Published in IET Systems Biology
Received on 12th January 2008
Revised on 5th May 2008
doi: 10.1049/iet-syb:20080091

Special Issue – Selected papers from the First q-bio Conference on Cellular Information Processing

ISSN 1751-8849

Protein–protein/DNA interaction networks: versatile macromolecular structures for the control of gene expression

L. Saiz[1] *J.M.G. Vilar*[2]

[1]*Department of Biomedical Engineering, University of California, 451 East Health Sciences Drive, Davis, CA 95616, USA*
[2]*Integrative Biological Modeling Laboratory, Computational Biology Program, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, Box #460, New York, NY 10021, USA*
E-mail: lsaiz@ucdavis.edu

Abstract: The assembly of macromolecular structures consisting of proteins and DNA lies at the core of many fundamental cellular processes, such as transcription, recombination and replication. A common theme to all these processes is DNA looping, which provides the backbone for the required long-range interactions on DNA and results in further complexity that is exceptionally difficult to tackle with traditional quantitative approaches. Here, recent advances in mathematical and computational methods to study the assembly of protein–protein/DNA complexes with loops and their effects in the cellular behaviour through gene regulation are reviewed. The interplay between multisite DNA looping and DNA bending regulatory proteins, such as the catabolite activator protein (CAP), and on its physiological consequences is focused on. It has become clear in the last few years that the complexity that looping brings about can actively control transcriptional noise and cell-to-cell variability. Here, it is shown that the DNA looping, through the effects of CAP, can also control the balance between robustness and sensitivity of the induction of gene expression.

1 Introduction

Networks of protein–protein and protein–DNA interactions mediated by the presence of DNA looping are deeply involved in many cellular processes, such as transcription, recombination and replication [1–4]. They are especially prominent in the regulation of gene expression, where proteins bound far away from the genes they control can be brought to the initiation of transcription region by looping the intervening DNA. The interplay between DNA looping and gene regulation was suspected early on to be present in eukaryotic enhancers [5] and was first identified in the *Escherichia coli ara* operon [6]. Since then, it has been experimentally studied in detail in many other systems, including *gal*, *lac* and *deo* operons in *E. Coli* [2, 4], the lysogenic to lytic switch in phage λ [7], the human β-globin locus [8], the nuclear hormone receptor RXR [9] and the tumour suppressor protein p53 [10].

Understanding of macromolecular assembly on looped DNA, especially when multiple binding sites and loops are involved [11, 12], is challenging at both biochemical and mathematical levels. From a biochemical point of view, looping introduces flexibility into the macromolecular assembly of protein-DNA complexes, which are no longer restricted to have a fixed rigid structure [13–15]. From a mathematical point of view, looping leads to the possibility of establishing simultaneous interactions between many components, which results in a large number of potential states of the protein-DNA complex [16]. Typically, the number of potential states scales exponentially with the number of components. This type of scaling, usually referred to as 'combinatorial complexity' [17] makes traditional state-based approaches impracticable for systems with more than just a few proteins. A notable example in which these issues are present is the prototypical *lac* operon in *E. coli*, which is still not completely understood despite being one of the two systems that led to the discovery of gene regulation [18].

In this review, we use the *lac* operon to illustrate how both of these biochemical and mathematical challenges have been addressed in recent years to provide an effective framework to faithfully study protein–protein/DNA interaction networks in realistic gene regulation setups. The general picture emerging from these quantitative analyses reveals that the presence of multiple binding sites and looped structures provides avenues to control cellular variability and to combine robust repression with sensitive induction, two seemingly mutually exclusive properties that are required for optimal functioning of metabolic switches.

2 The *lac* operon

The *E. coli lac* operon is the genetic system that regulates and produces the enzymes needed to metabolise lactose [18]. The response to lactose is controlled by the *lac* repressor [19], which can bind to O_1, the main operator, and prevent the RNA polymerase from binding to the promoter and transcribing the genes. There are also two auxiliary operators, O_2 and O_3, to which the repressor can also bind but not prevent transcription (Fig. 1*a*). Elimination of either one auxiliary operator has only minor effects; yet simultaneous elimination of both of them reduces the repression level by a factor 100 [20]. The reason behind this effect is that the *lac* repressor molecule has two DNA binding sites and thus can bind simultaneously to two operators and loop the intervening DNA.

3 Traditional quantitative approach to transcription regulation

The most widely used quantitative approaches to study DNA-protein assembly are based on thermodynamics [21]. Thermodynamics allows for a straightforward connection of the molecular properties of the system with the effects that propagate up to the cellular physiology.

The use of thermodynamic concepts applied to gene regulation was pioneered by Shea and Ackers [22, 23] and applied subsequently to a wide variety of gene regulation systems with simple binding events [24, 25] and with binding involving DNA looping [1, 26].

In the traditional framework, the probability for the macromolecular complex to be in a state k is given by $P_k = n^{j_k} e^{-\Delta G_k^{\circ}/RT}/Z$, where RT is the gas constant times the absolute temperature, ΔG_k° the standard (molar) free energy of the state k, j_k the number of molecules present in the complex in the state k and n the concentration of the molecular species. The partition function $Z = \sum_k n^{j_k} e^{-\Delta G_k^{\circ}/RT}$, where the summation is taken over all the states, is the normalisation factor.

Thus, to describe the system with the traditional approach, one has to provide the standard free energy for each state

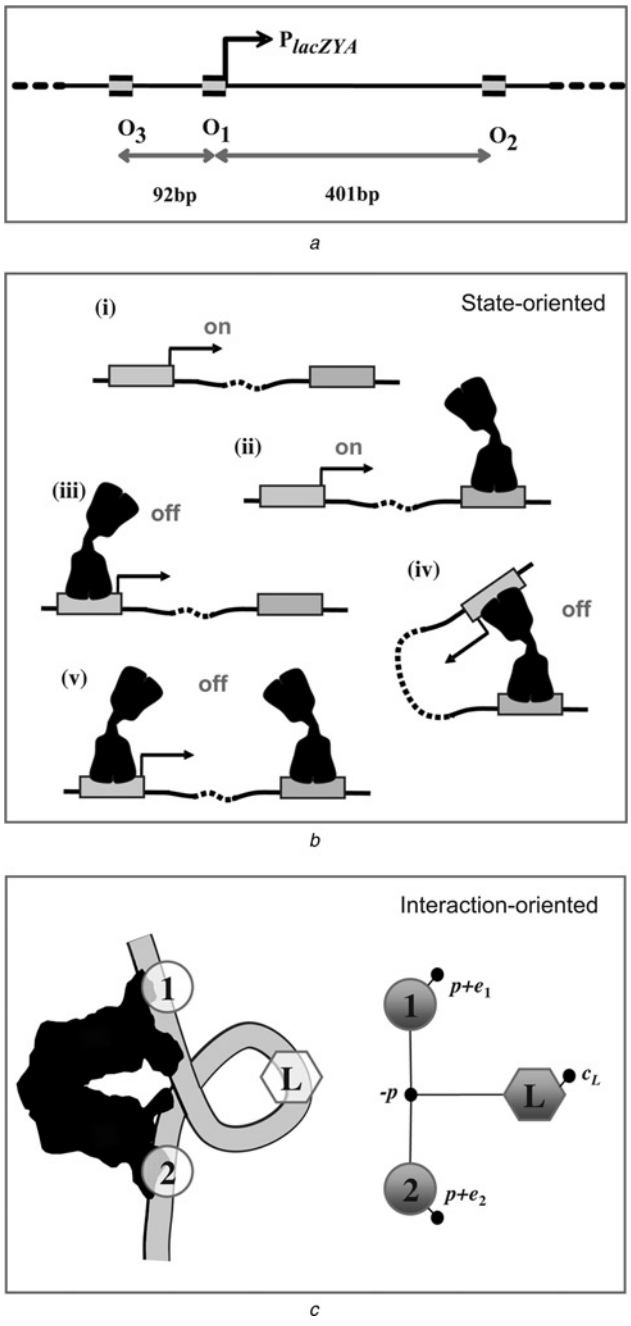

Figure 1 *The lac operon and its quantitative descriptions*

a The wild-type *lac* operon has three binding sites for the repressor: the main (O_1) and the two auxiliary (O_2 and O_3) operators, shown as grey rectangles on the thick black segment representing DNA. Binding of the *lac* repressor to O_1 prevents transcription of the three *lacZYA* genes

b The state-based description [26] of the *lac* repressor (in black) binding to two operators has five representative states. The thick black line represents DNA with the two *lac* operators shown as boxes and the start of transcription indicated by an arrow

c The interaction-based description [16] accounts for the binding of the *lac* repressor (black) to DNA (grey) using binary variables. The circles and the polyhedron correspond to the binary variables for the repressor-operator DNA binding sites and the DNA loop, respectively. The dots in the network diagram link the variables of a factor in the expression of the free energy (2) and their labels indicate the corresponding contribution to the free energy when all the linked variables are equal to one

doi: 10.1049/iet-syb:20080091

(ΔG_k°) as well as the number of molecules bound (j_k). Typically, this is done in the form of a table, which has as many entries as the number of states of the system [22]. In this way, for a macromolecular complex with N elements that can either be present or absent in the complex, one would need a table with 2^N entries for the standard free energy and another 2^N for the number of molecules. In general, if there are different conformational states, the number of states can be higher than 2^N. This is the case of the complex formed by two-operator DNA and the *lac* repressor (Fig. 1*b*) [26]. In this case, n is the concentration of free repressor in the cell right before the binding event takes place, and j_k is the number of molecules bound specifically to operator DNA in the state k. Thus, for the states k= (i), (ii), (iii), (iv) and (v) in Fig. 1*b* one has $j_{(i)}=0$, $j_{(ii)} = j_{(iii)} = j_{(iv)} = 1$ and $j_{(v)} = 2$.

The connection of macromolecular assembly with gene regulation is done simply by assuming that each state k has a well-defined transcription rate, Γ_k, which is used to compute the effective transcription rate as the average $\bar{\Gamma} = \sum_k \Gamma_k P_k$ [26].

4 Thermodynamic binary-variable approach to transcription regulation

Straightforward application of the traditional thermodynamic approach in a general framework is of limited use because the number of states that must be considered, as we have discussed, typically increases exponentially with the number of components. It has become clear recently that it is possible to efficiently overcome this limitation with a new approach that expresses the free energy of all the states in a compact form in terms of binary variables [16].

This approach is based on four major premises [16].

1. The specific configuration or state of the macromolecular complex can be described by a set of M binary variables, denoted by $s = (s_1, \ldots s_i, \ldots s_M)$, whose values indicate whether a particular molecular component is present ($s_i = 1$) or absent ($s_i = 0$) at a specific position within the complex [16]. The use of binary variables provides a concise method to describe all the potential complexes without explicitly enumerating them. This type of approach has been used in a wide range of interesting biological situations, such as diverse allosteric processes [27], binding of molecules to a substrate [28, 29], binding of multi-state proteins to receptor docking sites [30] and signalling through clusters of receptors [31–33].

2. The free energy of the complex can be decomposed into different modular contributions so that the free energy of all the possible configurations of the complex can be expressed as a function of both the different contributions to the free energy and the binary variables s. These contributions can be divided into positional, interaction and conformational free energies [1]. Briefly, the positional free energy, p, accounts for the cost of bringing one component to the protein-DNA complex, for instance, bringing the *lac* repressor to its DNA binding site. Its dependence on the component concentration, n, is given by $p = p^\circ - RT \ln n$, where p° is the positional free energy at 1 M. Interaction free energies, e, arise from the physical contact between components (e.g. electrostatic interactions) and conformational free energies, c, account for changes in conformation (e.g. looped vs unlooped states).

3. The transcription rate, as well as other quantities of interest, can also be expressed in terms of the binary variables.

4. The different expressions in terms of the binary variables s can be used to compute the quantities of interest without having to instantiate explicitly all the potential states of the complex [16].

With this approach, taking together premises 1–4, the effective transcription rate is obtained from

$$\bar{\Gamma} = \frac{1}{Z}\sum\nolimits_s \Gamma(s) e^{-\Delta G(s)/RT} \quad (1)$$

by computing the thermodynamic average over all the representative states. Here, $\Delta G(s)$ and $\Gamma(s)$ are the free energy and the transcription rate for each state of the complex $s = (s_1, \ldots s_i, \ldots, s_M)$, respectively, and $Z = \sum_s e^{-\Delta G(s)/RT}$ is the partition function used as a normalisation factor.

5 Application to a two-operator *lac* operon setup

In the case of the *lac* operon with two operators, O_1 and O_2, this new approach has a straightforward implementation (Fig. 1*c*) [16].

The free energy of the protein-DNA complex can be expressed as

$$\Delta G(s) = (p + e_1)s_1 + (p + e_2)s_2 + (c_L - p s_1 s_2)s_L \quad (2)$$

Here, e_1 and e_2 are the interaction free energy between the repressor and O_1 and O_2, respectively; and c_L is the conformational free energy of looping DNA. The binary variables s_1 and s_2 indicate whether ($s_i = 1$; for $i = 1,2$) or not ($s_i = 0$; for $i = 1, 2$) the repressor is bound to O_1 and O_2, respectively; and s_L is a variable that indicates the conformational state of the DNA, either looped ($s_L = 1$) or unlooped ($s_L = 0$). Thus, it is possible to write a global concise expression [16], instead of one for each of the five states [26], to specify the thermodynamic properties of the system.

```
In[1]:= p := p0 - RT Log[n]

In[2]:= ΔG := (p + e1) s1 + (p + e2) s2 + (cL - p s1 s2) sL

In[3]:= Z := Σ_{sL=0}^{1} Σ_{s2=0}^{1} Σ_{s1=0}^{1} e^{-ΔG/RT}

In[4]:= Γ̄ := (1/Z) Σ_{sL=0}^{1} Σ_{s2=0}^{1} Σ_{s1=0}^{1} Γmax (1 - s1) e^{-ΔG/RT}

In[5]:= Parameters := {RT → 0.6, p0 → 15, e1 → -27.8,
                       e2 → -26.3, cL → 23.35}

In[6]:= RepressionLevel = Simplify[Γmax / Γ̄] /. Parameters

Out[6]= (2.31023×10^-8 (0.286505 + 7.25755×10^10 n + 7.96698×10^16 n^2)) / (6.61892×10^-9 + n)
```

Figure 2 *Calculation of the repression level*

This figure illustrates how to compute the repression level with the software package *Mathematica* 6 (http://www.wolfram.com/) for the two-operator *lac* operon. The repression level is defined as the inverse of the normalised effective transcription: $\Gamma_{max}/\bar{\Gamma}$. The result obtained (Out[6]) is plotted in Fig. 3

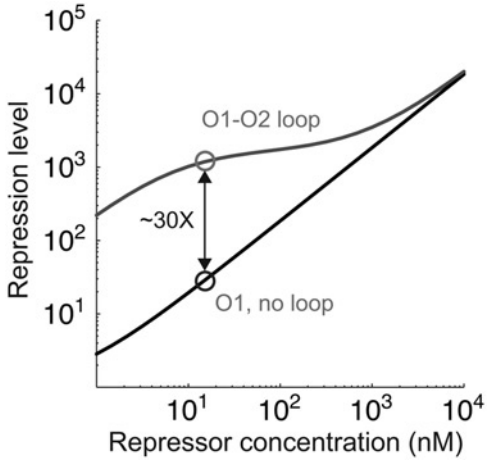

Figure 3 *Repression with and without looping in the lac operon*

The repression level ($\Gamma_{max}/\bar{\Gamma}$) is shown as a function of the *lac* repressor concentration for the *lac* operon with two (grey curve, labelled '01-02 loop') and one (black curve, labelled '01, no loop') operators. The two-operator case corresponds to the output (Out[6]) of the *Mathematica* 6 code shown in Fig. 2. The one-operator case was calculated in a similar fashion using a modified version of the *Mathematica* 6 code, adapted for a single binding site. Adding another binding site and DNA looping (O_1–O_2 loop) to the one-operator set-up (O_1, no loop) increases the repression level by a factor ~30 for wild-type concentrations of the repressor (grey circles; ~10 repressors/cell) [26]

The transcription rate is expressed in terms of these binary variables as

$$\Gamma(s) = \Gamma_{max}(1 - s_1) \tag{3}$$

where Γ_{max} is the maximum transcription rate.

The effective transcription rate, $\bar{\Gamma}$, is obtained by computing the thermodynamic average of $\Gamma(s)$ over the representative states, as discussed previously. This average can be computed straightforwardly using standard computer algebra software (Fig. 2) or if the system is too large, using Monte Carlo algorithms [16].

Note that the fact that there are usually multiple copies of the *lac* operon in a single cell is straightforwardly taken into account by considering that the different DNA regions containing a copy of the *lac* operon, which are located in different chromosomes, behave independently of each other and therefore that the average transcription rate is proportional to the number of copies. For practical purposes, this dependence with the number of copies is already incorporated in the value of Γ_{max} when it is identified with the maximum transcription rate per cell.

6 Effects of DNA looping

DNA looping has many obvious effects because of its role in mediating long-range interactions on DNA. It allows two, or more, DNA regions that are far apart to come close to each other, which is needed, for instance, to allow the transfer of genetic information that happens during recombination [34, 35]. DNA loops are also used to tie the end of chromosomes and regulate the length of telomeres [36].

Beyond these systems in which it is strictly required, DNA looping is also used to increase the binding of regulatory molecules to their cognate sites. In the case of the *lac* operon, such increased binding results in repression of transcription that is 30 times stronger (Fig. 3). DNA looping has also other more subtle roles, which are strongly interrelated with the inherent stochastic nature of cellular processes. Computational modelling of the *lac* operon [26] together with experimental data [20] showed that DNA looping can also be used to decrease the sensitivity of transcription to changes in the number of regulatory proteins. The transcription rate in the *lac* operon for the looping case shows a plateau-like behaviour, which is not present in the regulation with just a single operator (Fig. 3). The low sensitivity obtained with DNA looping in this region can be used to achieve fairly constant transcription rates among cells in a population, irrespective of the fluctuations in the numbers of *lac* repressor molecules. In contrast, using a single operator just propagates the fluctuations proportionally [1, 26].

It has also been shown [26] that DNA looping can reduce the intrinsic fluctuations of transcription [37–41]. If transcription switches slowly between active and inactive, there are long periods of time in which proteins are produced constantly and long periods without any production. Therefore, the number of molecules fluctuates strongly between high and low values. In contrast, if the switching is very fast, the production happens in the form of short and frequent bursts. This lack of long periods of time with either full or null production gives a narrower distribution of the number of molecules. DNA looping

doi: 10.1049/iet-syb:20080091

naturally introduces a fast time scale: the time for the repressor to be recaptured by the main operator before unbinding the auxiliary operator, which is much shorter than the time needed by a new repressor to bind to the main operator. Therefore, DNA properties are also important for controlling transcriptional noise [26].

7 Application to the wild-type, three-operator *lac* operon

The wild-type *lac* operon has three operators for specific binding of the repressor (Fig. 1*a*). In this case, the thermodynamic binary-variable approach has to consider the presence of multiple loops by the *lac* repressor between different combinations of pairs of DNA sites. The free energy of the protein-DNA complex of the wild-type *lac* operon can be expressed as [12]

$$\begin{aligned}\Delta G(s) = {} & (p+e_1)s_1 + (p+e_2)s_2 + (p+e_3)s_3 \\ & + (c_{L12} - ps_1s_2)s_{L12} + (c_{L13} - ps_1s_3)s_{L13} \\ & + (c_{L23} - ps_2s_3)s_{L23} \\ & + \infty(s_{L12}s_{L13} + s_{L12}s_{L23} + s_{L13}s_{L23}) \end{aligned} \quad (4)$$

where s_1, s_2 and s_3 are the binary variables that indicate whether or not the repressor is bound to O_1, O_2, and O_3, respectively; and s_{L12}, s_{L13} and s_{L23} are the variables that indicate whether or not DNA forms the loops O_1–O_2, O_1–O_3, and O_2–O_3, respectively. The subscripts of the different contributions to the free energy have the same meaning as those of the corresponding binary variables. In this case, with three interaction and three conformational free energies, it is possible to obtain the free energy of 14 states for different repressor concentrations. An important advantage of the binary variable description is that it can straightforwardly implement 'logical conditions'. For instance, the infinity in the last term of the free energy implements that two loops that share one operator cannot be present simultaneously by assigning an infinite free energy to those states.

Expression of the wild-type *lac* operon is completely abolished when the repressor is bound to O_1; otherwise, transcription takes place either at an activated maximum rate Γ_{max} when O_3 is free or at basal reduced rate $\chi\Gamma_{max}$ when O_3 is occupied. This reduction by a factor χ arises because binding of the repressor to O_3 prevents the catabolite activator protein (CAP) from activating transcription [42]. Activation is achieved when CAP bound to cyclic adenosine monophosphate (cAMP) binds between O_3 and O_1 and stabilises the binding of the RNA polymerase to the promoter [43]. The transcription rate $\Gamma(s)$ can be expressed in terms of binary variables as

$$\Gamma(s) = \Gamma_{max}(1 - s_1)(\chi s_3 + (1 - s_3)) \quad (5)$$

which provides a mathematical expression for the observed cis-regulatory transcription control [42, 44].

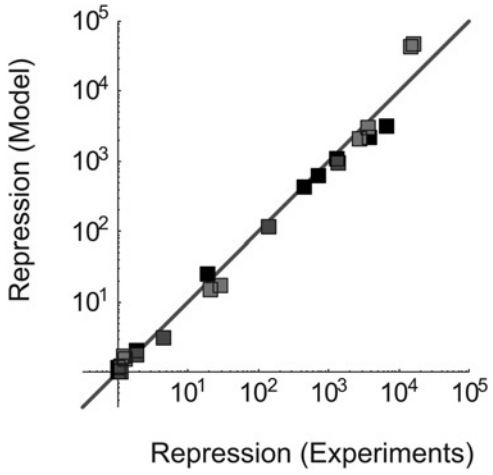

Figure 4 *Model vs. experimental repression levels*

The repression levels ($\Gamma_{max}/\bar{\Gamma}$) obtained from the three-operator wild-type *lac* operon model [12] are plotted against their corresponding experimental values [42], showing an excellent quantitative agreement over five orders of magnitude for wild-type and seven mutants accounting for all the combinations of deletions of the three operators. Three different repressor concentrations are considered, which lead to a total of 24 data points, shown as grey squares with different shades indicating increasing repressor concentrations from wild-type (darkest) with 10 repressors per cell, to 50 (medium shade) and 900 (lightest) repressors per cell. The black line corresponds to the identity between model and experimental values. The values of the parameters used are [12]: $e_1 = -27.8$ kcal/mol, $e_2 = -26.3$ kcal/mol, $e_3 = -24.1$ kcal/mol, $c_{L12} = 23.35$ kcal/mol, $c_{L13} = 22.05$ kcal/mol, $c_{L23} = 23.50$ kcal/mol, $p^\circ = 15$ kcal/mol, and $\chi = 0.03$. A deleted operator is modelled by increasing its free energy by 5 kcal/mol

It has been shown [12] that this approach accurately reproduces without free parameters the observed behaviour of the *lac* operon in quantitative detail over five orders of magnitude of the repression level for three repressor concentrations and eight strains with all the possible combinations of operator deletions (Fig. 4). In addition, this approach is able to reproduce the observed induction curves [45] of the system for different cellular conditions, such as in the presence or absence of active CAP [12].

8 Molecular gear for robust repression and sensitive induction

In the *lac* operon, the main operator and at least one auxiliary operator suffice to form DNA loops that substantially increase the ability of the repressor to bind the main operator and provide robust repression levels (Fig. 3). To what extent do the properties of transcription regulation depend on the molecular complexity that multiple DNA loops bring about?

The analysis of the model for the wild-type *lac* operon has revealed that the three-operator setup provides an efficient mechanism to combine robust repression with sensitive induction (Fig. 5) [46]. A key element is CAP, which

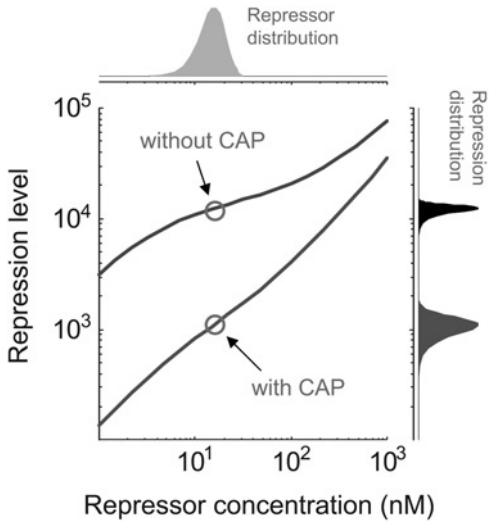

Figure 5 *Robust repression and sensitive induction*

The repression level in the presence (curve labelled 'with CAP') and in the absence (curve labelled 'without CAP') of active CAP obtained in [46] is plotted as a function of the repressor concentration. In the absence of active CAP, in addition to a reduced transcription $\bar{\Gamma} = (1/Z)\sum_s \Gamma_{max}\chi\,(1-s_1)e^{-\Delta G(s)/RT}$, the formation of the $O_1 - O_3$ loop is $\sim$1 kcal/mol more costly ($c'_{L13} = 23.05$ kcal/mol) than in the presence of active CAP, which leads to an almost flat profile of the repression level around wild-type repressor concentrations (grey circle). In the presence of active CAP, repression is reduced at the same time that the system becomes sensitive to changes in repressor concentration. Variability in repressor concentration over a population of cells, with a distribution such as that shown in light grey (top), would lead to a much wider distribution of repression levels with active CAP (in dark grey) than without active CAP (in black)

controls the stability of one of the DNA loops besides activating transcription [12]. When CAP is bound to its DNA site, the multi-loop system is sensitive to changes in the repressor concentration. The resulting sensitivity would make the system ready for induction when the repressor is inactivated by inducers such as allolactose or isopropyl-β-D-thiogalactoside (IPTG) [18]. This result strongly contrasts with the previous studies of regulation by single loops [26] which show that repression is highly insensitive to changes in repressor concentration (Fig. 3). Such robustness is recovered when active CAP is absent and the O_1–O_3 loop is not stabilised [12, 46]. Under these conditions, repression relies only in the O_1–O_2 loop and the system displays a lack of sensitivity to changes in repressor concentration. A three-operator system is thus able to put two apparently contradictory properties such as robustness and sensitivity together into a functional metabolic switch [46].

9 Stochastic dynamics of macromolecular assembly networks

It has been shown [16] that the binary-variable thermodynamic approach can be used as an efficient starting point to study the kinetics. We outline here the main ideas for the case in which only one component can change at a given time: either the component gets into or out of the complex. For each component i, we can define *on* (k^i_{on}) and *off* (k^i_{off}) rates for the 'association' and 'dissociation' rates, respectively, which, in principle, would depend on the pre-transition and post-transition states of the complex.

The explicit dynamics can be obtained by considering the change in binary variables as reactions

$$s_i \rightarrow (1 - s_i) \tag{6}$$

with rates

$$r_i = (1 - s_i)k^i_{on}(s) + s_i k^i_{off}(s) \tag{7}$$

The reaction changes the variable s_i to 1 when it is 0 and to 0 when it is 1, representing that the element gets into or out of the complex. The mathematical expression of the transition rate reduces to k^i_{on} when the element is outside the complex ($s_i = 0$) and to k^i_{off} when the element is inside the complex ($s_i = 1$). Typically, the *on* rate does not depend as strongly on the state of the complex as the *off* rate. The *on* rate is essentially the rate of transferring the component from solution to the complex. The *off* rate, in contrast, depends exponentially on the free energy. The principle of detailed balance can be used to obtain the *off* rates from the *on* rates:

$$k^i_{off}(s) = k^i_{on}(s')e^{-(\Delta G(s') - \Delta G(s))/RT} \tag{8}$$

As this expression shows, the thermodynamic model alone does not contain sufficient information to describe transition rates between all states but it can efficiently be used to relate the rates with each other and to explicitly obtain multiple *off* rates from a single *on* rate.

Note that, in general, not only the *off* rates but also the generalised *on* rates depend on the variables s. For instance, in the case of looping with two operators (Fig. 1), we have $k^L_{on} = k_{looping}(1 - s_1 s_2)$, where $k_{looping}$ is the rate of loop formation. The term $(1 - s_1 s_2)$ accounts for the fact that no loop can be formed when different repressors are bound to each operator.

The implementation of this kinetic framework considers the compact expressions for the transition rates between different states of the complex together with other reactions that affect or depend on the state of the complex. The whole set of reactions can thus be simulated using kinetic Monte Carlo algorithms [47, 48]. For instance, this methodology has allowed, for the first time, the study of the effects that DNA looping has in the induction kinetics of the lysogenic-to-lytic switch in the phage λ genetic system [16]. The processes considered included also, among others, binding and unbinding of repressors to DNA, conformational changes, production and degradation

doi: 10.1049/iet-syb:20080091

of mRNA, production and degradation of proteins, protein association and dissociation [16]. In this way, it has been possible to integrate the stochastic dynamics of macromolecular assembly into networks of chemical reactions and move the effects of macromolecular assembly up to the properties of cellular processes.

It is important to emphasise that the use of binary variables provides a very efficient avenue to tackle the combinatorial complexity even in highly cooperative systems. For instance, if the interactions are pairwise, as usually assumed, the energetics of all possible interactions between N components can be taken into account with just $N(N-1)/2$ terms in the free energy, which scales quadratically with the number of components instead of following the exponential increase of the number of states. The resulting expression for the free energy can in turn be used together with the detailed balance principle to specify an exponentially large number of *off* rates from a single *on* rate. In general, there can be situations in which the free energy and the transitions are inherently combinatorially complex. These situations, however, can only be characterised experimentally with an effort that grows exponentially with the number of components.

10 Conclusions

In this review, we have illustrated how recent statistical thermodynamics approaches based on binary variables are able to naturally incorporate the underlying molecular complexity into gene regulation models and to provide an avenue to accurately infer the effects of multiple DNA loops between different DNA sites. With this methodology, it has been possible to show that, in the *lac* operon, escalating complexity from one to two operators introduces stronger repression; and from two to three operators, concurrent robustness and sensitivity. Thus, the complexity of multiple repeated distal DNA binding sites, far from being just a remnant of evolution or a backup system as often assumed [18], can confer subtle, yet important, properties that are not present in simpler setups. These results indicate that key design principles that have been shown to play important roles in shaping the structure of biochemical networks [49–51] are also operating at the molecular level in the design and structure of protein–protein/DNA interaction networks. To bring forward these extra levels of regulation, it is crucial to have efficient methodologies, as the ones we have reviewed here, for comprehensively characterising and accurately predicting the collective properties of macromolecular complexes in terms of the properties of their constituent elements.

11 References

[1] VILAR J.M.G., SAIZ L.: 'DNA looping in gene regulation: from the assembly of macromolecular complexes to the control of transcriptional noise', *Curr. Opin. Genet. Dev.*, 2005, **15**, pp. 136–144

[2] ADHYA S.: 'Multipartite genetic control elements: communication by DNA loop', *Annu. Rev. Genet.*, 1989, **23**, pp. 227–250

[3] MATTHEWS K.S.: 'DNA looping', *Microbiol. Rev.*, 1992, **56**, pp. 123–136

[4] SCHLEIF R.: 'DNA looping', *Annu. Rev. Biochem.*, 1992, **61**, pp. 199–223

[5] MOREAU P., HEN R., WASYLYK B., EVERETT R., GAUB M.P., CHAMBON P.: 'The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants', *Nucleic Acids Res.*, 1981, **9**, pp. 6047–6068

[6] DUNN T.M., HAHN S., OGDEN S., SCHLEIF R.F.: 'An operator at −280 base pairs that is required for repression of araBAD operon promoter: addition of DNA helical turns between the operator and promoter cyclically hinders repression', *Proc. Natl. Acad. Sci. USA*, 1984, **81**, pp. 5017–5020

[7] PTASHNE M.: 'A genetic switch: phage lambda revisited' (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2004, 3rd edn.)

[8] TOLHUIS B., PALSTRA R.J., SPLINTER E., GROSVELD F., DE LAAT W.: 'Looping and interaction between hypersensitive sites in the active beta-globin locus', *Mol. Cell*, 2002, **10**, pp. 1453–1465

[9] YASMIN R., YEUNG K.T., CHUNG R.H., GACZYNSKA M.E., OSMULSKI P.A., NOY N.: 'DNA-looping by RXR tetramers permits transcriptional regulation "at a distance"', *J. Mol. Biol.*, 2004, **343**, pp. 327–338

[10] STENGER J.E., TEGTMEYER P., MAYR G.A.: 'P53 oligomerization and DNA looping are linked with transcriptional activation', *Embo J.*, 1994, **13**, pp. 6011–6020

[11] VILAR J.M.G., SAIZ L.: 'Multiprotein DNA looping', *Phys. Rev. Lett.*, 2006, **96**, p. 238103

[12] SAIZ L., VILAR J.M.G.: 'Ab initio thermodynamic modeling of distal multisite transcription regulation', *Nucleic Acids Res.*, 2008, **30**, pp. 726–731

[13] SEMSEY S., VIRNIK K., ADHYA S.: 'A gamut of loops: meandering DNA', *Trends Biochem. Sci.*, 2005, **30**, pp. 334–341

[14] SAIZ L., VILAR J.M.G.: 'DNA looping: the consequences and its control', *Curr. Opin. Struct. Biol.*, 2006, **16**, pp. 344–350

[15] SAIZ L., VILAR J.M.G.: 'Multilevel deconstruction of the in vivo behavior of looped DNA-protein complexes', *PLoS ONE*, 2007, **2**, p. e355

[16] SAIZ L., VILAR J.M.G.: 'Stochastic dynamics of macromolecular-assembly networks', *Mol. Syst. Biol.*, 2006, **2**, p. 2006 0024

[17] GOLDSTEIN B., FAEDER J.R., HLAVACEK W.S.: 'Mathematical and computational models of immune-receptor signalling', *Nat. Rev. Immunol.*, 2004, **4**, pp. 445–456

[18] MÜLLER-HILL B.: 'The lac operon: a short history of a genetic paradigm' (Walter de Gruyter, Berlin, New York)

[19] LEWIS M., CHANG G., HORTON N.C.: 'Crystal structure of the lactose operon repressor and its complexes with DNA and inducer', *Science*, 1996, **271**, pp. 1247–1254

[20] OEHLER S., AMOUYAL M., KOLKHOF P., VON WILCKEN-BERGMANN B., MUELLER-HILL B.: 'Quality and position of the three lac operators of *E. coli* define efficiency of repression', *Embo J.*, 1994, **13**, pp. 3348–3355

[21] HILL T.L.: 'An introduction to statistical thermodynamics' (Addison-Wesley Pub. Co, Reading, MA)

[22] ACKERS G.K., JOHNSON A.D., SHEA M.A.: 'Quantitative model for gene-regulation by lambda-phage repressor', *Proc. Natl. Acad. Sci. USA, Biol. Sci.*, 1982, **79**, pp. 1129–1133

[23] SHEA M.A., ACKERS G.K.: 'The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation', *J. Mol. Biol.*, 1985, **181**, pp. 211–230

[24] ARKIN A., ROSS J., MCADAMS H.H.: 'Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells', *Genetics*, 1998, **149**, pp. 1633–1648

[25] AURELL E., BROWN S., JOHANSON J., SNEPPEN K.: 'Stability puzzles in phage lambda', *Phys. Rev. E, Stat. Nonlinear Soft Matter Phys.*, 2002, **65**, p. 051914

[26] VILAR J.M.G., LEIBLER S.: 'DNA looping and physical constraints on transcription regulation', *J. Mol. Biol.*, 2003, **331**, pp. 981–989

[27] BRAY D., DUKE T.: 'Conformational spread: the propagation of allosteric states in large multiprotein complexes', *Annu. Rev. Biophys. Biomol. Struct.*, 2004, **33**, pp. 53–73

[28] DI CERA E.: 'Thermodynamic theory of site-specific binding processes in biological macromolecules' (Cambridge University Press, Cambridge, 1995)

[29] KEATING S., DI CERA E.: 'Transition modes in Ising networks: an approximate theory for macromolecular recognition', *Biophys. J.*, 1993, **65**, pp. 253–269

[30] BORISOV N.M., MARKEVICH N.I., HOEK J.B., KHOLODENKO B.N.: 'Signaling through receptors and scaffolds: independent interactions reduce combinatorial complexity', *Biophys. J.*, 2005, **89**, pp. 951–966

[31] BRAY D., LEVIN M.D., MORTON-FIRTH C.J.: 'Receptor clustering as a cellular mechanism to control sensitivity', *Nature*, 1998, **393**, pp. 85–88

[32] DUKE T.A., BRAY D.: 'Heightened sensitivity of a lattice of membrane receptors', *Proc. Natl. Acad. Sci. USA*, 1999, **96**, pp. 10104–10108

[33] MELLO B.A., TU Y.: 'Quantitative modeling of sensitivity in bacterial chemotaxis: the role of coupling among different chemoreceptor species', *Proc. Natl. Acad. Sci. USA*, 2003, **100**, pp. 8223–8228

[34] BROACH J.R.: 'Making the right choice – long-range chromosomal interactions in development', *Cell*, 2004, **119**, pp. 583–586

[35] RADMAN-LIVAJA M., BISWAS T., ELLENBERGER T., LANDY A., AIHARA H.: 'DNA arms do the legwork to ensure the directionality of lambda site-specific recombination', *Curr. Opin. Struct. Biol.*, 2006, **16**, pp. 42–50

[36] DE LANGE T.: 'T-loops and the origin of telomeres', *Nat. Rev. Mol. Cell Biol.*, 2004, **5**, pp. 323–329

[37] ELOWITZ M.B., LEVINE A.J., SIGGIA E.D., SWAIN P.S.: 'Stochastic gene expression in a single cell', *Science*, 2002, **297**, pp. 1183–1186

[38] PAULSSON J.: 'Summing up the noise in gene networks', *Nature*, 2004, **427**, pp. 415–418

[39] ELF J., LI G.W., XIE X.S.: 'Probing transcription factor dynamics at the single-molecule level in a living cell', *Science*, 2007, **316**, pp. 1191–1194

[40] RASER J.M., O'SHEA E.K.: 'Control of stochasticity in eukaryotic gene expression', *Science*, 2004, **304**, pp. 1811–1814

[41] THATTAI M., VAN OUDENAARDEN A.: 'Intrinsic noise in gene regulatory networks', *Proc. Natl. Acad. Sci. USA*, 2001, **98**, pp. 8614–8619

[42] OEHLER S., EISMANN E.R., KRAMER H., MULLER-HILL B.: 'The three operators of the lac operon cooperate in repression', *Embo J.*, 1990, **9**, pp. 973–979

[43] BUSBY S., EBRIGHT R.H.: 'Transcription activation by catabolite activator protein (CAP)', *J. Mol. Biol.*, 1999, **293**, pp. 199–213

[44] SETTY Y., MAYO A.E., SURETTE M.G., ALON U.: 'Detailed map of a cis-regulatory input function', *Proc. Natl. Acad. Sci. USA*, 2003, **100**, pp. 7702–7707

[45] KUHLMAN T., ZHANG Z., SAIER M.H. JR., HWA T.: 'Combinatorial transcriptional control of the lactose operon of

doi: 10.1049/iet-syb:20080091

Escherichia coli', *Proc. Natl. Acad. Sci. USA*, 2007, **104**, pp. 6043–6048

[46] SAIZ L., VILAR J.M.G.: 'Efficiency and versatility of distal multisite transcription regulation' 2007. arXiv:q-bio.SC/0704.3264, (http://arxiv.org/abs/0704.3264)

[47] BORTZ A.B., KALOS M.H., LEBOWITZ J.L.: 'New algorithm for Monte-Carlo simulation of Ising spin systems', *J. Comput. Phys.*, 1975, **17**, pp. 10–18

[48] GILLESPIE D.T.: 'General method for numerically simulating stochastic time evolution of coupled chemical-reactions', *J. Comput. Phys.*, 1976, **22**, pp. 403–434

[49] ALON U., SURETTE M.G., BARKAI N., LEIBLER S.: 'Robustness in bacterial chemotaxis', *Nature*, 1999, **397**, pp. 168–171

[50] HARTWELL L.H., HOPFIELD J.J., LEIBLER S., MURRAY A.W.: 'From molecular to modular cell biology', *Nature*, 1999, **402**, C47–C52

[51] SAVAGEAU M.A.: 'Parameter sensitivity as a criterion for evaluating and comparing the performance of biochemical systems', *Nature*, 1971, **229**, pp. 542–544