Transcription Regulation and Genome Organization

Hermsen, R.

2008

document version
Publisher's PDF, also known as Version of record

Summary

A defining property of living systems is their ability to respond to signals. These signals are of a physical or chemical nature: for instance, many organism detect light intensities (*seeing*), mechanical forces (*feeling*), and the presence of certain molecules in the environment (*smelling* and *tasting*). Their responses to such clues are not arbitrary. They have evolved to allow organisms to adjust their behavior to varying environments and circumstances, and ultimately to increase their chances to survive and create offspring.

It is no surprise that *humans* can see, smell, taste and feel, nor that they adjust their behavior on the basis of such sensory information. But it is less obvious how micrometer-sized single-cellular creatures such as bacteria can obtain, weight and exploit knowledge of a multitude of physico-chemical quantities. Yet, they obviously do. Bacteria swim towards food sources and to warmer places, they synchronize their biological clocks to circadian rhythms, monitor the density of their colony, measure the osmolarity of their environment and assess which types of sugars are available.

Bacteria often have to make *logical* decisions. A famous example is the sugar utilization system in the bacterium *Escherichia coli*. In order to uptake and digest different sugars, such as glucose, lactose, galactose and arabinose, *E. coli* needs to produce particular sets of proteins that catalyze the required metabolic reactions. However, these sugars are not always present in the environment. As the production of the proteins requires an investment in terms of energy and other resources, it would be quite inefficient to produce them constitutively. Hence, *E. coli* decides when to make these enzymes and when not to, depending on the availability of the sugars. As it turns out, glucose is *E. coli*'s preferred source of energy, because it allows for the highest growth rate. Therefore, *E. coli* produces the proteins necessary for the digestion of other sugars only if these sugars are present and no glucose is found in the environment. This illustrates that the bacteria integrates several input signals (sugar availabilities) to make the decision. In this example, the decision procedure can be described by the Boolean logic function ANDN (A **AND N**ot B). In general, many of the decisions taken by cells can be categorized using the language of Boolean logic.

Cells implement many decisions at the level of *transcription*. Transcription is the molecular process by which genes (stretches of DNA containing the information required for the synthesis of one protein[6]) are copied (transcribed) to a different molecular medium: RNA. These RNA copies are produced by a multi-subunit

[6]Some genes actually code for a stable RNA molecule instead of a protein.

molecular machine called RNA polymerase (RNAp). Each of these RNA molecules, so-called *messenger* RNAs or mRNAs, is subsequently used as a template for the assembly of one particular protein. As a consequence, the rate at which a gene is transcribed determines (to a large extent) how many copies of the corresponding protein are produced.

The term *transcription regulation* refers to the processes used by cells to regulate the rate at which specific genes are transcribed. The resulting changes in protein concentrations are of crucial importance, because proteins determine many of the cell's properties. Many proteins serve as enzymes, which modulate the rate at which chemical processes occur in the cell; as structural elements, they constitute many of the cellular structures; and other proteins operate as tiny motors converting chemical energy to mechanical motion. In conclusion, by modulating the transcription rates of sets of genes, cells can drastically alter their protein content, and consequently their behavior and appearance.

The regulation of transcription is mediated by a special family of proteins. These proteins are called *transcription factors* and function by virtue of their ability to bind rather specifically to particular DNA sequences. Such sequences are typically located close to the starting points of genes, where RNAp initiates the transcription process. When transcription factors bind to their binding sites, they can influence the efficiency of the first steps of the transcription process and hence change the transcription rate.

The work described in this dissertation concerns with, on the one hand, the mechanisms of transcription regulation, and on the other hand, the consequences of these processes for the organization of genomes. The analyses are based on theoretical models, but published experimental data are being used to test these models and their predictions. In our work, we mainly focus on prokaryotes, featuring the bacteria *Escherichia coli* in the leading role — even though various other organisms play a supporting part.

Mechanisms of transcription regulation

Experiments have identified several mechanisms by which bacteria regulate transcription rates. Most of these mechanisms rely on fact that transcription factors are able to either recruit other molecules to the DNA or, conversely, to prevent them from binding. For instance, if a transcription factor recruits RNAp to its binding site on the DNA (called the promoter), it activates transcription. If, on the other hand, it obstructs the binding of RNAp, then transcription is inhibited. This way, the transcription rate of genes can be made to depend on the concentrations of certain (active forms of) transcription factors.

Interestingly, transcription factors need relatively simple physical properties in order to function. In bacteria, transcription factors freely diffuse through the cell; because of the small diameter of a bacterial cell (in the order of a micrometer), a protein needs about 0.1 s to diffuse from one end of the cell to the

other, which makes active transport unnecessary. Transcription factors interact with the DNA and with other molecules (including other transcription factors) due to electrostatic interactions and hydrogen bonds. They can preclude another molecule (*e.g.* a transcription factor or RNAp) from binding to the DNA by binding strongly to a site that overlaps with the binding site of the another molecule.

In the Chapters 2 and 3 of this thesis, we ask what kind of functionality can in principle be obtained with these mechanisms, recruitment and hindrance, only. For this purpose, we formulate a model of transcription regulation. In this model, transcription factors can bind specifically to sites on the DNA. They can also recruit other molecules to the DNA if they bind sufficiently close together. Binding sites can overlap; molecules thus compete for binding to particular sites on the DNA. Subsequently, we use the formalism of statistical mechanics to calculate which fraction of the time molecules are bound. We use this model and an evolutionary algorithm to design transcription-regulatory systems that perform a pre-defined function and we thus explore the space of possible mechanism.

The simple mechanisms of recruitment and hindrance turn out to be immensely versatile. We show that, using rather complex distributions of binding sites, transcription regulation can perform all possible Boolean logic operations with two inputs. The functional designs often consist of modules of tandem binding sites to which transcription factors can recruit each other. This cooperative behavior leads to sharp responses of the transcription rate as a function of TF concentrations. But more intricate effects can be obtained if the modules (partly) overlap with each other, introducing competition for binding at the level of these complexes. Which module dominates in such a competition can depend strongly on the concentrations of the different transcription factors, which can be exploited to integrate different signals.

The complex designs that we describe are not unrealistic. We demonstrate that many real promoters in the bacterium *Escherichia coli* contain large numbers of transcription factor binding sites, and that overlap between these sites is extremely common. Also, transcription factors often bind to more than one binding site in a given promoter region — in exceptional cases, up to eleven sites for a single transcription factor have been documented.

Another world of possibilities is entered if we allow the systems to use feedback. In the simplest case, this means that the gene that is being regulated codes for a transcription factor that influences its own transcription rate (called auto-regulation). It has been shown that this allows for fine-tuning of the dynamical properties of these systems — *e.g.* their robustness to noisy signals or their response speed. We demonstrate that auto-regulation can also be used to achieve a more efficient repression mechanism and that it allows for alternative ways to integrate signals.

Again, the mechanisms we find are realistic. Auto-regulation is very common in *E. coli*: 59% of the transcription factors are known to bind to their own

promoter region, and this is likely a lower bound since our current knowledge of regulatory interactions is far from complete. The mechanisms we discovered shed new light on the possible functions of these feedback systems, most of which are not yet elucidated.

Chromosome organization

The processes of transcription and transcription regulation also has a considerable impact on the way genes are distributed on chromosomes. To start with, all regulatory sequences, such as binding sites for RNA polymerase and transcription factors, take up space on the DNA and thereby influence the spacing between genes. Indeed, regulatory sequences directly before and after genes leave footprints on the frequency distribution of distances between genes. (Here distances are measured in base pairs.) Therefore the statistical properties of the distances between genes reveal many properties of the gene regulatory mechanisms used in the organisms.

Distances between genes

In order to study the frequency distributions of distances between genes in detail, we compare them with random models. We exploit that, mathematically, these random models are equivalent to models of one-dimensional gases. In this analogy, genes correspond to gas particles and the DNA acts as a one-dimensional, finite space.

In the most naive model, the genes are distributed completely at random. This is formally equivalent to an ideal gas. This model does not describe the data well, because genes usually do not overlap. Therefore, a second model is proposed; here we assume that genes are distributed at random, except that they do not overlap. This model is analogous to the so-called Tonks gas: a gas of hard particles in one dimension. The Tonks gas model offers a better description of the gene distributions, but fails to explain why genes have a tendency to keep a certain minimal distance from each other. This inspires the final random model, called the Constant-Force model. In the Constant-Force model, we assume that the genes are accompanied by regulatory sequence that occupy space and therefore "push" the genes apart. This leads to a picture in which the genes are distributed at random, except that they do not overlap and repel each other at short distances. The Constant-Force model provides a good fit to the gene distributions in species such as *E. coli* and *Saccharomyces cerevisiae* (Baker's yeast).

The typical lengths of upstream and downstream regulatory regions are fit parameters of our model. This means that, by fitting our model to the distribution of genes in a particular organism, we can estimate the lengths of these upstream and downstream regulatory sequences using only the positions of genes as input. We use this to estimate lengths for various organisms.

The genomic data deviate from the Constant-Force model on several points. These deviations lead to interesting biological predictions. For instance, in most *fungi*, the distribution of distances between divergent gene pairs — neighboring genes that are transcribed from opposite DNA strands and in diverging directions—is bi-modal, strongly suggesting that their genomes contain many bi-directional promoters. Similarly, in *E. coli* we find a significant access of convergent gene pairs — neighboring gene pairs that are transcribed from opposite strands and in a converging orientation — that are unusually closely spaced; we predict that these gene pairs share a bi-directional terminator. We test all these predictions using expression data, Gene Ontology annotations and terminator predictions; the results indeed corroborate our hypotheses.

Operons

A special feature of most (if not all) prokaryotes and a few eukaryotes, is that their genes are organized in so-called *operons*. An operon is a cluster of several genes that are in one transcription unit. This means that the transcription machinery produces one long messenger RNA that contains all these genes. Genes in one operon are usually very closely spaced, and are always coded on the same strand—in so-called *tandem* orientation. As a result, the set of tandem neighboring gene pairs in such genomes consist of two populations: those pairs of genes that are in the same operon, and those that are in different operons. Correspondingly, the sequences between those genes (intergenic regions) are either inside an operon or between two operons. This division is visible in the distribution of the distances between tandem gene pairs: it is largely consistent with our random model, except that a considerable excess of gene pairs is found at short distances. Thus, the distribution of distances reveals the presence of operons.

A subject of ongoing debate is *why* genes are organized in operons. One school of thought argues that operons are used to co-regulate genes. Indeed, if several genes need to be expressed in a correlated fashion — perhaps because they have a related function — this could be achieved by placing them in one operon. Others argue that operon formation relies on horizontal gene transfer: the exchange of genes between organisms of different species. The horizontal transfer of a set of functionally interdependent genes may be more successful if they are organized in one cluster (an operon) than if they are dispersed on the genome. Therefore, operons may be "selfish" structures: their abundance could be due to their reproductive success rather than due to their added value for the organism.

These two arguments have one thing in common: they both silently assume that operons would not exist in the absence of any selective pressure to create them. In Chapter 5 we suggest quite the opposite: even if operons do not have any selective advantage (neither at the level of organisms, nor at the level of clusters of genes), they are expected to emerge. The reason is that two neighboring tandem genes are naturally in the same operon, *unless* there is a transcriptional terminator sequence in the intergenic region between them. This means that, in a sense,

operons are the default: only if there is sufficient and persistent evolutionary pressure to regulate genes independently, transcriptional terminators and private promoters are expected to emerge in the course of evolution. Moreover, existing terminators are continually challenged by myriad mutations. On evolutionary time scales they will survive only if they are under constant and sufficient purifying selection. Whenever this is not the case, the terminator will be lost, and operons form immediately.

The above picture holds only for prokaryotes. In prokaryotes, all that is required to produce a new operon is the removal of a transcriptional terminator between tandem genes. In eukaryotes, this is generally not enough, because the eukaryotic ribosomes, which use the mRNA templates to assemble proteins, cannot deal with mRNAs containing more than one gene, unless the mRNA contains an internal ribosome entry site (IRES). Without such a special sequence, the ribosome translates the first gene on the mRNA only, and ignores the other genes. This explains why operons are much more rare in eukaryotes than in prokaryotes.

To prove the concept, we present a simplified model of genome evolution and developed a novel simulation scheme based on population genetics. In simulations of this model, operons and shared terminators indeed emerge spontaneously. Moreover, the model reproduces the spacing of genes in the model prokaryotes *Escherichia coli* and *Bacillus subtilis*, including the characteristic close spacing of genes in operons and the differences in spacing between convergent, divergent and tandem gene pairs. As a side effect, it also explains why promoters and terminators usually tightly flank the genes they regulate.

Evolution of intergenic distances

Intergenic regions grow and shrink due to insertions and deletions. In intergenic regions, these mutations are typically only a few base pairs long. As the occurrence of mutations is a stochastic process, one would expect that the lengths of intergenic regions perform a "random walk" on evolutionary time scales. In Chapter 6 we propose a stochastic model for this evolutionary "diffusion" of intergenic regions.

This idea can be tested using data from closely related species. If a speciation event occurs, giving rise to two different species, the intergenic distances in the two resulting species are initially expected to be equally long. However, on evolutionary time scales, random insertions and deletions will lead to a decorrelation of these distances. This process is the combined effect of the random walks performed by the two species independently; therefore we can test our model by comparing the intergenic distances of two related species.

We compare our model to the divergence of the intergenic distances in *Escherichia coli* and *Salmonella enterica subsp. enterica serovar Typhi*. We focus on intergenic regions between tandem gene pairs. As we described, such intergenic regions come in two kinds: those "between" operons and those "inside" operons. The intergenic regions in the these two groups have rather different compositions: for instance, those of type "between" contain a transcriptional terminator, whereas

those “inside” do not. Also, the typical lengths of the two kinds of intergenic regions are very different. This implicates that the evolutionary diffusion of these different types is also not the same.

The main ingredients of the model are as follows. We assume that intergenic regions consist of, on the one hand, elements that have a fixed length (*e.g.* transcriptional terminators and promoters), and, on the other hand, “spacers”, whose length can change substantially. The rate at which mutations occur in spacers is assumed to depend linearly on their length, since a longer spacer contains more places where an insertion or deletion can take place. Therefore, longer intergenic regions are expected to change faster. These considerations can be formalized using Master equations and Fokker–Planck equations, which allow for quantitative predictions.

The diffusion model fits the data of *E. coli* and *S. Typhi* well. The data clearly show bigger changes in longer intergenic regions, and the diffusion of the two types of spacers is indeed different. The fit parameters also show that the divergence between *E. coli* and *S. Typhi* cannot be explained by insertions and deletions of single base pairs only. Indeed, calculations show that larger mutations can have a large influence on the speed of the evolution of the lengths of intergenic regions, even if they occur at a low rate.

The model can also be used to compute what happens if an operon splits in two, or if two operons merge. This is particularly relevant in the light of our suggestion in Chapte 5 that operons may form by merging processes that result from the loss of terminator sequences. If an operon splits in two, one intergenic region has to switch from type “inside” to type “between”. Conversely, if two operons merge, an intergenic region has to change from type “between” to type “inside”. Consequently, also the mode of diffusion of this intergenic region changes. Calculation of these processes allow for the identification of intergenic regions in which such a merging or splitting event may have taken place. We discuss these candidate regions in detail.