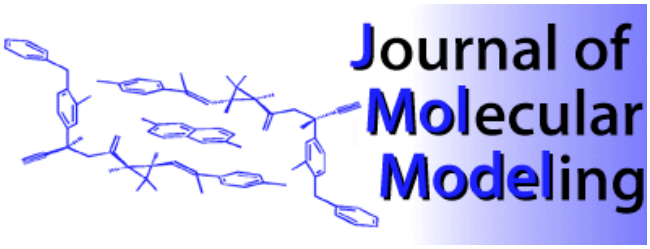
J. Mol. Model. **1999**, *5*, 125 - 133



FULL PAPER

Prediction of Protein Binding to DNA in the Presence of Water-Mediated Hydrogen Bonds

Yuefan Deng[1], James Glimm[1], Yuan Wang[1], Alex Korobka[1], Moshe Eisenberg[2], and Arthur P. Grollman[2]

[1]Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, NY 11794, USA. Tel: 516-632-8614; Fax: 516-632-8490; E-mail: deng@ams.sunysb.edu

[2]Department of Pharmacological Sciences, State University of New York at Stony Brook, Stony Brook, NY 11794, USA

Received: 26 October 1998/ Accepted: 27 May 1999/ Published: 4 August 1999

Abstract We extend our previous analysis of binding specificity of DNA-protein complexes to complexes containing water-mediated bridges. Inclusion of water bridges between phosphate and base, phosphate and sugar, as well as proteins and DNA, improves the prediction of specificity; six data sets studied in this paper yield correct predictions for all base pairs that have two or more hydrogen-bonds. Beside massive computation, our approach relies highly on experimental data. After deriving protein structures from DNA-protein complexes in which coordinates were established by X-ray diffraction techniques, we analysed all possible DNA sequences to which these proteins might bind, ranking them in terms of Lennard-Jones potential for the optimal docking configuration. Our prediction algorithm rests on the following assumptions: (1) specificity comes mainly from direct hydrogen bonding; (2) electrostatic forces stabilise DNA-protein complexes and contribute only weakly to specificity since they occur at the charged phosphate groups; (3) Van der Waals forces and electrostatic interactions between positively charged groups on the protein and phosphates on DNA can be neglected as they contribute primarily to the free energy of stabilisation as opposed to specificity.

Keywords Hydrogen Bonds, Protein-DNA Binding, Specificity, Water

Introduction

DNA is a long, thread-like macromolecule composed of two helical polynucleotide chains. The chains are coiled around a common axis and are antisymmetric with respect to the helical axis. The most frequent conformation is a so-called B-DNA in which the diameter of the double helix is approximately 20 Å.

Correspondence to: Y. Deng

Normally, this deoxyribose is bound to one of four different bases, adenosine, guanosine, cytosine and thymidine. Each helix' backbone is formed by the sequential linking of the phosphate-sugar monomers. The base protrudes into the helix with its plane lying perpendicular to the helical axis. Adjacent bases are separated by approximately 3.4 Å along the helix axis and related by a rotation of nearly 36 degrees. Hence, the helical structure repeats after ten residues on each chain. The two chains are held together by the stacking interactions between adjacent bases and to a lesser extent by the hydrogen-bonds (H-bonds) between opposing bases. Ad-

enosine is always paired with thymidine and guanosine is always paired with cytosine.

Proteins are polymers formed from a long chain of sequentially linked amino acids that are held together by peptide bonds. All proteins are built up from twenty different amino acids, as compared with only four nucleotides in a DNA molecule. Amino acids are capable of forming H-bonds; they play a role in stabilising different conformations and intermolecular binding. This leads to a much greater diversity in the possible conformations of proteins. The three dimensional structure of proteins is much more complex than that of DNA [1] and so far there is no general methodology to predict the shape of a folded protein from the amino acid sequence of its polypeptide chain.

Water profoundly influences the interactions of proteins and DNA [2]. The polarity and hydrogen-bonding (H-bonding) capability of water make it a highly interacting molecule. Water can (1) greatly weaken the electrostatic forces and H-bonds between polar molecules by competing for their attraction; (2) play both roles of an H-bond donor and, by lending electron pairs of the oxygen atom, an acceptor; and (3) form a variety of bridges between molecular donors and acceptors. For protein-DNA complexes, these bridges may be intra- or inter-molecular, e.g., phosphate-water-sugar, phosphate-water-base, base-water-base and various protein-water-DNA water bridges.

In nucleic acids, tertiary structure is the result of an equilibrium between electrostatic forces due to the negatively charged phosphates, stacking interactions between the bases due partially to hydrophobic and dispersion forces, H-bonding interactions between the polar substitutents of the bases, and the conformational energy of the sugar phosphate backbone [3]. The polynucleotide backbone exposes the negatively charged phosphates to dielectric screening by the solvent and promotes the stacked helical arrangement of adjacent bases in its preferred conformations. In this way, a hydrophobic core is produced where H-bonds between bases as well as additional sugar-base, sugar-sugar and H-bonds of protein-DNA are created. Thus, water molecules can increase the stability of helical conformations of nucleic acids by screening the charges of the phosphates and by bonding to the polar exocyclic atoms of the bases [3]. Geiduschek and Gray [4] first suggested that hydration plays a role in the stability of nucleic acid helices and later proposed that the base stacking forces together with H-bonding between complementary bases were responsible for double helical structures in solution. In 1967, Lewin proposed that water bridges contribute greatly to the stability of the DNA double helix in the solution. Saenger and Westhof [5] emphasised the importance of water molecules and water bridges in nucleic acid stability on the basis of crystallographic data. Westhof and Beveridge [6] published simulation results sizing the importance of water molecules and water bridges in nucleic acids [7].

Docking of oligonucleotides to proteins presents a formidable problem because the large number of atoms involved and the flexibility of DNA molecules and the side chains of the protein make the explicit search of binding space very difficult if not impossible even for a single DNA sequence. Therefore, most studies start with rigid-body matching involving a system of filters designed to eliminate unfavourable conformations. Scoring is based on a surface

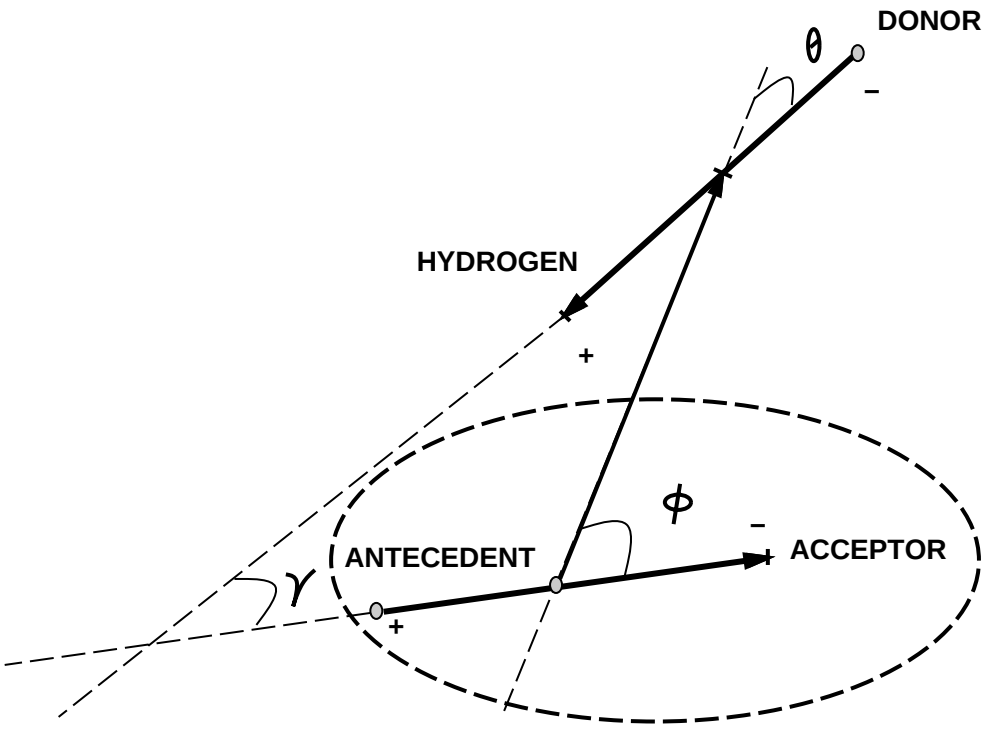


Figure 1 *Angles used in the definition of the LJ potential*

complementarity (for instance, the DOCK algorithm [8]) or an energy function or a combination of both. The energy function evaluates standard potentials such as electrostatic, van der Waals, hydrophobic, and hydrogen bonding. The methods used to explore binding space range from simulated annealing [9], to genetic algorithms [10], Monte Carlo [11, 12] and graph theory approaches [13]. In a recent study, Aloy et. al [14] developed an algorithm capable of performing a global search of binding space using both shape complementarity and electrostatic potential criteria. Advances in computing also allowed structural flexibility to be incorporated in the refinement of the results of rigid-body docking [11, 15]. An extended overview of the recent developments in the simulation of protein-protein and protein-DNA binding is given in [16]. We should mention that most of these studies aim to simulate a docking of several molecules into a complex that matches the data observed in crystallography and NMR experiments. The issue addressed here is the specificity of DNA recognition by the proteins.

In our previous paper [17], we presented an analysis of specificity of H-bonding in protein-DNA complexes without involving water interactions. A Monte Carlo method was employed to dock rigid DNA sequences (deformed in accordance with experimental data) to the established H-bonding sites on the protein molecule. We showed that base pairs with two or more hydrogen bonds are predicted correctly. Here we give some further results showing that water molecules and water bridges play an important role in protein binding to DNA. The water bridge in this paper refers to the bridge between protein and DNA. We start out with the experimentally determined structures of several complexes containing water bridges. Six protein-DNA complexes are analysed in this paper; zinc finger Zif268-DNA complex [18], λ-receptor-DNA complex, MyoD transcription activation domain-DNA complex [19], human-chicken estrogen receptor-DNA complex [20], MAT A1-ALPHA2-DNA ternary complex [21], and TRAMTRACK protein-DNA complex [22] or Zif268-DNA, LAMBDA-DNA, MyoD-DNA, ERDBD-DNA, MAT-DNA, and TRAMTRACK-DNA. The sets of data for all complexes are available in the PDB format from the Protein Data Bank [23]. We utilised the molecular dynamics simulation package XPLOR [24] when experimental data for hydrogen atoms was not available in the X-ray crystal structures. The program Midas Plus [25] provided us with means for visual examination of conformations. The remaining sections of the paper are: §2 data analysis; §3 our prediction of protein bonding to DNA; §4 discussion.

Data Analysis

This section presents the analysis of water mediated bonds in Protein-DNA complexes from crystallographic data. Water is layered between the protein and the DNA, and in some cases, one side of the water molecule is attached to the protein and the other to the DNA. We call such a configuration a water bridge. In addition, there are water clusters that include molecules bound to the DNA and molecules bound to the protein. We also regard these clusters as water bridges. However, these types of clusters are relatively scarce with most clusters attached only to the protein or only to the DNA.

LJ Potential For Hydrogen Bond

We define the length of H-bond to be the distance between the centres of the acceptor atom and the hydrogen connected to the donor. First, we search for the H-bonds to protein, DNA, water bridges, and water clusters using an LJ potential [17] between prospective atoms. We consider the H-bonds only if they are shorter than 3.2Å. The formula for the LJ potential is given below. The minimum of the potential is -0.5 kcal/mol. H-bonds with LJ energy higher of equal to -0.2 kcal/mol are regarded as weak.

For a system with N active H-bonds whose associated donors and acceptors span distances r_i, $i = 1, 2,\ldots, N$, the total system binding energy can be expressed as the sum of a modified pair-wise LJ potential

$$V = \sum_{i=1}^{N} V_i\left(r_i, \gamma_i, \theta_i, \phi_i\right) \tag{1}$$

where

$$V_i\left(r,\gamma,\theta,\phi\right) = \begin{cases} c_0\left[\left(\frac{\sigma}{r}\right)^{\alpha} - \frac{\alpha}{\beta}\left(\frac{\sigma}{r}\right)^{\beta}\right] & \text{if } r < r_0 \\ c_0\left[\left(\frac{\sigma}{r}\right)^{\alpha} - \frac{\alpha}{\beta}\left(\frac{\sigma}{r}\right)^{\beta}\right] \cdot & \\ \cdot\left[\frac{\cos(\gamma) + 3\cos(\theta)\cos(\phi)}{4}\right] & \text{if } r > r_0 \end{cases} \tag{2}$$

A and c_0 is a normalising coefficient. Here $r = r_0$ is the solution to the equation

$$\left(\frac{\sigma}{r}\right)^{\alpha} - \frac{\alpha}{\beta}\left(\frac{\sigma}{r}\right)^{\beta} = 0 \tag{3}$$

σ is a weakly pair-dependent coefficient that determines the location of the minimum of the potential, the angles τ, π and γ are defined in Figure 1 and r is the distance between the donor and the acceptor. The form of the LJ potential $V_i(r, \gamma, \theta, \phi)$ we used above for H-bonds is an approximation. The radial Lennard-Jones potential factor $((\sigma/r)^6–1.5(\sigma/r)^4)$ follows Brunger, which reflects the interaction between the two charged particles (the donor and the acceptor). The angle factor $\cos(\gamma) + 3\cos(\theta)\cos(\phi)$ is derived from the interaction energy between two point dipoles (the donor-hydrogen and

Table 1 *H-bonds on Major Groove*

Protein	Base Pair	Type	H-Bond 1		H-Bond 2		H-Bond 3	
			DNA Site	LJ (kcal/mol)	DNA Site	LJ (kcal/mol)	DNA Site	LJ (kcal/mol)
Zif268	2	Direct	N7	-0.48	O6	-0.38		
	3	Cluster	N7	-0.47				
		Bridge			O6	-0.31		
	4	Direct	N7	-0.47	O6	-0.27		
		Cluster					H41	-0.12
	5	Direct					H61	-0.20
	6	Direct	N7	-0.47				
		Cluster					H41	-0.09
	7	Direct	N7	-0.49	O6	-0.37		
		Bridge					H61	-0.12
	8	Direct	N7	-0.48	O6	-0.36	H41	-0.38
	9	Bridge	N7	-0.43	O6	-0.29		
	10	Direct			O6	-0.38		
	11	Direct					H61	-0.25
MAT	13	Cluster			O4	-0.27		
	15	Cluster	N4	-0.34	O6	-0.43	H42	-0.11
	17	Direct	N7	-0.49	O4	-0.41	H61	-0.22
	18	Direct	N7	-0.45	O6	-0.24		
ERDBD	12	Direct	N7	-0.43				
	13	Bridge	N7	-0.35	O4	-0.32		
	14	Direct	N7	-0.34			H41	0.45
		Bridge			O6	-0.37		
	15	Direct			O6	-0.28		
	16	Bridge	N7	-0.40				
		Cluster			O4	-0.29		
TRAMTRACK	6	Direct	N7	-0.48				
	7	Direct	N7	-0.49	O6	-0.25		
		Bridge					H41	-0.42
	8	Direct	N7	-0.44	O6	-0.35	H41	-0.40
	9	Direct	N7	-0.48			H61	-0.36
	10	Direct			O4	-0.38		
		Cluster	N7	-0.47				
LAMBDA	10	Direct	N7	-0.40	O6	-0.30		
	11	Direct					H41	-0.47
	12	Direct	N7	-0.17	O6	-0.16	H41	-0.18
		Cluster	N7	-0.48				
	13	Direct	N7	-0.37				
	15	Cluster					H41	-0.20
	17	Direct	N7	-0.49			H61	-0.37
		Bridge			O4	-0.29		
MyoD	5	Bridge			O6	-0.34	H41	-0.13
		Cluster	N7	-0.14				
	6	Bridge	N7	-0.45				
	8	Cluster	N7	-0.14	O6	-0.46	H41	-0.40
		Cluster	N7	-0.16				
	9	Cluster	N7	-0.46				
	10	Direct					H41	-0.17
		Bridge	N7	-0.18				
	11	Bridge	N7	0.29	O4	-0.19		
	12	Cluster	N7	-0.23				

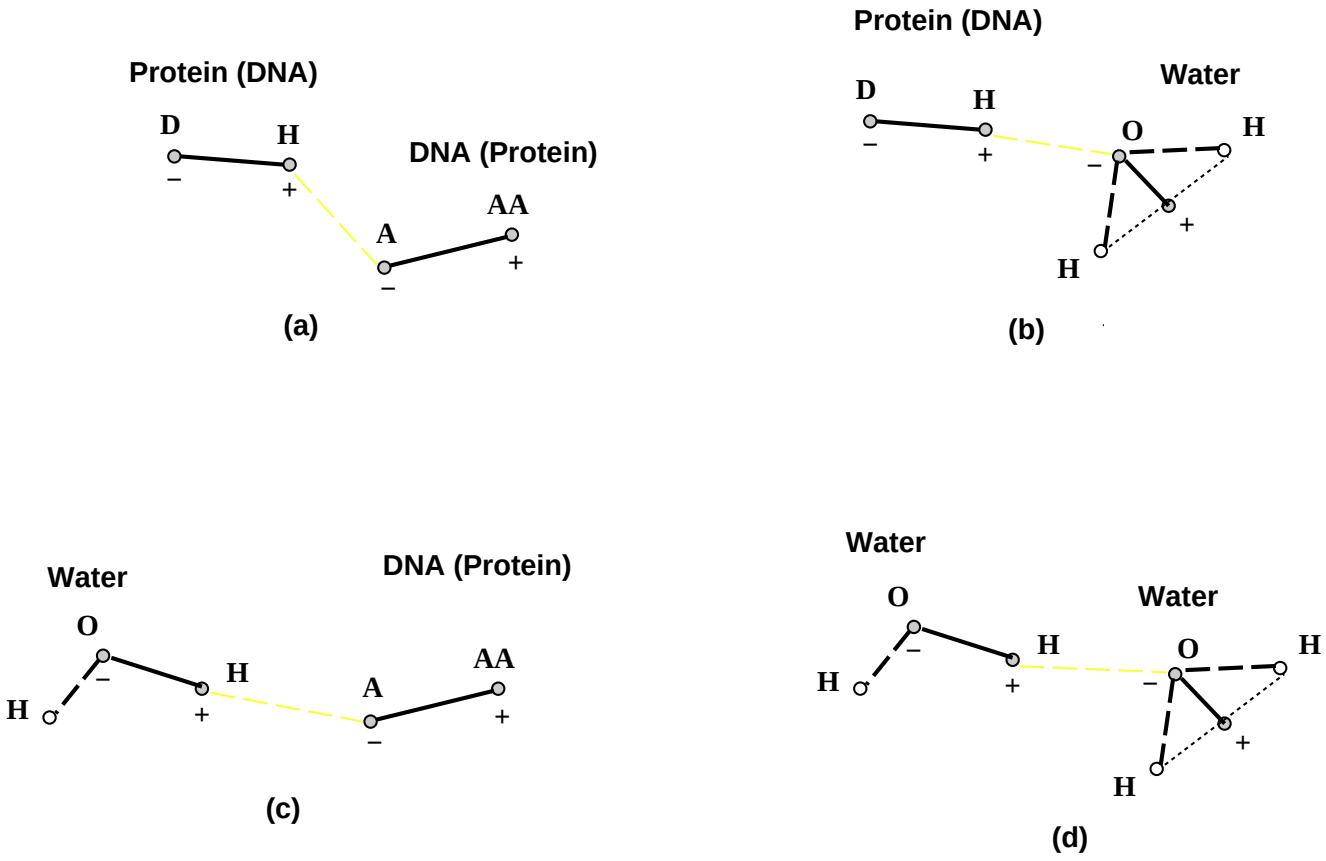


Figure 2 *Four kinds of dipole-dipole interactions in the DNA-Protein-Water system. D denotes Donor, A denotes Acceptor, and AA denotes Acceptor Antecedent*

the antecedent-acceptor) [17]. We emphasise that the energy minimisation is sensitive to the detailed form of the potential, including specifically the angular factors. In this section, we use the LJ potential formula above. We set $c_0 = 1$, $\alpha = 6$, $\beta = 4$.

Interactions in the analysed complexes for which LJ potential is evaluated are shown in Figure 2.

Six Protein-DNA Complexes

In this subsection, we will describe six different protein-DNA complexes: (1) Zinc finger Zif268-DNA complex, (2) MAT A1-ALPHA2-DNA ternary complex, (3) human-chicken estrogen receptor-DNA complex, (4) TRAMTRACK protein-DNA complex, (5) the λ-repressor mutant-DNA complex, and (6) transcription activation MyoD BHLH domain-DNA complex. All hydrogen bonds for each base pair in each complex used in the calculations are shown in Table 1.

Zinc finger Zif268: The zinc finger is a motif that is repeated in tandem to recognise DNA sequences of different lengths. Each finger is based on a similar framework, and each interacts with a small number of base pairs. The strength of the interaction can be varied by changes in the sequence of both the protein and the DNA and by varying the length of the spacing between the fingers. These changes allow a high level of specificity in recognition, and this modular design offers a large number of combinatorial possibilities for specific recognition of DNA [26].

The Zif268-DNA considered here consists of 129 water molecules, the protein with 85 residues and ten base pairs of DNA. The protein has 3 zinc finger motifs within its sequence. Its crystal structure shows that the α-helix with each zinc finger fits directly into the major groove of DNA and that residues from the NH2-terminal portion of each α-helix contact the base pairs in the major groove [18].

The water molecules stay on the concave surfaces of the protein and DNA [5]. Water molecules serve mainly as space fillers but they also form H-bond bridges between protein and DNA. There are 20 H-bonds in this structure (seven water bridges) spanning all 10 base pairs. To minimise the time of computation we chose only base pairs with two or more H-bonds.

The DNA in Zif268-DNA has a B-form DNA structure with a slight distortion [18]. The base pairs 2 — 4 and 8 — 10 have the same sequence (GCG) and they are more distorted than the others.

Table 2 *Prediction results for six complexes*

Protein	Base Pair	H-bond Number	Lab Sequence	Predictions A1	A2	B
Zif268	2	2	G	G		G
	3	2	C	C		C
	4	3	G	G		G
	6	2	G	G		G
	7	3	G	G		G
	8	3	G	G		G
	9	2	C	C		C
MAT	13	1	T	C		G
	15	3	C	C		C
	17	3	T	T		T
	18	2	C	C		C
ERDBD	12	1	G	G		G
	13	2	A	A		A
	14	3	C	C		C
	15	1	C	C		C
	16	2	T	T		T
TRAMTRACK	6	1	A	G		G
	7	3	G	G		G
	8	3	G	G		G
	9	2	A	A		A
	10	2	T	T		T
LAMBDA	10	2	G	G		G
	11	1	C	C		C
	12	4	G	G		G
	13	1	G	G		G
	15	1	G	G		A
	17	3	A	A		A
MyoD	5	3	C	C	C	C
	6	1	A	G	G	A
	8	3	C	C	C	C
	9	1	T	T	C	T
	10	2	G	G	G	G
	11	2	T	T	T	T

MAT A1-ALPHA2-DNA ternary complex: The MAT A1-ALPHA2-DNA ternary complex consists of two protein domains attached to a DNA fragment 19 base pairs long. There are 57 water molecules that form five water bridges between the DNA and the protein.

Human-chicken estrogen receptor-DNA complex: This is a dimer of the complex with each monomer consisting of two identical zinc-finger sequences 74 amino acids long bound to the 17 base pairs long DNA fragment. There are 158 water molecules that form five water bridges.

TRAMTRACK protein-DNA complex: This is a transcription regulation protein. The complex consists of two zinc-finger peptides bound to the DNA. There are two protein domains 66 amino acids long, two DNA duplexes 19 base pairs long, and 57 water molecules grouped into 36 clusters. To reduce the amount of computation we analysed only a monomer structure. This monomer contains two water bridges.

The λ-repressor mutant-DNA complex: This is another transcription regulation protein with three substituted amino acids. The complex was refined as a dimer consisting of two protein domains 92 amino acids long and two DNA molecules 20 base pairs long. There are 92 water molecules grouped in 63 clusters. We chose a monomer that included three water bridges.

Transcription activation MyoD BHLH domain-DNA complex: MyoD proteins are a family of myogenic factors that control the development of skeletal muscle cells. This structure contains the basic helix-loop-helix domain of MyoD complexes with the 14-base pair DNA fragment. The protein component consists of two polypeptides 68 and 62 amino

acids long. Contacts between the protein and the DNA are facilitated by 17 water molecules. There is a single direct H-bond and 13 water bridges spanning seven base pairs. To minimise the time of computation we omitted the twelfth base pair which has a single relatively weak H-bond (-0.233 kcal/mol). We did however, include three other base pairs, with each having a single strong H-bond.

Prediction

The general idea for our computer simulations [17] is that we generate a DNA sequence, fix it, and then carry on translation and rotation of the protein configuration to minimise the total LJ potential of every combination of the protein binding to the DNA. Figure 3 shows the general idea for our program. The parallel version of the program consists of a master process that keeps track of the current sequence and a number of client processes that request new sequences from the master process, compute their binding energies and store results in the log file. This simple parallelisation technique allows us to achieve linear scalability for any number of processing nodes. Upon completion of the program we sort each proposed binding configuration in terms of its LJ potential and check the rank of the experimentally defined structure with this list. The algorithm itself consists of several stages designed to reduce the number of Monte Carlo minimisations; (1) geometric hashing matching pairs of H-bonding sites, using the square-well potential; (2) least-squares minimisation of pairwise distances to rank the prediction given by the above step, which is then used to further filter out insignificant matches; and (3) Monte Carlo searching to stochastically minimise the system's LJ energy. The first two elements determine rigid body motions that attempt to bring serial pairs of atoms, one from the DNA and one from the protein, into coincidence. In this paper we extended this stage to allow oxygen atoms to have up to two H-bonds. As in our previous paper, only unique matches were allowed. We provide results with (Method B) and without (Method A)

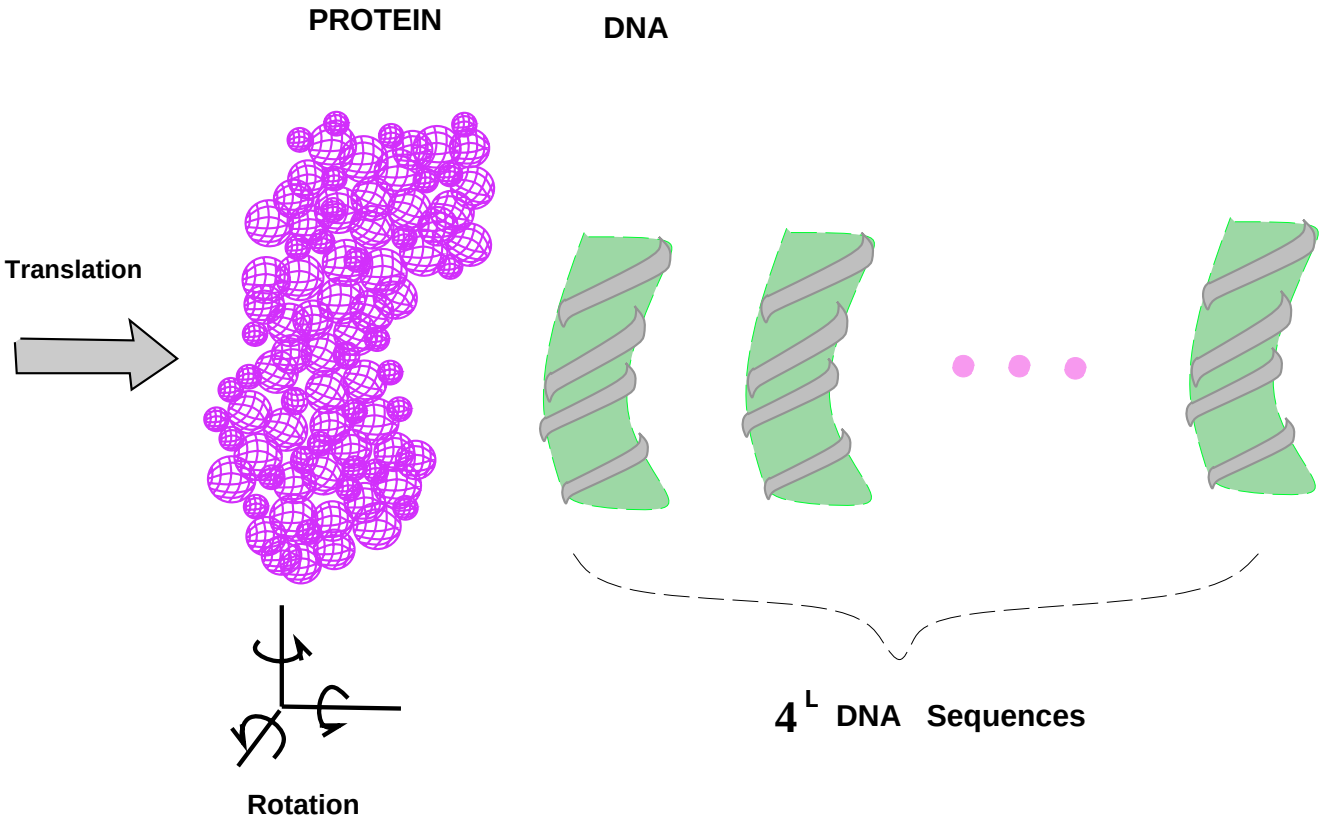


Figure 3 *Schematic diagram of the binding and docking algorithm. The algorithm has three steps: filter, comparison, and optimisation. There are* 4^{L} *different combinations of DNA sequences, where* L *is the number of base pairs. The deformations of B-DNA are induced by the experimental DNA sequence. Given a protein with P H-bond potential bonding sites (acceptor or donor). We try to predict which sequence of DNA to which the protein binds. For Zif268-DNA complex, we consider 20 potential binding sites on the protein and 8 base pairs of DNA. For 8 base pairs with 4 choices each, we have* $4^8 = 65536$ *possible DNA sequences. There are 32 potential H-bond binding sites on each 8-bp DNA sequence. We look for 16 H-bonds match. Thus, the total number of possible matches is* $4^8 \times \binom{32}{16} \times \binom{20}{16} \times 16! \approx 10^{28}$.

this feature. To reconstruct the geometry of H-bonds between pairs of points, we fix one end of the H-bond in the data and compute an "ideal" position for the other by extending, from the first point, in the ideal direction predicted for a H-bond. We then use this extended ideal point as the centre of the square well or quadratic potential when locating the other end of the H-bond in the data. A more detailed description of our algorithms can be obtained from our previous paper [17].

Computations were conducted on a Beowulf-type [27] parallel cluster on 4 to 18 300Mhz Pentium II nodes. Search times varied from several hours to several days depending on the number of base pairs considered and the number of hydrogen bonding sites on the protein.

Conceptually, for the present analysis, we regard the water molecules which form the water bridges to be rigidly connected to the protein. We will present the prediction for the six complexes introduced in the previous section. The results from both, unique matching (Prediction A) and multiple matching (Prediction B), methods are summarised in the Table 2.

Zif268-DNA: There are H-bonds in the major groove on every base pair of the DNA in Zif268-DNA. To reduce the computational time we selected the seven base pairs which contain two or more H-bonds.

The predicted pattern completely matches the experimental one. In this case, Method B achieved a better separation between the best and the next best predicted patterns (-6.326 and -6.215 kcal/mol with B vs. -5.998 and -5.997 kcal/mol with A).

MAT A1-ALPHA2-DNA ternary complex: This structure has a small number of H-bonds all of which where used in the computation. The base pair with a single H-bond was predicted incorrectly. The experimental structure was ranked third by the Method B while Method A rejected it at the geometric filtering stage.

Human-chicken estrogen receptor-DNA complex: In this case a perfect match was achieved despite the presence of two single bond base pairs. Both methods produced the same results.

TRAMTRACK-DNA complex: There is a single mismatch occurring at the base pair with a single H-bond. Both methods produced the same top pattern.

The λ-repressor mutant-DNA complex: This structure features rather uneven distribution of H-bonds and we decided to include base pairs with a single bond into the computation to provide more data on the performance of the algorithm on such base pairs.

The match appeared to be perfect with Method A but we had to increase the tolerance of the first filtering stage to prevent the experimental pattern from being excluded from the calculation. Method B mispredicted a base pair with a single H-bond, the experimental structure was ranked second.

MyoD-DNA complex: This structure also features base pairs with a single H-bond. In addition, there are multiple weak bonds (defined as having a Lennard-Jones potential higher than -0.20).

Method A had two structures tied for the first place. The one that was marginally better has one mismatch, the second one has two. Both mismatches occur at the base pairs with a single H-bond. Method B ranked the experimental structure on top followed by several structures, including those predicted by Method A, with insignificant differences in the total LJ potential.

Discussion

Sequence recognition involves direct H-bonding and van der Waals interactions between protein side chains and edges of base pairs exposed in the grooves of duplex DNA [28]. Electrostatic forces are involved in stabilising DNA-protein complexes but make a lesser contribution to sequence specificity since these occur at the charged phosphate groups. Van der Waals forces and electrostatic interactions between positively charged groups on the protein and phosphates on DNA are excluded from the computations since these forces contribute primarily to the free energy of stabilisation as opposed to sequence specificity. However, these interactions result in the distortion of the DNA geometry and we take them indirectly into account by modelling the distortion by the experimental conformation of DNA. The modelling algorithm was improved from the one used in our previous paper to include intra base pair distortions like "propeller twist". Using protein structures derived from DNA-protein complexes in which coordinates were established by X-ray diffraction techniques, we have analysed all possible DNA sequences to which these proteins might bind, ranking them in terms of Lennard-Jones potential for the optimal docking configuration. Water molecules, which can form hydrogen-bonded bridges between phosphate and base, phosphate and sugar, as well as proteins and DNA, are included in the analysis. Results of our study support the view (review by Berg and von Hippel, 1988 and Steitz, 1990) [29, 30] that H-bonding between side chains of proteins and sites exposed in the major groove of DNA are the critical determinants of recognition for proteins that bind in a sequence-specific manner to DNA.

A number of simplifying assumptions were made in formulating the algorithms used in this study. First, a potential energy function was constructed which involved only H-bonds. Second, where bidentate bonds formed between side chains of proteins and DNA may be relevant to specificity [28], we considered only one to one donor-acceptor type H-bonding with the binding sites on the major groove of the DNA. Third, the protein and the DNA were treated as rigid bodies during the matching process. Finally, the choice of a Lennard-Jones potential for H-bonds was necessarily approximate; moreover, the angular component used for thus equation was based on a dipole-dipole interaction [31], an assign-

ment that proved superior to the angular factor used for molecular mechanics simulations [25].

The algorithm developed to predict the binding specificity of DNA-protein complexes was successfully extended to complexes containing water-mediated bridges. Six data sets yielded correct predictions for all base pairs that had two or more H-bonds (23 out of 23). Although predictions for base pairs with a single H-bond were unstable (5.5 out of 9) they were better than the random result. Evidently, the inclusion of water bridges in the analysis increases the number of base pairs with two or more H-bonds, thus improving the results. A variation of the algorithm that allowed up to two matches for a single oxygen atom proved to be useful, producing improvements in three cases and a mismatch for a base pair with a single H-bond in one case. Further refinements of our project will use molecular dynamics to simulate the protein binding interaction to DNA [32, 33]. The availability of additional experimentally determined data sets should help to further validate the model and simulations.

References

1. Branden, C.; Tooze, J. *Introduction to protein structure*; Garland: New York, 1991.
2. Stryer, L. *Biochemistry*, 3rd ed.; W.H. Freeman: New York, 1988.
3. Vasilescu, D.; Jaz, J.; Packer, L.; Pullman, B. *Water and ions in biomolecular systems*; Plenum: New York, 1990, 11.
4. Geiduschek, E. P.; Gray, I. *J. Amer. Chem. Soc*. **1956**, *78*, 879.
5. Jeffrey, G. A.; Saenger, W. *Hydrogen bonding in biological structures*; Springer: New York, 1991.
6. Franks, F. *Water Science Reviews*; Cambridge University: Cambridge, 1990, 5.
7. Zhang, R. G.; Joachimiak, A.; Lawson, C. L.; Schevitz, R. W.; Otwinowski, Z.; Sigler, P, B. *Nature* **1987**, *327*, 591.
8. Shoichet, B.; Kuntz, I. *J. Mol. Biol.* **1991**, *221*, 327.
9. Stoddard, B.; Koshland, D. *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 1146.
10. Jones, G.; Willett, P.; Glen, R. G.; Leach, A. R.; Taylor, R. *J. Mol. Biol.* **1997**, *267*, 727.
11. Hart, T.; Read, R. *Proteins* **1992**, *13*, 206.
12. Trosset, J. Y.; Scheraga, H. A. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 8011.
13. Webster, D.; Rees, A. *Protein Eng*. **1993**, *65*, 94.
14. Aloy, P.; Moont, G.; Gabb, H. A.; Querol, E.; Aviles, F. X.; Sternberg, M. *Proteins* **1998**, *33*, 535.
15. Knegtel, R.; Boelens, R.; Kaptein, R. *Protein Eng*. **1994**, *7*, 761.
16. Sternberg, M.; Gaab, H.; Jackson, R. *Curr. Opin. Struct. Bio.* **1998**, *8*, 250.
17. Campbell, G.; Deng, Y.; Glimm, J.; Wang, Y.; Yu, Q.; Eisenberg, M.; Grollman, A. *J. Comput. Chem.* **1996**, *17*, 1712.
18. Pavletich, N. P.; Parbo, C. O. *Science* **1991**, *252*, 809.
19. Ma, P. C. M.; Rould, M. A.; Weintraub, H.; Pabo, C. O. *Cell* **1994**, *77*, 451.
20. Schwabe, J. W. R.; Chapman, L.; Finch, J. T.; Rhodes, D. *Cell* **1993**, *75*, 567.
21. Li, T.; Stark, M.; Johnson, A. D.; Wolberger, C. *Science* **1995**, *270*, 262.
22. Fairall, L.; Schwabe, J. W. R.; Chapman, L.; Finch, J. T.; Rhodes, D. *Nature* **1993**, *366*, 483.
23. http://www.rcsb.org/pdb/
24. Brunger, A. T. *X-plor, a system for crystallography and NMR*; Yale University, CT: New Haven, 1992.
25. *USCF Midas Plus*; MIDAS Software Distribution, Computer Graphics Laboratory, School of Pharmacy, University of California: San Francisco, CA 94143-0446, 1988.
26. Klug, A.; Rhodes, D. *Trends Biochemistry Science* **1987**, *12*, 464.
27. http://www.beowulf.org
28. Seeman, N. C.; Rosenberg, J. M.; Rich, A. *Proc. Nat. Acad. Sci. USA* **1976**, *73*, 804.
29. Berg, O. G.; Von Hippel, P. H. *TIBS* **1988**, *13*, 207.
30. Steitz, T. A. *Quarterly Reviews of Biophysics* **1990**, *23*, 205.
31. Jackson, J. D. *Classical electrodynamics*. **1998**, 143.
32. Saito, M. *Molecular Simulation* **1992**, *8*, 321.
33. Saito, M. *J. Chem. Phys.* **1994**, *101*, 4055.