Accurate Prediction of Gene Expression by Integration of DNA Sequence Statistics with Detailed Modeling of Transcription Regulation

Jose M. G. Vilar[†‡*]
[†]Biophysics Unit (CSIC–UPV/EHU) and Department of Biochemistry and Molecular Biology, University of the Basque Country, Bilbao, Spain; and [‡]IKERBASQUE, Basque Foundation for Science, Bilbao, Spain

ABSTRACT Gene regulation involves a hierarchy of events that extend from specific protein-DNA interactions to the combinatorial assembly of nucleoprotein complexes. The effects of DNA sequence on these processes have typically been studied based either on its quantitative connection with single-domain binding free energies or on empirical rules that combine different DNA motifs to predict gene expression trends on a genomic scale. The middle-point approach that quantitatively bridges these two extremes, however, remains largely unexplored. Here, we provide an integrated approach to accurately predict gene expression from statistical sequence information in combination with detailed biophysical modeling of transcription regulation by multidomain binding on multiple DNA sites. For the regulation of the prototypical *lac* operon, this approach predicts within 0.3-fold accuracy transcriptional activity over a 10,000-fold range from DNA sequence statistics for different intracellular conditions.

INTRODUCTION

In a now classic article proposing the *lac* operon model, Jacob and Monod put forward the very basic principles of gene regulation (1). They reasoned that there are molecules that bind to specific sites in nucleic acids to control whether or not genes are expressed. Since then, a major challenge in biology has been to understand how site-specific regulatory factors function and the effects that they have on gene regulation. Thus, over the last decades, there has been a large effort to produce reliable and efficient computer algorithms for the analysis and prediction of DNA binding sites (2).

These algorithms now have an extraordinary ability to predict with high accuracy how proteins bind single sites (3,4). At the same time, use of these highly accurate models to predict where additional binding sites might occur typically finds a wealth of sites that are not physiologically relevant (2). A rule of thumb to predict actual binding is that relevant sites often are positioned close to each other to act cooperatively (5). Clever refinement of this idea has led to heuristic approaches that have proved very successful at predicting the main gene expression trends on a genomic scale (6–12). The middle ground between detailed single-site and broad genomic predictions, however, still remains largely unexplored.

Here, we develop a quantitative framework that accurately integrates sequence statistics with a biophysical model for multidomain binding on nonadjacent DNA sites using as a prototype system the *lac* operon. This choice is motivated by two key features of the *lac* operon.

First, the very simple, yet extremely powerful, original idea of the *lac* repressor preventing transcription upon binding to the operator DNA in the promoter region has continued to evolve over the years to uncover a highly sophisticated mechanism that goes beyond simple binding events (13). It now incorporates an activator and two additional binding sites for the repressor outside the promoter region. These two additional sites are orders of magnitude weaker than the main site and by themselves do not affect transcription substantially. In combination with the main site, however, they can increase repression of transcription by a factor of ~100 (14,15).

Second, there is extremely detailed information about the *lac* operon that offers the possibility of considering the actual mode of binding. This point is important, because the precise sequence has been shaped by evolution through the actual biophysical mechanism. The available information includes detailed quantitative models of how the *lac* repressor binds to two sites simultaneously (16,17) and to the three sites for the repressor together with the effects of the catabolite activator protein (CAP) (18,19). The molecular and cellular parameter values needed by the models are also available, including the in vivo free energy of binding, the energetic costs of bending and twisting DNA upon two-site binding, and the effective transcription rate as a function of the binding state of the repressor (13,20).

Therefore, the *lac* operon provides an efficient platform to accurately test multisite models. In this classical example, without considering the two additional sites, no matter how good the single site model is, it would be off by a factor of ~100.

The focus here is to provide an avenue to extend traditional biophysical single-domain-binding models (21–23) to incorporate the details of multidomain binding, which are inherently different from those of single-domain binding of multiple transcription factors. The traditional approach considers the interaction of a transcription factor (TF) with a DNA site (*S*1) as a binding reaction of the type

Submitted June 8, 2010, and accepted for publication August 4, 2010.
*Correspondence: j.vilar@ikerbasque.org
Editor: Herbert Levine.
© 2010 by the Biophysical Society
0006-3495/10/10/2408/6 $2.00
doi: 10.1016/j.bpj.2010.08.006

$TF + S1 \Leftrightarrow TF \cdot S1$. The strength of the binding is typically assessed through position weight matrix (PWM) scores, which are directly related to the binding energy of the DNA-protein interaction (3,24). The extension to multidomain transcription factors in the presence of additional binding sites, denoted $S2$ and $S3$, has to consider also reactions of the type $TF \cdot S1 + S2 \Leftrightarrow S2 \cdot TF \cdot S1$ and $S2 \cdot TF \cdot S1 + S3 \Leftrightarrow S3 \cdot TF \cdot S1 + S2$. These more complex reactions account for binding of one domain of the TF while its other domain is still bound to DNA, and they usually involve looping the DNA between each pair of simultaneously bound sites.

The multisite approach is explicitly implemented by first considering the three *lac* operators as DNA signals. They are used to construct a probabilistic model that provides PWM scores for binding of a *lac* repressor domain to these and similar mutated sequences. The scores are subsequently linked parametrically to binding free energies and incorporated directly into a detailed biophysical model of transcription regulation that takes into account multidomain binding to multiple binding sites. The model considers a decomposition of the free energy of the protein-DNA complex into different modular contributions. The link between scores and free energies is calibrated by fitting the model to a subset of experimental transcription data. The calibrated model is then tested with different sets of data (Fig. 1).

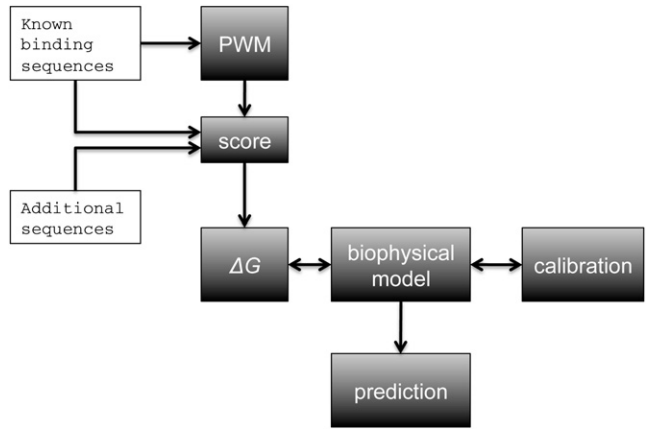

FIGURE 1 Integration of sequence statistics into predictive biophysical multidomain models. The approach is implemented by first considering the three operators as DNA signals. They are used to construct a probabilistic model that provides binding scores for these and similar mutated sequences. The scores are subsequently linked parametrically to binding free energies and incorporated directly into a detailed biophysical model of transcription regulation. The link between scores and free energies is calibrated by fitting the model to a subset of experimental data. The calibrated model is then tested with different sets of data.

METHODS

From sequence to score

The PWM method is used to describe repressor-operator binding (3,24). It assigns a score, S, to the sequence $X = x_1 x_2 \ldots x_w$ according to

$$S = \sum_{i=1}^{w} \ln \frac{p_{xi}}{q_x}, \tag{1}$$

where p_{xi} is the estimated probability of having the nucleotide x at position i of the binding site and q_x is the background frequency of that nucleotide. Taking into account small sample size, p_{xi} is estimated from the observed positional frequency as

$$p_{xi} = \frac{n_{xi} + 1}{N + 4}, \tag{2}$$

where n_{xi} is the number of sites having a nucleotide x at position i and N is the total number of sites in the training set. In our case, we have only three sequences in the training set corresponding to the three operators.

From score to free energy

We assume a linear relationship to transform the score, S, of each sequence into the interaction free energies, e, between the *lac* repressor domain and the DNA site:

$$e = aS + b, \tag{3}$$

where a and b are constants to be inferred from experiments. With this linear assumption, a selects the energy units and b the reference zero of energy.

Multidomain binding

The *lac* repressor is a tetramer consisting of two dimeric DNA binding domains. Multidomain binding is taken into account by decomposing the free energy of the protein-DNA complex into different modular contributions, including positional, interaction, and conformational free energies (19,25).

The positional free energy, p, accounts for the cost of bringing the *lac* repressor to its DNA binding site. Its dependence on the repressor concentration, n, is given by $p = p^\circ - RT\ln n$, where p° is the positional free energy at 1M. Interaction free energies, e, arise from the physical contact between a binding domain and DNA site. Thus, when only a single domain is involved, the free energy of binding is given by $\Delta G = e + p$. For two domains, denoted by subscripts 1 and 2, the free energy of binding is given by $\Delta G = e_1 + e_2 + c + p$. Conformational free energies, c, account for changes in DNA and repressor conformation, which are needed to accommodate multiple simultaneous interactions (Fig. 2).

All these contributions to the free energy, taking into account the three operators for specific binding of the *lac* repressor, can be expressed in mathematical terms as

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

where s_1, s_2, and s_3 are state variables that can take the values 0 and 1 to indicate whether (= 1) or not (= 0) the repressor is bound to the operators O_1, O_2, and O_3, respectively; and s_{L12}, s_{L13}, and s_{L23} are variables that indicate whether (= 1) or not (= 0) DNA forms the loops O_1-O_2, O_1-O_3, and O_2-O_3, respectively. The subscripts of the different contributions to the free energy have the same meaning as those of the corresponding binary variables. The infinity in the last term of the free energy implements that two

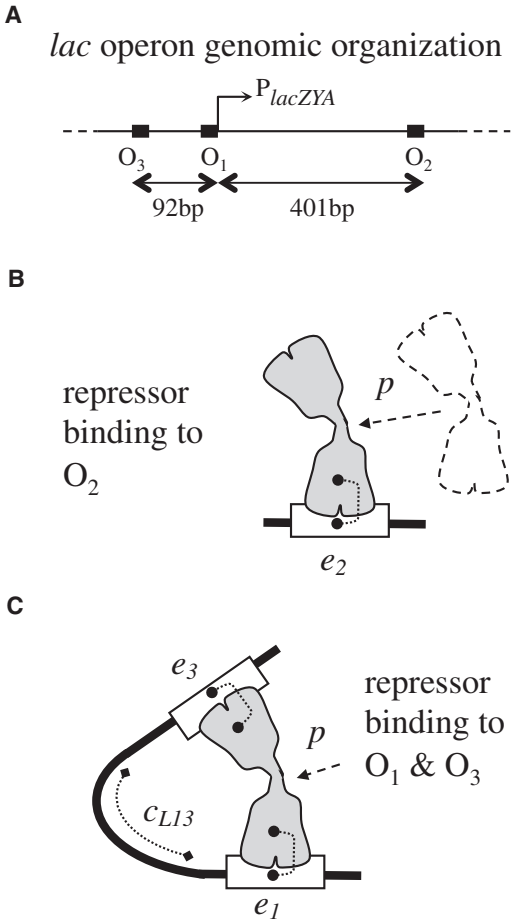

FIGURE 2 Operator locations on DNA and binding of the *lac* repressor. (*A*) The main (O_1) and two auxiliary (O_2 and O_3) operators are shown as black rectangles on the black line representing DNA. Binding of the *lac* repressor to O_1 prevents transcription of the three *lac*ZYA genes. (*B*) A repressor is shown bound to O_2. The free energy of binding is $\Delta G = e_2 + p$. (*C*) A repressor is shown looping DNA by binding simultaneously to O_1 and O_3. The free energy of this binding configuration is $\Delta G = e_1 + e_3 + c_{L13} + p$.

loops that share one operator cannot be present simultaneously by assigning an infinite free energy to those states (18).

The set of six state variables, denoted by $s = (s_1, s_2, s_3, s_{L12}, s_{L13}, s_{L23})$, describes the specific binding configuration of the repressor-DNA complex. For instance, a repressor bound to O_2 is specified by s = (0, 1, 0, 0, 0, 0); a repressor bound to O_1 and O_3 looping the intervening DNA, by s = (1, 0, 1, 0, 1, 0); and three repressors bound, one to each operator, by s = (1, 1, 1, 0, 0, 0). The specific value of the free energy is obtained by substituting the values of the state variables in the expression of the free energy. This description in terms of state variables can be visualized as a factor graph (Fig. 3).

The probability of any of these states depends exponentially on its free energy and is obtained from statistical thermodynamics as

$$P_s = \frac{e^{-\Delta G(s)/RT}}{Z}, \tag{5}$$

where RT is the gas constant times the absolute temperature. The partition function, $Z = \sum_s e^{-\Delta G(s)/RT}$, is used as a normalization factor.

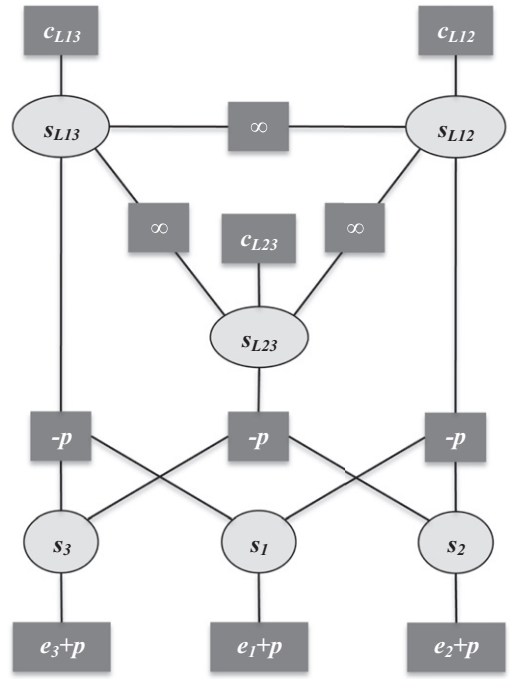

FIGURE 3 Factor graph for the free-energy components of the multisite *lac* repressor-operator binding. The free energy of the system, $\Delta G(s)$, as a function of the state variables, $s = (s_1, s_2, s_3, s_{L12}, s_{L13}, s_{L23})$, has a graphical representation in the form of a factor graph. The round nodes represent state variables and the rectangular nodes represent contributions to the free energy. The quantity in the rectangular node is present in the free energy when all its connecting state variables are equal to 1. The experimental values for wild-type parameters are $e_1 = -27.8$ kcal/mol, $e_2 = -26.3$ kcal/mol, $e_3 = -24.1$ kcal/mol, $c_{L12} = 23.35$ kcal/mol, $c_{L13} = 22.05$ kcal/mol, and $c_{L23} = 23.50$ kcal/mol. The dependence on the *lac* repressor concentration, n, is given by the positional free energy, $p = p^\circ - RT\ln n$, with $p^\circ = 15$ kcal/mol.

Transcriptional control

Gene expression in the *lac* operon is completely abolished when the repressor is bound to O_1; otherwise, transcription takes place either at an activated maximum rate, Γ_{max}, when O_3 is free or at a basal reduced rate, $\chi\Gamma_{max}$, when O_3 is occupied. This reduction by a factor χ arises because binding of the repressor to O_3 prevents CAP from activating transcription (13,18).

The transcription rate $\Gamma(s)$ can be expressed in terms of state variables as

$$\Gamma(s) = \Gamma_{max}(1 - s_1)(\chi s_3 + (1 - s_3)). \tag{6}$$

With this approach, the effective transcription rate,

$$\overline{\Gamma} = \sum_s \Gamma(s) P_S = \frac{1}{Z}\sum_s \Gamma(s) e^{-\Delta G(s)/RT}, \tag{7}$$

is obtained by computing the thermodynamic average over all the representative states, namely, by performing the sum above over all possible combinations of values of s.

Model calibration

The overall model has only two free parameters: the constants a and b that relate scores to free energies of binding. Their values are inferred by minimizing the square logarithmic error between measured and model

normalized transcription ($\overline{\Gamma}/\Gamma_{max}$). The values of the other four parameters, three conformational free energies and CAP activation, are taken from the experimental data. Explicitly, the value $\chi = 0.03$ was reported by Oehler et al. (26); the value $c_{L12} = 23.35$ kcal/mol was obtained by Saiz and Vilar (20) from experimental data in the Oehler et al. study (26); the values $c_{L13} = 22.05$ kcal/mol and $c_{L23} = 23.50$ kcal/mol were obtained from the value of c_{L12} by taking into account the dependence of the conformational free energy on the distance between operators (20,27,28) and the stabilization of the O_1-O_3 loop by CAP (29,30).

RESULTS AND DISCUSSION

We applied the multisite approach to classic experiments on the *lac* operon that considered gene expression for different repressor concentrations in *E. coli* strains covering all eight possible combinations of operator deletions (14). The sequences of the three wild-type (WT) operators O_1, O_2, and O_3 were used to compute the PWM from which we obtained the scores for these three operators and their respective deletions, O_{1M}, O_{2M}, and O_{3M} (see Table 1). The scores correctly ranked the three WT operators according to their measured strength and consistently ranked all the deletions below all the WT operators.

The values of parameters a and b were obtained by fitting the model to the experimental transcription data using

$$\begin{aligned}\Delta G(s) = {} & (p + aS_1 + b)s_1 + (p + aS_2 + b)s_2 \\ & + (p + aS_3 + b)s_3 + (c_{L12} - ps_1s_2)s_{L12} \\ & + (c_{L13} - ps_1s_3)s_{L13} + (c_{L23} - ps_2s_3)s_{L23} \\ & + \infty(s_{L12}s_{L13} + s_{L12}s_{L23} + s_{L13}s_{L23}) \qquad (8)\end{aligned}$$

as the free energy of the system. This expression is obtained after substitution of the relation $e = aS + b$ in Eq. 4. In this way, the binding is described by the PWM scores, S_1, S_2, and S_3, for each site together with the conformational contributions to the free energy from DNA looping (28).

The model, with just a and b as free parameters, is able to fit the experimental data (14) within 0.29-fold accuracy over a 10,000-fold range of transcriptional activity (Fig. 4 *A*). In total, there are 22 experimental points, accounting for eight operator configurations, three different repressor concentrations, and different functional forms of the transcription curves. The value *FA* that quantifies the ability of the model to capture the experimental data within *FA*-fold accuracy is explicitly defined for a set of N experimental, Γex, and computed, Γcp, transcription rates through the expression $N\log(1 + FA)^2 = \sum_{i=1}^{N}\log(\Gamma ex_i/\Gamma cp_i)^2$, and it indicates that typically measured and computed values differ from each other by a factor of $1 + FA$.

The interaction free energies obtained from the model for the best-fit a and b parameters and the corresponding experimental in vivo values (18) are shown in Table 1. The results of the model exhibit good agreement with the available experimental data. In terms of dissociation constants, the differences between the predicted and observed values are within the twofold range (Table 1). An advantage of the approach we have followed is that the in vivo free energies, and the corresponding dissociation constants, take into account implicitly the effects of nonspecific binding. The reason is that their values are measured with respect to the reference state with no repressor bound to the operators, which includes the repressors in solution in the cytosol as well as the repressors bound nonspecifically to DNA (for a detailed quantitative discussion, see Appendix II of Vilar and Leibler (16)).

To test the predictive potential of the multisite model, we used experimental data sets for two operator configurations to infer the values of parameters a and b and then used the calibrated model to predict the transcriptional activity for the other six configurations (Fig. 4 *B*). The accuracy of the model at predicting new data decreases only slightly with respect to the all-fit accuracy. In principle, only two experimental data points would be needed to calibrate the model, because there are only two free parameters. Indeed, just two experimental points can be used to calibrate the model with just a slight additional decease in global accuracy (Fig. 4 *C*). Therefore, without using any free energy of binding, the multisite model is able to accurately predict gene expression curves over a 10,000-fold range for eight different *E. coli* strains covering all possible combinations of operator deletions from just two experimental calibration data points and the sequences of the six DNA sites involved.

There is an important prediction that goes beyond the experimentally observed free energies of binding. The deletion O_{1M} of the main operator O_1 involved the mutation of just three DNA basepairs. As a consequence, the model

TABLE 1 Operator sequences and their statistical and binding properties

Name	Sequence	S	$aS + b$ (kcal/mol)	e (kcal/mol)	K_D^{sc} (nM)	K_D^{ex} (nM)
O_1	AATTGTGAGCGGATAACAATT	−13.38	−27.62	−27.8	0.728	0.54
O_2	AAATGTGAGCGAGTAACAACC	−12.17	−25.94	−26.3	12.1	6.62
O_3	GGCAGTGAGCGCAACGCAATT	−10.95	−24.25	−24.1	201	259
O_{1M}	AATTGTTAGCGGAGAAGAATT	−9.51	−22.26	N/A	5600	N/A
O_{2M}	GAAGGTTAATGAATAGCACCC	−5.12	−16.16	N/A	1.44×10^8	N/A
O_{3M}	TCGATCGAGCTCAACGCAATT	−4.71	−15.60	N/A	3.37×10^8	N/A

The PWM score, S, for a given operator sequence is used to estimate its interaction free energy with the *lac* repressor as $aS + b$, with $a = 1.387$ kcal/mol and $b = -9.064$ kcal/mol. The experimental values of these free energies, e, are from Saiz and Vilar (18). Dissociation constants are computed as $K_D^{sc} = e^{(aS+b+p^\circ)/RT}$ for the predictions from PWM scores and as $K_D^{ex} = e^{(e+p^\circ)/RT}$ for the experimental data. N/A stands for data not available.

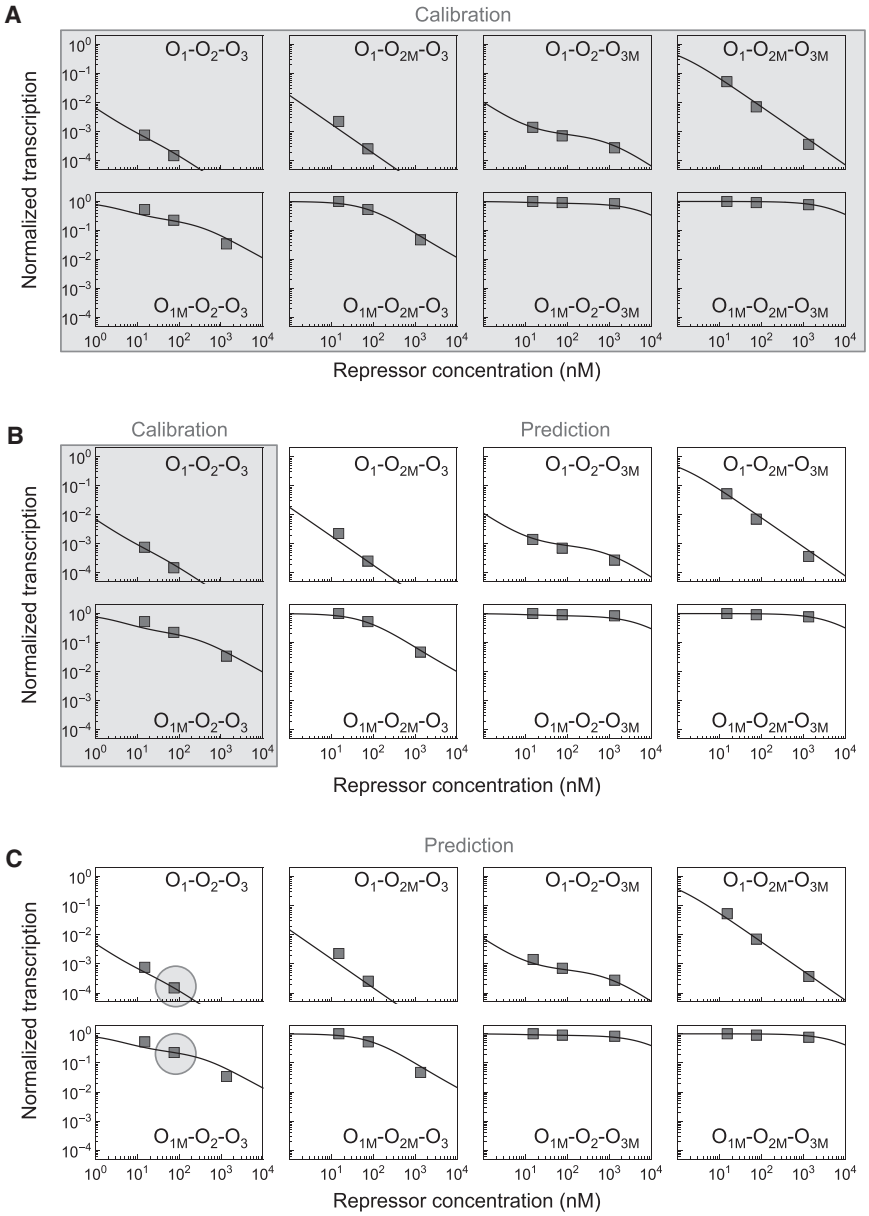

FIGURE 4 Model calibration and prediction of the transcriptional activity as a function of the repressor concentration. The normalized transcription ($\bar{\Gamma}/\Gamma_{max}$) was obtained for WT and seven mutants accounting for all the combinations of deletions of the three operators. For each of the eight cases, the results of the model (*solid lines*) as a function of the repressor concentration are compared with the experimental data from Oehler et al. (14) (*squares*). The particular set of WT or deleted operators is indicated for each curve; for instance, O_1-O_2-O_3 corresponds to the WT *lac* operon and O_{1M}-O_{2M}-O_{3M} to the mutant with all three operators deleted. The values of the experimental parameters used are $c_{L12} = 23.35$ kcal/mol, $c_{L13} = 22.05$ kcal/mol, $c_{L23} = 23.50$ kcal/mol, and $\chi = 0.03$. The PWM scores, S, for each site are as shown in Table 1. (*A*) Parameter values $a = 1.387$ kcal/mol and $b = -9.064$ kcal/mol, which connect interaction free energies with scores, $e = aS + b$, were obtained by fitting the model to all the experimental transcription data. (*B*) Parameter values $a = 1.348$ kcal/mol and $b = -9.531$ kcal/mol were obtained by fitting the model to the experimental data for operator configurations O_1-O_2-O_3 and O_{1M}-O_2-O_3. The model accurately predicts the normalized transcription for the other six operator configurations. (*C*) Only two experimental points (*large gray circles*) are used to obtain the parameter values $a = 1.462$ kcal/mol and $b = -8.208$ kcal/mol. The model is still able to accurately predict the normalized transcription for the remaining 20 experimental points.

predicts for O_{1M} an increase in free energy of 5.4 kcal/mol with respect to O_1, or, equivalently, an ~8000-fold increase of the dissociation constant, which is substantial but still remains relatively close to the free energy of binding to O_3, the weakest WT operator (Table 1). We found that such a decrease has transcriptional consequences that make it distinguishable from a complete deletion (Fig. 5). Thus, the multisite approach is able not only to both accurately predict gene expression and recover known free energies but also to obtain precise affinity estimates for very weak sites that were assumed not to bind the *lac* repressor.

Typically, the effects of a given sequence depend on the context. This dependence has been noted explicitly as one of the main limiting factors for identifying physiologically relevant sites and for linking statistical sequence information, such as PWM scores, to transcriptional activity (31). This fundamental problem in gene regulation is believed to result from the interplay among multiple DNA sites in orchestrating the binding patterns of transcription factors that control gene expression (2). The approach presented here overcomes this limitation by using detailed biophysical modeling of multidomain binding to directly connect statistical sequence information with transcriptional activity. We have shown that for the prototypical *lac* operon, which relies on a cluster of three nonadjacent sites over a 0.5-kb DNA region to control transcription, this multisite approach accurately recapitulates the observed transcriptional activity over a 10,000-fold range for all the possible combinations of operator deletions.

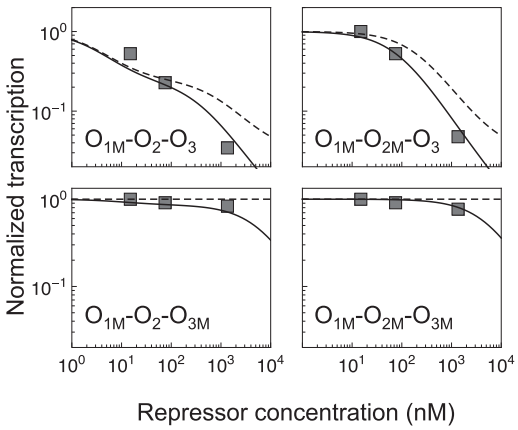

FIGURE 5 Complete deletions versus weak binding. The normalized transcription ($\overline{\Gamma}/\Gamma_{max}$) for the four configurations with O_{1M} is shown for the model as in Fig. 4 *A* (*solid line*); for the model assuming that the free energy of binding to O_{1M} is infinite, as in a complete deletion (*dashed line*); and for the experimental data from Oehler et al. (14) (*squares*).

This work was supported by the Ministerio de Ciencia e Innovación under grant FIS2009-10352.

REFERENCES

1. Jacob, F., and J. Monod. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* 3:318–356.
2. Wasserman, W. W., and A. Sandelin. 2004. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* 5:276–287.
3. Stormo, G. D. 2000. DNA binding sites: representation and discovery. *Bioinformatics.* 16:16–23.
4. Zhao, Y., D. Granas, and G. D. Stormo. 2009. Inferring binding energies from selected binding sites. *PLOS Comput. Biol.* 5:e1000590.
5. Tronche, F., F. Ringeisen, …, M. Pontoglio. 1997. Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome. *J. Mol. Biol.* 266:231–245.
6. Liu, R., R. C. McEachin, and D. J. States. 2003. Computationally identifying novel NF-kappa B-regulated immune genes in the human genome. *Genome Res.* 13:654–661.
7. van Batenburg, M. F., H. Li, …, O. C. Meijer. 2010. Paired hormone response elements predict caveolin-1 as a glucocorticoid target gene. *PLoS ONE.* 5:e8839.
8. Bussemaker, H. J., B. C. Foat, and L. D. Ward. 2007. Predictive modeling of genome-wide mRNA expression: from modules to molecules. *Annu. Rev. Biophys. Biomol. Struct.* 36:329–347.
9. Tavazoie, S., and G. M. Church. 1998. Quantitative whole-genome analysis of DNA-protein interactions by in vivo methylase protection in *E. coli*. *Nat. Biotechnol.* 16:566–571.
10. Bussemaker, H. J., H. Li, and E. D. Siggia. 2001. Regulatory element detection using correlation with expression. *Nat. Genet.* 27:167–171.
11. Markstein, M., R. Zinzen, …, M. Levine. 2004. A regulatory code for neurogenic gene expression in the *Drosophila* embryo. *Development.* 131:2387–2394.
12. van Nimwegen, E. 2007. Finding regulatory elements and regulatory motifs: a general probabilistic framework. *BMC Bioinformatics.* 8 (*Suppl 6*):S4.
13. Müller-Hill, B. 1996. The *lac* Operon: A Short History of a Genetic Paradigm. Walter de Gruyter, Berlin; New York.
14. Oehler, S., E. R. Eismann, …, B. Müller-Hill. 1990. The three operators of the *lac* operon cooperate in repression. *EMBO J.* 9:973–979.
15. Mossing, M. C., and M. T. Record, Jr. 1986. Upstream operators enhance repression of the lac promoter. *Science.* 233:889–892.
16. Vilar, J. M. G., and S. Leibler. 2003. DNA looping and physical constraints on transcription regulation. *J. Mol. Biol.* 331:981–989.
17. Alberts, B., A. Johnson, …, P. Walter. 2008. Molecular Biology of the Cell. Garland Science, New York.
18. Saiz, L., and J. M. G. Vilar. 2008. Ab initio thermodynamic modeling of distal multisite transcription regulation. *Nucleic Acids Res.* 36: 726–731.
19. Vilar, J. M. G., and L. Saiz. 2005. DNA looping in gene regulation: from the assembly of macromolecular complexes to the control of transcriptional noise. *Curr. Opin. Genet. Dev.* 15:136–144.
20. Saiz, L., and J. M. G. Vilar. 2006. DNA looping: the consequences and its control. *Curr. Opin. Struct. Biol.* 16:344–350.
21. Djordjevic, M., A. M. Sengupta, and B. I. Shraiman. 2003. A biophysical approach to transcription factor binding site discovery. *Genome Res.* 13:2381–2390.
22. Liu, X., and N. D. Clarke. 2002. Rationalization of gene regulation by a eukaryotic transcription factor: calculation of regulatory region occupancy from predicted binding affinities. *J. Mol. Biol.* 323:1–8.
23. Roider, H. G., A. Kanhere, …, M. Vingron. 2007. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics.* 23:134–141.
24. Berg, O. G., and P. H. von Hippel. 1987. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.* 193:723–750.
25. Saiz, L., and J. M. G. Vilar. 2006. Stochastic dynamics of macromolecular-assembly networks. *Mol. Syst. Biol.* 2, 2006.0024.
26. Oehler, S., M. Amouyal, …, B. Müller-Hill. 1994. Quality and position of the three *lac* operators of *E. coli* define efficiency of repression. *EMBO J.* 13:3348–3355.
27. Müller, J., S. Oehler, and B. Müller-Hill. 1996. Repression of *lac* promoter as a function of distance, phase and quality of an auxiliary *lac* operator. *J. Mol. Biol.* 257:21–29.
28. Saiz, L., J. M. Rubi, and J. M. G. Vilar. 2005. Inferring the in vivo looping properties of DNA. *Proc. Natl. Acad. Sci. USA.* 102:17642–17645.
29. Hudson, J. M., and M. G. Fried. 1990. Co-operative interactions between the catabolite gene activator protein and the *lac* repressor at the lactose promoter. *J. Mol. Biol.* 214:381–396.
30. Saiz, L., and J. M. G. Vilar. 2007. Multilevel deconstruction of the in vivo behavior of looped DNA-protein complexes. *PLoS ONE.* 2:e355.
31. Veprintsev, D. B., and A. R. Fersht. 2008. Algorithm for prediction of tumour suppressor p53 affinity for binding sites in DNA. *Nucleic Acids Res.* 36:1589–1598.