Protein–DNA binding specificity: a grid-enabled computational approach applied to single and multiple protein assemblies†

Krystyna Zakrzewska,* Benjamin Bouvier, Alexis Michon, Christophe Blanchet and Richard Lavery

Received 2nd June 2009, Accepted 24th September 2009
First published as an Advance Article on the web 7th October 2009
DOI: 10.1039/b910888m

We use a physics-based approach termed ADAPT to analyse the sequence-specific interactions of three proteins which bind to DNA on the side of the minor groove. The analysis is able to estimate the binding energy for all potential sequences, overcoming the combinatorial problem *via* a divide-and-conquer approach which breaks the protein–DNA interface down into a series of overlapping oligomeric fragments. All possible base sequences are studied for each fragment. Energy minimisation with an all-atom representation and a conventional force field allows for conformational adaptation of the DNA and of the protein side chains for each new sequence. As a result, the analysis depends linearly on the length of the binding site and complexes as large as the nucleosome can be treated, although this requires access to grid computing facilities. The results on the three complexes studied are in good agreement with experiment. Although they all involve significant DNA deformation, it is found that this does not necessarily imply that the recognition will be dominated by the sequence-dependent mechanical properties of DNA.

Introduction

Protein–DNA complexes play an undeniably major role in cellular function. Transcription factors control gene expression, enzymatic proteins repair and chemically annotate DNA, while other proteins and protein complexes package DNA, and control complex processes such as replication and recombination. The sequence-specificity of these interactions is very variable, many depend on locating precise binding sites within the genome, while others, such as the nucleosome,[1,2] have more subtle, but still biologically-vital specificities. Determining, and if possible predicting, binding specificity is therefore of considerable importance. Specificity can be determined with a wide variety of assays ranging from gel footprinting, and microcalorimetry to DNA chip experiments.[3–5] A sufficient number of such studies provide enough data to accurately define the size and base content along the binding site.

Analysing such data already poses a number of challenges.[6] Binding site information can be encoded in a so-called position-specific weight matrix,[7] which can then be used to predict a binding score for any given base sequence. This approach assumes that the binding potential of each position along the site is independent of all the others. In the case of conventional lock and key (direct) recognition, composed of specific, pairwise protein–DNA interactions (steric contacts, hydrogen bonds, salt-bridges, *etc.*), this is a reasonable assumption. However, when proteins induce significant DNA deformation, we can expect that these deformations (and, most notably, modified base stacking) will also play a role in (indirect) recognition and lead to coupling between adjacent positions. This can potentially make a simple weight-matrices inaccurate, as has indeed been shown both theoretically[8,9] and experimentally.[5] Taking even nearest-neighbour coupling into account naturally requires more binding data. This is still costly for large numbers of proteins, and simpler corrections to standard weight matrices have been proposed.[10]

Beyond simply predicting the specific binding sites of a given protein it would be very useful to be able to predict how even punctual mutations in the protein would influence binding specificity. This would enable data on a single transcription factor to be adapted to many, often high-homologous, family members. This seems to imply a structural, rather than a purely sequence-based approach in order to understand the molecular mechanisms underlying binding specificity.

A last difficulty occurs in the case of multi-protein complexes where the size of the binding site on DNA makes conventional assays impossible. This is notably the case for the nucleosome where 147 base pairs are wrapped around a protein core formed from eight histones. The combinatorial sequence possibilities for this length of DNA ($4^{147} = 3.2 \times 10^{88}$) are so vast that even SELEX approaches can only be applied to limited segments of the binding site. Recent experiments have therefore concentrated on defining the preferred positions of nucleosomes on genomic DNA.[12] While this data is very valuable there are still significant disagreements between the results of different studies.[11]

In an attempt to better understand protein–DNA binding specificity and to contribute to overcoming the difficulties discussed above, a number of groups are developing theoretical approaches to predicting binding. These include

Institut de Biologie et Chimie des Protéines, CNRS UMR 5086/Université de Lyon, 7 passage du Vercors, 69367 Lyon, France. E-mail: k.zakrzewska@ibcp.fr
† Electronic supplementary information (ESI) available: Tables S1–2; Fig. S1–2; links to view 3D visualisations of protein structures using FirstGlance. See DOI: 10.1039/b910888m

from knowledge-based methods, which parameterize effective potentials on the basis of experimental data (eventually combined with a coarse-grain representation of the protein–DNA complex),[12–16] to all-atom approaches which use conventional force fields, eventually modified to improve the fit with experimental data.[17–21] All these methods have to deal with the challenge of treating a large number of potential binding sequences and consequently must be able to make individual estimates very rapidly. Even if the aim is not to compare all possible binding sequences, scanning a eukaryotic chromosome will typically require hundreds of thousands of binding predictions. They also ideally have to allow for conformational changes at the protein–DNA interface and for the indirect effects involving the change of DNA conformation between its free and bound states.[22,23] In the case of the nucleosome, DNA deformation is believed to play a major role and the predictions of nucleosome positions made so far have been based on simple models of DNA deformability.[11,24]

Our studies in this field began with a method termed ADAPT[25,26] which made binding energy calculations rapid by pre-calculating and storing energy terms for all possible bases at each position within the binding site. These terms could subsequently be added to obtain the binding energy for any chosen sequence. This method was used to analyse direct *versus* indirect recognition in a variety of protein complexes[27] and also to demonstrate that non-negligible coupling between adjacent binding site positions often occurs when protein binding leads to DNA deformation.[8] The main limitation of our early studies was that computational speed was achieved at the cost of not allowing conformational adaptation. Binding energy estimates were all based on effectively threading a sequence into a single, average structure of the protein–DNA complex (obtained by energy minimisation using an average sequence with equal contributions of all four base pairs at every position). This constraint has been avoided in other approaches with good results, but at the cost of limiting the number of sequences which can be compared.[19–21] Recently, we have been able to extend our approach to allow for a reasonable degree of sequence-dependent flexibility without sacrificing the possibility of comparing all possible binding sequences.[28] The new version of ADAPT uses a divide-and-conquer approach to linearize the sequence combinatorial problem and to allow conformational optimisation of the protein–DNA interface as a function of the varying base sequence.

We now apply this method to a group of three protein–DNA complexes which share the relatively uncommon feature of having the protein–DNA interactions concentrated on the side of the minor groove of the double helix.[29] They also all involve significant DNA deformation. The complexes in question are firstly two important transcription factors, the TATA-box binding protein (TBP) which is a key element in the first stages of forming the transcription complex and the sex-determining protein (SRY) which belongs to the HMG box group of proteins, is critical in sex determination, and is involved in a number for gender-related pathologies.[30] Both of these proteins induce DNA bending away from the minor groove binding site and feature the partial intercalation of protein side chains between consecutive base pairs. The SRY complex also exhibits a number of cationic side protein chains placed within the DNA minor groove. The third complex we consider is the nucleosome, which poses a particular challenge due to its size and its low level of binding specificity. The main protein–DNA interactions in this case again involve cationic protein side chains penetrating the minor groove side of the double helix, but no partial intercalation. DNA bending is again a major feature of binding, but in contrast to SRY and TBP, bending now occurs towards the protein, rather than away from it.

We will show that the new version of ADAPT can successfully predict binding specificities for these three systems and can be used to analyse the degree of specificity of binding, the extent to which indirect recognition *via* DNA deformation is important. While the method we have developed makes studies of typical single protein complexes feasible with the in-house computer resources available to many molecular modelling groups, multi-protein complexes such as the nucleosome require a much larger volume of computations, but can be dealt with effectively using grid computing resources.

Methodology

Starting conformations for all the protein–DNA complexes treated here were taken from crystallographic or NMR results available in the PDB database:[31] crystallographic entry 1CDW for the human TATA-box binding protein,[32] NMR entry 1J46 for the human male sex-determining protein SRY,[33] and crystallographic entry 1KX5 for the xenopus/human nucleosome.†[34,35]

These conformations were used to build the corresponding complexes within the program JUMNA.[36] JUMNA is an all-atom modelling program applicable to nucleic acids and to nucleic acid complexes with ligands, which can be either small molecules or macromolecules such as proteins. It represents flexibility using a combination of helical coordinates, which control the position of each nucleotide or ligand in space, and internal coordinates, which allow for single bond rotation and valence angle deformation. Valence angle deformations are limited to flexible rings and to the principal angles of the phosphodiester backbone. All bond lengths are fixed, with the exception of those used to close the sugar rings (C4′–04′) and the junctions between successive nucleotides along each phosphodiester strand (O5′–C5′). This approach reduces the number of variables by roughly an order of magnitude compared to standard Cartesian coordinate approaches and creates an energy landscape which is not only of lower dimensionality, but also much smoother and better adapted to energy minimisation.[36,37]

All energy calculations used the AMBER parm99 force field,[38] with the parmbsc0 modifications to correct the behaviour of the α (P–O5′) and γ (C5′–C4′) torsions.[39] Solvent electrostatic damping was implemented with a sigmoidal distance-dependent dielectric function and counterion effects were taken into account by reducing the net phosphate charges to -0.5 e. All of the complexes studied were initially energy minimised using a conjugate gradient algorithm. In the case of the crystallographic coordinates 1CDW and 1KX5, an initial

minimisation was carried out to optimise the orientation of added hydrogen atoms, fixing the position of all the heavy atoms. Further minimisations allowed full conformational freedom for DNA and for protein side chains, but constrained the protein polypeptide backbone.

In order to test the sensitivity of our binding analysis to the initial conformation of the TBP and SRY complexes, we also performed an unrestrained 10 ns molecular dynamics simulation using NAMD,[40] starting from the experimental coordinates, and employing the parmbsc0 force field[39] in the presence of roughly 12 000 hydrating TIP3P water molecules[41] and neutralising Na^+ ions. Further details of the simulation protocol are given in an earlier publication.[42] Four conformations of each complex were extracted from the molecular dynamics trajectories at equally-spaced intervals and analysed in the same way as the experimental complexes.

In order to be able to study sequence-specificity, it is necessary to compare the formation energies of complexes with the full range of possible DNA sequences. This is a significant combinatorial problem since there are 4^N possible sequences for N base pairs. In the cases treated here we took into account base pairs belonging to binding sites with N equal to 7 for SRY, 8 for TBP and 143 for the nucleosome. This implies calculating energies 16 384, 65 536 and 1.24×10^{86} sequences respectively, which is clearly not feasible. In order to overcome this problem we have adopted a divide-and-conquer approach which breaks each binding site down into a series of overlapping oligonucleotides. We then calculate the energies of all possible sequences for each oligonucleotide and their interactions with neighbouring portions of the bound protein. For each base sequence, the oligonucleotide and the neighbouring protein side chains are energy-optimised, allowing DNA and the protein–DNA interface to adapt to the change of sequence. The protein backbone and the DNA not belonging to the oligonucleotide being optimised are, however, fixed in space. At the end of the optimisation we store all internal and interaction energy components in a matrix from which it is possible to rapidly reconstruct the internal energy of the protein, the complete DNA fragment and the total protein–DNA interaction energy. A similar calculation is performed on an isolated DNA fragment. By subtracting the energy of the isolated DNA, we can then estimate the binding energy for any given sequence. Earlier studies have shown that using overlapping pentanucleotide fragments is a good compromise between precision and computational speed, enabling unfragmented binding energies to be reproduced to within roughly 1–2 kcal mol^{-1}.[28] In this case, since each pentanucleotide has $4^5 = 1024$ possible sequences, a binding site of N base pairs requires (N-4) $\times$ 1024 energy optimisations. If the isolated DNA is treated in a similar way, the same number of optimisations are again required. Using this approach the combinatorial problem is reduced from exponential to linear as a function of the length of the binding site and the numbers given above for SRY, TBP and the nucleosome are reduced to 3072, 4096 and 142 336 respectively. This makes it feasible to treat single protein complexes on a simple computer cluster, but still poses a problem for the multi-protein complexes such as the nucleosome. However, since each energy optimisation is an independent process, this problem can be solved using grid computing resources[43] as described in the following section. Finally, all conformational analyses were carried out using Curves+ (http:gbio-pbil.ibcp.fr/Curves_plus)[44] and all molecular graphics were prepared using VMD.[45]

Results and discussion

We will now analyse the nature of the binding interface and the corresponding binding specificity of the three protein–DNA complexes we have considered. In the case of the nucleosome, we also test the binding predictions by scanning DNA sequences for which the nucleosome locations have been determined experimentally.

TBP

We begin with the TATA-box binding protein (TBP). As shown in Fig. 1, TBP causes significant deformation of DNA upon binding. The minor groove is opened by a combination of untwisting and base pair unstacking. In the crystal structure of the TBP complex,[32] this unstacking is particularly strong at the positions where pairs of phenylalanine residues are partially intercalated at the first TpA step of the TATA-box (at the bottom of Fig. 1 and at the ApG step six base pairs away in the $5' \rightarrow 3'$ direction. These partial intercalation sites have roll/rise values of 53°/4.6 Å and 40°/4.4 Å, respectively. The minor groove opening is presumably stabilised by a number of salt bridges between Arg or Lys residues and the DNA phosphate groups located between the partial intercalation sites. There are no charged residues positioned within the minor groove.

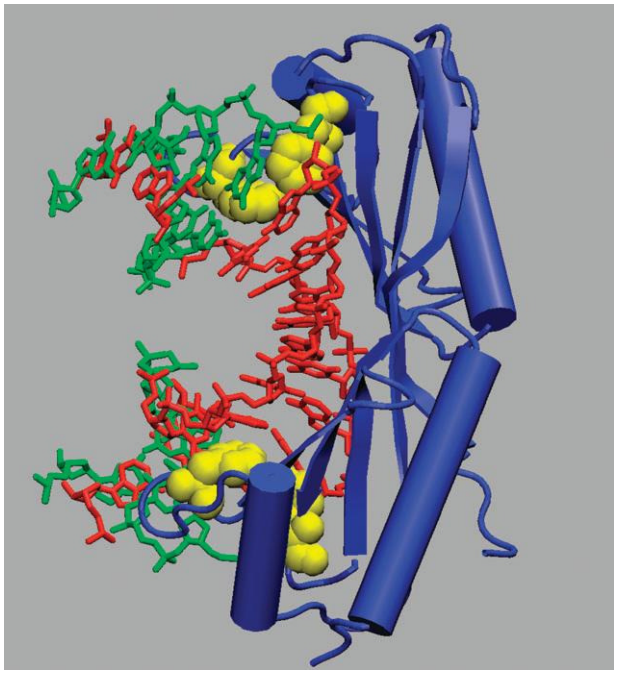

Fig. 1 TBP–DNA complex[32] showing the intercalating residues phenylalanine 193 and 210 (top) and 284 and 301 (bottom) as yellow van der Waals spheres. The protein backbone is shown in cartoon mode in blue and the DNA is coloured red for AT pairs and green for GC pairs.

Fig. 2 shows the sequence-specificity obtained with ADAPT. We have analysed an eight base pair segment at the TBP binding site, beginning and ending with the bases flanking the partial intercalation sites discussed above. In computational terms, this segment corresponds to four overlapping pentanucleotides, each with 1024 possible sequences, that is a total 4096 energy optimisations, compared to 65 536 which would have been necessary without using the divide-and-conquer algorithm. The resulting sequence logo is shown in Fig. 2A. This logo is based on the total binding energy which comprises the protein–DNA interaction, DNA deformation and protein deformation. Comparison with the experimental consensus for TBP (TATAWAWR, where W implies A or T and R implies A or G, see http://www.epd.isb-sib.ch/promoter_elements/)[46] shows very good agreement, although we note that the calculated selectivity for adenine in the sixth position is very low. The logo however does not fully represent the sequence information for the positions 4–6 since there are non-negligible correlations between neighbouring base pairs, as shown by the blue bars in the upper logo in Fig. 2.

The origin of specific binding in the case of TBP is very clear, as shown in the lower logo in Fig. 2, which presents the sequence logo derived from the DNA deformation energy alone. The result is almost indistinguishable from the total binding energy, showing that TBP finds its target site almost exclusively *via* indirect recognition, that is, the sequence-dependence of DNA mechanics. The positions of strongest sequence selectivity can nominally be correlated with strong DNA deformations including global bending and unwinding of the TBP binding site, strong inter-base pair roll (41° and 50°) and increased rise (4.6 Å and 4.4 Å) where the phenylalanine side chains intercalate (the first TpA of the TATA motif and following A in position 7 of the logo shown in Fig. 2). However, as we shall see for SRY predicting selectivity from geometry alone is not possible.

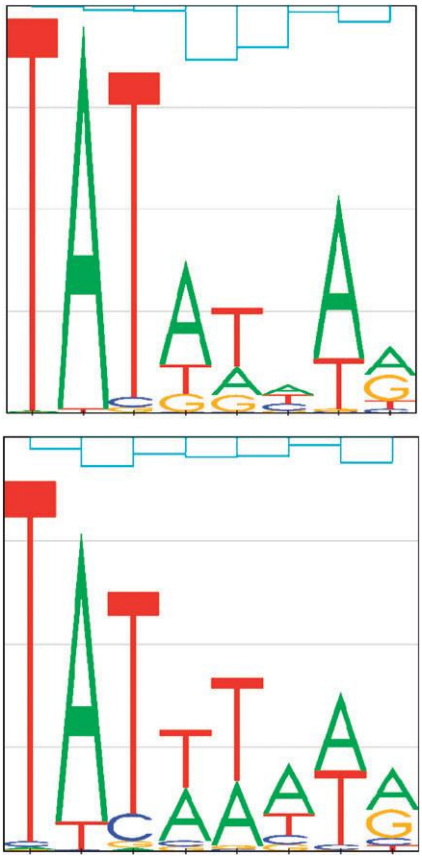

Fig. 2 TBP sequence logos: (above) total binding energy; (below) contribution from DNA deformation. The bars above each inter-base pair step in the binding site indicate the degree of nearest-neighbour coupling.

SRY

We now consider the sex-determining region Y protein (SRY). Fig. 3 shows that SRY again binds in and opens the minor groove of DNA, intercalating an isoleucine side chain at an ApA step towards one end of the binding site. In the NMR-derived experimental structure,[33] the intercalation site produces less perturbation than those seen in the TBP complex, with roll and rise values of 22° and 3.2 Å respectively. In contrast to the TBP–DNA interface, there are three cationic arginine residues bound in the minor groove.

Starting from the experimental structure, the sequence-dependence of the SRY complex was obtained for the seven base pairs constituting the main part of the binding site using three overlapping pentanucleotides and a total 3072 energy optimisations. The results are shown in Fig. 4 (top left). Compared to the experimental consensus, WAACAAW (based on 29 binding sequences),[47] the agreement is rather poor, although it should be noted that the binding site of the

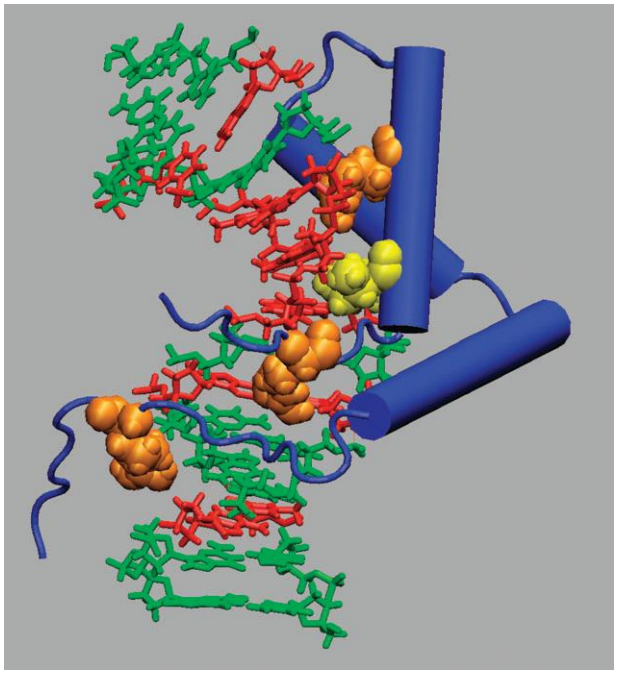

Fig. 3 SRY–DNA complex[33] showing the intercalating residue isoleucine 13 as yellow van der Waals spheres and the arginines 78, 7 and 20 (from bottom to top) in the minor groove in orange CPK. The protein backbone is shown in cartoon mode in blue and the DNA is coloured red for AT pairs and green for GC pairs.

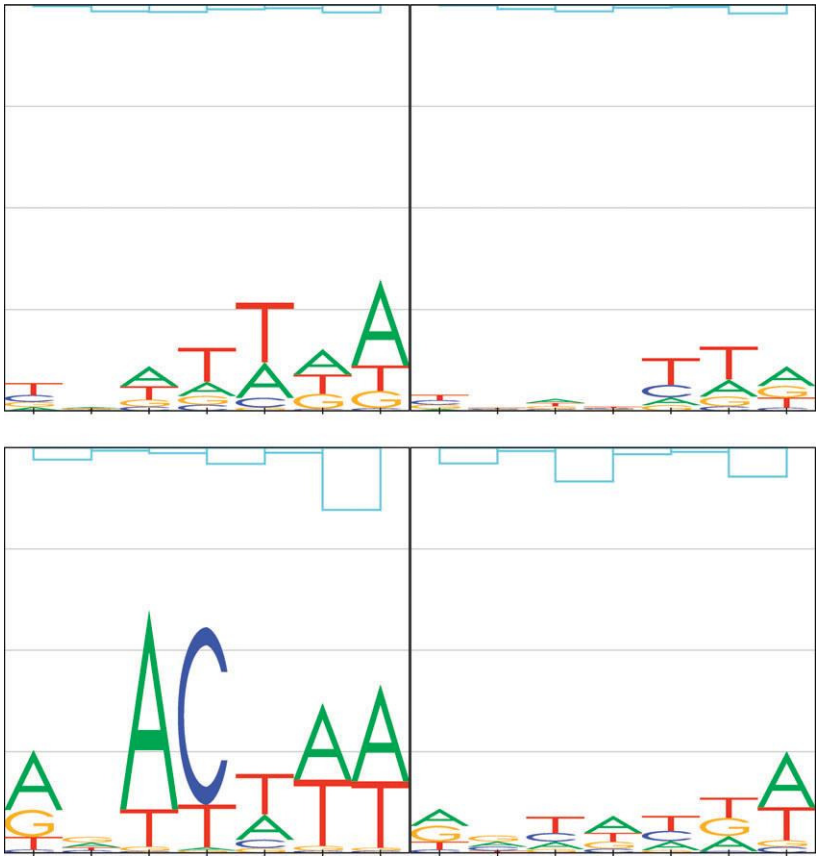

Fig. 4 SRY sequence logos. Above: using the NMR-derived experimental structure; (left) total binding energy; (right) contribution from DNA deformation. Below: using an energy-minimised structure following equilibration with molecular dynamics in explicit solvent; (left) total binding energy; (right) contribution from DNA deformation.

oligonucleotide used for the structural studies GCACAAA, also does not respect the consensus for two 5′-base pairs.

Since all our earlier studies of selectivity were based on crystallographic starting conformations, we decided to test the impact of relaxing the experimental structure using an unrestrained molecular dynamics simulation with an explicit solvent and counterion representation. After 10 ns of simulation we extracted and energy minimised a conformation which was then used to recalculate the sequence specificity. The results shown in the lower, left logo of Fig. 4 correspond much better to the consensus, notably showing a clear ApC preference at positions 3 and 4 and an enhanced preference for adenine at positions 1, 6 and 7. Position 2 is still virtually non-selective, which may well be correct given the difference between the consensus and the binding sequence used in the NMR studies.

If we look at the difference in structure between the NMR and MD-relaxed structures of the SRY complex, there are small changes which can nevertheless explain the enhanced selectivity seen with the latter structure. As shown in Fig. 5, there is little change in the DNA base pairs bound to SRY, although the overall bending of the DNA fragment increases by 30°. There is nevertheless a rotation and a displacement of the protein backbone. This movement leads to a stronger intercalation of Ile 13 at position 6-7 with and increase in roll and rise at this step to 49° and 6.2 Å. This may explain the enhanced preference for adenine at position 7. The protein movement also displaces the N terminal tail and, in turn, Arg 7 which rotates to be more aligned with the DNA axis. This new position maintains a hydrogen bond with the second strand thymine at position 3, while removing a steric hindrance with the H2 proton of adenine at position 3 and forming a hydrogen bond with O2 of cytosine at position 4. These changes support the suggestion that it may indeed be advisable to allow some optimisation of the relative position of the two partners in a complex.[21]

We now consider the origin of the sequence specificity of SRY. Following the approach used for TBP, the sequence logos resulting from the DNA deformation energy alone are shown in the right-hand logos of Fig. 4 for the NMR-derived starting structure (top) and for the unrestrained MD structure (bottom). In either case, the surprising conclusion is that DNA deformation contributes only weakly to the overall selectivity. Thus opening the minor groove and even partially intercalating a protein side chain, with associated inter-base pair roll and helical unwinding, does not necessarily favour a given DNA sequence.

Sensitivity to starting conformation

Given the results of calculating SRY binding selectivity using an MD-derived conformation we decided to compare results

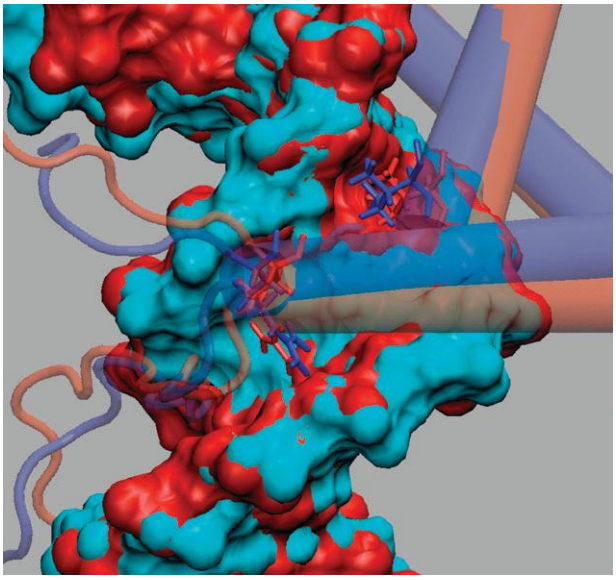

Fig. 5 Change in the SRY–DNA interface between the NMR structure (red) and the MD relaxed structure (blue). Two critical residues are shown in detail, Ile 13 (above) and Arg 7 (below). The remaining protein is shown in transparent cartoon mode and DNA is shown as a solvent accessible surface.

from four different snapshots extracted from the SRY–DNA MD trajectory and also repeat this test for an MD simulation of the TBP–DNA complex, starting, in this case, from the crystallographic coordinates. The results are presented in the supplementary material to this article† and can be compared with the data in Fig. 2 for TBP and in the lower part of Fig. 4 for SRY. For the TBP complex, Fig. S1† shows similar sequence logos for all four MD snapshots. Although the quantitative contributions of individual bases vary to some extent along the binding site, the TATAWAWR is reproduced for the total binding selectivity, with the exception of the last snapshot where T and A are almost equiprobable in position 2. All snapshots also clearly indicate that the recognition is largely "indirect", being dominated by DNA deformation. The results are encouraging given that the individual snapshot conformations vary by up to almost 3 Å RMSD and up to 3.9 Å for the bound DNA molecule (see Table S1†). The snapshots of the SRY complex also show very similar sequence logos for the four snapshots (see Fig. S2†) and concur in the analysis of the recognition as being mainly "direct", that is, due to protein–DNA interaction terms. As for TBP, the MD snapshots of the SRY complex again show significant conformational variability with total RMSD values up to 2.8 Å. In this case, the largest changes (2.5 Å) involve the protein which has a long, flexible tail, but DNA also changes by up to 2.5 Å RMSD. It is interesting to note that the interface area of the TBP complex in the MD snapshots generally stays close to that of the crystallographic structure, while all of the SRY snapshots have an interface area which is larger by roughly 100 $Å^2$. This is compatible with the weaker sequence selectivity seen on the NMR structure of the SRY complex. We also note that, for both TBP and SRY, there is an increased penetration of water into the protein–DNA interface in the MD snapshot conformations so that roughly 500 $Å^2$ of area now involves water contacts rather than direct protein–DNA contacts. This is to be expected given the thermal fluctuations of the complexes in solution at room temperature. But it is interesting that this increased "wetting" of the interface does not lead to visibly decreased recognition.

Nucleosome

We finally consider the sequence selectivity of nucleosome binding. As discussed above, generating the matrix of energy contributions for all overlapping pentanucleotides requires a significant amount of computation. Our approach began with energy optimising the conformation of DNA and of the histone side chains at the protein–DNA interface using roughly 20 000 steps of conjugate gradient minimisation. Given the internal/helical parameter representation used in JUMNA,[36] energy minimisation can lead to much larger changes in conformation than are common with Cartesian coordinate representations. This was confirmed in our first attempts at optimising the nucleosome conformation using parm99 parameters. The underestimated twist in this case resulted in the DNA becoming severely unwrapped from the histone core. The modifications to the backbone in the parmbsc0 force field[39] corrected this problem and resulted in a starting structure (shown in Fig. 6) very close to that of the crystal data.

In preparing the sequence-dependence study of the nucleosomal DNA, we chose to ignore two base pairs at either end, leaving 143 base pairs to scan. This enabled us to avoid having to deal with special cases for terminal pentanucleotides. The remaining 139 overlapping pentanucleotides then had to be

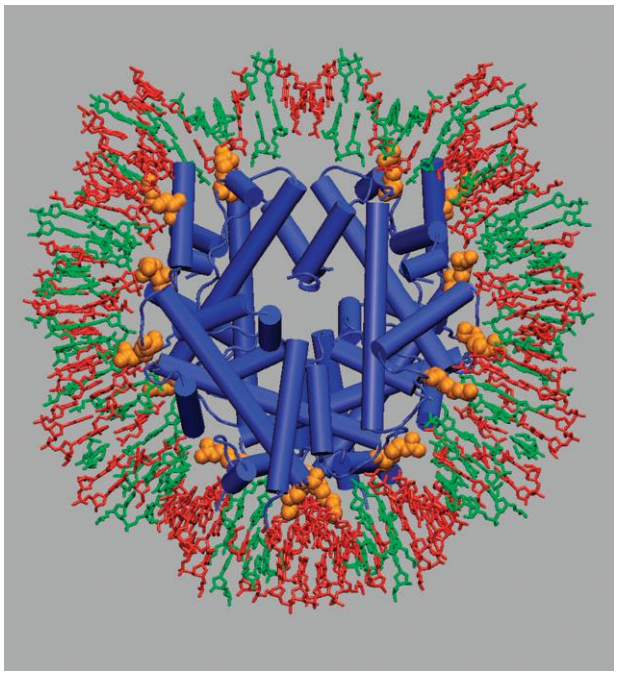

Fig. 6 Nucleosome–DNA complex[35] showing the 14 arginines (H2A chains C/G 29, 42 and 77, H3 chains A/E 49, 63 and 83, and H4 chains B/F 45) in the minor groove in orange CPK. The protein backbone is shown in cartoon mode in blue and the DNA is coloured red for AT pairs and green for GC pairs.

energy minimised with respect to the conformation of the DNA fragment and the neighbouring histone side chains for all possible base sequences. This implies a total of $139 \times 1024 = 142\,336$ optimisations, each with roughly 250 variables and requiring of the order of 500–1000 steps of conjugate gradient minimisation. A similar amount of computation is required to treat the isolated DNA fragment. In total, this task would require roughly 21 years on a single processor and is significantly too big for a typical linux cluster.

This problem was solved by running the computations on the EGEE grid facilities, which currently group together in excess of 114 000 cores in 267 centres around the world (http://eu-egee.org/). In order to avoid difficulties with the heterogeneous environments at different EGEE nodes, the code was included in each submission and local compilation was performed before beginning computations. Each submission required limited amounts of data (the nucleosome conformation, data on atom types, charges and connectivities, the force field parameters and indications of which pentanucleotide fragment and which sequence was to be treated. Output was similarly restricted to short list and data files. The list files were simply used for checking convergence before being deleted, while the necessary energy components were extracted from the data files to build the total energy matrix. Using the EGEE grid enabled the computations to be finished in 11 days. The number of cores running concurrently during the computations was typically more than 800 (reaching peak value of 1850), yielding a so-called "crunching factor" of 766 for 92% of the jobs submitted.

Before using the resulting energy matrix, we had to take into account the asymmetry which exists within the starting conformation of the nucleosome. The nucleosomal complex should show pseudodyadic symmetry, that is a segment such as 5′-TCCAG-3′ centred on a position L base pairs upstream of the nucleosome dyad, should be identical to a segment 5′-CTGGA-3′ (or 5′-TCCAG-3′ reading the complementary strand) L base pairs downstream. In practice, differences in the conformations of symmetry-related segments of the DNA–histone interface (mainly due to modifications in the conformation of side chains and of coupled changes in interfacial water molecules) are significant. Overcoming this problem during energy-optimisation would require a thorough search of side chain rotamers and probably a better treatment of solvent effects, both of which were beyond our current possibilities. We therefore imposed symmetry by comparing all symmetry-related energy locations in the energy matrix and systematically adopting the lowest of the two energies to both locations. A similar approach has been adopted in deriving sequence preference from experimental binding data.[11] This choice not only symmetrizes the energy matrix, but also helps in refining the energy minimisation, since it implies that two minimisations were carried out from somewhat different starting conformations and only the most stable final conformation was conserved.

For the energy matrix of isolated DNA we built the 143 base pair duplex in a B-DNA conformation and performed a full pentanucleotide scan. However, for some sequence fragments this led to unrealistic conformations with unusually narrow minor grooves. This is a problem we have encountered earlier and is related to using simple electrostatic damping to represent solvent effects. The narrow groove state is a local energy minimum, which for some sequences becomes marginally more stable than the conventional B-like minimum. This problem can be corrected using Poisson–Boltzmann electrostatics, but this is not possible for this large system. We consequently built an energy-minimised B-DNA conformation using an average base sequence (JUMNA allows constructions where all base pairs make a 25% energy contribution at each position) and then scanned the 1024 sequences of a single pentanucleotide within this construct. This simpler calculation provided the reference matrix which is subtracted from the nucleosome energy matrix to yield the DNA deformation energy for each chosen sequence

The sequence logo for the entire nucleosome DNA is not easily readable. Consequently, we divide the logo into 14 "windows" of ten base pairs (or 8 in the case of the DNA termini) centred on the positions where arginines from the H2A, H3 and H4 histones penetrate the DNA minor groove. Because of the symmetrisation carried out, there are only seven unique windows, as we move from the pseudodyad position, $5' \rightarrow 3'$ along one turn of the nucleosomal DNA to its terminus. These logos are shown in the left hand column of Fig. 7. Note that because of the imperfect repetition symmetry of the nucleosome, the windows are not necessarily contiguous (see the caption to Fig. 7 for details). It is also remarked that because of dyad symmetry any base selectivity in the windows shown will be reflected in the same selectivity for the complementary strand in the DNA turn preceding the pseudodyad.

There are several striking things to note with these results. Firstly, the sequence recognition is concentrated over a relatively small number of base pairs and very variable from window to window. Windows 2 and 3 (which interact with arginines from H3) show very little selectivity, while windows 1 (interacting with H4), 4, 5, 6 (interacting with H2A) and 7 (interacting with H3) carry most of the signals. Secondly, significant selectivity always implies selectivity in the centre of each window, close to the position where the arginine enters the minor groove, although, in the windows with the strongest signals (windows 4 and 6), some selectivity also involves the DNA whose major groove faces, but is distant from, the histone core. Thirdly, while adenine and thymine dominate the logos, we also see one strong selectivity for cytosine in window 4 and several weaker cytosine or guanine signals. When we look at the DNA deformation contribution alone (data not shown) we find that although this term is important in generating the overall selectivity it is far from dominant and we see a closer correlation between the total signals and those coming from the interaction term. This is contrary to common assumptions about the nature of nucleosome recognition, and merits further studies, notably studies in the absence of the minor groove penetrating histone arginines. This study is in progress.

As a test of the nucleosome binding energy predictions, we turn to data on the preferential positioning of nucleosomes around eukaryotic transcription start sites (TSS). Based on the TSS locations in the yeast genome,[48] Lee *et al.*[2] have recently determined preferential nucleosome positions. Their results

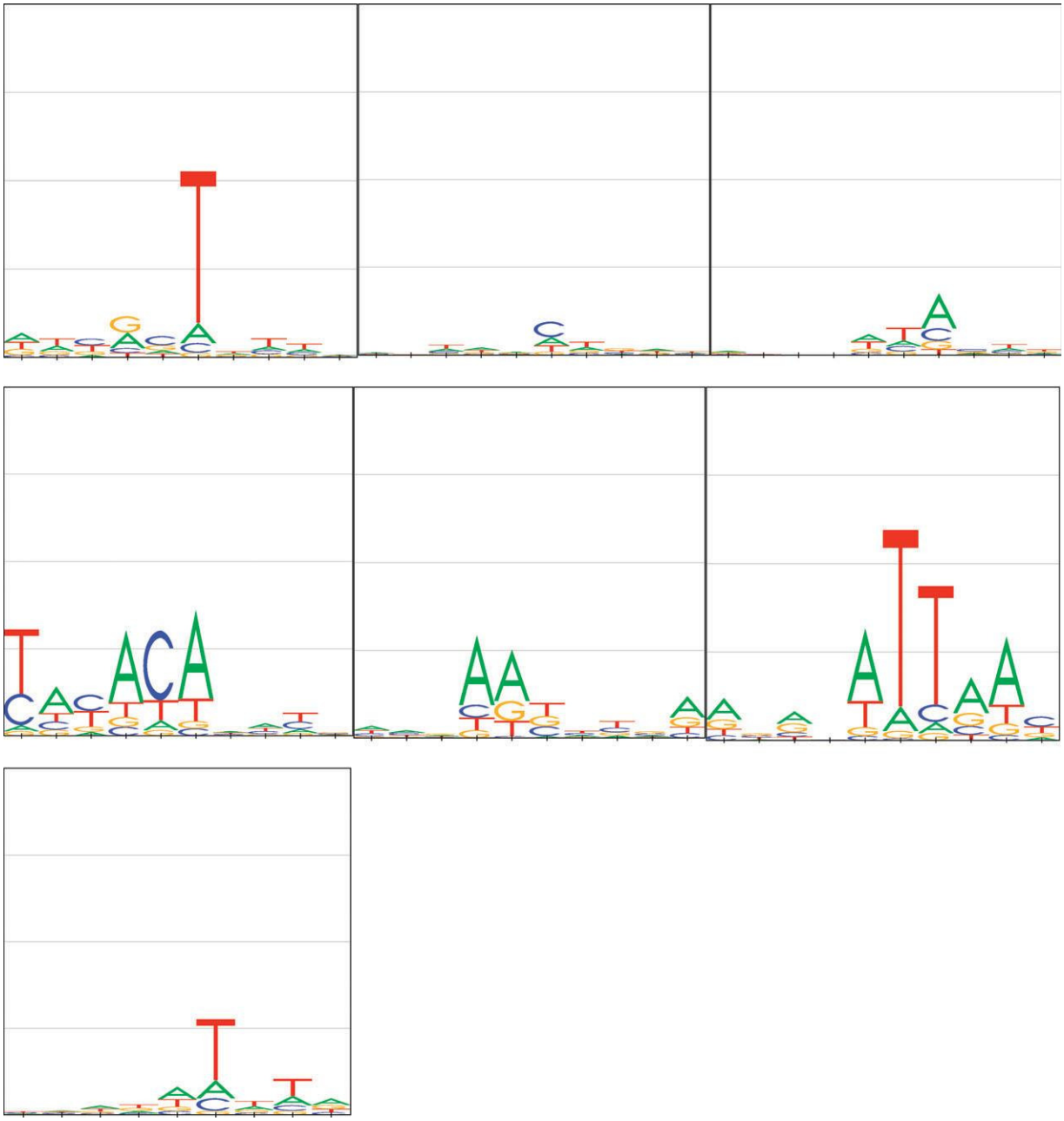

Fig. 7 Sequence logos for seven windows along one half of the nucleosomal DNA starting from the pseudodyad and moving in the 5′ → 3′ direction of one strand. The windows cover the base positions: 72–82 (1st row left), 82–91 (1st row centre) , 92–101 (1st row right), 104–113 (2nd row left), 115–124 (2nd row centre), 124–133 (2nd row right) and 135–143 (3rd row).

aligned for a total of 5015 TSS are shown in Fig. 8 as a dotted curve. It is clear from this data that nucleosomes avoid a zone just upstream of the TSS locations, but show enhanced probabilities just before and after this zone. There is also a distinct oscillation in the probabilities moving away from the TSS either upstream or downstream, with a spacing between peaks of roughly 160 base pairs. We used the nucleosome total binding energy matrix described above to scan 2000 base pairs centred on each of the 5015 TSS. We remark that this scan is extremely rapid, since estimating the nucleosome binding energy at a given site only requires adding several hundred energy components from the pre-calculated matrix. It is consequently possible to scan roughly 50 000 bp s^{-1} on a single 3 GHz processor. The results are shown in terms of predicted binding energies as the solid line in Fig. 8 (using an inverted axis to reflect the experimental binding probability curve). The correspondence with the experimental results is good concerning the nucleosome depleted zone and the enhanced probabilities around this zone, however the correspondence with the small oscillations moving further away up- or downstream is not seen. These oscillations may be simply due to the regular packing of successive nucleosome with respect to one strongly enforced position, following the so-called "parking lot model".[49]

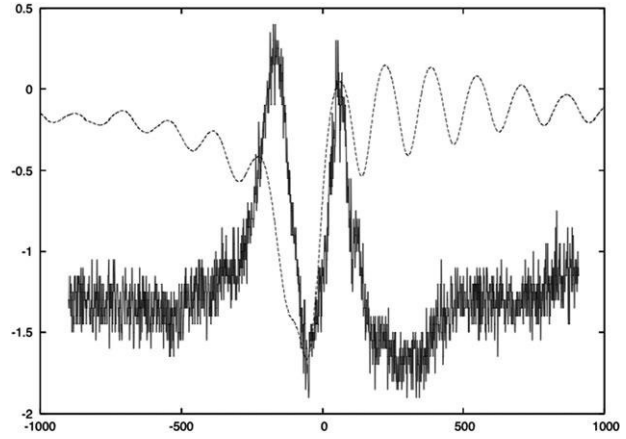

Fig. 8 Nucleosome binding energies calculated with ADAPT (solid line), compared to experimentally determined binding probabilities (dotted line)[2] in the vicinity of 5015 yeast transcription start sites,[48] aligned at position 0 on the abscissa. The ADAPT energies are plotted with an inverted axis to facilitate the comparison.

Conclusions

We have described the application of an all-atom, physics-based approach to predicting protein–DNA binding specificity which allows an exhaustive scan of all possible nucleic acid sequences within the binding site. We have obtained encouraging results using a standard force field and a simple treatment of solvent effects without any special reparameterisation. The only experimental data required for this approach is the atomic structure of a single complex between DNA and the target protein (or protein complex). Thanks to a divide-and-conquer algorithm, the approach is applicable to binding sites ranging from a few base pairs to well over a hundred with linear scaling.

Compared to the earlier version of our approach, the introduction of flexibility at the protein–DNA interface is a major step forward. However, there is still room for improvement, with the prime targets being a better optimisation of the amino side chains at the interface, *via* a rotamer search, and an better treatment of solvent effects. Since our method only optimised short fragments of the protein–DNA interface, the initial conformation of the complex remains important. Earlier work has suggested that unrestrained energy optimisation of complexes generally makes the results worse. It is consequently encouraging to see that a molecular dynamics relaxation of SRY with an explicit solvent and counterion environment actually leads to better predictions of sequence specificity. Similar treatment of TBP leads to results very close to those obtained with the crystallographic coordinates. We have also demonstrated that, for both SRY and TBP, the results in terms of sequence selectivity are robust with respect to the conformational fluctuations that occur during room temperature molecular dynamics simulations.

Concerning extensions to the conformational ADAPT search procedure, we must keep in mind the computational effort behind the overlapping oligonucleotide scans. Nevertheless, the grid-adapted nature of these computations means that we could support at least an order of magnitude increase at this stage without the approach becoming unfeasible, even for multi-macromolecular complexes as large as, or larger than, the nucleosome.

Finally, the results obtained for three different protein complexes which share binding on the minor groove face of DNA and significant induced-deformation of DNA provide some interesting contrasts. The sequence logos generated for TBP and SRY agree well with experimental results, as do the nucleosome position prediction around yeast transcription start sites. However, the study of these three systems suggests that significant DNA deformation upon protein binding does not necessarily translate into a significant indirect contribution to sequence selectivity. This is particularly surprising in the case of the nucleosome where DNA deformation does not appear to play a dominate role, as generally assumed. We are currently pushing this analysis further by looking at the role of the initial conformation of the nucleosome core particle, and at the impact of histone modifications.

Acknowledgements

R. L. and K. Z. acknowledge funding from the ANR grant HUGOREP/NT05-3_41825. B. B., A. M. and C. B. acknowledge funding from the ANR grant HIPCAL/ANR-06-CIS6-005. A. M. acknowledges funding from EU-FP6 grant EMBRACE/LHSG-CT-2004-512092. This work makes use of results produced with the EGEE grid (http://www.eu-egee.org) infrastructure co-funded by the European Commission (INFSO-RI-222667).

References

1 G. C. Yuan, Y. J. Liu, M. F. Dion, M. D. Slack, L. F. Wu, S. J. Altschuler and O. J. Rando, *Science*, 2005, **309**, 626.
2 W. Lee, D. Tillo, N. Bray, R. H. Morse, R. W. Davis, T. R. Hughes and C. Nislow, *Nat. Genet.*, 2007, **39**, 1235.
3 E. Roulet, S. Busso, A. A. Camargo, A. J. Simpson, N. Mermod and P. Bucher, *Nat. Biotechnol.*, 2002, **20**, 831.
4 T. H. Kim and B. Ren, *Annu. Rev. Genomics Hum. Genet.*, 2006.
5 S. J. Maerkl and S. R. Quake, *Science*, 2007, **315**, 233.
6 J. B. Kinney, G. Tkacik and C. G. Callan, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 501.
7 G. D. Stormo, *Bioinformatics*, 2000, **16**, 16.
8 A. O'Flanagan R, G. Paillard, R. Lavery and A. M. Sengupta, *Bioinformatics*, 2005, **21**, 2254.
9 E. Sharon, S. Lubliner and E. Segal, *PLoS Comput. Biol.*, 2008, **4**, e1000154.
10 J. Liu and G. D. Stormo, *BMC Bioinformatics*, 2005, **6**, 176.
11 E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thastrom, Y. Field, I. K. Moore, J. P. Wang and J. Widom, *Nature*, 2006, **442**, 772.
12 S. Selvaraj, H. Kono and A. Sarai, *J. Mol. Biol.*, 2002, **322**, 907.
13 M. M. Gromiha, J. G. Siebers, S. Selvaraj, H. Kono and A. Sarai, *J. Mol. Biol.*, 2004, **337**, 285.
14 M. J. Arauzo-Bravo, S. Fujii, H. Kono, S. Ahmad and A. Sarai, *J. Am. Chem. Soc.*, 2005, **127**, 16074.
15 W. Ge, B. Schneider and W. K. Olson, *Biophys. J.*, 2005, **88**, 1166.
16 B. Contreras-Moreira and J. Collado-Vides, *Bioinformatics*, 2006, **22**, e74.
17 J. J. Havranek, C. M. Duarte and D. Baker, *J. Mol. Biol.*, 2004, **344**, 59.
18 A. V. Morozov, J. J. Havranek, D. Baker and E. D. Siggia, *Nucleic Acids Res.*, 2005, **33**, 5781.
19 J. Ashworth, J. J. Havranek, C. M. Duarte, D. Sussman, R. J. J. Monnat, B. L. Stoddard and D. Baker, *Nature*, 2006, **441**, 656.
20 T. W. Siggers and B. Honig, *Nucleic Acids Res.*, 2007, **35**, 1085.
21 S. Jamal Rahi, P. Virnau, L. A. Mirny and M. Kardar, *Nucleic Acids Res.*, 2008, **36**, 6209.

22 N. R. Steffen, S. D. Murphy, L. Tolleri, G. W. Hatfield and R. H. Lathrop, *Bioinformatics*, 2002, **18**(Suppl. 1), S22.
23 Y. Zhang, Z. Xi, R. S. Hegde, Z. Shakked and D. M. Crothers, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 8337.
24 V. Miele, C. Vaillant, Y. d'Aubenton-Carafa, C. Thermes and T. Grange, *Nucleic Acids Res.*, 2008, **36**, 3746.
25 I. Lafontaine and R. Lavery, *Biophys. J.*, 2000, **79**, 680.
26 I. Lafontaine and R. Lavery, *Biopolymers*, 2000, **56**, 292.
27 G. Paillard and R. Lavery, *Structure*, 2004, **12**, 113.
28 C. Deremble, R. Lavery and K. Zakrzewska, *Comput. Phys. Commun.*, 2008, **179**, 112.
29 C. A. Bewley, A. M. Gronenborn and G. M. Clore, *Annu. Rev. Biophys. Biomol. Struct.*, 1998, **27**, 105.
30 P. Koopman, J. Gubbay, N. Vivian, P. Goodfellow and R. Lovell-Badge, *Nature*, 1991, **351**, 117.
31 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Res.*, 2000, **28**, 235.
32 D. B. Nikolov, H. Chen, E. D. Halay, A. Hoffman, R. G. Roeder and S. K. Burley, *Proc. Natl. Acad. Sci. U. S. A.*, 1996, **93**, 4862.
33 E. C. Murphy, V. B. Zhurkin, J. M. Louis, G. Cornilescu and G. M. Clore, *J. Mol. Biol.*, 2001, **312**, 481.
34 C. A. Davey, D. F. Sargent, K. Luger, A. W. Maeder and T. J. Richmond, *J. Mol. Biol.*, 2002, **319**, 1097.
35 T. J. Richmond and C. A. Davey, *Nature*, 2003, **423**, 145.
36 R. Lavery, K. Zakrzewska and H. Sklenar, *Comput. Phys. Commun.*, 1995, **91**, 135.
37 I. Lafontaine and R. Lavery, *Curr. Opin. Struct. Biol.*, 1999, **9**, 170.
38 T. E. 3. Cheatham, P. Cieplak and P. A. Kollman, *J. Biomol. Struct. Dyn.*, 1999, **16**, 845.
39 A. Perez, I. Marchan, D. Svozil, J. Sponer, T. E. 3. Cheatham, C. A. Laughton and M. Orozco, *Biophys. J.*, 2007, **92**, 3817.
40 J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale and K. Schulten, *J. Comput. Chem.*, 2005, **26**, 1781.
41 W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *J. Chem. Phys.*, 1983, **79**, 926.
42 S. B. Dixit, D. L. Beveridge, D. A. Case, T. E. 3. Cheatham, E. Giudice, F. Lankas, R. Lavery, J. H. Maddocks, R. Osman, H. Sklenar, K. M. Thayer and P. Varnai, *Biophys. J.*, 2005, **89**, 3721.
43 C. Blanchet, R. Mollon, D. Thain and G. Deleage, *2006 7th IEEE/ACM International Conference on Grid Computing*, 2006, p. 120.
44 R. Lavery, M. Moakher, J. H. Maddocks, D. Petkeviciute and K. Zakrzewska, *Nucleic Acids Res.*, 2009.
45 W. Humphrey, A. Dalke and K. Schulten, *J. Mol. Graphics*, 1996, **14**, 33.
46 P. Bucher, *J. Mol. Biol.*, 1990, **212**, 563.
47 V. R. Harley, R. Lovell-Badge and P. N. Goodfellow, *Nucleic AcidsRes.*, 1994, **22**, 1500.
48 L. David, W. Huber, M. Granovskaia, J. Toedling, C. J. Palm, L. Bofkin, T. Jones, R. W. Davis and L. M. Steinmetz, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 5320.
49 R. Kiyama and E. N. Trifonov, *FEBS Lett.*, 2002, **523**, 7.