Published online 25 May 2007

Nucleic Acids Research, 2007, Vol. 35, No. 11 ***e80***
doi:10.1093/nar/gkm268

General transfer matrix formalism to calculate DNA–protein–drug binding in gene regulation: application to O_R operator of phage λ

Vladimir B. Teif*

Institute of Bioorganic Chemistry, Belarus National Academy of Sciences, Street Kuprevich 5/2, 220141, Minsk, Belarus

Received February 2, 2007; Revised and Accepted April 9, 2007

ABSTRACT

The transfer matrix methodology is proposed as a systematic tool for the statistical–mechanical description of DNA–protein–drug binding involved in gene regulation. We show that a genetic system of several *cis*-regulatory modules is calculable using this method, considering explicitly the site-overlapping, competitive, cooperative binding of regulatory proteins, their multilayer assembly and DNA looping. In the methodological section, the matrix models are solved for the basic types of short- and long-range interactions between DNA-bound proteins, drugs and nucleosomes. We apply the matrix method to gene regulation at the O_R operator of phage λ. The transfer matrix formalism allowed the description of the λ-switch at a single-nucleotide resolution, taking into account the effects of a range of inter-protein distances. Our calculations confirm previously established roles of the contact CI–Cro–RNAP interactions. Concerning long-range interactions, we show that while the DNA loop between the O_R and O_L operators is important at the lysogenic CI concentrations, the interference between the adjacent promoters P_R and P_{RM} becomes more important at small CI concentrations. A large change in the expression pattern may arise in this regime due to anticooperative interactions between DNA-bound RNA polymerases. The applicability of the matrix method to more complex systems is discussed.

INTRODUCTION

Motivation

Gene regulation is governed by a number of biomolecules competing for DNA-binding sites, recognizing each other, assembling on the double helix, binding ligands on their 'backs', forming sophisticated DNA structures, etc. This picture is further complicated because DNA is tightly packed *in vivo*, and because proteins or drugs may link DNA segments separated by large distances along the sequence. Knowing the information about all molecular players and the rules of their interaction, Nature 'calculates' the transcription level for each gene. Is this biological LEGO game solvable on a computer? Let us take this as a working hypothesis.

Statistical mechanics of gene regulation

It is now believed that most of the binding events involved in gene regulation are reversible and governed by the thermodynamic equilibrium (1–3). Nowadays, high-throughput microarray technology allows us to determine thousands of thermodynamical parameters from a single experiment (4). In addition, the bioinformatics sequence analysis methods provide a way to predict the protein–DNA-binding affinities (5–7). There is a growing understanding now that a statistical–mechanical methodology is required to predict gene regulation based on this large amount of data (8–14). Some methods consider just several predefined binding sites to find a solution for comparatively simple gene regulatory systems (8). For example, the *Escherichia coli*'s lac-operator containing several binding sites for LacI and C-reactive proteins (CRPs) that may multimerize and assist DNA loop formation, can even be described analytically (9). The combinatorial regulation at a single eukaryotic enhancer is also a solvable task (15,16). More complex systems may be accessed with the help of different network approaches (17). However, once we identify all the important states (which is not a trivial task), a huge number of states, parameters and computation time make many interesting systems practically incalculable without special tricks.

Fortunately, although the binding events are very complex, everything is still centered on the DNA, which

*To whom correspondence should be addressed. Tel: +375 17 267 82 63; Fax: +375 17 267 86 47; Email: teif@iboch.bas-net.by

provides a one-dimensional template for protein binding. We show here that it is possible to describe it mathematically as a one-dimensional system even with DNA looping and multilayer protein assembly involved. One-dimensionality significantly simplifies theoretical description. Instead of jumping from node to node in the multidimensional reaction space, we screen all possible binding reactions, going in one direction along the DNA. This allows us not to skip seemingly unimportant states (e.g. non-specific binding), which gives the method more predictive power.

Lattice models for DNA-ligand binding

The principles of calculation of macromolecule binding to a one-dimensional lattice were formulated in the second part of the 20th century. The purpose of this field was initially to take into account some of the following features of DNA–protein binding: (i) binding site overlapping; (ii) competitions between different protein types, or different binding modes; (iii) site specificity determined by the DNA sequence; (iv) contact interactions between proteins bound to the DNA (e.g. when the protein is assembled from several subunits with 'sticky ends'); (v) long-range interactions (e.g. the DNA conformational transitions, changes in the DNA charge density or topology).

Several methods of solving one-dimensional lattice models have been developed in the past, including the generating functions (18,19), the transfer matrix method (20,21), the combinatorial approaches (22,23) and other modifications (24–26). In the case of non-site-specific binding, the problems of site overlapping, competitions and contact interactions may be solved analytically by any of these methods. The McGhee–von Hippel (MvH) approach is probably most widely used for the description of typical DNA–protein and DNA–drug experiments (22). However, site-specificity requires calculations according to the real polymer sequence, which rules out any analytical solutions (22,23,26). Taking into account the long-range interactions between the proteins bound to the DNA, poses additional difficulties that cannot be easily resolved by the combinatorial approaches (23,26). For example, the recent GOMER algorithm allows treating long-range interactions between a protein and a DNA promoter, but not the long-range interactions between two DNA-bound proteins (11). The generating functions method is a more general tool that has been extensively tested for many kinds of one-dimensional problems (18,19). At first, this method seemed inapplicable for the case of long-range interactions (27). Later studies have showed that the generating functions method still allows treating long-range cooperativity, but it fails if more then one type of large protein exists in the system (28). On the other hand, the transfer matrix method allows treating site-specificity (20), long-range interactions (27) and multiple binding (29). Yet there are other basic binding features such as the multilayer assembly, DNA looping, nucleosome sliding, etc. for which none of these methods have been tested. It seems from the literature analysis that only the transfer matrix method is left as a potential approach to solve the whole complexity of DNA–protein–drug lattice models in a unified systematical way. Up to now, there were no attempts to apply this method to the biophysical characterization of complex gene regulatory systems. On the other hand, a complementary field of mathematical analysis of DNA sequences now actively uses matrix methods. This provides an additional argument for choosing the transfer matrix formalism as a general systematic tool.

The legacies of Markov and Ising

At this point, we have to make several methodological comments. All models for DNA-ligand binding mentioned above belong to the class of the so-called Ising models. In his doctoral thesis in 1924, Ernest Ising was studying ferromagnetism and introduced the model of a linear chain of magnetic moments, which are only able to take two positions, 'up' and 'down', and which are coupled by interactions between the nearest neighbors. Later the Ising model became popular in many fields of physics. Naturally, when a number of physicists moved to biophysics inspired by Schrödinger's definition of DNA as a one-dimensional aperiodic crystal, they brought the Ising model to the new field, in particular to the study of DNA melting and DNA-ligand binding (24). At that time, the field of bioinformatics did not yet exist.

When bioinformatics emerged later, it was mostly driven by the biologically inspired mathematicians who came with their own concepts such as the Markov chains. In the 1920s, a pioneer of cybernetics, Norbert Wiener, performed a first rigorous study of a continuous Markov process. This work and the later Kolmogorov's probability theory popularized the ideas of the 19th-century mathematician Andrei Markov, who studied the sequences of random variables in which the future variable is determined by the present variable but is independent of the way in which the present state arose from its predecessors. The Markov chains are now widely used in bioinformatics, in particular in the DNA sequence analysis (5–7, 30–32).

Both the Markov chains and the Ising lattices may be formulated with the help of the transfer matrices containing the probabilities of transition of a system between different states. Consequently, the Ising model may be converted into the Markov model, and vice versa. The general transfer matrix formalism is evidently the ancestor of both the Markov chains and the Ising lattices. However, the differences between the transfer matrices employed in the biophysical and bioinformatical studies of DNA–protein interaction are not just historical. Bioinformatics is mostly interested in DNA sequence analysis, motif finding, etc. Therefore the different states in the Markov chains are either (A, T, G, C) or the occurrences of dinucleotides, trinucleotides, etc. or some other 'words' composed from the nucleotide dictionary (5–7, 30–32). On the other hand, in the biophysical models, the DNA sequence is fixed, and the different states are 'free' or 'bound' depending on the presence or absence of a protein at a given DNA site (19–29). In the matrix models considered below, the states

are also divided into 'free' or 'bound', but these states are further subdivided into a number of microstates allowing us to treat complex models of DNA–protein–drug interaction. Hopefully our work will help to join the efforts of biophysics and bioinformatics in the description of gene regulation.

In the current work, we provide the unified matrix formalism to calculate the multiprotein assembly on the DNA. We consider the most important experimental features—the site-specificity, competitions, multilayer binding, nucleosomes and DNA loops—and show that all these essential constituents of gene regulation are computable using the transfer matrix formalism. Then we test the method on gene regulation of phage λ. The matrix formalism not only allowed a correct description of this well-documented system, but also suggested new biologically relevant predictions.

GENERAL METHODOLOGY

The idea of the transfer matrix method is to consider the DNA molecule as a 1D lattice of units, each unit being characterized by a matrix of statistical weights corresponding to all its possible states (20,29). The matrix of statistical weights is called the 'weight matrix' or the 'transfer matrix'. The weight matrix depends on the DNA sequence and the chosen model of DNA–protein and protein–protein interactions. The partition function is given by the product of the matrices corresponding to all DNA units. The probabilities of the binding events may be calculated from the partition function. The general methodology consists of choosing the elementary DNA unit, enumerating all its possible states, constructing the corresponding transfer matrices, applying the boundary conditions and finally calculating the maps of binding or the binding curves.

Choosing the elementary unit

Let us consider the DNA molecule as a linear lattice of N units numbered by index n, $n = 1 \ldots N$ (Figure 1A). In the case of independent ligand binding (non-interacting binding sites) it is convenient to choose the elementary unit coinciding with the binding site (29). However, when the binding is sequence-specific and the binding sites may overlap, one may use some physically distinguished units instead (the nucleotide, base pair, nucleosome, etc.). Throughout this article, we will assume that the elementary unit is the base pair.

States enumeration

We have to list all available states for each elementary DNA unit. This may be done in several ways. If we forget a state, this will lead to an error in the partition function. If we enumerate a state more than once, this will add unnecessary parameters and increase the computation time. Therefore, it is important to find the shortest complete number of states, R. The transfer matrix will then contain $R \times R$ elements corresponding to all possible combinations of states of a given unit and its nearest neighbor.

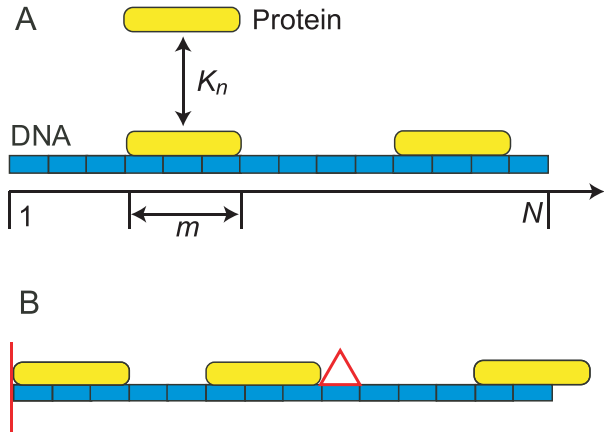

Figure 1. Schematic description of the method. (**A**) The DNA is shown as a 1D lattice of units numbered by index n, $n = 1, \ldots N$. Protein binding to the m DNA units starting at unit n is characterized by the binding constant K_n. Each unit is assigned the transfer matrix Q_n which consists of statistical weights $Q_n(i, j)$ giving the probabilities for unit n to be in state i provided the unit $n+1$ is in state j. (**B**) The boundary conditions are applied to the m units at the DNA ends and close to the other physical obstacles.

Transfer matrix construction

The transfer matrices Q_n consist of the elements $Q_n(i, j)$ equal to statistical weights corresponding to the nth DNA unit being in state i followed by the next unit in state j. The physical meaning of each element of the weight matrix is the conditional probability of having the unit n in state i provided the next unit is in state j. Evidently only several combinations of states i and j are allowed. For example, although the binding sites might overlap, the bound proteins cannot overlap if they are in the same layer. The allowed states are characterized by statistical weights given as a combination of the concentrations and energetic parameters. The prohibited states are characterized by zero statistical weights.

Boundary conditions

The transfer matrices constructed at the previous stage are the 'regular matrices' corresponding to the DNA units far from any obstacles. All regular matrices keep the same locations of zero elements. Close to the DNA ends or close to the physical obstacles (Figure 1B) the transfer matrices change according to the boundary conditions. Previous studies have showed that boundary conditions may set strong constraints on sequence-specific target location by proteins on DNA (33–35). Our calculations confirm this conclusion. In general, protein hanging out from DNA ends may be either prohibited or allowed. In the former case, the transfer matrices corresponding to the DNA ends have more zero elements then the regular matrices. If protein hanging out from DNA ends is allowed, then the matrix mask is not changed, but the binding constants depend on the distance from the DNA end. The special boundary conditions may be also set inside the DNA segment.

Calculation of the binding probabilities

The final output of the calculations is either the equilibrium protein distribution along DNA (the map of binding), or the dependencies of the binding probabilities on protein concentrations (the binding curves). The intermediate results necessary for these calculations are the partition function and its derivatives. The standard expression for the partition function Z of a linear lattice of N units is given by Equation (1) (20).

$$Z = (1 \quad 1 \quad \ldots \quad 1) \times \prod_{n=1}^{N} Q_n \times \begin{pmatrix} 1 \\ 1 \\ \ldots \\ 1 \end{pmatrix} \qquad \mathbf{1}$$

If we deal with a homopolymer, it is possible to diagonalize the matrices and turn matrix multiplication into multiplication of the diagonal elements. However, in the case of gene regulation the DNA sequence specificity is a significant feature, and this trick will not work out. All our matrices are different and should be multiplied according to the sequence of the DNA units. The straightforward partition function calculation according to Equation (1) leads to extensive matrix–matrix manipulations. It is easier to replace it by the vector-matrix multiplication and calculate the partition function Z and its derivatives recursively according to Equations (2) and (3) (21).

$$Z = A_N \times \begin{pmatrix} 1 \\ 1 \\ \ldots \\ 1 \end{pmatrix}, \quad A_i = A_{i-1} \times Q_n, \quad A_0 = (1 \quad 1 \quad \ldots \quad 1) \qquad \mathbf{2}$$

$$\frac{\partial Z}{\partial X} = \frac{\partial A_N}{\partial X} \times \begin{pmatrix} 1 \\ 1 \\ \ldots \\ 1 \end{pmatrix}, \quad \frac{\partial A_n}{\partial X} = \frac{\partial A_{n-1}}{\partial X} \times Q_n + A_{n-1} \times \frac{\partial Q_n}{\partial X}, \quad A_0 = (1 \quad 1 \quad \ldots \quad 1) \qquad \mathbf{3}$$

Here X is any parameter explicitly entering at least one of the statistical weights. Since we have all elements of the matrix Q_n in the analytical form, the matrix $\partial Q_n/\partial X$ may be found analytically. This saves computational time and decreases errors in the numerical calculations.

The probability of a given state is, as with all partition functions, the term in the partition corresponding to that state divided by the sum of all of the terms (24). Suppose we have a parameter X uniquely entering the statistical weight of a given state of a given DNA unit. Then differentiating the partition function by X will filter all configurations of the system, which contain a given DNA unit in a given state. If X enters the partition function linearly, then the probability of this state is equal to the corresponding derivative of the partition function, multiplied by X and divided by the partition function. In particular, the probability that the nth DNA unit is covered by the protein of type g is given by Equation (4):

$$c_{ng} = \frac{\partial Z}{\partial K_{ng}} \times \frac{K_{ng}}{Z} \qquad \mathbf{4}$$

Here K_{ng} is the binding constant for a protein of type g binding to the DNA site starting at the nth unit. The whole set of c_{ng} values determines the complete map of protein binding to DNA. Having in mind that the proteins may also assemble on DNA in a multilayer fashion, the term 'map of binding' may be generalized to a multi-dimensional plot giving the probabilities for all binding events.

The binding curve c_g gives the average degree of protein g binding to DNA:

$$c_g = \frac{1}{N} \sum_{n=1}^{N} c_{ng} \qquad \mathbf{5}$$

When protein binding to DNA is coupled to other reactions in the solution (e.g. protein dimerization, modification, activation, etc.), the equilibrium concentrations c_{0g} should be found from the corresponding laws of mass action. In most cases, a small number of proteins bound to DNA site-specifically does not affect the bulk concentration of free proteins. On the other hand, the non-specific binding to DNA should be taken into account to correct the concentration of free proteins. *In vivo*, a large fraction of regulatory proteins binds DNA non-specifically (36). The non-specific binding may be viewed as one of the simple equilibriums to be solved before calculating the maps of binding. Once the effective concentrations of free proteins in the solution have been determined, the maps of protein binding to DNA may be calculated from the matrix method. Thus, the calculation of the protein arrangements on the DNA may be decoupled from the simple equilibrium calculations in solution. While calculating the maps of binding, we treat all binding events as sequence-specific, including the non-specific binding.

CONSTRUCTION OF THE BASIC MATRIX MODELS

The methodology described in the previous section allows construction of the weight matrices for different models of multimolecular interaction. Now we have to show that it is possible to split the complex gene regulatory processes into a number of well-defined physical events solvable using the transfer matrix formalism. Firstly, we extract the basic binding features that may be used as the building blocks for more complex interaction models. Figure 2 summarizes the models, which we solve in the current section using the transfer matrix formalism. The other models may be constructed either as derivatives or as combinations of the basic ones.

Sequence-specific binding

Let us first consider a single-protein sequence-specific binding to DNA (Figure 2A). The protein may be either bound or not. If the protein covers m DNA units upon

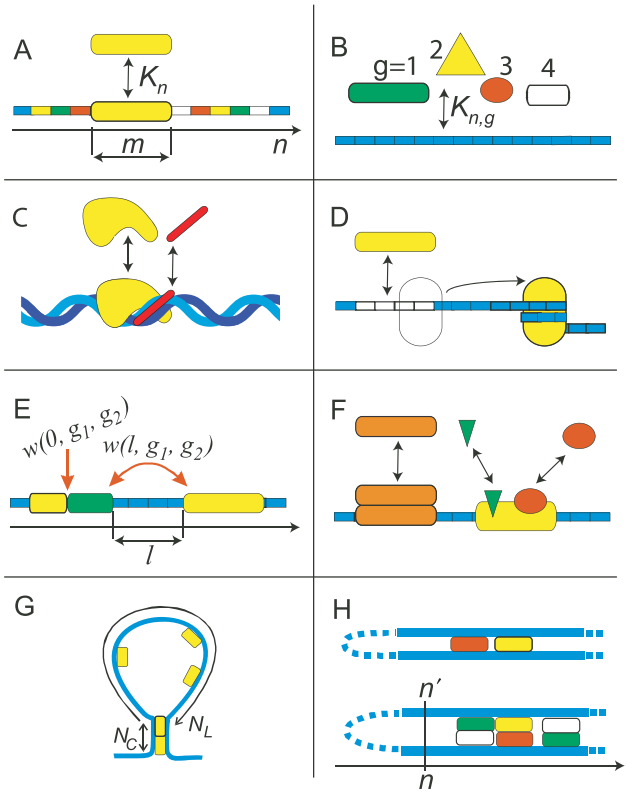

Figure 2. The schemes of the basic binding models. (**A**) Sequence-specific binding of a single protein. (**B**) Competitive binding of several protein spices or several modes of binding of the same protein. (**C**) Protein–small drug competition. (**D**) Protein–nucleosome competition. (**E**) Cooperative binding (includes contact interactions between the proteins bound to adjacent DNA units, and long-range interactions between the proteins separated by l DNA units). (**F**) Multilayer binding (includes piggy-back binding of small ligands on the backs of DNA-bound proteins, and the multilayer assembly of proteins of similar size). (**G**) Small DNA loops induced by protein cross-linking. (**H**) Large DNA loops maintained by protein bridging.

binding, then each unit may be in $R_A = m+1$ states. Let us number the states of the DNA unit by index i. States $i = 1, \ldots, m$ correspond to the DNA unit being covered by the protein, depending on where the protein starts respectively to the given DNA unit, from the protein's left ($i=1$) to its right ($i=m$) end. State $i=m+1$ corresponds to the free DNA unit.

The allowed combinations of states of the nth unit and the $(n+1)$th unit correspond to the following non-zero elements of the transfer matrix $Q_n(i, j)$: the protein starts at the nth DNA unit ($i=1$, $j=2$), the protein starts before the nth unit and covers the nth unit ($i=2 \ldots m-1$, $j=i+1$), the protein ends at the nth unit and is followed by a free unit ($i=m$, $j=m+1$), the protein ends at the nth unit and is followed by the next protein ($i=m$, $j=1$), and the nth unit is free from proteins ($j=i=m+1$). The other elements of the transfer matrix are equal to zero.

We want to deal with the binding constants available from experiments. These values cannot be easily localized among the multiple protein contacts with individual DNA units. Therefore, we assign the whole energy of protein–DNA binding to the first protein contact with the DNA. In the case of a single-protein binding this corresponds to the matrix element $Q_n(1, 2)$. We set $Q_n(1, 2) = K_n \times c_0$, where K_n is the binding constant for the frame of m DNA units starting at the nth unit, c_0 is the molar protein concentration. All the rest non-zero matrix elements are units. Equation (6) shows an example of the transfer matrix constructed according to this algorithm for $m=3$.

$$Q_n = \begin{pmatrix} 0 & K_n c_0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \quad \mathbf{6}$$

Similar weight matrices have been used in the early DNA-ligand studies (20). Unlike this simple example, below we will deal with large matrices that are not easy to represent in a journal page. The site-specificity of binding will be preserved in the same way in all our following models.

Competitive binding

Competitive binding (Figure 2B) is another basic feature, which is important for many experimental systems. Competitions of several types may be encountered in gene regulation. Two most commonly used models are the competition of different proteins and the competition of different modes of binding of the same protein. Although the biology is different, the transfer matrices for these two competitive models are described by the same mathematics.

Let the protein binding to DNA be characterized by f different complexes numbered by index g, $g = 1, \ldots, f$. Each protein–DNA complex of type g involves m_g DNA units. Then each DNA unit may be in R_B states given by Equation (7):

$$R_B = m_1 + \ldots + m_f + 1 \quad \mathbf{7}$$

The detailed states enumeration for this model and the algorithm of transfer matrix construction is given in the Supplementary Data. Analogously to the single-protein matrix, we describe the binding of a protein of type g to a frame of m_g DNA units starting at unit n, by the statistical weight $K_{ng} \times c_{0g}$. Here K_{ng} is the binding constant, and c_{0g} is the bulk concentration of g-type protein. The number of protein types may be less then the number of types of the protein–DNA complexes. In the case of the competition between different binding modes of the same protein the concentrations c_{0g} are the same.

Small drugs

Many anticancer and antimicrobial drugs exert their activity through direct DNA binding. Leaving aside the compounds that induce covalent DNA modifications, let us look here at the reversible minor groove binders. This is a wide class of drugs such as netropsin, distamycin and their derivatives, many of which bind DNA sequence-specifically (37). A number of potential drugs are now

being developed based on synthetic polyamides, recognizing the regulatory protein-binding sites on DNA, and thus interfering with gene regulation. This field has opened several years ago and is already becoming pharmaceutically important (38). For example, it was shown recently that sequence-specific polyamides alleviate the transcription inhibition associated with long GAA·TTC repeats in Friedreich's ataxia (39).

The competition of proteins and small drugs may be better described by the allosteric competition model shown in Figure 2C rather then the simple competition in Figure 2B. The allosteric competition model allows 'overlapping' of molecules bound to DNA. This may be realized if a small drug slides along the minor groove, while the protein slides along the major groove (40). The drug changes the DNA conformation, widening the minor groove and narrowing the major groove. In this case, the states enumeration of the basic competitive model (Table S1) will be changed. The description of the states should now indicate whether a unit is bound to a small drug in the minor groove and whether it is bound to a protein in the major groove. The change in the algorithm for transfer matrix construction is also straightforward. Those states that include the nearby binding of both the protein and the drug are charged an additional allosteric cooperativity constant.

Nucleosomes

In eukaryotes, many regulatory sites are covered by the nucleosomes. A nucleosome may free DNA in two ways. The first possibility is that the nucleosome dissociates from the DNA and goes to the solution. This situation may be described by the basic protein–protein competition model (Figure 2B). The second possibility is that the nucleosome slides along the DNA to a new position without being completely dissociated (Figure 2D). The nucleosome binding is site-specific as well as the protein binding. It is known experimentally that there are specific DNA sequences that are more bendable and phased to favor nucleosome binding (41,42). The nucleosomes are probabilistically positioned in the chromatin according to the DNA sequence. Moving a nucleosome along the DNA requires some work. All nucleosomes are the same with respect to DNA binding, and therefore the assignment of the nucleosome-binding constants from the DNA sequence analysis is even simpler and more effective then in the general DNA–protein case.

When protein binding is strong enough to compensate for the nucleosome dissociation energy, it will result in the competition of a protein and a nucleosome octamer described by a standard competitive model [Equation (7)]. When protein binding is weak, the protein–nucleosome competition will result in the nucleosome sliding. In the latter case, the order and the total number of nucleosomes on the DNA is fixed, and the competitive model should be modified to keep track of all nucleosomes within a given DNA sequence. Due to a large number of states, such a model is hardly applicable to the whole-genome studies, but it may be important for short DNA segments containing a regulatory sequence covered by just several nucleosomes. In such a system, proteins help each other bind DNA by displacing nucleosomes; hence, their 'collaborative competitions' may cause experimentally observable binding cooperativity (43). Apart from nucleosomes, the sliding/dissociation model may be also useful for a broad class of ring-like proteins that assemble on the double helix and slide along DNA, e.g. helicases.

Cooperative protein assembly

Let us consider the cooperative multiprotein binding within a single layer along DNA (Figure 2E). This is a conventional class of models, which includes both the short- and long-range interactions between the proteins bound to the DNA (44). The proteins may interact either through direct contacts or through the DNA. The binding of proteins on the backs of other proteins is not allowed here (see below for the multilayer binding).

Contact cooperativity. The contact cooperativity is easy to incorporate in the basic competitive matrix considered above. Since we have already defined the matrix elements for the protein–protein contacts, we just have to set additional statistical weights for these elements. These values are usually called the contact cooperativity parameters. They are denoted in our formalism as $w(0, g_1, g_2) = \exp(\varepsilon(0, g_1, g_2)/RT)$, where $\varepsilon(0, g_1, g_2)$ is the free energy of a contact between the proteins of types g_1 and g_2 bound to the adjacent DNA units, $RT \approx 0.6$ kcal.

The f-mer assembly model. The simplest and most commonly used model of the contact protein–protein interactions is known as the McGhee-von Hippel (MvH) cooperativity (22). This model corresponds to the situation when each of the bound proteins may interact with two of its nearest neighbors. A more general case is the multiprotein assembly on the DNA forming f-mer complexes with f ranging from two (dimer assembly) to infinity. Many proteins involved in gene regulation through f-mer assembly lay between the two extreme cases of the classical MvH contact cooperativity (which may lead to f-mers of infinite size, e.g. RecA assembly into large filaments along DNA), and the pairwise cooperativity (which leads to f-mers composed of two proteins, e.g. homodimerization). An example of the pairwise cooperativity is the phage λ repressor CI. This protein binds DNA in a dimeric form. The CI dimers bound to adjacent DNA sites recognize each other due to direct protein–protein contacts formed by their C-terminal domains. Each CI dimer may interact only with one nearest neighbor. Once a pair of dimers is formed, a third dimer binds DNA non-cooperatively (45,46). Such interactions may be described by the extended MvH model, where two different orientations of the bound proteins are taken into account (47,48). The f-mer assembly model provides a further generalization.

In the general case of f-mer assembly with asymmetrical interactions, we distinguish f different complex types, even if the proteins are identical. The first protein in f-mer, the second protein in f-mer, ..., the last protein in f-mer. Each protein-DNA complex (not each protein) should be

assigned its contact cooperativity parameters for interactions with other complexes. The matrix formalism is thus very convenient for this model. The problem of the hetero-*f*-mer assembly is treated in the same way as the homo-*f*-mer assembly. In particular, for the MvH cooperativity we use a single complex type ($f=1$) for each protein.

Long-range interactions. A further generalization of the cooperative models is to take into account the interactions beyond direct protein–protein contacts. There are several possibilities to take this into account in the matrix formalism (27,49). Let us assume that the interaction of the proteins of types g_1 and g_2 depends on the distance l between them. Analogously to the contact interactions, we characterize it by the cooperativity constants $w(l, g_1, g_2) = \exp(\varepsilon(l, g_1, g_2)/RT)$. The contact interactions model is thus a particular case of long-range interactions with $l=0$. The states enumeration includes now V_g states for a DNA unit being between bound proteins, where V_g is the maximum length of g-type protein interaction. Three additional states correspond to the units belonging to the left and right free DNA ends, and the units between non-interacting isolated proteins (Supplementary Data has the detailed states enumeration). Each DNA unit may be in R_E states:

$$R_E = \sum_{g=1}^{f} (m_g + V_g) + \max(V_g) + 3 \quad \textbf{8}$$

How large is the interaction range *in vivo*? The DNA conformational transitions or changes in the DNA charge distribution induced by protein binding may propagate for up to several dozens of base pairs (50). On the other hand, the changes in DNA topology or DNA looping may cover very large distances. In the worst case, long-range interactions involve the whole DNA molecule. The problem with long-range interactions in the matrix formalism is that the number of states R_E increases linearly with the maximum interaction length $\max(V_g)$. A straightforward calculation for $\max(V_g)=300$ takes several hours on a Pentium M 725 laptop. Special algorithms for sparse matrix handling allow accelerating the calculations.

Multilayer binding

Although DNA provides a one-dimensional template for protein binding and we are using the one-dimensional mathematics, the binding events are not confined to one dimension (Figure 2F). Proteins may form multilayer structures on the DNA. Once bound to the DNA, a protein by itself may provide a lattice for new binding events. Proteins may bind other proteins and small ligands such as metal ions and ATP. One important case belonging to this class of models is the 'piggy-back' binding model which was initially formulated for the DNA-dependent ATPase of DNA gyrase (51). This model describes binding of small ATP 'riders' to the 'backs' of the DNA-bound proteins. Another important example of the vertical assembly is the multimerization of the proteins of comparable size with the possibility of formation of bridge-like structures. For example, a CI dimer bound to the DNA may bind another CI dimer on its 'back' to form a tetramer (12).

Let each protein may form from zero to *fff* additional 'vertical' complexes with other proteins or drugs. Then each DNA unit may be in R_F states:

$$R_F = (fff + 1) \times R_E \quad \textbf{9}$$

The algorithm for transfer matrix construction is analogous to the algorithm for the competitive binding model (Supplementary Data). The second layer of proteins is characterized by the binding constants in the same way as the convenient single layer protein–DNA binding.

DNA loops

DNA looping is an essential component of gene regulation (1,12). DNA loops *in vivo* may range from tens or hundreds of nucleotides in the case of the proximal promoters, to thousands in the case of the distant enhancers and to millions in the case of the chromosome structure maintenance elements. Our calculations indicate that the loops larger than 1000 units are practically incalculable in the frame of the long-range interactions model. Thus, we divide the loops into two computable classes. Small loops (Figure 2G) may be calculated taking into account the long-range interactions between all the units. Large loops cannot be calculated like this. However, the behavior of a very large loop depends mainly on the binding events at the segments brought together by the protein crosslinks, since the system 'forgets' what happens deep inside the loop (Figure 2H). Between the small and large loops, stands a class of intermediate loops that are difficult to calculate in the frame of the matrix formalism. Fortunately, most of the loops involved in gene regulation belong to one of the two calculable classes of 'small' or 'large' loops. For example, one may encounter sophisticated loops of intermediate classes in single-stranded RNA folding (52), but hardly in gene regulatory systems of our interest where large transcription factors bind stiff double-stranded DNA.

Small loops

In many cases, gene regulatory events are limited to a relatively small unpacked DNA segment accessible for the regulatory proteins. In this case, only simple loops shown in Figure 2G may be formed. Let the DNA unit may be either inside the loop, or outside of the loop. The looped segment consists of the 'necklace' (the two polymer segments brought together by protein crosslinks) and the 'loop' itself (the polymer units between the starting and ending crosslinks, numbered according to their position in the loop). Let the largest loop consist of N_L units, and the longest necklace has N_c crosslinks. Then for a situation shown in Figure 2G, a DNA unit may be in R_G states:

$$R_G = 6 + 2 \times N_c + 2 \times N_L \quad \textbf{10}$$

The first crosslink starting a loop of length l is assigned an additional statistical weight $w(l)$. For a loop much larger then the DNA persistence length, the statistical

weight may be taken in the general form $w(l) = \text{const} \times l^{-\alpha}$ (49). For smaller loop lengths we may either use the tabulated experimental values (12,56), or try to calculate protein-dependent DNA looping on the basis of a rigorous statistical mechanics (54). In the small loops model, we do not assign the loop length before calculations. Knowing the sequence and the binding constants, we may predict, which loop configuration is the most probable one. However, as mentioned above, this model poses large computational difficulties, which are a hard nut to crack for the loop lengths larger than several hundreds of units.

Large loops

There are several reasons for distinguishing large loops from the previous case. First, when the loop is much larger then the persistence length, the energy of its formation almost does not depend on the local bending, twisting, etc. as in the case of small loops. Second, due to the compact DNA packing *in vivo*, large loops are also not that sensitive to the entropic contributions since the large loop is not given the whole space it would require for an arbitrary set of conformations. The loop may be formed only at a specific predefined position, compartmentalized both in the 3D space of the chromatin and in the 1D space of the genome sequence. The last but not the least argument is that the class of large loops is actually abundant in gene regulation (55).

The large loops may be formed by different mechanisms: either by a single protein cross-linking two DNA segments, as in the case of LacI repressor of *E. coli* (Figure 2H, top), or by a multilayer assembly of several proteins creating a 'bridge' between two DNA segments, as in the phage λ CI repressor (Figure 2H, bottom). In both cases, the loop is formed by connecting the DNA regions that have specific affinity for the cross-linking proteins. The two DNA sequences are predefined for a potential loop closure. We may still consider this 'sandwich' as a 1D system. Since we know where the complementary bottom and top segments start and end, the unit n belonging to the bottom DNA segment uniquely determines the unit n' in the top DNA segment in front of the unit n (Figure 2H, bottom).

Let us construct the enumeration of states of a DNA unit based on the concept of multiple protein layers. The states of the DNA unit indicate whether a protein is bound at a given position in each layer. The crosslink between the two DNA segments is formed when proteins fill the corresponding vacant places in both layers. The state corresponding to the first crosslink is assigned an additional statistical weight characterizing the probability of the loop formation ($w_{\text{loop}} = \exp(-\triangle G_{\text{loop}}/(RT))$, where $\triangle G_{\text{loop}}$ is the energy of the DNA loop formation). The states corresponding to the protein of type g bound at the first layer are characterized by the binding constants K_{ng}. The l-bp gap between the proteins g_1 and g_2 belonging to the same layer is assigned a statistical weight $w(l, g_1, g_2)$ as in the long-range interactions model. The vertical contact between the proteins g_1 and g_2 belonging to the first and second layers is assigned a statistical weight $w_{\perp}(g_1, g_2)$. The contacts of the second layer proteins with the DNA are characterized by the binding constants $K_{n'g}$. When we have more then two protein layers, the description is analogous. The detailed states enumeration for the large loops model is given in Table S3 (Supplementary Data). A double-layer loop shown in Figure 2H, is characterized by R_H states for a DNA unit:

$$R_H = 6 \times R_E + 2 \quad \textbf{11}$$

The behavior of the DNA loops has been analyzed in detail for the case of non-specific protein binding (2,56,57). The general feature of the models G and H in Figure 2 is the following. For a large number of potential cross-linking sites, DNA looping occurs abruptly at a critical protein concentration. The loop is stable in a large interval of protein concentrations. At very high concentrations, DNA again unloops because the multiprotein structures are assembled at both DNA segments instead of joining two DNA segments. In the case of site-specific binding, the system bears these general non-specific features, but also provides a possibility of a delicate combinatorial control of gene regulation, as we will see in the next section.

Thus, we have solved all basic models in Figure 2, and showed how to construct the other models as modifications or combinations of the basic ones. Now let us perform calculations for a concrete genetic system.

CALCULATION OF GENE REGULATION AT O_R OPERATOR OF PHAGE λ

The binding events at O_R operator of bacteriophage λ control the famous genetic switch from the lysogenic state (when λ peacefully lives inside the infected *E. coli*) to the lytic state (when λ duplicates itself in a large number of copies leading to the lysis of the host cell). Two regulatory proteins, the CI and Cro repressors, act at O_R operator. In the lysogenic state, the *cro* gene coding the Cro protein is ~95% suppressed, while the *cI* gene coding the CI protein is on. The CI protein aims to maintain its own expression and to switch off all other genes. CI domination determines that the phage is in the lysogenic state. When the host cell is damaged or irradiated, its SOS system activates RecA protein that stimulates self-cleavage of CI. This leads to induction of the lytic state of phage λ. In the lack of CI, Cro dominates at O_R, maintaining its own expression, switching on early lytic genes, and suppressing the *cI* gene (46).

The CI and Cro proteins homodimerize in solution due to C-terminal domains and bind DNA as dimers using a helix-turn-helix motif in the N-terminal domains. The dimers may adsorb on DNA non-specifically (36) or bind sequence-specific sites (45). The specific binding constants are orders of magnitude larger then the non-specific ones. The CI and Cro dimers cover 17 bp upon binding to DNA. The structure of O_R operator is shown in Figure 3. The O_R operator consists of three 17-bp specific binding sites for Cro and CI proteins, enumerated O_R1, O_R2 and O_R3. Each site may bind Cro and CI with different affinities. The O_R site overlaps with the

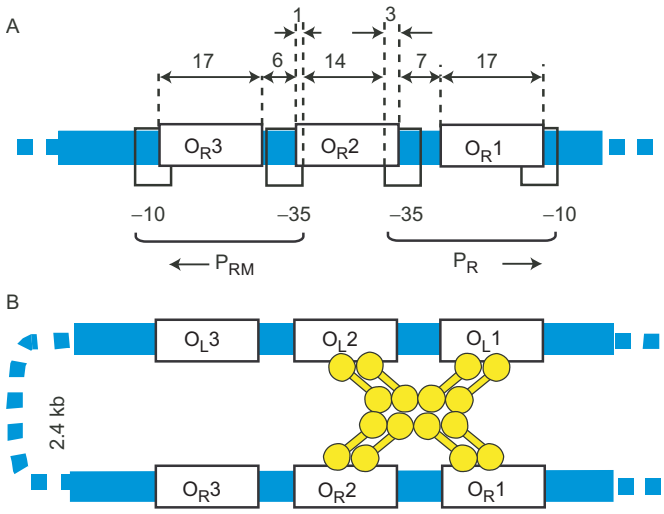

Figure 3. (**A**) The scheme of transcription regulation at O_R operator of phage λ. The regulatory proteins Cro and CI bind DNA in dimeric form, covering 17 bp upon binding. The O_R operator consists of three binding sites, O_R1, O_R2 and O_R3. RNAP covers 35 bp upon binding to the promoters P_{RM} and P_R. P_{RM} overlaps with O_R3. P_R overlaps with O_R1 and O_R2. (**B**) The CI dimers may assemble in two layers to form tetramers and/or octamers bridging the O_R and O_L operators separated by ~2.4 kb.

promoters P_{RM} and P_R. The σ subunit of the RNA polymerase (RNAP) binds to the –10 and –35 recognition regions at the promoter, making ~35 bp inaccessible for binding by other proteins. RNAP binding to P_{RM} starts transcription of CI protein (direction to the left from O_R in Figure 3A). RNAP binding to P_R starts transcription of Cro and other yearly lytic proteins (to the right from O_R). P_R overlaps with O_R1 and O_R2. Thus, binding of repressor proteins to any of O_R1 and O_R2 sites precludes RNAP binding at P_R. P_{RM} overlaps with O_R3 and borders upon O_R2. RNAP bound at P_{RM} contacts with CI dimer bound at O_R2. This contact activates transcription from P_{RM}. The Cro–Cro and CI–CI contacts are also energetically favorable (45).

The binding scheme shown in Figure 3A allows 40 distinguishable arrangements of three proteins, Cro, CI and RNAP among three binding sites O_R1, O_R2 and O_R3 (45). Several years ago it seemed natural that this picture completely describes the regulatory events at O_R operator (58,59). However, then it was shown that it should be further complicated to take into account the interaction of O_R with another operator O_L situated ~2.4 kb from O_R. O_L may be linked to O_R thought DNA looping due to bridging by CI proteins (Figure 3B). The O_L operator consists of three binding sites O_L1, O_L2, O_L3 similar to O_R, and overlaps with promoter P_L symmetrical to P_R.

Most of the binding energies for this system are known from the experiments. The energies of non-specific binding of Cro and CI are –4.2 and –4.1 kcal/mol correspondingly (36). Cro binds O_R1, O_R2 and O_R3 sites with the energies –12.0, –10.8 and –13.4 kcal/mol (45). CI binds O_R1, O_R2 and O_R3 sites with the energies –12.5, –10.5 and –9.5 kcal/mol correspondingly (45).

The cooperativity of repressor binding

Let us first look into the competitive cooperative binding of Cro and CI at the region of lambda-phage (λ-phage) sequence 37 930–38 030 containing the O_R operator and the flanking regions including P_{RM} and P_R promoters (58). In order to simplify the system, in the first series of calculations (Figures 4–6) we are considering a 'mutant' without the spacers between the O_R1, O_R2 and O_R3 binding sites shown in Figure 3A. (A detailed treatment of the intact lambda (λ) sequence will be provided later in Figure 7). On the other hand, a small 3-bp overlapping of O_R1 site and P_R promoter is surely critical for the behavior of the system, and we have to consider this in all calculations.

Figure 4 shows the maps of binding calculated for the set of energies chosen above and the concentrations characteristic to the lysogenic state of phage lambda (phage λ): [CI] = 0.2 μM, [Cro] = 0.02 (60). Figure 4A shows how the system of two regulatory proteins, Cro and CI would behave in the absence of cooperativity. Since Cro-binding constants for O_R2 and O_R3 are larger then CI binding constants, Cro dominates at these sites when Cro and CI are at comparable concentrations. That is not what is observed in the experiments where Cro domination would mean the end of the lysogenic state.

Figure 4B shows the results of the calculations for the same system taking into account protein–protein interactions in the MvH model (22). The proteins interact with their left and right neighbors symmetrically. The energies of CI–CI and Cro–Cro interactions are set as –2.5 and –0.3 kcal/mol correspondingly. These values are in line with the experimental data (45). Taking into account the MvH cooperativity dramatically changes the non-cooperative map of binding, the O_R1 and O_R2 sites are now covered by CI proteins with almost the same probability, while only O_R3 is left for Cro binding. This is already closer to the experimental situation in the lysogenic state of phage λ. However, now the cooperativity is 'too crude': CI proteins not only bind O_R1 and O_R2, but also tend to occupy the adjacent promoter regions, which should be the targets for RNAP. A small but not negligible Cro presence at O_R2 is also not compatible with the discrete nature of the lambda-switch (λ-switch).

Experimentally, the interactions between CI dimers are not symmetric as in Figure 4B. CI dimers bound to adjacent DNA sites interact with each other due to a direct contact between C-terminal domains. There is only one recognition domain per dimer and hence, once a pair of interacting dimers is formed, the third dimer binds the adjacent site non-cooperatively (Figure 4C). On the other hand, Cro–Cro interactions are weaker and do not show such asymmetry. Darling and coauthors (45) determine different energies of Cro–Cro interaction for the proteins bound to O_R1–O_R2 and O_R2–O_R3 sites. However, most of the literature treats Cro–Cro interactions as symmetric and does not provide any structural information for the asymmetry as in the case of CI dimers (46). The different cooperativity for different Cro-binding sites may just reflect the fact that the interactions are length-dependent (the O_R1–O_R2 and O_R2–O_R3 spacers are equal to 7 and

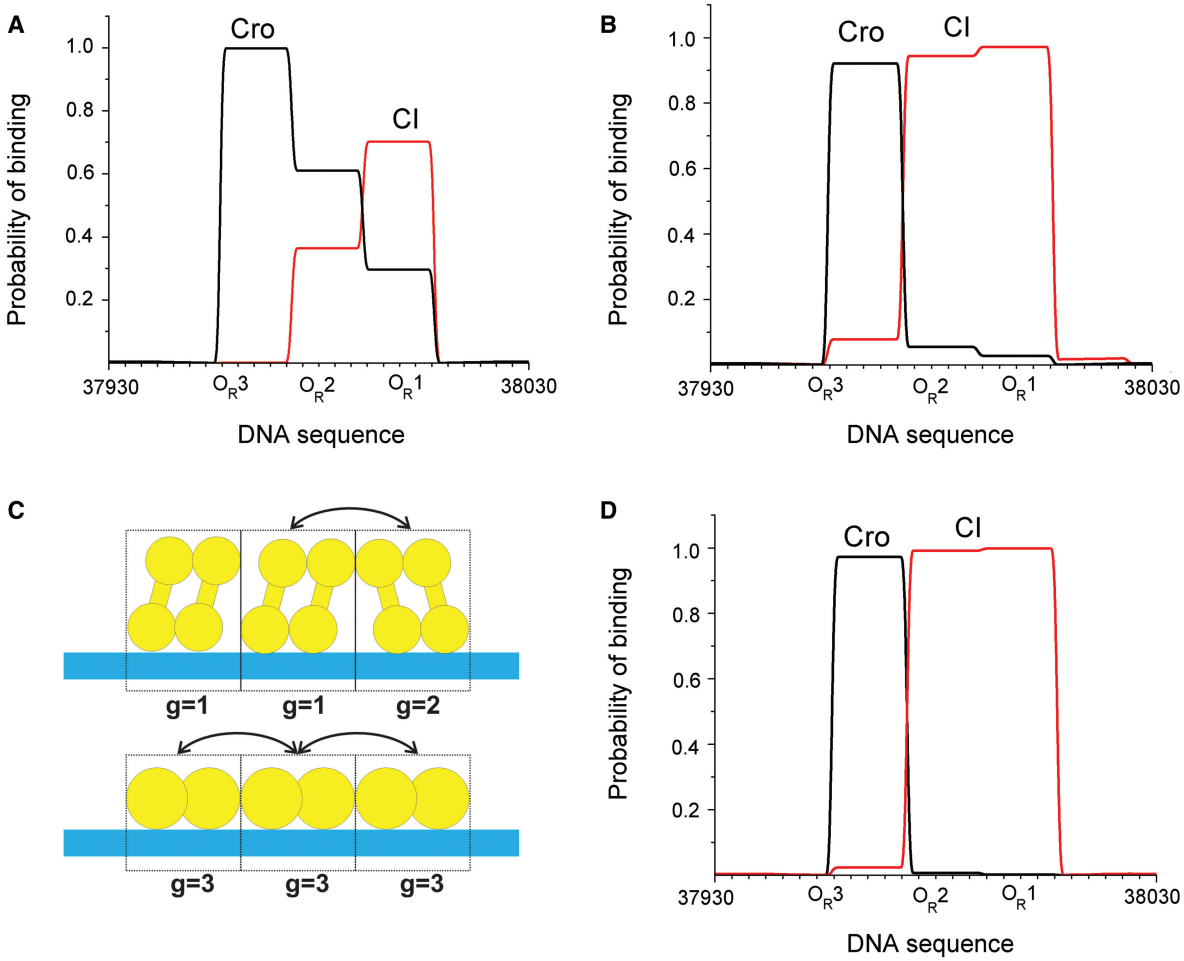

Figure 4. The maps of binding of CI (black) and Cro (red) calculated in the frame of the matrix formalism. Here and in Figures 5–6, the λ O_R sequence does not contain spacers between the O_R1, O_R2 and O_R3 sites. [CI] = 0.2 μM, [Cro] = 0.02 μM. (**A**) Non-cooperative competitive binding of CI and Cro dimers. (**B**) The Cro–Cro and CI–CI interactions are taken in the form of the MvH cooperativity (20). (**C**) Three types of complexes ($g = 1, 2, 3$) formed by asymmetric protein interactions. The CI dimers bound to adjacent DNA sites interact by a direct contact between the recognition domains, while the third dimer binds DNA non-cooperatively. The Cro–Cro interactions are weaker and do not show such asymmetry. (**D**) The map of binding calculated for the model in Figure C.

6 bp correspondingly). Since we cannot distinguish length dependence from asymmetry in the model with deleted spacers, in Figures 4–6 we will assume that Cro–Cro interactions are weak and symmetrical. In these calculations, we set the Cro–Cro interactions equal to –0.3 kcal/mol based on the cooperativity value –0.9 kcal/mol for a complete saturation of three O_R sites by Cro (45). Later in Figure 7, we will take care of the length-dependent characteristics of all protein–protein interactions.

Thus, in the frame of the matrix formalism, we assign two types of complexes to CI binding ($g = 1$, 2) and one type to Cro binding ($g = 3$). The different complexes are indicated in Figure 4C. This situation corresponds to the following set of the contact cooperativity parameters: $w_{02} = 0$, $w_{22} = 0$, $w_{33} = 1.7$ (calculated from the experimental Cro–Cro interaction energy –0.3 kcal/mol), $w_{12} = 74.5$ (calculated from the experimental CI–CI interaction energy –2.5 kcal/mol) (45). All the rest contact cooperativity parameters w_{ij} are units. According to Equation (7), the transfer matrix constructed for each DNA unit has $R \times R$ elements, $R = 3 \times 17 + 1 = 52$.

Figure 4D shows the map of binding calculated according to the scheme in Figure 4C. The energies of interaction used in Figure 4D, are the same as in Figure 4B. We see that a subtle change in the model (the asymmetry in CI–CI interactions) leads to the distinguishable changes in the map of binding. Substantial efforts have been undertaken to construct a proper lattice model for cooperative interactions in this system (61,62). The treatment of such effects is quite natural in the matrix method. We construct the transfer matrix for a general situation and then just set unique weights for all conceivable protein–protein contacts.

The combinatorial control of RNAP

Now let us add the RNAP to the system of λ O_R, Cro and CI. RNAP forms an additional complex, $g = 4$, which covers $m_4 \approx 35$ bp. According to Equation (7), the system

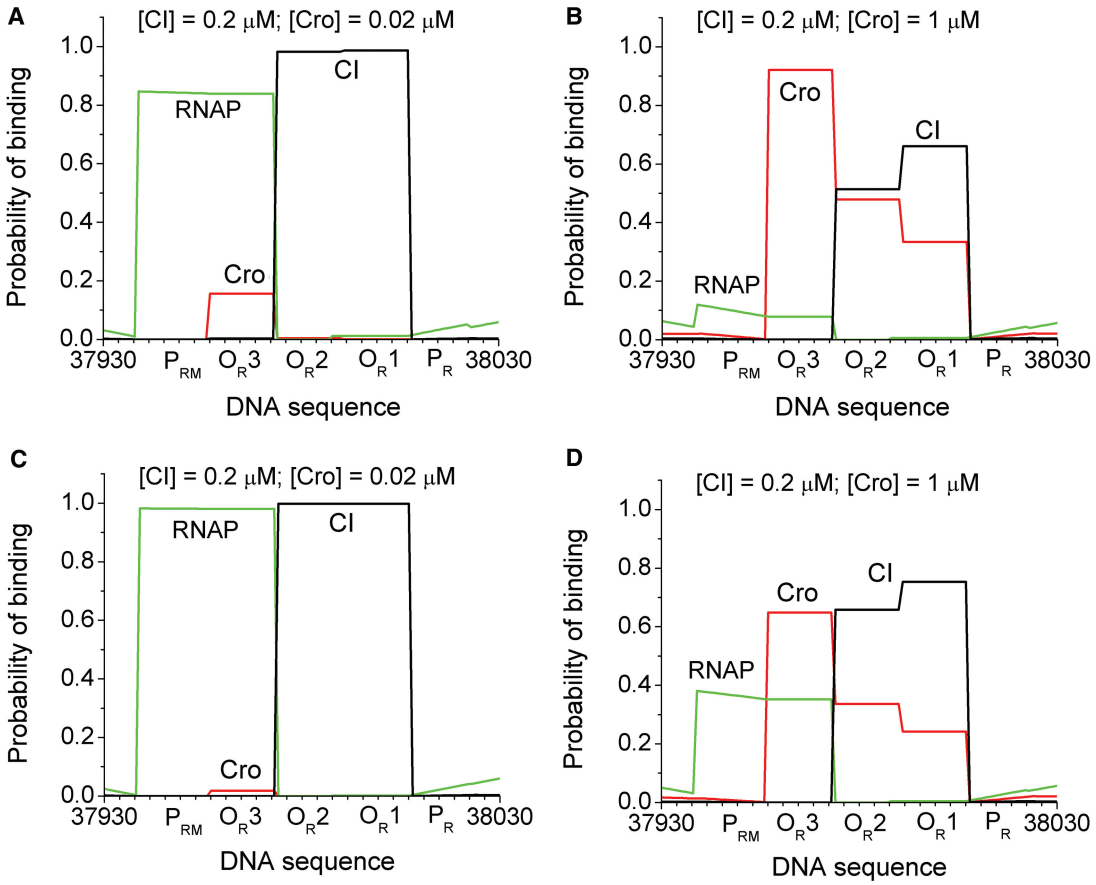

Figure 5. The maps of binding of CI (black), Cro (red) and RNAP (green) calculated for the experimental energies reported for site-specific (45) and non-specific (36) binding, and the concentrations corresponding to the lysogenic state of phage λ. [RNAP] = 3 μM (55); CI and Cro concentrations are indicated in the figure. The interactions between RNAP and CI are neglected in (**A**) and (**B**) (w(0, RNAP, CI) = 1) and taken into account in (**C**) and (**D**) (w(0, RNAP, CI) = 10). The O_L–O_R loop formation is not taken into account.

is now characterized by the transfer matrices with $R = 3 \times 17 + 35 + 1 = 87$ states.

In all following calculations, we take [RNAP] = 3 μM, which corresponds to the lysogenic λ state (63). The energies of RNAP binding to P_{RM} and P_R promoters are equal to –11.5l and –12.5 kcal/mol correspondingly (45).

The non-specific RNAP-binding constant differs from 10^2 to 10^7 depending on the ionic conditions (10,64). In our calculations, we set it equal to $1 \times 10^3\ M^{-1}$, which is close to the experimental value of non-specific holoenzyme binding to double-stranded DNA at 0.01 M $MgCl_2$, 0.2 M NaCl (64). This binding constant determines that ~10% of RNAP is bound to DNA non-specifically at 3 μM concentration. However, our calculations indicate that the λ-switch is quite robust concerning the choice of the non-specific binding constants. The pattern of site-specific binding is almost unaffected since the site-specific binding constants are more than five orders of magnitude larger then the non-specific ones. RNAP binding to DNA is required but is not enough to start transcription. Two other events are important as well: (i) the RNAP-binding site should contain a promoter, and (ii) the activator binding may be required to stimulate RNAP.

Early studies have showed that P_{RM} is maximally stimulated (~10-fold) only when O_R1 and O_R2 are both occupied by CI dimers (58). The role of CI bound at O_R1 in stimulating P_{RM} is primarily to promote repressor binding to O_R2 through cooperative CI–CI interaction (58). Thus, the activating action of CI on RNAP comes from a single RNAP–CI contact. The RNAP–CI contact does not alter the initial RNAP binding to DNA, but it induces the conformational changes in the DNA–RNAP complex helping the open complex formation (65). The open complex formation is one of the compulsory steps in transcription initiation. This multistep reaction may be characterized by a pseudo first-order equilibrium (66). It is this change in the rate of the open complex formation that we may roughly associate with the experimentally observed 10-fold P_{RM} stimulation by CI. In the frame of our formalism, it is equivalent to setting $w(0, \text{RNAP, CI}) = 10$.

Figure 5 shows the maps of binding of CI, Cro and RNAP at O_R operator and its flanking sequences, calculated for the set of parameters chosen above. Figure 5A and B corresponds to the situation when RNAP and CI do not interact with each other. Figure 5A is calculated for the typical lysogenic concentrations: [CI] = 0.2 μM, [Cro] = 0.02 μM. Under these conditions,

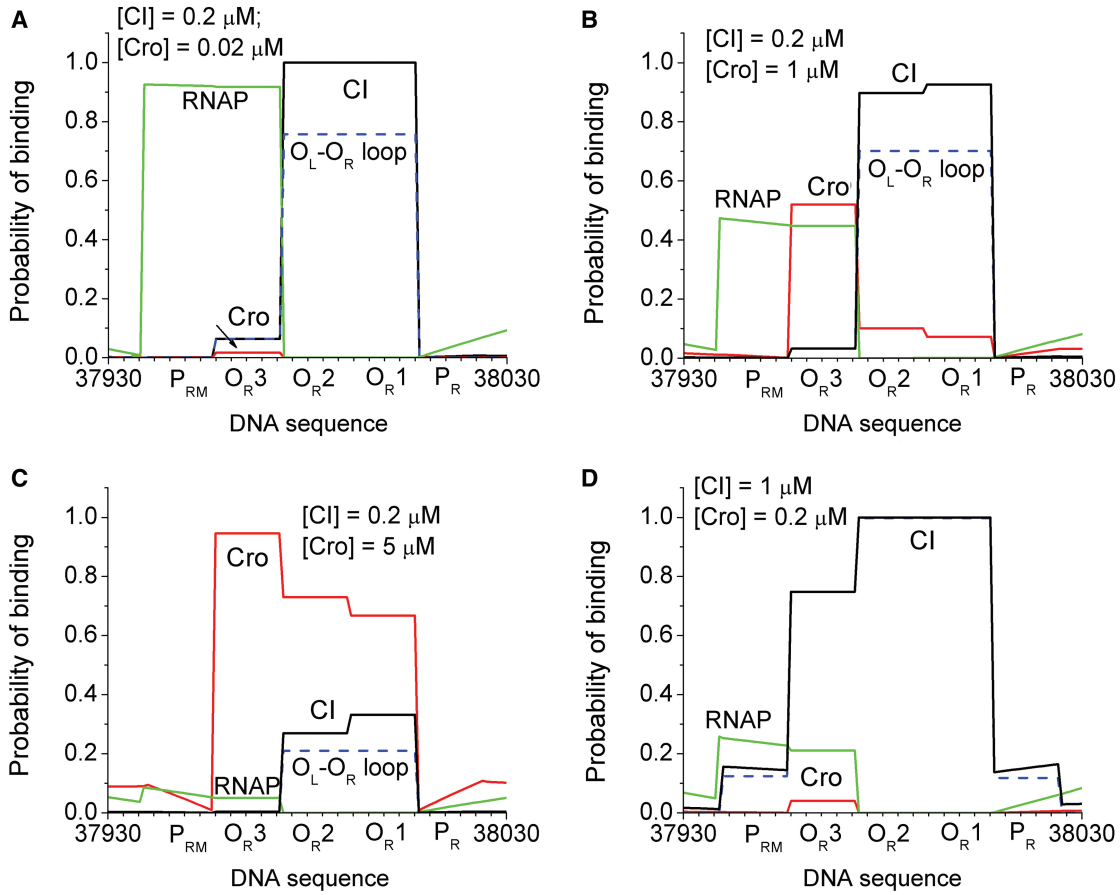

Figure 6. The effect of DNA looping in gene regulation at λ O_R. The colors for CI, Cro and RNAP-binding maps are the same as in Figure 5. The blue dashed line shows the probability of the O_L–O_R loop formation. When dashed line is not seen, it coincides with the solid CI line. The energetic parameters for the single-layer protein binding are the same as in Figure 5. The energies of the vertical contacts between proteins of the first and second layer are equal to the energies of the corresponding horizontal protein–protein interactions. $w_{loop} = 1 \times 10^{-9}$; [RNAP] = 3 μM.

CI repressors are always bound to O_R1 and O_R2 sites, and RNAP covers P_{RM} promoter with ~80% probability. The probability of RNAP binding to P_R promoter is vanishingly small, as it should be expected for the lysogenic state. The *cI* gene is switched on by the P_{RM} promoter, while the *cro* gene regulated by P_R promoter is off. In this situation, CI provides a positive regulation of its own gene. The more we have CI proteins, the higher is the probability of RNAP binding to P_{RM}, the higher is the level of *cI* expression and the more we have new CI proteins. At a higher Cro concentration (Figure 5B) RNAP binding to P_{RM} is suppressed. A 20-fold increase in Cro concentration determines 8-fold decrease of RNAP binding to P_{RM}, as seen from the comparison of Figure 5A and B.

Figure 5C and D is calculated taking into account interactions between RNAP and CI with w(0, RNAP, CI) = 10. A favorable RNAP–CI contact allows these proteins to collaborate against Cro. Cro binding to DNA is almost completely suppressed at 0.02 μM (Figure 5C). Figure 5D corresponds to Cro = 1 μM. It shows that although increasing Cro concentration helps competing with RNAP, the probability of RNAP binding to P_{RM} operator is still ~30%. Thus, the CI–RNAP contact makes the λ switch less sensitive to the Cro influence.

The role of DNA looping

Now let us consider the possibility of the DNA loop formation between the operators O_R and O_L (Figure 3B). According to Table S3, we need additional parameters to construct the transfer matrix for the large loops model. We have to set the energies of CI, Cro and RNAP interaction with O_L operator, the energy of the loop formation and the energies of the vertical CI–CI contacts. Three binding sites at O_L, O_L1, O_L2, O_L3 are symmetrical and almost energetically equivalent to the corresponding binding sites at O_R operator. The structure of the promoters surrounding O_L and O_R operators differs. O_L has only one promoter P_L (symmetrical to P_R) which binds RNAP with the energy –12.5 kcal/mol. There is no direct experimental data on the energy of the O_L–O_R loop formation. We set the energy of the loop formation equal to $\triangle G_{loop} = -12$ kcal/mol which results in $w_{loop} = 1 \times 10^{-9}$. This value is close to the asymptotic value for length-dependent DNA loops studied for the constructs based on the *lac*-repressor system (50). The experimental data on the vertical CI–CI interactions in CI octamerization and DNA loop formation are not that clear as the values for the horizontal CI–CI interactions (67,68). We use here the estimate that the energy of the vertical CI–CI contact is

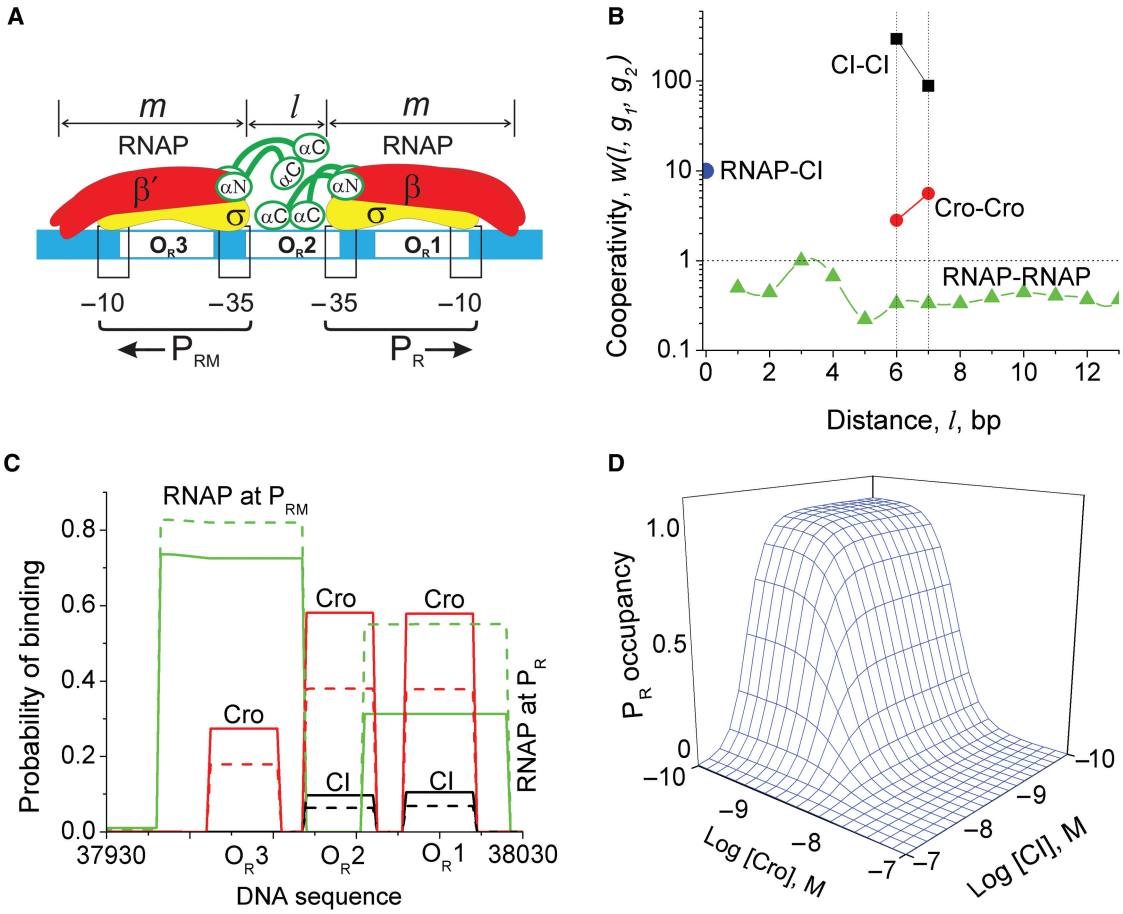

Figure 7. The effect of a range of inter-protein distances in gene regulation at λ O_R. (**A**) The scheme of promoter interference. Two RNAPs do not impede each other's initial binding to the promoters P_R and P_{RM} but interfere at the subsequent stage of the open complex formation (65). CI and Cro also participate in this binding (data not shown). (**B**) The length-dependent interactions between bound proteins. The RNAP–RNAP interaction potential is adopted from Table 1 of (72), the CI–CI and Cro–Cro cooperativity constants are recalculated from (45) and the estimate of RNAP–CI cooperativity is taken from (58). (**C**) The maps of binding calculated in the frame of the matrix formalism taking into account long-range protein–protein interactions. The colors are the same as in Figures 4–6. The dashed line corresponds to the absence of the RNAP–RNAP interactions (the CI–CI, Cro–Cro and RNAP–CI interactions are taken as before). The solid line is calculated taking into account the RNAP–RNAP interactions according to the potential in (B). The concentrations used: [CI] = 2 nM, [Cro] = 0.05 μM, [RNAP] = 3 μM. (**D**) The P_R occupancy by RNAP calculated for different concentrations of free CI and Cro.

equal to the energy of the horizontal CI–CI contact. The single-layer protein interactions are still described by the model in Figure 4C and the parameter set above. According to Equation (11), taking into account DNA looping between the O_L and O_R operators results in the weight matrices of $R \times R$ elements, with $R = 518$.

Figure 6 shows the maps of binding at O_R operator calculated in the frame of the large loops model taking into account the possibility of O_R–O_L contact due to bridging by CI proteins. Figure 6A is calculated for the typical lysogenic concentrations ([CI] = 0.2 μM, [Cro] = 0.02 μM). The CI proteins occupy the O_R1 and O_R2 sites with almost 100% probability, while the probability of the O_L–O_R loop formation is ~70%, and the probability of RNAP binding at P_{RM} promoter is ~80%. When we add 1 μM Cro, the level of P_{RM} expression is still above 40% (Figure 6B). Only a five-micromolar Cro concentration is enough to suppress RNAP binding to P_{RM} (Figure 6C). This is much larger then the typical concentrations *in vivo*. Yet, this would suppress RNAP binding not only to P_{RM}, but also to P_R. Thus, we cannot switch λ from lysogeny to lysis just playing with Cro concentrations. Even a large number of Cro proteins occasionally trapped inside the cell will not trigger the λ-switch. Only the cleavage of CI repressors may act as a trigger.

These calculations provide an argument in support to the recent mutation studies, which show that *cro* gene might be unimportant for the lysogenic to lytic switch, and that Cro's primary role in induction is instead to mediate the weak repression of the early lytic promoters (69). We see that one of the functions of the O_R–O_L loop is to make the lysogenic state more stable by decreasing its dependence on Cro fluctuations. That is not the only function of the O_L–O_R loop.

Figure 6D shows what happens if CI is overexpressed during the lysogenic λ state. At high CI concentrations, CI occupies O_R3 instead of RNAP. This is because the

vertical CI–CI assembly may take place not only at O_R1 and O_R2 but also at O_R3. An additional vertical contact between CI dimers bound to O_R3 and O_L3 sites may recruit a CI tetramer that strongly competes with RNAP at P_{RM}. At lower concentrations CI occupies only O_R1 and O_R2 while O_R3 is left for Cro/RNAP competition (Figure 6A–C).

The situation shown in Figure 6D corresponds to the decrease of CI synthesis from P_{RM} by the O_L–O_R loop due to a too high CI concentration. Thus at high CI concentrations this protein implements a negative regulation of its gene *cI*. Taken together with the positive regulation at small CI concentrations (70), this provides a stable mechanism of the lysogeny maintenance: the concentration of the regulatory proteins always tends to be at the lysogenic level. Only external signals leading to the degradation of CI may switch the phage from the lysogeny to lysis, where P_R promoter becomes active instead of P_{RM}, and Cro proteins start to increase their own synthesis.

Figure 6 shows that the O_L–O_R loop is formed only at high-enough lysogenic CI concentrations. The probability of the O_L–O_R loop formation is sufficiently lower in the lytic state. For example, our calculations give just 0.1% to the probability of the loop formation at $[CI] = 0.01\,\mu M$, $[Cro] = 0.1\,\mu M$, $[RNAP] = 3\,\mu M$. Thus at small CI concentrations in the lytic state the O_L–O_R loop may be neglected. Still, the switch from the lysogenic to lytic λ state should be discrete. What, if not the loop, contributes to this discreteness? In order to answer this question, in the next section we look at the fine structural details of the intact λ O_R.

The role of a range of inter-protein distances

Now let us perform calculations for the intact λ O_R sequence. Remember, that O_R1 and O_R2 are separated by a 6-bp spacer, while O_R2 and O_R3 are separated by a 7-bp spacer, the details that we have skipped in our previous calculations. Therefore, we now have to use the long-range interaction model for the CI–CI and Cro–Cro interactions instead of the contact interactions model. We do not know the complete law of the length-dependence for these interactions. The experiments on the intact λ O_R give us just the values of the cooperativity parameters for the proteins separated by 6 and 7 bp (45). This corresponds to $\varepsilon(6, CI, CI) = -3.3$ kcal, $\varepsilon(7, CI, CI) = -2.6$ kcal, $\varepsilon(6, Cro, Cro) = -0.6$ kcal, $\varepsilon(7, Cro, Cro) = -1.0$ kcal (37). We will treat now all CI–CI and Cro–Cro interactions as asymmetric and pairwise.

What about the RNAP–RNAP interactions? The P_R and P_{RM} promoters are separated by 13 bp. Most of the quantitative models for the λ-switch do not take the RNAP–RNAP interactions into account (59–63,68,71), although it is known experimentally, that two RNAP molecules do interact even when they are bound to the promoters separated by up to 23 bp (72). It has been proposed that the long-range RNAP–RNAP interactions are through C-terminal domains of the α-subunits, which hang out from RNAP trying to bind the DNA next to the σ-subunit to assist the open complex formation (Figure 7A). Although two RNAPs do not impede each other's initial binding to the promoters P_R and P_{RM}, they interfere at the subsequent stage of the transcriptionally active open complex formation. A recent experimental study has determined the rate of the open complex formation as a function of the RNAP–RNAP separation along DNA (72). In particular, it was shown that a 10-bp deletion in the DNA sequence between the P_R and P_{RM} promoters leads to the suppression of promoter interference. Let us assign the RNAP–RNAP interactions a length-dependent potential w(*l*, RNAP, RNAP), which is normalized so that to be equal to unit at the point of the 10-bp deletion. This potential recalculated from Table 1 of Strainic and co-authors (72) is plotted in our Figure 7B. For a comparison, we have also plotted the values for the CI–CI and Cro–Cro cooperativity parameters. The RNAP–RNAP interaction potential is always equal or lower then unit. This means that the RNAP–RNAP interactions are anticooperative unlike the CI–CI, RNAP–CI and Cro–Cro interactions. Binding of RNAP to λ O_R in the absence of regulatory proteins is indeed anticooperative (72). What happens in the presence of the regulatory proteins is not evident before the calculations.

Figure 7C shows the maps of binding calculated for the intact λ O_R at small concentrations of repressors (corresponds to the lytic state or the boundary between the two states). The dashed lines are calculated when the RNAP–RNAP interference is not taken into account. The solid lines are calculated when the RNAP–RNAP interference is taken into account in the form of the experimental interaction potential shown in Figure 7B. Under these conditions, RNAP binding is allowed at both promoters. However the situation when both promoters are highly active is not compatible with the discrete logic of the λ-switch. The O_L–O_R loop formation is unfavorable at these small [CI] concentrations, and therefore the discreteness of the λ logic is maintained due to the RNAP–RNAP interactions. Once RNAP binds to one of the promoters, it decreases the probability of another RNAP binding to the second promoter. The comparison between the dashed and the solid lines in Figure 7C shows that the RNAP–RNAP interference may account for a large change in the pattern of transcription expression.

Figure 7D shows the occupancy of P_R promoter by RNAP as a function of the concentrations of free Cro and CI. The calculations are performed taking into account all distance-dependent protein–protein interactions shown in Figure 7B. In this regime, the highest level of P_R activity is at the smallest CI and Cro concentrations. At very low CI concentrations or in the absence of CI, increasing Cro concentration yields a monotonous decrease of P_R activity, leading to a subsequent switching off the *cro* gene and the other yearly lytic genes. Figure 7D is consistent with the experiments (45). The comparison of this figure with the previous calculations based on the '40-state model' (45), which does not take into account distance-dependent protein–protein interactions, shows that the effect of RNAP interference is basically to shift the P_R activity towards lower CI and Cro concentrations. This changes the concentrational threshold required for the λ-switch to lysis.

DISCUSSION

We have showed that a genetic system consisting of several *cis*-regulatory modules is calculable using the transfer matrix formalism considering explicitly site-overlapping, competitive and cooperative binding of regulatory proteins, their multilayer assembly and DNA looping. We have developed the matrix models for the basic types of interactions between regulatory proteins, drugs and nucleosomes and have showed how more complex models may be derived starting from the basic ones. Then we applied the matrix methodology to the problem of gene regulation at O_R operator of phage λ. The matrix method not only allowed a correct description of the available experimental data for this well-documented genetic system, but also suggested new biologically relevant predictions.

The λ-switch at a single-nucleotide resolution

Have we brought anything new to the λ-switch? Several types of quantitative descriptions of the λ-switch are available in the literature. They work by separating the system into a number of energetic states and performing the stochastic kinetic analysis (71) or finding equilibrium from the system of the corresponding laws of mass action (45,59,60,68). Such descriptions require distinguishing all the relevant states of the system manually before starting the analysis. Usually these states are associated with the individual binding sites (free or bound). When protein–protein interactions come into play, each combination of states of the binding sites gives birth to a new energetic state of the whole system. The more we know experimentally, the more states appear in the new quantitative descriptions. Thus, having initially started from eight states for CI and Cro rearrangements at λ O_R (59), the quantitative models for the λ-switch now deal with 40 states for CI–Cro–RNAP binding (45), or 64 states taking into account the O_R–O_L loop (68).

On the other hand, the transfer matrix formalism allows a description at the level of the elementary DNA units. We have chosen one base pair as an elementary unit for the λ system. Although this tremendously increases the number of states (while calculating the binding maps shown in Figure 6, we have rendered through 518 states for each of 100bp comprising the DNA sequence of our interest), the systematic nature of the matrix formalism allows us to be much more general. When new experimental data reveals some previously unknown interactions, we still have the same matrices; we just have to change the statistical weights. Working at the level of individual base pairs gives the method more predictive power.

The application of the matrix formalism to the λ-switch allowed us to take into account both the short-range interactions (including protein overlapping and protein–protein contacts) and the long-range interactions (including the interference between the neighboring P_R and P_{RM} promoters, and the P_L promoter laying 2.4 kb away). The long-range interactions were split into two types of effects. Firstly, the gene regulatory effects on the level of *cis*-regulatory modules communicating through DNA looping (Figure 6). Secondly, the fine effects such as the single nucleotide mutations and the interference of the adjacent promoters (Figure 7). The latter effects are experimentally important (72,73) but have not been taken into account by the previous models. In particular, we have showed that while the DNA loop between the O_R and O_L operators is important at the lysogenic CI concentrations, the interference between the adjacent promoters becomes more important at small CI concentrations. According to these calculations, a significant change of the expression pattern may arise in this regime due to anticooperative interactions between RNA polymerases bound at P_R and P_{RM}. As we see from Figure 7, the distance-dependent protein–protein interactions cannot be neglected in the quantitative treatment of this system.

The systematic nature of the transfer matrix formalism

The results of the calculations reported in the previous section allow us to enlarge our knowledge of the functioning of phage λ. However, more important are the general methodological issues raised by the current work. When we first approached the transfer matrix formalism, it was not evident at all, whether it is applicable to the biophysical characterization of the complex gene regulatory systems. While the transfer matrices are widely used in bioinformatics for DNA sequence analysis (5–7,30–32), in biophysics the method has been used only for the classical biopolymer problems such as the DNA-ligand binding (20), DNA melting (49) or actin–miosin interactions (29). As mentioned in the introduction, it seemed from the literature analysis, that the transfer matrix method was the only lattice approach left for our purposes. However, there were still concrete doubts, mainly because the method is computationally costly, and because there were no examples of its applications to the multiprotein binding events overstepping the limits of one dimension. Our study has showed that these tasks are still solvable. We were able to split the concrete gene regulatory system into a number of binding events, to characterize them by the experimental thermodynamical parameters, to construct the matrix formalism, to calculate the probabilities of all binding events and to link them with the decisions of the gene regulatory system. Furthermore, it appeared that the transfer matrix formalism possesses several nice features that allow us to consider it as a general systematic tool for statistical-mechanical description of DNA–protein–drug binding involved in gene regulation. These features are the scalability, extendibility and compatibility with the matrix approaches of bioinformatics, as detailed below.

Scalability. The proposed matrix approach is easily scalable. This is achieved by redefining the elementary DNA unit. For example, a 2.4-kb loop between the O_L and O_R operators of phage λ is 'large' when the elementary DNA unit is associated with one base pair. However, if we associate an elementary DNA unit with the length of the repressor binding site (17 bp) the O_L–O_R loop will contain just 150 of these units, which is already a computationally 'small' loop.

Extendibility. The extendibility of the matrix method means that the complex matrix models may be derived based on the simpler models incorporated in the existing transfer matrices. For example, we have showed how the small drug and nucleosome models arise as the modifications to the basic competition model; the large loops model arises as a combination of the models for the cooperative competitive assembly and the multilayer binding. Although we are not using transfer matrices as direct building blocks to construct matrices for the complex models, we do use the states enumerations as the building blocks [Equations (7–11)]. A tempting idea of using a direct matrix product for new matrix construction might be examined in future (74).

Compatibility with bioinformatics approaches. The sequence analysis allows calculating the affinities of protein binding to DNA based on statistical mechanics (75). These methods are now being extensively developed, and most of them are based on the weight matrices (5–7, 30–32). In some cases, sequence analysis already provides a good prediction of the protein–DNA binding constants (5–7). The common matrix formalism would allow using the output of the sequence analysis as the input for our calculation of the maps of protein binding. Then one may use our calculated maps of binding as input for the matrix approaches of networks analysis to predict gene expression in even more complex systems (14).

How far can we progress?

The concentrations of all transcription factors and their interaction energies with DNA and other proteins must be known in advance in the transfer matrix formalism. This reflects a fundamental nature of the combinatorial control of gene regulation, not the demands of our algorithm. As we have showed, the λ-switch is quite robust to the small changes in the concentrations of transcription factors. Most of genetic systems have a discrete logic of this type: the switching event is either the lowering of a concentration below a threshold level (as with CI degradation upon UV-irradiation of phage λ) or changing the concentration from a less-then-threshold-level to a threshold level (76). The autoregulation then helps to maintain the concentrations of transcription factors in a fixed range required for a proper gene functioning in the given regime. The eukaryotic systems are even more robust. If just one of the regulatory proteins is missing (or its concentration is below a threshold), the enhanceosome complex will not assemble and the transcription machinery will not be recruited to a given promoter.

The lack of biophysical interaction energy data is still a large problem. At present, only several genetic systems are well characterized in terms of the binding constants and the cooperativity parameters. One of the by-products of our study is a collection of the energetic parameters that help to characterize phage λ more completely. This set of parameters may be used as a basis for modeling a number of the λ mutations and the analogous phages. Indeed, in the absence of a detailed knowledge about, say, a *Lactobacillus casei's* bacteriophage A2 CI protein (77), we may substitute it with the characteristics of the homologous λ CI protein. Future studies will show how good such an approximation is. There are also a number of energetic parameters, which have a general character not confined to a concrete genetic system. For example, the RNAP–DNA, and the RNAP–RNAP interactions as discussed in our Figure 7. Such interactions, as well as the DNA interactions with histones and abundant non-histone proteins should be accurately systematized and tabulated.

A growing number of other important systems are now being characterized in terms of thermodynamics on a level comparable with that of the λ-switch. The well-known examples are the *lac* operon and a number of other bacterial genes, the early developmental genes of simple eukaryotes, the Epstein–Barr virus and the human interferon gene. Thus at present, we may be busy characterizing such systems with the help of the transfer matrix formalism. In future, however, we expect that the high-throughput experiments and/or theory will be able to provide the necessary energetic constants for the arbitrary sequences, and the proposed method will become an increasingly useful framework to study gene regulation in new pharmaceutically important systems.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

I thank Dmitri Lando, Peter H. von Hippel, Terrence Hwa, Alexander Fridman and the anonymous referees for valuable comments. I am grateful to Avinoam Ben-Shaul for a kind hosting of my stay in the Hebrew University of Jerusalem. This work was partly supported by Belarusian National Foundation of Fundamental Research, grant B06M-127.

Conflict of interest statement. None declared.

REFERENCES

1. Vilar,J.M.G. and Leibler,S. (2003) DNA looping and physical constraints on transcription regulation. *J. Mol. Biol.*, **331**, 981–989.
2. Teif,V.B. (2005) Ligand-induced DNA condensation: choosing the model. *Biophys. J.*, **89**, 2574–2587.
3. Widom,J. (2005) Target site localization by site-specific, DNA-binding proteins. *Proc. Natl Acad. Sci. USA*, **102**, 16909–16910.
4. Ho,S.-W., Jona,G., Chen,C.T.L., Johnston,M. and Snyder,M. (2006) Linking DNA-binding proteins to their recognition sequences by using protein microarrays. *Proc. Natl Acad. Sci. USA*, **103**, 9940–9945.
5. Morozov,A.V., Havranek,J.J., Baker,D. and Siggia,E.D. (2005) Protein–DNA binding specificity predictions with structural models. *Nucleic Acids Res.*, **33**, 5781–5798.
6. Sarai,A. and Kono,H. (2005) Protein-DNA recognition patterns and predictions. *Ann. Rev. Biophys. Biomol. Struct.*, **34**, 379–398.
7. Gershenzon,N.I., Stormol,G.D. and Ioshikhes,I.P. (2005) Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites. *Nucleic Acids Res.*, **33**, 2290–2301.

8. Bintu,L., Buchler,N.E., Garcia,H.G., Gerland,U., Hwa,T., Kondev,J., Kuhlman,T. and Phillips,R. (2005) Transcriptional regulation by the numbers: applications. *Curr. Opin. Gen. Dev.*, **15**, 125–135.
9. Kuhlman,T., Zhang,Z., Saier,M.Jr and Hwa,T. (2007) Quantitative characterization of combinatorial transcriptional control of the lactose operon of E. coli. *Proc. Natl Acad. Sci. USA*, **104**, 6043–6048.
10. Grigorova,I.L., Phleger,N.J., Mutalik,V.K. and Gross,C.A. (2006) Insights into transcriptional regulation and σ competition from an equilibrium model of RNA polymerase binding to DNA. *Proc. Natl Acad. Sci. USA*, **103**, 5332–5337.
11. Konishi,T. (2005) A thermodynamic model of transcriptome formation. *Nucleic Acids Res.*, **33**, 6587–6592.
12. Saiz,L. and Vilar,J.M.G. (2006) DNA looping: the consequences and its control. *Cur. Opin. Struct. Biol.*, **16**, 344–350.
13. Granek,J.A. and Clarke,N.D. (2005) Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome Biol.*, **6**, 87.
14. Gianchandani,E.P., Papin,J.A., Price,N.D., Joyce,A.R. and Palsson,B.O. (2006) Matrix formalism to describe functional states of transcriptional regulatory systems. *PLoS Comp. Biol.*, **2**, 902–917.
15. Veitia,R.A. (2003) A sigmoidal transcriptional response: cooperativity, synergy and dosage effects. *Biol. Rev.*, **78**, 149–170.
16. Wang,J., Ellwood,K., Lehman,A., Carey,M.F. and She,Z.S. (1999) A mathematical model for synergistic eukaryotic gene activation. *J. Mol. Biol.*, **286**, 315–325.
17. Saiz,L. and Vilar,J.M.G. (2006) Stochastic dynamics of macromolecular-assembly networks. *Mol. Sys. Biol.*, **2**, doi: 10.1038/msb4100061.
18. Lifson,S. (1964) Partition functions of linear-chain molecules. *J. Chem. Phys.*, **40**, 3705–3710.
19. Schellman,J.A. (1974) Cooperative multisite binding to DNA. *Isr. J. Chem.*, **12**, 219–238.
20. Crothers,D.M. (1968) Calculation of binding isotherms for heterogeneous polymers. *Biopolymers*, **6**, 575–584.
21. Gurskii,G.V., Zasedatelev,A.S. and Vol'kenshtein,M.V. (1972) Theory of one-dimensional adsorption. II. Adsorption of small molecules on a heteropolymer. *Mol. Bio. (Moscow)*, **6**, 479–489.
22. McGhee,J.D. and von Hippel,P.H. (1974) Theoretical aspects of DNA-protein interactions: cooperative and non-cooperative binding of large ligands to a one-dimensional homogeneous lattice. *J. Mol. Biol.*, **86**, 469–489.
23. Zasedatelev,A.S., Gurskii,G.V. and Vol'kenshtein,M.V. (1971) Theory of one-dimensional adsorption. I. Adsorption of small molecules on a homopolymer. *Mol. Biol. (Moscow)*, **5**, 194–198.
24. Poland,D. (1979) *Cooperative Equilibrium in Physical Biochemistry*. Clarendon, Oxford.
25. Di Cera,E. (1995) *Thermodynamic Theory of Site-Specific Binding Processes in Biological Macromolecules*. Cambridge University, Cambridge.
26. Nechipurenko,Yu.D., Jovanovic,B., Riabokon,V.F. and Gursky,G.V. (2005) Quantitative methods of analysis of footprinting diagrams for the complexes formed by a ligand with a DNA fragment of known sequence. *Ann. N.Y. Acad. Sci.I.*, **1048**, 206–214.
27. Chen,Y. (1987) Binding of *n*-mers to one-dimensional lattices with longer than close-contact interactions. *Biophys. Chem.*, **27**, 59–65.
28. Chen,Y. (1990) A general secular equation for cooperative binding of *n*-mer ligands to a one-dimensional lattice. *Biopolymers*, **30**, 1113–1121.
29. Chen,Y. (2004) Multiple binding of ligands to a linear biopolymer. *Methods Enzymol.*, **379**, 145–152.
30. Bussemaker,H.J., Li,H. and Siggia,E.D. (2000) Building a dictionary for genomes: Identification of presumptive regulatory sites by statistical analysis. *Proc. Natl Acad. Sci. USA*, **97**, 10096–10100.
31. Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge.
32. Sinha,S., van Nimwegen,E. and Siggia,E.D. (2003) A probabilistic method to detect regulatory modules. *Bioinformatics*, **19**, i292–i301.
33. Epstein,I.R. (1978) Cooperative and noncooperative binding of large ligands to a finite one-dimensional lattice. A model for ligand-oligonucleotide interactions. *Biophys. Chem.*, **8**, 327–339.
34. Di Cera,E. and Phillipson,P.E. (1996) Map analysis of ligand binding to a linear lattice. *Biophys. Chem.*, **61**, 125–129.
35. Flyvbjerg,H., Keatch,S.A. and Dryden,D.T.F. (2006) Strong physical constraints on sequence-specific target location by proteins on DNA molecules. *Nucleic Acids Res.*, **34**, 2550–2557.
36. Bakk,A. and Metzler,R. (2004) *In vivo* non-specific binding of λ *cI* and *cro* repressors is significant. *FEBS Lett.*, **563**, 66–68.
37. Grokhovsky,S.L., Surovaya,A.N., Burckhardt,G., Pismensky,V.F., Chernov,B.K., Zimmer,Ch. and Gursky,G.V. (1998) DNA sequence recognition by bis-linked netropsin and distamycin derivatives. *FEBS Lett.*, **439**, 346–350.
38. Dervan,P.B., Doss,R.M. and Marques,M.A. (2005) Programmable DNA binding oligomers for control of transcription. *Curr. Med. Chem. - Anti-Cancer Agents*, **5**, 373–387.
39. Burnett,R., Melander,C., Puckett,J.W., Son,L.S., Wells,R.D., Dervan,P.B. and Gottesfeld,J.M. (2006) DNA sequence-specific polyamides alleviate transcription inhibition associated with long GAA·TTC repeats in Friedreich's ataxia. *Proc. Natl. Acad. Sci. USA*, **103**, 1497–11502.
40. Nguyen-Hackley,D.H., Ramm,E., Taylor,C.M., Joung,J.K., Dervan,P.B. and Pabo,C.O. (2004) Allosteric inhibition of zinc-finger binding in the major groove of DNA by minor-groove binding ligands. *Biochemistry*, **43**, 3880–3890.
41. Mohammad-Rafiee,F., Kulić,I.M. and Schiessel,H. (2004) Theory of nucleosome corkscrew sliding in the presence of synthetic DNA ligands. *J. Mol. Biol.*, **344**, 47–58.
42. Segal,E., Fondufe-Mittendorf,Y., Chen,L., Thåström,A., Field,Y., Moore,I.K., Wang,J.-P.Z. and Widom,J. (2006) A genomic code for nucleosome positioning. *Nature*, **442**, 774–778.
43. Miller,J.A. and Widom,J. (2003) Collaborative competition mechanism for gene activation in vivo. *Mol. Cell. Biol.*, **23**, 1623–1632.
44. Teif,V.B., Haroutunian,S.G., Vorob'ev,V.I. and Lando,D.Y. (2002) Short-range interactions and size of ligands bound to DNA strongly influence adsorptive phase transition caused by long-range interactions. *J. Biomol. Struct. Dynam.*, **19**, 1103–1110.
45. Darling,P.J., Holt,J.M. and Ackers,G.K. (2000) Coupled energetics of λ *cro* repressor self-assembly and site-specific DNA operator binding II: cooperative interactions of *cro* dimers. *J. Mol. Biol.*, **302**, 625–638.
46. Ptashne,M. (1992) *A Genetic Switch*. Cell Press and Blackwell Science, Cambridge, MA.
47. Nechipurenko,Y.D. and Gursky,G.V. (1986) Cooperative effects on binding of proteins to DNA. *Biophys. Chem.*, **24**, 195–209.
48. Wolfe,A.R. and Meehan,T. (1992) Use of binding site neighbor-effect parameters to evaluate the interactions between adjacent ligands on a linear lattice. Effects on ligand-lattice association. *J. Mol. Biol.*, **223**, 1063–1087.
49. Poland,D. (2004) DNA melting profiles from a matrix method. *Biopolymers*, **73**, 216–228.
50. Schurr,J.M., Delrow,J.J., Fujimoto,B.S. and Benight,A.S. (1997) The question of long-range allosteric transitions in DNA. *Biopolymers*, **44**, 283–308.
51. Chen,Y., Maxwell,A. and Westerhoff,H.V. (1986) Co-operativity and enzymatic activity in polymer-activated enzymes. A 1D piggy-back binding model and its application to the DNA-dependent ATPase of DNA gyrase. *J. Mol. Biol.*, **190**, 201–214.
52. Grilley,D., Soto,A.M. and Draper,D.E. (2006) Mg^{2+}-DNA interaction free energies and their relationship to the folding of RNA tertiary structures. *Proc. Natl Acad. Sci. USA*, **103**, 14003–14008.
53. Saiz,L., Rubi,J.M. and Vilar,J.M.G. (2005) Inferring the in vivo looping properties of DNA. *Proc. Natl Acad. Sci. USA*, **102**, 17642–17645.
54. Zhang,Y., McEwen,A.E., Crothers,D.M. and Levene,S.D. (2006) Statistical-mechanical theory of DNA looping. *Biophys. J.*, **90**, 1903–1912.
55. Arnosti,D.N. and Kulkarni,M.M. (2005) Transcriptional enhancers: intelligent enhanceosomes or flexible billboards? *J. Cell. Biochem.*, **94**, 890–898.
56. Lando,D.Y. and Teif,V.B. (2002) Modeling of DNA Condensation and decondensation caused by ligand binding. *J. Biomol. Struct. Dynam.*, **20**, 215–222.
57. Vilar,J.M.G. and Saiz,L. (2006) Multiprotein DNA looping. *Phys. Rev. Lett.*, **96**, 238103.

58. Meyer,B.J., Maurer,R. and Ptashne,M. (1980) Gene regulation at the right operator (O_R) of bacteriophage λ. II. O_Rl, O_R2, and O_R3: their roles in mediating the effects of repressor and cro. *J. Mol. Biol.*, **139**, 163–194.
59. Ackers,G.K., Johnson,A.D. and Shea,M.A. (1982) Quantitative model for gene regulation by phage repressor. *Proc. Natl Acad. Sci. USA*, **79**, 1129–1133.
60. Bakk,A., Metzler,R. and Sneppen,K. (2004) Sensitivity of O_R in phage λ. *Biophys. J.*, **86**, 58–66.
61. Saroff,H.A. (1993) Individual-site binding data and the energetics of protein-DNA interactions. *Biopolymers*, **33**, 1327–1336.
62. Ben-Naim,A. (1998) Cooperativity in binding of proteins to DNA. II. Binding of bacteriophage λ repressor to the left and right operators. *J. Chem. Phys.*, **108**, 6937–6946.
63. Santillán,M. and Mackey,M.C. (2004) Why the lysogenic state of phage λ is so stable: a mathematical modeling approach. *Biophys. J.*, **86**, 75–84.
64. deHaseth,P.L., Lohman,T.M., Burgess,R.R. and Record,T.M.Jr. (1978) Nonspecific interactions of *Escherichia coli* RNA polymerase with native and denatured DNA: differences in the binding behavior of core and holoenzyme. *Biochemistry*, **17**, 1612–1622.
65. Nickels,B.E., Dove,S.L., Murakami,K.S., Darst,S.A. and Hochschild,A. (2002) Protein-protein and protein-DNA interactions of σ70 region 4 involved in transcription activation by λcI. *J. Mol. Biol.*, **324**, 17–34.
66. Saecker,R.M., Tsodikov,O.V., McQuade,K.L., Schlax,P.E.Jr, Capp,M.W. and Record,M.T.Jr. (2002) Kinetic studies and structural models of the association of *E.coli* σ^{70} RNA polymerase with the λP_R promoter: large scale conformational changes in forming the kinetically significant intermediates. *J. Mol. Biol.*, **319**, 649–671.
67. Rusinova,E., Ross,J.B.A., Laue,T.M., Sowers,L.C. and Senear,D.F. (1997) Linkage between operator binding and dimer to octamer self-assembly of bacteriophage λ cI repressor. *Biochemistry*, **36**, 12994–13003.
68. Dodd,I.B., Shearwin,K.E., Perkins,A.J., Burr,T., Hochschild,A. and Egan,J.B. (2004) Cooperativity in long-range gene regulation by the λ CI repressor. *Genes Dev.*, **18**, 344–354.
69. Svenningsen,S.L., Costantino,N., Court,D.L. and Adhya,S. (2005) On the role of Cro in λ prophage induction. *Proc. Natl Acad. Sci. USA*, **102**, 4465–4469.
70. Michalowski,C.B. and Little,J.W. (2005) Positive autoregulation of *c*I is a dispensable feature of the phage λ gene regulatory circuitry. *J. Bacteriol.*, **187**, 6430–6442.
71. Arkin,A., Ross,J. and McAdams,H.H. (1998) Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected *Escherichia coli* cells. *Genetics*, **149**, 1633–1648.
72. Strainic,M.G.Jr, Sullivan,J.J., Collado-Vides,J. and deHaseth,P.L. (2000) Promoter interference in a bacteriophage lambda control region: effects of a range of interpromoter distances. *J. Bacteriol.*, **182**, 216–220.
73. Davis,C.A., Capp,M.W., Record,M.T.Jr and Saecker,R.M. (2005) The effects of upstream DNA on open complex formation by Escherichia coli RNA polymerase. *Proc. Natl Acad. Sci. USA*, **102**, 285–290.
74. Woodbury,C.P.Jr. (1988) Direct product-matrix method treatment of macromolecular binding. *Biopolymers*, **27**, 1305–1317.
75. Berg,O.G. and von Hippel,P.H. (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, **193**, 723–750.
76. Buchler,N.E., Gerland,U. and Hwa,T. (2003) On schemes of combinatorial transcription logic. *Proc. Natl Acad. Sci. USA*, **100**, 5136–5141.
77. García,P., Ladero,V., Alonso,J.C. and Suárez,J.E. (1999) Cooperative interaction of CI protein regulates lysogeny of *Lactobacillus casei* by bacteriophage A2. *J. Virol.*, **73**, 3920–3929.