahrenkrog, B. and Aebi, U. (2003). The nuclear pore complex: nucleocytoplasmic transport and beyond. *Nat Rev Mol Cell Biol* **4**, 757-66. Review

.

Fantom Consortium and Riken Genome Groups. (2005). The transcritional landscape of tyhe mammalian genome. *Science* **309**, 1559-63.

Felsenfeld, G. (1992). Chromatin as an essential part of transcription mechanism. *Nature* **355**, 219-223.

Felsenfeld, G. and Groudine, M. (2003). Controlling the double helix. *Nature* **421**, 448-53. Review.

Filipowicz, W. and Pogacic, V. (2002). Biogenesis of small nucleolar ribonucleoproteins. Review. *Curr Opin Cell Biol* **14**, 319-27.

Flint, J., Tufarelli, C., Peden, J., Clark, K., Daniels, R. J., Hardison, R., Miller, W., Philipsen, S., Tan-Un, K. C., McMorrow, T. et al. (2001). Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the α globin cluster. *Human Mol. Genet.* **10**, 371-382.

Foucrier, J., Grand, M.-C., De Conto, F., Bassaglia, Y., Géraud, G., Scherrer, K. and Martelly, I. (1999). Dynamic distribution and formation of a para-sarcomeric banding pattern of prosomes during myogenic differentiation of satellite cells *in vitro*. *J. Cell Science* **112**, 989-1001.

Foucrier, J., Y., B., Grand, M.-C., Rothen, B., Perriard, J.-C. and Scherrer, K. (2001). Prosomes form sarcomere-like banding patterns in skeletal, cardiac, and smooth muscle cells. *Exp. Cell Res.* **266**, 193-200.

Fried, M. and Crothers, D. M. (1981). Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nulceic Acids Research* **9**, 6505-6524.

Fulton, A. B. and Alftine, C. (1997). Organization of protein and mRNA for titin and other myofibril components during myofibrillogenesis in cultured chicken skeletal muscle. *Cell Struct. Funct.* **22**, 51-58.

Furuichi, Y. and Shatkin, A. (2000). Viral and cellular mRNA capping: past and prospects. *Adv Virus Res* **55**, 135–184.

Gasser, S. (2002). Visualizing chromatin dynamics in interphase nuclei. *Science* **296**, 1412-6. Review.

Gaszner, M. and Felsenfeld, G. (2006). Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet* **7**, 703-13. Epub 2006 Aug 15. Review.

Georgiev, G., Samarina, O., Lerman, M., Smirnov, M. and Svertzov, A. (1963). Biosynthesis of Messenger and Ribosomal Ribonucleic Acids in the Nucleochromosamal Apparatus of Animal Cells. *Nature* **200**, 1291-1294.

Gerstein, M., Bruce, C., Rozowsky, J., Zheng, D., Du, J., Korbel, J., Emanuelsson, O., Zhang, Z., Weissman, S. and Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. *Genome Res* **17**, 669-681: doi:10.1101/gr.6339607.

Gingeras, T. (2007). Origin of phenotypes: Genes and transcripts. *Genome Res* **17**, 682-690; doi:10.1101/gr.6525007.

Glover, D., Zaha, A., Stocker, A., Santelli, R., Pueyo, M., De Toledo, S. and Lara, F. (1982). Gene Amplification in Rhynchosciara Salivary Gland Chromosomes. *PNAS* **79**, 2947-2951.

Goldenberg, S., Vincent, A. and Scherrer, K. (1979). Evidence for the protection of specific RNA sequences in globin messenger ribonucleoprotein particles. *Nuc. Acids Res.* **6**, 2787-2797.

Gorgoni, B. and Gray, N. (2004). The roles of cytoplasmic poly(A)-binding proteins in regulating gene expression: a developmental perspective. *Brief Funct Genomic Proteomic* **3**, 125-41. Review.

Grewal, S. and Jia, S. (2007). Heterochromatin revisited. *Nat Rev Genet* **8**, 35-46. Review.

Griffiths, P. and Stotz, K. (2006). Genes in the postgenomic era. *Theor Med Bioeth* **27**, 499-521.

Grossbach, U. (1974). Chromosome puffs and gene expression in polytene cells. In *Cold Spring Harb Symp Quant Biol*, vol. 38, pp. 619-27. Cold Spring Harbour: Cold Spring Harb Lab.

Grossi de Sa, M.-F., Standart, N., Martins de Sa, C., Akhayat, O., Huesca, M. and Scherrer, K. (1988). The poly(A)-binding protein facilitates in vitro translation of poly(A)-rich mRNA. *Eur J Biochem.* **176**, 521-526.

Groudine, M. and Weintraub, H. (1981). Activation of globin genes during chicken development. *Cell* **24**, 393-401.

Grummt, I. (2006). Actin and myosin as transcription factors. *Curr Opin Genet Dev* **16**, 191-6.

Haeusler, R. and Engelke, D. (2006). Spatial organization of transcription by RNA polymerase III. *Nucleic Acids Res. 2006;34(17):4826-36. Epub 2006 Sep 13. Review* **34**, 4826-36. Epub 2006 Sep 13. Review.

Handwerger, K. and Gall, J. (2006). Subnuclear organelles: new insights into form and function. *Trends Cell Biol* **16**, 19-26. Epub 2005 Dec 1. Review.

Hernandez-Verdun, D. (2006). The nucleolus: a model for the organization of nuclear functions. *Histochem Cell Biol* **126**, 135-48. Epub 2006 Jul 12. Review.

Hershey, A. and Chase, M. (1955). An upper limit to the protein content of the germinal substance of bacteriophage T2. *Virology* **1**, 108–127.

Holtzer, H., Rubinstein, N., Fellini, S., Yeoh, G., Chi, J., Birnbaum, J. and Okayama, M. (1975). Lineages, quantal cell cycles, and the generation of cell diversity. *Q Rev Biophys* **8**, 523-57. Review.

Holtzer, H., Weintraub, H., Mayne, R. and Mochan, B. (1972). The cell cycle, cell lineages, and cell differentiation. *Curr Top Dev Biol* **7**, 229-56. Review.

Horn, P. and Peterson, C. (2006). Heterochromatin assembly: a new twist on an old model. *Chromosome* **14**, 83-94.

Hough, B., Smith, M., Britten, R. and Davidson, E. (1975). Sequence complexity of heterogeneous nuclear RNA in sea urchin embryos. *Cell* **5**, 291-9.

Hube, F., Guo, J., Chooniedass-Kothari, S., Cooper, C., Hamedani, M., Dibrov, A., Blanchard, A. A. A., Wang, X., Deng, G., Myal, Y. et al. (2006). Alternative Splicing of the First Intron of the Steroid Receptor RNA Activator (SRA) Participates in the Generation of Coding and Noncoding RNA Isoforms in Breast Cancer Cell Line. *DNA Cell Biol* **25**, 418–428

Hughes, T. (2006). Regulation of gene expression by alternative untranslated regions. *Trends Genet* **22**, 119-22. Epub 2006 Jan 23. Review.

Iarovaia, O., Razin, S. V., Linares-Cruz, G., Sjakste, N. and Scherrer, K. (2001). In chicken leuklemia cells globin genes are fully transcribed but their RNAs are retained in the perinucleolar area. *Exp. Cell. Res.* **270**, 159-165.

Imaizumi-Scherrer, M.-T., Maundrell, K., Civelli, O. and Scherrer, K. (1982). Transcriptional and post-transcriptional regulation in duck erythroblasts. *Dev. Biol.* **93**, 126-138.

Ioudinkova, E., Razin, S., Borunova, V., De Conto, F., Rynditch, A. and Scherrer, K. (2005). RNA-dependent nuclear matrix contains a 33 kb globin full domain transcript as well as prosomes but no 26S proteasomes. *J. Cell. Biochem.* **94**, 445-457.

Jackson, R. and Standart, N. (2007). How do microRNAs regulate gene expression. *Sci STKE* **2007(367)**, re1. Review.

Jacob, F. and Monod, J. (1961). Genetic Regulatory Mechanisms in the Synthesis of Proteins. *J. Mol. Biol.* **3**, 318-356.

Johannsen, W. (1909). Elemente der exakten Erblichkeitslehre. Jena; Quoted by Nils Roll-Hansen (1989).

Judd, B., Shen, M. and Kaufman, T. (1972). The anatomy and function of a segment of the X chromosome of Drosophila melanogaster. *Genetics* **71**, 139-56.

Kaeser, M. and Emerson, B. (2006). Remodeling plans for cellular specialization: unique styles for every room. *Curr Opin Genet Dev* **16**, 508-12. Epub 2006 Aug 14.

Kaufman, B., Mcdonald, M., Gay, H., Wilson, K., Wyman, R. and Okuda, N. (1948). Organisation of the Chromosome. *Canergie Inst. Year Book* **47**, 144-155.

Kelly, R., Alonso, S., Tajbakhsh, S., Cossu, G. and Buckingham, M. (1995). Myosin light chain 3F regulatory sequences confer regionalized cardiac and skeletal muscle expression in transgenic mice.
. *J Cell Biol* **129**, 383-96.

Kepes, F. and Vaillant, C. (2003). Transcription-Based Solenoidal Model of Chromosomes
. *Complexus* **1**, 171– 180 (DOI:10.1159/000082184).

Khandjian, E., Matter, J., Leonard, N. and Weil, R. (1980). Simian virus 40 and polyoma virus stimulate overall cellular RNA and protein synthesis. *Proc Natl Acad Sci U S A* **77**, 1476-80.

Kim, V. and Dreyfuss, G. (2001). Nuclear mRNA binding proteins couple pre-mRNA splicing and post-splicing events. *Mol Biol Cell* **12**, 1-10. Review.

Kindler, S., Wang, H., Richter, D. and Tiedge, H. (2005). RNA transport and local control of translation - Review. *Annu Rev Cell Dev Biol* **21**, 223-45.

Kioussis, D. (2005). Gene regulation: kissing chromosomes. *Nature* **435**, 579-80.

Kiss, T. (2006). SnoRNP biogenesis meets Pre-mRNA splicing. *Mol Cell* **23**, 775-6. Review.

Kleckner, N. (2006). Chiasma formation: chromatin/axis interplay and the role(s) of the synaptonemal complex. *Chromosoma* **115**, 175-94. Epub 2006 Mar 23. Review.

Koslowsky, D. (2004). A historical perspective on RNA editing: how the peculiar and bizarre became mainstream. *Methods Mol Biol* **265**, 161-97. Review.

Kouzarides, T. (2007). Chromatin modifications and their function. *Cell* **128**, 693-705. Review.

Kutay, U. and Guttinger, S. (2005). Leucine-rich nuclear-export signals: born to be weak. Review. *Trends Cell Biol* **15**, 121-4.

Képès, F. (2003). Periodic epi-organization of the yeast genome revealed by the distribution of promoter sites. *J Mol Biol* **329**, 859-65.

Lanctot, C., Cheutin T, Cremer M, Cavalli G, Cremer T. (2007). Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nat Rev Genet* **8**.

Lander, E., Linton, L., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. et al. (2001.). Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.

Lara, F. (1987). Gene amplification in Rhynchosciara (1955-1987). *Mem Inst Oswaldo Cruz.* **82 Suppl 3**, 125-8.

Latchman, D. S. (1990). Eukaryotic transcription factors. *Biochemical Journal* **270**, 281-289.

Lawrence, J. and Singer, R. (1991). Spatial organization of nucleic acid sequences within cells. *Semin Cell Biol* **2**, 83-101.

Lawrence, J., Singer, R. and Marselle, L. (1989). Highly localized tracks of specific transcripts within interphase nuclei visualized by in situ hybridization. *Cell* **57**, 493-502.

Lennon, G. and Perry, R. (1990). The temporal order of appearance of transcripts from unrearranged and rearranged Ig genes in murine fetal liver. *J Immunol* **144**, 1983-7.

Lima de Faria, A. (1979). Prediction of gene location and classification of genes accoding to the chromosome field. In *"Specific eukaryotic genes", Alfred Benzon Symposium 13*: Munksgaard, 1979.

Lima de Faria, A. (1983). Experimental demonstration of interactions between chromosomes. Interchromosomal effects. In *Molecular evolution and organisation of the chromosomes*, pp. p. 641. Amsterdam: Elsevier.

Lima-de-Faria, A. (1980). Classification of genes, rearrangements and chromosomes according to the chromosome field. *Hereditas* **93**, 1-46.

Maco, B., Fahrenkrog, B., Huang, N. and Aebi, U. (2006). Nuclear pore complex structure and plasticity revealed by electron and atomic force microscopy. *Methods Mol Biol* **322**, 273-88. Review.

Mangus, D., Evans and MCJacobson, A. (2003). Poly(A)-binding proteins: multifunctional scaffolds for the post-transcriptional control of gene expression. *Genome Biol* **4**, 223. Epub 2003 Jul 1. Review.

Maniotis, A., Bojanowski, K. and Ingber, D. (1997). Mechanical continuity and reversible chromosome disassembly within intact genomes removed from living cells. *J Cell Biochem* **65**, 114-30.

Martin, K. J. (1991). The interactions of transcription factors and their adaptors, coactivators and accessory proteins. *BioEssays* **13**, 499-503.

Martins de Sa, C., Grossi de Sa, M. F., Akhayat, O., Broders, F., Scherrer, K., Horsch, A. and schmid, H. P. (1986). Prosomes: ubiquity and inter species structural variation. *J. Mol. Biol.* **187**, 479-493.

Matzke, M. and Birchler, J. (2005). RNAi-mediated pathways in the nucleus. Review. *Nat Rev Genet* **6**, 24-35.

Maundrell, K., Maxwell, E. S., Civelli, O., Vincent, A., Goldberg, S., Buri, J.-F., Imaizumi-Scherrer, M.-T. and Scherrer, K. (1979). Messenger RNP complexes in avian erythroblasts: Carriers of post-transciptional regulation? *Mol. Biol. Rep.* **5**, 43-51.

Maundrell, K., Maxwell, S., Puvion, E. and Scherrer, K. (1981). The nuclear matrix of duck erythroblasts is associated with globin mRNA coding sequences but not with the major proteins of 40S nuclear RNP. *Exptl. Cell. Res.* **136**, 435-445.

Maundrell, K. and Scherrer, K. (1979). Characterization of Pre-messenger-RNA Containing Nuclear Ribonucleoprotein Particles from Avian Erythroblasts. *Eur J Biochem.* **99**, 225-238.

Maxwell, E. and Fournie, M. (1995). The small nucleolar RNAs. *Annu Rev Biochem* **64**, 897-934. Review.

Mayer, C. and Grummt, I. (2006). Ribosome biogenesis and cell growth: mTOR coordinates transcription by all three classes of nuclear RNA polymerases. *Oncogene.* **25**, 6384-91.

Mendel, J. (1866). Versuche über Pflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brünn* **4 Abhandlungen**, 3–47. Cited by Robert C. Olby (1997) on http://www.mendelweb.org/MWolby.html, accessed 2007-03-16.

Missler, M. and Sudhof, T. (1998). Neurexins: three genes and 1001 products. *Trends Genet* **14**, 20-6. Review.

Morgan, T., Sturtevant, A., Muller, H. and Bridges, C. (1915). The mechanism of Mendelian heredity. New York: Holt Rinehart & Winston.

Munroe, D. and Jacobson, A. (1990). Tales of poly(A): a review. **91**, 151-8.

NCBI Map Viewer. (2006). Drosophila melanogaster (fruit fly) genome view: www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=7227.

Nickerson, J. (2001). Experimental observations of a nuclear matrix. *J Cell Sci.* **114**, 463-74.

Nickerson, J., Blencowe, B. and Penman, S. (1995). The architectural organization of nuclear metabolism. *Int Rev Cytol* **162A**, 67-123.

Nordheim, A., Pardue, M., Weiner, L., Lowenhaupt, K., Scholten, P., Moller, A., Rich, A. and Stollar, B. (1986). Analysis of Z-DNA in fixed polytene chromosomes with monoclonal antibodies that show base sequence-dependent selectivity in reactions with supercoiled plasmids and polynucleotides. *J Biol Chem* **261**, 468-76.

Odartschenko, N. and Keneklis, T. (1973). Localization of paternal DNA in interphase nuclei of mouse eggs during early cleavage. *Nature* **241**, 528-529.

Ohno, S. (1972). So much “junk” DNA in the genome. In *In Evolution of genetic systems; Brookhaven Symposia in Biology*, vol. 23 (ed. e. H. H. Smith), pp. pp. 366–370. Brookhaven: Gordon & Breach, New York.

Olink-Coux, M., Huesca, M. and Scherrer, K. (1992). Specific types of prosomes are associated to subnetworks of the Intermediate Filaments in PtK1 cells. *Europ J Biochem* **59**, 148-159.

Osheim, Y., Miller, O. and Beyer, A. (1985). RNP particles at splice junction sequences on Drosophila chorion transcripts. *Cell* **43**, 143-151.

Pal, J. K., Gounon, P., Grossi de Sa, M.-F. and Scherrer, K. (1988). Presence and Distribution of Specific Prosome Antigens Change as a Function of Embryonic Development and Tissue-Type Differentiation in Pleurodeles Walti. *J. Cell Sci.* **90**, 555-567.

Pearson, P. (2006). Historical development of analysing large-scale changes in the human genome. *Cytogenet Genome Res* **115**, 198-204. Review.

Penman, S., Fulton, A., Capco, D., Ben Ze'ev, A., Wittelsberger, S. and Tse, C. (1982). Cytoplasmic and nuclear architecture in cells and tissue: form, functions, and mode of assembly. In *Cold Spring Harb Symp Quant Biol.*, vol. 46; Pt 2, pp. 1013-28. Cold Spring Harbour: Cold Spring Harbour Lab. Press.

Pennisi, E. (2003). A Low Number Wins the GeneSweep Pool. *Science* **300**, 1484.

Perry, R. and Kelley, D. (1975). Methylated constituents of heterogeneous nuclear RNA: presence in blocked 5' terminal structures. *Cell* **6**, 13-9.

Perry, R. and Scherrer, K. (1975). The methylated constituents of globin mRNA. *FEBS Lett* **57**, 73-8.

Razin, S., Borunova, V., Rynditch, A., Ioudinkova, E., Smalko, V. and Scherrer, K. (2004). The 33kb Transcript of the Chicken alpha-Globin Gene Domain is Part of the Nuclear Matrix. *J. Cell. Biochem.* **92**, 445-457.

Razin, S., Rzeszowska-Wolny, J., Moreau, J. and Scherrer, K. (1985). Localization of sites of DNA attachment to the nuclear matrix in the domain of the chicken alpha-globin genes in functionally active and inactive nuclei. *Mol. Biol.* **19**, 376-385.

Rees, H., Jenkins, G., Seal, A. and Hutchinson, J. (1982). Assays of the phenotypic effects of changes in DNA amounts. In *In Genome Evolution*, (ed. G. Dover, and Flavell, RB): Academic Press.

Rich, A. (1961). The Transfer of Information Between the Nucleic Acids. In *Molecular and Cellular Synthesis*, vol. 19th Growth Symposium (ed. S. f. t. S. o. D. a. Growth), pp. cf. page 10: The Ronald Press Company.

Rodriguez, M., Dargemont, C. and Stutz, F. (2004). Nuclear export of RNA. *Biol Cell* **96**, 639-55.

Roll-Hansen, N. (1989). The crucial experiment of Wilhelm Johannsen. *Biol Philos* **4**, 303–329.

Rosbash, M., Ford, P. and Bishop, J. (1974). Analysis of the C-value paradox by molecular hybridization. *Proc Natl Acad Sci U S A* **71**, 3746-50.

Rout, M. and Blobel, G. (1993). Isolation of the yeast nuclear pore complex. *J Cell Biol* **123**, 771-83.

Santelli, R., Siviero, F., Machado-Santelli, G., Lara, F. and Stocker, A. (2004). Molecular characterization of the B-2 DNA puff gene of Rhynchosciara americana. *Chromosoma. 1976 Jun 23;55(4):349-57.* **113**, 167-76. Epub 2004 Jul 29.

Scheer, U. and Benavente, R. (1990). Functional and dynamic aspects of the mammalian nucleolus. *Bioessays* **12**, 14-21.

Scheer, U. and Hock, R. (1999). Structure and function of the nucleolus. *Curr Opin Cell Biol.* **11**, 385-90.

Scherrer, K. (1967). Pattern of Messenger RNA in Animal Cells and the Concept of "Cascade Regulation". In *Int. Symp. Biochemistry of Ribosomes and Messenger-RNA (1967)*, vol. 1968 (1) (ed. P. L. R. Lindigkeit, J. Richter), pp. 259-277. Castle Reinhardsbrunn: Akademie Verlag (Berlin).

Scherrer, K. (1980). Cascade regulation: a model of integrative control of gene expression in eukaryotic cells and organisms. In *Eukaryotic Gene Regulation*, vol. 1 (ed. Kolodny), pp. 57-129. Boca Raton, Florida: CRC press Inc.

Scherrer, K. (1989). A unified matrix hypothesis of DNA-directed morphogenesis, protodynamism and growth control. *Bioscience Reports* **9**, 157-188.

Scherrer, K. (2003). Historical Review: The discovery of "giant" RNA and of RNA processing: 40 years of enigma. *TIBS* **28**, 566-571.

Scherrer, K. and Bey, F. (1994). The Prosomes (Multicatalytic Proteinase - Proteasomes) and their relation to the untranslated messenger ribonucleoproteins, the cytoskeleton and cell differentiation. *Progr. Nucl. Acids Res. Mol. Biol.* **49**, 1-64.

Scherrer, K. and Darnell, J. E. (1962). Sedimentation Characteristics of Rapidly Labelled RNA from HeLa Cells. *Biochem. Biophys. Res. Comm.* **7**, 486-490.

Scherrer, K. and Jost, J. (2007). The gene and the genon concept: A functional and information-theoretic analysis. In *Mol Syst Biol*, vol. 3:87, pp. Epub 2007 Mar 13: EMBO and Nature Publishing Group.

Scherrer, K., Latham, H. and Darnell, J. E. (1963). Demonstration of an Unstable RNA and of a Precursor to Ribosomal RNA in HeLa Cells. *Proc. Natl. Acad. Sci. U.S.* **49**, 240-248.

Scherrer, K. and Marcaud, L. (1968). Messeger RNA in avian erythroblasts at the transcriptional and translational levels and the problem of regulation in animal cells. *J. Cell. Physiol.* **72**, 181-212.

Scherrer, K., Marcaud, L., Zajdela, F., London, I. M. and Gros, F. (1966). Patterns of RNA metabolism in a differentiated cell: a rapidly labeled, unstable 60S RNA with messenger properties in duck erythroblasts. *Proc. Natl. Acad. Sci. USA* **56**, 1571-1578.

Schmid, H. P., Akhayat, O., Martins de SA, C., Puvion, F., Koehler, K. and Scherrer, K. (1984). The Prosome: an ubiquitous morphologically distinct RNP particle associated with repressed mRNPs and containing specific ScRNA and a characteristic set of proteins. *EMBO Journal* **3**, 29-34.

Shapiro, T. A. and Englund, P. T. (1995). The Structure and Replication of Kinetoplast DNA. *Annual Review of Microbiology* **49**, 117-143.

Shatkin, A. and Manley, J. (2000). The ends of the affair: capping and polyadenylation. *Nat Struct Biol.* **7**, 838-42.

Shimizu, R. and Yamamoto, M. (2005). Gene expression regulation and domain function of hematopoietic GATA factors. *Semin Cell Dev Biol.* **16**, 129-36. Epub 2004 Dec 10. Review.

Sims, R., Mandal, S. and Reinberg, D. (2004). Recent highlights of RNA-polymerase-II-mediated transcription. *Curr Opin Cell Biol* **16**, 263-71. Review.

Snyder, M. and Gerstein, M. (2003). Genomics. Defining genes in the genomics era. *Science* **300**, 258-60.

Soller, M. (2006). Pre-messenger RNA processing and its regulation: a genomic perspective. *Cell Mol Life Sci* **63**, 796-819. Review.

Sontheimer, E. (2005). Assembly and function of RNA silencing complexes. Review. *Nat Rev Mol Cell Biol* **6**, 127-38.

Sontheimer, E. and Carthew, R. (2005). Silence from within: endogenous siRNAs and miRNAs. Review. *Cell* **122**, 9-12.

Spilianakis, C., Lalioti, M., Town, T., Lee, G. and Flavell, R. (2005). Interchromosomal associations between alternatively expressed loci. *Nature* **435**, 637-45.

Spohr, G., Granboulan, N., Morel, C. and Scherrer, K. (1970). Messenger RNA in HeLa Cells: An Investigation of Free and Polyribosome-bound Cytoplasmic Messenger Ribonucleoprotein Particles by Kinetic Labelling and Electron Microscopy. *Eur J Biochem.* **17**, 296-318.

Spohr, G., Imaizumi, T. and Scherrer, K. (1974). Synthesis and processing of nuclear precursor-messanger RNA in avian erythroblasts and Hela cells. *Proc. Natl. Acad. Sci. USA* **71**, 5009-5013.

Spohr, G., Mirault, M.-E., Imaizumi, M.-T. and Scherrer, K. (1976). Molecular-Weight Determination of Animal-Cell RNA by Electrophoresis in Formamide under Fully Denaturing Conditions on Exponential Polyacrylamide Gels. *Europ. J. Biochem.,* **62**, 313-322.

Spohr, G. and Scherrer, K. (1972). Differential Turnover of Two Messengers in One Cell Type: 9S Globin mRNA and 12S mRNA in Differentiating Avian Erythroblasts. *Cell Differentiation* **1**, 53-61.

Stadler, S., Schnapp, V., Mayer, R., Stein, S., Cremer, C., Bonifer, C., Cremer, T. and Dietzel, S. (2004). The architecture of chicken chromosome territories changes during differentiation. *BMC Cell Biol* **5**, 44.

Stalder, J., Larsen, A., Engel, J. D., Dolan, M., Groudine, M. and Weintraub, H. (1980). Tissue-specific DNA cleavage in the globin chromatin domain introduced by DNase I. *Cell* **20**, 451-460.

Steitz, T. and Moore, P. (2003). RNA, the first macromolecular catalyst: the ribosome is a ribozyme. Review. *Trends Biochem Sci* **28**, 411-8.

Sumner, A. (1982). The nature and mechanisms of chromosome banding
. *Cancer Genet Cytogenet* **6**, 59-87. Review.

Tang, G. (2005). siRNA and miRNA: an insight into RISCs. Review. *Trends Biochem Sci.* **30**, 106-14.

Therwath, A. and Scherrer, K. (1982). Precursors of distinct size for chicken alpha A, alpha D and beta globin mRNA. *FEBS Letters* **142**, 12-16.

Thiele, B., Belkner, J., Andree, H., Rapoport, T. and Rapoport, S. (1979). Synthesis of non-globin proteins in rabbit-erythroid cells. Synthesis of a lipoxygenase in reticulocytes. *Eur J Biochem* **96**, 563-9.

Thomson, A., Rogers, J. and Leedman, P. (1999). Iron-regulatory proteins, iron-responsive elements and ferritin mRNA translation. Review. *Int J Biochem Cell Biol* **31**, 1139-52.

Tiedemann, H., M, A., H, G. and W, K. (2001). Pluripotent cells (stem cells) and their determination and differentiation in early vertebrate embryogenesis. *Dev Growth Differ.* **43**, 469-502. Review.

Tonegawa, S. (1983). Somatic generation of antibody diversity. *Nature* **302**, 575-81. Review.

Travers, A. (1999). Chromatin modification by DNA tracking. *Proc. Natl. Acad. Sci. USA* **96**, 13634-37. Review.

Tschochner, H. and Hurt, E. (2003). Pre-ribosomes on the road from the nucleolus to the cytoplasm. *Trends Cell Biol* **13**, 255-63.

Tuan, D. Y. H., Solomon, W.B., London, I.M. and Lee, D.P. (1989). An erythroid-specific, developmental-stage-independent enhancer far upstream of the human "beta-like globin" genes. *Proceedings of the National Academy of Sciences USA* **86**, 2554-2558.

Twiss, J. and van Minnen, J. (2006). New insights into neuronal regeneration: the role of axonal protein synthesis in pathfinding and axonal extension. *J Neurotrauma***;23**, 295-308. Review.

Valadkhan, S. (2005). snRNAs as the catalysts of pre-mRNA splicing. Review. *Curr Opin Chem Biol* **9**, 603-8.

Venkatesan, S., Gershowitz, A. and Moss, B. (1980). Modification of the 5' end of mRNA. Association of RNA triphosphatase with the RNA guanylyltransferase-RNA (guanine-7-)methyltransferase complex from vaccinia virus. *J Biol Chem* **255**, 903-8.

Venter, C. and (et al). (2001). The Sequence of the Human Genome. *Science* **291**, 1304-1351.

Vincent, A., Akhayat, O., Goldenberg, S. and Scherrer, K. (1983). Differential repression of specific mRNA in erythroblast cytoplam: possible role for free mRNP proteins. *EMBO Journal* **2**, 1869-1876.

Vincent, A., Civelli, O., Buri, J. F. and Scherrer, K. (1977). Correlation of specific coding sequences with specific proteins associated in untranslated cytoplamic messenger ribonucleoprotein complexes of duck erythroblasts. *FEBS Letters* **77**, 281-286.

Vincent, A., Civelli, O., Maundrell, K. and Scherrer, K. (1980). Identification and Characterization of the Translationally Repressed Cytoplasmic Messenger-Ribonucleoprotein Particles from Duck Erythroblasts. *Eur J Biochem.* **112**, 617-633.

Vincent, A., Goldenberg, S., Standart, N., Civelli, O., Imaizumi-Scherrer, M.-T., Maundrell, K. and Scherrer, K. (1981). Potential role of mRNP proteins in cytoplasmic control of gene expression in duck erythroblasts. *Mol. Biol. Rep.* **7**, 71-81.

von Kries, J., Buck, F. and Stratling, W. (1994). Chicken MAR binding protein p120 is identical to human heterogeneous nuclear ribonucleoprotein (hnRNP) U. *Nucleic Acids Res* **22**, 1215-20.

Von Kries, J. P., Buhrmester, H. and Strätling, W. (1991). A Matrix/attachment region binding protein: identification, purification and mode of binding. *Cell* **64**, 123-135.

Warner, J. R. and al, e. (1963). A multiple ribosomal structure in protein synthesis. *Proc. Nat. Acad. Sci. U.S.A.* **49**, 122-129.

Warocquier, R. and Scherrer, K. (1969). *RNA Metabolism in Mammalian Cells at Elevated Temperature. Eur J Biochem.* **10**, 362-370.

Weintraub, H. and Groudine, M. (1976). Chromosomal subunits in active genes have an altered conformation. *Science* **193**, 848-56.

Wick, K. and Matthews, K. (1991). Interactions between lac repressor protein and site-specific bromodeoxyuridine-substituted operator DNA. Ultraviolet footprinting and protein-DNA cross-link formation. *J Biol Chem* **266**, 6106-12.

Will, C. and Luhrmann, R. (2005). Splicing of a rare class of introns by the U12-dependent spliceosome. Review. *Biol Chem* **386**, 713-24.

Wu, Z., Shi, Y., Tibbetts, R. and Miyamoto, S. (2006). Molecular linkage between the kinase ATM and NF-kappaB signaling in response to genotoxic stimuli. *Science* **311**, 1141-6.

Zhu, J. and McKeon, F. (2000). Nucleocytoplasmic shuttling and the control of NF-AT signaling. *Cell Mol Life Sci* **57**, 411-20. Review.

Figure legends

Figure 1 Definition of the gene: a functional polypeptide basis of a unit function. By genetic analysis, the gene is identified as a phenotypic function. An individual function is based on co-operating proteins or polypeptides; the latter represent, hence, the basic unit functions. At nucleic acid levels, the closest equivalent is the coding sequence for such a polypeptide, inserted into the mRNA. In the general case, such a coding sequence - gene equivalent - is fragmented in the DNA, which constitutes the genotype, basis of a specific phenotype.

Figure 2 The Jacob and Monod Model of the operon: In the bacterial operon, several coding sequences (cistrons) are coupled together to secure a metabolic pathway as, e.g., in case of the *lac* operon. When activated, such an operon is transcribed as a unit and, prior to termination of transcription, a polyribosome is formed on the mRNA, and the products, the enzymes Z and Y as well as an acetylase are made. DNA, mRNA and the translation machinery form, hence, a tightly linked physical complex; therefore (as in a timepiece), arrest at any level stops the entire machinery. In the repressed state, in the upstream operator/promoter sequence where the RNA polymerase attaches and transcription has to start, the repressor may attach on the basis of a sequence-specific protein-DNA interaction, prohibiting transcription. The repressor is the product of a distant gene coding for a polypeptide. Once attached to the DNA, the repressor may become the target of an inducer, in the case cited a small Mr chemical compound reducing the affinity constant of the DNA-repressor interaction. Regulation operates thus primarily at transcriptional level, controlling types and amounts of polypeptides formed; in this case it acts in a negative manner via the repressor, but positive regulation via peptides acting as inducers exists as well. Note that the operon arrangement implies already an expression program including the operator in the sense of the genon concept.

Figure 3 From Gene to Phen in space and time: Once the unit physical complex of the bacterial translation machinery got disrupted, when during evolution the genome-DNA was removed from the polyribosomes and stored away in the nucleus, by necessity a time delay results since, prior to gene expression, the transcripts have first to be transported in space. Thus, two inter-dependant vectors in space and time result which, ensemble, govern gene expression. Furthermore, transport of transcripts may be interrupted and considerable time delay may result (up to 30 years, e.g., in case of the human maternal histone mRNA laid down in the unfertilised egg); the corresponding mRNA forms repressed mRNP complexes to be activated upon specific signals, and constitute *peripheral memories* of genetic information. But transcripts may be stored during earlier stages of their processing from primary pre-mRNA to mRNA; these unspliced or partially spliced pre-mRNAs may still contain individual exons rather than finally constituted coding sequences or genes. The gene, which has to be reconstituted each time an mRNA is formed, springs up, thus, during RNA processing. It is subject to terminal controls which may bear on its nature (final splicing), its cellular site and time of expression. Nature, timing and site of gene expression are hence largely subject to *post-transcriptional regulation* (Scherrer, 1980).

Figure 4 Genon and Transgenon: (box 1) The equivalent of the polypeptide-gene at RNA level is the coding sequence which is inserted in the mRNA and framed by the 5-side and 3'-side UTRs. In the latter and superimposed onto the coding sequence is an ensemble of signals constituting the *Genon.* The genon represents a program in cis of sequence oligomotifs, eventual binding sites (oligomotifs may form hairpins - as shown - or not) for regulatory proteins (or si/miRNAs - not shown). (Box 2) When present, protein factors interact with the oligomotifs (empty coloured circles) in cis forming RNPs (insert B); the ensemble of the factors (filled circles) picked up by an mRNA constitutes its specific *transgenon*. (Box 3) The *Holo-Transgenon* of a given cell is constituted of by all these factors, which eventually will recognise an oligomotif in the cis-genon. (Grey box) A subset of factors (filled circles) interacting with a specific mRNA constitute the latter's transgenon. Insert (A): dark field EM picture of globin mRNA showing its compact non-random nature due to secondary structure. Insert (B): dark field EM picture of a globin mRNP constituted by globin mRNA and 3 times its mass of specific associated proteins (Civelli et al., 1980). Notice, that proteins are attached all along the mRNA chain interacting within the coding sequence. The latter contains, hence, two types of information relating (1) to the genetic code and (2), to sequence oligo-motifs recognising specific RNA-binding proteins (or interfering RNAs) acting as vehicles of post-transcriptional controls. (For experimental details see (Dubochet et al., 1973)).

Figure 5 From DNA to pre-mRNA and mRNA expression: Proto-, Pre- and Genon: The genomic domain (line A) with exons (light green) and fragments of coding sequences (dark green), as well as inter-genic (not shown) and intra-genic non coding DNA, contains instructions for remodelling and activation of chromatin; this constitutes the *proto-genon* (A'). From these a pre-mRNA (B) or a full domain transcript (FDT) with its *pre-genon* (B') may spring off. The latter may contain gene fragments subject to differential splicing; shown is the case of a pre-mRNA containing the two ORFs 1 and 2. Below are show the two mRNAs created with their respective genons and, thereafter, the two gene equivalents, the coding sequence in mRNAs (1) and (2) with their products, peptide 1 and 2 securing two functions. Insert: To the *genon* signals (oligomotifs) carrying distinct instructions for specific steps of processing and gene expression (left) correspond factors from the *transgenon* (right), in active or inactive states, which may - or not (when inactive or absent) - implement the corresponding control.

Figure 6 Transcription size and genomic domains: **(A)** The size of giant transcripts (up to 50-100.000 nt) corresponds - by order of magnitude - to the genomic domains observable in specific types of chromosomes (B, C), or the "Christmas trees" of primary transcripts observable in the EM after spreading of nucleoli (D1, 2) or non-ribosomal chromatin (D3). (For exp.details see (Scherrer and Darnell, 1962; Scherrer et al., 1963), reporting the original observation of "giant" RNA and RNA processing; c.f. also Fig 9 in (Scherrer and Marcaud, 1968) and Fig.6 in (Spohr et al., 1976)) **(B)** Lampbrush chromosomes of *Pleurodeles waltl* stained for IIF with anti-prosome monoclonal antibodies (for exp. details see (Pal et al., 1988)). Lampbrush chromosomes are characteristic of the transcription of the entire genome during the diplotene stage of oogenesis in *amphibia* and birds. Projecting from the chromosome axis are the chromatin loops corresponding to genomic domains, which carry the "christmas trees" of DNA in maximal transcription (comparable to those shown in panel D3). Prosomes (insert) are protein particles (built of 2x14 subunits in 4 superposed rings of 7) found associated to chromatin and (pre-)mRNP complexes; they constitute also the core of the 26S proteasomes (Scherrer and Bey, 1994). Notice their association to the loops (maximal at their basis), and also their shedding (arrows) from the chromosomes into the nucleoplasm. **(C)** Polytene chromosomes of *Rynchsciara americana* in specific stages of larval development and differentiation (c.f. F. Lara (Glover et al., 1982; Lara, 1987)). Polytene chromosomes represent interphase chromosomes generated by DNA replication without cell division; about 10'000 DNA strands stay associated and form the bands visible in the light microscope due to chromatin hyper-condensation. These physical bands correspond to the meiotic genes in cytogenetics of, e.g. drosophila, to units of transcription and, in *sciaridae*, of DNA replication. Notice the development of transcriptional "puffs" at specific stages of differentiation. (D) Transcription and formation of nucleoli (relation of transcription and nuclear architecture). (1) Organised nucleolus with its fibriller centre (F) where transcription takes place and the granular zone (G) constituted by already processed ribosomal subunits. (2) Hamkelo-Miller spreads of dissociated nucleoli allow to see consecutive ribosomal DNA domains in transcription: the ribosomal transcripts form RNPs, which, eventually, are organised, into the nucleolar dynamic architecture. (3) Transcripts of non-ribososomal genomic domains of various sizes.

Figure 7 Transcription, (pre-)mRNA transport and prosome-specific (PS) nuclear matrix and cytoskeleton. **(a)** In situ hybridisation with a globin riboprobe on transformed Avian Erythroblasts (AEV cells) showing 3 cells; the lower two are partially (left) and fully (right) induced for hemoglobin production (exp. details in (Iarovaia et al., 2001)). Notice accumulation of globin RNA around the nucleolus (NO) in the un-induced cell, and the presence of 2 nuclear processing centres (PC) and of mRNA in the cytoplasm after induction. **(b)** A partially induced AEV cell *in situ* hybridized with a globin riboprobe (red) as in (a), counterstained by IIF with a 23K-subunit-specific anti-PS monoclonal Ab (23K p-mAb) serving as a marker for nuclear and cytoplasmic (pre-)mRNPs (green); white dots indicate a 1 / 1 ratio of the two markers and, hence, co-localisation of globin RNA with the 23K-type PS (exp. details in (De Conto et al., 1999)). Notice the abundance of globin mRNA-23K PS complexes at the periphery of the PCs extending to the nuclear membrane, as well as their presence at specific sites in the cytoplasm where repressed globin mRNPs accumulate, whereas the 23K PS distribute throughout the cytoplasm, similar to globin mRNA in (a). **(c, d)** Nuclear matrix preparations of mouse myoblasts stained with the 23K-specific p-mAb, prior and after RNase treatment (exp. details in (ref FdC (??)). Notice the presence of about 50% of the 23K PS-mRNP complexes on the nuclear matrix and the appearance, after RNase, of PS-specific networks within the matrix engulfing the nucleoli (black craters). **(e, f)** Two types of Prosome-specific cytoskelettal networks co localising both with cytokeratins (exp. details in (Olink-Coux et al., 1992)). Epithelial cell stained with a p25K-specific (e)

and a p33K-specific p-mAb (f). Notice that different networks are occupied by the two types of PS (although both corresponding to the cytokeratin type of IF), as well as the peri-nuclear staining and filamentous links in between cells; in (f) the PS are on a network starting at the Golgi centre and ending at the plasma membrane on desmosome-like patches.

Figure 8 The physical supports of gene expression and storage. Not only proteins, but also DNA and RNA are organised in space. In proteins, "spacer" peptides place active sites in precise positions and intra- and intermolecular interactions create the 3D structure necessary for function as enzymes or structural building blocks. DNA and RNA interact with proteins not only for control of gene expression at genon level but secure the also the nuclear constitutive and dynamic architecture: DNPs and pre-RNPs constitute the skeleton of the nuclear matrix. The relatively stable 3D DNA network is modified during differentiation and physiological change. The RNA in processing, as the secondary backbone of the nuclear matrix, permanently controls the dynamic nuclear architecture securing transport of the integrated information of gene and genon. This primary transport system is prolonged into the cytoplasm by the 3 cytoskelettal systems of actin, intermediate filaments and tubulin. Thus, gene fragments are in defined 3D positions where transcripts are generated, migrate to nuclear processing centres and export systems to end up in defined cellular sectors or structures where genes are delivered to the places of their function. All these mechanisms are highly controlled in the 3D space; breakdown of the underlying systems leads to malfunction and pathology as particularly visible in cancer cells which, quite generally, show modifications, and even breakdown of matrix and cytosquelettal organisation.

Figure 9 The Unified Matrix Hypothesis (Scherrer, 1989) postulates the existence of a 3D network of Chromatin primed by intrinsic properties of the genomic DNA. This constitutes a third type of genetic information based essentially on the distance of sites where two DNA strands interact, at distant sites on the same and/or on different chromosomes; mere DNA length becomes a genetic information. (A, B) The network of *Ectopic Pairing* shows the existence of such a 3D chromatin system, as observed for the 4 polytene chromosomes in *drosophila* salivary gland cells (A) which are genuine interphase cells (micrograph courtesy V. E. Barsky; c.f. (Ananiev et al., 1981)). Notice intra- and inter-chromosomal as well as telomeric links. The cables suspend the nucleolus in a fixed position; since it contains the highly amplified genomic domains for ribosomal RNA; notice that the DNA must pass through some of these cables. (B) the position of these cables linking interbands is genetically fixed (Kaufman et al., 1948). (D, E) The formation of the matrix network: The DNA in normal interphase cells being flexible, it may directly interact at specific sites (A_1 - A_n in C) within and in between the chromosomes, eventually forming a 3D network (D) of euchromatic chromatin and, secondarily, the matrix protein network (dashed lines) binding to the Matrix Attachment Regions (MARs; small dots). Condensed heterochromatin (fat dots) can not participate to this system; the DNA network is modified mainly during differentiation by conversion of hetero- and euchromatic and by epigenetic modifications. (E, F) Correlations of UMH and the Chromosome Field theory (Lima de Faria, 1979). Aligning (by increasing length) chromosome arms (centromers vertical to the left, telomeres right on a borderline at 45° angle) carrying the ribosomal DNA of same and neighbouring species, it appears that the rDNA is always at an identical chromosome position relative to centromer and telomere (E). The nucleolus being in a fixed position in the ectopic network (see A), this fact might be explained according to the UMH (F): a specific position in space would imply a specific position along the DNA and, as the result, in the derived 3D network.

Figure 10 The Cascade of Regulation (Scherrer, 1967; Scherrer, 1980; Scherrer and Marcaud, 1968): The information content of the zygotic genome is gradually reduced to that expressed in a differentiated cell. In Homo s., information for an estimated 500.000 polypeptide-genes are reduced to a few hundred in gradual steps; as few as 3 genes may account for up to 90 % protein output, as is the case in red blood cells (Imaizumi-Scherrer et al., 1982). The Holo-Cascade (not shown) includes additional steps, leading upstream from the information content of an entire species to that of populations and individuals, and downstream from the polypeptide to the assembled, functional protein including all post-translational modifications (Scherrer, 1980). Under the direction of holo-Genon and holo-Transgenon, the DNA reduces the genomic information by DNA rearrangements to that of an individual cell, and then by individual steps of processing to that necessary for the expression of an individual function, as shown here and outlined in the text. These may include: (1-2) Chromatin modification and activation (proto-genon-dependant); (3) transcription and formation of pre-mRNP (pre-genon); (4-6) gradual processing and splicing (pre-genon); (7) export and formation of

cytoplasmic mRNP (genon); (8-9) activation (de-repression) of mRNP (genon); (10) translation of mRNA (genon) followed by peptide formation (genon has expired).

Figure 11 Endo- and Exo-cascade. The information guiding gene expression stems not only from the genome but also from the outside of cell and organism. Genon and transgenon are directly or indirectly modified by input from the Exo-system (for organisms, possibly, the *ecosystem*). **(A)** Information Processing. From the DNA to the individual gene and phenotype, the genomic information decreases, eliminated by selection of domains and RNA processing. Concomitantly, external input is integrated into the expression process, guiding selection, specific processing and activation of specific genons and mRNA, mainly via the holo-transgenon, composed of factors encoded either by the genome or else imported from the outside of cell and organisms. **(B)** Within the cell, the genomic cascade of regulation (Endo-cascade) is infiltrated by the information from outside cell and organism (Exo-cascade). This input is highest at the periphery of the cellular systems: the organism, the cellular membrane, the mRNA-genon, but may reach the pre-genons, as well as the genomic DNA, as detailed in **(C)**.

Figures 1 - 6, Scherrer and Jost (2007) submitted to Theory Biosci.

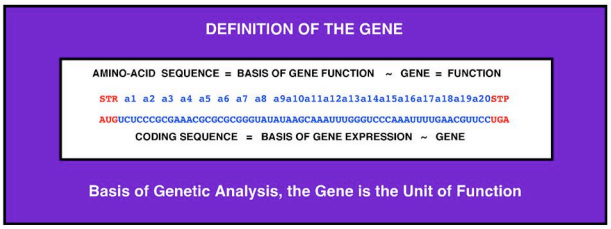

Figure 1

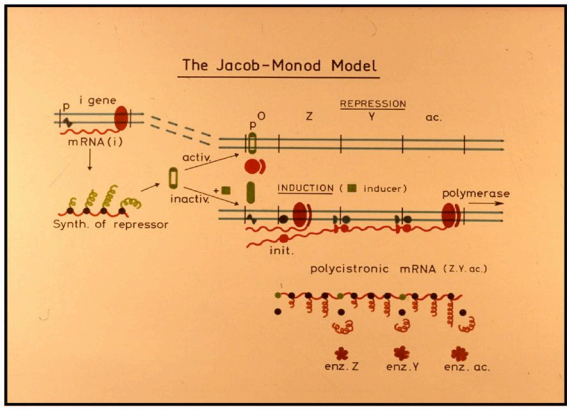

Figure 2

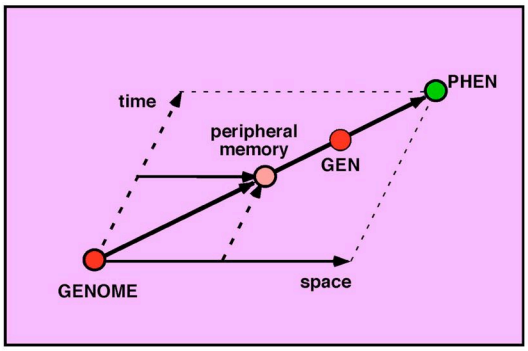

Figure 3

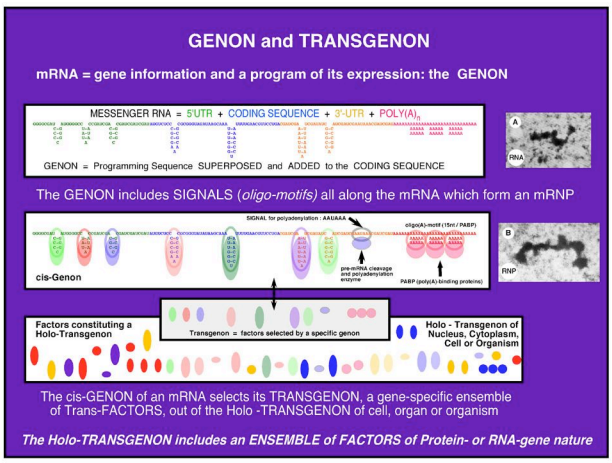

Figure 4

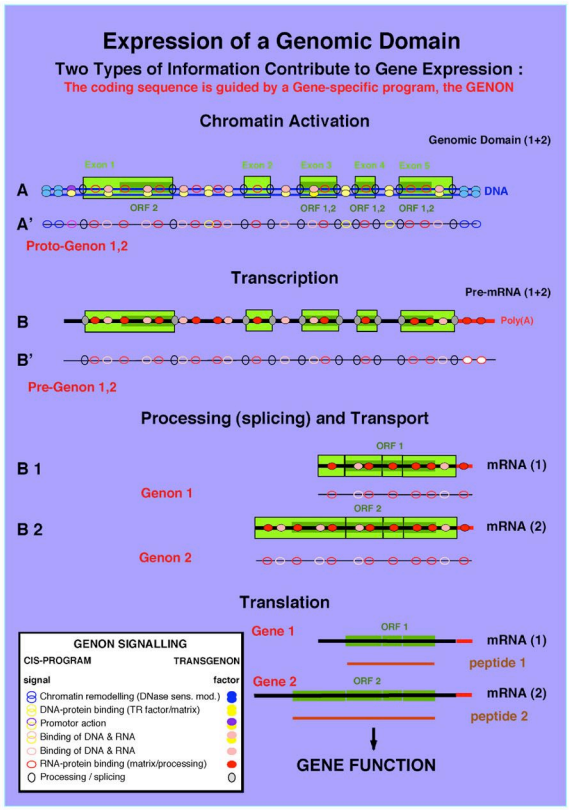

Figure 5

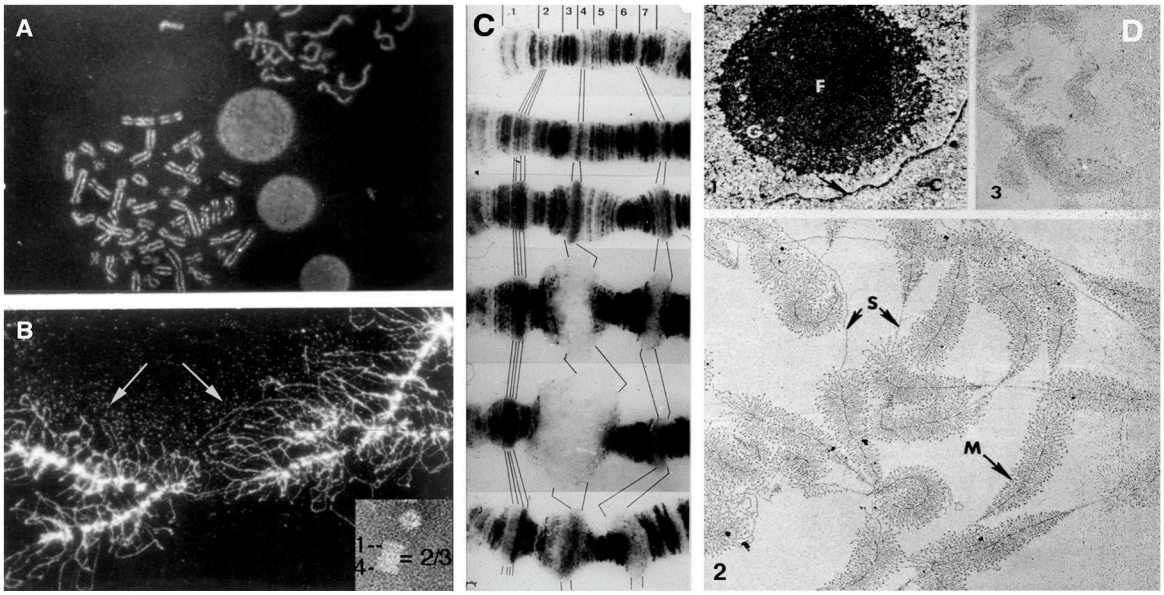

Figure 6

Figures 7 - 8 Scherrer and Jost (2007)
submitted to Theory Biosciences

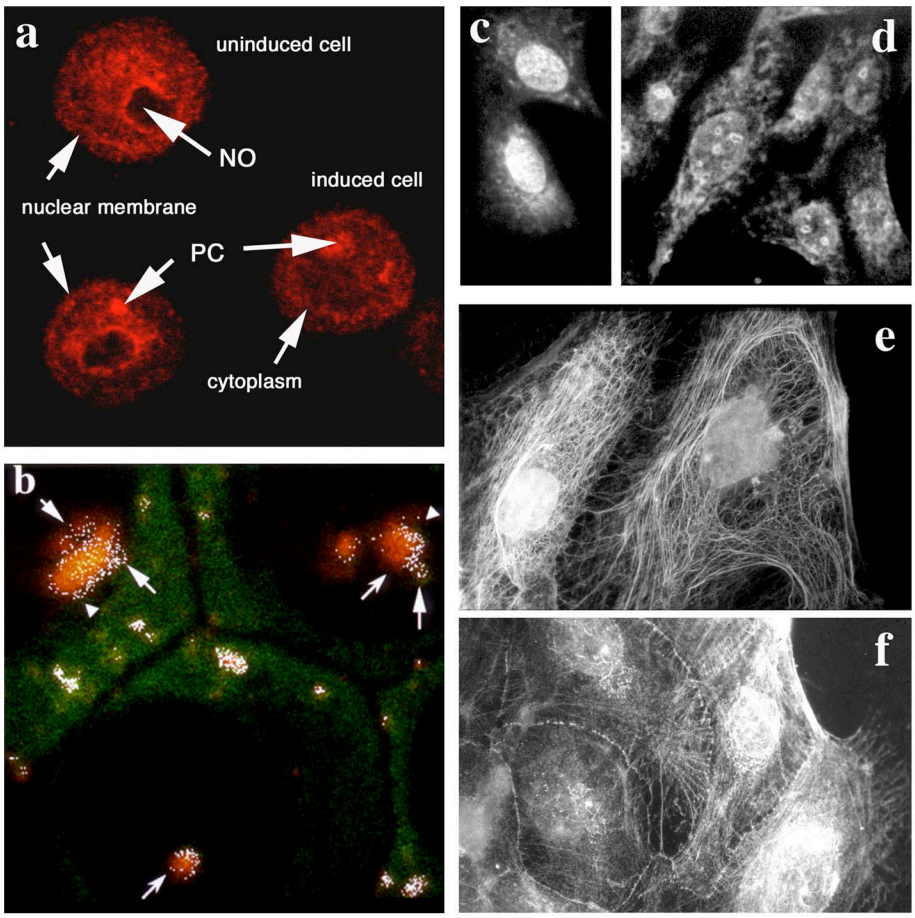

Figure 7

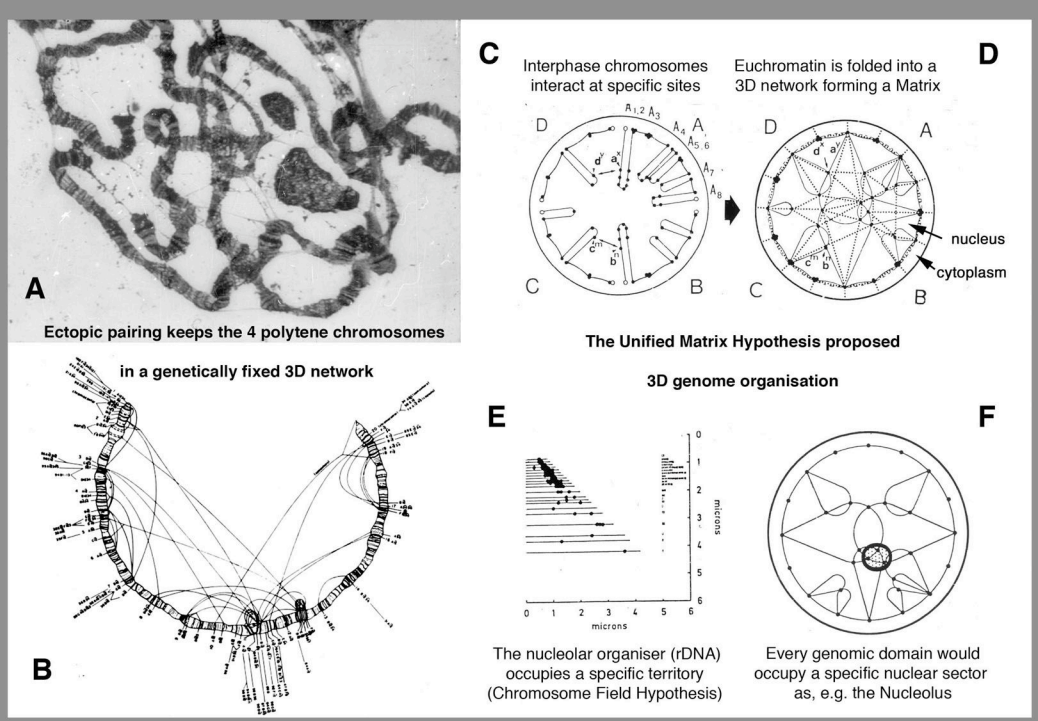

Figure 8

Figures 9 - 11, Scherrer and Jost (2007)
submitted to Theory in Biosciences

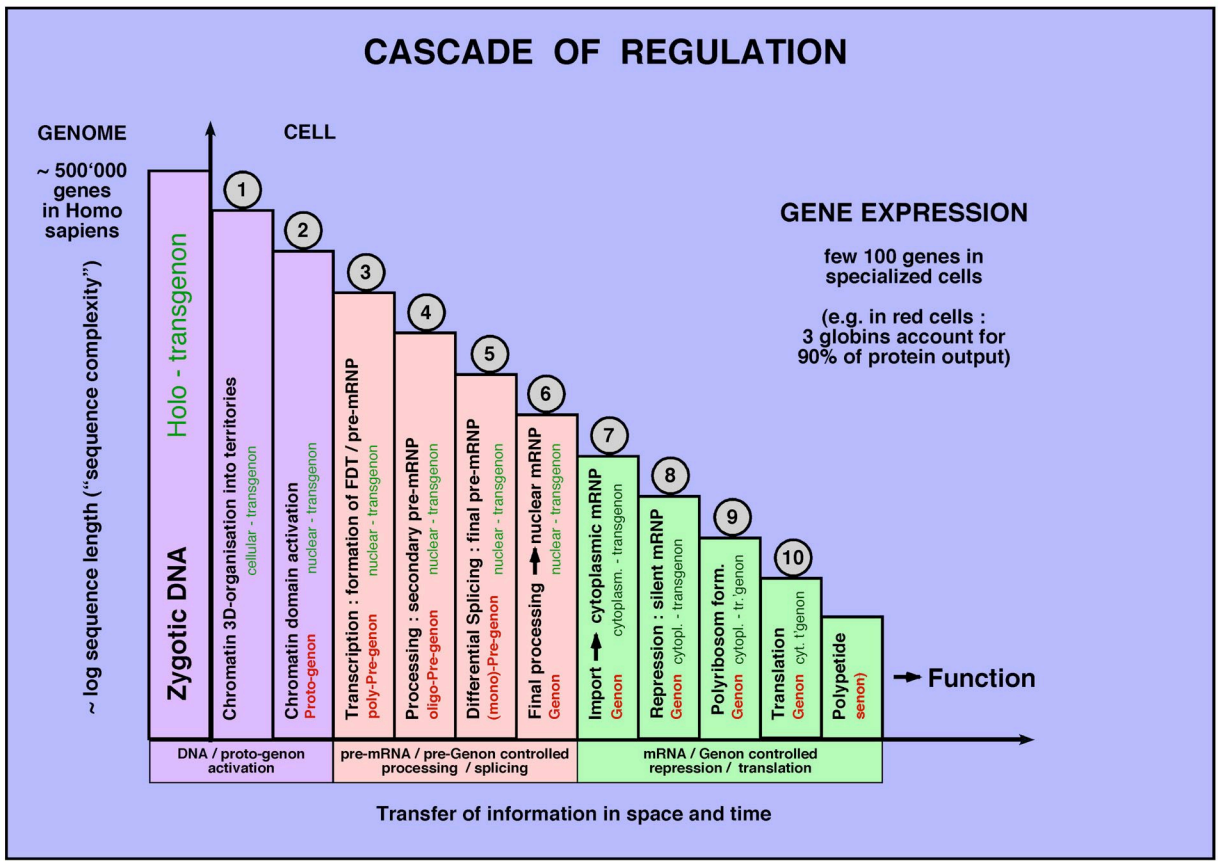

Figure 10

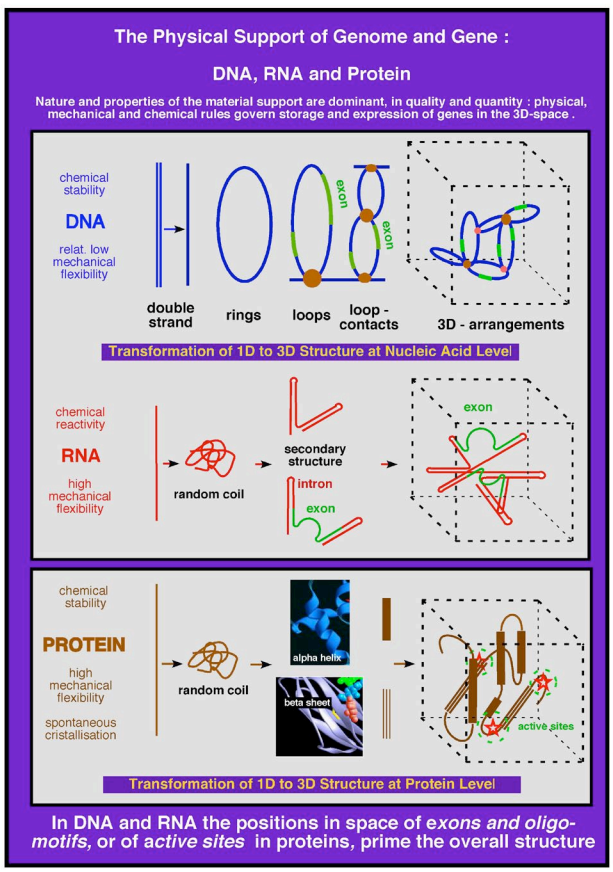

Figure 9

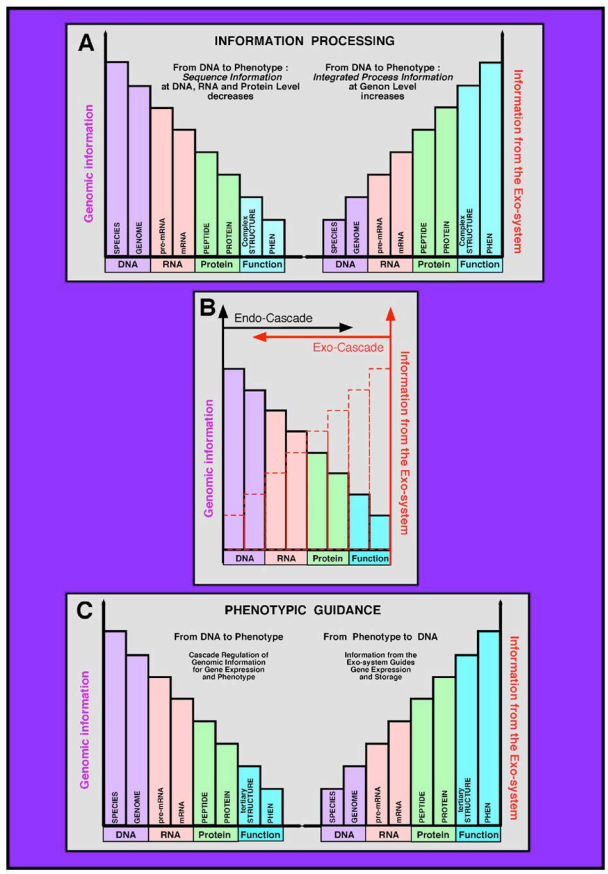

Figure 11