Preferred Binding Sites of Gene-Regulatory Proteins Based on the Deterministic Dead-End Elimination Algorithm

Robert G. Endres[1], Thomas C. Schulthess[1], and Ned S. Wingreen[2] *

[1] Center for Computational Sciences and Computer Science & Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6164, USA
[2] NEC Laboratories America, Inc., 4 Independence Way, Princeton, NJ 08540, USA

Abstract. Identifying the specific DNA-binding sites of regulatory proteins is essential to understanding gene expression in the biological cell. Rather than employing statistical bioinformatics tools, we choose an atomistic approach based on the protein-DNA co-crystal structure and binding energies obtained from the AMBER force field. Our method uses the deterministic dead-end elimination and A* algorithms to search a library of discrete amino-acid sidechain rotamers and DNA basepairs for the optimal binding conformations. As an example, we search all possible 10-basepair DNA sequences (4^{10}) of Zif268, a well-studied mouse regulatory protein.

1 Introduction

Gene expression is regulated by proteins which bind to the DNA in the vicinity of the gene. However, these proteins can regulate many different genes by binding to different, albeit similar DNA sequences. Fast methods to determine binding sites are demanded in order to keep up with the sequencing of new genomes, and the desire to find new therapeutics for diseases.

While experimental methods based on DNase I footprinting [1] and DNA microarrays [2] are laborious and error-prone[3], standard statistical bioinformatics tools like the consensus sequence and the weight-matrix require prior knowledge of a lot of binding sites to succeed [4]. Both methods make the approximation of treating each DNA-sequence position as independent[5].

In contrast to these methods, we emphasize that protein-DNA binding is based on physical interactions at the atomic scale, *i.e.* electrostatic, van-der Waals, and hydrophobic interactions as well as (water mediated) hydrogen bonds. The deformability of DNA is also an important part of the binding specificity, since a protein-induced deformation of DNA upon binding will occur more readily for certain DNA sequences [6]. Previous atomistic methods have focused on this aspect of binding, but either neglected the protein completely [7] or modeled it as a rigid object[8]. We were therefore motivated

* Present address: Department of Molecular Biology, Princeton University, Princeton, NJ 08544-1014, USA

to develop a predictive theory of protein-DNA interactions which includes protein-sidechain flexibility in a fully atomistic description.

Our model consists of the fixed protein and DNA backbones obtained from a relevant co-crystal structure as shown in Fig. 1. For the unbound system we take a canonical B-DNA form, widely separated from the protein. No explicit water molecules are included in the current approach. To find preferred binding sites, we search the combined space of all DNA sequences and a library of discrete sidechain rotamers for the DNA-contacting amino acids[9,10]. Since temperature-dependent statistical algorithms often fail to find the global minimum on rough potential energy surfaces[11], we resort to the deterministic dead-end elimination (DEE) [12] and A* algorithms [13]. Binding energies are evaluated using the molecular-mechanics package AMBER 7 (parm99 force field) [14]. An initial fast search for preferred DNA sequences is made using vacuum energies, but the final ranking is done with the generalized-Born implicit-solvent model (GB/SA) [15]. The AMBER force field is able to capture the fine structure of DNA, including the A- to B-DNA transition, even with the implicit-solvent model [16]. The same force field and conformational-search algorithm have been used successfully for protein design [13].

2 Numerical Methods

To find the highest-affinity binding sites, we use the dead-end-elimination (DEE) [12]. If it converges, DEE deterministically finds the global energy minimum. As a necessary condition to apply DEE, the total binding energy is decomposed in a pairwise fashion,

$$\Delta E^{\mathrm{vac}} = \sum_{i=1}^{P} [\ \Delta\epsilon(i_r) + \sum_{j=i+1}^{P} \Delta\epsilon(i_r, j_s)], \tag{1}$$

where $P = N_{aa} + N_{bp}$ is the sum of the total numbers of protein amino acids N_{aa} and DNA basepairs N_{bp}. The quantities $\Delta\epsilon(i_r)$ and $\Delta\epsilon(i_r, j_s)$ are the self-energies and interaction energies of rotamers/basepairs of the bound minus unbound structures, where r and s describe rotamers/basepairs, while i and j are protein/DNA sequence positions. The resulting number of all possible conformations is extremely large, $O(n^P)$, if n is the average number of rotamers/basepairs at each position. Since a pairwise expression for the binding energy is strictly only correct in vacuum [17], we **pre-calculated** the self-energies and interaction energies without the solvent.

We implemented several levels of DEE algorithms as recently described in Refs. [18,19], including the so-called super-rotamer approach [20]. Just to mention two illustrative examples of DEEs, there is Desmet's original dead-end elimination criterion [12] which is the least-efficient but best scaling DEE algorithm regarding CPU-time, $\sim O(P^2 n^2)$. It uses the fact that a potential

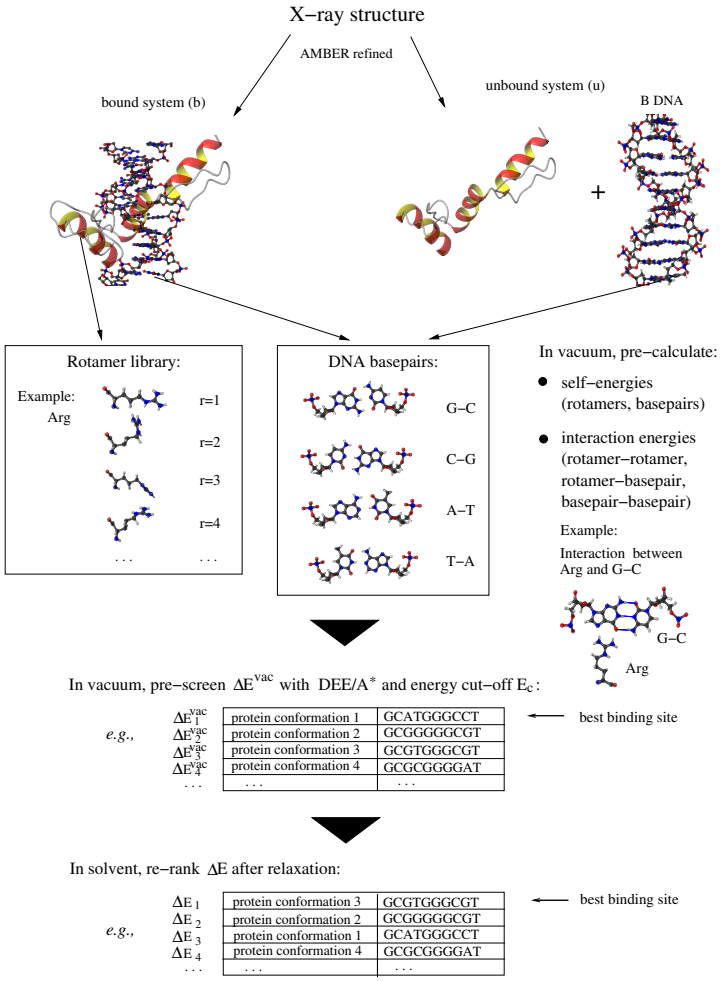

Fig. 1. Overview of atomistic method to identify DNA binding sites. (As an example, we choose the protein Zif268 whose "ribbon"-structure is shown.)

rotamer/basepair i_r is inconsistent with binding energies within E_c of the global minimum and hence can be eliminated (*i.e.* is a "dead-ending single" or DES), if its best possible energy contribution to the total binding energy is still worse than the worst possible energy contribution of an alternative rotamer/basepair t at the same position i by more than E_c,

$$\Delta\epsilon(i_r) + \sum_{j\neq i}^{P} \min \Delta\epsilon(i_r, j_s) > \Delta\epsilon(i_t) + \sum_{j\neq i}^{P} \max \Delta\epsilon(i_t, j_s) \; + \; E_c. \qquad (2)$$

Here, min (max) is the minimum (maximum) interaction energy between rotamer/basepair i_r (i_t) and all possible rotamers/basepairs at position j. Our

second example is Goldstein's [20] sharpest but worst scaling, $\sim O(P^3 n^5)$, criterion for identifying so-called "dead-ending pairs" (DEPs). Pair $\Delta\epsilon(i_r, j_s)$ is dead-ending, if

$$\Delta\epsilon(i_r, j_s) - \Delta\epsilon(i_u, j_v) + \sum_{k \neq i,j}^{P} \min\{\Delta\epsilon[(i_r, j_s), k_t] - \Delta\epsilon[(i_u, j_v), k_t]\} > E_c, \quad (3)$$

where $\Delta\epsilon[(i_{r/u}, j_{s/v}), k_t] = \Delta\epsilon(i_{r/u}, k_t) + \Delta\epsilon(j_{s/v}, k_t)$. These DEPs cannot be eliminated directly, but do not need to be included in subsequent calculations of the minimum in Eq. 2. As illustrated in Fig. 2, the DEE/E_c algorithms for DESs and DEPs are applied iteratively until no further DESs can be eliminated. Note, however, that according to Eq. 2 the energy cut-off E_c is applied to each position i separately. Consequently, even with perfect convergence of the DEEs, the resulting total binding energies can still be many times E_c above the global minimum. The number of unwanted, high-energy rotamer and basepair combinations can be further reduced by combining rotamers/basepairs from different positions to super-rotamers [20]. For instance, pairing rotamers/basepairs from two different positions eliminates all of their previously flagged DEPs, and effectively applies the energy cut-off to the two positions simultaneously, leading to a stricter enforcement of the cut-off criterion. Unfortunately, the memory requirement for storing the increased number of interaction energies is a limiting factor. After several rounds of DEE algorithms and of forming super-rotamers, the remaining combinatorial space can be searched using the graph-search A* algorithm [13]. In principle, this **pre-screening** procedure returns an exhaustive list of rotamer-DNA sequence combinations with vacuum binding energies within E_c of the best.

Finally, we **re-rank** the best vacuum protein-DNA conformations using the solvated binding energy $\Delta E = E_b - E_u$, where E_b and E_u are the total energies of the bound and unbound protein-DNA systems in implicit solvent. Since the re-ranking in solvent is likely to change the order of the binding energies compared to vacuum, the cut-off E_c has to be chosen large enough to include the top solvated configurations.

3 Results and Discussion

As an example, we apply our method to the mouse protein, Zif268 (PDB file 1AAY, Fig. 1). Zif268 has a tandem repeat of three specific binding motifs, so-called zinc fingers, each stabilized by a zinc ion (for a review see Ref. [3]). A high resolution co-crystal structure is available [21] with the protein bound to its consensus sequence, GCGTGGGCGT [22].

The best calculated binding sequence in solvent was TG**GTGGGCGG**, the 10th was **GCGTGGGCG**A, while the 17th matches the experimental consensus **GCGTGGGCGT**, as indicated by the underlining. Note that the 10th (last) base position of the binding sequence is experimentally only

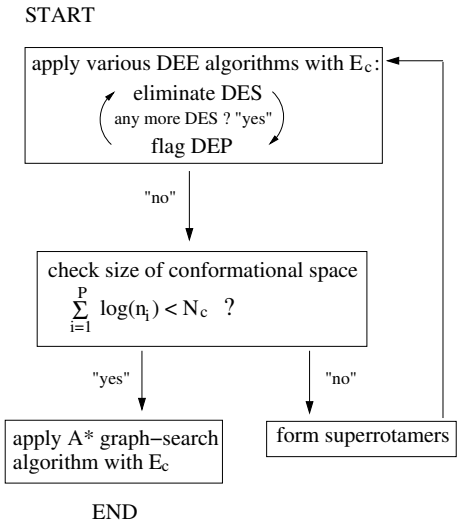

Fig. 2. Flowchart of the dead-end elimination (DEE) and A* graph-search algorithms. DES and DEP are defined in Eq. 2 (3). The parameters E_c and N_c are the energy cut-off and the upper limit of conformational size managable by the A* algorithm, respectively. Furthermore, $P = N_{aa} + N_{na}$ (see Eq. 1), and n_i is the number of rotamers/basepairs at position i. Due to convergence problems of the DEE/A*, we restricted ourselves to only a single (best) set of rotamers for each vacuum DNA sequence that was not eliminated.

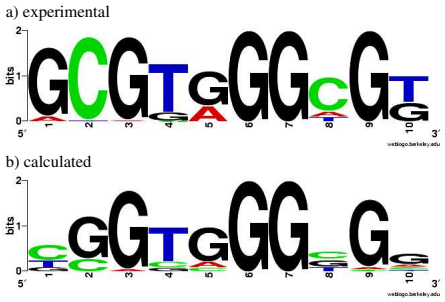

Fig. 3. (a) Zif268 10-base sequence logo from experiment [22], and (b) from our calculation. Figure prepared with the WebLogo [24] from http://weblogo.berkeley.edu/ .

weakly conserved, and was not considered part of a previously derived consensus sequence [23].

Fig. 3 compares the experimental logo (a graphical presentation of the weight-matrix [24]) (a) with our calculations using the 30 best binding sequences (b). The logos clearly show strong similarities, but the lack of explicit water molecules in our model, end-effects caused by the short DNA segment, and an additional approximation (see caption of Fig. 2) lead to some disagreement - particularly at the DNA-positions 1 and 2 contacted by the zinc finger which has the most water-mediated contacts [25].

In summary, we have introduced a novel methodology for identifying potential protein-DNA binding sites through a global search of DNA sequences and DNA-contacting-sidechain conformations. We tested performance by calculating the DNA-sequence logo for the mouse transcription factor Zif268, and found good agreement with experiment.

Acknowledgement. Research sponsored by the Division of Mathematical Information and Computational Sciences, U.S. Department of Energy, under Contract DE-AC05-00OR22725 with UT-Battelle LLC.

References

1. M.L. Bulyk, X. Huang, Y. Choo, and G.M. Church: Proc. Natl. Acad. Sci. USA **98**, 7158 (2001)
2. T.I. Lee, N.J. Rinaldi, F. Robert, D.T. Odom, Z. Bar-Joseph, G.K. Gerber, N.M. Hannett, C.T. Harbison, C.M. Thompson, and I. Simon I. *et al.*: Nature **298**, 799 (2002)
3. C.O. Pabo, E. Peisach, and R.A. Grant: Annu. Rev. Biochem **70**, 313 (2001)
4. G.D. Stormo: Bioinformatics **16**, 16 (2000)
5. P.V. Benos, M.L. Bulyk, and G.D. Stormo: Nucleic Acids Res. **30**, 4442 (2002)
6. W.K. Olson, A.A. Gorin, X.J. Lu, L.M. Hock, and V.B. Zhurkin: Proc. Natl. Acad. Sci. USA **95**, 11163 (1998)
7. N.R. Steffen, S.D. Murphy, L. Tolleri, G.W. Hatfield, and R.H. Lathrop: Bioinformatics **18**, S22 (2002)
8. I. Lafontaine and R. Lavery: Biopolymers (Nucleic Acid Sciences) **56**, 292 (2001)
9. R.L. Dunbrack, Jr. : Curr. Opin. Struct. Biol. **12**, 431 (2002)
10. All rotamers in the library *bbdep02.May.sortlib* (downloaded from http://dunbrack.fccc.edu/bbdep) were employed, except for Arg and Glu. For these long-sidechain amino acids, we used only the 20 statistically most significant rotamers (probability $\gtrsim 0.1\%$).
11. C.A. Voigt, D.B. Gordon, and S.L. Mayo: J. Mol. Biol. **299**, 789 (2000)
12. J. Desmet, M. de Maeyer, B. Hazes, and I. Lasters: Nature **356**, 539 (1992)
13. A.R. Leach and A.P. Lemon: Proteins **33**, 227 (1998)
14. D.A. Pearlman, D.A. Case, J.W. Caldwell, W.S. Ross, T.E. Cheatham III, S. Debolt, D. Ferguson, G.L. Seibel, and P.A. Kollman: Comp. Phys. Commun. **91**, 1 (1995)
15. A. Onufriev, D. Bashford, and D.A. Case: J. Phys. Chem. B **104**, 3712 (2000)
16. V. Tsui and D.A. Case: J. Am. Chem. Soc. **122**, 2489 (2000)
17. A.G. Street and S.L. Mayo: Fold. Des. **3**, 253 (1998)
18. D.B. Gordon and S.L. Mayo: J. Comp. Chem. **19**, 1505 (1998)
19. N.A. Pierce, J.A. Spriet, J. Desmet, and S.L. Mayo: J. Comp. Chem. **21**, 999 (2000)
20. R. Goldstein: Biophys. J. **66**, 1335 (1994)
21. M. Elrod-Erickson, M.A. Rould, L. Nekludova, and C.O. Pabo: Structure **4**, 1171 (1996)
22. A. Swirnoff and J. Milbrandt: Mol. Cell. Biol. **15**, 2275 (1995)
23. B. Christy and D. Nathans: Proc. Natl. Acad. Sci. USA **86**, 8737 (1989)
24. T.D. Schneider and R.M. Stephens: Nucleic Acids Res. **18**, 6097 (1999)
25. G. Roxström, I. Velázquez, M. Paulino, and O. Tapia: J. Biomol. Struct. Dyn. **16**, 301 (1998)