Protein Engineering vol.7 no.6 pp.761–767, 1994

Monte Carlo docking of protein–DNA complexes: incorporation of DNA flexibility and experimental data

Ronald M.A.Knegtel, Rolf Boelens and Robert Kaptein[1]

Bijvoet Center for Biomolecular Research, Utrecht University, Padualaan 8, 3584 CH, Utrecht, The Netherlands

[1]To whom correspondence should be addressed

A Monte Carlo simulation program (MONTY) has been developed to dock proteins onto DNA. Protein and DNA interact via square-well potentials for hydrogen bond and van der Waals interactions. The effect of the inclusion of DNA flexibility and experimentally derived restraints has been tested on members of the helix–turn–helix family of DNA binding proteins. Unwinding and bending the DNA double helix improves the number of correctly retrieved hydrogen bonds in simulations starting from the 434 *cro* protein monomer complexed with a standard B-DNA O_R1 half-site. Agreement with phosphate ethylation interference and mutagenesis data is rewarded with energy bonuses. This protocol was tested on protein–DNA complexes of 434 *cro*, *lac* headpiece and a mutant *lac* headpiece resembling the *gal* repressor headpiece with the recognition helices in correct and reversed orientations in the DNA major groove. The inclusion of experimental data gives an improved convergence of the correctly oriented structures and allows for an easier discrimination between correctly and incorrectly docked complexes.

Key words: DNA flexibility/docking/molecular surface complementarity/protein–DNA interaction/structure prediction

Introduction

The structural basis of the specific recognition of DNA by proteins remains an important, yet unsolved, problem in molecular biology [for reviews see von Hippel and Berg (1989), Harrison and Aggarwal (1990), Steitz (1990) and Pabo and Sauer (1992)]. In recent years high resolution structures of DNA-binding proteins and their complexes with DNA target sites have become available through X-ray crystallography and multidimensional NMR spectroscopy which have provided a view of the recognition process in atomic detail. The conclusion to be drawn from these studies is that there exists no general one-to-one code for protein–DNA recognition. Despite the lack of simple recognition rules combined with the large number of motional degrees of freedom in both proteins and DNA, it seems worthwhile exploring computational docking methods to predict the optimal conformation of protein–DNA complexes.

Until now, most effort in the computational study of molecular association and recognition has focused on the binding of small organic molecules or peptides to protein surfaces [for reviews see Wodak *et al.* (1987) and Cherfils and Janin (1993)]. In these studies the entire protein is usually kept rigid and the ligand is translated and rotated to find the optimal binding site in terms of some simple intermolecular energy function. The results obtained so far are encouraging; in many cases complexes close to the native complex structure are retrieved with good interaction energies. In virtually all studies, however, so-called 'false positives' are found which have equally good or better energies and contact surfaces (even if more sophisticated potential energy functions are applied) which differ markedly from the native complex. The occurrence of such complexes could be interpreted as our current methods being too crude or, alternatively, that these complexes represent other possible modes of binding. The inclusion of external information from biophysical experiments has been shown to improve the chances of finding correct solutions in docking studies (Yue, 1990; Weber *et al.*, 1992).

We have attempted a similar computational approach to the problem of protein--DNA recognition by means of a Monte Carlo docking program called MONTY, which has been described in detail elsewhere (Knegtel *et al.*, 1994). When docking a protein onto DNA an advantage is that usually the binding site is globally known, i.e. the major groove of DNA. An exhaustive search of the entire receptor surface can thus be avoided and the computer time gained can be invested in allowing more atomic detail and flexibility in the simulations when starting from complexes not too remote from the expected native complex. MONTY searches the surface of the major groove of a rigid DNA molecule for the optimal binding site by rotating and translating a protein molecule whose core is kept rigid but surface side chains are freely rotatable. Side-chain flexibility is important because an uncomplexed protein structure is unlikely to have its surface side chains already in a conformation suitable to make specific contacts with DNA bases and backbone. A simple intermolecular square-well potential for van der Waals and hydrogen bond interactions is applied to evaluate the molecular complementarity.

One of the conclusions of our earlier study (Knegtel *et al.*, 1994) was that an improvement in the number of correctly retrieved intermolecular interactions is to be expected by allowing the DNA to adapt its structure to the bound protein and by making use of the wealth of biochemical data available on protein–DNA interactions. In this paper we describe the effects of a simple model for DNA flexibility and the application of experimentally obtained information on the capabilities of MONTY to retrieve protein–DNA complexes resembling the native complex. We have chosen as model systems three members of the helix–turn–helix (HTH) family of DNA binding proteins, the phage 434 *cro* and the bacterial *lac* and *gal* repressor headpiece protein–DNA complexes. The DNA binding properties of 434 *cro* and *lac* repressor headpiece have been well studied by biochemical methods (Wharton *et al.*, 1984; Koudelka *et al.*, 1987, 1988; Lehming *et al.*, 1987a,b, 1988; Sartorius *et al.*, 1989, 1991; Zhang and Gottlieb, 1993), as well as structural methods such as X-ray crystallography (Aggarwal *et al.*, 1988; Mondragón and Harrison, 1991) and NMR spectroscopy (Boelens *et al.*, 1987; Lamerichs *et al.*, 1990; Chuprina *et al.*, 1993). On the *gal* repressor protein–DNA complex only biochemical data are currently available (von Wilcken-Bergmann and Müller-Hill, 1982; Majumdar and Adhya, 1987; Weickert and Adhya, 1992). The *gal* repressor headpiece is, however, homologous to the *lac* repressor headpiece (von Wilcken-Bergmann and Müller-Hill, 1982) and thus an approximate structure of its complex with DNA can be derived from the *lac* headpiece complex by 'mutating'

residues in the *lac* repressor headpiece structure to those of the *gal* repressor. Although the protein structure obtained by this procedure is in fact a *lac* headpiece mutant, we will refer to it as the *gal* repressor headpiece.

An interesting point of discussion in the past has been the orientation of the recognition helix of the *lac* repressor headpiece in the major groove. The helix orientation has now been established by both NMR spectroscopy (Boelens *et al.*, 1987) and mutagenesis experiments (Lehming *et al.*, 1987a, 1988) to be opposite to that of other HTH proteins (among which is 434 *cro*). This provides an interesting test case for MONTY to attempt to reproduce the preference for a particular helix orientation of each protein with or without the use of experimentally obtained information. The *gal* system serves as an example of the ability of MONTY to predict the helix orientation and specific protein−DNA interactions for a protein for which no structural data are available to date.

Materials and methods

Monte Carlo simulations with MONTY

The essence of the Monte Carlo docking method implemented in the Fortran 77 program MONTY (Knegtel *et al.*, 1994) has remained unchanged. After randomization of the dihedral angles of user-selected protein surface side chains an initial rough minimization is performed by randomly rotating these side chains to remove any existing van der Waals overlap. If overlap remains, the protein is translated away from the DNA and the procedure is repeated until all intermolecular collisions have been resolved. Then the protein is randomly rotated and translated by small amounts and the dihedral angles of the rotatable side chains are varied in random steps of $\pm 10°$. During a subsequent minimization phase, only moves lowering the energy are accepted to relax the starting structure. After the minimization phase the Metropolis Monte Carlo (MC) algorithm (Metropolis *et al.*, 1953) is applied to decide on the acceptance or rejection of attempted moves of the protein and DNA. This means that moves improving the interaction energy are accepted, while those decreasing the number of interactions are accepted only with a probability equal to the Boltzmann factor $\exp(-\Delta E/kT)$, where ΔE is the energy difference between the previously accepted and the new attempted move, k is the Boltzmann constant and T is the absolute temperature. In this way the system is capable of overcoming small energy barriers, while moves giving a large increase in energy will be rejected.

Several modifications have been made to the program. One is the inclusion of polar protons as used in the GROMOS simulation package (van Gunsteren and Berendsen, 1987), whereas the older version of MONTY used only heavy atoms. MONTY coordinate input files are now identical to those used by GROMOS. In addition, an angle-dependent hydrogen bond potential has been included which recognizes hydrogen bonds between polar protons and acceptors when their angle is $>110°$ and their distance is between 0.17 and 0.28 nm. These conditions have been chosen deliberately as less stringent than usual to detect hydrogen bonds (for instance, angles $\geq 135°$ and distances ≤ 0.25 nm) to compensate for inaccuracies in intermolecular interactions caused by the use of mutually unadapted structures and the geometrically less restricted nature of charge−charge interactions.

The well-depths of the square-well potential have remained unchanged. In case a hydrogen bond is present, a bonus of −5 kcal/mol is added to the total energy. Intermolecular van der Waals interactions are now counted for all atom types, instead of only the carbon atoms when they are found within a distance range of 0.23−0.38 nm (a lower bound of 0.17 nm is used for atom pairs involving protons), and a bonus of −0.5 kcal/mol is given for such contacts. In the case where the van der Waals contact involves the thymine methyl group and a carbon atom of the protein, it is currently given a bonus of −5 kcal/mol, equal to that for hydrogen bonds. These contacts are weighted heavier in the new version of MONTY because hydrophobic interactions involving the thymine methyl are believed to contribute in a similar way as hydrogen bonds to the specificity of protein−DNA recognition (Pabo and Sauer, 1984; Harrison and Aggarwal, 1990). Van der Waals collisions, established when two atoms are closer than the lower bounds mentioned above, are given a penalty of +1000 kcal/mol. During all simulations a value of kT of 7.7 kcal/mol was applied which corresponds to a chance of accepting the loss of a hydrogen bond of ~50% and for a van der Waals contact ~90%.

Furthermore, an additional optimization phase at the end of the simulation has been added where only the rotatable side chains of surface residues which are not yet involved in intermolecular hydrogen bonds are rotated 5000 times by random amounts to search for additional possible hydrogen bonds.

An important change in MONTY is the inclusion of DNA flexibility. When B-DNA coordinates are read in an ideal helix is fitted through the positions of the C3′ atoms of both strands according to the parametric representation given in Equation 1.

$$\mathbf{r}(t) = \mathbf{r}(0) + R\cdot\cos(t)\cdot\mathbf{a} + R\cdot\sin(t)\cdot\mathbf{b} + k\cdot t\cdot\mathbf{c} \quad (1)$$

The ideal positions of the DNA C3′ atoms are represented by the vector $\mathbf{r}(t)$ with an offset in space of $\mathbf{r}(0)$. The principal axes lying in the plane and along the axis of the helix are denoted by the vectors **a**, **b** and **c**, respectively, t is the helix twist angle measured in cycles, R is the radius and k is the pitch of the helix.

DNA flexibility in MONTY is currently implemented such that first the DNA is unwound or overwound by changing the value of k by a small amount. Then each base pair which is not in van der Waals contact with the protein is shifted with a small step size towards the protein, whereby the outer base pairs are forced to undergo larger shifts than the inner ones. The maximum step size is defined by the user and will be referred to as the DNA bending parameter. Its magnitude is a measure for the allowed bendability of the DNA. This procedure is executed every 200th step of the simulation (which was empirically determined to give good results) to generate several bent DNA structures. The protein is then allowed to adapt itself to the new DNA structure during the following 200 moves.

Another new feature is the possibility of incorporating experimentally obtained information in the simulations by giving bonuses for correctly formed interactions. It was already possible to restrain two atoms by means of a quadratic distance restraining potential which could be used to include nuclear Overhauser effect (NOE)-derived interatomic distances obtained from NMR experiments. A new option is to give extra energy bonuses for biochemically determined interactions where only one of the two partners involved in an interaction is known while the other remains unspecified. An example is the phosphate ethylation interference experiment where the phosphates hindering binding upon ethylation are known but not the amino acids contacting them. Mutagenesis studies from which certain amino acids can be identified as being crucial to complex formation but the contacted base remains unknown can be treated in a similar fashion. In these cases energy bonuses can be given for any

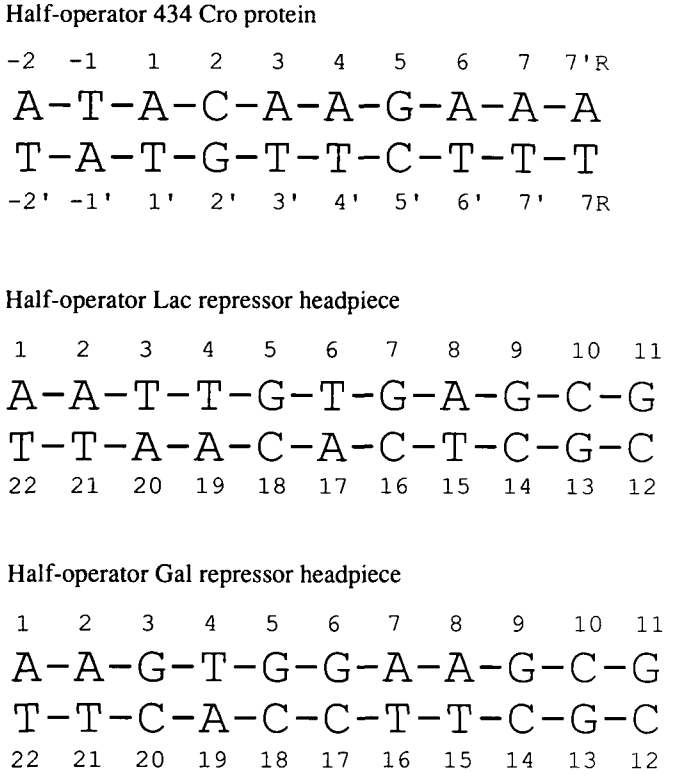

Fig. 1. DNA target sites of the 434 *cro*, *lac* and *gal* repressor proteins. Numbering is as used in the original papers. In all three cases the other protein monomer in the dimer binds to the right of the depicted half-site.

hydrogen bond to a set of specified phosphates or for van der Waals contacts between selected protein residues and any DNA base. When docking unknown complexes such restraints could provide a way to discriminate between correct structures and 'false positives' which have equally good interaction energies but do not agree as well with experimental data.

All manual modeling of starting structures, the mutation of the *lac* headpiece to the *gal* headpiece and the building of B-DNA half-sites, were performed with the Insight II software from Biosym Technologies, San Diego, CA. Standard B-DNA was superimposed on the phosphorus atoms of the crystal DNA. In all cases the protein surface side-chain dihedrals were initially randomized. All simulations and analyses were performed on Silicon Graphics Indigo 2 and Power Challenge computers on which a MONTY simulation of 200 000 MC moves takes ~3–4 h, depending on the CPU involved.

Results and discussion

Simulations with flexible DNA

We have tested the options for DNA flexibility in MONTY on protein–DNA complexes of the left O_R1 half-site of the 434 *cro* protein–DNA complex crystal structure (PDB entry 3CRO; Bernstein *et al.*, 1977; Mondragón and Harrison, 1991) by comparing the percentage of correctly retrieved interactions in simulations starting from the native complex, a complex with the crystal DNA replaced by canonical B-DNA and complexes starting from B-DNA complexes while applying DNA bending. The O_R1 half-site DNA sequence used for 434 *cro* and its numbering are shown in Figure 1. The pitch parameter k in Equation 1 was changed every 200 steps by a random amount $\leq \pm 0.2\%$ which was empirically found to give reasonable

Table I. Effect of DNA bendability on retrieval of protein–DNA interactions in the 434 *cro* complex averaged over 20 structures

Protein–DNA contact		X-ray	0.0	0.2
Lys7–Ade1	Nζ–O1P	25	0	5
Arg10–Thy-1	Nε–O1P	40	0	10
Gln17–Thy-1	H–O1P	90	0	5
Gln17–Ade1	Nε2–O1P	50	35	25
Lys27–Thy4′	H–O1P	95	80	85
Lys27–Thy3′	Nζ–O2P	20	15	5
Gln28–Ade1	Oε1–H62	95	0	25
Gln28–Ade1	Hε21–N7	60	10	15
Gln29–Gua2′	Hε22–O6	55	10	10
Ser30–Thy4′	Hγ1–O1P	80	65	45
Gln32–Cyt2	Oε1–H41	50	10	15
Thr39–Cyt5′	Hγ1–O1P	80	60	60
Arg41–Thy6′	H–O2P	95	40	50
Arg43–Thy6′	H–O3′	90	35	60
Phe44–Cyt5′	H–O2P	100	40	65

Retrieval of hydrogen bonds is expressed as the percentage of structures in which the listed interactions were observed. The DNA bendability factor is either 0 (B-DNA) or 0.2 and 200 000 MC steps were performed in all cases. For hydrogen bonds involving equivalent atoms/protons the interaction with the highest occurrence was counted.

deformations of the DNA. The bending parameter was set to 0.2, which was found to give best results in test simulations using values in the range of 0.00–0.25. For each complex, 20 MONTY simulations were performed using different random number seeds. In all simulations 200 000 attempted MC steps were performed. Prolonged test simulations involving 500 000 Monte Carlo moves with DNA bending showed virtually no improvement.

Table I shows the effect of including DNA flexibility on the percentage of retrieval of correct hydrogen bonds in the 434 *cro* protein–DNA complex. When the simulation is started with the native complex, as determined by X-ray crystallography, most of the specific protein–DNA interactions are retrieved. In this case protein and DNA are structurally optimally adapted to each other. When the crystal DNA is replaced by standard B-DNA several contacts are lost or their retrieval percentage is diminished markedly. The total number of retrieved hydrogen bonds is larger than that reported in our previous study of 434 *cro* complexes with B-DNA, but in that case the protein had to move up or down one base pair to find its binding site and thus had less time to re-establish all contacts, while in this simulation we started from structures close to the native complex.

When the DNA is allowed to adapt itself structurally to the protein, several previously lost contacts are found again or their retrieval percentage is increased markedly. Hydrogen bonds involving long flexible side chains show generally lower retrieval percentages. It is interesting that in the molecular dynamics simulation of the *lac* headpiece–DNA complex in solution (Chuprina *et al.*, 1993) a similar behavior is observed. In that case long side chains are observed to exchange between different hydrogen bonding sites on the DNA surface. Also, in our Monte Carlo simulations residues with long side chains like lysines and arginines are able to reach hydrogen bond acceptors on a DNA surface spanning approximately three subsequent base pairs. These side chains are flexible by their nature and these observations suggest that the recognition of DNA by protein might not be as static as it is often thought. Experimental evidence for such dynamic behavior was recently found in NMR studies

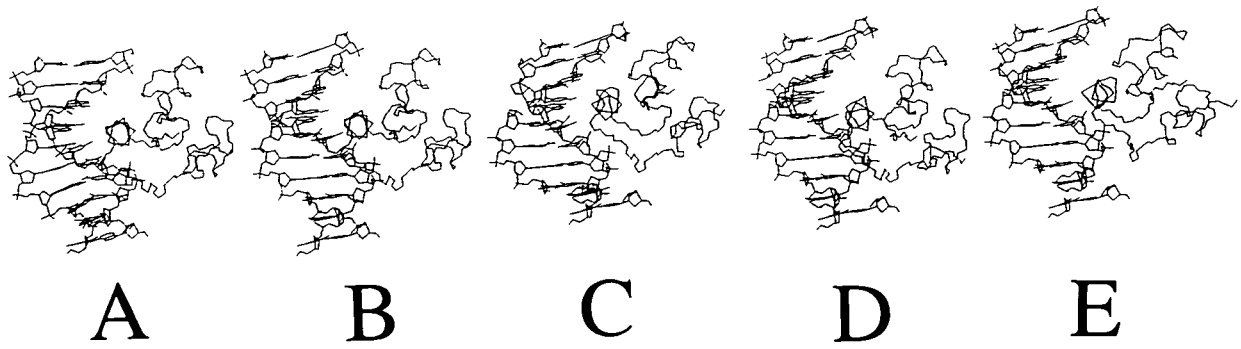

Fig. 2. The effect of DNA bending on protein–DNA complexes of 434 *cro* repressor. (**A**) The native crystal structure; (**B**) the starting structure for MONTY with standard B-DNA; (**C–E**) representative low energy structures generated by MONTY using a DNA bending parameter of 0.2.

of the *Antennapedia* homeodomain complexed with DNA (Qian *et al.*, 1993). In the NMR spectra, Asn51 showed significant line broadening upon binding which was attributed to local exchange of this side chain between different contact sites on the DNA. Mutagenesis data are more difficult to interpret in this respect because binding studies may not reflect, for instance, protein–DNA contacts which exist only a small percentage of the time or exchange between DNA backbone and bases.

Within a family of docked complexes no correlation was found between the energy of the structures and their deviation from the native structure, even after energy minimization with GROMOS (results not shown). This is probably due to the fact that the starting complexes were already quite close to the energy minimum. Current force fields are not able to discriminate between such small differences in molecular complementarity in terms of interaction energy.

Figure 2 shows the X-ray structure of the 434 *cro* protein–DNA complex as a reference, the starting structure with standard B-DNA instead of the crystal DNA and three representative low-energy structures taken from the simulation series performed with a bending factor of 0.2. Bending the DNA gives a better molecular surface complementarity between protein and DNA, especially at the edges of the DNA. This is a logical consequence of the bending algorithm which will only bend base pairs not in contact with the protein and these are bound to be located at the edges of the operator. This also means, however, that best results will be obtained when the simulations are started with the protein already close to the correct binding site because the DNA will start curling around the protein at its starting position. When this is too far removed from the correct binding site the protein might become trapped and will be unable to locate the correct binding site.

Including experimental data in MONTY simulations

The effect of the incorporation of biochemically obtained information was tested on both the 434 *cro* repressor complex crystal structure and the complex of the *lac* repressor headpiece with an 11 bp operator, whose structure was determined on the basis of 2-D NMR data and restrained molecular dynamics using GROMOS (Chuprina *et al.*, 1993). In the case of the *lac* repressor headpiece, the lowest-energy structure of the complex obtained after 40 ps of equilibrium restrained molecular dynamics refinement was used. In addition, we generated a '*gal* headpiece' from the structure of the *lac* repressor headpiece by replacing the unconserved residues with those of the *gal* repressor and built complexes with standard B-DNA of its consensus operator (Weickert and Adhya, 1992). The DNA sequences used for 434 *cro*, *lac* headpiece and *gal* headpiece and their numberings are shown in Figure 1. MONTY simulations were performed for the three protein–DNA complexes with two series of manually built complexes with the recognition helices lying in the major groove in opposite orientations. For all proteins, three hand-built models with the reversed helix orientation were generated initially. The complex which gave the largest interaction energies in test simulations was selected for the simulations presented here.

Table II. Experimentally determined contacts and their MONTY energy bonuses for 434 *cro*, *lac* and *gal* headpiece protein–DNA interactions

434 *cro* protein–DNA complex	Energy bonus (kcal/mol)
Phosphate ethylation interference	
Thy6′, Cyt5′, Thy4′, Thy3′, Thy-1	−50
lac repressor headpiece protein–DNA complex	
Phosphate ethylation interference	
Thy4, Cyt14, Thy15	−50
Gua5	−25
Mutation studies	
Arg22 NH1, NH2 with the bases of:	
Gua5 or Cyt 18	−30
Gln18 Oε1, Nε2 and Tyr17 Cε1, Cε2 with the bases of:	
Thy6, Gua7, Cyt16 or Ade17	−15
gal repressor headpiece protein–DNA complex	
Mutation studies	
Arg22 NH1, NH2 with the bases of:	
Gua5 or Cyt18	−30
Ala18 Cβ and Val17 Cγ1, Cγ2 with the bases of:	
Gua6, Ade7, Thy16 or Cyt17	−15

In the case of phosphate ethylation interference a bonus is given for any hydrogen bond from the protein to the phosphates of the listed bases. In the case of the mutation data a bonus is given for a van der Waals contact between the protein side-chain atom and the mentioned nucleotide base.

Table II lists the nature of the experimental data used in these simulations. The phosphate ethylation interference data give information about which phosphates of the DNA are in contact with the protein. If a hydrogen bond is made to such a phosphate group the energy bonus is added to the total energy. In the case of the *lac* repressor, the phosphate of base 5 was reported to be less strongly protected by the protein (Barkley and Bourgeois, 1978) and was given a smaller bonus. The mutation data used for the *lac* repressor gives information about three amino acids in the recognition helix which are known to recognize specifically three base pairs in the operator (Sartorius *et al.*, 1989). These

Table III. Average MONTY energies of 20 protein–DNA complexes of 434 *cro* with the recognition helix placed in opposite directions in the major groove

Starting structure	Helix orientation	Total energy	Bonus energy
X-ray	correct	−286 ± 30	−
X-ray	reversed	−238 ± 35	−
B-DNA	correct	−231 ± 34	−
B-DNA	reversed	−206 ± 38	−
X-ray	correct	−611 ± 60	−378 ± 34
X-ray	reversed	−501 ± 56	−238 ± 48
B-DNA	correct	−521 ± 47	−348 ± 44
B-DNA	reversed	−395 ± 71	−255 ± 56

All energies are in kcal/mol and are given for simulations with and without the application of phosphate bonus energy terms.

findings were based on double mutations of the *lac* repressor recognition helix and operator to the corresponding amino acids and bases of the *gal* repressor (Sartorius *et al.*, 1989), which allow the mutagenesis information to be used for MONTY simulations of both complexes. Which base of base pair 5 (cf. Figure 1) is contacted by Arg22 cannot be concluded on the basis of the biochemical data, and in the case of Tyr17 and Gln18 it is not known which residue recognizes which base of the two base pairs 6 and 7. Similar bonuses were given in the case of the *gal* repressor, but now for amino acids Val17, Ala18 and the conserved Arg22. MONTY gives a bonus for any van der Waals contact made by terminal side-chain atoms of these residues to the bases involved. The bonuses for contacts made by residues 17 and 18 are smaller because there are more possibilities to establish a contact with one of the four bases than for Arg22 which supposedly contacts only one base pair. Energy bonuses for correctly formed interactions were chosen to be at least comparable with the noise in our simulations which ranged from 20 to 40 kcal/mol. The results of free MONTY simulations and those of simulations applying phosphate ethylation interference or mutation data were compared.

Table III shows the MONTY energies of 434 *cro* protein–DNA complexes with the recognition helix in opposite orientations in the major groove. Results are given for simulations with and without using the phosphate ethylation interference experiments data which identified the contacted phosphate groups in the complex. From Table III it is clear that the correctly oriented complexes already have better intermolecular interaction energies, even without applying additional phosphate contact bonuses. The differences between correctly oriented 434 *cro* and the reversed versions are however still comparable with the standard deviation in the total energy. Replacing the crystal structure DNA with standard B-DNA reduces the energy difference and the total interaction energy further due to the smaller surface complementarity between the two molecules, although the correct orientation is still preferred. When giving −50 kcal/mol bonuses for hydrogen bonds formed with phosphates identified in ethylation interference experiments, the differences become larger and are still observed when B-DNA is used. Apparently 434 *cro* is able to make on average two hydrogen bonds more to the selected phosphates in the correct orientation than in the reversed orientation, judging from the average 100 kcal/mol differences in bonus energies. In particular, the number of hydrogen bonds to phosphates 4′ and 5′ is diminished and in the case of phosphate 3′ reduced to zero in the reversed orientation.

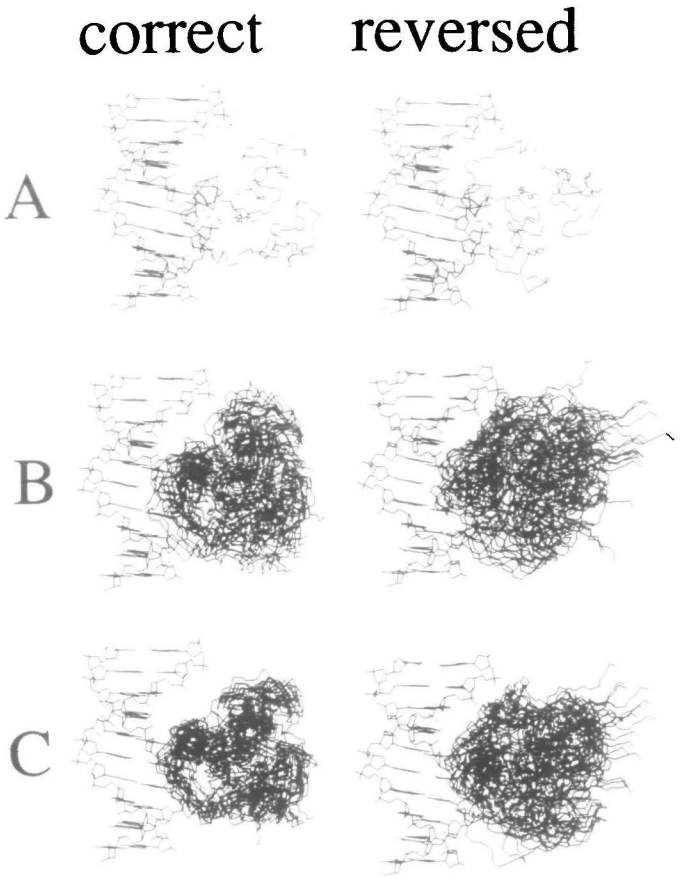

Fig. 3. The use of experimental phosphate ethylation interference data in MONTY simulations of 20 434 *cro* protein–DNA complexes. (**A**) The starting structures with the crystal DNA replaced by B-DNA. The orientation of the recognition helix in the figure on the left is the correct orientation. Protected phosphates are indicated by their van der Waals surfaces. (**B**) Complexes generated by MONTY without the use of experimental data. (**C**) Complexes generated by MONTY with the application of phosphate ethylation interference data.

Figure 3 shows for the simulations performed with B-DNA the resulting 20 structures for the correct and reversed orientation with and without phosphate bonuses applied. Starting with the B-DNA structures, the molecular surface complementarity is low and the spread in structures generated with MONTY is larger than in the simulations starting from the crystal DNA. From Figure 3 it is clear that the use of experimental restraints gives a better convergence resulting in reasonably well defined structures close to the native conformation. In the case of the reversed orientation the application of experimental data does not improve the convergence and a large spread in complex structures remains.

For the *lac* repressor headpiece we have performed similar simulations of complexes with oppositely oriented recognition helices with and without incorporating phosphate ethylation interference and mutation data. The results of these calculations are reported in Table IV. From the simulations without experimental restraints there appears to be already a preference for the correct helix orientation. Again the energy difference becomes smaller when the original DNA is replaced by B-DNA. When phosphate ethylation interference data are used in the simulation the results are less unambiguous. In terms of the total energy, the correct orientation is still preferred over the reversed orientation, although when B-DNA is used the difference is relatively small. In fact, whereas the bonus energy reflects the

Table IV. Average MONTY energies of 20 protein–DNA complexes of *lac* repressor headpiece and an 11 bp operator with the recognition helix placed in opposite orientations in the major groove

Starting structure	Helix orientation	Total energy	Bonus energy
NMR	correct	−290 ± 44	–
NMR	reversed	−229 ± 27	–
B-DNA	correct	−261 ± 35	–
B-DNA	reversed	−232 ± 36	–
Phosphate ethylation interference data			
NMR	correct	−549 ± 39	−316 ± 19
NMR	reversed	−396 ± 47	−239 ± 33
B-DNA	correct	−459 ± 50	−260 ± 41
B-DNA	reversed	−436 ± 42	−289 ± 38
Mutation data			
NMR	correct	−407 ± 65	−159 ± 45
NMR	reversed	−301 ± 59	− 98 ± 61
B-DNA	correct	−334 ± 62	−133 ± 42
B-DNA	reversed	−287 ± 48	−107 ± 44

All energies are in kcal/mol and are given for simulations with and without the application of experimental data bonus energy terms.

preference for the correct orientation when starting from the NMR structure, this is not the case for the simulations involving B-DNA where the bonus energy is slightly larger for the reversed orientation. When complexes with correct and reversed orientations are analyzed in more detail it turns out that particularly the phosphate group of Gua5 takes part in more hydrogen bonds in the reversed orientation than in the correct orientation. The average number of hydrogen bonds to this phosphate group is 1.0 in the correct orientation and 1.8 in the reversed orientation. The average number of hydrogen bonds to the other selected phosphates remains approximately the same in both cases. From the original phosphate ethylation interference data it is known, however, that the phosphate of Gua5 is in fact less contacted by the protein than the other phosphates, which is only in agreement with the situation in the correctly oriented complexes. The ambiguity is probably due to the fact that the phosphates identified by ethylation interference are symmetrically placed around the major groove, i.e. two on either site, which makes a good discrimination on the basis of only these data more difficult.

Figure 4 shows the 20 resulting structures for the simulations of the *lac* headpiece docked with standard B-DNA in both orientations, with and without phosphate ethylation interference bonuses. While the *lac* headpiece structures in the correct orientation converge to a native-like complex when the bonuses are applied (Figure 4C), the reversed orientation in fact diverges while attempting to fulfil as many restraints as possible. This behavior agrees with the increase in bonus energy at the cost of a decrease in total interaction energy as listed in Table IV.

When the mutagenesis data are used, the correct orientation is selected both in terms of the total as well as the bonus energy (cf. Table IV). This is to be expected because the orientation of the recognition helix in the major groove is basically fixed by the two contact points between protein and DNA. In this case the use of experimental data merely enhances and confirms the energy differences found in the free simulations and reduces the sampled conformational space to regions in agreement with experimental data. Recently it was shown by chemical modification of guanine bases in the *lac* operator that both the N^7 nitrogen and the O^6 oxygen atoms of Gua5 are involved in

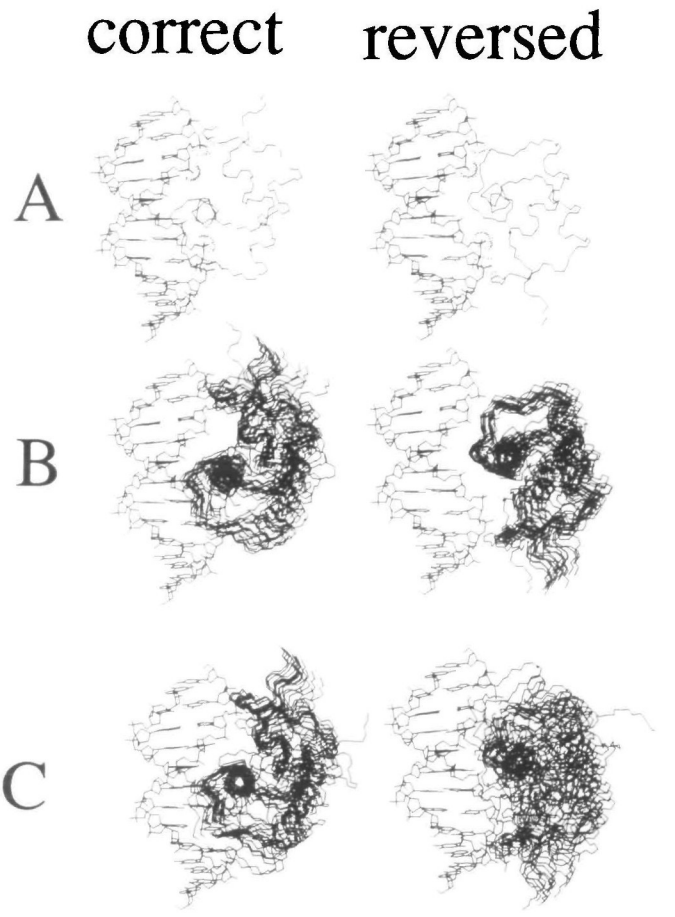

Fig. 4. The use of experimental phosphate ethylation interference data in MONTY simulations of 20 *lac* headpiece protein–DNA complexes. (**A**) The starting structures with the original DNA replaced by B-DNA. The orientation of the recognition helix in the figure on the left is the correct orientation. Protected phosphates are indicated by their van der Waals surfaces. (**B**) Complexes generated by MONTY without the use of experimental data. (**C**) Complexes generated by MONTY with the application of phosphate ethylation interference data.

Table V. Average MONTY energies of 20 protein–DNA complexes of *gal* repressor headpiece derived from the *lac* repressor headpiece structure and an 11 bp operator with the recognition helix placed in opposite orientations in the major groove

Starting structure	Helix orientation	Total energy	Bonus energy
B-DNA	correct	−209 ± 27	–
B-DNA	reversed	−174 ± 27	–
Bend DNA	correct	−241 ± 35	–
Bend DNA	reversed	−200 ± 24	–
Mutation data			
B-DNA	correct	−340 ± 47	−134 ± 36
B-DNA	reversed	−186 ± 38	−12 ± 36
Bend DNA	correct	−375 ± 49	−152 ± 35
Bend DNA	reversed	−224 ± 34	−33 ± 31

For DNA bending a factor of 0.2 was applied. All energies are in kcal/mol and are given for simulations with and without the application of experimental data bonus energy terms.

a hydrogen bond between protein and DNA (Zhang and Gottlieb, 1993). In the refined NMR structure of the complex (Chuprina *et al.*, 1993) a hydrogen bond donated by the side chain of Arg22 to the O^6 atom was observed, but none to the N^7 atom. In the

simulations described here an additional hydrogen bond from Arg22 to the N^7 position is observed with ~40% of the occurrence percentage of the hydrogen bond to O^6, which might explain the importance of this acceptor site for complex formation.

In the case of the *gal* repressor headpiece–DNA complex, no structural data are available. This situation thus represents a test case for MONTY's abilities to predict an as yet unsolved complex. We began our simulations with standard B-DNA complexed with the *gal* headpiece which was built from the *lac* headpiece by mutating the non-conserved residues. Also in this case the correct orientation is preferred energetically, both with and without the use of mutation data bonuses. When in addition DNA bending with a bending factor of 0.2 is applied, the intermolecular and bonus energies for both orientations are increased by similar amounts, leaving the energy differences the same. A similar behavior was observed in the case of *lac* and 434 *cro* (results not shown). This suggests that DNA bending with MONTY increases the surface complementarity and the number of intermolecular interactions in both complexes with different orientations, but does not allow for an easier selection of the correct one.

The correctly oriented complexes of the *gal* headpiece have van der Waals contacts between the side chains of Ala18 and to a lesser degree also Val17 and the methyl group of Thy16, which is the base which differentiates the *gal* operator from the *lac* operator. Also, the methyl group of Thy15 is recognized in the correct orientation by both amino acids, although Val17 appears to make the closest contact. These contacts, which determine the specificity of the *gal* for its target site, are absent, however, when the recognition helix is reversed. In that case, Val17 and Ala18 are close to the charged phosphates of Thy4 and Thy15, while the methyl groups of Thy15 and Thy16 face the charged residues Arg22 and Lys29.

We have described here two enhancements of Monte Carlo protein–DNA docking: the incorporation of DNA flexibility and the use of biochemical information. Adding DNA flexibility to Monte Carlo docking simulations improves the number of correctly retrieved contacts when starting from complexes with the DNA in a standard B-DNA conformation. Because the bending algorithm folds the DNA around the protein, the starting structure should already be close to the correct complex structure to prevent the protein from becoming trapped. This means that the inclusion of DNA flexibility is only warranted in the final phases of a docking study when some reliable starting model, for instance based on homology, biochemically obtained information or docking studies with rigid DNA, is available. In the case of the simulations with different helix orientations, DNA bending as implemented in MONTY improves both types of complexes equally well and does not allow for an easier differentiation between correct and incorrect complexes.

The inclusion of biochemically or biophysically obtained data improves the convergence of the simulations and confirms the results obtained from unrestrained simulations. In our study we have made use of only a limited amount of the available experimental information on 434 *cro* and the *lac* and *gal* headpieces which already gave good results in selecting the correct binding orientation. It is clear, however, that including more of such data as restraints in simulations will raise the chances of obtaining the correct answers with a higher degree of certainty. Although phosphate ethylation interference data in itself may in some cases be insufficient to be decisive, in combination with information — for instance from mutational and genetic studies, NMR spectroscopy, footprinting and photocross-linking — the conformational space to be searched will be reduced drastically and reliable models for protein–DNA complexes can be obtained.

Acknowledgements

The authors would like to thank Dr S.C.Harrison for sending us the coordinates of the refined 434 *cro*– and repressor–DNA complexes and Dr J.A.C.Rullmann for many helpful suggestions and discussions. This research was supported by the Netherlands Organization for Chemical Research (SON) and the Netherlands Organization for Scientific Research (NWO).

References

Aggarwal,A.K., Rodgers,D.W., Drottar,M., Ptashne,M. and Harrison,S.C. (1988) *Science*, **242**, 899–907.

Barkley,M.D. and Bourgeois,S. (1978) In Miller,J.H. and Reznikoff,W.S. (eds), *The Operon*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 177–220.

Bernstein,F.C. *et al.* (1977) *J. Mol. Biol.*, **112**, 535–543.

Boelens,R., Scheek,R.M., van Boom,J.H. and Kaptein,R. (1987) *J. Mol. Biol.*, **193**, 213–216.

Cherfils,J. and Janin,J. (1993) *Curr. Opin. Struct. Biol.*, **3**, 265–269.

Chuprina,V.P., Rullmann,J.A.C., Lamerichs,R.M.J.N., van Boom,J.H., Boelens,R. and Kaptein,R. (1993) *J. Mol. Biol.*, **234**, 446–462.

Harrison,S.C. and Aggarwal,A.K. (1990) *Annu. Rev. Biochem.*, **59**, 933–969.

Knegtel,R.M.A., Rullmann,J.A.C., Boelens,R. and Kaptein,R. (1994) *J. Mol. Biol.*, **235**, 318–324.

Koudelka,G.B., Harrison,S.C. and Ptashne,M. (1987) *Nature*, **326**, 886–888.

Koudelka,G.B., Harbury,P., Harrison,S.C. and Ptashne,M. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 4633–4637.

Lamerichs,R.M.J.N., Boelens,R., van der Marel,G.A., van Boom,J.H. and Kaptein,R. (1990) *Eur. J. Biochem.*, **194**, 629–637.

Lehming,N., Sartorius,J., Niemöller,M., Genenger,G., von Wilcken-Bergmann,B. and Müller-Hill,B. (1987a) *EMBO J.*, **6**, 3145–3153.

Lehming,N., Sartorius,J., Kisters-Woike,B., von Wilcken-Bergmann,B. and Müller-Hill,B. (1987b) *EMBO J.*, **9**, 615–621.

Lehming,N., Sartorius,J., Oehler,S., von Wilcken-Bergmann,B. and Müller-Hill,B. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 7947–7951.

Majumdar,A. and Adhya,S. (1987) *J. Biol. Chem.*, **262**, 13258–13262.

Metropolis,N.A., Rosenbluth,A.W., Rosenbluth,M.N., Teller,A.H. and Teller,E.J. (1953) *J. Chem. Phys.*, **21**, 1087–1092.

Mondragón,A. and Harrison,S.C. (1991) *J. Mol. Biol.*, **219**, 321–334.

Pabo,C.O. and Sauer,R.T. (1984) *Annu. Rev. Biochem.*, **53**, 293–321.

Pabo,C.O. and Sauer,R.T. (1992) *Annu. Rev. Biochem.*, **61**, 1053–1095.

Qian,Y.Q., Otting,G., Billeter,M., Müller,M., Gehring,W. and Wüthrich,K. (1993) *J. Mol. Biol.*, **234**, 1070–1083.

Sartorius,J., Lehming,N., Kisters,B., von Wilcken-Bergmann,B. and Müller-Hill,B. (1989) *EMBO J.*, **8**, 1265–1270.

Sartorius,J., Lehming,N., Kisters-Woike,B., von Wilcken-Bergmann,B. and Müller-Hill,B. (1991) *J. Mol. Biol.*, **218**, 313–321.

Steitz,T.A. (1990) *Q. Rev. Biophys.*, **23**, 205–280.

van Gunsteren,W.F. and Berendsen,H.J.C. (1987) *Groningen Molecular Simulation (GROMOS) Library Manual*. Biomos BV, Groningen, The Netherlands.

von Hippel,P.H. and Berg,O.G. (1989) In Saenger,W. and Heinemann,U. (eds), *Protein–Nucleic Acid Interaction*. Macmillan Press Ltd, Houndmills, UK, pp. 1–18.

von Wilcken-Bergmann,B. and Müller-Hill,B. (1982) *Proc. Natl Acad. Sci. USA*, **79**, 2427–2431.

Weber,D.J., Gittis,A.J., Mullen,G.P., Abeygunawardana,C., Lattman,E.E. and Mildvan,A.S. (1992) *Proteins*, **13**, 275–287.

Weickert,M.J. and Adhya,S. (1992) *J. Biol. Chem.*, **267**, 15869–15874.

Wharton,R.P., Brown,E.L. and Ptashne,M. (1984) *Cell*, **38**, 361–369.

Wodak,S.J., De Crombrugghe,M. and Janin,J. (1987) *Prog. Biophys. Mol. Biol.*, **49**, 29–63.

Yue,S.-Y. (1990) *Protein Engng*, **4**, 177–184.

Zhang,X. and Gottlieb,P.A. (1993) *Biochemistry*, **32**, 11374–11384.

Received December 29, 1993; revised February 21, 1994; accepted March 7, 1994