J Mol Model (2001) 7:164–168
DOI 10.1007/s008940100018

SHORT COMMUNICATION

Lawrence M. Brown · Robert E. Bruccoleri
Jiri Novotny

Empirical free energy analysis of the engrailed Q50K homeodomain mutant-DNA complex and its mutants

Received: 24 January 2001 / Accepted: 26 February 2001 / Published online: 16 May 2001

Abstract Empirical free energy calculations were performed for the engrailed homeodomain mutant protein in complex with its consensus promoter. The program, CONGEN™, was used to generate the atomic coordinates, which were missing in the original crystal structure of the Q50K mutant protein. To generate the said atomic coordinates more accurately, the initial CONGEN-generated model was subjected to 300 ps of belly dynamics in AMBER™. Complex formation energies were calculated for the 'wild-type' complex as well as ten computer-generated mutants that parallel mutants/substitutions studied by Sauer's group for each of three models. The hydrophobic, electrostatic, and entropic contributions were calculated. The non-linearized finite difference Poisson–Boltzmann equation was solved for the complexes at the ionic strength under which dissociation constant measurements were taken by the aforementioned group. A good overall agreement existed between calculated and experimental $\Delta\Delta G$ estimates. Discrepancies between absolute experimental and calculated values for $\Delta\Delta G$ are hypothesized to be due to the missing conformational entropic contribution of the operator DNA molecule.

Keywords Empirical free energy calculations · Engrailed Q50K–DNA complex · Homeodomain–DNA interactions · Protein–DNA interactions

L.M. Brown (✉) · J. Novotny
Department of Molecular Biology,
Princeton University, Princeton, NJ 08544, USA
e-mail: lbrown@phoenix.princeton.edu
Tel.: +1-609-258-5414, Fax: +1-609-258-4575

J. Novotny
Victor Chang Cardiac Research Institute,
Darlinghurst, NSW 2010, Australia

R.E. Bruccoleri
Congenomics, Inc., Pennington, NJ 08534 USA

Introduction

The structural basis of biological specificity is a central problem in molecular biology, one which is key to some of the most important phenomena that occur in living organisms, including antibody–antigen interaction, enzyme–substrate complex formation, and transcriptional activation. The specificity with which one macromolecule binds to another can be described as the difference in Gibbs free energy of binding (ΔG, in a constant pressure system) between the bound and free states of the given components of the complex. Since the specificity with which a macromolecule binds to another can help explain the etiology of some observed phenotypes in protein substitutions and DNA mutations, biochemists often seek to quantify ΔG by determining the dissociation constant (K_d) of the complex and relating it to Gibbs free energy by:

$$\Delta G = -RT \ln K_d$$

(where R is the ideal gas constant, and T is temperature in degrees Kelvin).

Using atomic coordinates derived from X-ray crystallographic or NMR methods as the sole input, empirical free energy calculations have been used successfully to delineate contributions to specificity in protein–protein and protein–ligand complexes at the atomic level (reviewed in [1, 2]) [3]. However, until very recently, serious technical difficulties precluded the application of this technique to DNA–protein interactions [4]. In this communication, we report the results obtained from detailed calculations using, as our model, the engrailed Gln50Lys homeodomain mutant complexed to its consensus DNA promoter [5]. The dissociation constants for this protein and its operator were published by Ades and Sauer [6], work that sought to determine in a semiquantitative manner the key components of specificity within the homeodomain–DNA complex. Using mutagenesis and gel shift assays to study several protein substitutions and DNA mutants, Ades and Sauer derived K_d values for the protein and its operator, which we have correlated with our estimates.

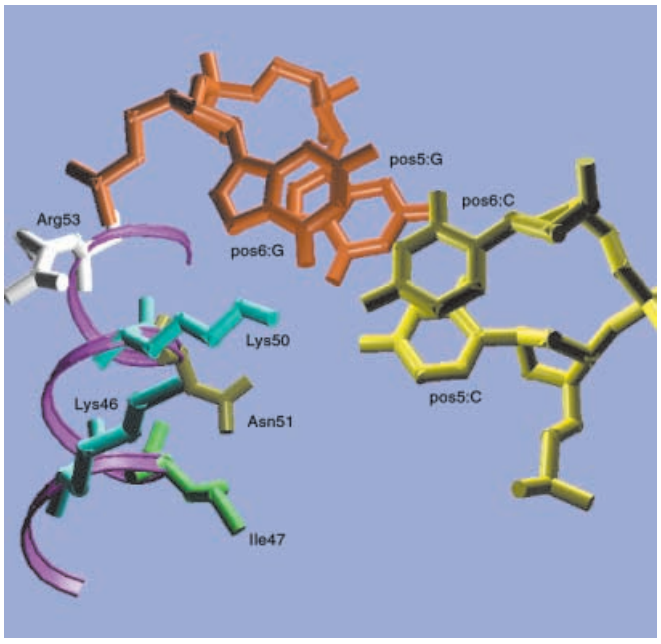

Fig. 1 Raster3D™ rendering of the time-averaged engrailed Q50K model in complex with its consensus operator DNA molecule. View skew to the helical axis of the major groove interactions between Lys 46, Ile 47, Lys50, Asn51, and Arg53 with the basepairs in positions 5 and 6 of the consensus operator core. All residues are rendered in a 'stick' representation of their heteroatoms

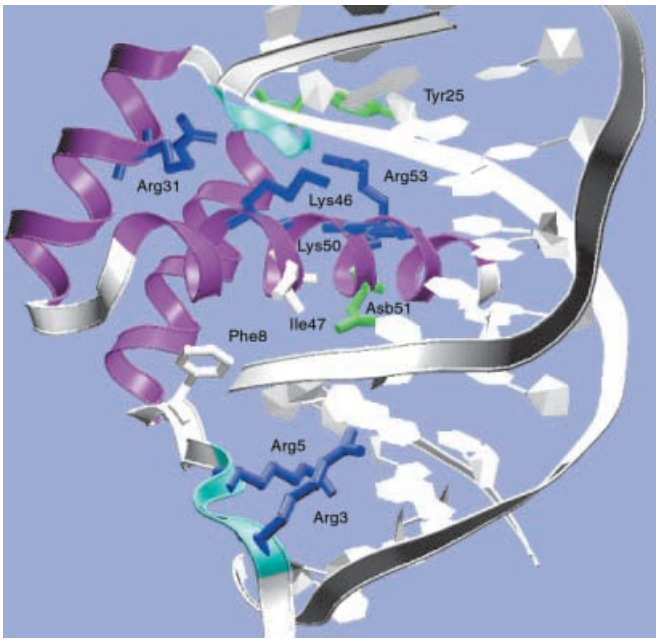

Fig. 2 Raster3D™ rendering of the time-averaged engrailed Q50K model in complex with its consensus operator DNA molecule. Whole complex rendering with all residues predicted to contribute ≥1 kcal mol^{-1} to $\Delta\Delta G_{binding}$ illustrated in a 'stick' rendering of their heteroatoms

The homeodomain family of transcription factors provides a simple model for the observation of protein–DNA interactions. Many homeodomains interact with a consensus DNA sequence bearing the highly conserved core sequence TAATNN. The two basepairs 3′ of the core TAAT, along with the side chain at position 50, appear to be critical determinants of specificity (Figure 1), as has been observed by several groups [7, 8, 9]. Alpha-helix 3 (recognition helix) of the homeodomain and its N-terminal arm form a network of major and minor groove contacts with the operator DNA (Figure 2). Several groups have noted that the N-terminal arm of engrailed is disordered in the free state, becoming ordered only upon DNA binding, leading some groups to ignore those residues for which there is no electron density [10]. Sauer's group showed that an alanine substitution for arginine within this arm resulted in a tenfold decrease in affinity with the consensus operator [6]. Our calculations show that, although the correlations between calculated and experimental values are similar in an N-terminally-truncated and complete model subjected to molecular dynamics, the confidence coefficient of the statistics is significantly different. In this communication, we seek to complement these observations with a more quantitative derivation based upon the available structure of the altered-specificity mutant of the engrailed homeodomain (2hdd, [5]) using empirical free energy calculations and belly dynamics.

Methods

The same formal concepts employed in attributing components of free energy in noncovalent complex formation were used [3, 4, 11]. Gibbs free energy (or ΔG) is considered to be made up of desolvation (hydrophobic effect), electrostatic interactions, and entropic changes (both associational/cratic and conformational entropy as determined by torsional counting for protein sidechains). All electrostatics calculations were conducted in CONGEN™ [12] using the nonlinear finite difference Poisson–Boltzmann scheme with uniform charging, antialiasing, and harmonic smoothing in a continuum dielectric at ionic strength 0.05 [13].

Mutations made in the given structures were generated by editing the PDB file from which they came. Nucleic acid mutations were made by truncating the targeted 'wild-type' residue to its phosphate, renaming the residue according to the target transition, and allowing CONGEN's IC SETUP/IC BILD facility to use its internal coordinate parameter set to build in the atoms of the new molecule, with subsequent rounds of adopted-basis Newton–Raphson (ABNR) energy minimization with the AMBER94 potential. For side-chain substitutions, alanine substitutions involved just truncating the residue to the beta-carbon and allowing it to optimize its position during ABNR minimization. For the Lys50Gln mutant, the lysine's β-, γ-, and δ-carbon positions were all maintained for the glutamine; therefore no radical repositioning was necessary and no conformational sampling was

undertaken beyond the calling of IC BILD and subsequent energy minimization. Inosine parameters were generated in AMBER's RTF by copying guanine parameters, assigning H8 character to the hydrogen at position 2 (replacing the amine) and, maintaining the same bond angle amongst C6, N1, C2, and H2 as existed in the guanine. The RTF and parameters sets were then recompiled within CONGEN and run normally.

To model the N-terminal arm of engrailed, we subjected the constituent residues to 300 ps of simulation in explicit solvent and salt, using the time-averaged coordinates from the production run for subsequent empirical free energy calculations. Simulations were run in AMBER6™ [14] on a 16-processor SGI Origin™ server over a single processor. The complex was neutralized and then minimized by conjugate gradients in four steps: (1) in vacuo, (2) explicit solvent only, (3) ions only, and (4) over the whole system, followed by 300 ps of production. Belly dynamics was invoked, placing harmonic constraints on the DNA molecule and all residues of the protein save the N-terminal-most four residues. Particle-mesh Ewald summation was used for electrostatics estimation during production, and trajectories were monitored for convergence in RMSD every 100 ps.

Results and discussion

Calculations were initially pursued with a truncated model based upon 2hdd, the engrailed Q50K variant hereafter referred to as the 'wild-type', which lacks coordinates for the six N-terminal-most amino acids. Another engrailed structure, 1hdd, which has coordinates for both proline 4 and arginine 3, was used as a template for the manual addition of the α-, β-, and γ-, carbon traces of these two residues to the 2hdd-based model. Correlation between the experimental and predicted $\Delta\Delta G$ values is quite high at r=0.7 and with a confidence coefficient of r^2=0.5. These values improve to 0.8 and 0.6 respectively, when the outlier R3A is ignored (data not shown). The low confidence value indicated the need to model the N-terminal arm of engrailed accurately. This was accomplished as described in the Methods section. After recalculation of the complex's ΔG of binding, the correlation and confidence coefficients improved to r=0.8 and r^2=0.7 (minus the R3A outlier), respectively (Figure 3). We noted that our data can be divided into two smaller subsets: (1) a small cluster comprising the DNA mutations with r=0.8 and (2) a larger data cluster comprising the estimates of ΔG for the amino-acid substitutions and the 'wild-type' with r=0.7 (Figure 3). As observed by Ades and Sauer [6], the experimental ΔG estimates for the DNA mutations cluster in a range of approximately 1.3 kcal mol^{-1} (and around 2 kcal mol^{-1} for our calculated estimates), a range equivalent to the experimental error (an average probable error of 0.64 kcal mol^{-1} or ≈1.3 kcal mol^{-1} for all measurements taken). Experimental estimates of $\Delta G_{binding}$ for the protein substitutions, on the other hand, have a range of around 3.6 kcal mol^{-1}, only slightly higher than the average experimental error (≈2 kcal mol^{-1}), with our range of calculated results ranging around 8.3 kcal mol^{-1}. Such a small range of free energy differences amongst the mutations indicates that transitions alter neither the solvent-accessible nor the electrostatic-potential surface of the DNA radically, changes that our empirical free energy analysis could easily quantify. Instead, it is likely that differences in local stacking interactions that affect high-frequency modes within the operator are responsible for the differences in $\Delta G_{binding}$. Calculations that will allow us to quantify this internal conformational entropy are under way. Be that as it may, we have found that our residue-by-residue contribution estimates are in agreement with experimental and structural observations.

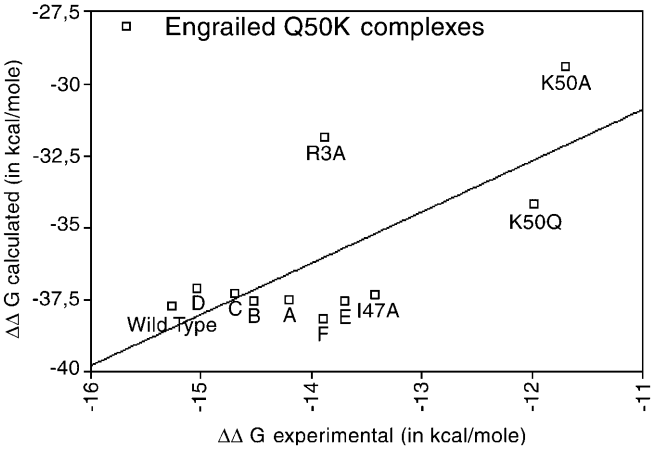

Figure 3 Scatter plot comparing experimental versus calculated $\Delta\Delta G$s for the complete engrailed Q50K model after 300 ps of belly simulation and subsequent time averaging. Labels for transition mutants within the six basepair consensus sequence core are as follows: **A**: position 1 T:A / C:G; **B**: position 1 T:A / C:I; **C**: position 2 A:T / G:C; **D**: position 2 A:T / I:C; **E**: position 5 C:G / T:A; **F**: position 6 C:G / T:A. See text for r and r^2 values

Figure 2 depicts those residues which were determined to contribute ≥1 kcal mol^{-1} of free energy to the binding of the complete engrailed model to the consensus TAATCC-containing operator of the Q50K variant. Minor and major groove interactions appear to contribute equally to affinity, as seen by nearly equal residue contributions in the wild type for arginine 5 and lysine 50 (Table 1), an observation made by others [6, 15]. What is more, our numbers also indicate that Lys50 is the largest single contributor to $\Delta G_{binding}$, in agreement with its perceived role as the key determinant of specificity (Table 1). Lys 50 is proximal to the carbonyl oxygen of guanines at positions 5 and 6 in the core consensus sequence element, with crucial hydrogen bonding predicted to exist between its amine and these moieties (Figure 1). Our calculations indicate that the hydrogen bond to N7 and O6 of the position 5 guanine is the more important with a significant loss in ΔG only with the adenine mutation at that position and not at position 6 (Table 2). Our residue-by-residue $\Delta\Delta G$ estimates also agree with Ades and Sauer's alanine substitution experi-

Table 1 Individual residue ΔG contributions of the wild-type engrailed Q50K homeodomain bound to its consensus operator. Total ΔG is determined by the weighted equation: total ΔG=–ΔG(hb-scaled)+$T\Delta S$(scaled)(cratic and conformational)+ΔG(elec.) (scaled); where ΔG(hb) is the hydrophobic contribution to binding free energy, $T\Delta S$ the entropic, and ΔG(elec.) the electrostatic. ΔG(hb) is a product of molecular surface area, hence these estimates are included. See [3, 4] for further details

Residue	Surface area	ΔG(hb)	$T\Delta S$	ΔG (elec.)	Total ΔG
Sense Strand	**305.10**	**21.40**		**3.10**	**7.90**
α Sense Strand	**355.00**	**24.90**		**–4.90**	**2.60**
Engrailed	**675.90**	**47.30**		**–4.80**	**–10.80**
ARG 3	88.00	6.20	0.19	1.70	–2.80
ARG 5	101.90	7.10	0.33	1.40	–3.30
PHE 8	30.10	2.10	0.00	–0.20	–1.30
TYR 25	22.50	1.60	0.00	–1.10	–1.40
ARG 31	20.40	1.40	0.00	–1.30	–1.40
LYS 46	40.00	2.80	0.00	0.50	–1.50
ILE 47	50.10	3.50	0.09	0.00	–2.00
LYS 50	85.40	6.00	0.00	0.00	–3.60
ASN 51	34.70	2.40	0.00	0.30	–1.30
ARG 53	44.60	3.10	0.03	–3.10	–3.10
Totals	**1336.00**	**93.50**	**1.20**	**–6.60**	**–37.70**
Observed					**–15.27**

ments wherein a loss of affinity of approximately 1.3–1.7 kcal mol^{-1} is seen for R3A (–2 kcal mol^{-1}) and I47A (–1 kcal mol^{-1}), and >2.7 kcal mol^{-1} for K50A (–3 kcal mol^{-1}) [6]. Our numbers point to a less prominent role by Asn 51 in the 'wild type' in binding than others have predicted while Arg53's interaction with the phosphodiester backbone of the DNA molecule appears as the third most crucial contribution in agreement with Frankel's 'indirect readout' hypothesis for engrailed binding specificity (Figure 2) [16]. A favorable increase in $\Delta\Delta G$ of –1.6 kcal mol^{-1} for Asn 51 is observed with the loss of the major groove guanine amine at position 5, corresponding to the inosine mutation (Figure 1). This indicates that Asn 51 would have a more favorable electrostatic contact (rather than only a van der Waals contact) in the TAATAA operator found most favorable for engrailed Q50 binding (Tables 1 and 2). Ades and Sauer [6] observe the diminution of affinity with mutations to positions 1 and 2 of the operator and note that a C:I substitution for A:T reduces binding to a greater extent than T:A even though C:I and T:A have similar functional groups in the minor groove. Our numbers also show a dramatic difference between C:I mutations and C:G or wild-type operator sequences at positions 1 and 2, with the difference in arginines 3 and 5 residue $\Delta\Delta G$ largely coming from an unfavorable increase in $\Delta G_{\text{electrostatic}}$. This implies that the loss in affinity comes with the loss of water-mediated contacts between the arginines and the guanine (or adenine) amines (Table 2). Our statistics also indicate that the proximal amine at position 1 is the more important for arginine 3 association to the operator, since an inosine mutation at that position results in a large (–0.6 kcal mol^{-1}) decrease in $\Delta\Delta G$. (Tables 1 and 2). This is in parallel with a large (–0.5 kcal mol^{-1}) decrease in $\Delta\Delta G$ for arginine 5 with respect to the equivalent guanine to inosine mutation at position 2. This would suggest that these two residues are 'responsible' for read-out from these two positions as might be suggested from their positioning in Figure 2.

The sole outlier for our statistics was the arginine 3 to alanine (R3A) substitution, whose statistics are curiously

Table 2 Total complex and selected residue ΔG contributions of the Ades and Sauer [2, 15] generated substitutions and mutations to the engrailed Q50K homeodomain bound to its operator as determined by empirical free energy calculations. Column headings are same as for Table 1 above

Complex	Residue	ΔG (hb)	$T\Delta S$	ΔG (elec.)	Total ΔG
K50Q					
GLN 50	5	0.11	0.7	–2.5	
Total		*92.9*	*1.41*	*0.6*	*–34.2*
Observed					*–11.99*
I47A					
ALA 47	1.9	0	0	–1.1	
LYS 50	6.1	0	0	–3.7	
Total		*92.6*	*1.15*	*–6.9*	*–37.3*
Observed					*–13.43*
K50A					
ALA 50	0.8	0	0	–0.5	
Total		*82.9*	*1.34*	*–2.2*	*–29.4*
Observed					*–11.7*
R3A					
ALA 3	1.1	0	–0.4	–0.8	
ARG 5	6.4	0.26	0.6	–3.3	
LYS 50	6	0	–0.2	–3.7	
Total		*82.9*	*0.89*	*–6.9*	*–31.9*
Observed					*–13.9*
Position 1	**T:A ->C:G**				
ARG 3	6.2	0.19	1.9	–2.7	
ARG 5	7.1	0.33	1.9	–3.1	
Total		*93.6*	*1.21*	*–6*	*–37.5*
Observed					*–14.2*
	T:A ->C:I				
	ARG 3	6.2	0.19	2.2	–2.6
	ARG 5	7.1	0.37	2.6	–2.7
Total		*93.7*	*1.24*	*–6.1*	*–37.5*
Observed					*–14.52*
Position 2	**A:T ->G:C**				
ARG 3	6.1	0.2	2	–2.6	
ARG 5	7.1	0.35	1.4	–3.3	
Total		*93.3*	*1.23*	*–6*	*–37.3*
Observed					*–14.7*
	A:T ->I:C				
	ARG 3	6.1	0.19	2.9	–2.3
	ARG 5	7.1	0.37	2.2	–2.9
Total		*93.7*	*1.24*	*–5*	*–37.1*
Observed					*–15.04*
Position 5	**C:G ->T:A**				
LYS 50	5.8	0.01	0.7	–3.2	
ASN 51	3.2	0.11	–2.9	–2.9	
Total		*93.6*	*1.29*	*–6.4*	*–37.6*
Observed					*–13.7*
Position 6	**C:G ->T:A**				
LYS 50	6	0	0.2	–3.5	
Total		*94.1*	*1.2*	*–6.9*	*–38.1*
Observed					*–13.9*

skew to the correlation of the rest of the mutants and substitutions. Having calculated the entropic contributions of free and bound sidechains, electrostatic interactions, and the contribution due to the 'hydrophobic effect', we postulated that we had ignored the difference between the wild type and the alanine substitution in the entropic degrees of freedom enjoyed by the N-terminal arm's backbone while calculating the substitution's ΔG. If indeed the N-terminal arm is only 'organized' subsequent to DNA binding, we hypothesized there would be a difference in the 'organization penalty' between the wild type and R3A. For semiquantitative evidence, we used CONGEN to perform backbone conformational searches with modified Go-Scheraga chain closure for both proteins in both the bound and free states with regard to the operator DNA molecule. We estimated backbone entropic contributions of 0.62 and 0.42 for the wild type and R3A substitution, respectively, therefore illustrating an overestimation of the entropic penalty of binding. Properly scaled as a fifth contribution to $\Delta G_{\text{binding}}$, this entropic overestimation could account for the lower $\Delta\Delta G$ value predicted for the R3A substitution. Furthermore, this is in agreement with the observation of Ades and Sauer [6] that, although arginines 3 and 5 do not interact with each other directly, they most likely interact by being crucial to a DNA-binding-dependent positioning of the N-terminal arm.

Conclusion

Our empirical free energy analysis has corroborated the experimental observations made by several groups and has illustrated the importance of the minor-groove interactions of engrailed's N-terminal arm. Our data show discrepancies with observed ΔGs because the entropic contributions of internal DNA molecule motions and homeodomain termini ordering have yet to be calculated. Yet the molecular dynamics simulations and subsequent coordinate and covariance analyses required for the determination of these contributions is now in progress and will be summarized in a forthcoming communication.

References

1. Vajda, S.; Novotny, J.; Sippl, M. *Curr. Opin. Struct. Biol.* **1997**, *7*, 222–8.
2. Ajay; Murcko, M. A. *J. Med. Chem.* **1995**, *38*, 4953–67.
3. Novotny, J.; Bruccoleri, R. E.; Saul, F. A. *Biochemistry* **1989**, *28*, 4735–49.
4. Brown, L.; Bruccoleri, R. E.; Novotny, J. *Pac. Symp. Biocomput.* **1998**, *3*, 339–48.
5. Tucker-Kellogg, L.; Rould, M. A.; Chambers, K. A.; Ades, S. E.; Sauer, R. T.; Pabo, C. O. *Structure* **1997**, *5*, 1047–54.
6. Ades, S. E.; Sauer, R. T. *Biochemistry* **1995**, *34*, 14601–08.
7. Hanes, S. D.; Brent, R. *Cell* **1989**, *57*, 1275–83.
8. Hanes, S. D.; Brent, R. *Science* **1991**, *251*, 426–30.
9. Treisman, J.; Gonczy, P.; Vashishta, M.; Harris, E.; Desplan, C. *Cell* **1989**, 59, 553–62.
10. Kissinger, C. R.; Liu, B. S.; Martin-Blanco, E.; Kornberg, T. B.; Pabo, C. O., *Cell* **1990** *63* 579–90.
11. Novotny, J.; Bruccoleri, R. E.; Davis, M.; Sharp, K. A. *J. Mol. Biol.* **1997**, *268*, 401–11.
12. Bruccoleri, R. E. *Molec. Simulations* **1993**, *10*, 151–174.
13. Bruccoleri, R. E.; Novotny, J.; Davis, M.; Sharp, K. A. *J. Comput. Chem.* **1997**, *18*, 268–76.
14. Case, D. A.; Pearlman, D. A.; Caldwell, J. W.; Cheatham, T. E.; Ross, W. S.; Simmerling, C.; Darden, T.; Merz, K. M.; Stanton, R. V.; Chen, A.; Vincent, J. J.; Crowley, M.; Tsui, V.; Radmer, R.; Duan, Y.; Pitera, J.; Massova, I.; Seibel, G. L.; Singh, U. C.; Weiner, P.; Kollman, P. A.; *AMBER 6.0*, University of California, San Francisco.
15. Ades, S. E.; Sauer, R. T. *Biochemistry* **1994**, *33*, 9187–9194.
16. Frankel, E.; Rould, M. A.; Chambers, K. A.; Pabo, C. O. *J. Mol. Biol.* **1998**, *284*, 351–61.