Weight matrices for protein-DNA binding sites from a single co-crystal structure

Robert G. Endres[1,2,*] and Ned S. Wingreen[2]
[1]*NEC Laboratories America, Inc., Princeton, New Jersey 08540, USA*
[2]*Department of Molecular Biology, Princeton University, Princeton, New Jersey 08544-1014, USA*
(Received 21 November 2005; revised manuscript received 31 January 2006; published 28 June 2006)

Transcription-factor proteins bind to specific DNA sequences to regulate gene expression in cells. DNA-binding sites are often identified using weight matrices calculated from multiple known binding sites. However, in many cases the number of examples is limited. Here, we report on an atomistic method that starts from an x-ray co-crystal structure of the protein bound to one particular DNA sequence, and infers other binding sites, which are used to construct a weight matrix. The emphasis of the paper is on using the Wang-Landau Monte Carlo algorithm to efficiently sample high-affinity binding sites, which demonstrates that sampling can produce accurate weight matrices in analogy to bioinformatics approaches. For cases of low complexity, we compare to the exhaustive (but slow) dead-end elimination algorithm. To recover crystal binding sites, it is important to include bound water in the protein-DNA interface. Our approach can, in principle, even be applied when no native protein-DNA co-crystal structure is available, only the structure of a closely related homologous protein whose amino-acid sequence is changed to the protein of interest.

DOI: 10.1103/PhysRevE.73.061921 PACS number(s): 87.15.Cc, 87.10.+e

I. INTRODUCTION

Identification of DNA-binding sites of transcription-factor proteins is essential to understand how cells respond to stimuli like nutrient availability and stress. In a bioinformatics approach, experimentally known binding sites are used to construct a weight matrix [1], which is then used to find further binding sites. The weight matrix depends on the frequencies $f_{i,b}$ of the four DNA bases (b=A,C,G,T) at each binding-site position i. Successful use of a bioinformatics weight matrix to identify new binding sites generally requires $\sim$10 or more examples of known binding sites. To identify new binding sites from fewer known examples, atomistic methods start from a relevant co-crystal structure and evaluate protein binding energies to some or all possible DNA sequences (up to 4^L, where L is the binding-site length). These atomistic approaches either neglect the protein completely [2], freeze the protein in its crystal structure [3,4], or allow protein flexibility through variable protein sidechain conformations (rotamers) within the limited search capacity of the exact dead-end elimination (DEE) algorithm [5]. The DEE algorithm iteratively eliminates high-energy residues (rotamers and base pairs) inconsistent with some chosen energy ϵ above the initially unknown ground state [6]. While the DEE algorithm works well for densely packed protein cores, it performs poorly for protein exteriors [7] and loosely bound interfaces, e.g., between a protein and DNA [5]. Furthermore, the DEE algorithm scales inefficiently with system size [8] and may fail to converge when a set range of conformations with a range of energies above the ground state is required [9].

In contrast to the exhaustive atomistic approach using the DEE algorithm, the bioinformatics approach uses a small sample of binding sites to construct a weight matrix. The success of weight matrices based on small samples suggests that sampling could be employed to accelerate atomistic approaches. However, temperature-dependent Monte Carlo (MC) algorithms and genetic algorithms are expected to perform poorly due to the roughness of the energy landscape [8]. In contrast, the Wang-Landau (WL) [10] algorithm overcomes these limitations: The WL algorithm uses a sampling acceptance probability independent of temperature and inversely proportional to the density of states (DOS). As a result, sampling of conformations follows a biased random walk in energy space until the histogram of visited energies becomes flat. This biases sampling toward regions of low DOS, and can be used to efficiently obtain conformations near the ground state. The WL algorithm has been applied successfully to protein-folding models [11].

In this work, we apply the WL algorithm to sample transcription-factor binding sites within a fixed energy of the ground state. For the case of a frozen protein or when only a very small number of rotamers is used, we compare with exact results from the DEE algorithm. Identification by the WL algorithm of a sample of low-energy binding sites proves sufficient to construct a highly accurate weight matrix. For our tests, we considered three "zinc-finger" DNA-binding proteins [12–14]. Transcription factors of this class are common in eukaryotes [15], comprising $\sim$2% of the human genome [16], and are promising candidates for gene therapy [17].

II. MATERIALS AND METHODS

A detailed atomistic model based on an x-ray co-crystal structure and a rotamer library [18] has been introduced by us previously [5] (see Fig. 1). Some practical improvements that we have implemented are outlined below. As required by the DEE algorithm, the binding energy was decomposed to be pairwise in sidechains and base pairs. We used the CHARMm29 package with force field *par_all27_prot_na.prm* [19]. Solvent effects were included based on solvent-accessible surface area and atomic solvation parameters [20]. Solvent effects introduce corrections to the pairwise energy due to multiple burial of surfaces. To

*Corresponding author. Email address: rendres@princeton.edu

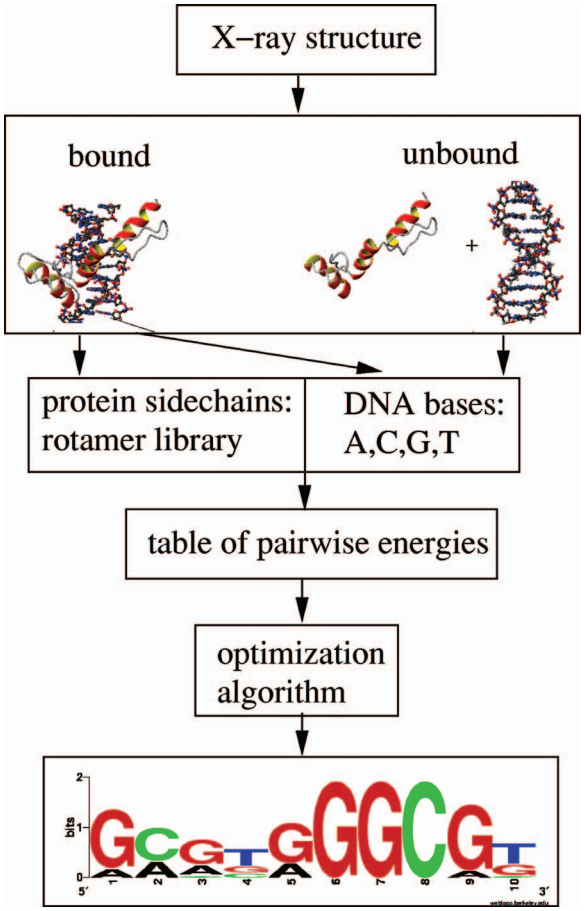

FIG. 1. (Color) Flowchart of algorithm.

keep these corrections small, i.e., to ensure that the pairwise approximation is close to the true energy for each configuration, we employed generic sidechains (Ala residues) and base pairs (G-C base pairs without partial charges or methyl groups). The use of generic sidechains captures much of the multiple-burial effect within the pairwise energy [21]. The binding energy ΔE defined as the difference between bound (b) and unbound (u) configurations for a given set of rotamers and base pairs $\{r\}$ is given by

$$\Delta E(\{r\}) = \Delta E_0 + \sum_{i=1}^{N} \left([\Delta E_1(i_r) - \Delta E_0] + \sum_{i'=i+1}^{N} [\Delta E_2(i_r, i'_{r'}) - \Delta E_1(i_r) - \Delta E_1(i'_{r'}) + \Delta E_0] \right), \quad (1)$$

where N is the sum of the number of amino acids (M) and number of DNA base pairs (L) being optimized. Equation (1) is an expansion around a protein-DNA template binding energy ($\Delta E_0 = E_0^b - E_0^u$). The template has all flexible residues replaced by generic ones. $\Delta E_1(i_r) = E_1^b(i_r) - E_1^u(i_r)$ and $\Delta E_2(i_r, i'_{r'}) = E_2^b(i_r, i'_{r'}) - E_2^u(i_r, i'_{r'})$ are the binding energies where one and two generic residues are replaced by specific conformations at position i and position pair i, i', respectively. If i_r (or $i'_{r'}$) is an amino acid, the conformation from the crystal structure is used for the unbound compound. Corrections beyond second order are generally small ($\sim 0.1\%$), and, hence, are neglected. To model unbound DNA, we do not use canonical B DNA, since this may introduce a bias toward certain DNA sequences (destabilized B DNA structures lead to stabilized binding). Instead, we use precalculated energies $\Delta E_1(i_r)$ and $\Delta E_2(i_r, i'_{r'})$ when i_r and $i'_{r'}$ are DNA base pairs. These energies are averages from 15 relaxed 10-base-pair-long DNA structures of random sequences.

In this study, we consider three cases of different complexity. Case I uses a maximum of five rotamers per optimized amino acid. Case II uses about 20 rotamers per amino acid (up to 80 for Arg and Lys). In both cases, these rotamers are used in addition to the set of native rotamers measured from the crystal structure. Case III freezes the protein into its crystal structure. In a variant of case III, we include the crystal water molecules, enabling us to study the effects of specific water-mediated interactions. The resulting weight matrices are calculated from binding sites within 20 kcal/mol of the strongest binding site in order to include at least ten binding sites.

Our primary reference co-crystal x-ray structure consists of Zif268 [12] [Protein Data Bank (PDB) code 1aay] bound to its experimental consensus sequence GCGTGGGCGT ($L = 10$) with $M = 25$ DNA-contacting amino acids [5]. We also consider variants of Zif268, PDB structures 1g2f [13] and 1mey [14]. Co-crystal structure 1g2f is used to find the weight matrix of Zif268 based on homology modeling, i.e., if the crystal structure of the native protein-DNA complex (1aay) was not available. For this purpose, we replaced the amino-acid sequences of the three α-helices in structure 1g2f by the corresponding amino-acid sequences of Zif268. Specifically, we replaced in the first α-helix QLTNLDT by RSDELTR, in the second α-helix QQASLNA by RSDHLTT, and in the third α-helix TLHTAT by ASDERL from the N to the C terminus. Several variants of the DEE algorithm were implemented [7] including the super-residue approach [22], which combines the remaining residues of two or more positions into new super-residues facilitating new rounds of elimination. While the super-residue scheme ultimately leads to a converged result, i.e., identifying all conformations with binding energies within energy ϵ of the ground state, the growing memory requirements rapidly make the approach impracticable. For instance, we were able to use the DEE algorithm for case I even for a relatively large $\epsilon = 30$ kcal/mol, but not for case II even for $\epsilon = 5$ kcal/mol. (Only the ground state was obtainable by DEE for case II.)

As an alternative to the DEE approach, the WL algorithm computes the DOS $g(E)$ iteratively on a discrete energy grid (bins) [10]. Due to its statistical nature, the WL algorithm is not guaranteed to find the exact ground state, but returns multiple conformations within a few kcal/mol in minutes for the cases studied here. Convergence to an accurate DOS is controlled by a factor $f > 1$ which is reduced in steps toward 1. Starting from a conformation with binding energy E_1, a new conformation is generated at random. The probability of accepting a move to the new conformation depends on its binding energy E_2 according to

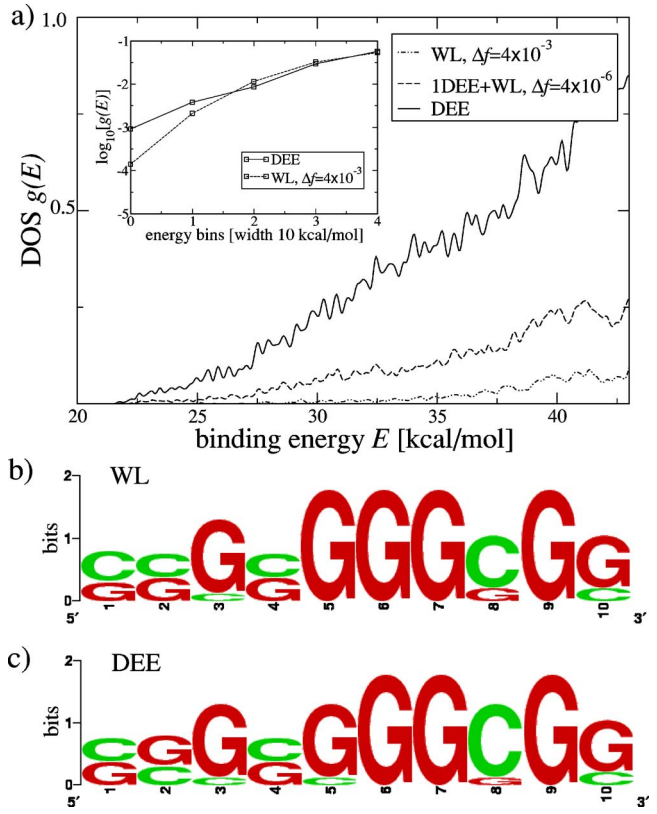

FIG. 2. (Color online) Comparison of Wang-Landau and dead-end elimination algorithms for case I (up to 5 rotamers per sidechain) using Zif268 (1aay). (a) Unnormalized DOS $g(E)$ calculated from binding energies using Gaussian broadening 0.1 kcal/mol. Inset: $\log[g(E)]$ of 5 lowest energy bins, where both $g(E)$ were independently normalized. (b) and (c) Sequence LOGOS based on 10 best binding sequences. DEE energy cut-off: $\epsilon = 30$ kcal/mol. CPU times based on single 3.6 GHz Intel Xeon CPU: 7 h for both WL and 1DEE+WL curves.

$$p(E_1 \to E_2) = \min\left(\frac{g(E_1)}{g(E_2)}, 1\right). \tag{2}$$

If the move is accepted (not accepted), the DOS $g(E)$ and the histogram $H(E)$ are both updated in energy bin E containing E_2 (E_1), $g(E) \to g(E) \times f$, and $H(E) \to H(E)+1$. Once the histogram is sufficiently flat [$|H(E)-\langle H\rangle| < 0.2$ for all energy bins E], the histogram is reset [$H(E) \to 0$] while keeping $g(E)$, and the convergence factor is reduced ($f_{i+1} = \sqrt{f_i}$ where $f_0 = e \approx 2.718\,28\ldots$).

III. RESULTS

A. Comparison of Wang-Landau and DEE algorithms

For case I, Fig. 2(a) compares $g(E)$ obtained from the WL algorithm (dash-dotted line) with results from the exact DEE algorithm using super-residues (solid line). Also shown is the result from one round of the DEE algorithms without super-residues (dashed line) followed by the WL algorithm, which leads to a drastic acceleration (smaller $\Delta f = f - 1$ for the same computational time). The WL algorithm misses many binding sites (or DOSs) compared to the DEE algorithm, especially for low energies (see inset), but for both algorithms the sequence LOGOs, i.e., graphical representations of weight matrices [23], are remarkably similar [see Figs. 2(b) and 2(c)].

For case II, Fig. 3(a) shows $g(E)$ obtained from the WL algorithm after a single round of DEE without super-residues for various limits of convergence Δf. The weight matrix obtained for $\Delta f = 10^{-4}$ is shown in Fig. 3(b). This example shows that even when the DEE algorithm is not capable of converging for $\epsilon > 0$, the WL algorithm is still able to produce a converged sequence LOGO. Since the LOGO or weight matrix is generally the desired final result, the WL algorithm presents a fast and powerful alternative to the DEE algorithm.

B. Potential improvements and effects of crystal water

How good is the predictive power of our atomistic model? Using Zif268 as an example, our model does not identify the full consensus sequence when decoy rotamers are allowed [in Fig. 3(b), a nucleotide of the crystal binding site is underscored if it is predicted correctly by the largest letter of the LOGO at the same position]. There are many possible reasons for an incorrectly predicted nucleotide. The list of modeling simplifications includes frozen backbones for protein and DNA, classical force fields, and an implicit water model, as well as neglect of both crystal waters and entropy changes upon binding. In order to identify possible sources of discrepancy, we used the fact that the protein and DNA interfacial surfaces are stereochemical complements of each other, which leads to a strong bias of the frozen protein toward its crystal binding site. We used this bias of the frozen protein to test modeling improvements without the additional uncertainties of the limited rotamer library and incomplete sampling of binding sites. For the frozen protein model, the WL and DEE algorithms produced identical sequence LOGOs.

Figures 4(a) and 4(b) show the minor effect of allowing protein backbone flexibility. This was done by relaxing both the bound and unbound structures to a local minimum and recalculating the binding energies. Care was taken to include enough high-energy structures, since relaxation can result in significant reranking. Next, we tested the role of water-mediated hydrogen bonds. The importance of bound water at the protein-DNA interface is well established [24]. Figure 4(c) shows the effect of including crystal waters that mediate contacts at base pair positions 2, 4, 8, and 10. At this level of modeling, the sequence of largest letters (bases) from the LOGO reproduces the crystal binding site and consensus sequence.

C. Other Zif268-like proteins and homology modeling

The model with backbone flexibility and explicit water molecules correctly predicts the binding site from the co-crystal structure when this structure corresponds to the optimal binding site (as for 1aay). In order to see how robust the predictive power is, we examined two cases when the co-crystal sequence does not correspond to the consensus se-

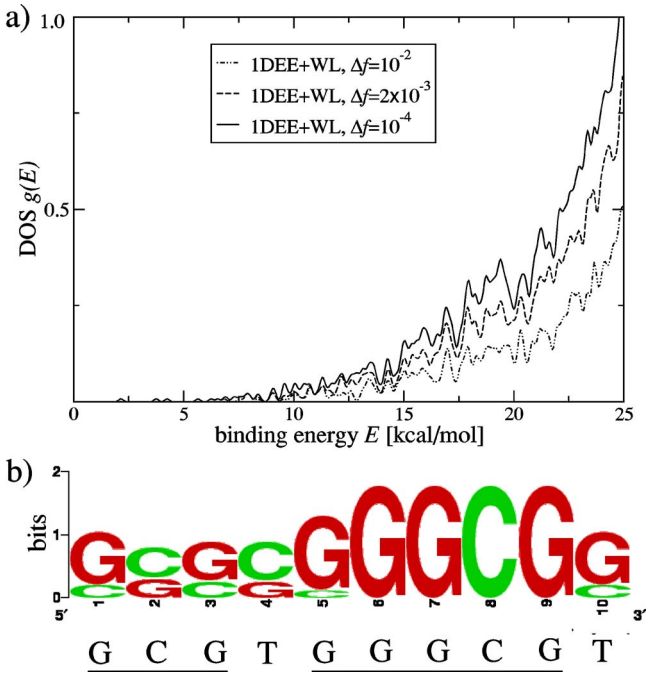

FIG. 3. (Color online) Results for case II (≈20 rotamers per sidechain) using Zif268 (1aay) after an initial round of DEE followed by the WL algorithm. (a) DOS $g(E)$ for binding energies (Gaussian broadening 0.1 kcal/mol) for different convergence limits Δf. The true ground-state energy obtained from the DEE algorithm was set to zero. CPU times are 1 h ($\Delta f = 10^{-2}$), 2 h ($\Delta f = 2 \times 10^{-3}$), and 3.5 h ($\Delta f = 10^{-4}$). (b) Sequence LOGOs based on the 10 best binding sequences for $\Delta f = 10^{-4}$, compared to experimental consensus sequence (black letters). Correctly predicted nucleotides are underscored.

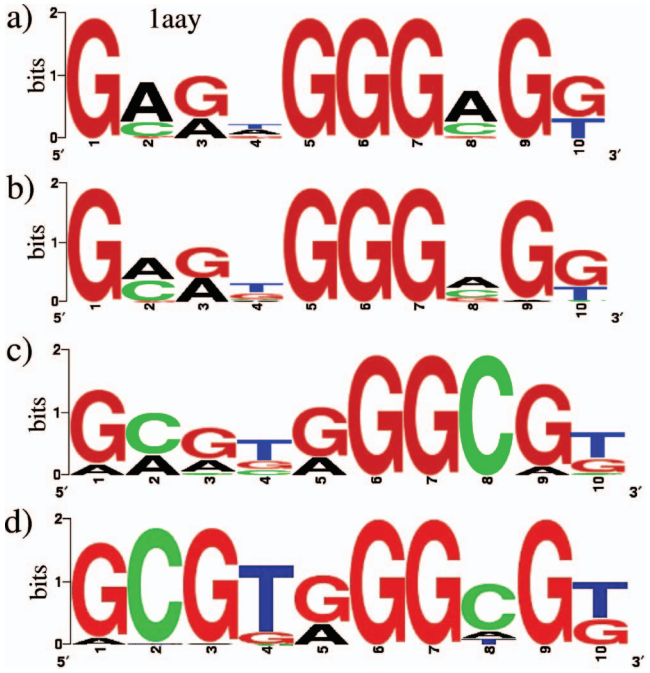

FIG. 4. (Color) Sequence LOGOs based on the 30 best binding sequences of Zif268 (1aay) without sidechain flexibility. Several levels of modeling (a)-(c) lead to step-wise improvement toward experimental result (d): (a) as returned from the DEE or WL algorithm; (b) re-ranked binding energies through relaxation of 100 best bound and unbound structures with 30 steepest decent steps using CHARMm; (c) same, but crystal water molecules are included; (d) based on experimentally known binding sites [25]. The sequence of the largest letters (bases) in (c) and (d) is the crystal binding site and consensus sequence.

quence. Figure 5(a) shows the LOGO obtained from a Zif268 variant bound to the eukaryotic TATA promoter region (1g2f) while Fig. 5(b) shows the LOGO for a designed zinc-finger protein bound to a suboptimal binding site (1mey). In both cases we included explicit waters and allowed relaxation [cf. Fig. 4(c)]. Most but not all nucleotides were correctly predicted (underscored). Incorrectly predicted sites may indicate remaining modeling limitations such as the classical force field and the neglect of entropic contributions to binding, but more likely indicate that only a subset of base pairs is actually used for binding.

The application of our approach could be greatly extended if instead of the native protein-DNA co-crystal structure, the co-crystal structure of a homologous protein could be used. Such an approach is possible provided the structure and docking arrangement are sufficiently similar. To obtain a weight matrix through homology modeling, we start from the co-crystal structure of 1g2f [13], a Zif268 variant, differing in the amino-acid sequences of the α-helices. For modeling, we replace the corresponding α-helix sequences of 1g2f with the sequences from Zif268 (1aay) as outlined in Sec. II. The sidechain conformations were simply taken from 1aay, but generally would result from an energy optimization of the protein. As shown in Fig. 6(a), the frozen protein has a significantly different weight matrix compared to frozen native Zif268 [cf. Fig. 4(a)]. The differences presumably originate from the slightly different orientations of the α-helices [13]. In contrast, Fig. 6(b) shows the weight matrix obtained using the flexible protein. This allows us to recover a GC-rich weight matrix characteristic of Zif268 [cf. Fig. 3(b)].

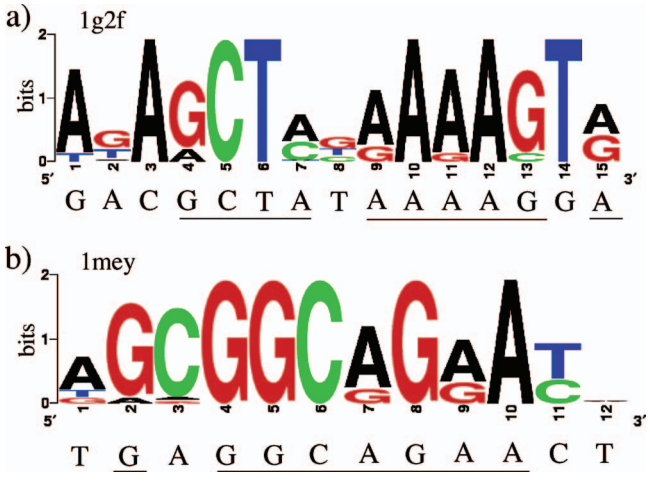

FIG. 5. (Color) Sequence LOGOs based on the 30 best binding sequences of two frozen proteins including crystal waters after relaxation. (a) 1g2f (variant of Zif268) and (b) 1mey (redesigned zinc finger protein). Crystal binding sites are given beneath the LOGOs; correctly predicted nucleotides are underscored.

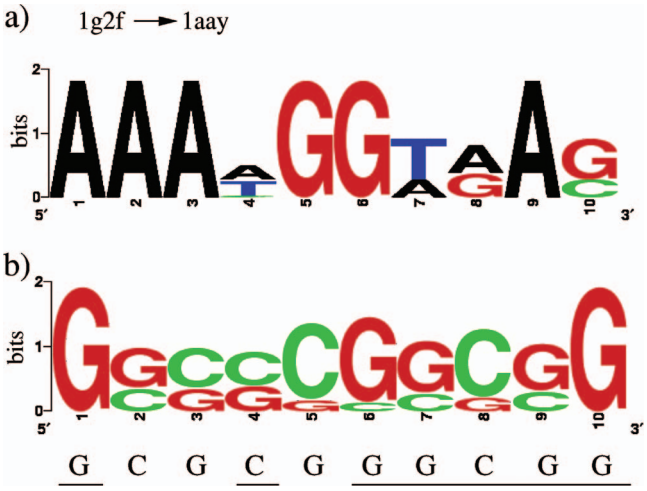

FIG. 6. (Color) Homology modeling based on structure 1g2f and amino-acid sequence 1aay, (a) frozen protein and (b) flexible protein. For the DNA-binding site search, only basepairs 4-15 of 15-basepair long DNA segment in 1g2f are optimized [cf. Fig. 5(a)]. Sequence LOGOs based on the 30 best binding sequences; only basepairs 4-13 are shown. Consensus sequence from Fig. 3(b) for flexible native Zif268 (1aay) is shown for comparison; correctly predicted nucleotides are underscored.

IV. CONCLUSIONS

DNA-binding sites are often identified using weight matrices calculated from multiple known binding sites. However, in many cases the number of known binding sites is limited. Our atomistic method starts from a co-crystal structure of the protein bound to one particular DNA sequence, e.g., obtained from x-ray diffraction or NMR, and infers other binding sites, which are used to construct a weight matrix. However, we do not aim to find the truly best binding sites, which is a difficult task considering the rough potential energy surface in the high-dimensional conformational space of rotamers and DNA base pairs. Instead, we follow bioinformatics approaches and sample high-affinity binding sites with the efficient temperature-independent Wang-Landau Monte Carlo algorithm. We demonstrate that this sampling produces accurate weight matrices when compared to the slow but exact dead-end elimination algorithm. In particular, the Wang-Landau algorithm is most effective when combined with initial elimination of energetically unfavorable conformations using the dead-end elimination algorithm. The dead-end elimination algorithm by itself is not likely to converge, particularly when binding energies within some energy window above the ground state are required. Although very good weight matrices have recently been obtained with the temperature-dependent Metropolis MC algorithm with an improved energy function [26], use of the WL algorithm should offer significant computational improvements.

Often the native protein–DNA co-crystal structure may not be available, but a co-crystal structure of a related, homologous protein may be. Using our approach, we demonstrate homology modeling by changing the amino-acid sequence in a co-crystal structure of a variant of Zif268 to the amino-acid sequence of Zif268, and recover a GC-rich weight matrix typical of Zif268 when the protein is allowed to be flexible. Further improvements of the proposed method will require improved classical force fields [26,27], in particular for water, as well as prediction of the locations of bound water [28].

[1] G. D. Stormo, Bioinformatics **16**, 16 (2000).
[2] N. R. Steffen, S. D. Murphy, L. Tolleri, G. W. Hatfield, and R. H. Lathrop, Bioinformatics **18**, S22 (2002).
[3] G. Paillard and R. Lavery, Structure (London) **12**, 113 (2004).
[4] G. Paillard, C. Deremble, and R. Lavery, Nucleic Acids Res. **32**, 6673 (2004).
[5] R. G. Endres, T. C. Schulthess, and N. S. Wingreen, Proteins **57**, 262 (2004).
[6] J. Desmet, M. de Maeyer, B. Hazes, I. Lasters, Nature (London) **356**, 539 (1992).
[7] D. B. Gordon and S. L. Mayo, J. Comput. Chem. **19**, 1505 (1998); N. A. Pierce *et al.*, *ibid.* **21**, 999 (2000).
[8] C. A. Voigt, D. B. Gordon, and S. L. Mayo, J. Mol. Biol. **299**, 789 (2000).
[9] A. R. Leach and A. P. Lemon, Proteins **33**, 227 (1998).
[10] F. Wang and D. P. Landau, Phys. Rev. Lett. **86**, 2050 (2001).
[11] N. Rathore and J. J. de Pablo, J. Chem. Phys. **116**, 7225 (2002).
[12] M. Elrod-Erickson, M. A. Rould, L. Nekludova, and C. O. Pabo, Structure Fold. Des. **4**, 1171 (1996).
[13] S. A. Wolfe, R. A. Grant, M. Elrod-Erickson, and C. O. Pabo, Structure (London) **9**, 717 (2001).
[14] C. A. Kim and J. M. Berg, Nat. Struct. Biol. **3**, 940 (1996).
[15] R. R. Beerli and C. F. Barbas III, Nat. Biotechnol. **20**, 135 (2002).
[16] R. Tupler, G. Perini, and M. R. Green, Nature (London) **409**, 832 (2001).
[17] A. C. Jamieson, J. C. Miller, and C. O. Pabo, Nat. Rev. **2**, 361 (2003).
[18] R. L. Dunbrack, Jr., Curr. Opin. Struct. Biol. **12**, 431 (2002).
[19] B. R. Brooks *et al.*, J. Comput. Chem. **4**, 187 (1983); http://www.charmm.org
[20] L. Wessen and D. Eisenberg, Protein Sci. **1**, 227 (1992).
[21] N. Zhang, C. Zeng and N. S. Wingreen, Proteins **57**, 565 (2004).
[22] R. Goldstein, Biophys. J. **66**, 1335 (1994); L. L. Looger and H. W. Hellinga, J. Mol. Biol. **307**, 429 (2001).
[23] T. D. Schneider and R. M. Stephens, Nucleic Acids Res. **18**, 6097 (1999).
[24] K. Nadassy, S. J. Wodak, and J. Janin, Biochemistry **38**, 1999 (1998); B. Jayaram and T. Jain, Annu. Rev. Biophys. Biomol. Struct. **33**, 343 (2004).
[25] A. Swirnoff and J. Milbrandt, Mol. Cell. Biol. **15**, 2275 (1995).
[26] A. V. Morozov, J. J. Havranek, D. Baker, and E. D. Siggia, Nucleic Acids Res. **33**, 5781 (2005).
[27] A. V. Morozov, T. Kortemme, K. Tsemekhman, and D. Baker, Proc. Natl. Acad. Sci. U.S.A. **101**, 6946 (2004).
[28] L. Jiang, B. Kuhlman, T. Kortemme, and D. Baker, Proteins **58**, 893 (2005).