BMC Structural Biology

Research article

Open Access

Identification of DNA-binding protein target sequences by physical effective energy functions: free energy analysis of lambda repressor-DNA complexes.

Elisabetta Moroni*[1,2], Michele Caselle[1] and Federico Fogolari[3]

Address: [1]Dipartimento di Fisica Teorica, Università di Torino and INFN, Via P. Giuria 1, 10125 Torino, Italy, [2]Dipartimento di Fisica G. Occhialini, Università di Milano-Bicocca and INFN, Piazza delle Scienze 3, 20156 Milano, Italy and [3]Dipartimento di Scienze e Tecnologie Biomediche, Università di Udine, P.le Kolbe 4, 33100 Udine, Italy

Email: Elisabetta Moroni* - moroni@to.infn.it; Michele Caselle - caselle@to.infn.it; Federico Fogolari - ffogolari@makek.dstb.uniud.it

* Corresponding author

Published: 27 September 2007

BMC Structural Biology 2007, **7**:61 doi:10.1186/1472-6807-7-61

Received: 20 February 2007

Accepted: 27 September 2007

This article is available from: http://www.biomedcentral.com/1472-6807/7/61

Abstract

Background: Specific binding of proteins to DNA is one of the most common ways gene expression is controlled. Although general rules for the DNA-protein recognition can be derived, the ambiguous and complex nature of this mechanism precludes a simple recognition code, therefore the prediction of DNA target sequences is not straightforward. DNA-protein interactions can be studied using computational methods which can complement the current experimental methods and offer some advantages. In the present work we use physical effective potentials to evaluate the DNA-protein binding affinities for the λ repressor-DNA complex for which structural and thermodynamic experimental data are available.

Results: The binding free energy of two molecules can be expressed as the sum of an intermolecular energy (evaluated using a molecular mechanics forcefield), a solvation free energy term and an entropic term. Different solvation models are used including distance dependent dielectric constants, solvent accessible surface tension models and the Generalized Born model. The effect of conformational sampling by Molecular Dynamics simulations on the computed binding energy is assessed; results show that this effect is in general negative and the reproducibility of the experimental values decreases with the increase of simulation time considered. The free energy of binding for non-specific complexes, estimated using the best energetic model, agrees with earlier theoretical suggestions. As a results of these analyses, we propose a protocol for the prediction of DNA-binding target sequences. The possibility of searching regulatory elements within the bacteriophage λ genome using this protocol is explored. Our analysis shows good prediction capabilities, even in absence of any thermodynamic data and information on the naturally recognized sequence.

Conclusion: This study supports the conclusion that physics-based methods can offer a completely complementary methodology to sequence-based methods for the identification of DNA-binding protein target sequences.

Background

Protein-DNA recognition plays an essential role in the regulation of gene expression. Although a significant number of structures of DNA binding proteins have been solved in complex with their DNA binding sites increasing our understanding of recognition principles, most of the questions remain unanswered. Several studies showed that protein-DNA recognition could not be explained by a simple one-to-one correspondence between amino acids and bases [1-3], even if hypothesized hydrogen bonding patterns and definite preferences have been actually found in experimentally solved structures [1]. Moreover regulatory proteins are known to recognize specific DNA sequences directly through atomic contacts between protein and DNA and/or indirectly through water-mediated contacts and conformational changes [1,4,5]. The degree of redundancy and flexibility seems to suggest that the recognition mechanism is ambiguous, therefore the prediction of DNA target sequences is not straightforward [6].

DNA protein interactions can be studied using several different computational methods, which could offer several advantages compared to the current experimental methods, more laborious and slow. In the following we will indicate, for simplicity, DNA-binding protein target sequences with the more specific term "transcription factor binding sequences", although the first term is more general.

Computational tools for the identification of Transcription Factors (TF) binding sequences can be organized in two main approaches:

- "sequence based methods" in which a central role is played by the statistical properties of the base distribution in the DNA regions which are expected to be involved in transcriptional regulation (see [7,8] for a general review on the subject).

- "structure based tools" which use the structural information on protein-DNA complexes derived from X-ray crystallography and Nuclear Magnetic Resonance.

The main focus of this paper is on the second approach, although the best results will likely be obtained by tools able to combine in a clever way these two approaches.

Sequence based methods

This type of algorithms can in turn be divided into two broad groups:

i) enumerative methods, which explore all possible motifs up to a certain length (see e.g. [9-16]).

ii) local search algorithms, including expectation maximization and various flavours of Gibbs sampling (see e.g. [17-20]).

It is important to stress that this type of studies cannot be based exclusively on the statistical features of the DNA regions presumably involved in transcriptional regulation, but must be complemented with independent information about gene regulation. In this respect three important sources of information may be used: the functional annotations collected in public databases, gene expression data on a global scale, and the so called 'phylogenetic footprinting'. In particular this last approach, thanks to the increasing number of sequenced genomes, has proved to be very effective in these last few years (see e.g. [21-30]). The major problem of all these tools is the large number of false positives, above all in the case of higher eukaryotes (for a thorough analysis of this problem see the interesting assessment of TF binding sites discovery tools reported in [31]). It is exactly to cope with this type of problem that it could be important to resort to structure based approaches.

Structure based methods

These methods can be broadly divided into two classes according to a nomenclature adopted in the context of protein structure prediction [32]:

i) those based on knowledge based potentials (mostly statistical effective energy functions, SEEFs);

ii) those based on physical potentials (or physical effective energy functions, PEEFs).

SEEFs are energy functions derived from a dataset of known protein-DNA structures. A set of features is selected (e.g. nucleotide-amino acid contacts, roll angles for DNA bases, interatomic distances, etc.); the process often involves parameter choices, like threshold on distances or interval binning. The statistical properties of these features are compared with a-priori expectations and log-odd scores are derived [6,33-38]. At the most basic level, structures may be used to define contacts among DNA bases and protein amino acids and, for each pair of positions, the occurrences of nucleotides and amino acids contacts are used to derive effective potentials [36]. Moreover a statistical potential, taking into account contact geometry and spatial arrangement of contacting residues can be derived [6]. Recently interesting developments of this approach have been proposed ([38,35,5,39]). The approach suffers from theoretical and practical problems. From the theoretical point of view potentials of mean force are not in general additive and the exact modelization of a-priori expectations (or so-called reference state) may be difficult for complex systems (see e.g. [40]). The main practical problem is the requirement of a large number of sequences or binding experimental data since the available data may be biased towards specific classes of protein-DNA complexes. Moreover datasets generally do not contain unfavourable interactions between amino acids and bases since they entail protein-DNA complexes that occur naturally. Thus the statistical potential may predict correctly the wild type targets as opposed to incorrect ones, but it may not be as good at distinguishing among mutants.

Notwithstanding all caveats usage of SEEFs are widespread in the field of structural predictions. Provided that sufficient data are available these methods are reasonably fast and accurate, as demonstrated for instance in the field of protein structure prediction (see e.g [41]).

A more radical approach is to estimate the free energy of binding starting directly from the available (or homology built) protein-DNA complexes using physical effective energy functions (PEEFs). This approach has been successfully used in many contexts, ranging from estimation of DNA- or protein-ligand binding free energy to estimation of protein-DNA binding free energy (see e.g. [42,43]). There are, however, many problems connected with the approach which are mainly due to:

i) difficulties in estimating entropic effects;

ii) difficulties in properly estimating solvation effects;

No consensus has emerged on the choice of parameters (e.g. inner dielectric constant, surface tension coefficient, forcefield parameters) and on the protocols that should be applied;

iii) difficulties in estimating gas-phase energy with available forcefields which are derived from the analysis of small compounds at equilibrium and do not take into account electrostatic polarization.

In order to get rid as far as possible of all these problems, binding free energies are expressed relative to a reference system and in most computational studies optimal parameters have been chosen for matching experimental data.

As far as protein-DNA complexes are concerned attempts to compute binding free energies using physics based approaches have started in the 1990s. The electrostatic component of the binding free energy has been studied according to continuum methods and its dependence on temperature and salt concentration has been computed [44-46]. Integration of electrostatics with other components including DNA conformational free energy has been extended from DNA-ligand complexes [47] and protein-peptide complexes [48] to protein-DNA complexes [49]. Recently Wojciechowski et al. [50] studied the complex of telomerase end binding protein with single stranded DNA optimizing the weights of different contributions in order to reproduce binding data. The availability of the successful analytical generalized Born model treatment of electrostatics solvation effects enabled computation of binding energies with hybrid molecular mechanics/Generalized Born surface accessibility methods by Jayaram et al. [51]. The group of Kollman developed the molecular mechanics/Poisson Boltzmann surface accessibility (MM/PBSA) methodology and applied it extensively to biomolecular systems (see for a review of these applications [42,43] and [42,43,43,52] for important extensions of these ideas).

However, when MM/GBSA or MM/PBSA energy versus time plots are presented for explicit solvent molecular dynamics simulation snapshots, fluctuations in the range of tens to hundreds of kcal/mol are found, thus posing an issue on the reliability of averages. In this respect SEEFs appear much more robust energy estimation methods.

In a few very recent reports interesting results have been reported concerning the capability of hybrid methods to predict protein-DNA binding sites [53-57]. In this paper we focus on the application of PEEFs to a single DNA binding protein in complex with many different DNA sequences.

The availability of high resolution X-ray crystal structure [58] and suitable experimental data makes the λ repressor-operator complex an interesting system for computational analysis of protein-DNA interaction. The bacteriophage λ repressor protein is a small, 92 amino acid, protein that binds the DNA as a dimer. Each monomer binds to an operator half site. The amino-terminal domain of λ repressor is responsible for DNA binding and the carboxy-terminal domain is primarily responsible for dimerization [59]. Each monomer contains a typical helix-turn-helix motif found in a variety of DNA binding proteins [4,60]. The free energy of binding of λ repressor for wild-type O_R1 operator DNA and of all possible single base-pair substitutions within the operator have been experimentally measured using the filter binding assay technique and changes in the free energy of binding caused by the mutations have been determined [61]. Besides being a perfect playground to test our methods, the so called "λ-switch" in which the λ repressor is involved is very interesting in itself (for a review see [62]). This "genetic" switch is tightly regulated by the λ repressor and the *Cro* proteins. In these last years this system, due to its relative simplicity and to the availability of rather precise experimental data attracted a lot of interest and various models (see for instance [63-65] and references therein) have been proposed to describe its behaviour. Despite these efforts in all these models there are still a few open problems which need to be understood. In particular it has been recently realized that in order to ensure the remarkable stability of the λ switch one should require a very high non-specific affinity both for the λ repressor and for *Cro* [65,66]. Such a prediction is very difficult to test experimentally but could rather directly be evaluated with the tools which we shall discuss in this paper. In fact one of the main goal of the test which we shall perform on the λ repressor will be the evaluation of its non-specific binding energy and the comparison with the prediction of the model discussed in [66].

In the present work we apply different techniques to evaluate the binding affinities by means of computational methods. It is assumed that the relative free energy of binding of a protein to different DNA sequences may be expressed as the sum of a molecular mechanical term, that includes the non-bonded electrostatic and Van der Waals contributions, and a hydration term that can be further split in a polar and a hydrophobic contribution. Due to the peculiar nature of hydrogen bonds similar alternative models are tested where an energy term proportional to the number of hydrogen bonds is added.

The systems studied here differ only in one or two base-pairs and therefore the inaccuracies implicit in the assumption of rigid docking, of the solvation model, of the treatment of entropy and in lack of a complete conformational search for side chains at protein-DNA interface should mostly cancel out in comparison. The aims of this paper are:

1) to provide an assessment of the accuracy of different methods and protocols by comparison with experimental data;

2) to provide a reliable estimate of non-specific binding energies;

3) to propose a protocol for the prediction of DNA-binding target sequences which makes no use of sequence information.

To pursue these objectives we make use of extensive computations and address several specific issues. In particular:

i) we estimate optimal weights for different contributions to DNA-protein binding free energies using different solvation models;

ii) for 52 single base-pair mutants we perform 1 ns molecular dynamics (MD) runs and we assess the effect of MD on the computed binding energies;

iii) we compute MM/GBSA binding energies for one thousand complexes where the bases of the double stranded DNA are substituted according to randomly generated DNA sequences in order to estimate non-specific binding free energy;

iv) we scan the entire bacteriophage λ genome with the scoring profiles obtained from free energy computations. One of the profiles is obtained making use only of the structural data available for a single molecular complex, with no sequence information.

The statistical analysis of the results show that computational methods may offer a predictive tool truly complementary to sequence-based identification of DNA-binding protein target sequences. This is particularly important in view of the emergence of consensus protocols where the independence of the different methods is a prerequisite.

Results and discussion

Binding free energy changes between the λ repressor dimer and the DNA operator mutants have been calculated using different methodologies, as described in the *Methods* section.

MM/DDDC-OONS

We calculated the binding free energy between the λ repressor dimer and the DNA operators, after having energy-minimized every complex using a distance dependent dielectric constant (1r, 2r, 4r, 8r, respectively, in order to match subsequent energy evaluation). The interaction energy between the protein and the DNA, $\Delta U(\vec{r}_1,...,\vec{r}_n)$, has been evaluated using four values for ε (1r, 2r, 4r, 8r) then the solvation term ΔG_{solv} has been determined according to the model of Oobatake et al. [54] using Eq. 2. The best scaling factors have been determined (together with the standard deviation computed according to Eq. 9) fitting the set of experimentally measured protein-DNA binding affinities and are reported in Table 1. The addition of a specific hydrogen bond term reduces the coefficients of the electrostatic term. The RMSD and the correlation coefficient r have been computed and a leave-one-out scheme has been adopted, in order to verify the performance of the model (Table 2). The same analysis has been performed for 5000 replicas of the dataset with one third of the set left out and used for cross-validation. The average RMSD and correlation are essentially the

Table 1: Optimal scaling factors for the MM/DDDC-OONS model and the MM/DDDC-HP model. Standard deviations (see Methods section) are given in parentheses.

MM/DDDC-OONS	**1r**	**2r**	**4r**	**8r**
x_{vdW}	0.075 (0.071)	0.041 (0.100)	0.043 (0.111)	0.025 (0.116)
x_{DDDC}	0.083 (0.019)	0.184 (0.037)	0.359 (0.076)	0.802 (0.159)
x_{OONS}	0.072 (0.070)	-0.020 (0.075)	-0.084 (0.061)	-0.043 (0.076)
x_{const}	65.254 (20.806)	73.523 (21.615)	75.782 (23.697)	64.715 (26.116)
MM/DDDC-OONS (+HB)	**1r**	**2r**	**4r**	**8r**
x_{vdW}	0.075 (0.072)	0.040 (0.100)	0.042 (0.109)	0.034 (0.113)
x_{DDDC}	0.068 (0.030)	0.154 (0.046)	0.286 (0.089)	0.637 (0.179)
x_{OONS}	0.075 (0.071)	-0.019 (0.075)	-0.066 (0.061)	-0.010 (0.076)
x_{HB}	-0.151 (0.249)	-0.166 (0.152)	-0.226 (0.145)	-0.269 (0.144)
x_{const}	57.867 (24.223)	68.819 (21.992)	69.397 (23.718)	58.463 (25.680)
MM/DDDC-HP	**1r**	**2r**	**4r**	**8r**
x_{vdW}	0.144 (0.065)	0.215 (0.106)	0.085 (0.115)	0.133 (0.128)
x_{DDDC}	0.076 (0.017)	0.221 (0.033)	0.402 (0.075)	0.844 (0.153)
x_P	-0.018 (0.008)	-0.027 (0.011)	-0.007 (0.011)	-0.012 (0.012)
x_H	-0.028 (0.007)	-0.021 (0.007)	-0.013 (0.008)	-0.015 (0.008)
x_{const}	-4.023 (26.223)	21.075 (24.822)	38.294 (32.091)	31.927 (29.069)
MM/DDDC-HP + (HB)	**1r**	**2r**	**4r**	**8r**
x_{vdW}	0.144 (0.065)	0.221 (0.104)	0.105 (0.111)	0.175 (0.123)
x_{DDDC}	0.059 (0.027)	0.185 (0.040)	0.281 (0.089)	0.623 (0.172)
x_P	-0.018 (0.008)	-0.029 (0.011)	-0.011 (0.011)	-0.018 (0.011)
x_H	-0.029 (0.007)	-0.022 (0.007)	-0.017 (0.008)	-0.017 (0.007)
x_{HB}	-0.181 (0.218)	-0.210 (0.138)	-0.327 (0.143)	-0.332 (0.137)
x_{const}	-13.355 (28.624)	12.784 (25.076)	21.219 (31.635)	20.937 (28.041)

same reported for the leave-one-out scheme reported in Table 2. From the same analysis variances of the coefficients have been estimated with essentially the same results as those reported in Table 1.

The best correlation coefficient (r = 0.703 for MM/DDDC-OONS(+HB) model) has been obtained for $\varepsilon = 4r$, although values of $\varepsilon = 2r$ and $\varepsilon = 8r$ gave very similar results for both MM/DDDC-OONS and MM/DDDC-OONS(+HB) models. Except for the MM/DDDC-OONS(+HB) model with $\varepsilon = 1r$, the F-statistic shows that the model is significant (p < 0.001). The dielectric constant $\varepsilon = 1r$, which gives the worst results tends in many cases to overestimate binding free energy changes lower than 1.0 kcal/mol whereas binding free energy changes greater than 2 kcal/mol are underestimated. A similar behaviour has been observed for $\varepsilon = 2r$, $4r$, $8r$, even if these models are able to better reproduce binding free energy changes, in particular improvements have been obtained for values lower than 1.0 kcal/mol (Figure 1). The analysis of the best scaling coefficients is not straightforward because there is a strong correlation between the energy terms. For instance, for all ε models the electrostatic term is strongly anticorrelated with the OONS solvation term.

Moreover the estimated variance of coefficients is often very large. Notwithstanding these difficulties it is worth noting that some terms appear to be particularly important. For instance each protein-DNA hydrogen bond (when explicitly included in the model) appears to contribute -0.15 to -0.27 kcal/mol, depending on the electrostatic model assumed.

As expected the electrostatic term is reduced when hydrogen bonds are taken into account separately. For $\varepsilon = 2r$ the best scaling coefficient x_{DDDC} changes from 0.154 to 0.182 upon removal of the term proportional to the number of hydrogen bonds.

The correlation between the different contributions is reflected in the changes, with changing dielectric model, of the OONS term scaling factor, which is always strongly reduced by the scaling factor ranging from 0.075 to -0.066. Finally the constant term which takes into account

Table 2: RMSD and correlation coefficients (r) between calculated and experimental values using all available data (all) and the leave-one-out cross validation technique (loo)

	all				loo	
	rmsd	**r**			**rmsd**	**r**
MM/DDDC-OONS			**F(3,48)**	**P**		
1r	0.990	0.538	6.527	< 0.001	1.086	0.406
2r	0.886	0.656	12.108	< 0.001	0.967	0.575
4r	0.857	0.684	14.081	< 0.001	0.926	0.619
8r	0.969	0.673	13.253	< 0.001	0.943	0.600
MM/DDDC-OONS (+HB)			**F(4,47)**	**P**		
1r	0.986	0.543	4.923	0.002	1.109	0.375
2r	0.875	0.667	9.420	< 0.001	0.967	0.576
4r	0.836	0.703	11.471	< 0.001	0.918	0.629
8r	0.838	0.701	11.332	< 0.001	0.922	0.624
MM/DDDC-HP			**F(4,47)**	**P**		
1r	0.863	0.678	10.007	< 0.001	0.954	0.591
2r	0.803	0.730	13.416	< 0.001	0.890	0.658
4r	0.850	0.690	10.695	< 0.001	0.944	0.601
8r	0.840	0.699	11.243	< 0.001	0.933	0.614
MM/DDDC-HP (+HB)			**F(5,46)**	**P**		
1r	0.857	0.684	8.089	< 0.001	0.967	0.578
2r	0.783	0.745	11.500	< 0.001	0.888	0.662
4r	0.805	0.728	10.369	< 0.001	0.903	0.645
8r	0.791	0.739	11.099	< 0.001	0.889	0.659
MM/GBSA			**F(5,46)**	**P**		
	0.992	0.664	7.258	< 0.001	1.109	0.551
MM/GBSA (+HB)			**F(6,45)**	**P**		
	0.782	0.746	9.413	< 0.001	0.928	0.630

common entropic terms (which can be estimated to be in the range 20 to 40 kcal/mol) and the free energy of binding of the reference complex (which implies the addition of 11.3 kcal/mol), expected to be in the range of 30 to 50 kcal/mol, is slightly larger than expected.

MM/DDDC-HP

The OONS solvation term is accounting for both apolar and electrostatic solvation terms which should be already taken into account, at least partly, in the distance dependent dielectric constant. The same calculations described above have been performed using a similar approach in which the solvation term of the binding free energy is taken to be proportional to the polar/apolar accessible surface area of the molecule (see Eq. 6). The best scaling factors have been determined fitting the set of experimentally measured protein-DNA binding affinities (Table 1). The quality of the computed binding free energies ΔG_{calc} has been assessed evaluating the linear correlation coefficient r and the root mean square deviation (RMSD) between calculated and experimental values. In order to verify the performance of the model, a leave-one-out scheme has been adopted (Table 2). The F-statistic shows that the model is significant ($p < 0.001$).

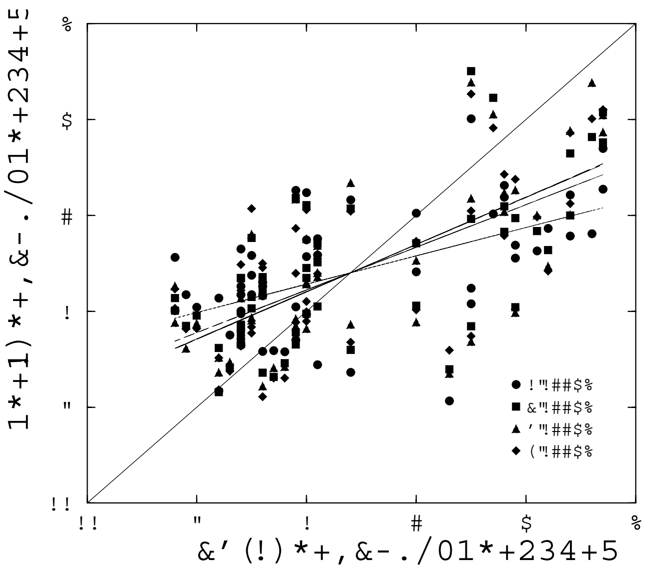

Figure 1
Computed binding free energies (MM/DDDC-OONS(+HB) model) versus experimental measurements, using a distance dependent dielectric constant ($\varepsilon = 1r$, $2r$, $4r$, $8r$). The correlation coefficients between calculated and experimental values are 0.543, 0.667, 0.703 and 0.701 for $\varepsilon = 1r$, $2r$, $4r$, $8r$, respectively.

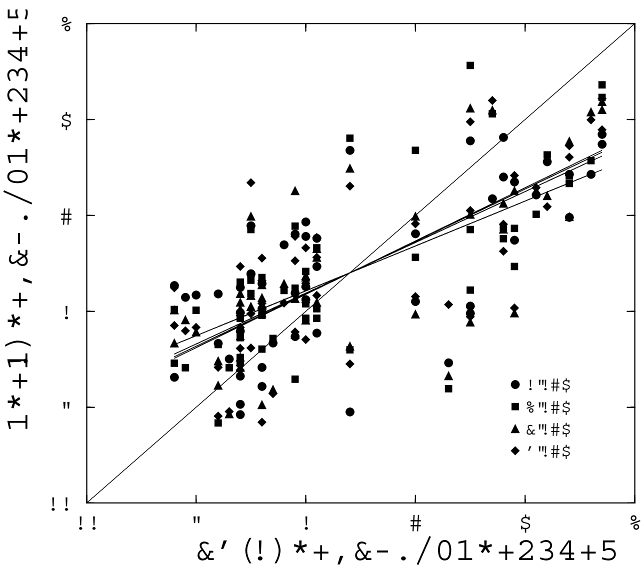

Figure 2
Computed binding free energies (MM/DDDC-HP(+HB) model) versus experimental measurements, using a distance dependent dielectric constant ($\varepsilon = 1r$, $2r$, $4r$, $8r$). The correlation coefficients between calculated and experimental values are 0.684, 0.745, 0.728 and 0.739 for $\varepsilon = 1r$, $2r$, $4r$, $8r$, respectively.

All the values of the distance dependent dielectric constant which have been tested gave a quite high and similar linear correlation coefficient. The highest correlation value (r = 0.745) was obtained for $\varepsilon = 2r$ for the MM/GBSA(+HB) model and the lowest ones for $\varepsilon = 1r$ similar to the MM/DDDC-OONS model. Generally, binding free energy changes lower than 1.0 kcal/mol are overestimated whereas binding free energy changes greater than 2 kcal/mol are underestimated in all the cases. More accurate predictions have been obtained for $\varepsilon = 2r$, in particular for values lower than 1.0 kcal/mol (Figure 2).

The optimal scaling coefficients are in the expected range (Table 1), in particular for $\varepsilon = 2r$, $4r$, $8r$ the constant term x_{const} is in the range 10–20 kcal/mol, moreover the coefficients x_H and x_P have the right order of magnitude of typically used surface tension coefficients for water biomolecular interface, even if the sign is incorrect. It should be noted however that there is a strong correlation (ranging in this case from 0.2 to 0.6) between the coefficients of most terms and the coefficient of the constant term.

Also for the present model the addition of an explicit hydrogen bond term reduces the coefficient of the electrostatic term as could be expected.

These results support the conclusion that, in general, there is no advantage in using the detailed solvation models compared to the simpler polar/apolar model, as far as the binding free energy is concerned. Based on the range of the scaling coefficients the two models appear of similar quality. Scaled free energy components for the MM/DDDC-HP(+HB) model are reported in Table 3.

MM/GBSA

In this approach all structures have been energy-minimized using the Generalized Born solvent model, then the binding free energy for every molecule has been calculated according to the MM/GBSA model using the Eq. 7. As in the previous cases, we determined the best scaling factors (and standard deviations according to Eq. 9) fitting the set of experimentally measured protein-DNA binding affinities (Table 4), then we assessed the quality of ΔG_{calc} predictions evaluating the linear correlation coefficient r and the root mean square deviation (RMSD) between calculated and experimental values. Finally we verified the performance of the model, using the leave-one-out scheme (Table 2). The same analysis has been performed for 5000 replicas of the dataset with one third of the set left out and used for cross-validation. The average RMSD and correlation are essentially the same reported for the leave-one-out scheme reported in Table 2. The standard deviations of the coefficients are essentially the same as reported in

Table 3: Components of the free energies (in kcal/mol) calculated according to the MM/DDDC-HP(+HB) model. The distance dependent dielectric constant is 1r. The constant term x_{const} is -13.355 kcal/mol

Original base	Mutated	ΔG_{exp}	ΔG_{calc}	vdW	Coul	H	P	HB
T3-A39	C-G	0.400	0.311	-50.460	-12.925	36.028	47.532	-6.508
	G-C	0.600	0.204	-50.326	-12.684	35.875	47.203	-6.508
	A-T	0.400	-0.091	-50.384	-12.669	35.965	46.861	-6.508
A4-T38	C-G	1.000	1.245	-49.982	-13.112	35.965	48.057	-6.327
	T-A	1.400	2.666	-49.455	-13.227	35.992	48.858	-6.147
	G-C	1.100	1.456	-49.704	-13.341	36.114	48.070	-6.327
C5-G37	T-A	0.800	1.681	-50.791	-13.042	36.270	49.108	-6.508
	A-T	0.600	0.404	-50.786	-12.985	36.360	47.860	-6.689
	G-C	0.700	0.659	-50.921	-12.952	36.417	47.978	-6.508
C6-G36	T-A	0.900	1.176	-50.637	-12.833	36.068	48.622	-6.689
	A-T	1.000	1.918	-50.517	-12.560	35.698	49.161	-6.508
	G-C	0.600	1.016	-50.216	-12.680	35.837	47.939	-6.508
T7-A35	G-C	0.400	0.018	-50.525	-13.005	35.845	47.387	-6.327
	A-T	-0.200	0.301	-50.387	-13.076	35.731	47.715	-6.327
	C-G	0.900	0.725	-50.472	-12.987	36.267	47.781	-6.508
C8-G34	T-A	0.500	1.878	-50.138	-13.246	36.020	48.924	-6.327
	A-T	3.600	2.417	-49.584	-12.945	35.415	49.213	-6.327
	G-C	3.400	1.965	-49.762	-13.259	36.086	48.583	-6.327
T9-A33	A-T	2.800	2.387	-49.570	-12.933	36.187	48.386	-6.327
	G-C	1.000	1.111	-49.390	-13.552	36.308	47.427	-6.327
	C-G	0.300	0.489	-50.729	-12.988	36.235	47.834	-6.508
G10-C32	A-T	0.400	0.974	-50.711	-12.531	36.073	48.187	-6.689
	T-A	0.500	1.377	-50.195	-13.338	36.243	48.530	-6.508
	C-G	0.200	0.654	-51.175	-12.775	36.602	48.227	-6.870
G11-C31	C-G	1.400	-0.061	-50.889	-13.098	36.203	47.768	-6.689
	A-T	1.100	0.760	-50.512	-13.375	35.925	48.766	-6.689
	T-A	2.300	0.451	-51.217	-12.921	36.481	48.332	-6.870
C12-G30	G-C	2.500	2.764	-48.826	-13.604	36.091	48.425	-5.966
	A-T	2.500	1.043	-50.496	-12.687	36.059	48.030	-6.508
	T-A	2.700	2.161	-49.468	-13.250	35.898	48.845	-6.508
G13-C29	A-T	2.900	2.337	-50.000	-12.972	36.226	48.766	-6.327
	C-G	3.400	2.413	-49.823	-12.729	36.460	48.188	-6.327
	T-A	3.200	2.546	-49.982	-12.803	36.574	48.621	-6.508
G14-G28	C-G	3.700	2.728	-49.462	-12.929	36.226	48.215	-5.966
	T-A	3.700	2.834	-49.692	-12.379	35.851	48.556	-6.147
	A-T	3.100	2.202	-50.081	-13.153	36.145	48.793	-6.147
T15-A27	G-C	2.900	1.728	-50.522	-12.607	36.887	47.834	-6.508
	A-T	2.000	1.089	-50.515	-12.812	36.617	47.663	-6.508
	C-G	2.500	0.966	-50.540	-12.811	36.544	47.637	-6.508
G16-C26	C-G	0.900	1.793	-49.675	-12.665	36.140	47.676	-6.327
	T-A	1.000	1.767	-50.112	-12.163	35.993	47.912	-6.508
	A-T	0.600	1.276	-50.287	-13.133	36.279	48.280	-6.508
A17-T25	C-G	0.200	1.168	-50.790	-12.598	36.796	47.624	-6.508
	T-A	-0.100	1.134	-50.710	-12.660	36.638	47.729	-6.508
	G-C	-0.200	1.256	-50.358	-12.788	36.760	47.506	-6.508
T18-A24	C-G	2.000	1.796	-49.628	-13.518	36.648	47.978	-6.327
	A-T	2.800	2.799	-49.290	-13.722	36.428	48.885	-6.147
	G-C	1.100	1.749	-49.774	-13.455	36.261	48.399	-6.327
A19-T23	T-A	0.400	1.239	-50.111	-12.649	36.212	47.650	-6.508
	C-G	0.500	1.378	-50.196	-12.587	36.216	47.808	-6.508
	G-C	0.400	0.780	-50.410	-12.892	36.243	47.703	-6.508
wild- type	--	0.000	1.158	-50.404	-13.063	36.234	48.254	-6.508

Table 4. Our calculation shows that the MM/GBSA(+HB) model gives the best performances (r = 0.746), although the linear correlation coefficient between calculated and experimental values differs slightly from the best values obtained from the other models. The F-statistic shows that the model is significant (p < 0.001). Computed values ver-

Table 4: Optimal scaling factors for the MM/GBSA model and the MM/GBSA(+HB) model: analysis of the λ repressor-operator system (λ), of the *Cro*-operator system (*Cro*) and the joint analysis (λ + *Cro*). Standard deviations (see Methods section) are given in parentheses.

MM/GBSA	λ	Cro	λ + Cro
x_{Coul}	0.157 (0.045)	0.132 (0.067)	0.153 (0.037)
x_{vdW}	0.252 (0.124)	0.332 (0.132)	0.313 (0.086)
x_{GB}	0.142 (0.046)	0.121 (0.069)	0.145 (0.038)
x_P	-0.010 (0.012)	0.019 (0.012)	-0.009 (0.006)
x_H	-0.029 (0.008)	-0.009 (0.011)	-0.031 (0.005)
x_{const}	-11.744 (28.025)	70.791 (37.615)	-19.519(11.903)
MM/GBSA(+HB)	λ	**Cro**	λ **+ Cro**
x_{Coul}	0.042 (0.053)	0.016 (0.066)	0.051 (0.041)
x_{vdW}	0.206 (0.113)	0.219 (0.118)	0.262 (0.079)
x_{GB}	0.025 (0.054)	0.002 (0.068)	0.043 (0.041)
x_P	-0.016 (0.011)	-0.001 (0.012)	-0.024 (0.006)
x_H	-0.023 (0.008)	-0.005 (0.010)	-0.029 (0.005)
x_{HB}	-0.533 (0.156)	-0.879 (0.224)	-0.561 (0.125)
x_{const}	13.699 (26.308)	65.011 (32.867)	-21.595 (10.908)

sus experimental data are reported in Figure 3. As far as the scaling coefficients are concerned (see Table 4), it is worth noting that addition of an explicit hydrogen bond term has a dramatic effect on the coefficients of van der Waals and electrostatic terms, as could be expected, because the latter terms already take into account hydrogen bond energetics. For the MM/GBSA model (with no explicit term for hydrogen bonds) the coefficients of the electrostatic and GB solvation terms are 0.16 and 0.14 which correspond to a dielectric constant of ~6. Surface tension coefficients x_P and x_H (-0.010 and -0.029 respectively) have the same order of magnitude of the commonly used surface tension coefficient (ca 0.02 kcal/mol Å^{-2}), but opposite sign. However the terms proportional to the solvent accessible surface area are strongly correlated to each other and to the constant term.

The constant term is -11.7 kcal/mol, lower than what expected, probably as a consequence of the correlation of this term with the polar and hydrophobic surface area terms (the linear correlations of the coefficients are 0.51 and 0.75, respectively). The standard deviation of this term is however very large (28.0 kcal/mol).

Scaled free energy components for the MM/GBSA(+HB) model are reported in Table 5.

The analysis of Table 5 shows that the most important feature for computing the binding free energy is the number of intermolecular hydrogen bonds. The correlation of the associated energy term with the experimental free energy of binding is 0.58. Other terms are strongly correlated among each other and therefore it is difficult to single out specific contributions. The correlation between different energetic terms range from -0.99, for GB solvation energy and Coulombic energy, to 0.44, for GB solvation energy and polar area burial energy term.

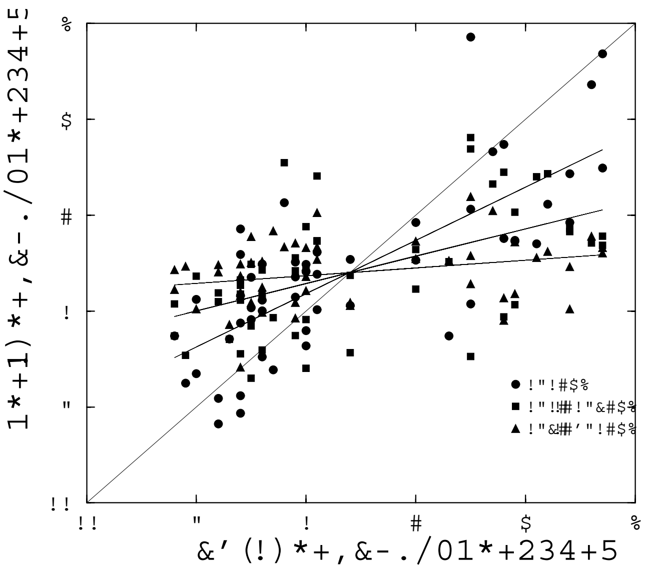

Figure 3
Binding free energy values calculated using the MM/GBSA(+HB) model versus experimental values Structures at 0.0 ns refer to the minimized complexes. The other two sets of data have been obtained by averaging over the MD simulation times 0.0 to 0.5 ns and 0.5 to 1.0 ns. The correlation coefficients between calculated and experimental values are 0.746, 0.534 and 0.284 for the minimized complexes, for the averages over time 0.0 to 0.5 ns and for the averages over time 0.5 to 1 ns, respectively.

Table 5: Components of the free energies (in kcal/rnol) calculated according to the MM/GBSA model. The constant term x_{const} is 13.699 kcal/mol

Original base	Mutated	ΔG_{exp}	ΔG_{calc}	Coul	vdW	GB	P	H	HB
T3-A39	C-G	0.400	0.115	-79.691	-33.547	50.408	30.607	37.296	-18.658
	G-C	0.600	0.525	-79.292	-33.467	50.226	30.690	37.327	-18.658
	A-T	0.400	1.165	-78.394	-33.432	49.648	30.588	37.181	-18.125
A4-T38	C-G	1.000	0.777	-81.129	-33.803	51.302	30.367	37.934	-17.592
	T-A	1.400	1.036	-81.246	-33.723	51.482	30.201	38.215	-17.592
	G-C	1.100	1.613	-79.949	-33.852	50.681	30.609	38.017	-17.592
C5-G37	T-A	0.800	2.136	-80.893	-33.864	51.109	30.552	39.125	-17.592
	A-T	0.600	0.996	-79.955	-33.842	50.645	30.484	38.090	-18.125
	G-C	0.700	0.377	-81.302	-33.810	51.382	30.589	37.944	-18.125
C6-G36	T-A	0.900.	1.362	-80.476	-33.730	50.877	30.661	38.456	-18.125
	A-T	1.000	0.622	-79.641	-33.667	50.411	30.314	38.163	-18.658
	G-C	0.600	1.110	-80.186	-33.579	50.678	30.491	38.132	-18.125
T7-A35	G-C	0.400	-0.082	-80.781	-33.792	51.084	30.579	37.254	-18.125
	A-T	-0.200	0.318	-80.382	-33.960	50.847	30.200	38.038	-18.125
	C-G	0.900	1.146	-80.387	-33.817	50.873	30.729	37.641	-17.592
C8-G34	T-A	0.500	1.350	-80.849	-33.862	51.113	30.593	38.247	-17.592
	A-T	3.600	3.360	-77.402	-33.305	49.165	30.110	38.152	-17.059
	G-C	3.400	2.435	-78.703	-33.834	49.920	30.582	37.829	-17.059
T9-A33	A-T	2.800	2.732	-78.808	-33.818	50.131	30.297	38.289	-17.059
	G-C	1.000	1.489	-79.210	-34.213	50.394	30.812	37.599	-17.592
	C-G	0.300	0.714	-80.866	-33.839	51.216	30.778	37.851	-18.125
G10-C32	A-T	0.400	1.868	-79.605	-33.008	50.347	30.450	38.110	-18.125
	T-A	0.500	1.034	-80.384	-34.230	50.938	30.722	38.414	-18.125
	C-G	0.200	-0.182	-81.436	-34.074	51.523	30.789	37.975	-18.658
G11-C31	C-G	1.400	1.533	-78.037	-33.806	49.387	30.408	38.007	-18.125
	A-T	1.100	1.009	-79.919	-33.894	50.529	30.555	38.163	-18.125
	T-A	2.300	0.731	-78.332	-33.617	49.511	30.299	38.362	-19.191
C12-G30	G-C	2.500	3.881	-78.666	-33.416	50.051	30.575	38.163	-16.526
	A-T	2.500	2.061	-78.455	-33.360	49.873	30.306	37.588	-17.592
	T-A	2.700	2.656	-78.889	-33.693	50.116	30.223	38.257	-17.059
G13-C29	A-T	2.900	1.744	-79.600	-33.700	50.381	30.624	37.933	-17.592
	C-G	3.400	1.941	-80.240	-33.646	50.825	30.805	38.090	-17.592
	T-A	3.200	2.131	-79.467	-33.720	50.343	30.737	38.665	-18.125
G14-G28	C-G	3.700	3.688	-78.059	-33.822	49.562	30.408	38.425	-16.526
	T-A	3.700	2.495	-78.334	-33.204	49.700	30.231	37.996	-17.592
	A-T	3.100	1.703	-80.426	-33.916	50.878	30.593	38.467	-17.592
T15-A27	G-C	2.900	1.755	-79.475	-33.371	50.375	30.884	37.766	-18.125
	A-T	2.000	1.946	-79.417	-33.359	50.311	30.820	37.484	-17.592
	C-G	2.500	1.088	-80.371	-33.422	50.848	30.775	37.683	-18.125
G16-C26	C-G	0.900	1.507	-79.347	-33.801	50.310	30.514	37.724	-17.592
	T-A	1.000	1.411	-79.452	-33.016	50.268	30.302	37.735	-18.125
	A-T	0.600	1.490	-79.733	-33.915	50.492	30.521	38.550	-18.125
A17-T25	C-G	0.200	0.100	-81.783	-33.375	51.717	30.952	37.014	-18.125
	T-A	-0.100	0.258	-81.595	-33.370	51.621	30.858	37.171	-18.125
	G-C	-0.200	0.741	-79.839	-33.332	50.528	30.608	37.202	-18.125
T18-A24	C-G	2.000	1.549	-81.158	-33.893	51.430	30.971	38.090	-17.592
	A-T	2.800	1.747	-81.528	-33.712	51.592	30.487	37.735	-16.526
	G-C	1.100	1.382	-80.583	-33.977	51.012	30.627	38.195	-17.592
A19-T23	T-A	0.400	1.602	-78.329	-33.474	49.625	30.680	37.526	-18.125
	C-G	0.500	0.917	-79.871	-33.472	50.549	30.684	37.453	-18.125
	G-C	0.400	0.878	-80.353	-33.573	50.784	30.669	37.777	-18.125
wild- type	--	0.000	1.120	-79.937	-33.811	50.610	30.593	38.090	-18.125

18 single base-pair mutants exhibit large (greater than 2.0 kcal/mol) unfavourable free energy of binding. Loss of hydrogen bonds contributes for 1 or 1.5 kcal/mol for mutants 14CG, 8AT, 8GC, 9AT, 18AT, 12TA, 12GC, while for other mutants the most important unfavourable contributions come mostly from Coulombic and van der

Waals terms. It should be noted, however, that solvation terms are correlated with Coulombic and van der Waals terms.

This analysis is in general in line with the detailed analysis reported by Oobatake et al. [54], although the exact values of energy contributions differ.

Analysis of molecular dynamics trajectories

The procedure used for computing binding energies may suffer from incomplete relaxation and incomplete conformational sampling. An approach that has been used in the past for sampling more conformations and reduce the effect of fluctuations is to analyse snapshots from molecular dynamics runs. In many studies no scaling factor was applied at all, with good results.

We performed 1 ns of MD simulations for every structure in order to test the effectiveness of a first principles computation of binding free energies and to check the effect of molecular dynamics relaxation on the computed energies. We calculated the average value of every component of the binding free energy using snapshots taken every 50 ps, then we used the same set of fitting equations using average values to determine the best scaling factors. We chose to use the MM/GBSA(+HB) model for computing binding free energies because it gave good results on the starting structures and the coefficients can be used to monitor the quality of the fitting. Figure 3 The quality of the computed binding free energies ΔG_{calc} has been assessed evaluating the linear correlation coefficient r and the root mean square deviation (RMSD) between calculated and experimental values (see Table 6). Results show that MD simulations do not improve the prediction capabilities of the model. Actually the linear correlation coefficient calculated averaging over 1.0 ns is 0.356, much lower than the correlation at t = 0.0 ns. Results obtained averaging over the time interval 0.0–0.5 ns, gave a linear correlation coefficient comparable to what obtained with other models on the starting structures (r = 0.534) but lower than the linear correlation coefficient obtained at t = 0.0 ns. The linear correlation coefficient between experimental values and the results obtained averaging over the time interval 0.5–1.0 ns, is r = 0.284, indicating that MD causes the loss of any correlation.

Table 6: RMSD and correlation coefficients (r) between experimental and calculated values obtained averaging over different time intervals, using the MM/GBSA(+ HB) model.

	MM/GBSA+(HB)	
	rmsd	r
0.0–0.5 ns	0.993	0.534
0.5–1.0 ns	1.126	0.284
0.0–1.0 ns	1.098	0.356

Moreover optimal scaling factors obtained averaging over the time interval 0.5–1.0 ns have the tendency to lose any physical meaning (Table 7). When optimal scaling factors obtained on the starting structure are used to compute binding free energies using average values, no correlation is detectable with experimental data. The value of the binding free energy change of every complex across 1 ns of simulation has been observed to strongly fluctuate, making it difficult to obtain an accurate estimate of it.

In order to verify whether this problem could have been circumvented using a larger conformational sampling, the simulations of 10 mutants have been extended to 4 ns, obtaining a total of 400 snapshots for every simulation. In particular we extended the simulations of the wild type complex and the best and the worst mutants (G17-C25 and T14-A28 respectively) with negative results. Although the system is most probably not fully equilibrated, it is reasonable to suspect that even longer (in the range of few tens ns) molecular dynamics simulations will not improve the results obtainable on the starting structures. The main reasons of failure of this approach are probably the large conformational fluctuations developing in MD simulations and the combination of relatively short molecular dynamics simulations with snapshots energy evaluation using the MM/GBSA(+HB) continuum model. Large conformational fluctuations observed in MD simulations are reflected in energetic fluctuations in the range of tens of kcal/mol, thus posing an issue on the reliability of the free energy average values. Moreover, since we observed that the results could not be improved extending the simulation time, it is reasonable to ascribe the failure of the method, at least partially, to inaccuracies in the force field parametrization. Actually, all force fields are based on numerous approximations, in particular nucleic acid force fields could suffer from two main problems which could give rise to inaccuracies. The first is that the target experimental data used in the optimization process

Table 7: Optimal scaling factors for the MM/GBSA(+ HB) model obtained averaging over different time intervals. Standard deviations (see Methods section) are given in parentheses.

MM/ GBSA+(HB)	0.0–0.5 ns	0.5–1.0 ns	0.0–1.0 ns
x_{Coul}	0.052 (0.028)	0.012 (0.024)	0.039 (0.029)
x_{vdW}	0.118 (0.049)	-0.039 (0.055)	0.060 (0.064)
x_{GB}	0.054 (0.029)	0.014 (0.025)	0.040 (0.031)
x_{H}	-0.001 (0.005)	0.007 (0.005)	0.004 (0.006)
x_{P}	-0.013 (0.004)	0.001 (0.004)	-0.007 (0.005)
x_{HP}	-0.039 (0.120)	0.078 (0.107)	0.027 (0.145)
x_{const}	-10.184 (6.002)	3.055 (4.486)	-2.953 (6.527)

are typically crystal structures of DNA and RNA. However, the presence of the lattice environment in crystals is known to influence the structure of DNA, limiting the transferability of crystal data to solution. The second is the treatment of electrostatics which is crucial in these simulations, given the polyanionic nature of DNA. In particular, the electrostatic polarization, which is an effect that can significantly reduce electrostatic interactions of partial atomic charges, is very important for accurate treatment of interactions in different environments, since significant structural changes of DNA may occur in response to environment.

Correct predictions

Table 8 shows the number of correct predictions, according to the criteria described in the *Methods* section. In the last column the number of cases in which the difference $D = |\Delta G_{exp} - \Delta G_{calc}|$ is lower than 0.3 kcal/mol, that is the number of the more accurate predictions, has been reported. It should be noted that the fitting of coefficients aims at minimizing the RMSD between calculated and experimental values and not at maximizing the number of "correct" predictions. When a simple simulated annealing procedure is applied to the coefficients the number of correct predictions can be increased by several units. It is instructive for instance to consider the MM/GBSA(+HB) model, where 41 "correct" predictions can be achieved with minor (mostly less than 10%) variations relative to the starting values of coefficients. From this qualitative point of view, the prediction capabilities of the different models can be compared. The best performing models appear to be the MM/DDDC-HP model with $\varepsilon = 2r$. On average the DDDC-HP model performs better than the similar DDDC-OONS model. For $\varepsilon = 1r$ results are worst than for higher ε values.

Molecular dynamics trajectories were analysed similarly, using average values for the different contributions to the free energy of binding. In particular the lowest number of correct predictions has been obtained averaging over the time interval 0.5–1.0 ns, actually there is no cases in which both ΔG_{exp} and ΔG_{calc} are <1.0 kcal/mol and the number of cases in which ΔG_{exp} and ΔG_{calc} are separated by less than 0.3 kcal/mol has been strongly reduced.

Table 8: Number of correct predictions of every model. Parenthetical data correspond to the number of prediction separated by less than 0.3 kcal/mol from the experimental data.

	$\Delta G_{calc} < 1.0$, $\Delta G_{exp} < 1.0$	$\Delta G_{calc} > 1.0$, $\Delta G_{exp} > 1.0$	ΔG_{calc} 1.0, ΔG_{exp} 1.0, $D < 0.5$	tot	($D < 0.5$)
MM/DDDC-OONS (1r)	10	23	2	**35**	(7)
MM/DDDC-OONS (2r)	15	24	0	**39**	(11)
MM/DDDC-OONS (4r)	16	22	1	**39**	(13)
MM/DDDC-OONS (8r)	14	23	1	**38**	(16)
MM/DDDC-OONS(+HB) (1r)	8	23	2	**33**	(9)
MM/DDDC-OONS(+HB) (2r)	14	23	1	**38**	(8)
MM/DDDC-OONS(+HB) (4r)	17	21	1	**39**	(10)
MM/DDDC-OONS(+HB) (8r)	15	22	2	**39**	(13)
MM/DDDC-HP (1r)	11	23	1	**35**	(12)
MM/DDDC-HP (2r)	16	21	4	**41**	(11)
MM/DDDC-HP (4r)	13	21	1	**35**	(12)
MM/DDDC-HP (8r)	13	24	0	**37**	(12)
MM/DDDC-HP(+HB) (1r)	12	23	1	**36**	(11)
MM/DDDC-HP(+HB) (2r)	15	22	3	**40**	(11)
MM/DDDC-HP(+HB) (4r)	13	21	2	**36**	(10)
MM/DDDC-HP(+HB) (8r)	16	24	1	**38**	(13)
MM/GBSA	13	24	1	**38**	(10)
MM/GBSA(+HB)	13	24	2	**39**	(13)

Generally, we observed that the number of cases in which ΔG_{exp} and ΔG_{calc} are both <1.0 kcal/mol decreases while the number of cases in which ΔG_{exp} and ΔG_{calc} are both > 1.0 kcal/mol remains nearly constant; however at the same time the number of cases in which ΔG_{exp} and ΔG_{calc} are separated by less than 0.3 kcal/mol strongly decreases, indicating that there is a reduction of the accuracy in reproducing experimental binding energies. Overall this analysis is consistent with the analyses reported in the previous sections.

Validation of the MM/GBSA model using the Cro-OR1 complex

The optimal scaling coefficients are likely to depend on the complex and mutants studied. In order to verify that such coefficients do not produce wild results when applied on different complexes with similar binding features we considered the *Cro* OR1 complexes which was obtained from crystallographic structure (PDB id. code: 6CRO) after mutation of 14 bases. The *Cro* protein belongs to the same family of λ repressor but to a different domain, according to SCOP classification [67] and it has very limited similarity with λ repressor although they bind DNA in a similar fashion. This system is therefore suited for testing the overall quality of the scaling procedure. Also for *Cro* a set of measurements for each mutant of the OR1 sequence is available [68]. When all contributions to the binding free energy, computed according to the MM/GBSA(+HB) model, are scaled by the coefficients determined on the λ repressor complexes the computed energies show a remarkable correlation coefficient of 0.62 with the experimental values, although the binding energies are overestimated by approximately 10 kcal/mol. This fact could reflect differences in the entropic contribution to binding (arising from restriction in side chain and backbone mobility) that are likely to be different for the two systems. Indeed, the crystallized *Cro* protein is roughly only two thirds of the repressor sequence. Notwithstanding the differences in overall binding energy, the binding differences for the mutants are on average reproduced by the energetic model.

As a further test, we performed the reverse analysis where the scaling coefficients are obtained on the *Cro*-OR1 complex and validated on the λ repressor-operator complex. Also in this case the computed energies show a remarkable correlation coefficient of 0.69 with the experimental values, although they are all underestimated by approximately 16 kcal/mol.

In order to verify how sensitive the scaling coefficients are to the experimental data used in the fit, we calculated the binding free energies of *Cro* and each mutant of the OR1 sequence according to the MM/GBSA model, using Eq. 7. Finally we combined the two experimental datasets of λ repressor and *Cro* and we refitted the model.

As in the previous cases, we calculated the best scaling factors fitting the set of experimentally measured protein-DNA binding affinities (Table 4), then we assessed the quality of ΔG_{calc} predictions evaluating the linear correlation coefficient r and the root mean square deviation between calculated and experimental values. Finally we verified the performance of the model, using the leave-one-out scheme. The best performance has been obtained for the MM/GBSA(+HB) model, which gives a correlation coefficient r of 0.69 and a rmsd of 0.74 for *Cro* and a correlation coefficient r of 0.67 and a rmsd of 0.83 for the two combined systems. The same analysis has been performed for 5000 replicas of the dataset with one third of the set left out and used for cross-validation. The average RMSD and correlation are essentially the same reported for the leave-one-out scheme. From the same analysis variances of the coefficients have been estimated with essentially the same results as those reported in Table 4. As far as the scaling coefficients are concerned (see Table 4), by comparing the results obtained for λ, *Cro* and the two combined systems, we can observe that the sets of values obtained for λ and λ + *Cro* are all in the same range except for the constant term, probably as a consequence of the fact that the entropic contribution to binding are likely different for the two systems. However it is worth noting that the standard deviation of this term is very large in both cases. As far as the scaling coefficients obtained for *Cro*, they are rather different from the others, except for x_{vdw} and x_{HB}, which scale the Van der Waals and H-bonds contributions respectively. However we observed that the electrostatic and GB solvation terms are strongly correlated to each other (the linear correlation coefficient is 0.998), as well as the constant term and the polar and hydrophobic surface area terms (the linear correlation of the coefficients is 0.645 and 0.784). The standard deviation of the constant term is also very large (see Table 4). Overall these results validate the approach for predicting binding free energies for similar protein-DNA complexes.

Analysis of non-specific protein-DNA binding

In order to study non-specific protein-DNA binding one thousand random DNA sequences have been generated and each sequence has been threaded onto the DNA phosphate backbone of the crystal structure in order to obtain a set of structural models with new DNA sequences. Minimization was performed according to the protocol described in the *Methods* section. We refer to to this set of complexes as to the "non-specific" set.

Binding free energies for each member of the generated non-specific set have been computed according to the MM/GBSA(+HB) model, using the optimal scaling factors determined by fitting the 52 experimental data (see Table 4).

We calculated the Z-scores of both the random structures and the single base-pair mutants, i.e. the distribution of the difference between the binding free energy of a complex and the average energy of the non-specific set, normalized by the standard deviation of the computed energies. Z-scores represent the specificity of a complex, with larger negative values corresponding to higher specificity. Figure 4 shows the distribution of computed energies. The distribution of the Z-scores of the single base-pair mutant complexes, is found at the negative tail of the non-specific distribution, indicating that these complexes are more stable than the complexes formed with a DNA random sequence, as one expects. The computed energies have an average difference of 4.8 kcal/mol and a standard deviation of 2.2 kcal/mol, giving thus an average z-score for the single base-pair mutants of 2.14 and 2.87 for the lowest computed energy in the set. The average non-specific binding energy seems surprisingly low (meaning that it implies that a rather large fraction of λ repressors present in the cell is actually non-specifically bound to DNA) but, remarkably enough, it agrees within the errors with the value proposed in [66] as a way to explain the impressive stability of the λ-switch.

It is interesting to compare the computed free energies of binding for the non-specific DNA complexes with those expected based on single mutants binding energies under the assumption of additivity. The expected free energies are higher than those computed by optimal scaling of contributions. The average difference, with respect to the specifically bound sequence, are 18.4 kcal/mol and 6.1 kcal/mol, respectively. This has been interpreted as a consequence of the fact that adjacent multiple substitutions may introduce additional energy minima compared to single mutations in a tight complex. This result is in line with the saturation effect in observed vs. predicted binding energy that has been described by Stormo and co-workers [69,70] and recently experimentally demonstrated [71]. It is also interesting to note that the non-specific binding energy is comparable to the energy computed by Northrup and co-workers for loosely docked complex of *Cro* to non-cognate DNA [72], which implies that the mode of binding may substantially change for non-specifically bound DNA sequences. This would be consistent with the capability of the protein of sliding along DNA, which would not be feasible for a tight complex.

Identification of putative transcription factor binding sites

The aim of this section is to understand whether the methods described here can be used for searching genomes for candidate transcription factor binding sites.

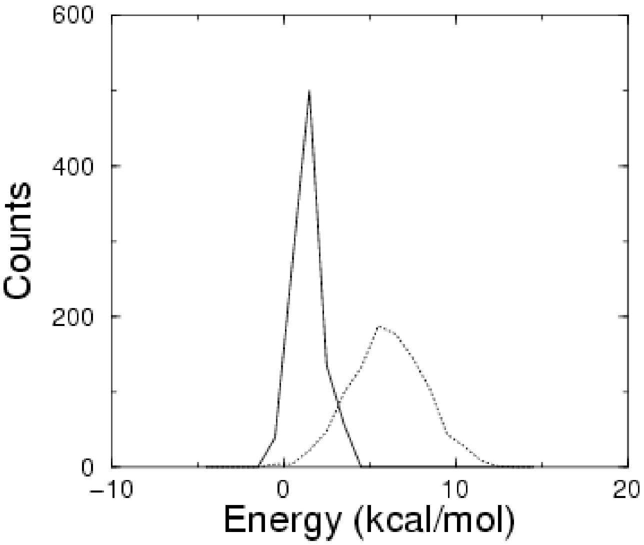

Figure 4
Distributions of the calculated binding free energy (bin width = 1 kcal/mol) for the "non-specific" set (dotted line) and for the single base-pair mutants (continuous line). The distributions are normalized to the same total number of counts.

In particular we aim at verifying:

i) whether the MM/GBSA(+HB) model is able to identify transcription factor binding sites in the absence of thermodynamic data about single base-pair mutants, but just knowing the recognized sequence;

ii) whether some predictions can still be afforded in the absence of thermodynamic data and of any information on recognized sequences. The latter situation could be encountered when a model of the complex is built by homology and differences in protein DNA-contacting residues imply a different specificity.

Identification of putative transcription factor binding sites knowing the bound sequence

The analysis in the previous section used knowledge about single base-pair mutants which is rarely available. Here we ask what predictions can be made when no thermodynamic data on mutants or wild-type sequence binding is available, but the cognate sequence is available. One thousand random DNA sequences were generated and the corresponding structural models were built by performing mutations on the double stranded DNA in the complex crystal structure using the program WHATIF [73]. Structures were energy minimized using the same protocol used for the MM/GBSA(+HB) methodology. Assuming that random sequences will have a larger free energy of binding compared to the bound sequence, optimal scaling parameters were sought in order to make the free energy difference in binding with respect to the naturally occurring complex equal to 10.0 kcal/mol. This value is arbitrary, albeit not unrealistic. Eq. 7 is solved (in a least square sense) subtracting the row corresponding to the wild-type complex from all other rows, and fixing all the energy differences equal 10.0. The differences in Coulombic, van der Waals, GB solvation energy, polar and apolar surface area and number of hydrogen bonds, with respect to wild-type complex, have been tabulated and the optimal scaling parameters have been determined.

The free energies computed on the random sequences have been used to compute single base-pair mutant free energies as described in the *Methods* section. The single base-pair mutant energies for the wild type sequence have been reset to 0.0 (this assumes that the specific bound sequence is known) and the lowest computed single base-pair mutant binding energy has been subtracted from all other values.

The plot of computed single base-pair mutant energies vs. experimental energies (computed under the hypothesis of additivity) shows a good correlation (0.58) but seems insufficient for predictive purposes. However, when the bacteriophage λ genome is scanned using the corresponding free energy matrix (see Methods), high-affinity binding sites are correctly recognized, and in general the energies computed using the matrix and those predicted based on addition of single base-pair mutation effects are well correlated (corr. coeff. 0.74).

We asked what is the advantage of such computation compared to the simpler model that assigns a constant energy penalty to each mutation over the specific bound sequence. In such case the correlation between the computed and reference binding energies is slightly lower, but still significant (0.72). The advantage of using computational results over a much simpler single parameter approach seems therefore very limited, although the 1% best sites predicted by the MM/GBSA(+HB) energy and the simple mutation models display only 15% common sites, proving that the two methods are largely uncorrelated.

Identification of putative transcription factor binding sites without knowing the bound sequence

As a last test we simulate a realistic situation in which no thermodynamic data or information on the recognized sequences are available. We considered the set of one thousand random DNA sequences and the corresponding structural models built by performing mutations on the double stranded DNA in the complex crystal structure as the only information available. Obviously the crystallographic complex does contain information on the specific sequence because protein and DNA conformations are fitting each other in the complex. If non-specific complexes were to be built by homology without knowing the exact DNA sequence bound, it is likely that side chains would be placed differently with different results. Finally, structures were energy minimized using the same protocol used for the MM/GBSA(+HB) methodology. As in the tests above we found optimal scaling factors in order to make all (non-specific) binding free energies equal to 10.0 kcal/mol.

In order to avoid a trivial solution to the fitting problem with all coefficients equal 0.0 except the constant term equal 10.0, we follow a two-step procedure. In the first step we assume a reasonable value (30 kcal/mol) for the constant term which must be brought to the left-hand side of Eq. 7. Coulombic, van der Waals, GB solvation energy, polar and apolar surface area and number of hydrogen bonds have been evaluated, Eq. 7 is then solved (in a least square error sense) and the optimal scaling parameters have been determined. The lowest binding energy sequence according to the scaling parameters is determined. The row corresponding to this complex is subtracted from all other rows thus removing the constant term. In the second step the newly obtained matrix, which does not include the constant term anymore is used to find the best coefficients to make all the energy differences equal 10.0 kcal/mol. Therefore all energies are expressed relative to the lowest computed energy at the first step.

The free energies computed on the random sequences have been used to compute single base-pair mutant free energies as described in the *Methods* section. At variance with the test performed above we do not set to 0.0 the energies of specific bound sequence (which is assumed here to be unknown). The correlation coefficient between computed and experimental energies (computed under the assumption of additivity) for the bacteriophage λ genome is 0.50 (Figure 5).

As a further test of the performance of the approach we generated the logo [74] of the 10 best binding sequences according to the thermodynamic data on single base-pair mutants and those found with the present approach (Figure 6). An overall agreement between the two logos is apparent.

Conclusion

In the present work physical effective energy functions are used to estimate the free energy of binding of λ repressor to the DNA operator and single base-pair mutants, for which thermodynamic data are available. Thermodynamic data allow one to study the best results achievable, with the modeling approach and energy functions pre-

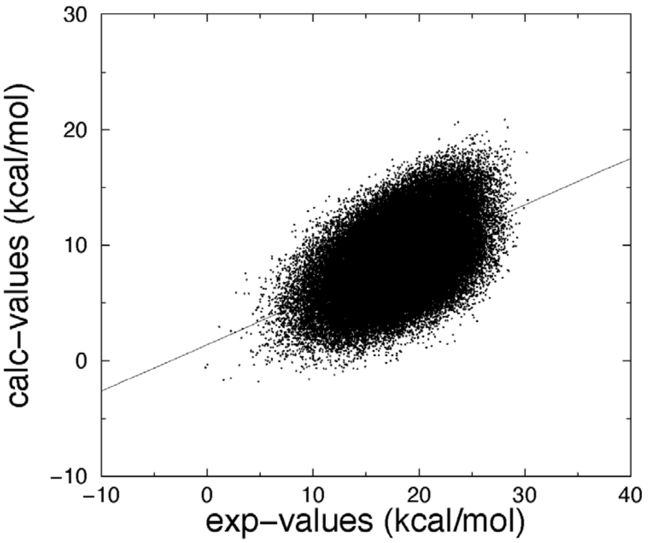

Figure 5
Binding free energies predictions without using specific bound sequence nor thermodynamic information versus binding free energy values obtained under the hypothesis of additivity [69,70] using experimental data on single base-pair mutants. The correlation coefficient is 0.50.

sented here, with models that assume that the binding energy is a linear combination of different contributions.

Simple models that use a distance dependent dielectric constant and simple terms for surface area proportional energy contributions and for hydrogen bonding perform surprisingly well for values of ε ranging from $2r$ to $8r$.

A two-parameter model for surface area proportional energy contributions performs better than the more complex model of Oobatake et al. [55], which was however not derived for usage in the more complex energy functions employed here.

The performance of MM/GBSA(+HB) and to a lesser extent MM/GBSA model is comparable to or superior to other models. A conclusion for the MM/GBSA model is that electrostatic energies should be reduced by a proper scaling factor corresponding to dielectric constants in the range of 6. This conclusion is reached also by a similar analysis of protein *Cro*-operator mutants.

The effect of molecular dynamics on the computed binding free energies is in general negative and the reproducibility of the experimental values decreases with the increase of simulation time considered. This may be a consequence of the large fluctuations developing in MD simulations which probably would require a much longer simulation time. Moreover it is reasonable to take into account that the poor performance of the method can be partially caused by the errors in the force field used in MD simulations. Another plausible source of inaccuracy is the mismatch between the energy model and system representation used in MD simulation and those used for minimization and energy evaluation. It appears therefore that it is worth to invest more time in optimizing the starting structure, rather than for sampling the conformational space by molecular dynamics simulations, or, alternatively, to adopt different strategies for sampling protein and DNA flexibility [75].

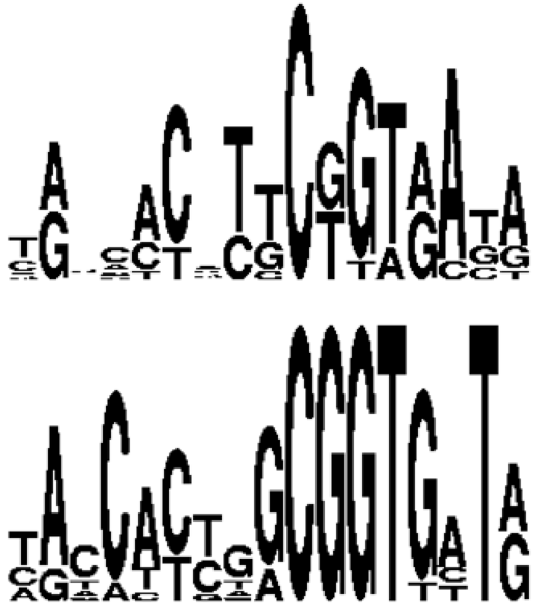

Figure 6
Logos obtained from the ten best binding sequences according to the experimental data of Sarai et al. (ref. [61]) (lower panel) and according to the computations on "non-specific" complexes complexes with no sequence or thermodynamic data information (upper panel).

The analysis of non-specific complexes using the best performing energetic model with properly scaled coefficients allows to evaluate a non-specific binding energy difference, with respect to the specific bound sequence, of 6.06 ± 2.17 kcal/mol, definitely lower than what expected based on an additive model (18.1 kcal/mol for the single base-pair mutants computed energies). This result is in line with the saturation effect described by Stormo and coworkers [69,70] and with the theoretical analysis of Bakk and Melzer [66].

Although the results presented on single base-pair mutants are not exciting, using computational methods may be very useful for identifying transcription factor binding sites.

When no thermodynamic data are available but the specific bound sequence is known the computed MM/GBSA(+HB) free energies are slightly more predictive than a simple substitution profile which assigns a penalty for any point mutation.

The most interesting test performed here considers a realistic scenario where no information on the bound sequence is available. Even in this case MM/GBSA(+HB) energies are predictive.

This result has important consequences for the prediction of transcription factor binding sites which often use consensus methods. A prerequisite for the usefulness of consensus methods is that these are as independent of each other as possible. Since most methods use common prior knowledge and often related statistical methods, independence is not guaranteed. Methods which are based on completely independent principles, like those based on physical effective energy functions and free energy computations, offer a completely complementary methodology for deriving profile matrices for scanning entire genomes. The results reported here, with much caution because the structural model for the specific bound sequence is known and not modeled by homology or other methods, support usage of these methods for the identification of DNA-binding protein target sequences. In view of the very recent impressive results reported by the group of Baker [76] it is apparent that significant improvements to the approach described in this paper may be obtained by extensive refinement and screening of protein side chain conformation at protein-DNA interface.

Methods

Model building

Atomic coordinates of the λ repressor dimer bound to O_L1 DNA operator were taken from the 1.8 Å resolution X-ray crystal structure deposited in the Protein Data Bank [77] (PDB code 1LMB). The operator is 17 base-pairs in length and is composed by two approximately symmetric parts, the "consensus half" (maintaining the notation of the PDB file, base-pairs A19-T23 to G11-C31) and the "non-consensus half" (base-pairs T3-A39 to G10-C32) (see Figure 7). Since the coordinates of the NH_2-terminal arm of the repressor bound to the non-consensus half operator were not available, the lacking amminoacids were added using the protein bound to the consensus half operator. Using the program ProFit V2.2 [78], the C_α carbons of the proteins have been superimposed and afterward the amino acids of the rotated structure have been added to the other one. Since the detailed X-ray crystal structure is made up of λ repressor dimer and O_L1 operator DNA while the experimental data concern the O_R1 site, the WHATIF [73] program was used to substitute the base-pair at position 5 to obtain the wild-type O_R1 operator. All possible single base-pair substitutions within the DNA sequence were generated using the program WHATIF [73].

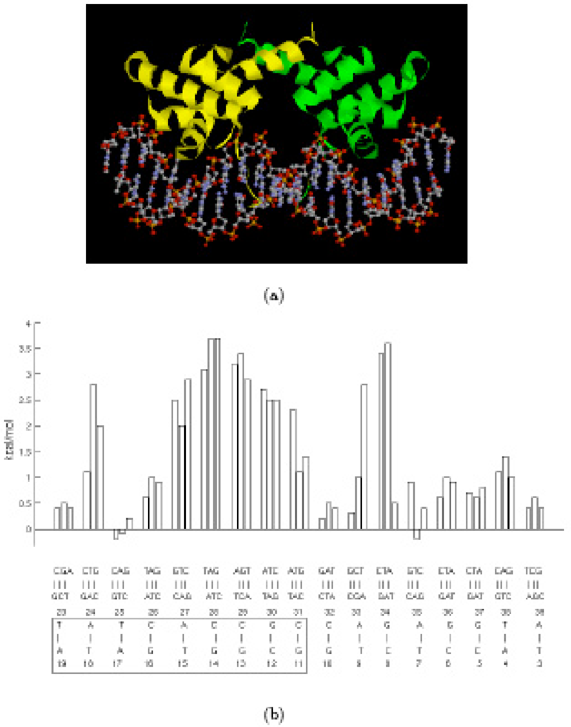

Figure 7

(a) Structure of the complex of λ repressor [58] with operator DNA. The protein was crystallized with a 19-bp duplex of which the central 17 bps are shown. The consensus half is to the left, (b) Relative free energy changes in the binding of λ repressor to O_R1 on base substitutions. The figure shows the change in affinity that results from each of the three possible substitution at all 17 sites. The left part represents the consensus half-operator (solid box) and the right half the non-consensus half-operator (redrawn from ref. [61]).

Molecular Dynamics Simulations

Hydrogen atoms have been added using the program pdb2gmx of the GROMACS package [79,80]. Every structure has been optimized performing 200 steps of energy minimization using the NAMD program, fixing all C_α carbons and phosphate groups coordinates. A dielectric constant of 10 has been employed with a cut-off of 12 Å for non-bonded interactions.

The net charge of the system (-36) has been neutralized placing a corresponding number of sodium counterions in energetically favourable positions. The electrostatic potential was calculated via numerical solutions of the Poisson-Boltzmann equation using the University of Houston Brownian Dynamics (UHBD, version 6.x) program [81,82]. A counterion was placed at the lowest potential position at 7.0 Å from any heavy atom of the solute. The cycle was repeated until the net charge of the system was 0.

The complex and counterions were solvated in a box of TIP3P [83] water molecules using the solvate module in the program VMD [84]. The resulting system contained about 4200 solute atoms and 50400 solvent atoms. The coordinates of the solute were fixed and the solvent was energy minimized using 100 steps of conjugate gradient. A solvent equilibration was carried out by performing molecular dynamics for 50 ps using a 1 fs time step to let the water molecules move to adjust to the conformation of the solute. The system was then energy minimized using 100 steps of conjugate gradient and, after 100 ps equilibration, 1-ns MD simulations was performed using a 2-fs timestep. A snapshot of the trajectory was stored every 10 ps for later analysis. The shakeH algorithm was used in order to fix bond length between each hydrogen and its mother atom to its nominal value and to extend the simulation time-step [85]. All molecular dynamics simulations of the complex were run under constant NPT conditions using the NAMD program [86]. The pressure of the system was coupled, through a Berendsen-thermostat [87], to a pressure bath with target pressure 1.01325 bar and time constant 100 fs. The temperature has been kept to 300 K by simple velocity rescaling every picosecond. Long-range electrostatic interactions were treated by particle mesh Ewald (PME) method [88] employing a grid of 128 × 128 × 128 points. The cut-off was 12 Å and the tolerance was 10^{-6} which resulted in an Ewald coefficient of 0.257952. The order for PME interpolation was 4.

The simulations were performed on a cluster composed by ten dual-processor nodes based on Intel XeonTM 2.8 GHz, with hyper-threading technology.

Free energy calculations

The free energy of binding for each structure has been computed according to the framework reviewed by Gilson et al. [89] who derived the expression of the free energy of binding in terms of the microscopic properties of the two molecules involved, using standard statistical thermodynamics. Here, similar to other works employing continuum methods several simplifications are adopted. The free energy of binding for each complex minus the entropic contribution is expressed as the sum of the interaction energy between the protein and the DNA $\Delta U(\vec{r}_1,\ldots,\vec{r}_n)$ and a solvation free energy term ΔG_{solv}:

$$\Delta G = \Delta U(\vec{r}_1,\ldots,\vec{r}_n) + \Delta G_{solv}(\vec{r}_1,\ldots,\vec{r}_n) \tag{1}$$

It has been assumed that the entropy restriction in internal degrees of freedom and overall rotation and translation degrees of freedom is the same for all complexes.

The effect that association has on intramolecular energy has been neglected. Moreover no extended conformational search has been performed for protein side chains and DNA, partly because this task is not easily accomplished and partly because large conformational changes often result in large molecular mechanics energy changes, so we aimed at keeping the systems to be compared as close as possible. The free energy of binding has been calculated using different methodologies detailed below. For all models alternative versions in which an energy term proportional to the number of hydrogen bonds has been added have been considered.

Except where noted, all contributions to the free energy of binding have been optimally scaled in order to best reproduce available experimental data (see later).

MM/DDDC-OONS

In this method [54] electrostatic interactions have been estimated using a distance dependent dielectric constant (DDDC) while the solvation energy is proportional to the solvent accessible surface area through the atomic solvation parameters of Oobatake, Ooi, Nemethy and Scheraga (OONS) [55].

All structures have been energy minimized with 200 conjugate gradient steps, using a distance-dependent dielectric constant (four values have been tested: 1*r*, 2*r*, 4*r*, 8*r*, with the distance *r* expressed in Å) and a cut-off of 12 Å. The molecular mechanics interaction energy $U(\vec{r}_1,\ldots,\vec{r}_n)$ was evaluated using CHARMM (version 27b2), a classic and well-tested molecular mechanics force-field [90,91]. This term includes the nonbonded electrostatic and Van der Waals contributions. The solvation free energy term G_{solv} has been calculated according to the model developed by Oobatake et al. [54]. This model consists in assigning every atom to one of 9 classes of chemical groups and assuming that the hydration free energy of every group i in a solute is proportional to its solvent accessible surface area (SASA) A_i because the group can directly interact only with water molecules at the surface.

$$G_{solv} = \sum_i g_i^{hyd}(A_i(complex) - A_i(protein) - A_i(DNA)) \tag{2}$$

The proportionality constants g_i^{hyd} have been determined from thermodynamic data on the transfer of small molecules from the gas phase into aqueous environment assuming the additivity of contributions from individual groups [55].

MM/DDDC-HP

In a very similar approach the OONS 9-parameter solvation model has been replaced by a simpler 2-parameter hydrophobic, polar (HP) solvation model. Energy minimization protocol and tested values are the same as for the MM/DDDC-OONS for proper comparison.

MM/GBSA

In this method the solvation free energy term is split in a polar (electrostatic) and a non-polar (hydrophobic) term.

$$\Delta G_{solv}(\vec{r}) = \Delta G^{polar}(\vec{r}) + \Delta G^{non-polar}(\vec{r}) \qquad (3)$$

The polar term is computed using the Generalized Born approach [92]. All complexes have been energy minimized by 200 conjugate gradients minimization steps using the generalized Born model as implemented in the CHARMM program, then the solute and solvation energy terms have been computed for both the complex and the isolated molecules. The binding energy was then computed by subtraction. Doubling the number of minimization steps does not affect significantly the results.

The non-polar term $G^{non\text{-}polar}$, which takes into account the tendency of the non-polar parts of the molecule to collapse, is taken to be proportional to the solvent-accessible surface area A, i.e. $G^{non\text{-}polar} = \gamma A$, where the surface tension coefficient γ has been empirically determined to be equal to 20 cal Å $^{-2}$ mol^{-1} for this kind of applications [93].

A variant of this methodology including splitting the solvent accessible surface area into a polar and a hydrophobic contribution (i.e. using two different surface tension coefficients), and including a term proportional to the number of hydrogen bonds has been considered here.

Finding optimal scaling factors

The choice of methods and parameters in molecular mechanics/implicit solvent methods is subject to large uncertainties. In order to explore the best performance achievable with these methodologies, optimal scaling factors for the different contributions were searched that could best reproduce the experimental data. This approach is not new and it has been used successfully by other groups (see e. g. [57]). In practice it is expected that proper scaling is able to compensate for the many inaccuracies of the model. In general terms, the free energy of binding has been computed as a linear combination of contributions E_i, with corresponding coefficients x_i, i.e.:

$$\Delta G_{binding} = \sum_i x_i \Delta E_i \qquad (4)$$

where Δ represents the difference between the complex and the isolated protein and DNA molecules. Coefficients $x_1,\ldots,x_n$ have been found in order to best reproduce the 52 experimentally available free energies of binding. Contributions have been arranged in a 52 × n matrix A where each row corresponds to each structural model and each column corresponds to a different contribution to the free energy of binding. The experimental binding free energies have been arranged in a 52-component vector ΔG_{exptl}. The linear system

$$Ax = \Delta G_{exptl}$$

where x is the n-component vector of coefficients, has been solved (in a least square sense) using singular value decomposition [94] and the best x_i coefficients have been used to calculate binding energies ΔG_{calc}. A constant term takes into account the entropy loss upon complexation and other possible contributions identical for all complexes.

A linear model, compared to more sophisticated methods, has the advantage that the number of adjustable parameters is limited and easily interpretable in physical terms.

In the following we detail the contributions considered for each energetic model.

The free energy of binding has been computed for the MM/DDDC-OONS model according to the following equation:

$$\Delta G_{MM/DDDC\text{-}OONS} = x_{vdW}\Delta E_{vdW} + x_{DDDC}\Delta E_{elec,DDDC} + x_{OONS}\Delta G_{OONS} + x_{const.}(+x_{HB}N_{HB}) \qquad (5)$$

where ΔE_{vdW} is the van der Waal contribution, $\Delta E_{elec,DDDC}$ is the Coulombic energy, computed with a distance dependent dielectric constant, ΔE_{OONS} is the solvation energy according to the Oobatake et al. model [55] and N_{HB} is the number of intermolecular hydrogen bonds.

As mentioned above, the coefficients bear physical meaning. For instance the term x_{const} should account for rotational and translational entropy loss upon binding and it can be expected to be in the range 20–40 kcal/mol.

The term proportional to the number of hydrogen bonds was alternatively added in order to take into account possible inaccuracies in the treatment of these interactions by molecular mechanics and solvation terms. In practice every time this term is added the coefficients of molecular mechanics and solvation terms are greatly reduced thus avoiding double counting of hydrogen bond interactions.

A similar expression for the free energy of binding has been used for the MM/DDDC-HP model:

$$\Delta G_{MM/DDDC\text{-}HP} = x_{vdW}\Delta E_{vdW} + x_{DDDC}\Delta E_{elec,DDDC} + x_H\Delta A_H + x_P\Delta A_P + x_{const.}(+x_{HB}N_{HB}) \qquad (6)$$

Here the coefficients x_H and x_P represent the surface tension coefficients multiplying hydrophobic and polar solvent accessible surface areas ΔA_H and ΔA_P, respectively. We expect these coefficients to be in the range of tens of cal Å^{-2} mol^{-1}.

The solvent accessible area has been also splitted in polar and hydrophobic area for finding optimal scaling parameters for the MM/GBSA methodology:

$$\Delta G_{MM/GBSA} = x_{vdW}\Delta E_{vdW} + x_{Coul}\Delta E_{Coul} + x_{GB}\Delta G_{GB} + x_H\Delta A_H + x_P\Delta A_P + x_{const.}(+x_{HB}N_{HB}) \qquad (7)$$

where ΔG_{GB} is the generalized Born solvation energy. The coefficients x_{Coul} and x_{GB} are exactly and roughly, respectively, inversely proportional to the effective dielectric constant and are thus expected to be in the range 0.05 to 1.0.

Possible pitfalls of the method

Scaling energy terms for free energy evaluation of models which have been minimized without scaling such terms is clearly inconsistent. A correct procedure would be to iteratively find the optimal scaling factors, minimizing the energy using such scaling factors and repeating these two steps until convergence. This procedure faces some difficulties because an important term like the hydrogen bond term is discrete and does not have a counterpart in standard forcefields, where such interactions are described typically through electrostatic and van der Waals terms. Similarly the minimization of terms proportional to the solvent accessible surface area requires algorithms which are rarely available in molecular mechanics packages. A further difficulty is that any unbalance among forcefield terms might introduce distortions in molecular structure, notably of hydrogen bond lengths. Although the issue of iteratively fitting optimal scaling factors is worth being further investigated, here the approach of scaling factors has been applied in a more rough way. We have matched as far as possible the energetic model used for minimization with that used for fitting scaling factors, as mentioned above, but we have not minimized again the models using the scaling factors. A similar mismatch between conformational sampling and energy evaluation is implicit in the analysis of molecular dynamics snapshots. Other sources of error in this case are the large conformational (and energetic) fluctuations molecules undergo during simulation and in general the inaccuracy of implicit solvent methods (used in energy evaluation) where small energy differences arise from subtraction of rather large values. It should be noted that for molecular dynamics snapshots inaccuracies do not cancel out because there are no restrained parts in the molecules.

DNA sequence dependent deformability

An important aspect of protein-DNA interaction, addressed quantitatively by Olson and co-workers [39], is the capability of DNA sequences to adopt specific local conformations. The statistics of parameters and pairwise parameter correlations shows definite preferences. In the approach described above, changes in intramolecular energy terms are disregarded altogether by the assumption of rigid docking. The strains introduced in complex molecular structures, however, are typically relaxed over the structure and should have consequences on the intermolecular energy terms too. In order to assess the effect of DNA sequence dependent deformability we followed the approach of Olson and co-workers [39], who made available average parameters for the six parameters describing local geometry of a base-pair step in B-DNA, the force constant parameters for all pairwise deviation from equilibrium values and a program to analyse DNA structures [95].

The analysis was performed for the native structure parameters, simply replacing the identity of the base-pair mutated, and on the mutated structures, minimized using the generalized Born model. For both cases poor correlation with experimental binding data was found. Remarkably, however, the native sequence was the third lowest energy sequence among all 52 sequences. Energy minimization in general increases the energy associated with the deformability of DNA. Computation of the fitness of a sequence to local geometry parameters gives important informations although it is likely that the computed energy is not accurate for conformations far from equilibrium. Inclusion of the DNA sequence dependent deformability energy in the analyses detailed below did not improve results significantly, notwithstanding the additional scaling parameter introduced for this purpose. For this reason this term was not considered further.

Performance analysis

After fitting scaling factors to experimental data, the root mean square difference between calculated and experimental data was computed. This quantity can provide however a poor evaluation of the predictive power of the calculations when the test systems are very similar. Therefore the correlation coefficient between calculated and experimental data was also computed. Optimal scaling factors were computed taking all the data available.

Fitting 52 experimental data with up to 7 parameters will always results in a positive correlation coefficient. In order to make sure that the results obtained are significant we performed different kind of analyses:

i) a leave-one-out scheme has been adopted. All but one of the data were taken and the root mean square difference and correlation coefficient were computed using the set of data not used in the fitting procedure. The same scheme has been applied to 5000 replicates with one third of the data left out of the fitting procedure and used for RMSD and correlation coefficient computation.

ii) the variance of each linear coefficient has been estimated from the multiple regression analysis using the variance/covariance matrix and the square error of computed data, according to standard linear regression procedures [96]. In practice the standard deviation of experimental data has been estimated as

$$\sigma = \sqrt{\frac{\sum_i (G_{i,calc} - G_{i,exptl})^2}{n-m}} \qquad (8)$$

Then the variance of each coefficient has been estimated from the variance/covariance matrix of coefficients:

$$\sigma^2(x_j) = (A^T A)^{-1}_{jj}\sigma^2 \qquad (9)$$

The different models considered employ a different number of fitting parameters and therefore different performances are expected. Although these parameters are often correlated, the analysis of the variance gives an immediate clue as to which variables are more important.

iii) analysis of variance (ANOVA) calculations have been performed and a significance test based on the F-statistic and the corresponding confidence level has been computed [96].

iv) one thousand replicates of the original data has been generated with the column elements containing the experimental data randomly swapped. The average of the correlation coefficient between swapped experimental data and fitted data has been computed together with the standard deviation. The results of this computation (not reported) fully supports the results of the statistical analyses described above; Finally, a useful alternative to assess the quality of ΔG_{calc} predictions and to compare the different models from a qualitative point of view, consists in determining the number of "correct predictions", defined as the number of cases in which both ΔG_{exp} and ΔG_{calc} are <1.0 kcal/mol, or >1.0 kcal/mol, or else separated by less than 0.3 kcal/mol. The threshold value of 1.0 kcal/mol requires some explanations. The experimental values of the free energy change relative to the wild-type operator O_R1 have been calculated using the equation ΔG_{exp} = -0.546 ln (K_d of substituted sequence)/(K_d of O_R1) after having determined the dissociation constant of every mutant. It is simple to verify that the threshold value of 1.0 kcal/mol corresponds to a remarkable reduction in the dissociation constant of the mutant (ca. 5-fold), with respect to the dissociation constant of the wild-type operator (K_d of O_R1 = 10^{-9}), whereas values of ΔG higher than 1.0 kcal/mol correspond to a reduction in the dissociation constant from 5 (ΔG_{exp} = 1.0 kcal/mol) to 25-fold (ΔG_{exp} = 3.4, which is the maximum value of ΔG_{exp}). Therefore it is reasonable to define ΔG_{calc} as correct, if both ΔG_{exp} and ΔG_{calc} are in one of the defined intervals or even if the difference $D = |\Delta G_{exp} - \Delta G_{calc}|$ is lower than 0.3 kcal/mol, which corresponds to a ratio between the dissociation constant of a mutant and the dissociation constant of the wild-type complex lower than 2.0.

Analysis of non-specific protein-DNA binding

One thousand random DNA sequences were generated and the corresponding structural models were generated by performing mutations on the double stranded DNA in the complex crystal structure using the program WHATIF [73]. The resulting dataset of complexes was assumed to be representative of non-specific protein-DNA complexes. We are interested in understanding how reliable is the method for predicting putative binding sites. The so-called Z-score of the specific bound sequence compared to random sequences has been considered. The Z-score is defined here as the distance of the free energy computed for the specific bound 17-mer (ΔG) from the average non-specific binding energy (< ΔG >), normalized by the standard deviation of the computed non-specific binding energies (σ_G).

$$Z-score = \frac{\Delta G - <\Delta G>}{\sigma_G} \qquad (10)$$

Averages are performed over the one thousand random sequences. A large Z-score implies that the specific bound sequence can be distinguished from other non-specific bound sequences. The structures were energy minimized using the same protocol used for MM/GBSA free energy estimation. For all minimized complexes the Coulombic energy, van der Waals energy, GB solvation energy, polar and apolar surface accessible area and intermolecular hydrogen bonds number were tabulated. For each model i of the 1000 random DNA sequence complexes the binding energy $G(i)$ has been computed using different amounts of the experimental information available. Different analyses, detailed in the Results section, were performed. The possibility of using the data computed on the set of non-specific complexes for defining a profile of the recognized DNA sequences has been explored as follows. The calculated binding energy values for the set of non-specific complexes were summarised in a set of 68 values corresponding to the average contribution to the binding free energy of each possible of the 4 bases at each of the possible 17 bound sequence positions. These 68 values have been derived as follows. Possible substitutions are indexed from 1 to 4 for A, C, G and T, respectively. A 1000 × 68 matrix A was set where each element $A(i, j)$ is 1.0 or 0.0 if the base at position $j/4$ (rounded at the closer upper integer) has index j mod 4 in sequence i. The set of 68 substitution free energies $x(j)$ were found by solving (in a root mean square error sense) the overdetermined equation $Ax = G$. The resulting 68-element vector x was arranged in a 17 × 4 matrix. Variants on this procedure are described in the Results section according to the level of information available included in the analysis.

Scanning of bacteriophage λ genome

The free energy matrix derived from the analysis of non-specific protein-DNA complexes was used to score all 17-mer subsequences in the bacteriophage λ genome (Accession number: NC_001416.1, 48502 base-pairs) on both strands. In principle the score represents the free energy of binding of the 17-mer considered.

Reference "experimental" binding free energy values, for comparison with computed data, were obtained under the hypothesis of additivity [69,70] using experimental data on single base-pair mutants.

Competing interests

The author(s) declares that there are no competing interests.

Authors' contributions

EM designed and performed most tests and analyses, and wrote part of the code used for the analyses. MC and FF conceived the project, supervised the work and designed some of the analyses. FF wrote part of the code used for the analyses. All authors have read and approved the final version of the manuscript.

Acknowledgements

We wish to thank Drs. G. Tecchiolli and P. Zuccato of Exadron, the HPC Division of the Eurotech Group, for providing hardware and expert technical assistance.

Dr. M. Isola of the University of Udine is gratefully acknowledged for helpful discussions on statistical aspects of multiple regressions.

EM wishes to thank Profs. C. Destri, G. Marchesini and F. Rapuano for helpful discussions.

Part of the research was funded by FIRB grant RBNE03B8KK from the Italian Ministry for Education, University and Research.

References

1. Pabo C, Sauer RT: **Transcription factors: structural families and principles of DNA recognition.** *Ann Rev Biochem* 1992, **61**:1053-1095.
2. Matthews B: **Protein-DNA interaction. No code for recognition.** *Nature* 1998, **335**:294-5.
3. Pabo CO, Sauer RT: **Protein-DNA recognition.** *Ann Rev Biochem* 1984, **53**:293-321.
4. Harrison SC, Aggarwal AK: **DNA recognition by proteins with the helix-turn-helix motif.** *Ann Rev Biochem* 1990, **59**:933-969.
5. Gromiha MM, Siebers JG, Selvaraj S, Kono H, Sarai A: **Intermolecular and intramolecular readout mechanisms in protein-DNA recognition.** *J Mol Biol* 2004, **337**:285-294.
6. Kono H, Sarai A: **Structure-based prediction of DNA target sites by regulatory proteins.** *Proteins* 1999, **35**:114-131.
7. Wassermann WW, Sandelin A: **Applied bioinformatics for the identification of regulatory elements.** *Nat Rev Genet* 2004, **5**:276-287.
8. Pennacchio LA, Rubin EM: **Genomic strategies to identify mammalian regulatory sequences.** *Nat Rev Genet* 2001, **2**:100-109.
9. Sinha S, Tompa M: **Discovery of novel transcription factor binding sites by statistical overrepresentation.** *Nucleic Acids Res* 2002, **30**:5549-5560.
10. Sinha S, Tompa M: **YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation.** *Nucleic Acids Res* 2003, **31**:3586-3588.
11. Birnbaum K, Benfey PN, Shasha DE: **cis element/transcription factor analysis (cis/TF): a method for discovering transcription factor/cis element relationships.** *Genome Res* 2001, **11**:1567-1573.
12. Wolfsberg TG, Gabrielian AE, Campbell MJ, Cho RJ, Spouge JL, Landsman D: **Candidate regulatory sequence elements for cell cycle-dependent transcription in Saccharomyces cerevisiae.** *Genome Res* 1999, **9**:775-792.
13. Caselle M, Di Cunto F, Provero P: **Correlating overrepresented upstream motifs to gene expression: a computational approach to regulatory element discovery in eukaryotes.** *BMC Bioinformatics* 2002, **3**:7.
14. Cora' D, Di Cunto F, Provero P, Silengo L, Caselle M: **Computational identification of transcription factor binding sites by functional analysis of sets of genes sharing overrepresented upstream motifs.** *BMC Bioinformatics* 2004, **5**:57.
15. van Helden J, Andre B, Collado-Vides J: **Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies.** *J Mol Biol* 1998, **281**:827-842.
16. Jensen LJ, Knudsen S: **Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation.** *Bioinformatics* 2000, **16**:326-333.
17. Lawrence CE, Reilly AA: **An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences.** *Proteins* 1990, **7**:41-51.
18. Thijs G, Marchal K, Lescot M, Rombauts S, Moor BD, Rouze P, Moreau Y: **A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes.** *J Comput Biol* 2002, **9**:447-464.
19. Thompson W, Rouchka EC, Lawrence CE: **Gibbs Recursive Sampler: finding transcription factor binding sites.** *Nucleic Acids Res* 2003, **31**:3580-3585.
20. Hughes JD, Estep PW, Tavazoie S, Church GM: **Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae.** *J Mol Biol* 2000, **296**:1205-14.
21. Hardison R: **Conserved non-coding sequences are reliable guides to regulatory elements.** *Trends Genet* 2000, **16**:369-372.
22. Duret L, Dorkeld F, Gautier C: **Strong conservation of non-coding sequences during vertebrates evolution: potential involvement in post-transcriptional regulation of gene expression.** *Nucleic Acid Res* 1993, **21**:2315-2322.
23. Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, Frazer KA: **Identification of a coordinate regulator of**

interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* 2000, **288:**136-140.

24. Goettgens B, Barton L, Gilbert J, Bench A, Sanchez M, Bahn S, Mistry S, Grafham D, McMurray A, Vaudin M, Amaya E, Bentley D, Green A, Sinclair A: **Analysis of vertebrate scl loci identifies conserved enhancers.** *Nat Biotechnol* 2000, **18:**181-186.
25. Flint J, Tufarelli C, Peden J, Clark K, Daniels R, Hardison R, Miller W, Philipsen S, Tan-Un K, McMorrow T, Frampton J, Alter B, Frischauf A, Higgs D: **Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the alpha globin cluster.** *Hum Mol Genet* 2001, **10:**371-382.
26. Lenhard B, Sandelin A, Mendoza L, Engström P, Jareborg N, Wasserman WW: **Identification of conserved regulatory elements by comparative genome analysis.** *J Biol* 2003, **2:**13.
27. Zhang Z, Gerstein M: **Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements.** *J Biol* 2003, **2:**11.
28. Cora' D, Herrmann C, Dieterich C, Di Cunto F, Provero P, Caselle M: **Ab initio identification of putative human transcription factor binding sites by comparative genomics.** *BMC Bioinformatics* 2005, **6:**110.
29. Sandelin A, Wasserman WW, Lenhard B: **ConSite: web-based prediction of regulatory elements using cross-species comparison.** *Nucleic Acids Res* 2004, **32:**W249-W252.
30. Prakash A, Blanchette M, Sinha S, Tompa M: **Motif discovery in heterogeneous sequence data.** *Pac Symp Biocomput* 2004:348-59.
31. Tompa M, Li N, Bailey TL, Church GM, Moor BD, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Regnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z: **Assessing computational tools for the discovery of transcription factor binding sites.** *Nat Biotechnol* 2005, **23:**137-144.
32. Lazaridis T, Karplus M: **Effective energy functions for protein structure prediction.** *Curr Op Struct Biol* 2000, **10:**139-45.
33. Sippl M: **Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins.** *J Mol Biol* 1990, **213:**859-883.
34. Lustig B, Jernigan RL: **Consistencies of individual DNA base-amino acid interactions in structures and sequences.** *Nucleic Acids Res* 1995, **23:**4707-4711.
35. Kaplan T, Friedman N, Margalit H: **Ab initio prediction of transcription factor targets using structural knowledge.** *PLoS Comput Biol* 2005, **1(1):**e1.
36. Mandel-Gutfreund Y, Margalit H: **Quantitative parameters for amino acid base interaction: implications for prediction of protein DNA binding sites.** *Nucleic Acids Res* 1998, **26:**2306-2312.
37. Liu Z, Mao F, Guo J, Yan B, Wang P, Qu Y, Xu Y: **Quantitative evaluation of protein-DNA interactions using an optimized knowledge-based potential.** *Nucleic Acids Res* 2005, **33:**546-558.
38. Zhang C, Liu S, Zhu Q, Zhou Y: **A knowledge-based energy function for protein-ligand, protein-protein and protein-DNA complexes.** *J Med Chem* 2005, **48:**2325-2335.
39. Olson WK, Gorin AA, Lu X, Hock LM, Zhurkin VB: **DNA sequence-dependent deformability deduced from protein-DNA crystal complexes.** *Proc Natl Acad Sci USA* 1998, **95:**11163-11168.
40. Zhou H, Zhou Y: **Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction.** *Protein Sci* 2002, **11:**2714-2726.
41. Schueler-Furman O, Wang C, Bradley P, Misura K, Baker D: **Progress in modeling of protein structures and interactions.** *Science* 2005, **310:**638-642.
42. Wang W, Donini O, Reyes C, Kollman P: **Simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions.** *Ann Rev Biophys Biomol Struct* 2001, **30:**211-43.
43. Kollman P, Massova I, Reyes C, Kuhn B, Huo S, Chong L, Lee M, Duan Y, Wang W, Donini O, Cieplak P, Srnivasan J, Case D, Cheatham T: **Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models.** *Acc Chem Res* 2000, **33:**889-897.
44. Zacharias M, Luty B, Davis M, McCammon J: **Poisson-Boltzmann analysis of the lambda repressor-operator interaction.** *Biophys J* 1992, **63:**1280-1285.
45. Misra V, Hecht J, Sharp K, Friedman R, Honig B: **Salt effects on protein-DNA intractions. The lambda cI repressor and EcoRI endonuclease.** *J Mol Biol* 1994, **238:**263-80.
46. Fogolari F, Elcock A, Esposito G, Viglino P, Briggs J, McCammon J: **Electrostatics in the homeodomain-DNA interaction.** *J Mol Biol* 1997, **267:**368-381.
47. Baginski M, Fogolari F, Briggs J: **Electrostatic and non-electrostatics contributions to the binding free energies of anthracycline antibiotics to DNA.** *J Mol Biol* 1997, **274:**253-267.
48. Froloff N, Windemuth A, Honig B: **On the calculation of binding free energies using continuum methods: application to MHC class I protein-peptide interactions.** *Protein Sci* 1997, **6:**1293-1301.
49. Misra V, Hecht J, Yang A, Honig B: **Electrostatic contributions to the binding free energy of the lambdacI repressor to DNA.** *Biophys J* 1998, **75:**2262-2273.
50. Wojciechowski M, Fogolari F, Baginski M: **Thermodynamic and electrostatic properties of ternary Oxytricha nova TEBP-DNA complex.** *J Struct Biol* 2005, **152:**169-184.
51. Jayaram B, McConnell K, Dixit SB, Beveridge DL: **Free energy analysis of protein-DNA binding: the EcoRI endonuclease-DNA complex.** *J Comput Phys* 1999, **151:**333-357.
52. Gorfe AA, Jelesarov I: **Energetics of sequence-specific protein-DNA association: computational analysis of integrase Tn916 binding to its target DNA.** *Biochemistry* 2003, **42:**11568-11576.
53. Endres RG, Schulthess TC, Wingreen NS: **Toward an atomistic model for predicting transcription-factor binding sites.** *Proteins* 2004, **57:**262-268.
54. Oobatake M, Kono H, Wang Y, Sarai A: **Anatomy of specific interactions between lambda repressor and operator DNA.** *Proteins* 2003, **53:**33-43.
55. Oobatake M, Ooi T: **Hydration and heat stability effects on protein unfolding.** *Prog Biophys Mol Biol* 1993, **59(3):**237-84.
56. Havranek JJ, Duarte CM, Baker D: **A simple physical model for the prediction and design of protein-DNA interactions.** *J Mol Biol* 2004, **344:**59-70.
57. Morozov A, Havranek J, Baker D, Siggia E: **Protein-DNA binding specificity predictions with structural models.** *Nucleic Acids Res* 2005, **33:**5781-5798.
58. Beamer L, Pabo C: **Refined 1.8 Åcrystal structure of the *lambda*-repressor operator complex.** *J Mol Biol* 1992, **227:**177-196.
59. Johnson AD, Poteete AR, Lauer G, Sauer RT, Ackers GK, Ptashne M: **λ-repressor and cro components of an efficient molecular switch.** *Nature* 1981, **294:**217-223.
60. Brennan R, Matthews B: **The helix-turn-helix DNA binding motif.** *J Biol Chem* 1989, **264:**1903-1906.
61. Sarai A, Takeda Y: **λ-repressor recognizes the approximately 2-fold symmetric half-operator sequences asymmetrically.** *Proc Natl Acad Sci USA* 1989, **86:**6513-6517.
62. Ptashne M: *A Genetic Switch: Phage Lambda and Higher Organisms* 2nd edition. Cambridge, MA: Cell Press and Blackwell Scientific Publications; 1992.
63. Ackers G, Johnson A, Shea M: **Quantitative model for gene regulation by λ phage repressor.** *Proc Natl Acad Sci USA* 1982, **79:**1129-1133.
64. Shea MA, Ackers GK: **The O_R control system of bacteriophage lambda. A physical-chemical model for gene regulation.** *J Mol Biol* 1985, **181:**211-230.
65. Aurell E, Brown S, Johanson J, Sneppen K: **Stability Puzzles in Phage Lambda.** *Phys Rev E Stat Nonlin Soft Matter Phys* 2002, **65(5 Pt 1):**051914.
66. Bakk A, Melzer R: **In vivo non-specific binding of lambda CI and Cro repressors is significant.** *FEBS letters* 2004, **563:**66-68.
67. **Structural Classification of Proteins** [http://scop.mrc-lmb.cam.ac.uk/scop/]
68. Takeda Y, Sarai A, Rivera VM: **Analysis of sequence-specific interaction between Cro repressor and operator DNA by systematic base substitution experiments.** *Proc Natl Acad Sci USA* 1989, **86:**439-443.
69. Benos PV, Bulyk ML, Stormo GD: **Additivity in protein-DNA interactions: how good an approximation is it?** *Nucleic Acids Res* 2002, **30(20):**4442-51.
70. Benos PV, Lapedes AS, Stormo GD: **Is there a code for protein-DNA recognition? Probab(ilistical)ly...** *Bioessays* 2002, **24:**466-475.

71. Maerkl SJ, Quake SR: **A systems approach to measuring the binding energy landscapes of transcription factors.** *Science* 2007, **315:**233-237.
72. Thomasson KA, Ouporov IV, Baumgartner T, Czaplinski J, Kaldor T, Northrup SH: **Free energy of nonspecific binding of Cro repressor protein to DNA.** *Phys Chem* 1997, **101:**9127-9136.
73. Vriend G: **WHAT IF: A molecular modeling and drug design program.** *J Mol Graph* 1990, **8:**52-54.
74. Crooks GE, Hon G, Chandonia JM, Brenner SE: **WebLogo: A sequence logo generator.** *Genome Res* 2004, **14:**1188-1190.
75. van Dijk M, van Dijk ADJV, Hsu V, Boelens R, Bonvin AMJJ: **Information-driven protein-DNA docking using HADDOCK:it is a matter of flexibility.** *Nucleic Acids Res* 2006, **34:**3317-3325.
76. Ashworth J, Havranek JJ, Duarte CM, Sussman D, Monnat RJ, Stoddard BL, Baker D: **Computational redesign of endonuclease DNA binding and cleavage specificity.** *Nature* 2006, **441:**656-659.
77. Berman H, Westbrook J, Feng Z, Gilliand G, Bhat T, Weissig H, Shindyalov I, Bourne P: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28:**235-242.
78. **Dr. Andrew C.R. Martin's Group** [http://www.bioinf.org.uk/software/]
79. Berendsen H, der Spoel D, van Drunen R: **GROMACS: a message-passing parallel MD implementation.** *Comp Phys Comm* 1995, **91:**43-56.
80. Lindahl E, Hess B, van der Spoel D: **GROMACS 3.0: A package for molecular simulation and trajectory analysis.** *J Mol Mod* 2001, **7:**306-317.
81. Madura J, Davis M, Gilson M, Wade R, Luty B, McCammon J: **Biological applications of electrostatics calculations and Brownian dynamics simulations.** *Rev Comp Chem* 1994, **5:**229-267.
82. Madura J, Briggs J, Wade R, Davis M, Luty B, Ilin A, Antosiewicz J, Gilson M, Bagheri B, Scott SLR, McCammon J: **Electrostatics and diffusion of molecules in solution: simulations with the University of Houston Brownian Dynamics program.** *Comp Phys Comm* 1995, **91:**57-95.
83. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML: **Comparison of simple potential functions for simulating liquid water.** *J Chem Phys* 1983, **79:**926-935.
84. Humphrey W, Dalke A, Schulten K: **VMD Visual Molecular Dynamics.** *J Mol Graph* 1996, **14:**33-38.
85. Andersen HC: **Rattle: a velocity version of the shake algorithm for molecular dynamics calculations.** *J Comp Phys* 1983, **52:**24-34.
86. Kale L, Skeel R, Bhandarkar M, Brunner R, Gursoy A, Krawetz N, Phillips J, Shinozaki A, Varadarajan K, Schulten K: **NAMD2: greater scalability for parallel molecular dynamics.** *J Comp Phys* 1999, **151:**283-312.
87. Berendsen HJC, Postma JPM, van Gunsteren WF, Di Nola A, Haak JR: **Molecular dynamics with coupling to an external bath.** *J Chem Phys* 1984, **81:**3684-3690.
88. Darden T, York D, Pedersen L: **Particle Mesh Ewald. An N.log(N) method for Ewald sums in large systems.** *J Chem Phys* 1993, **98:**10089-10092.
89. Gilson M, Given JA, Bush BL, McCammon J: **The statistical-thermodynamic basis for computation of binding affinities: a critical review.** *Biophys J* 1997, **72:**1047-1069.
90. Brooks B, Bruccoleri R, Olafson B, States D, Swaminathan S, Karplus M: **CHARMM: a program for macromolecular energy minimization and dynamics calculations.** *Comput Chem* 1983, **4:**187-217.
91. MacKerell A, Bashford D, Bellott M, Dunbrack R, Evanseck J, Field M, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau F, Mattos C, Michnick S, Ngo T, Nguyen D, Prodhom B, Reiher W, Roux B, Schlenkrich M, Smith J, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M: **All-atom empirical potential for molecular modeling and dynamics studies of proteins.** *J Phys Chem B* 1998, **102:**3586-3616.
92. Qiu D, Shenkin P, Hollinger F, Still W: **The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii.** *J Phys Chem* 1997, **101:**3005-3014.
93. Fogolari F, Brigo A, Molinari H: **Protocol for MM/PBSA molecular dynamics simulations of proteins.** *Biophys J* 2003, **85:**159-166.
94. Press WH, Teukolsky SA, Vetterling WT, Flannery BP: *Numerical Recipes in C* 2nd edition. Cambridge University Press; 1995.
95. Lu XJ, Olson WK: **3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures.** *Nucleic Acids Res* 2003, **31:**5108-5121.
96. Berenson ML, Levine DM, Goldstein M: *Intermediate statistical methods and applications* NJ, Prentice-Hall, Inc., Englewood Cliffs; 1983.