Theory Biosci. (2014) 133:1–21
DOI 10.1007/s12064-013-0182-7

ORIGINAL PAPER

Information theory, gene expression, and combinatorial regulation: a quantitative analysis

Jürgen Jost · Klaus Scherrer

Received: 2 September 2011 / Accepted: 19 April 2013 / Published online: 15 May 2013

Abstract According to a functional definition of the term "gene", a protein-coding gene corresponds to a polypeptide and, hence, a coding sequence. It is therefore as such not yet present at the DNA level, but assembled from possibly heterogeneous pieces in the course of RNA processing. Assembly and regulation of genes require, thus, information about when and in which quantity specific polypeptides are to be produced. To assess this, we draw upon precise biochemical data. On the basis of our conceptual framework, we also develop formal models for the coordinated expression of specific sets of genes through the interaction of transcripts and mRNAs and with proteins via a precise putative regulatory code. Thus, the nucleotides in transcripts and mRNA are not only arranged into amino acid-coding triplets, but at the same time may participate in regulatory oligomotifs that provide binding sites for specific proteins. We can then quantify and compare product and regulatory information involved in gene expression and regulation.

Keywords Gene expression · Gene regulation · RNA · Combinatorial regulation scheme · Information theory

Introduction

In the terminology of present molecular genetics, the term "gene" can refer to genes conceived either as protein products, or else as DNA stretches with grouped genetic information including regulation. The about 20,000 protein-coding "genes" assumed to exist in the human genome [13] relate clearly to the latter. In this situation, in the literature, frequently a new clarification of the gene definition is called for; see for instance [10]. In fact, in [30, 31], we have already developed a gene definition from the perspective of the functional products produced in the cell. Our gene definition clearly distinguishes between the product and the regulatory aspects. This will provide us with the conceptual framework to analyze the relationship between DNA stretches and functional products in a quantitative information theoretical manner.

In the present paper, we shall only discuss protein-coding genes, and leave out the issue of genes coding for functional RNAs, although a similar analysis is possible. Neither will pseudogenes be considered here. More precisely, a polypeptide is represented by an mRNA coding sequence prior to translation, together with the choice of a translation start site, as in some cases, several such start sites exist and consequently different polypeptides can be produced from the same mRNA coding sequence. In eukaryotic cells in particular, such an mRNA coding sequence cannot readily be identified with a stretch of DNA in the genome. Rather, it is assembled in the course of a complex regulation process from various pieces (exons) of pre-mRNA that are in turn built up from DNA transcripts.

Many details are well known, but to set up the frame for our analysis, we shall now briefly recall some of the essential steps of this process. First of all, there are various alternatives in the expression pathway, ranging from non-expression to different types of polypeptides created by alternative splicing from one DNA domain or transcript. The regulation of individual genes is determined by the various possibilities for regulatory factors controlling the expression pathway that bind at sites at the DNA, pre-mRNA or mRNA level. These regulatory factors, which

J. Jost (✉) · K. Scherrer
Leipzig, Germany
e-mail: jjost@mis.mpg.de

could be of protein or RNA nature, typically bind at specific nucleotide sequences, so-called oligomotifs.

To account for this, in [30, 31] we have coined the term "genon" (and various derivations of it, like "pregenon" or "protogenon", referring to the pre-mRNA and DNA level, respectively) for the program at the mRNA level controlling in cis the expression of a gene. This is materialized in those oligomotifs, that is, specific factor binding sites, and it is thus superimposed onto and added to the coding sequence in cis. Therefore, these genons are encoded already in the DNA in the same strand as the coding sequence. In particular, one and the same stretch of DNA or RNA can have both a coding and a regulatory function. While these functions may share the same material substrate, they need to be conceptually distinguished. In particular, this may break the redundancy of the genetic code. That is, different nucleotide triplets coding for the same amino acid may contribute to different regulatory functions within distinct genons. Consequently, mutations that are neutral with respect to polypeptide coding may well have deleterious consequences because they may interfere with gene regulation. Concerning the issue of the different possible translational start sites, the genon corresponding to the mRNA with the full coding sequence may pick up alternative factors exposing different start codons.

However, genes are not functional and hence not expressed in isolation. Rather, the essential aspect is the coordinated activation of specific sets of genes according to the state of the cell and its environment. Since the number of factors (of polypeptide or RNA nature) a genome can accommodate is necessarily limited, these factors have to operate by combinatorial rules to coordinate the activation of specific combinations of genes.

It is the purpose of this paper to evaluate and quantify these aspects in information theoretical terms. We shall do this on the basis of presently available numbers for some eukaryotic genomes, in particular the human genome. Because of the present uncertainty about numbers and roles of directly functional or regulatory RNAs, we restrict our analysis here to protein-forming genes. In particular, we do not enter here into the mechanisms of RNA interference (RNAi), which are still not fully understood. Nevertheless, our general scheme is developed in such a way as to be also applicable in principle to such RNAi.

Thus, we shall distinguish between regulatory and product information, i.e., the genon on one hand and the coding triplet sequence on the other. We shall attempt to quantify information about ensembles of sequences, regulation and coregulation, and individual expression pathways. Let us first consider the classical aspect. The information in a sequence quantifies the reduction of uncertainty when we obtain a specific sequence from an ensemble of sequences with the given symbol frequencies and correlations. In our situation, the sequences are either composed of nucleotides or of amino acids as symbols. Because of the redundancy of the genetic code, a coding sequence underlying a gene contains more information than the polypeptide, that is, the amino acid sequence the gene is coding for. Let us secondly consider the regulatory aspect. The genon and its precursors, the proto- and pregenon, as materialized by sequence fragments or pieces like oligomotifs, also carry a certain information. The information concerning regulation of individual genes and networks is determined by the various possibilities for regulatory factors controlling the expression pathway that bind at the sites provided by proto-, pre- and genon. These regulatory factors implement decision about expression versus non-expression or select one of several alternative splicings or at a later stage in the expression pathway decide which genes are translated and how many copies of them are produced. As already discussed, here, we are concerned with the coordinated activation of specific sets of genes that are appropriate given the current state of the cell and its environment. In particular, we shall propose and analyze a particular combinatorial model for the interaction of RNA binding proteins with oligomotifs in the mRNA. This model will flexibly account for the selection of specific ensembles of genes.

Our numerical estimates are often tentative and may need to be considerably revised on the basis of more accurate future data than available at present. So, what do we gain from this exercise?

1. A theoretical principle for quantifying the contribution of various steps and interferences in the regulatory process. Whenever we have precise data about the numbers of factor binding oligomotifs and the complementary regulatory factors, we can compute the number of different possibilities that can be realized from one and the same coding region. This will offer new possibilities to bring insights from such fields as cybernetics or control theory into molecular biology. For instance, we may quantitatively estimate the ability of a cell to compensate for external disturbances and to maintain a stable function in the presence of such disturbances.
2. A framework for quantitatively addressing not only which genes are expressed, but also how many copies of them. While this is obviously important, it does not seem to have received much attention so far in theoretical models.
3. A simple formal model for the coregulation, that is, the coordinated activation of large, but very specific sets of genes. Again, that latter model makes some simplifying assumption, but hopefully this will serve to best bring out the underlying principle. In particular,

we shall see that there is an astounding number of possibilities to select specific combinations of large numbers of genes through the combinatorial interaction of a few oligomotifs and regulatory factors. This power of combinatorics seems to have not been appreciated so far in theoretical molecular biology.

Information theory

Since we shall make repeated use of Shannon's theory of information, we repeat here its basic aspects. Shannon's theory is concerned with a sender that transmits symbols to a receiver. The receiver does not know which symbol out of n possible ones she will receive at each step. Since the nature of the sender is not so important for our purposes, we can also consider a receiver that simply observes events. She only knows that the symbols or events occur with frequencies $p_v, v = 1, \ldots, n$, the information obtained by observing a particular such event is the average (negative) logarithm of these probabilities,

$$I = -\sum_{v=1}^{n} p_v \log p_v. \tag{1}$$

We note that for quantifying this information obtained by observation of an event, it does not matter which event actually happens. Only the probabilities of the events matter. I measures the amount of uncertainty before an event was observed. That is, the information I quantifies the amount by which the uncertainty is reduced by the observation of which one of the possible events actually does occur.

Also, when we observe N independent ensembles, with information content $I_k, k = 1, \ldots, N$ each, the total information obtained by observing one event from each ensemble simultaneously then simply is the sum

$$I_{\text{total}} = \sum_{k} I_k. \tag{2}$$

Before proceeding, we make a general remark. As just explained, the quantity of information depends on the choice of an ensemble of events ν with their relative frequencies or probabilities p_ν. If we change the ensemble or vary the probabilities, the information obtained by observing a specific event also changes. For instance, when we look at the ensemble of all combinatorially possible strings of 300 amino acids, and if we assume that each of the 20 different amino acids occurs with the same probability $\frac{1}{20}$ which is somewhat smaller than 2^{-4}, we have 20^{300} different possibilities and the information I in (3.3) then is more than $4 \times 300 = 1{,}200$ bits. We can reduce the uncertainty, however, already by restricting only to those amino acid sequences that are potentially biologically useful, that is, lead to properly folded polypeptides (ignoring exceptions like casein which does not fold, but is nevertheless biologically useful). These are much fewer sequences, and most of the above figure of (more than) 1,200 bits is already diminished by that reduction. That is, as soon as we know that an amino acid sequence leads to polypeptides properly folded into 3D structures, we have already gained a lot of information. On the other hand, then finding out which particular polypeptide is actually realized by our sequence yields comparatively little additional information. Thus, the above 1,200 bits can be broken down into two summands, one containing the information gained by knowing that an amino acid sequence corresponds to some polypeptide, that is, is a sequence that can fold properly, and the other specifying which particular polypeptide it actually is (this example will be taken up again in Sect. 4.1 below).

As in the example of amino acid sequences just discussed, we are usually interested in ensembles of sequences. In fact, the sequences to which we wish to apply information theory can be of rather different natures. Therefore, we develop some mathematical framework first. We employ the formal notion of an alphabet A consisting of letters α, and we shall now discuss different types of such alphabets. When we consider nucleotide sequences, the letters stand for the four different bases. When we consider coding sequences, we have 64 letters, one for each triplet. For amino acid sequences, we have 20 different amino acids, i.e., letters in the corresponding alphabet. Below, we shall also be interested in sequences of regulatory oligomotifs ("oligos" for short). In our examples, there will be typically about 1,000 different such oligos. Thus, we then have an example of an alphabet with about 1,000 letters. Proceeding with the terminology, strings of letters are called words. Let us again spell this out for our examples. When a nucleotide sequence consists of 40 kbp, we have a word of 40,000 letters from our nucleotide alphabet. When a coding sequence consists of 300 triplets, or a polypeptide as an amino acid sequence contains 300 amino acids, we have a word of length 300 in the corresponding alphabet. Finally, when we consider an RNA molecule containing 100 oligos, we consider it as a word of length 100 in the oligo alphabet, whereas as a nucleotide word, it will be much longer, say 10,000 letters when it is composed of that many nucleotides.

An important issue is whether all words that can be formed by the letters of our alphabet are possible or relevant to the biological situation at hand, or whether there exist constraints. For the amino acid words, we have already discussed such a constraint above, the folding criterion. On the other hand, we could also take the list of

all proteins recorded for human cells. If such a complete list were known, we could work with a much smaller set of amino acid sequences than those that properly fold. Obviously, such a scheme would miss out new sequences generated by genetic mutations.

Having clarified how the amount of information depends on the choice of the ensemble, we approach another issue. When applied to the molecular biology of the cell, information theory can be used in different ways, possibly with different results, and this brings us to the core of the present article. On one hand, we can compare the starting point, the genome, with the final result, the proteome, and quantify the relation between DNA segments and polypeptides in information theoretic terms. That is, we can ask how much control is needed for the regulation process to produce the protein mix in the cell from its DNA according to external and internal circumstances. The corresponding computation does not take into account the actual details of the regulation process, that is, the expression pathway, but only compares the input and the output of the process to assess its complexity. On the other hand, we can also quantify the number of decisions and their relative frequencies within the expression pathway in information theoretic terms. We can then compare the results of the two computations, for the input–output relationship on one hand and for the sequence and network of regulatory steps on the other. The question is whether the amount of information from these two computations coincides or not. If we had an optimally implemented, non-redundant regulation process that does not utilize any external information, the two information quantities should be the same. If they are not the same, we should try to interpret and understand their difference. The regulatory information might be much larger than the input–output information. A reason for the discrepancy might be that the control process could be redundant and non-efficient. It might also be the case, however, that the regulation process has to be stable in the presence of external perturbations, that is, that it has to compensate for variations in external conditions. According to Ashby's law [3], the amount of internal information needed to compensate such variations has to be at least as large as the entropy of these variations. This amount cannot be detected in the input–output relation. This effect increases the regulatory information. This, however, may be counterbalanced by another effect working in the opposite direction. That effect makes use of information available in the environment. Certain substances, such as vitamins, may participate in cellular processes and feed external information into them. Of course, for the analysis of the life cycle of a virus, the utilization of information that is provided by the host cell, and therefore external to the virus, is the key. Here, however, we are concerned with eukaryotic cells and for them externally provided information should play a less dominant role. Therefore, we should expect this effect to be smaller in information theoretic terms than the first one, the compensation of external perturbations. Therefore, computations of the regulatory information yield larger results than computations of the product information concerned with what polypeptides are produced in which quantities from which portions of the DNA.

In any case, on the basis of present molecular biological knowledge, we are not yet in a position to produce exact computations. Rather, we can only provide estimates for the amounts of information and that is what we shall attempt in the sequel.

Transcription

Genomic domains

In the following discussion, the definition of genomic domains will be of particular importance to estimate types and numbers of transcripts and genes and we therefore develop our definition here. As already discussed, most protein-coding genes as sequences coding for polypeptides do not exist at the genomic level as physical entities, but have to be composed during RNA processing and, in particular, by differential splicing. The latter mechanism allows for producing several genes, and thus genons, from a primary transcript and leads to a multiplication of functional products encoded in a single DNA region. The recent finding by transcript mapping that most of a (human) chromosome is actually transcribed [12] can be related to the old finding that the primary transcripts carrying information for protein biosynthesis are very large, as pre-mRNAs or full domain transcripts (FDTs). The latter best represent the division of the genomic DNA into domains as units of transcription.

Genomic "domains" have been extensively discussed in the past [24, 27]. It was an early observation that division of the Drosophila haploid genomic DNA by the number of bands observed in the polytene chromosomes [11] results in stretches of DNA length, compatible, by order of magnitude, with the observed size of primary transcripts in many types of cells [32]. The polytene chromosome bands are not only physical entities seen in the microscope, but correspond also to units of meiotic recombination. On this basis, genetics was done at the time of Morgan and his school; classical cytogenetics is based on the division of the genome into such units.

The most recent analysis of Drosophila chromosomes, as published within the NCBI *Drosophila melanogaster* fruit fly release 5.2, correlates well with these basic facts [1]. Figure 1a shows an actual representation of a

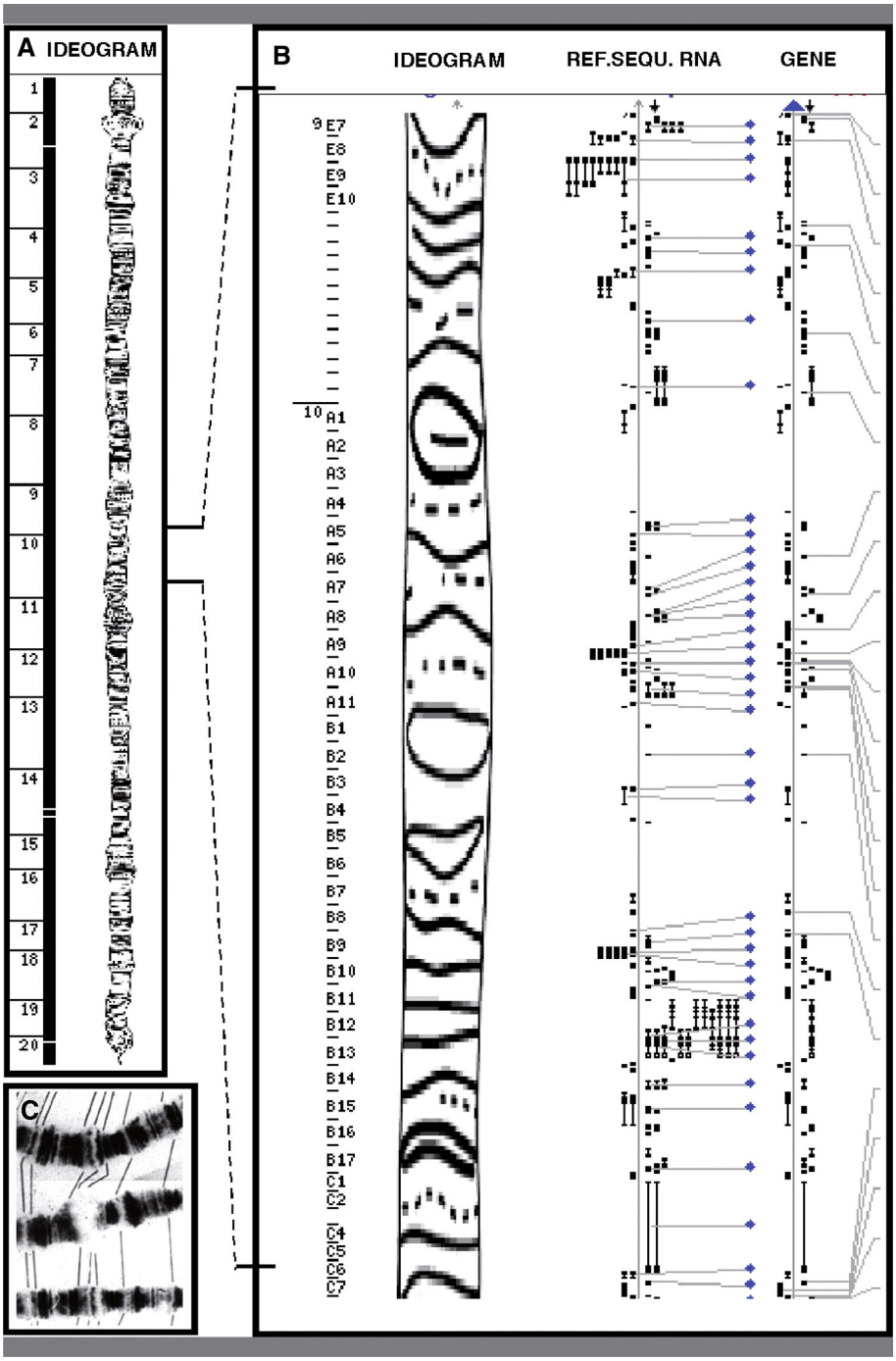


Fig. 1 Genomic domains in *Drosophila melanogaster*. Chromosome organization shown is chromosome X **a** full length and **b** enlargement of region 9/10 including about 1,000 kbp and **c** a typical puff of band in active transcription. **a** On the *left* is shown the subdivision into chromosome segments and the overall ideogram of the chromosome as seen by microscopy (*right*). **b** Enlargement of region 9E7 to 10C7. From *left* to *right* Chrom. region, microscopic ideogram, RNA map, and gene positions. Note the alignments with chromosome bands and the extent of mapped transcripts. In the ideograms, the *black band* corresponds to the hetereochromatic condensed chromatin. The *blank regions* in between are the interbands which generally are AT-rich [22] and accumulate RNA polymerase 2 (ref) which, when a band is transcribed after "puffing", spreads all over the puff. **c** A *Rynchosciara americana* chromosome band in puffing [20]; notice the puffing when in transcription and eventual regression into a heterochromatic band

Drosophila melanogaster chromosome banding pattern, illustrating such domains of chromosome structure and function. The bands seen are not only observable physical units and units of transcription, as already mentioned, but in the *Sciaridae* insect family (Fig. 1c), they are also units of replication. Indeed, the DNA in such bands must locally be amplified prior to transcription [20]. As shown in the annotated representation (see the URL in [1]) of these cytogenetically defined bands in Drosophila (Fig. 1b), the genes identified as products map to these bands. Particularly interesting is the fact that transcription mapping shows most of the DNA in these areas to be transcribed.

Table 1 gives a quantitative summary of these most recent data available based on the *Drosophila melanogaster* genome sequence, annotation and cytogenetic data illustrated in Fig. 1b. Excluding the Y chromosome, all together there are up to 5,792 bands including 14,560 genes (about 3 genes/domain); on average there are 20,768 bp/

Table 1 *Drosophila melanogaster* chromosome bands: summary of the NCBI Map Viewer presenting a graphical view of release 5.2 of the annotated *Drosophila melanogaster* genome (incorporating all available heterochromatic sequences into the assembly and including an annotation update; see Fig. 1)

Chrom./arm Names	Bands total number		Genes total number		Chrom. size (kbp)	Band size (kbp)
	Chrom.	Cont. reg.	Chrom.	Cont. reg.		
X1157	1,016	2,330	2,330	22,400	19,360	
2L	950	806	2,756	2,756	23,000	24,210
2R	1,282	1,138	3,025	3,025	21,100	16,453
3L	1,029	835	2,809	2,809	24,500	23,809
3R	1,318	1,178	3,549	3,549	27,900	21,168
4	56	44	91	91	1,350	24,107
All	5,792	5,067	14,560	14,560	120,250	20,768

Heterochromatic sequences are assembled into six unplaced contigs linked to specific chromosome arms and one unordered superscaffold that contains all unplaced contigs (source: http://www.ncbi.nlm.nih.gov/mapview/mapsearch.cgi?taxid=7227) (numbers of bands or genes in chromosome or contig-region, average)

chromosome band. Extrapolating to the size of the human genome, these units seem to correlate well with the basic facts and assumptions in terms of genome organization into domains in general. In human gene expression, assuming 80,000 protein-coding transcripts (gene-polypeptides) for 20,000 genomic domains (of about 100 kbp each) [13], most likely due to the multiplication factor of differential splicing and to a lesser extent to different translation starting sites, there might be more genes per domain (about 4 genes/domain). Thus, the classical Drosophila polytene chromosome organization still seems to represent best what we mean when speaking of genomic domains as related to genes.

The example of gene expression in the chicken alpha globin gene domain along the cascade of regulation

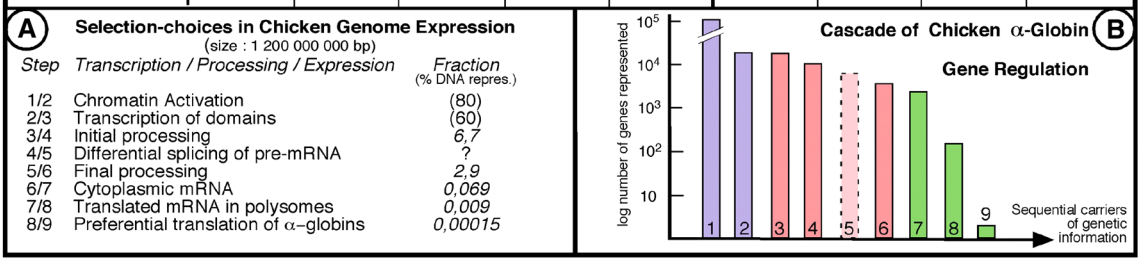
To start with, it must be emphasized that regulation of gene expression is not only a qualitative problem, but, in the first place, a quantitative one. To regulate about 3,000 genes in *E. coli* is basically a different problem from coping with about 500,000 genes in the human genome. The basic problem is that due to the inherent thermodynamic noise, no physical system can cope with choices better than 1:1,000, and biochemical systems are in fact limited to about 1:100. Direct choices of genes at the DNA level are conceivable for *E. coli* where genes are bundled in operons reducing the 3,000 genes to fewer operons. In eukaryotes, the direct choice of single genes at the DNA level is therefore a priori excluded. The solution is multiple sequential choices of ensembles of genes as contained, typically, at the level of genomic domains, of pre-mRNA and mRNA populations. This matter was dealt with in some publications about the cascade of regulation [27, 32].

Unfortunately, neither for Drosophila nor the human species there are, to our knowledge, data about the types and numbers of transcripts produced in course of RNA processing, determined by systematic fractionation of specific differentiated cells. Therefore, we resort hereafter to the example of the avian red blood cells for which such results are available. We may first recall quantitative data concerning the chicken alpha-globin gene domain which provides the example of a genomic domain that was subject to extensive biochemical analysis in terms of its expression pathway [18]. It stands for a highly specialized cell in terminal differentiation with a protein output consisting of 90 % of hemoglobins; there are three different, embryonic, adult minor and major hemoglobins, constituted by the embryonic (π) as well as the adult major (αA) and minor (αD) globins, combined to embryonic or adult beta globin ($\alpha_2 \beta_2$). The three alpha genes are contained in a genomic domain of 30–40 kbp which seems to be fully transcribed into an FDT [25]. This biological model represents a rare case where the length of (non-repetitive) the sequence (called sequence complexity in the biochemical literature) at the genomic as well as the many RNA processing and regulative steps of the cascade of regulation ([27]; see insert B in Table 2) was quantitatively determined by Cr_0t re-association kinetics (Cr_0t and C_0t analysis see [6]), after careful cell fractionation and isolation of the nucleic acids. Table 2 gives the summary of the analysis and insert B in Table 2 is a simplified graphic representation of the "cascade of regulation" [27, 32] (discussed recently in [30, 31]), which reduces the information content of the genome to that of an ensemble of genes actually expressed at a given time in a cell. Table 3 summarizes the actually available data based on recent genomics on the one hand and former results obtained by cell and RNA fractionation and reassociation kinetics [18], on the other.

The great surprise from this type of analysis was that in a highly specialized cell, on synthesizing just a few proteins in very large amounts, as in the immature red blood cells, a relatively large fraction of the genome was transcribed. Indeed, Cr_0t analysis of pulse-labeled RNA gave, for RNA of more than 1.5×10^6 Mr, a value of 11.0 and 4.2 % of the ($2N$) genome represented resp. for the highly unstable primary transcripts and slowly turning over RNA (before and after a 40 min chase by arrest of transcription with actinomycin D) and 12.4 % resp. 5.6 % for RNA larger than 5×10^4 Mr (Fig. 1 and Table 1 in [18]). In other types of cells, by order of magnitude, similar values were found for nuclear and polyribosomal mRNA sequence complexities. These figures were well in line with earlier estimates of 10–20 % of the genome being transcribed in avian erythroblasts obtained by saturation hybridization in

Table 2 The cascade of globin gene regulation in avian erythroblasts: (table) for DNA and the successive RNA fractions during pre-mRNA processing and translation, the α-globin mRNA concentration and the number of different types of mRNA represented were calculated (plain print) or else, determined (italics) by cell and RNA fractionation

COMPARTMENT	NUCLEUS					CYTOPLASM			
NA-protein complexes	chromatin		pre-mRNA - protein complexes			untranslated mRNA - protein complexes		translated mRNA in polyribosomes	
Carrier of Information	total DNA (100 %)	active DNA (60 %)	primary RNA	processed pre-mRNA initial	processed pre-mRNA final	total mRNA	repressed mRNA	total mRNA	globin mRNA
α-Globin-RNA Concentration (3 x 600 nt) fraction	$1.5x10^{-6}$	$7.2x10^{-5}$							
%			*0,01*	*0,05*	*0,10*	*44*	10	*70*	100
Fraction of DNA represented	1	(0.6)	(0.6)	*0.067*	*0,029*	$6,9x10^{-4}$	$6,0x10^{-4}$	$0,9x10^{-4}$	$1,5x10^{-6}$
Number of Genes represented	(250'000)	(150'000)	(150'000)	(15'000)	not appl.	*1'613*	*1'402*	*211*	*3*
SELECTION STEP	1	2	3	4	5	6	7	8	9



plain print : literature-based calculations; (parenthesis) : estimations; *italics* : experimental data (Imaizumi et al, 1980)

Sequence complexity (Cr_0t) and RNA concentration values were determined according to as cited in [18] by hybridisation kinetics of labeled unique sequence cDNA to specific fractions of RNA in excess, as described in [18]. The chicken genome has 1.2 x 10^9 bp [15] and the three mRNAs (aA, aD and p) sum up to 2,000 nt. Note that primary transcripts are highly unstable and could, hence, not be measured by (Cr_0t) (insert A and B). Successive selection of chromatin domains to be transcribed and of transcripts conserved after successive steps of transcription are indicated by systematic reduction of the fraction of genomic DNA represented (for detailed data see Table 3); (B) graphic representation of the Cascade of Regulation of globin gene expression (cf. discussion in [30]) in analogy to the theoretical model published earlier [27]

RNA excess [18, 32]. All these data compare favorably with the actually estimated 60 % of the genome being transcribed in a differentiated [12] cell, since, in those earlier determinations, the RNA analyzed represented only the relatively stable intermediary and final processing products, but not all of the highly unstable (<20 min half-life) primary transcripts.

In avian immature red blood cells, the first step of information processing leading to the selective activation of part of the genome for transcription—and thus the RNA represented in the primary transcripts—corresponds hence to a choice of about 6 in 10 (we replace here the 1982 data by today's estimates as mentioned above). About half of these transcripts are processed within 1–2 h; in other terms, the relatively stable, single sequence nuclear RNA represents about 6 % of the 1*N* genome (or 3 % of the 2*N* DNA), or a total choice of 1 in 20. Whereas the primary transcripts decay with a half-life of <20 min the relatively stable RNA shows, in erythroblasts, two components with half-lives of 3 and 12 h which correspond also to two MW classes of Mr >10^6 and Mr <10^6 [18]; in the latter the sequence length is further reduced.

Turning to the cytoplasm, we have to keep in mind that the terminally differentiating erythroblast represents a non-dividing cell (in g0) and that, therefore, most housekeeping genes are turned off. Thus, for a dividing differentiated cell, we should rather expect about 10,000 different mRNAs.

In the cytoplasmic RNA, the observed Cr_0t values correspond to 6.9, 6.0 and 0.9×10^{-4} of the DNA, or about 1,600, 1,400 and 210 genes of average size (1,000 nt) represented resp. in all of the cytoplasm, the repressed and the translated mRNA fractions. This amounts to a further choice of about 1 in 8 for the genes actually expressed. Finally, the selection of the three α-globin mRNAs (about 4,000 molecules/cell) for preferential translation corresponds to a further selection step of 1 in 70. We thus have to consider sequential reduction of genomic information of 2:3 and 1:20 at DNA and pre-mRNA levels, of 1:40 upon export to the cytoplasm, and of 1:8 plus 1:70 in the cytoplasm, to arrive at the preferential expression of the three α-globin genes. For the human genome (3.2×10^9 bp), this would amount, correspondingly, to the selection of 2×10^3 nt (the major globin mRNAs) in 3.2×10^9 bp, or a choice of 1.5×10^{-6} altogether.

Types of transcripts and genes

For the human genome, to our knowledge no systematic study of the type outlined above for the avian globin genes has been carried out. We thus have to resort to indirect data to estimate quantitatively the successive selection steps during gene expression. Homo sapiens have about 55,000 genomic domains, among which about 20,000 are protein coding [13], and most of them are transcribed in an organism at one time or another [12]. For simplicity, we take into account only the FDTs and thus consider only protein-forming genes derived from primary pre-mRNA (for more details, see the recent data produced within the ENCODE project, which indicate that a major part of the genome is transcribed [10, 12]). In particular, we shall neglect transcripts arising from smaller transcription units (TUs) interspersed in between the FDTs, or superposed onto them as alternative forms of transcription. This may in particular concern the small nuclear and cytoplasmic RNAs, including controlling RNAs acting in RNAi.

At present, the question of the transcription start is fully open: e.g., the assumed mode of operation of the so-called promoters as acting at initiation of transcription is being questioned; actually, they may as well be involved in RNA processing (cf. discussion in [31]). Furthermore, recent data taking into account sparse RNAs and allowing for the detection of transcripts over a concentration range of 1 in 1,000 indicates that about 30 % of transcripts mapping upstream of the conventional transcription start sites have escaped analysis thus far [33]. These data correlate well with the above-mentioned range in so-called sequence complexity (=non-repetitive sequence length) from the primary transcripts to mRNA.

In a typical differentiated cell, about 60 % of genomic domains may be transcribed, that is, a specific cell has about 12,000 different FDTs. 2/3 of those, that is, about 8,000, would be expressed by forming proteins sometime during the life of that cell. At a given time in a specific cell, however, only about 1,000 of these primary transcripts may be completely processed and expressed eventually. Through combination mechanisms like alternative splicing, about 5,000 different mRNAs can be produced from those processed transcripts and are found in the cytoplasm. Out of those, in turn, at a given time of the day, about 1,000 types of mRNA are translated in the polyribosomes, leading to as many different polypeptides. Even though a certain fraction, perhaps <20 % (in terms of mass), of the transcribed genomes correspond to functional products that are not of the protein type,[1] throughout its lifespan a differentiated cell can produce altogether about 50,000 different polypeptides (8,000 different transcripts leading up to 50,000 different mRNAs).

In the entire human organism, about 200,000 different mRNAs resp. polypeptides can be found (cf. discussion in [31]), a figure that may increase with progress in genomics and proteomics. Taking into account the actual state of transcriptome and proteome analysis, and adding the basic immunoglobulin gene rearrangements (excluding individual antibody peptides), perhaps about 500,000 different polypeptides may be found to be synthesized by, e.g., the human organism over its lifespan.

In general, these peptides are organized into multi-subunit proteins; this may either reduce or amplify the number of functional proteins relative to constitutive peptides. Strict homo- or heteromer association would reduce, whereas variable combinatorial subunit association might increase this, hence unpredictable, number. Current estimates may take into consideration the existence of 50,000–100,000 different protein types in a human or other mammalian cell over its lifespan.

For the analysis in the remainder of this paper, we decide to consider the following selection and combination steps; these constitute a subset of the cascade of regulation. To stay coherent, we refer to a specific differentiated cell, at a specific time and stage of physiological condition.

0. Chromatin activation: four out of five.
1. Selection for activation of genomic domains for transcription: three out of five.
2. Selection of those transcripts that are processed at some time during the life of the cell: one out of three.
3. Selection of those transcripts that are fully processed at a given time: one out of ten.

[1] Note, however, that one and the same genomic domain (on both DNA strands and sometimes in both directions) can produce different functional products, some of which may be of protein, but others of the RNA type.

Table 3 Selection-choices in genome transcription and α-globin gene expression in chicken erythroblasts

Step	Processing event	Sequ. complexity (number bp or nt)	Transcripts (no of diff. types)	Fraction DNA (% represented)	α-Globin genes[(d)] (fraction in DNA/RNA)	Choice
	Full genome(xx)	*1,200,000,000*	24,000	100	0.0000015	na
1–2	Chromatin activation	960,000,000	19,200	80	0.0000019	4/5
2–3	Transcription of domains	720,000,000	14,400	60	0.0000025	6/8
3–4	Initial processing pre-mRNA	130,200,000	2,640	*6.7*	0.0000138	0.67/6
4–5	Diff. splicing (ass. 5 mRNA)	na	13,200	na		na
5–6	Final processing	58,800,000	na	*2.9*	0.0000306[d]	2.9/6.7
6–7	Import to cytoplasm	1,613,000[a]	*1.613*	*0.069*	0.00124[d]	0.069/2.9
7–8	Translated	211,000[a]	*211*	*0.009*	0.009[d]	0.9/6.9
8–9	Preferential translation	1,200[b]	2	0.00015	0.5	0.15/9
9–10	Individual globin mRNA	600[c]	1		1	1/3

Data from [15]

Chicken genome size: 1,200,000,000 bp (about 24,000 domains of 50,000 bp average)

The α-globin genes (600 nt each): embryonic (π)/adult major (α^A) and minor (α^D)

Values in italics indicate experimental data (15), plain print indicates calculated or estimated on the basis of data in the literature

Only transcripts containing (exons of) protein-coding genes are taken into account, excluding directly functional or regulatory RNAs

na not applicable

[a] Assumed average mRNA = 1,000 nt

[b] α-Globin gene expression embryonic: π and α^A; adult: α^A and α^D

[c] Differential quantitative expression of the π, α^A or α^D genes

[d] 3 α-Globin genes of 3 x 600 nt (cf. legend Table 2)

Table 4 Figures for the human genome and its transcripts from Gencode version 14 [13]

Total number of genes	55,889
Protein-coding genes	20,078
Long non-coding RNA genes	12,933
Small non-coding RNA genes	9,173
Pseudogenes	13,341
Total number of transcripts	190,051
Protein-coding transcripts	80,413
Full length protein coding	56,728
Partial length protein coding	23,685
Nonsense mediated decay transcripts	12,421
Long non-coding RNA loci transcripts	21,271
Total no. of distinct translations	81,071
Genes that have more than one distinct translation	14,558

Note the difference in terminology. What is called a gene in the table means a genomic domain in the DNA, whereas the protein-coding genes in our sense correspond to the distinct translations in that table

4. Combination of exons in transcripts into mRNAs: fivefold increase.
5. Selection of mRNAs for translation: one out of five.
6. Combination of polypeptides into proteins: one to four polypeptides assembled into one protein.

Numbers of transcripts and gene products

We now ask how many copies of each type of mRNA exist in a cell. Neglecting repetitions (as ribosomal gene domains, e.g.), we may assume that each eventually transcribed genomic domain exists only once.

The number of given RNA molecules can vary widely, according to their type and position in the expression cascade. Concerning pre-mRNAs, spreads of transcription complexes visualized in the EM show domains with maximal polymerase loading, as the ribosomal domains, and others where only a few polymerases can be seen attached over a domain of several 1,000 bps [14]. For the avian globin domains (up to 40 kbp), we have calculated about 100 polymerases to suffice for the production of 20,000 mRNAs over the timespan of about 24 h necessary for red blood cell maturation. Actual concentration of the given pre-mRNAs and their processing products depends of course critically on their metabolic stability which varies widely (primary transcripts turn over with a half-life of about 20 min; global nuclear RNA has half-lives of 3 up to 12 h; most RNA never leave the nucleus but slowly turn over).

For many cytoplasmic mRNAs, one may only find one copy per cell on average at a given time, whereas others can exist in substantially larger numbers. For example, the

three globin mRNAs in an avian blood cell may be present in about 1,000 (α^D), 9,000 (α^A) and 10,000 (β) copies. There exist more than 1,000,000 ribosomes in a cell (they are composed of three types of rRNA (5S, 18S, 28S) and more than 80 different proteins).

The number of proteins of a given type in a cell varies between <20,000 and 100,000,000 (an occurence of >50,000 is considered as frequent). Thus, a cell would contain about half a billion (500,000,000) protein molecules (of 1,000–20,000 different types).

Ensembles of products

The human genome consists of 3.2×10^9 nucleotide base pairs of which only 5–10 % may not be transcribed (telomeres, centromeres, interdomain DNA) (In the 15 cell lines reported in [10], 74.7 % of the genome was covered by primary transcripts and 62.1 % by processed transcripts. Because of the considerable amount of variation between different cell lines, we expect larger fractions when more cell lines are evaluated). Excluding small TUs (tRNA, snoRNA, si/miRNA, etc.), according to present estimates on the basis of high throughput data, there are about 20,000 genomic domains that contain protein-coding genes, their length ranging between <1,000 and 2.4×10^6 base pairs, peaking at about 27,000 bps, a figure in line with the data given for Drosophila in Table 1 (in fact, experimental data taking into account natural [16, 17] or experimental [22] fragmentation of DNA into such putative domains show a modal size distribution corresponding to a size of 40,000 bps). Each of them contains between 1 and 178 exons, 8.8 on average (median 7). As the human genome sequencing data [19] show, an exon can contain between <10 and 17,000 base pairs, 145 on average (median 122), but with a modal peak at 30–40 bp [9]. The average intron length is 3,366 (median 1,023) (thus, the introns are considerably longer, and their length is more variable than the exons). Combining exons through splicing then leads to an average mRNA length of about 1,300 nucleotides (average 1,340, median 1,100), with a coding sequence of about 900 nucleotides, corresponding to an average length of about 300 amino acids in a polypeptide. Thus, 900 out of the 1,300 nucleotides contain coding information, whereas the remaining 400 ones have a purely regulatory function, as for example the 5′- and 3′-UTRs. At least half of the coding sequence, however, also has some regulatory function, in the sense that it contains oligomotifs as binding sites for regulatory proteins. Furthermore, RNAi via si/miRNAs may operate on not yet known oligomotifs; we will therefore exclude RNAi in our further discussion. In addition, those coding nucleotides that do not participate in the regulation process via protein binding may have more general structural roles, in addition to their coding function.

Product information

With the preceding figures, we can try to estimate the corresponding information theoretic quantities for selection of nucleotide and amino acid sequences. Let us first consider sequence information related to products, that is, polypeptides representing genes.

Amino acid sequences

We begin with a simple and well-known situation. When four nucleotides occur with frequencies p_i, $i = 1, 2, 3, 4$, the average information carried by one of them is

$$I_{\text{nuc}} = -\sum_i p_i \log p_i, \tag{3}$$

and we have $I_{\text{nuc}} = 2$ bits when all the frequencies are equal, $p_i = 1/4$. According to (2), the information content of a sequence of a gene S consisting of 900 base pairs then is estimated by

$$I_{\text{S}} \leq 900 I_{\text{nuc}} = 1{,}800 \text{ bits}. \tag{4}$$

For this estimate, it is assumed that at each position, a nucleotide is selected independently of the other positions. That means that (4) yields the value of the information in the absence of sequence correlations. When sequence correlations are taken into account, the value gets smaller, as will be discussed in more detail below.

Similarly, the 20 amino acids occur with frequencies $p_a, a = 1, \ldots, 20$, and so the average information carried by one of them is

$$I_{aa} = -\sum_a p_a \log p_a. \tag{5}$$

When all amino acids occur with equal frequency $p_a = 1/20$, each of them contributes

$$I_{aa} = \log 20 \approx 4.32. \tag{6}$$

The information of a polypeptide of length 300 then is bounded from above by the value for a sequence of 300 uncorrelated amino acids, that is, by the value for a random sequence,

$$I_{pp} \leq 300 \times 4.32 \approx 1{,}300 \text{ bits}. \tag{7}$$

These estimates, however, have fundamental deficiencies that we now wish to address. In one sense, they underestimate the coding information. We stated above that the average length of a polypeptide is about 300 amino acids, but in fact this length typically ranges from about 50 to 2,500. Thus, we should also take this possible length range into account instead of just estimating the information for one single length. In another and perhaps more important sense, however, the value in (7) constitutes a gross

overestimate. The point is that not every combinatorially possible amino acid sequence can become a biochemically feasible polypeptide, that is, lead to a folded protein. So, a more relevant estimate should start with the number of such biochemically feasible polypeptides. This, however, encounters certain difficulties, given the present state of biochemical knowledge.

In fact, it is not clear what determines whether a couple of amino acid sequences can fold into a protein. Some researchers think that this is only a question intrinsic to the sequence; it may depend on the possible configurations that the sequence can attain in space according to the physical attractions between different sites and their corresponding free energies. Thus, according to this approach, it is only required that the configuration of minimal free energy leads to a well-folded protein. The difficulties then arise from computing that minimal free energy configuration which is an unsolved bioinformatics problem. Other researchers maintain that for the correct folding of proteins, particular other helper proteins are needed, the so-called chaperons [21]. Thus, whether an amino acid sequence folds properly depends on the cellular environment it finds itself in. In particular, this would imply that the protein-folding problem cannot be solved by a computational approach on the basis of the sequence information alone. A striking example for this effect is the prion proteins that represent an alternative folding of an amino acid sequence with physiological properties that are drastically different (infectivity) from those of the normal folding structure (review in [2]). Actually, the presently most successful algorithm for predicting folding patterns of biological amino acid sequences, that is, ones derived from genes as opposed to coming from arbitrary combinatorial arrangements, circumvents this problem. Given the sequence, this algorithm simply searches data banks for the most similar sequence whose folding pattern is already known and uses that information for the prediction of the folding structure of the sequence in question. This approach leads to a drastic reduction of the estimated number of amino acid sequences to be taken into account as a solid basis of an information theoretic analysis.

Ensemble and sequence entropies

Mathematically, this issue can be formulated more abstractly. There exist two approaches for calculating the information content in ensembles of products, via ensemble or (when they are linearly arranged as sequences) via sequence entropy. We begin with the concept of *ensemble entropy*: we consider an ensemble of N items belonging to M types x with relative frequencies $p(x)$. For the present consideration, these items are biochemically feasible polypeptides, or when we look at a more special situation, perhaps the ensemble contains only those polypeptides that are present in a particular organism or cell, and the frequencies of types represent how often these polypeptides occur in the specified ensemble. The entropy of such an ensemble of polypeptides then is

$$I = -\sum_x p(x) \log p(x). \tag{8}$$

When all the $p(x)$ are equal, then

$$I = \log M. \tag{9}$$

This is the maximal value the entropy can attain for an ensemble of M types. Biochemical knowledge about our ensemble leads to different individual relative frequencies and reduces the entropy. In addition, considerations about physical aspects can decrease the entropy. For instance, a physicist might consider the minimal free energy of the folding of a protein as an example of a so-called Hamiltonian H, to obtain a so-called Boltzmann–Gibbs distribution with probabilities $p(x) = \frac{1}{Z} e^{-\beta H(x)}$, where the partition function Z is simply a normalization factor and β is a constant, the so-called inverse temperature. We then insert these probabilities into (8). When the Hamiltonian H is non-constant and β is non-zero, the entropy (8) is smaller than the upper bound (9). Thus, biochemical or physical insights can be used to come up with more realistic ensembles of polypeptides and lead to reduced estimates for the entropy of the ensemble.

There is an alternative approach for determining the entropy of our ensemble of polypeptides or other biochemical sequences, namely via its *sequence entropy*: we consider sequences S of length n of "symbols" *as* drawn from an "alphabet" A of size $|A|$, occurring with relative frequencies p_a. In our present application, where the sequences represent polypeptides, the alphabet consists of the 20 amino acids and a symbol is a particular such amino acid. Each position in the sequence has entropy $-\sum_a p_a \log p_a$. When there are no correlations between the positions in the sequences, the entropy of a sequence is

$$I_S = -n \sum_a p_a \log p_a. \tag{10}$$

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

$$I_S = n \log |A|. \tag{11}$$

Since there are $M := |A|^n$ such different sequences, this is the same as the ensemble entropy $\log M$ in (9) above. Again, there are refinements through additional knowledge, such as

- unequal distribution of the p_a, or
- sequence correlations, that is, correlations between the occurrences of symbols at different positions

that decrease the entropy. The concept of sequence entropy is designed to incorporate these aspects. The first one is rather simple. Concerning sequence correlations, the simplest ones are those between neighboring positions. To capture them, we consider the so-called block entropies

$$-\sum_{v} p_v \log p_v,\ v = \text{block of length } l. \tag{12}$$

Here, such a block consists of l adjacent positions. Thus, the block entropy is not based on the frequencies of individual symbols, but on the frequencies of blocks of l adjacent symbols. When there are correlations, these frequencies are different from the product of the frequencies of the individual symbols constituting such a block. The evaluation of these block entropies is only computationally feasible for relatively small values of l. For amino acid sequences, we have $|A| = 20$. It turns out that going beyond $l = 5$ (pentapeptides) or $l = 6$ (hexapeptides) yields little additional information and may rather obscure the patterns. On the other hand, the block entropies do not capture long range correlations because by definition they are restricted to correlations within blocks of the given (small) size. Examples are the complementarities between the 5′ and 3′ ends of certain messenger RNAs and the folding pattern of tRNA and rRNA. In general, long range correlations are computationally difficult to find without biochemical insights guiding the scheme.

In any case, in the ideal situation where all regularities are taken into account, the sequence entropy coincides with the ensemble entropy. In general, we should consider these two quantities as alternative methods to use biochemical restrictions on ensembles of polypeptides to estimate our uncertainty about individual members of the ensembles.

Regulatory information

Concerning gene expression at large, we have to deal with two types of information embodied in sequences:

1. the one related to product information, concerning arrangements of triplets within the sequence coding for a polypeptide, as outlined above, and the production of ensembles of products;
2. the regulative information embodied in the sequential arrangement of oligomotifs where regulative factors operate that has to be developed now.

Note that (2) applies also to the genon in the coding sequence and mRNA, to the extent that such oligomotifs are superposed onto the string of triplets and allow, due to the wobble in triplet usage for a given amino acid, to differentially regulate, at mRNA level, a given gene or polypeptide. Replacing the random possible amino acid sequences by some empirically derived rules of occurring polypeptide structure, the numbers of different polypeptides to enter our calculation can be reduced. As discussed, this allows us to apply ensemble entropies rather than sequence entropies to estimate the information content of an ensemble of polypeptides. However, for the arrangement of regulatory oligomotifs, no empirical rules are presently available, and therefore we have to resort to sequence entropies as refined by block entropies to take local correlations into account, once the relevant bioinformatics data become available.

Regulation is not only concerned with individual genes, but mainly with the coordinated activation of specific sets of genes, according to the requirements of the cell and depending on cell external conditions. In this section, we provide the first steps toward a quantification of this aspect of regulation.

Our analysis starts with the sequence information in cis relevant to the regulation of the gene expression process. This is realized in a (pre-)mRNA by the interaction of cis and trans, the cis (pre-)genon selecting, via oligomotifs and suitable protein factors.

The interaction between oligomotifs and proteins

In the frame of our analysis, the genome is the starting point of gene expression, and the genes with their genons are the final point, since we have decided to exclude for the time being all posttranslational regulation from our analysis [30, 31]. The gene–genon relations, as realized in specific mRNAs, yield the simplest model to apply information theory to regulation of gene expression. The insight gained may allow us to go upstream to the pregenons and eventually to the protogenons realized in the domains constituting the genome, that is, the programs at pre-mRNA or DNA level, resp., controlling the expression of a gene. We will therefore analyze the relations realized in the messenger RNA–protein complexes (mRNPs) of regulative significance. It should be recalled that, at this last pre-translational level, differential regulation of gene expression in the general case is negative, as will become important in Sect. 6. Specific mRNAs are repressed in mRNPs and must be released for translation. This was shown experimentally and follows from the fact that translation factors act non-discriminatively on many mRNAs and are plethoric, whereas repressed mRNPs have specific combinations of proteins which, in the general case, are not abundant in the cellular milieu [36].

As already discussed, regulation of gene expression by proteins implies the presence of arrays of binding sites (oligomotifs) in a nucleic acid; experimental data as well as theoretical considerations indicate that such sites are discrete and specific to some peptide motifs. This is in contrast to RNAi where, a priori, any nucleotide sequence, even in overlapping fragments, may theoretically act as a controlling element from the transgenon. Too little is known about RNAi at this moment. In contrast, the search for a code governing the limited ensembles of regulating proteins and oligomotifs seems amenable to information theoretic analysis (cf. recent attempts at defining a "splicing code" [5]).

A classical example of such proteins regulating specific mRNAs is the iron response element (IRE) in ferritin and transferrin receptor mRNA (see reviews in [34, 35]). The specific IRE oligomotif sequence is found in the 5′-UTR of the ferritin mRNA and in the 3′-UTR of transferrin receptor mRNA. It binds a specific 90 kD regulatory protein which represses ferritin mRNA translation but stabilizes transferrin receptor mRNA; this system regulates uptake and intracellular storage of iron. There are other cases documented (see references in [7, 34, 35]). As explained, specific oligomotifs in the mRNA provide binding sites for specific proteins that bind to the mRNA to form RNP complexes (or, alternatively, binds si- or miRNAs for final or temporary silencing). For the coarsest estimate again, we consider the collection of different mRNA binding proteins of perhaps 3,000 $\approx 2^{11.5}$ different types in the cytoplasm of a cell, assumed to be equally frequent in a first analysis. We also assume that there is one-to-one correspondence between mRNA oligomotifs and mRNA-binding proteins. In other words, each such oligomotif provides a binding site for precisely one such protein. Again, we hope that this simplification will best bring out the essential principle. Selecting one type of protein, or equivalently, of an oligomotif out of those, therefore contributes an information of about 11–12 bits. An mRNA has several different such oligomotifs; for instance, for globin mRNA we estimate about 20 sites/600 nucleotides. Since about 50 % of the mRNA is covered by proteins [36], about 300 nucleotides are available for carrying oligomotifs. When each site thus consists of about 15 nucleotides (the actual number may vary between 7 and 50 nucleotides), there are 4^{15} possibilities for the composition of oligomotifs. Since, however, there are only about 3,000 different binding proteins, it suffices to have 3,000 different oligomotifs when we neglect the effects resulting from varying binding affinities. Therefore, there are $\binom{3000}{20}$ possibilities for distributing the oligomotifs among the mRNAs in question, and since there are up to 10,000 different such mRNAs in a given cell (in fact, 1,600 in the globin example in Table 1), it is possible to distribute the oligomotifs in such a manner that each such mRNA can be uniquely identified from its collection of oligomotifs. On the other hand, in order to be able to select for release from repression specific sets of mRNAs, say 200 out of 1,600, via their oligomotifs, these 200 should share certain combinations of oligomotifs that distinguish them from the other mRNAs.

When the different protein binding sites operate independently, and when we assume that a single such protein out of 3,000 candidates can bind to each oligomotif, we get a total contribution of about $12 \times 20 = 240$ bits.[2] Again, this upper bound is not biochemically realistic as we should take constraints on the binding of proteins to sites into account. When we exercise such constraints, the upper bound is decreased. On the other hand, one and the same protein can also have different functions, and one and the same oligomotif may bind to different proteins, and such effects increase our figures.

The above-mentioned thus furnishes a flexible mechanism for the simultaneous action of several factors or, better, for specific factor combinations.

In any case, the preceding figure of 240 bits only concerns a single genon. In a similar manner, as outlined below, we can estimate the contribution of the pregenon at the pre-mRNA and the protogenon at the DNA level through the selection of RNA or DNA binding regulatory factors. The number of oligomotifs possible at each step may be considered to be of the same order, about 3,000. Since the protogenon and the pregenon are carried by different types of nucleic acids, and since the pregenon and the genon are operative in different cellular compartments and/or environments, the interacting protein factors differ between them. In particular, even though most of the oligomotifs of the genon are already present at the DNA or preRNA level, they become operative only at the genon level as only there the corresponding factors are thought to be present. Thus, when considering combinatorial regulation at protogenon, pregenon and genon level together, we have to reckon with about 9,000–10,000 oligomotifs, assuming for simplicity that the oligomotifs at the different levels are all different. This means that about 5 % of all genes in a cell might be involved in this combinatorial regulation scheme, a genetic load that seems reasonable in terms of molecular genetics.

Another aspect is the following. The mRNA molecules are not just linear sequences, but form a spatial structure by internal complementarity of sequence motifs and

[2] In certain cases, also more than one protein having the same amino acid motif for binding at an oligomotif, but different overall sequence, may bind to each site. Furthermore, to a protein bound to an oligomotif, other proteins can bind in turn by protein–protein interaction. When we thus consider the theoretical possibility that any combination of proteins could potentially bind (directly or indirectly) to any site, we get a much larger value.

hydrophobic interactions, controlling interaction via bonds between different parts. In the absence of proteins, they would form so-called secondary structures that are determined by competing complementarity relations between bases. In fact, besides the sequence identity of the oligomotifs, also their position in the secondary structure is relevant for the attachment of complementary proteins, because that position may influence their accessibility to such proteins. Generally, in an mRNP complex, in steady state, proteins are bound to at least 50 % of an mRNA (cf. [36]). The spatial structure is then determined by binding between those attached proteins rather than by RNA–RNA interactions; the secondary RNA structure that is determined by binding between complementary bases gets largely overridden by protein binding, forming the tertiary structure of the complex (however, the complementarity of 5′ and 3′ UTR sequences in, e.g., globin mRNA [26], indicates the importance of direct interaction within an mRNA). As the proteins are attached to oligomotifs in the mRNA, conceptually we thus obtain relations between oligomotifs, whether their attached protein complexes bind to each other or not. It is important to realize that this spatial interaction will thus depend on the sequential arrangement of the oligomotifs. Therefore, not only the biochemical identity of these oligomotifs that determines which proteins can bind to them is relevant, but also their order in the mRNA and their mutual distances in the mRNA sequence. Since this relational structure between oligomotifs is determined by the binding of proteins to oligomotifs and the sequential order of those oligomotifs, it cannot contain more information than the combinatorics of that binding and that sequential order. Again, in the simplest case, we consider only independent pairwise bindings and assume that this is a binary relationship, that is, the binding strength plays no role. We then have as many possibilities as we can form disjoint pairs between 20 oligomotifs in a globin size mRNA.

In the pregenon, as the resulting structure is relevant for the processing of pre-mRNA, this also contributes processing information arising from the binding of regulatory proteins to oligomotifs. An average primary pre-mRNA may contain about 10,000 nucleotides and hence about 300 oligomotifs. By protein–protein interaction, they form large complexes whose 3D organization is relevant for processing in time and space, but this issue is not explored here. However, it must be pointed out that there are steric restrictions to binary protein interaction in such huge complexes and certain rules of protein–protein binding have to be taken into account.

Also, interactions with other RNAs, cellular proteins or higher order structures like the cytoskeleton and the nuclear matrix [28] contribute to regulatory information. In particular, the primary transcripts are the organizing skeleton of the RNA-dependent nuclear matrix, the network which conditions the dynamic nuclear architecture and regulation of RNA processing and transport in time and space [25].

In any case, this processing information is not only concerned with a single gene, but with many of them simultaneously. For instance, 3,000 such (pre-)mRNA binding proteins might be involved in the regulation of 10,000 transcripts in a given cell. Therefore, from formal reasoning already, one should expect correlations both between the activities of different RNP-proteins and the affinities of binding sites in different mRNAs. In Sect. 5.3, we shall turn to this aspect of coregulation and analyze how the selection of specific sets of simultaneously active genes is quantifiable within our information theoretical framework, based on the genon. First, however, we shall now discuss the formal framework for the corresponding code.

A combinatorial regulatory code

A very important aspect is the complementarity between regulated elements, the mRNAs with their oligomotifs and the regulators, the protein factors that selectively address those oligomotifs. This is characteristic of eukaryotic gene regulation and distinguishes it from prokaryotic regulation schemes and is due essentially to the absence of operon-type sequence arrays of genes and the resulting need for combinatorial pleiotropy of regulative factors.

This requires a new type of code because it represents an emergent organization of regulation that operates at a higher level and is distinct from the genetic code. Therefore, for its operation, it needs its own specific rules. In order to keep the load for the genome under a reasonable level, these rules need to be of a combinatorial nature.

In formal terms, we have a new code whose letters now are regulatory proteins, or formally equivalently the corresponding oligomotifs (instead of the letters of the genetic code which were nucleotides, and where we might consider the triplets as words). We should point out that for our formal considerations here, it is not relevant that the oligomotifs in turn are composed of nucleotides. For the present purposes, we consider the oligomotifs as the basic irreducible units, that is, the letters of our code. As discussed in Sect. 5.1, altogether, there might be up to 10,000 such letters, about 3,000 each at DNA, pre-mRNA and mRNA level.

In our simplified model that was built around the example of globin (pre-)mRNAs (see Sect. 5.1), words in this code may consist of 20 letters within an mRNA of 600 nt. A typical mRNA might be about 1,000 nucleotides long and contain about 30 such oligomotifs, which then would be the word length. Thus, we can identify a word in our code with a specific type of (pre-)mRNA, assuming

that two biochemically different mRNAs also carry different oligomotif collections. For simplicity, at this point, we do not consider the order of the letters in a word, that is, consider words as equivalent that only differ by the ordering of their letters, and we also assume that all the letters in a word are different from each other. The first aspect will be addressed in more detail in a future paper, whereas the second one would simply make our combinatorial scheme somewhat more complicated without affecting its essence.

The mRNA includes or, in a somewhat metaphorical language, chooses letters (oligomotifs) to be affected by external regulators. Changing a letter has a small, local effect, as it only affects the cell states that the particular gene carried by that mRNA participates in. Therefore, mutations that change such an oligomotif into a different one have only a limited, local effect. Conversely, specific factors seek out such letters to affect specific groups of mRNAs. In our formalism, this means that all the words containing the selected letter are simultaneously identified. Since any such choice will affect many mRNAs simultaneously, here the resulting effect will be large and global, as will be analyzed in more detail in Sect. 5.3 Thus, the system can flexibly switch between different cell states. These regulating factors, however, should be largely protected from mutations in the course of evolution as the consequences might be drastic. A way around this is duplication of the genes coding for such global regulators. After such duplication, one of the copies is free to mutate and to explore new evolutionary possibilities as long as the other copy is kept fixed and continues to fulfill the original regulatory function; see e.g., [4, 23, 37, 38]. In any case, the combinatorial scheme reverses the roles of regulators and regulated elements. Whereas in prokaryotic schemes, control is directly exerted by the regulators and changes can be implemented by their modifications, here the regulators have to stay rigid and small changes in control arise from modifications of the regulated elements.

In a certain sense, the oligomotif code is analogous to the triplet code. Here, we propose that an mRNA carries oligomotifs each of which specifies one out of a set of regulatory proteins. Likewise, the mRNA contains coding triplets and, according to the triplet code, each such triplet specifies one particular tRNA (which in turn translates the triplet into an amino acid). An oligomotif recognizes a particular amino acid sequence in a polypeptide. A triplet recognizes its complementary triplet in the tRNA.

Similar principles should apply to the combinatorics of regulation at the proto- and pregenon level, although of course the specific molecules and motifs involved can be different. A few remarks should suffice. For instance, we have the splicing process at the pre-mRNA level. Introns are excised and exons are combined according to specific rules that control the decision between alternative possibilities. Biochemically, this again depends on the interactions of oligomotifs with specific proteins. Analysis of pre-mRNA-protein complexes showed that the ensembles of proteins are largely different from those found on cytoplasmic repressed mRNAs (see discussion in [36]). The proteins involved here are typically different from the mRNA-binding proteins, even though some of the oligomotifs could be the same at the pre-mRNA and mRNA level. Introns also carry certain oligomotifs, and the combinatorics of protein binding at these oligomotifs, as well as possibly at oligomotifs in neighboring exons, might decide about their excision. Introns can carry a single oligomotifs up to several hundreds of vastly different sizes (from a few bases to tens of thousands).

The regulatory roles of the various transcription factors at the DNA level have been much studied (recent review in [8]). It is at present not entirely clear, however, which of them are only active at and restricted to the DNA level and are not contained in the transcripts, and which may bind to mRNA as well and play decisive roles at the primary transcript processing stage. Furthermore, some of the functions of giant transcript might be the transmission of bound factors to egg or daughter cells, in meiosis and mitosis when the nucleus is disassembled and transcription stops [29].

The combinatorics of the selection of specific combinations of active genes

We now turn to the combinatorics of this regulation scheme. An example: as shown in Table 2, there are three globin mRNAs that contribute 90 % of proteins made in our red blood cell model, whereas about 200 mRNAs are in an active translation state (of course, not all cellular states have 200 mRNAs, but in the sequel we shall use that number to develop our estimates). In turn, 200 mRNAs are selected for translation among the about 1,600 present altogether in the cytoplasm of an erythroblast. However, it is not necessary to select completely arbitrary subsets of 200 out of 1,600 mRNAs. The corresponding number $\binom{1600}{200}$ would vastly exceed the combinatorial possibilities available for differential regulation. Rather, specific subsets are selected according to physiological state. During terminal differentiation of a red blood cell, roughly, there are possibly about ten different sequential physiological states of the cell that need to be distinguished. In addition, these states need to be stabilized across different external and internal conditions and it also requires the selective translation of specific sets of mRNAs. Clearly, it requires variations of transgenons.

This selection of 200 among 1,600 different mRNAs might work by affecting about 3 of the 20 oligomotifs present in a given mRNA, insofar as the binding proteins which act as repressors can be removed. This is caused, possibly, by external factors which may lead to enzymatic modification of the corresponding RNP proteins (phosphorylation, acetylation, glycosylation, etc.) and hence chemical modifications or alternatively by structure modifications due to allosteric factors, followed possibly by proteolysis.

Considering that all the mRNAs in a given cell looks combinatorially as follows, assuming 3,000 different oligomotifs, as argued above, there are $\binom{3{,}000}{3 \approx 4.5 \times 10^9}$ different sets of three oligomotifs, henceforth called a 3-tuple (rather than a triple, in order not to be confused with a triplet, that is, three nucleotides coding for an amino acid). Thus, by choosing a specific such 3-tuple, that many conditions can be distinguished. However, when we assume that each mRNA has 20 oligomotifs, it contains $\binom{20}{3=1{,}140}$ different 3-tuples and therefore it can participate in that many conditions. Thus, if one particular mRNA codes for such an essential gene that it needs to participate in every condition, then under that constraint only 1,140 different conditions can be realized.

In order that several mRNAs participate in the same condition, according to our model, we assume that they need to share at least three oligomotifs. More generally, when some mRNAs share m oligomotifs ($3 \leq m \leq 20$), they can simultaneously participate in $\binom{m}{3}$ conditions, and this number varies between 1 (for $m = 3$) and 1,140 (for $m = 20$). However, when $m = 20$, these mRNAs can no longer be distinguished through their oligomotifs and therefore m should be smaller than that, unless those mRNAs should always be simultaneously translated. The information related to this aspect would be quantified in terms of the number of combinatorially possible conditions (which still needs to be mathematically determined) and their probabilities. Let us consider some numerical examples. The general scheme is the following. Let K oligomotifs ("oligos" for short) be given. Then there are $\binom{K}{20}$ different possibilities to choose 20 among them, that is, we can distinguish that many mRNAs through their different endowments with 20 out of these K oligos (of course, there could be many different mRNAs equipped with the same set of oligos, which would simply mean that all of them are always simultaneously translated; for simplicity, we do not take that into account here). A condition for translation is specified by selecting three out of these K oligos for removal (of course, one could also select more and then have more mRNAs translated; for simplicity, we only analyze the basic situation of the removal of three different RNA binding proteins). Thus, for every choice of $\binom{K}{3}$, we have a different condition. The number of mRNAs participating in such a condition is the number of mRNAs that carry all those three oligos. Thus, 3 out of their 20 oligos are fixed, and 17 remain for free choice. That is, we have $\binom{K-3}{17}$ different possibilities. Thus, assuming that all the above $\binom{K}{20}$ possibilities are realized, by selecting three oligos we select $\binom{K-3}{17}$ different mRNAs. Here are some simple numerical examples.

- Distribute 21 oligos among 21 mRNAs (20 oligos/mRNA) so that each mRNA is identified by which oligo it does not contain. By specifying three oligos, any of the possible $\binom{21}{3} = \binom{21}{18} = 1330$ combinations of 18 mRNAS can then be selected. Here, we have only relatively few different mRNAs. Of course, one can take these as the core and amplify the scheme by adding other mRNAs that share some of their oligomotifs with this core, but not all of them. One could also consider other cores with completely disjoint collections of oligomotifs, in situations where non-overlapping combinations of mRNAs are to be selected according to physiological circumstances and cellular requirements.
- Distribute 23 oligos among $\binom{23}{3} = 1771$ mRNAs (20 oligos/mRNA) so that each mRNA is identified by which three oligos it does not contain. By specifying three oligos, any of the possible $\binom{23}{3} = \binom{23}{20} = 1771$ combinations of $\binom{20}{3} = 1140$ mRNAS can be selected. Here, the collection of selected mRNAs is perhaps somewhat large.
- Distribute 22 oligos among $\binom{22}{2} = 231$ mRNAs (20 oligos/mRNA) so that each mRNA is identified by which two oligos it does not contain. By specifying three oligos, any of the possible $\binom{22}{3} = \binom{22}{19} = 1440$ combinations of $\binom{19}{2} = 171$ mRNAS can be selected. Here, we select sets of mRNAs of reasonable size.

In any case, the preceding numerical examples indicate that the number of RNA-binding proteins, about 3,000, is far larger than what is required by the combinatorial scheme described. Of course, the scheme can be refined, and that might need some more proteins. We also observe that when an oligo is contained in a coding sequence, then, when a collection of mRNAs containing that specific oligo is selected for translation, all these mRNAs contain the same piece of coding sequence and hence produce polypeptides that share some short amino acid sequence. Thus, the scheme can ensure the selection of combinations of related polypeptides. More generally, the combinatorics of RNA–protein interactions could also implement some

hierarchical ordering of sets of related genes. A first oligo identifies a general group, and the second and third one then determine more specific subgroups.

The genons provide the operation sites for ensembles of transgenons, selected out of local holo-transgenons, to select specific sets of genes. These sets are determined by the combinatorics of the oligomotifs, that is, by the genons, as demonstrated in the above example. Which set is selected then is decided by the operation of transgenons.

The coordinated activation of specific genes is needed both to achieve a particular state of the cell and to maintain and stabilize that state under fluctuating external conditions. Thus, part of the information needed here is about the identity of those different states, that is, which particular state is selected from the available ensemble of states. That part of the information can therefore be quantified in terms of that ensemble. Another part of the information is utilized to stabilize the cell state against external fluctuations. According to Ashby's law of requisite variety [3], the internal variety of the system needs to be at least as large as the entropy of the external perturbations, ito compensate for them internally. That part of the information therefore is not visible in the collection of activated genes, but rather used up for that compensation of fluctuations in external conditions. For instance, the relative concentration of factors comprising the transgenon could play a role here.

The combinatorial model discussed here was built around globin mRNAs. In particular, for other mRNAs, according to their overall length, we might have more or $<$20 oligomotifs that constitute binding sites for regulatory proteins. The essential principle of our model, however, remains unaffected by this variation of the actual numbers.

In addition to the combinatorics of the selection of specific sets of genes, the binding between RNAs and proteins may also have some role for protein–protein interactions. An RNA could function as a scaffold that brings specific proteins that attach to its oligomotifs together so that they find each other and can interact within a higher order complex.

The information contained in the sequential arrangement of oligomotifs present is reduced by processing in the course of gene expression. This information needs to be complemented by decision-making information from outside the mRNA in question. This refers to coregulation and here finer distinctions are needed. The essential decision-making process is hence based on the combination of trans-acting factors present in a given cell compartment, a cell, a cell lineage or organism, from which a specific (pre-)genon picks up specific factors. The calculation of the corresponding information present is an arduous undertaking, since the factors involved are contributed by various sources: the genome of the specific cell or else the genome at large to the extent that various cells may contribute external factors to a given cell, as cytokines, etc.; furthermore, there are factors picked up from the outside of the organism in a given environment. A major ensemble of factors involved, but neglected here, are regulatory RNAs acting within the frame of RNAi, which, again, may be produced inside a cell compartment—nucleus or cytoplasm—or be channelled into a cell from the physiological environment. Here, however, we have confined our considerations to regulative factors of protein nature contributed by the genome. We have neglected factors from the exo-system, like organic and inorganic compounds as vitamins and hormones on one hand, and metals and other inorganic ions on the other, which act, in general, indirectly via interaction with regulatory proteins or enzymes present in a cell. Furthermore, there are physical factors as temperature, the presence or absence of light, etc.

We must stress, however, that the main information for decision making in gene expression must be given by the availability, the activity and the concentration of regulative factors in a given cell, at a given time and in a specific physiological situation.

Pathways of gene expression

In this section, we shall analyze and estimate the regulative information gained on a single gene expression pathway independently of the rest of the cell, and in particular without considering its interactions with other expression pathways. The network aspects then remain to be worked out in future work.

In order to put the sequel into perspective, we start with a general consideration. As a numerical example, when we assume m steps of regulation and when at each step we choose 1 out of $100 \approx 2^7$ possibilities, we gain information of $7m$ bits. For example, $m = 10$ would yield 70 bits. Both these figures, that is, the number of steps and the selectivity at each step, are within a biochemically realistic range. Thus, this figure provides a rough estimate against which we can now compare the figure obtained from more detailed biochemical knowledge. In contrast to the previous sections, we are in a position to provide a more or less realistic estimate, and we can then assess the discrepancy between the rough bound and the precise estimate.

In fact, there are two conceptually different schemes of accounting for regulative information, and the comparison of the different results may lead to insight into the functional role and effect of gene regulation. On one hand, we can investigate when, where and between how many alternatives decisions are taken, like whether or not to process further some transcript, or choosing between different possible splicing. We would then simply add the bits from the individual decisions that are taken during the

cascade of gene regulation. On the other hand, we can look at the final effects, that is, whether from a genomic domain a specific product, or which one of several products, is produced in the end. The amount of information computed that way will generally be smaller than the one from the first scheme, simply because several different decisions at different stages may in the end nevertheless lead to the same end result. Below, we shall discuss that difference in terms of Ashby's law of requisite variety, but for the moment we proceed to the accounting.

We refer to the steps in the cascade of regulation as outlined above in Chapter 3.3. We shall estimate the values of the probabilities to be inserted into our formula (3.3), to obtain a specific numerical value for the information provided or gained during the regulation of a single pathway. Our first accounting scheme evaluates the principal steps of decision, considering repression as the default state (note that these estimations concern *one* single genomic domain)

0. Probability $p_{00} = 1/5$ for chromatin remaining inactive; $p_{01} = 4/5$ selected to be activated. This decision yields $-(1/5\log 1/5 + 4/5\log 4/5) \approx 0.72$ bits according to formula (3.3).
1. Probabilities $p_{10} = 2/5$ that a genomic domain is not transcribed and $q_{11} = 3/5$ that it is transcribed.[3] This decision yields $-(2/5\log 2/5 + 3/5\log 3/5) \approx 0.97$ bits according to formula 3.3.
2. Probabilities $p_{20} = 1/3$ that a transcribed genomic domain is not processed into a final product during the life of the cell and $p_{21} = 2/3$ that it is. Yields ≈ 0.92 bits.
3. In fact, this can be further refined by breaking down the life cycle of a cell into specific instances and, taking into account that even if intermediate processing steps are carried out, it may not necessarily lead to a final product. We assume that when such a transcribed domain is to be processed at a given instance, it is not carried out with probability $p_{30} = 9/10$, while processing does take place at that given instance has $p_{31} = 1/10$. Therefore, at each instance the decision yields ≈ 0.47 bits. The question then is how many independent such instances we have. When their number is k, the total amount then is $\approx 0.47k$. We need to compare this with 2., i.e., relate the 9/10 of the present step to the 1/3 of 2. Since $(9/10)^{10.43} \approx 1/3$ (that is, after carrying out a bit more than ten steps with probability 9/10 each, we end up with a probability of 1/3), we might assume that we have about $k = 10$ such instances, or, with some more numerical precision, $10.43 \times 0.47 \approx 4.9$ bits, in place of the 0.97 bits of 2.
4. Splicing into one of five possible mRNAs. In the absence of more detailed information, the probabilities for these mRNAs are assumed to be equal: $p_{41} = \ldots = p_{45} = 1/5$. This yields $-\log(1/5) \approx 2.32$ bits.
5. Probability $p_{50} = 4/5$ that these mRNAs then are not translated at a given instance (cumulative for all five mRNAs), and $p_{51} = \ldots = p_{55} = 1/5$ for a genomic domain to lead to a specific RNA that then is translated into a polypeptide. The decision for each such mRNA then accounts for ≈ 0.72 bits, that is, altogether $\approx 5 \times 0.72 = 3.6$ bits

Taking the results of 0., 1., 3., 4., 5. together then yields for each geonomic domain

$$I_{\text{decision}} \approx 12.5 \text{ bits.} \tag{13}$$

This number then quantifies the amount of decision taken during the cascade of regulation in terms of the probabilities of the outcomes of the individual decision steps for a single domain possibly producing some final product. One reason that this figure is relatively small is that the probabilities at the different steps are not equal. For example, at a given instance, it is much more likely that a transcript is not processed than that it is. In formal terms, the choice between two alternatives with unequal probabilities yields <1 bit of information, and in particular, when one of these probabilities is very small, the information also becomes very small, as explained in Sect. 2.

Returning to the analysis on which 13 is based, we should also consider more detailed biochemical distinctions, for example at DNA level methylation, and furthermore, at chromatin level, histone acetylation, phosphorylation, etc., or their absence, but for the moment we keep the above figure of ≈ 12.5, though this may severly underestimate the complexity of the decision process. In contrast to the preceding, we shall now analyze these steps in terms of the effects on the final result, that is, whether a genomic domain leads to some functional product or not and more precisely to which of several possible ones it might lead.

0. Probability $q_{00} = 1/5$ for chromatin remaining inactive and $q_{01} = 4/5$ for it being activated eventually.
1. Probabilities $q_{10} = 2/5 \times 4/5 = 8/25$ that a genomic domain is not transcribed and $q_{11} = 3/5 \times 4/5 = 12/25$ that it is transcribed from the activated chromatin.
2. Probabilities $q_{20} = 1/3 \times 3/5 \times 4/5 = 4/25$ that a genomic domain is transcribed, but the transcript not processed to mRNA during the life of the cell and $q_{21} = 2/3 \times 3/5 \times 4/5 = 8/25$ that the transcript is also processed. In this step, 2 out of 3 of the transcripts are processed. The corresponding probability 2/3 is then multiplied with the previous probability 3/5 for transcription of a domain, to obtain the probability that

[3] Recall that in step 1, 3 out of 5 domains were transcribed.

a domain is not only transcribed, but that the transcript is also processed.

3. Probabilities $q_{30} = 9/10 \times 2/3 \times 3/5 \times 4/5 = 36/125$ that this processing does not take place at a given instance and $q_{31} = 1/10 \times 2/3 \times 3/5 \times 4/5 = 4/125$ that it does.
4. Probabilities $q_{41} = \ldots = q_{45} = 1/5 \times 1/10 \times 2/3 \times 3/5 \times 4/5 = 4/625$ for a genomic domain after transcription and processing leading to one of five possible mRNAs through alternative splicing (with the probabilities for these mRNAs assumed to be equal, as before).
5. Probabilities $q_{50} = 4/5 \times 1/10 \times 2/3 \times 3/5 \times 4/5 = 16/625$ that these mRNAs then are not translated (cumulative for all five mRNAs), $q_{51} = \ldots = q_{55} = 1/5 \times 1/5 \times 1/10 \times 2/3 \times 3/5 \times 4/5 = 4/3125$ for a genomic domain to lead to a specific RNA that then is translated into a polypeptide.

Thus, according to formula (3.3), in terms of its selection between different possible results, we can quantify the information provided by the genon as

$$\begin{aligned} I = -\Bigg(& q_{00}\log q_{00} + q_{10}\log q_{10} + q_{20}\log q_{20} + q_{30}\log q_{30} \\ & + \sum_{j=1}^{5} q_{4j}\log q_{4j} + 5q_{50}\log q_{50} + \sum_{j=1}^{5} q_{5j}\log q_{5j}\Bigg) \\ = & (1/5\log(5) + 8/25\log(25/8) + 4/25\log(25/4) \\ & + 36/125\log(125/36) \\ & + 4/125\log(125/4) + 5 \times 16/625\log(625/16) \\ & + 5 \times 4/3125\log(3125/4)) \approx 2.8 \end{aligned} \tag{14}$$

(not activated, not transcribed, not processed at all, not processed at given time, not translated or translated into a polypeptide from one out of five possible mRNAs that can be created by alternative splicing). Of course, this low figure arises because we are considering here a single genomic domain, whereas about 10,000 are transcribed at a given time in a given cell. In fact, what is essential is not the uncoordinated regulation of these genomic domains in parallel, but the coordinated regulation of specific subsets as we emphasize throughout this paper. This is because on one hand the choice of oligomotifs to constitute a pregenon is already determined by the primary DNA sequence, whereas, on the other, the essential regulatory decision is at the level of the composition of the transgenon, as will be detailed below.

From the perspective of information theory (Sect. 2), we have analyzed here the situation where the proto-, pre-, and the genon in cis, according to the various terms in (14), are considered as the receivers, and the events are specific transgenon factors that, for instance, induce or remove the processing inhibition. When, conversely, we consider the genon as the sender and the transgenon as the receiver, we are rather in the situation of Sect. 5.3, where we have discussed the coordinated regulation and expression of specific combinations of genes and the numbers given there apply.

We should point out already at this stage that the figure of Eqs. (13) or (14) does not capture the essential contribution of the genon. First of all, the combinatorics of oligomotif choice as described above (Sect. 5.3) underlying the regulation steps is based on more alternatives than the ones accounted for in the above steps, but many such decisions about the selection of 3-tuples of oligomotifs allowing for the unambiguous identification of an mRNA lead to the same result, and only those results enter into the computation presented here. More importantly, the essential question is not what is happening to a specific genomic domain, but rather which combinations of genes are simultaneously activated at a given time. For the latter, the required information is much larger, as described above in Sect. 5.3.

In any case, even if we confine our analysis to a single domain, this rather low figure of 2.8 bits does not contain the information about the number of identical polypeptides (or other functional products) derived from the genomic domain in question. This might be of the order of 2^{17} or 2^{18}, so that we should add 17 or 18 to the above quantity I to obtain a total value

$$I_{\text{genon}} \approx 20. \tag{15}$$

Analogously, we can increase the number given in 13 in that manner to obtain

$$I_{\text{decision}} \approx 29 \text{ bits}. \tag{16}$$

Again, the contribution of 17 or 18 from the polypeptide numbers seems to be an overestimate, insofar as the number of polypeptides produced need not be so precise; what is needed is an amount of polypeptides within a certain range, say within 10 % of some value. We recall, however, that in contrast (13) was an underestimate, so that it is not so clear whether (16) is too large or too small.

In any case, when we compare product and regulative information, we see that (15) and (16) are quite small in comparison with the sequence information (7) that refers to the possible products that a sequence could code for. In the absence of further specifications, that is, when we assume all amino acid sequences to be of the same length, and to be equally likely to occur, (7) then also gives the ensemble information for all such sequences of that length (assumed to be 300 in the above). Of course, when we look at the larger ensemble of amino acid sequences within the typical length range 50–2,000, the ensemble entropy becomes even larger. On the other hand, if we only consider the

ensemble of polypeptides present in the cell under consideration, we have perhaps 2^{16} different types, leading, when equidistributed, to an entropy of 16. We have a total number of perhaps 2^{30}, giving another contribution of 30 bits, that is, altogether 36 bits, which is of a magnitude comparable to (15) or (16).

Also, the preceding does not take the information about when and where some polypeptide is produced into account. For instance, in muscle cells, there may be about 500−1,000 different locations for the synthesis and assembly of specific proteins, contributing a spatial information of 9−10 bits. The relevant temporal resolution is perhaps more difficult to assess. This aspect is nevertheless very important. The timing within the lifespan of a cell or an organism, i.e., a physiological situation, is what matters for what products should be produced in which quantities. What is critical is not the overall regulation, but the regulative setting at a given time in a specific cell, possibly waiting for the next setting. This is, unfortunately, not reflected in our above figures.

Conclusion

We have seen that on one hand the cascade of regulation leads to a loss of information from a sequence at the DNA level to the functional product emerging from it. On the other hand, specific information is gained through the contributions of factors from trans that determine the different types of functional products produced from one and the same sequence at the DNA level and the numbers of these products. Thus, according to the preceding considerations and calculations, at the end of the cascade of regulation, less information can be extracted than at its beginning. This reduction of information is more drastic than the one caused by the degeneracy of the genetic code, that is, several triplets can code for the same amino acid. This can be quantified in two different ways:

1. The regulatory elements of the genon in cis as well as the trans elements derived from the DNA are also encoded as collections of sequences of base pairs (the oligomotifs), and we can evaluate the corresponding sequence information for a 900 nt long mRNA. This should be of a magnitude comparable to the product information of about 1,800 bits according to (4).
2. As discussed in 5, the genon in cis selects specific proteins that bind to the mRNA to form RNP complexes. We have estimated the total contribution as ≈ 250 bits for a given genon (noting, however, that the relevant mRNA binding proteins regulate many genes simultaneously). Likewise, the contribution of the pregenon at the RNA level, as well as the protogenon at the DNA level through the selection of DNA binding regulatory elements can be estimated. Altogether, this contributes the information for a given pathway.

Thus, we have a specific sequence of nucleotides (the oligomotifs) underlying the genon in cis and we have specific operations selecting binding proteins that lead to specific products selected out of the possibilities inherent in the genomic domain. In that order, the information encoded in the DNA decreases. For the genomic domain, however, we see an increase of information through the specification of its end product out of the various possibilities. This increase is much smaller than the above decrease.

In [31], we have interpreted this phenomenon in terms of Ashby's law of requisite variety. Ashby's law is concerned with stable regulation in the presence of external perturbations, and it says that the internal variety of the control must be at least as great as the external variety of disturbances to be compensated. Control reduces variety, in contrast to information transmission that conserves variety. In other words, there are positive external signals that should lead to responses of the cell, that is, specific state changes, and there are external disturbances that should not lead to changes of the internal state of the cell. We then interpret the difference between the sequence or process information on one hand and the pathway information on the other hand as the reduction in variety the cell needs to provide for compensating external disturbances and maintaining a stable function in the presence of such disturbances. For instance, if the cell produces a particular polypeptide when some external variable, like the temperature, exceeds some threshold, then the decision involves one bit of information (on/off) in response to one external bit of information (temperature above/below threshold). The concrete numbers we have provided here then allow for an estimate of the amount of compensation of external variety that the cell is capable of. In fact, some of this is contained in the difference between the quantity I_{decision} of (16) and the value of I_{genon} of (15). However, there are other compensation mechanisms not visible in the decision processes during the cascade of regulation, but for instance more generally represented in the protein composition of a cell and therefore in information theoretic terms hidden in the sequence information of the different polypeptides.

The essential conclusion of the present paper is that it is possible to try to evaluate quantitatively some of the control mechanisms of gene expression that are necessary because genes are not regulated individually. The fundamental aspect of eukaryotic gene regulation is the coordinated activation of specific sets of genes, which is basically governed by the sequential arrangement of oligomotifs in

the sequence of a genon and their selection by a transgenon. We have identified and quantified a combinatorial scheme of RNA–protein interactions underlying this coordinated regulation. However, the basic biochemical rules underlying the code of definition and selection of oligomotifs are not yet known (there should be a specific number of such oligos, substantially smaller than the combinatorial possibilities for 15–50 nucleotides and most likely varying among species and therefore not as universal as the genetic code). An example might be the recent proposition of a "splicing code" [5]. Furthermore, the mechanism of regulation by RNA–RNA interactions, which probably is at least as important in this regard, had to be left to a future investigation.

Acknowledgments 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. We thank the three referees for their insightful comments. The second author was supported by the French CNRS, the Universities Paris 6 and 7, and by bioMérieux SA, in particular by Dr. Alain Mérieux. He also thanks the Max Planck Institute for Mathematics in the Sciences for its hospitality and good working conditions.

References

1. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Rubin GM, Venter JC (2000) The genome sequence of *Drosophila melanogaster*. Science 287:2185–2195
2. Aguzzi A, Miele G (2004) Recent advances in prion biology. Curr Opin Neurol 17:337–342
3. Ashby WR (1956) An introduction to cybernetics. Chapman and Hall, London
4. Banerjee A, Jost J (2007) Laplacian spectrum and protein–protein interaction networks. eprint arXiv 0705.3373
5. Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, Blencowe BJ, Frey BJ (2010) Deciphering the splicing code. Nature 465:53–59
6. Birnstiel ML, Sells BH, Purdom IF (1972) Kinetic complexity of RNA molecules. J Mol Biol 63:21–39
7. Casarrubea D (2013) A new mouse model with gain of iron regulatory protein 1 function. PhD thesis, Heidelberg University
8. Chen K, Rajewsky N (2007) The evolution of gene regulation by transcription factors and microRNAs—review. Nat Rev Genet 8:93–103
9. Deutsch M, Long MY (1999) Intron and exon structures of eukaryotic model organisms. Nucleic Acids Res 27:3219–3228
10. Djebali S et al (2012) Landscape of transcription in human cells. Nature 489:101–108
11. Edstrom J (1964) The role of chromosomes in development. Academic Press, New York, p. 137
12. Project Consortium ENCODE (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447:779–961. doi:10.1038/nature05874
13. http://www.gencodegenes.org/releases/14.html
14. Hamkalo BA, Miller OL, Bakken AH (1974) Ultrastructure of active eukaryotic genomes. In: cold spring harbor symposia on quantitative biology, vol 38, pp 915–919
15. Hillier L, LaDeana W et al (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432:695–716
16. Hotta Y, Chandley AC, Stern H (1977) Biochemical analysis of meiosis in the male mouse. Chromosoma 62:255–268
17. Hotta Y, Chandley AC, Stern H (1977) Meiotic crossing-over in lily and mouse. Nature 269:240–242
18. Imaizumi-Scherrer M-T, Maundrell K, Civelli O, Scherrer K (1982) Transcriptional and post-transcriptional regulation in duck erythroblasts. Dev Biol 93:126–138
19. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921
20. Lara FJ (1987) Gene amplification in *Rhynchosciara* (1955–1987). Mem Inst Oswaldo Cruz 82(Suppl 3):125–128
21. Mayer MP (2010) Gymnastics of molecular chaperones—review. Mol Cell 39:321–331
22. Moreau J, Matyash-Smirniaguina L, Scherrer K (1981) Systematic punctuation of eukaryotic DNA by AT-rich sequences. Proc Natl Acad Sci USA 78:1341–1345
23. Pastor-Satorras R, Smith E, Sole RV (2003) Evolving protein interaction networks through gene duplication. J Theor Biol 222:199–210
24. Razin SV, Farrell CM, Recillas-Targa F (2003) Genomic domains and regulatory elements operating at the domain level. Int Rev Cytol 226:63–125
25. Razin SV, Rynditch A, Borunova V, Ioudinkova E, Smalko V, Scherrer K (2004) The 33 kb transcript of the chicken alpha-globin gene domain is part of the nuclear matrix. J Cell Biochem 92:445–457
26. Salser W, Browne F, Clarke P, Heindell H, Higuchi R, Paddock G, Roberts J, Studnicka G, Zakar P (1976) Determination of globin mRNA sequences and their insertion into bacterial plasmids. Prog Nucleic Acid Res Mol Biol 19:177–204
27. Scherrer K (1980) Cascade regulation: a model of integrative control of gene expression in eukaryotic cells and organisms. In: Kolodny GM (ed) Eukaryotic gene regulation, vol 1. CRC Press Inc., Boca Raton, pp 57–129
28. Scherrer K (1989) A unified matrix hypothesis of DNA-directed morphogenesis, protodynamism and growth control. Biosci Rep 9(2):157–188
29. Scherrer K (2010) The genon-concept and the transcription factor cycle (TFC) hypothesis. Abstr Book A 045:112
30. Scherrer K, Jost J (2007) The gene and the genon concept: a functional and information-theoretic analysis. Mol Syst Biol 3:87
31. Scherrer K, Jost J (2007) The genon concept: gene storage and expression in view of information theoretic analysis. Theory Biosci. doi:10.1007/s12064-007-0012-x.
32. Scherrer K, Marcaud L (1968) Messenger RNA in avian erythroblasts at the transcriptional and translational levels and the problem of regulation in animal cells. J Cell Physiol 72:181–212
33. Snyder M, Gerstein M (2003) Genomics. Defining genes in the genomics era. Science 300:258–260
34. Theil E (1990) Regulation of ferritin and transferrin receptor mRNAs. JBC 265:4771–4774
35. Thompson A, Rogers J, Leedman P (1999) Iron-regulatory proteins, iron-responsive elements and ferritin mRNA translation. Review. Int J Biochem Cell Biol 31:1139–1152
36. Vincent A, Goldenberg S, Standart N, Civelli O, Imaizumi-Scherrer MT, Maundrell K, Scherrer K (1981) Potential role of mRNP proteins in cytoplasmic control of gene expression in duck erythroblasts. Mol Biol Rep 7:71–81
37. Wagner A (2001) The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. Mol Biol Evol 18:1283–1292
38. Wagner A (2003) How the global structure of protein interaction networks evolves. Proc R Soc B: Biol Sci 270:457–466