Specificity in Protein-DNA Interactions: TATA Box-Binding Protein Recognition of Promoter Sequences

by

Nina Pastor Colon

A dissertation submitted to the Graduate Faculty in Biomedical Sciences in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

1997

This manuscript has been read and accepted for the Graduate Faculty in Biomedical Sciences in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

8/14/97
Date

R. Osman
Dr. Roman Osman
Chair of Examining Committee

8/16/97
Date

Dr. Terry A. Krulwich
Executive Officer

Members of the Examining Committee:

Dr. Roman Osman
Dr. Harel Weinstein
Dr. Carter Bancroft
Dr. Mihaly Mezei
Dr. David Beveridge

THE CITY UNIVERSITY OF NEW YORK

Abstract

SPECIFICITY IN PROTEIN-DNA INTERACTIONS: TATA BOX-BINDING PROTEIN RECOGNITION OF PROMOTER SEQUENCES

by

Nina Pastor Colon

Advisor: Professor Harel Weinstein, D. Sc.

The TATA box-binding protein (TBP) is an ancient basal transcription factor absolutely required for transcription in archaea and eukaryotes. In order to understand at the molecular level the mechanisms that might be used by TBP to select its binding sites (the TATA boxes), molecular dynamics simulations were carried out to explore the conformations of DNA oligomers that contain TATA boxes, of TBP, and of a TBP/DNA complex, all in aqueous solution. The free DNA sequences were chosen to include known binding sites for wild type TBP, binding sites for mutant TBPs, one inverted TATA box, and a sequence that is not bound by TBP and serves as a negative control. An inosine variant of the Major Late Promoter was included to test the equivalence of IC and AT basepairs. Direct comparison to experimental data showed that the simulation protocol is capable of reproducing known structural and dynamic properties of the systems studied. The counterion distribution around the DNA oligomers showed some sequence dependence, but the interpretation could be complicated as different measures indicate varying degrees of temporal

convergence.

A TATA box sequence ideal for TBP binding was derived from the analysis of specific average and transient structural properties of the free DNA oligomers that resemble the conformation imposed on DNA in complexes with TBP. Comparison to experimental data showed that the stronger the adherence to this ideal sequence, the greater the measured affinity of TBP for the particular DNA sequence. The simulations of TBP revealed little change in the backbone structure, but flexibility evidenced by the sampling of various rotamers by the side chains of internal residues. The distribution of the rotating side chains is different in the free protein and in the complex, and their location on the protein could be related to the directionality of TBP binding to DNA. Analysis of the water structure surrounding the macromolecules allowed for the quantification of the changes in the structure of water in the first hydration layer. These changes in the solvent structure are related to the measured decrease in heat capacity upon TBP binding to TATA sequences.

This thesis is dedicated to

my father; I think he would have liked it

my family in México, for their unwavering support

the friends in México and abroad who have kept in touch with me through these years

my acquired family in New York, wherever they might be

Acknowledgements

I have been very lucky to be guided through graduate school by Dr. Harel Weinstein. This work reflects the combination of freedom and obsession with perfection that characterized four years of collaboration.

I will be eternally indebted to Dr. Roman Osman for introducing me to the marvels of DNA conformation and the dynamics of the counterions around DNA. I thank Dr. Leonardo Pardo for the passion for mechanistic explanations, and for statistics. May the Iberoamerican collaboration last a long time.

The analysis presented in Chapters 6 and 7 would not have been possible without the invaluable help of Dr. Mihaly Mezei, and what he calls the bleeding edge of science.

I would like to thank Dr. Alex MacKerell for providing me with the CHARMM parameters for inosine bases, and for waiting patiently for me to get my act together. This work required many CPU hours and gigabytes of disk space. For their patience, I thank the system managers at Mount Sinai, especially Janne Ravantti and Kevin Kelliher. Some of the simulations were carried out at the Cornell National Supercomputer Facility (sponsored by the National Science Foundation and IBM). I thank Dr. Ernest Mehler for the multiple implementations of CHARMM in all the platforms I could use.

The staff in the main office of the Department of Physiology and Biophysics has made my life very easy during these years. Special thanks to Maureen Milici, Mildred Tolson, Miriam Rivera and Karen Perry.

This work was supported by a Fulbright/CONACyT (México) scholarship (number 80242) and by the Association for International Cancer Research.

Table of Contents

List of Tables

List of Figures

1 Introduction

The successful formation of a protein/DNA complex depends on multiple factors (Record 1991; Rhodes 1996; von Hippel 1986), including the generation of the appropriate contact surfaces between the protein and the DNA, the ease of deformation of the DNA and/or the protein, the dehydration of the interface and the release of counterions from the DNA and the protein. The TATA Box-binding protein (TBP) seems to select its binding sites through a delicate balance of all these contributions. This thesis is devoted to the exploration of the interplay between such determinant factors in the formation of specific complexes between DNA and TBP.

1.1 The Biology of TBP

TBP is an ancient transcription factor, existing in eukaryotes (Roeder 1996) and archaebacteria (Marsh 1994; Rowlands 1994). Its function is to recognize promoter sequences in the DNA, thereby directing RNA polymerase to start transcription at precise sites. It is the functional analogue of bacterial σ factors (Hellman 1988).

TBP is absolutely required for transcription by the three nuclear RNA polymerases (Cormack 1992). RNA polymerase II transcription of many protein coding genes and the transcription of some RNA genes by RNA polymerase III

requires direct contact of the DNA by TBP, at a sequence called the TATA box, which is located approximately 30 basepairs upstream of the transcription initiation site (Breathnach 1981; Burley 1996). This complex directs the assembly of the remaining basic transcription factors into a preinitiation complex (Orphanides 1996).

1.1.1 Molecular Biology

TBP has been cloned from many organisms, ranging from archaea (Marsh 1994; Rowlands 1994) to humans (Hoffmann 1990; Kao 1990), spanning 3 billion years of evolution. A list of the available TBP sequences at the National Center for Biotechnology Information (NCBI) is included in the Materials section (2.1), and an alignment of the 180-residue C-terminal domain is shown below in Table 1.1. This is the actual DNA binding domain. The alignment was done by hand given the paucity of insertions/deletions in the DNA binding domain of TBP. For the more divergent sequences, published alignments (DeDecker 1996; Marsh 1994; Nikolov 1994) were used to position the insertions/deletions. Residue numbering and protein names in this work are defined by this alignment.

Table 1.1 Sequence alignment of the conserved C-terminal domain of TBP

```
mutation       @   $    @    @@@ @ @@   @   $ @$ $    @    $ @ @  @ @@ # @ @@@
100%               N              L          Y P           P     LIF  G
DNA               **  * *                   *   ***   * *  !*  * * * * *
ATH2  ( 19) 1 SGIVPTLQNI VSTVNLDCKL DLKAIALQAR NAEYNPKRFA AVIMRIREPK TTALIFASGK
ATH1  ( 19) 1 SGIVPTLQNI VSTVNLDCKL DLKAIALQAR NAEYNPKRFA AVIMRIREPK TTALIFASGK
ZMA1  ( 19) 1 SGIVPTLQNI VSTVNLDCKL DLKAIALQAR NAEYNPKRFA AVIMRIREPK TTALIFASGK
ZMA2  ( 19) 1 SGIVPTLQNI VSTVNLDCKL DLKAIALQAR NAEYNPKRFA AVIMRIREPK TTALIFASGK
STU   ( 19) 1 SGIVPTLQNI VSTVNLDCKL DLKAIALQAR NAEYNPKRFA AVIMRIREPK TTALIFASGK
NTA   ( 19) 1 SGIVPTLQNI VSTVNLDCKL DLKAIALQAR NAEYNPKRFA AVIMRIREPK TTALIFASGK
MCR   ( 19) 1 SGIVPTLQNI VSTVNLDCKL DLKAIALQAR NAEYNPKRFA AVIMRIREPK TTALIFASGK
GMA   ( 19) 1 SGIVPTLQNI VSTVNLDCKL DLKTIALQAR NAEYNPKRFA AVIMRIRDPK TTALIFASGK
TAE1  ( 51) 1 SGIVPVLQNI VSTVNLDCRL DLKQIALQAR NAEYNPKRFA AVIMRIRDPK TTALIFASGK
TAE2  ( 20) 1 SGIIPTLQNI VSTVNLDCKL DLKAIALQAR NAEYNPKRFA AVIMRIREPK TTALIFASGK
ACA   ( 79) 1 SGIVPTLQNI VSTVNLGCKL DLKNIALHAR NAEYNPKRFA AVIMRIREPK TTALIFASGK
PCA1  ( 49) 1 SGIVPSLQNI VATVNLDCRL DLKTIALQRR NAEYNPKRFA AVIMRIREPK TTALIFASGK
PCA2  ( 50) 1 SGIVPTLQNI VATVNLDCRL DLKTIALHAR NAEYNPKRFA AVIMRIREPK TTALIFASGK
SPO   ( 52) 1 SGIVPTLQNI VATVNLDCRL DLKTIALHAR NAEYNPKRFA AVIMRIREPK STALIFASGK
SCE   ( 61) 1 SGIVPTLQNI VATVTLGCRL DLKTVALHAR NAEYNPKRFA AVIMRIREPK TTALIFASGK
GGA   (122) 1 SGIVPQLQNI VSTVNLGCKL DLKTIALRAR NAEYNPKRFA AVIMRIREPR TTALIFSSGK
XLA   (117) 1 SGIVPQLQNI VSTVNLGCKL DLKTIALRAR NAEYNPKRFA AVIMRIREPR TTALIFSSGK
MMU   (136) 1 SGIVPQLQNI VSTVNLGCKL DLKTIALRAR NAEYNPKRFA AVIMRIREPR TTALIFSSGK
MAU   (138) 1 SGIVPQLQNI VSTVNLGCKL DLKTIALRAR NAEYNPKRFA AVIMRIREPR TTALIFSSGK
HSA   (155) 1 SGIVPQLQNI VSTVNLGCKL DLKTIALRAR NAEYNPKRFA AVIMRIREPR TTALIFSSGK
TGR   (122) 1 SGIVPQLQNI VSTVNLGCKL DLKTIALRAR NAEYNPKRFA AVIMRIREPR TTALIFSSGK
PFL   (120) 1 SGIVPQLQNI VSTVNLGCKL DLKTIALRAR NAEYNPKRFA AVIMRIREPR TTALIFSSGK
ACL   ( 12) 1 SGIIPELQNV VSTVNLGCTL ELKEIAMQAR NAEYNPKRFA AVIMRIRDPK TTALIFGSGK
ENI   ( 88) 1 SGIVPTLQNI VATVNLDCRL DLKTIALHAR NAEYNPKRFA AVIMRIREPK TTALIFASGK
SPU   ( 85) 1 SGIVPQLQNI VSTVNLNCRL DLKKIALHAR NAEYNPKRFA AVIMRIREPR TTALIFSSGK
SFR   (128) 1 PGIVPQLQNI VSTVNLNCKL DLKKIALHAR NAEYNPKRFA AVIMRIREPR TTALIFSSGK
DME   (172) 1 PGIVPQLQNI VSTVNLCCKL DLKKIALHAR NAEYNPKRFA AVIMRIREPR TTALIFSSGK
DDI   ( 21) 1 SGIIPTLQNI VSTVNMATEL YLKAIALGAR NAEYNPKRFA AVIMRIREPK TTALIFKSGK
CEL   (161) 1 DIPMPALQNI VSTVNLGVQL DLKKIALHAR NAEYNPKRFA AVIMRIREPR TTALIFSSGK
OVO   (  1) 1 DVPSPALQNI VSTVNLGVPL DLKKIALHAR NAEYNPKRFA AVIMRIREPR TTALIFSSGK
TTH   ( 47) 1 EDLYPKLQNI VSTVNLSTKL DLKQIALRAR NAEYNPKRFA AVIMRLRDPK TTALIFASGK
DROT  ( 45) 1 AQHEIRLQNI VATFSVNCEL DLKAINSRTR NSEYSPKRFR GVIMRMHSPR CTALIFRTGK
EHI   ( 48) 1 TITHPEIVNV VSRFQLGVKL ELRKIVQKAI NAIYNPKRFA GAIMRISSPK STALIFQTGK
PFA   ( 45) 1 KNISLNIHNI ISSANLCIDI NLRLVAVSIR NAEYNPSKIN TLIIRLNKPQ CTALIFKNGR
SSH   (  8) 1 YKPIVNIENI VATVTLEQSL DLYAMERSIP NIEYDPDQFP GLIFRLEQPK VTALIFKSGK
SAC   (  8) 1 YKAVVNIENI VATVTLDQTL DLYAMERSVP NVEYDPDQFP GLIFRLESPK ITSLIFKSGK
HASA  (  4) 1 LADTIHIENV VASSDLGQEL ALDQLSTDLP GAEYNPEDFP GVVYRLQEPK SATLIFRSGK
MJA   (  4) 1 MEPEIKIVNV VVSTKIGDNI DLEEVAMILE NAEYEPEQFP GLVCRLSVPK VALLIFRSGK
PWO   (  5) 1 SKVKLRIENI VASVDLFAQL DLEKVLDLCP NSKYNPEEFP GIICHLDDPK VALLIFSSGK
PSP   (  5) 1 SNVKLRIENI VASVDLFTDL NLEKVIEICP SSKYNPEEFP GIICRFDDPK VALLIFSSGK
TCE   (  2) 1 SNVKLRIENI VASVDLFTQL NLERVIEMCP HSKYNPEEFP GIICRFDEPK VALLIFSSGK
              |------------|       |-----|    |-|       |---|       |---|   |
                    S1               H1       S2         S3          S4
```

```
mutation      S @ $@    @ @@@   @    @@@       @  @$          @  $@   @  @   @@
100%                                                  N
DNA         * ** *                                * ** *
ATH2  61  MVCTGAKSED FSKMAARKYA RIVQKLGF----PA KFKDFKIQNI VGSCDVKFPI RLEGLAYSHA
ATH1  61  MVCTGAKSEH LSKLAARKYA RIVQKLGF----PA KFKDFKIQNI VGSCDVKFPI RLEGLAYSHS
ZMA1  61  MVCTGAKSEQ QSKLAARKYA RIIQKLGF----PA KFKDFKIQNI VGSCDVKFPI RLEGLAYSHG
ZMA2  61  MVCTGAKSEQ QSKLAARKYA RIIQKLGF----PA KFKDFKIQNI VGSCDVKFPI RLEGLAYSHG
STU   61  MVCTGAKSEQ QSKLAARKYA RIIQKLGF----PA KFKDFKIQNI VGSCDVKFPI RLEGLAYAHG
NTA   61  MVCTGAKSEQ SSKLAARKYA RIIQKLGF----DA KFKDFKIQNI VGSCDVKFPI RLEGLAYSHG
MCR   61  MVCTGAKSEQ QSKLAARKYA RIIQKLGF----PA KFKDFKIQNI VGSCDVKFPI RLEGLAYSHG
GMA   61  MVCTGAKSEQ QSKLAARKYA RIIQKLGF----PA KFKDFKIQNI VGSCDVKFPI RLEGLAYSHG
TAE1  61  MVCTGAKSEE HSKLAARKYA RIVQKLGF----PA TFKDFKIQNI VASCDVKFPI RLEGLAYSHG
TAE2  61  MVCTGAKSEQ QSKLAARKYA RIIQKLGF----PA KFKDFKIQNI VASCDVKFPI RLEGLAYSHG
ACA   61  MVCTGAKSEE ASRLAARKYA RIIQKLGF----AA KFLDFKIQNI VGSCDVRFPI RLEGLAFAHN
PCA1  61  MVVTGAKSED DSKLASRKYA RIIQKLGF----NA KFTDFKIQNI VGSCDVKFPI RLEGLAYSHG
PCA2  61  MVVTGAKSED DSKLASRKYA RIIQKLGF----NA KFTDFKIQNI VGSCDVKFPI RLEGLAYSHG
SPO   61  MVVLGGKSED DSKLASRKYA RIIQKLGF----NA KFTDFKIQNI VGSCDVKFPI RLEGLAYSHG
SCE   61  MVVTGAKSED DSKLASRKYA RIIQKIGF----AA KFTDFKIQNI VGSCDVKFPI RLEGLAFSHG
GGA   61  MVCTGAKSEE QSRLAARKYA RVVQKLGF----PA KFLDFKIQNM VGSCDVKFPI RLEGLVLTHQ
XLA   61  MVCTGAKSEE QSRLAARKYA RVVQKLGF----PA KFLDFKIQNM VGSCDVKFPI RLEGLVLTHQ
MMU   61  MVCTGAKSEE QSRLAARKYA RVVQKLGF----PA KFLDFKIQNM VGSCDVKFPI RLEGLVLTHQ
MAU   61  MVCTGAKSEE QSRLAARKYA RVVQKLGF----PA KFLDFKIQNM VGSCDVKFPI RLEGLVLTHQ
HSA   61  MVCTGAKSEE QSRLAARKYA RVVQKLGF----PA KFLDFKIQNM VGSCDVKFPI RLEGLVLTHQ
TGR   61  MVCTGAKSEE QSRLAARKYA RVVQKLGF----PA KFLDFKIQNM VGSCDVKFPI RLEGLVLTHQ
PFL   61  MVCTGAKSEE QSRLAARKYA RVVQKLGF----PA KFLDFKIQNM VGSCDVKFPI RLEGLVLTHQ
ACL   61  MVCTGAKSEQ DSRTAARKYA KIVQKLGF----PA KFTEFKIQNI VGSCDVKFPI RMEPLAYQHQ
ENI   61  MVVTGAKSED DSKLASRKYA RIIQKLGF----NA KFTDFKIQNI VGSCDIKFPI RLEGLASRHH
SPU   61  MVCTGAKRED NSRLAARKYA RVVQKLGF----AA KFLDFKIQNM VGSCDVKFPI RLEGLVLTHG
SFR   61  MVCTGAKSEE DSRLAARKYA RIIQKLGF----TA KFLDFKIQNM VGSCDVKFPI RLEGLVLTHG
DME   61  MVCTGAKSED DSRLAARKYA RIIQKLGF----PA KFLDFKIQNM VGSCDVKFPI RLEGLVLTHC
DDI   61  MVCTGAKSED ASRFAARKYA RIIQKLDF----PA RFTDFKIQNI VGSCDVKFPI KLELLHNAHT
CEL   61  MVCTGAKSEE ASRLAARKYA RIVQKLGF----QA KFTEFMVQNM VGSCDVRFPI QLEGLCITHS
OVO   61  MVCTGAKSEE SSRLAARKYA RIVQKLGF----NA KFTEFKVQNM VGSCDVRFPI QLEGLCLTHT
TTH   61  MVCTGAKTEE DSNRAARKYA KIIQKIGF----PV QFKDFKIQNI VGSTDVKFPI NLDHLEQDHK
DROT  61  VICTGARNEI EADIGSRKFA RILQKLGF----PV KFMEYKLQNI VATVDLRFPI RLENLNHVHG
EHI   61  IVCTGTRSIE ESKIASKKYA KIIKKIGY----PI HYSNFNVQNI VGSCDVKFQI ALRTLVDSDL
PFA   61  IMLTGTRTKK DSIMGCKKIA KIIKIVTKD---KV KFCNFKIENI IASANCNIPI RLEVLAHDHK
SSH   61  MVVTGAKSTE ELIKAVKRII KTLKKYGIK---IM GKPKIQIQNI VASANLHVNV NLDKAAFLLE
SAC   61  MVVTGAKSTD ELIKAVKRII KTLKKYGMQ---LT GKPKIQIQNI VASANLHVIV NLDKAAFLLE
HASA  61  VVCTGAKSVD DVHEALGIVF GDIRELGID---VT SNPPIEVQNI VSSASLEQSL NLNAIAIGLG
MJA   61  MKIVNCTGAK SKEEAEIAIK KIIKELKDAGIDVI ENPEIKIQNM VATADLGIEP NLDDIALMVE
PWO   61  LVVTGAKSVQ DIERAVAKLA QKLKSIGVK---FK RAPQIDVQNM VFSGDIGREF NLDVVALTLP
PSP   61  LVVTGAKSVD DIKRAVYKLI EMLKKIGAK---FT REPQIDIQNM VFSGDIGMEF NLDAVALILP
TCE   61  LVVTGAKSVE DIERAVNKLI QMLKKIGAK---FS RAPQIDIQNM VFSGDIGMEF NLDAVALSLP
          -----|  |----------------|            |-----------|       |-----|
            S5             H2                      S1'              H1'
```

```
mutation      $ $ @@ @ @# @     #  @ #@# $ @@ @  @   $ @  @  @      $@
100%          YEPE F  G                 L F  G K    G
DNA              ** *  * *       ! * * * * *  * * ** *
ATH2 121 -AFSSYEPELF PGLIYRMK---VP KIVLLIFVSG KIVITGAKMR DETYKAFENI YPVLSEFRKI
ATH1 121 -AFSSYEPELF PGLIYRMK---LP KIVLLIFVSG KIVITGAKMR EETYTAFENI YPVLREFRKV
ZMA1 121 -AFSSYEPELF PGLIYRMK---QP KIVLLIFVSG KIVLTGAKVR EETYTAFENI YPVLAEFRKV
ZMA2 121 -AFSSYEPELF PGLIYRMK---QP KIVLLIFVSG KIVLTGAKVR EETYTAFENI YPVLSEFRKI
STU  121 -AFSSYEPELF PGLIYRMK---QP KIVLLIFVSG KIVITGAKVR DETYTAFENI YPVLTEFRKN
NTA  121 -AFSSYEPELF PGLIYRMK---QP KIVLLIFVSG KIVLTGAKVR DETYTAFENI YPVLTEFRKN
MCR  121 -AFSSYEPELF PGLIYRMK---QP KIVLLIFVSG KIVLTGAKVR EETYTAFENI YPVLTEFRKN
GMA  121 -AFSSYEPELF PGLIYRMK---QP KIVLLIFVSG KIVLTGAKVR DETYTAFENI YPVLTEFRKN
TAE1 121 -AFSSYEPELF PGLIYRMK---QP KIVLLVFVSG KIVLTGAKVR DEIYAAFENI YPVLTEYRKS
TAE2 121 -AFSSYEPELF PGLIYRMR---QP KIVLLIFVSG KIVLTGAKVR EETYSAFENI YPVLTEFRKV
ACA  121 -HYCSYEPELF PGLIYRMV---QP KIVLLIFVSG KIVLTGAKVR EEIYEAFENI YPVLTEYKKT
PCA1 121 -TFSSYEPELF PGLIYRMV---KP KIVLLIFVSG KIVLTGAKVR EEIYQAFEAI YPVLNEFRKS
PCA2 121 -TFSSYEPELF PGLIYRMV---KP KIVLLIFVSG KIVLTGAKVR EEIYQAFEAI YPVLSEFRKS
SPO  121 -TFSSYEPELF PGLIYRMV---KP KVVLLIFVSG KIVLTGAKVR EEIYQAFEAI YPVLSEFRKH
SCE  121 -TFSSYEPELF PGLIYRMV---KP KIVLLIFVSG KIVLTGAKQR EEIYQAFEAI YPVLSEFRKM
GGA  121 -QFSSYEPELF PGLIYRMI---KP RIVLLIFVSG KVVLTGAKVR AEIYEAFENI YPILKGFRKT
XLA  121 -QFSSYEPELF PGLIYRMI---KP RIVLLIFVSG KVVLTGAKVR AEIYEAFENI YPILKGFRKT
MMU  121 -QFSSYEPELF PGLIYRMI---KP RIVLLIFVSG KVVLTGAKVR AEIYEAFENI YPILKGFRKT
MAU  121 -QFSSYEPELF PGLIYRMI---KP RIVLLIFVSG KVVLTGAKVR AEIYEAFENI YPILKGFRKT
HSA  121 -QFSSYEPELF PGLIYRMI---KP RIVLLIFVSG KVVLTGAKVR AEIYEAFENI YPILKGFRKT
TGR  121 -QFSSYEPELF PGLIYRMI---KP RIVLLIFVSG KVVLTGAKVR GEIYEAFENI YPILKGFRKT
PFL  121 -QFSSYEPELF PGLIYRMI---KP RIVLLIFVSG KVVLTGAKVR GEIYEAFENI YPILKGFRKT
ACL  121 -QFCSYEPELF PGLIYRML---QP KIVLLIFVSG KVVLTGAKER TEIYRAFEQI YPVLTQFRKR
ENI  121 -NFSSYEPELF PGLIYRMM---KP KIVLLIFVSG KIVLTGAKVR EEIYQAFELI YPVLSDFRKV
SPU  121 -QFSSYEPELF PGLIYRMV---KP RIVLLIFVSG KVVLTGAKVR QEIYDAFNNI YPILKSFKKT
SFR  121 -QFSSYEPELF PGLIYRMV---KP RIVLLIFVSG KVVLTGAKVR EEIYEAFDNI YPILKSFKKQ
DME  121 -NFSSYEPELF PGLIYRMV---RP RIVLLIFVSG KVVLTGAKVR QEIYDAFDKI FPILKKFKKQ
DDI  121 -SFTNYEPEIF PGLIYKMI---QP KVLLLIFVSG KIVLTGAKVR EYIYEAFENI YPVLSAFKKV
CEL  121 -QFSTYEPELF PGLIYRMV---KP RVVLLIFVSG KVVITGAKTK RDIDEAFGQI YPILKGFKK
OVO  121 -QFSTYEPELF PGLIYRMV---KP RVVLLIFVSG KVVITGAKYK KDIDDAFNQI YPILKGFKK
TTH  121 -KFVQYEPEIF PGKIYREF---NT KIVLLIFVSG KIVLTGAKTR ENINKAFQKI YWVLYNYQKK
DROT 121 -QFSSYEPEMF PGLIYRMV---KP RIVLLIFVNG KVVFTGAKSR KDIMDCLEAI SPILLSFRKT
EHI  121 -AFCQYEPEVF PGLVYRMA---SP KVTLLVFSTG KVVLTGAKDE ESLNLAYKNI YPILLANRKE
PFA  121 -EYCNYEPELF AGLVYRYKPTSNL KSVILIFVSG KIIITGCKSV NKLYTVFQDI YNVLIQYKN
SSH  121 -N-NMYEPEQF PGLIFRMD---DP RVVLLIFSSG KMVITGAKRE DEVSKAVKRI FDKLAELDCV
SAC  121 -N-NMYEPEQF PGLIYRMD---EP RVVLLIFSSG KMVITGAKRE DEVHKAVKKI FDKLVELDCV
HASA 121 LEQIEYEPEQF PGLVYRLD---DP DVVVLLFGSG KLVITGGQNP DEAEQALAHV QDRLTELGLL
MJA  121 -G-TEYEPEQF PGLVYRLD---DP KVVVLIFGSG KVVITGLKSE EDAKRALKKI LDTIKEVQEL
PWO  121 -N-CEYEPEQF PGVIYRVK---EP KSVILLFSSG KIVCSGAKSE ADAWEAVRKL LRELDKYGLL
PSP  121 -N-CEYEPEQF PGVIYRVK---EP RAVILLFSSG KIVCSGAKSE QDAWEAVKKL LRELEKYGLI
TCE  121 -N-CEYEPEQF PGVIYRVK---EP RAVILLFSSG KIVCSGAKSE HDAWEAVRKL LRELEKYDLI
             |-|       |---|          |---|    |-----|  |-----------------|
             S2'       S3'            S4'       S5'           H2'
```

N-terminal residues indicated in parenthesis. Residue numbering (1 to 180) is defined by this table. (-) indicate insertions/deletions.

The secondary structure assignment is indicated below the last sequence in the alignment; 100% conserved residues are shown above the corresponding position. DNA: * residues contacting (within 5Å) DNA; ! = cis proline.

mutants: # DNA specificity relaxation (Arndt 1992; Arndt 1994; Strubin 1992) ; @ DNA binding impaired (Arndt 1992; Bryant 1996; Cormack 1992; Poon 1993; Reddy 1991; Yamamoto 1992); $ DNA binding not rescued by TFIIA and TFIIB (Bryant 1996).

While the N-terminal domain is highly divergent, both in size and in sequence, the DNA binding domain of TBP consists of two imperfect direct repeats. The archaeal sequences show the greatest symmetry, lending support to the hypothesis that this protein arose from gene duplication and fusion before archaea and eukaryotes diverged (Marsh 1994). The largest insertions-deletions have occurred in the loop connecting helix 2 and strand 1', which is the "seam" between the two repeats. Insertions/deletions have also occurred in the loops connecting helix 1' and strand 2', and strand 3' and strand 4'.

Twenty-one residues have been 100% conserved throughout evolution. These residues are marked in the alignment in the row starting with "100%". The relevance of these residues will be discussed below.

1.1.2 Structural Biology

1.1.2.1 Structure of free TBP and TBP/DNA complexes

Crystallographic determinations of the structures of DNA binding domain of TBP-from archeabacteria (PWO) (DeDecker 1996), yeast (SCE) (Chasman 1993), and plants (ATH2) (Nikolov 1994; Nikolov 1992) reveal the same architecture: a molecular saddle with a near twofold symmetry axis, reflecting the symmetry in the amino acid sequence (Figure 1.1). The structure is composed of an eight stranded β-sheet forming the seat of the saddle (S1(1') and S3-S5(3'-5')), and four α-helices, two on the top of the sheet (H2 and H2') and one at each end (H1 and H1'). S2(2') and the loop connecting it to S3(3')

form hairpins at the ends of the sheet, constituting the stirrups of the saddle. The underside of the saddle is the DNA binding site. In all observed cases, TBP crystallizes as a dimer, where the C-terminal stirrup interacts with the hydrophobic DNA binding surface of the dimerization partner.

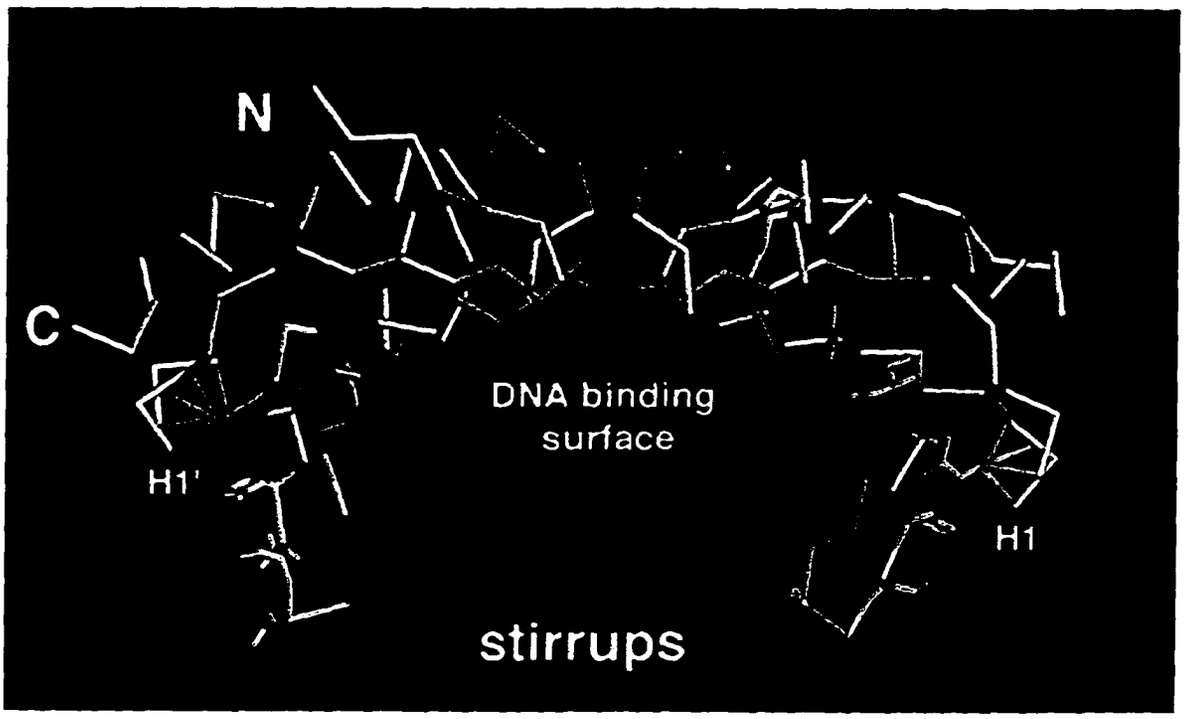

Figure 1.1 TBP architecture. Cα trace (in green) of the DNA binding domain of ATH2 (1VOK, (Nikolov 1994)). The N- and C-termini, helices 1 and 1', and the stirrups are indicated. Yellow: 100% conserved residues - clockwise: E128, P127, E126, Y125, G132, G150, G156, P49, I55, G59, L22, Y34, and P36. Red: 100% conserved residues involved in DNA binding - clockwise: F130, F147, K151, L145, N9, N99, L54, and F56. (Black and white version: red = darkest gray; green = gray; yellow = light gray).

The structural integrity of this saddle depends on some of the 100% conserved residues. For example, I55 and its symmetry related position I146 (substituted conservatively by V and L in other sequences) form the center of a hydrophobic core above the pairs of F residues that insert in the DNA (see below) and link the C-termini of helices 2 and 2' to the rest of the protein. The two stirrups are held extended by the stacking of an aliphatic side chain (L22/L112 — almost 100% conserved, only ACL has M instead) on Y34/Y125, which stack on P36/P127. The C-terminal stirrup is also part of the binding surface for TFIIB. There are two cis-P in the structure of TBP, located at the well exposed turn between strands 3(3') and 4(4'). One of them is 100% conserved (P49), while the C-terminal homologue is found substituted by a three residue insertion (PFA) or a T (TTH).

TBP has also been crystallized in complex with DNA (Juo 1996; Kim 1994; Kim 1993; Kim 1993; Nikolov 1996), and in ternary complexes with two other basic transcription factors, TFIIA (Geiger 1996; Tan 1996) and TFIIB (Nikolov 1995). In all the crystallized complexes, TBP binds to eight basepairs in the minor groove of the DNA (Figure 1.2), consistent with the ability of TBP to bind to DNA where AT basepairs were substituted by IC basepairs (Lee 1991; Starr 1991); these basepairs are sterically identical in the minor groove, but differ in the major groove. The bound sequences encode a characteristic TA repeat (Table 1.2). The C-terminal half of the DNA binding domain interacts with the first four basepairs of the TATA box, and the N-terminal half with the last four basepairs.

Table 1.2 TATA sequences crystallized in complexes with TBP

promoter	sequence	NDB ID numbers
mlp	T A T A A A A G	PDT025,PDT032,PDT034
CYC-1	T A T A T A A A	PDT012
CYC-1	T A T A A A A C	PDT036
E4	T A T A T A T A	PDT024
consensus	T A T A @ A @ X	
bp step	1 2 3 4 5 6 7	
bp	1 2 3 4 5 6 7 8	

@ = A or T; NDB = Nucleic Acid Database (Berman 1992)

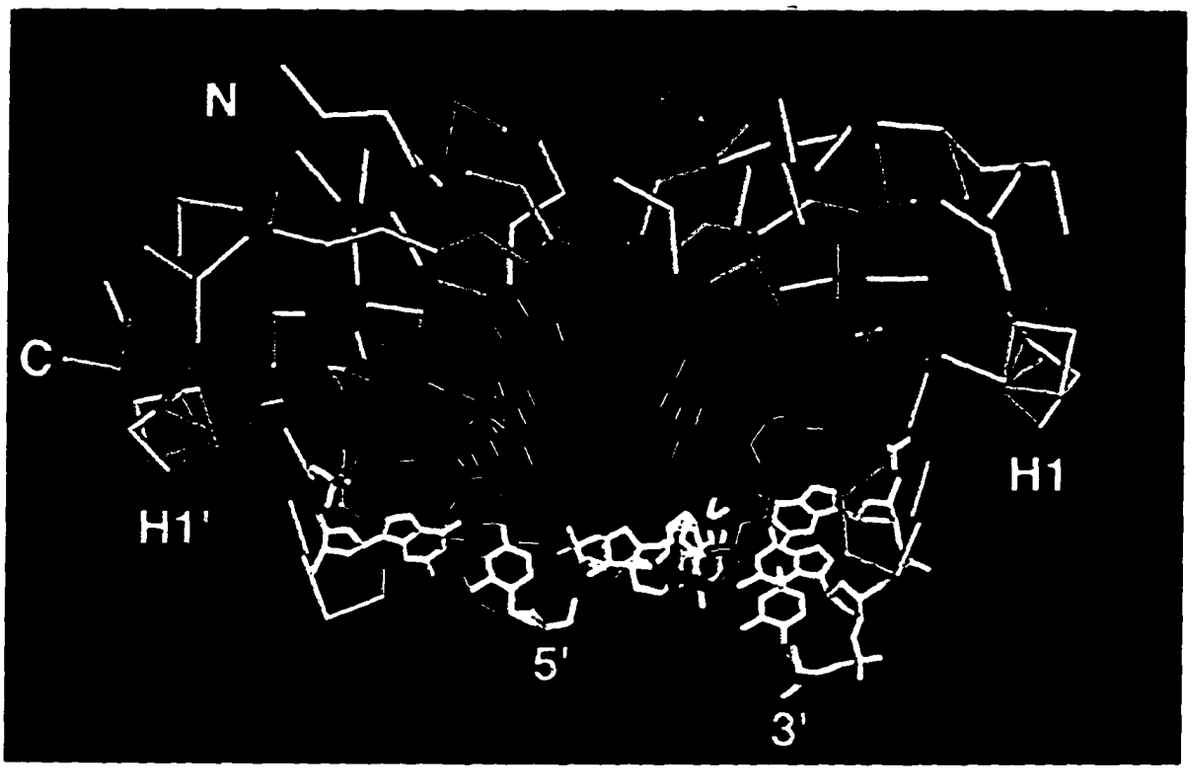

Figure 1.2 Structure of an ATH2/DNA complex (PDT025, (Kim 1994)). Cα trace (green) of the DNA binding domain of ATH2 complexed with the adenovirus 2 major late promoter (blue and white). The TATA box is indicated in blue. The N- and C-termini, helices 1 and 1', and 5' and 3' ends of the sense strand are indicated. Red: 100% conserved residues involved in DNA binding - clockwise: F130, F147, K151, L145, N9, N99, L54, and F56. (Black and white version: blue = darkest gray, red = dark gray; green = gray; white = white).

While TBP does not change its conformation drastically upon binding to DNA, the geometry of the DNA, which is practically identical in all the crystallized complexes, has two kinks that are caused by the partial insertion of two pairs of F residues (F39 and F56; F130 and F156) at the first and last basepair steps of the TATA box (bp steps 1 and 7). These insertions result in a bend of ~90° between the DNA preceding the TATA box and the stretch following it, confirming gel retardation assay measurements of the bending angle (Horikoshi 1992; Starr 1995) and electron microscopy studies (Griffith 1995). The DNA is unwound in these complexes, and the minor groove is widened. The unwinding is compensated by writhe, as suggested by the near-zero linking difference measured by Lorch and Kornberg (Lorch 1993). A summary of basepair step geometry parameters for all the available crystal structures is presented in Table 1.3.

Table 1.3 Basepair step geometric parameters for DNA crystallized in complexes with TBP

step	*shift*	*slide*	*rise*	*tilt*	*roll*	*twist*
1	0.2 ± 0.5	-1.9 ± 0.4	4.9 ± 0.4	-0.6 ± 3.3	40.9 ± 4.8	19.8 ± 2.5
2	-1.3 ± 0.5	-1.5 ± 0.1	3.4 ± 0.2	1.7 ± 2.0	16.5 ± 2.8	16.4 ± 1.3
3	0.2 ± 0.3	1.3 ± 0.1	3.3 ± 0.2	3.4 ± 1.8	7.6 ± 2.9	25.5 ± 1.7
4	0.3 ± 0.4	1.2 ± 0.3	3.4 ± 0.3	-1.8 ± 1.1	26.2 ± 3.2	8.1 ± 6.3
5	-0.4 ± 0.3	2.0 ± 0.3	3.5 ± 0.4	1.0 ± 2.5	25.9 ± 4.3	22.4 ± 3.1
6	0.4 ± 0.3	1.2 ± 0.5	3.3 ± 0.3	0.8 ± 4.5	24.2 ± 4.1	22.7 ± 3.7
7	-0.1 ± 0.7	0.8 ± 1.0	5.4 ± 0.6	1.7 ± 4.4	44.6 ± 6.3	22.5 ± 6.1

mean ± 99% confidence interval

The actual geometry of the DNA bound by TBP seems to be relevant for both the stability of the TBP/DNA complex (Starr 1995) and the assembly of the preinitiation complex, since TFIIB binds to DNA both upstream and downstream of the TBP binding site (Nikolov 1995). There is also evidence from solution studies that show that TBP can bind to non-optimal TATA boxes, but that the resulting complex cannot be recognized by TFIIB, and hence is transcriptionally inactive (Bernues 1996).

The underside of the TBP saddle is the binding interface with the DNA (Figure 1.1). This interface is mostly hydrophobic, but up to six H bonds have been identified in the crystal structures. They involve the interactions between T (T64 and T155) and N (N9 and N99) residues and the two central basepairs of the TATA box (bp step 4).

Some of the 100% conserved residues (N9, N99, F56, F130, F147, L54, L145) map to the interface with DNA, explaining the conservation. Notably, F39 is not invariant (PFA has an I at this position), and this could be related to the viability of F39L mutants in SCE (Reddy 1991). Surprisingly, the four V residues contacting base edges (V11, 62, 101 and 153) are not invariant either. They have been substituted by I (11, 101 and 153) in PFA, and V62 is I in DROT, M in PFA and K in MJA. The binding site preference for these proteins is unknown, so a rationalization for these mutations is impossible at present.

Also striking is the substitution pattern for the two symmetry related T residues. T64 is a L in SPO and a V in MJA; T155 is a S in PWO, PSP and TCE.

The pattern suggests a reason for the fact that T64 is not always engaged in H bonding with a DNA base, while T155 always is. This could be one of the instances of asymmetry in the binding surface of TBP, and could be relevant to choosing a particular direction of binding (see below).

1.1.2.2 The protein and its interaction with DNA

The central role of TBP in transcription guarantees a wealth of mutational data, exploring the interface both with DNA and with the rest of the transcription machinery (for a review, see (Hernandez 1993) and (Nikolov 1994)). Only the mutations related to DNA binding will be discussed here.

Table 1.4 shows all the positions that have been mutated resulting in the elimination of DNA binding. The numbering scheme is the one defined by the alignment in Table 1.1 where these positions are marked with an @ above.

Most of these studies were done before any structural information was available, so in some cases the explanation for the lack of DNA binding is unclear, and might be due to misfolding of TBP. For example, mutation of 100% conserved G59 and G150 is lethal to yeast (Arndt 1992); examination of the structure of TBP (Figure 1.1) shows that these Gs are located at a tight turn between strands 4 and 5, and that this turn is closely packed with the loop connecting strand 1 to helix 1. Mutation of the stirrup residues (L22/112, Y34/125, P127) eliminates DNA binding, despite the fact that they do not lie at the interface; they carry one of each of the pairs of the F residues (F39 and F130) that insert into the DNA. Also interesting are the temperature sensitive

mutants by Schultz (Schultz 1992) and Cormack (Cormack 1992). These mutations could probably be accommodated by the structure, but lead to suboptimal packing. While Bryant et al. (Bryant 1996) designed their mutational strategy to target only surface exposed residues, the drastic charge reversals they engineered may destabilize the protein severely, especially at positions L7, E69, E70 and R160, which lie close to one another at the interface between the two lobes of TBP.

Of the mutations that map to the interface with DNA, some are obvious problems of steric clashes or apposition of negative charges near the phosphates of the DNA. Also, elimination of a single positive charge (from R or K residues) impairs binding. More interesting are mutations to A at positions L54, L145 and F147, which suggest that direct contact between the side chains and the DNA base edges is critical for the formation of a stable complex. It would be interesting to see if these mutants are now capable of binding to sites containing guanine; the shorter side chain would now be able to accommodate the exocyclic amino group at position 2 in the purine ring.

Table 1.4 TBP mutants incapable of binding to DNA

TBP	mutation	reference
SCE	**V11E, F39L, R45C, T52K, F56Y, F56L, F(39,56)L V101E, R136C, V143K, F147Y, F147L**	(Reddy 1991)
SCE	L7K, L16K, L20K, L22K, **L54K**, L112K, L115K, L133K, L144K, **L145K**, L154K **K50L, K60L, K67L, K141L, K151L, K158L**	(Yamamoto 1992)
SCE	P5S, I83N	(Schultz 1992)
SCE	T51I, S76N	(Cormack 1992)
SCE	G59V, G59P, G150V, G150P **L54A, L54K, L45D L145A, L145K, L145D**	(Arndt 1992)
SCE	**L129P, F130Q, F130T**	(Poon 1993)
HSA	Y125A, **R136A**	(Tang 1996)
HSA	I3R, L7E, G17R, C18R, D21R, A26E, R30E, N31E, A32E, Y34E, **I43D, R45E,** R47E, **R50E, S58R, G65R,** E69R, E70A, L74E, R77E, K78Q, R81S, L86E, G87E, F88E, K91E, D94R, F95E, D105R, F108E, P109E, H119E, Q120E, Y125E, P127E, **R141E, F147A, S149F, K158E,** R160E, E162R, E165R, Y171E, P172R	(Bryant 1996)

underlined entries map to the core of TBP, and are likely to be misfolded or unstable; entries boldface map to the TBP/DNA interface (residues within 5Å of any DNA atom).

1.1.2.3 The DNA target: the definition of a TATA box

TBP binds the minor groove of AT-rich DNA with a clear sequence preference (the consensus TATA box has the sequence T A T A t/a A t/a X (Breathnach 1981)). This preference has been confirmed *in vivo* (Chen 1988),

with *in vitro* transcription assays (Wobbe 1990), with *in vitro* selection site assays (Wong 1994), and by X-ray crystallography (Table 1.2).

Mutations in TBP that change binding site specificity are marked with a # in the alignment (Table 1.1), and are detailed in Table 1.5. In all cases, a broadening of specificity has occurred, rather than a change to a new set of binding sites. This fact has interesting consequences, in that assuming that all TBP/DNA complexes impose similar geometrical constraints on DNA, the DNA properties of all these sequences appear to be equivalent as far as TBP is concerned.

Table 1.5 TBP mutants with altered DNA recognition properties

SCE mutant	recognized sequences	reference
I134F, V143T, L145V	TATAAA, TGTAAA	(Strubin 1992)
L145F	TATAAA, CATAAA, TATAAG	(Arndt 1994)
L54F	TATAAA, TATAAG	(Arndt 1994)

The triple mutant defined in the first row of Table 1.5 (called m3) has been used in numerous studies (see for example (Bryant 1996; Whitehall 1995)) because it offers the advantage of assaying the expression of a gene the transcription of which is not directed by endogenous, wild type TBP. Assuming that the m3 mutant will bind to DNA in the same fashion as the wild type TBP, the need for the L145V side chain shortening is easy to explain, as this residue lies directly above the C2 position of the adenine in basepair 2, which is where

the exocyclic amino group for the guanine would be; the V143T mutation would allow for the formation of a H bond to the thymine O3' in basepair 3. The I134F mutation is the hardest to understand; this side chain lies over the sugar backbone of basepair 2 adenine and basepair 3 thymine, and it would probably have to rotate to accommodate the phenylalanine ring (Kim 1993).

The Arndt mutations are more complicated to model directly on the available crystal structures. Examination of the SCE/CYC-1 complex (Kim 1993) shows that there is not much space left for the phenylalanine ring on top of the adenine in basepair 1, which is contacted by F147; the same happens at the 3' end of the TATA box. Another indication that these proteins do not recognize DNA in the same fashion as wild type TBP is that L145F shows a different DNase I footprint on DNA, shifted 3 bases upstream (Arndt 1992). L54F is also unstable.

While the overall fold and the residue layout in the underside of TBP are almost perfectly symmetrical, the charge distribution on the molecule is not (Nikolov 1994; Pastor 1995). The C-terminal stirrup is negatively charged (due in part to the two conserved glutamic acids, E126 and E 128), and the rest of the molecule is positively charged. This uneven charge distribution gives a meaning to the two different orientations in which TBP can bind to a DNA molecule, ultimately defining which strand of DNA will be transcribed and which RNA polymerase will be involved (Wang 1996; Wang 1995; Whitehall 1995). TATA boxes are also asymmetrical, showing a stronger consensus at the 5' end than at the 3' end (Breathnach 1981). In all the crystal structures obtained to

date, TBP always is found to bind in the same orientation, with the more negatively charged C-terminal domain binding to the 5' of the TATA box, and this arrangement is congruent with further interactions with TFIIB and TFIIA and the assembly of the preinitiation complex (Coulombe 1994; Leuther 1996; Robert 1996). While TBP can apparently bind in both directions to promoters with sequence TATATAAA or TATAAAAG, the reverse complexes have up to three times shorter lifetimes than the "correct" complexes (Kim 1997), indicating that the two orientations are not completely equivalent.

The TBP mutants with altered DNA binding abilities also have been used to probe whether TBP alone will bind preferentially in one direction, or if it has to be aided by other proteins of the transcription machinery. Strubin et al. (Strubin 1992) only tested mutations in the 5' half of the TATA box, and it has been assumed ever since that their mutant m3 only binds to TGTAAA boxes. While wild type TBP will promote bidirectional RNA polymerase III transcription from a TATAAATA promoter, this mutant only promoted transcription from TGTAAATA, and not from TATAAACA, again suggesting unidirectional binding (Whitehall 1995). The Arndt (Arndt 1994) mutants (Table 1.5) apparently can bind in both orientations, even though L145F was shown to promote transcription only from CATAAA boxes, and not from TATAAG boxes. The current biochemical evidence suggests that TBP alone will bind in both orientations, and that the TBP associated factors (TAFs) (Pugh 1991) and transcription machinery will determine the productive one by association with the flanking sequences (Li 1995; Wang 1996; Whitehall 1995; Xu 1991).

1.1.3 Physical Chemistry

1.1.3.1 Binding affinity and kinetics

TBP binds to TATA containing DNA with nanomolar affinity. The values that have been measured for the equilibrium binding constant are listed in Table 1.6. The spread in the data for the same TATA box could be due to the different TBP concentrations, binding buffers (for instance, TBP seems to be sensitive to glutamate (Brenowitz, personal communication)), the assays used for quantitating binding (footprinting *versus* filter binding, for example), and the different sources for TBP.

It is clear from this table that TBP prefers AT-rich DNA over genomic sequences, with a selectivity of 1000 to 10,000 fold. Coleman et al. (Coleman 1995) argue that this is very poor selectivity, and that TBP binds stably to noncognate sequences as well, an activity that is relevant to TBP recruitment to TATA-less promoters. The worst substrate tested is poly(dG-dC).

The table also shows that the ability of TBP to discriminate between various AT-rich DNA sequences is poor, on the order of 3 to 4 fold. This is reminiscent of the poor sequence discrimination power of other minor groove binding molecules that also prefer AT-rich DNA (see section 1.2 below).

Table 1.6 Equilibrium binding constants for TBP and DNA

TATA box	TBP	equilibrium constant (x 10^{-9} M)	reference
TATAAAAG	SCE	2.0	(Hahn 1989)
TATATATA	SCE	2.0	
TATATAAA	SCE	4.0	
TATTATTTA	SCE	3.0	
poly(dA-dT)	SCE	2.0	
ssDNA	SCE	20	
poly(dI-dC)	SCE	3000	
genomic	SCE	5000	
poly(dG-dC)	SCE	80000	
TATAAAAG	SCE	0.3	(Hoopes 1992)
TATATAA	ACA	1.1	(Wong 1994)
TATAAAA	ACA	3.7	
TATATATA	ACA	1.4	
TATAAATA	ACA	1.6	
TATATATA	SCE	3.1	(Petri 1995)
TATAAAAT	SCE	4.3	(Perez-Howard 1995)
TATAAAAG	HSA	0.5	(Coleman 1995)
genomic	HSA	5000	
TATAAAAG	SCE	6.6	(Parkhurst 1996)

TBP binds to DNA slowly ($t_{1/2}$ in the order of minutes), and comes off even more slowly ($t_{1/2}$ in the order of minutes to hours) (Coleman 1995; Hoopes 1992; Kim 1997; Perez-Howard 1995; Petri 1995). The slow off rate could be

biologically relevant, given that TBP remains bound at the promoter after the RNA polymerase starts the elongation cycle, thereby being able to participate in multiple rounds of transcription initiation (Burley 1996). The slow on rates have been interpreted as evidence for binding in two steps, the second one being rate limiting and corresponding to a molecular rearrangement (Coleman 1995; Coleman 1995; Hoopes 1992). There is still a debate on whether TBP dimerizes in solution or not, making the transition to monomer the time delaying step (Coleman 1995), or whether the slow binding is due to the drastic bending of the DNA (Hoopes 1992). Another proposal has been made, suggesting that binding is a one step process, and that the slow rate can be explained by the low concentration of DNA competent for binding (Parkhurst 1996; Perez-Howard 1995; Petri 1995). In agreement with this hypothesis, pre-bending a TATA box by superhelical stress enhances TBP binding up to 300 fold (Parvin 1995), especially for suboptimal TATA boxes such as that of the IgH promoter.

There is only one report (Petri 1995) of a detailed study of the thermodynamics of SCE TBP binding to the E4 promoter (TATATATA). Brenowitz and collaborators looked at two aspects of binding: the salt dependence and the temperature dependence.

1.1.3.2 Salt dependence of binding

The binding constant for the association reaction decreases as the salt concentration increases, as expected for protein-DNA interactions (Record

1976). From the slope of the plot of the association constant *versus* salt concentration, Brenowitz and collaborators estimate that 3.5 ± 0.2 cations are released from the surface of DNA (Petri 1995). The same analysis performed on the adenovirus 2 major late promoter (mlp: TATAAAAG) also indicates that 3.5 ± 0.2 cations are displaced from DNA (Brenowitz, personal communication). Such a small number is surprising in view of the fact that there are six lysines and three arginines close to the phosphates of the TATA box in the ATH2/mlp complex, and suggests that these side chains form very unstable salt bridges with the phosphates, if at all.

1.1.3.3 Heat capacity change upon binding

The thermodynamic analysis of the binding reaction (Petri 1995) showed the hallmarks of specific DNA binding, as defined by Spolar and Record (Spolar 1994): the free energy of binding is almost temperature-insensitive, but both the enthalpy and entropy are strong functions of temperature with an optimal temperature for binding (at 30° C, the optimal growing temperature for yeast). In fact, the calculated heat capacity change (ΔCp) is amongst the largest measured so far (-3.5 kcal mol^{-1} K^{-1}) for protein-DNA complexes.

The ΔCp has been interpreted as resulting from the exclusion of water molecules from hydrophobic surfaces (Spolar 1989), with the corresponding decrease in Cp for the water; this amounts to saying that the hydrophobic effect is the main driving force in complex formation. In the case of the TBP-DNA complex, the ΔCp is so large that the change in surface area can only account

for ≈ 20 % of the ΔCp (Kim 1994), indicating that processes other than dehydration are largely responsible for complex formation. A further indication that occlusion of solvent accessible surface is not the best model to account for the change in heat capacity is that ΔCp appears to be DNA sequence dependent (Brenowitz, personal communication). The nature of these processes remains a mystery so far, but Sturtevant (Sturtevant 1977) proposed other sources for the ΔCp: formation and dissociation of H bonds and salt bridges, and changes in the vibrational frequencies of the macromolecules; these are amenable to study with the kind of methods used in this thesis.

1.2 Sequence Specificity in the Minor Groove

TBP targets the minor groove of AT-rich DNA, one of the surface combinations in DNA with the least information content given the stereochemical similarity of AT and TA basepairs in the minor groove side (Seeman 1976). Numerous studies by Dervan's group have shown that it is possible to design specific ligands for the minor groove of the DNA, as long as there are GC basepairs in the target sequence (see (Geierstanger 1995) and (Wemmer 1997) for reviews); the exocyclic amino group of the guanine is the main distinguishable feature in an otherwise smooth minor groove floor. As a rule, small molecules such as netropsin cannot tell AT apart from TA basepairs efficiently. An interesting report by Abu-Daya et al. (Abu Daya 1995) shows that Hoechst 33258, a thin, crescent shaped molecule, will bind AATT 50 fold better

than TATA, and propose that the selection is for DNA sequences with narrow minor grooves, precisely the opposite of what TBP produces on DNA. Actually, molecules like netropsin and distamycin are inhibitors of transcription because they compete with TBP for AT-rich binding sites (Chiang 1994). Dervan's best design can produce a 10 fold preference for TTT over ATA, for a hairpin molecule that binds in the 2:1 mode (White 1996); these molecules prompted Dickerson and collaborators to propose that TBP is a super-lexitropsin (Juo 1996).

The list of proteins that bind predominantly in the minor groove for which there is detailed structural information includes TBP, HMG domains -SRY and LEF-1- (Love 1995; Werner 1995), DNase I (Suck 1988) and IHF (Rice 1996). All these proteins impose a major deformation on the DNA upon binding (Werner 1997; Werner 1996), through the partial intercalation of bulky side chains (F, M, I, P) at TA, AA and AG steps. Also, aromatic rings (F,Y) and aliphatic side chains are placed close to riboses to help open the minor groove. The absence of amino groups at the purine C2 position is monitored by hydrophobic side chains (M, L, I, V). The motif of intercalating aliphatic side chains between basepair steps is also shared by the LacI (Lewis 1996) and PurR (Schumacher 1994) repressors, but these proteins read both grooves of the DNA, and will not be discussed further.

The sequence specificity displayed by these minor groove binding proteins is not remarkable. DNase I will bind and cut DNA sequences with a wide minor groove, but is fairly non-specific, a fact easy to explain from the

dearth of contacts between the protein and the edges of the DNA bases. SRY and LEF-1 are the most specific, and bind to sites that contain GC base pairs, reading them through direct H bonding to the amino group from serine, glutamine and asparagine side chains. SRY actually provides what could be the first example of specific readout of an AT basepair: a tyrosine side chain is stacked against it such that there is an adenine N3 - aromatic interaction, close contact to the hydrogen at the adenine C2, and H bonding to the thymine O2. Transversion of this basepair eliminates binding, suggesting that this is a directional interaction, with the caveat that the transversion also might alter basepair stacking and hence the cost for bending the DNA.

1.3 Statement of the problem: a strong case for indirect readout

In summary, it is still unclear how it is that TBP choses its binding sites. A direct readout approach is untenable, given the stereochemical equivalence of AT and TA basepairs in the minor groove. This opens the possibility for more subtle characteristics of the different DNA sequences to become specificity determinants.

Part of the recognition mechanism of the proteins that distort DNA is a certain predisposition of the specific sequence to acquire the conformation imposed by the protein. This proposition makes sense energetically, because the cost of deformation (Drapper 1993) would be less for the cognate sequences, thus aiding the selection of the binding site. Because TBP distorts

the DNA, it is likely that sequence dependent DNA bendability plays an important role in determining the observed sequence specificity(Harrington 1994). Two manifestations of DNA bendability can be imagined that are not mutually exclusive: One is that certain basepair steps may have an average geometry that is biased towards the conformation induced by TBP. This is the same idea as Calladine's "static wedges" (Calladine 1986). The other option is that even for basepair steps with a structure that is straight on average, some might make more frequent transient excursions (Hagerman 1992) than others into the distorted conformation imposed by TBP, and hence offer a transient "prepared state" for TBP binding.

DNA sequence dependent bendabilities might also explain the directionality of binding, which was proposed to be due to a matching of opposite flexibilities between TBP and the TATA box (Kim 1993): the C-terminal half was found to have smaller B factors than the N-terminal half (suggesting greater rigidity), and alternating TA sequences are assumed to be more flexible than A tracts.

The contributions from the environment (water and counterions) are reflected in the salt dependence of binding and the huge decrease in heat capacity upon binding. Sturtevant (Sturtevant 1977) suggested processes that would contribute to ΔCp: the hydrophobic effect; the formation or dissociation of ion pairs; the formation or breaking of hydrogen bonds, and a change in the vibrational frequencies of the protein and/or the DNA, from the free to the complexed state.

In an attempt to understand at the molecular level the mechanisms that might be used by TBP to select its binding sites, microcanonical (NVE) molecular dynamics (MD) simulations were carried out to explore the conformations of double stranded DNA oligomers that contain TATA boxes, free TBP, and a TBP/DNA complex, all in aqueous solution. The sequences (identified by the coding strand only in Table 2.3) were chosen to include three known binding sites for wild type TBP (Wong 1994) (**mlp**, **at** and **6t**), two binding sites for mutant TBPs (Arndt 1992; Arndt 1994) (**2c** and **7g**), one inverted TATA box (**r28**), and a sequence that is not bound by TBP and serves as a negative control (**gc**). An inosine variant of **mlp** was included to test the equivalence of IC and AT basepairs (**i**).

The analysis of the simulations was geared towards measuring properties that relate to DNA bendability, protein flexibility, protein and DNA hydration, and the condensation of counterions. Whenever possible, a direct comparison to experimental data was performed, and showed that the simulation protocol is capable of reproducing known structural and dynamic properties of the DNA, protein and water. While the direct calculation of heat capacity changes requires simulations at various temperatures and under constant pressure, the simulations presented here allow for testing the molecular implications of such change, such as alterations in the water - water interaction around the macromolecules. Most significantly, the DNA sequence dependence of these properties has allowed for a rationalization of the binding preferences of TBP.

2 Materials and Methods

2.1 TBP sequences from The National Center for Biotechnology Information (NCBI http://www.ncbi.nlm.nih.gov/)

Table 2.1 Available TBP sequences at NCBI

NCBI #	SPECIES	ABBREVIATION
135626	Arabidopsis thaliana 2	ATH2
135627	Arabidopsis thaliana 1	ATH1
974216	Zea mays 1	ZMA1
1729907	Zea mays 2	ZMA2
135640	Solanum tuberosum	STU
1498162	Nicotiana tabacum	NTA
1351224	Mesembryanthemum crystallinum	MCR
1220522	Glycine max	GMA
135641	Triticum aestivum 1	TAE1
417882	Triticum aestivum 2	TAE2
135634	Acanthamoeba castellanii	ACA
540195	Pneumocystis carinii 1	PCA1
540193	Pneumocystis carinii 2	PCA2
135639	Schizosaccharomyces pombe	SPO
135643	Saccharomyces cerevisiae	SCE
2145310	Gallus gallus	GGA
135642	Xenopus laevis	XLA
135638	Mus musculus	MMU
1729911	Mesocricetus auratus	MAU
37066	Homo sapiens	HSA
1079365	Trimeresurus gramineus	TGR
1483195	Protobothrops flavoviridis	PFL
1174643	Acetabularia cliftonii	ACL
887880	Emericella nidulans	ENI
1840134	Strongylocentrotus purpuratus	SPU
1729912	Spodoptera frugiperda	SFR
135636	Drosophila melanogaster	DME
135635	Dictyostelium discoideum	DDI
417896	Caenorhabditis elegans	CEL
294034	Onchocerca volvulus	OVO
482247	Tetrahymena thermophila	TTH
11136	Drosophila melanogaster TRF	DROT
1729910	Entamoeba histolytica	EHI
417904	Plasmodium falciparum	PFA
1419209	Sulfolobus acidocaldarius	SAC
1361936	Sulfolobus shibatae	SSH
1070345	Halobacterium salinarium	HASA
2129299	Methanococcus jannaschii	MJA
2129423	Pyrococcus woesei	PWO
2129420	Pyrococcus sp.	PSP
498255	Thermococcus celer	TCE

2.2 Structures from NDB (Berman 1992) and PDB (Bernstein 1977)

Table 2.2 DNA oligomers used for validation of the simulations

SEQUENCE (NDB or PDB)	RESOLUTION (or NMR)	SEQUENCE (5′ - 3′)
adh008	1.8	GCCCGGGC
adh010	1.8	GGTATACC
adh026	1.7	GGGCGCCC
adh027	2.0	GGGCGCCC
adh033	1.5	ATGCGCAT
adh038	1.4	GTGTACAC
adh039	1.4	GTGTACAC
adh047	1.64	GTGCGCAC
adh070	1.9	ACGTACGT
adj049	1.65	CCCGGCCGGG
adj050	1.7	GCGGGCCCGC
adj051	1.8	GCGGGCCCGC
adj067	1.9	ACCGGCCGGT
adj069	2.0	CCGGGCCCGG
adj075	1.9	GCACGCGTGC
adl025	2.0	CCCCCGCGGGGG
bdf068	1.9	CTCGAG
bdj017	1.6	CCAGGCCTGG
bdj019	1.4	CCAACGTTGG
bdj025	1.5	CGATCGATCG
bdj031	1.5	CGATTAATCG
bdj036	1.7	CGATATATCG
bdjb44	1.3	CCAACITTGG
bdj051	2.0	CATGGCCATG
bdj052	1.9	CCAAGCTTGG
bdj060	1.7	CTCTCGAGAG
bdj081	1.9	CAAAGAAAAG
bdl020	1.9	CGCGAATTCGCG
1buf	NMR	CAATTG
1uqa	NMR	CATATG
1uqb	NMR	CAGCTG
1uqc	NMR	CACGTG
1uqd	NMR	CGATCG
1uqe	NMR	CGTACG
1uqf	NMR	CGGCCG
1uqg	NMR	CGCGCG
1d42	NMR	GTATATAC
1d70	NMR	GTATAATG
1d18	NMR	CATGCATG
1d19	NMR	GTACGTAC
1d20	NMR	TCTATCACCG
142d	NMR	AGCTTGCCTTGAG

The coordinates for free TBP were obtained from PDB entries 1VOK (ATH2), 1TBP (SCE) and 1PCZ (PWO). TBP/DNA complexes were obtained from NDB entries PDT025, PDT032, PDT034, PDT012, PDT036 and PDT024.

2.3 Assembly of the Simulation Systems

The simulation cell contents are detailed in Table 2.3 below.

For the pure water simulation, a 26 Å cubic box was generated in InsightII (Biosym Technologies 1993). The **sod3** system was assembled by placing three sodium ions along a line, each separated by 5Å, solvating them in InsightII (Biosym Technologies 1993) and trimming the water layer with Simulaid (Mezei 1997) to yield a hexagonal prism.

The free DNA dodecamers were built with B-DNA conformation using QUANTA (Molecular Simulations Inc. 1992); the 5' phosphate groups at the end of the strands were removed. The charge of all the DNA molecules was neutralized by adding 22 sodium ions, positioned originally at 5Å from the P atom along the O-P-O bisector.

ath includes the 180 C-terminal residues from ATH2 (starting at S19 in the original structure), taken from the copy in PDB entry 1VOK (Nikolov 1994) with the smallest B-factors. The three internal water molecules were included in the simulation. Hydrogen atoms were added in CHARMM23 with the HBUILD module. **athdna** includes the 180 C-terminal residues from ATH2 (starting at S19 in the original structure PDT025 (Kim 1994)), and a bound DNA

dodecamer with sequence identical to that of **mlp** (the first and last basepairs of the original DNA were deleted). The three internal water molecules were also included. Hydrogen atoms were added in CHARMM23 with the HBUILD module. 22 sodium ions were added, positioned originally at 3Å from the P atom along the O-P-O bisector. The optimal orientation in the simulation cell was obtained with Simulaid (Mezei 1997).

The macromolecules were surrounded by a 14Å water layer in InsightII (Biosym Technologies 1993) and the water layer was trimmed with Simulaid (Mezei 1997) to yield a hexagonal prism of the appropriate dimensions.

Table 2.3 Simulated systems

name	system	TIP3	Na^+	$[Na^+]$	ρ	cell	side	length
	pure water	588	0	0	1.00	c	26	26
sod3	Na^+ solution	480	3	.34	1.02	h	14	28
mlp	CTATAAAAGGGC	3401	22	.36	1.07	h	24	72
mlp-I	CTATAAAAGGGC	3401	22	.36	1.07	h	24	72
2c	CCATAAAAGGGC	3401	22	.36	1.07	h	24	72
6t	CTATATAAGGGC	3434	22	.36	1.08	h	24	72
7g	CTATAAGAGGGC	3437	22	.36	1.08	h	24	72
r28	CTTTTATAGGGC	3425	22	.36	1.08	h	24	72
i	CTITIIIIGGGC	3422	22	.36	1.08	h	24	72
at	ATATATATATAT	3498	22	.35	1.10	h	24	72
gc	GCGCGCGCGCGC	3459	22	.35	1.09	h	24	72
ath	ATH2 TBP	5965	0	0	1.05	h	30	87
athdna	ATH2 TBP + mlp	6165	22	.20	1.07	h	36	64

only the sequence of the coding strand is shown for the double stranded DNA molecules. TIP3: number of explicit water molecules included in the simulation; Na^+: number of sodium atoms added for electroneutrality; $[Na^+]$ in moles/liter; ρ: density in g/cm^3; cell: c=cubic, h=hexagonal prism. Side and length of the cell in Å.

2.4 Molecular Dynamics Protocol

The simulations were done with the CHARMM simulation package (Brooks 1983) and the CHARMM23 potential (MacKerell 1995) (the parameters for inosine were kindly provided by Dr. MacKerell), in the NVE ensemble, using periodic boundary conditions. SHAKE was applied to all hydrogen-containing bonds. A spherical cutoff of 13Å with a switching function for the van der Waals term and a shifting function for the electrostatic term were used. Except for **mlp-I**, which was included to test for the equivalence of measured properties for one particular sequence in two independent simulations (as in (Auffinger 1995; Auffinger 1996)), all the runs used the Verlet integrator, and the $\sqrt{2}$ default for assignment of velocities. Equilibration phases were done with a time step of 2 fs (except for **mlp-I**, which was equilibrated with a time step of 1.5 fs); the production phase was done with a time step of 1.5 fs.

The water was equilibrated first (6 ps heating from 0K to 300K or 600K, followed by 30 to 100 ps dynamics), keeping both the macromolecule (DNA or TBP) and sodium ions fixed. For **athdna**, water and sodium ions were equilibrated at 600K for 100 ps. Equilibration at 600K for **ath** and **athdna** was chosen because it allows the water and sodiums to explore more space in less time, without the risk of distorting internal geomety (SHAKE was used for the water molecules throughout the simulations). After energy minimization of the whole system, it was slowly heated to 300K (10 ps), and equilibrated for 30 ps. The production phase lasted from 510 ps to 2080 ps (see Table 2.4).

Table 2.4 Simulation protocol

name	system	I: (ps)	II: (ps)	III: (ns)
	pure water	0	30	2.04
sod3	Na^+ solution	0	30	2.04
mlp	CTATAAAAGGGC	30	30	2.04
mlp-l	CTATAAAAGGGC*	30	30	0.51
2c	CCATAAAAGGGC	30	30	0.51
6t	CTATATAAGGGC	30	30	0.51
7g	CTATAAGAGGGC	30	30	0.51
r28	CTTTTATAGGGC	30	30	0.51
i	CTITIIIIGGGC	30	30	1.02
at	ATATATATATAT	30	30	1.02
gc	GCGCGCGCGCGC	30	30	1.02
ath	ATH2 TBP	40 (600K)	30	1.02
athdna	ATH2 TBP + mlp	100 (600K)	30	1.02

I: number of ps for solvent equilibration at 300K unless indicated otherwise; II: number of ps simulated at 300K before data acquisition; III: ns for data acquisition; *:used the Leapfrog algorithm, $\sqrt{3}$ as the seed for velocity assignment, and a time step of 1.5 fs throughout.

2.5 Analysis

2.5.1 Conformational characterization

The conformational analysis for DNA was carried out with the CURVES algorithm (Lavery 1989; Lavery 1988) implemented in the Dials and Windows package (Ravishanker 1989). Since this algorithm performs a global fit to the DNA axis, the reported angles and displacements depend on the DNA length. To allow for a comparison of the local basepair step geometry between the simulated DNA dodecamers and the NMR and crystal structures of DNA

oligomers, which have different lengths, all the DNA oligomers analyzed in this work were disassembled into their constitutive basepair steps. Non-self-complementary steps were considered in both orientations, to account for the sign reversal in *shift* and *tilt*. Dihedral angles, which are independent of length, were collected on the whole oligomers. Data were collected for all basepairs except those at the ends of the oligomers.

The DNA axes shown in Figure 4.2 were calculated with CURVES from the average structures representing the longest time interval in the simulations with a < 1.8 Å rms difference between the instantaneous structures in that time interval: **mlp** 550-1630 ps, **2c** 265-550 ps, **6t** 280-550 ps, **7g** 210-550 ps, **r28** 250-550 ps, **i** 320-960 ps, **at** 120-490 ps, **gc** 220-550 ps. All the structures were aligned by superimposing the coordinates of all DNA atoms to the starting structure of each simulation.

The conformational analysis for TBP dihedral angles, both from **ath** and **athdna** was done with the PROCURVES algorithm (Sklenar 1989) implemented in the Dials and Windows package (Ravishanker 1989).

Statistical analyses for the DNA basepair step geometrical parameters were done with SAS (SAS Institute Inc. Box 8000, Cary, NC 27511-8000). Averages and standard deviations were calculated from all the geometries in the production phase of the simulations. The same data set was used to calculate the frequency of occurrence of basepair step geometries within the 99% confidence interval defined by the DNA oligomers found in the TBP/DNA

crystal structures or in the **athdna** simulation. As there are different numbers of structures for each basepair step, these frequencies were further scored by a χ^2 test to determine the basepair steps with the highest and the lowest frequency of occurrence.

2.5.2 Diffusion coefficients

The diffusion coefficients for sodium and water (oxygen atom used as tracer) were calculated from non-overlapping 102 ps intervals (1360 time frames). The mean square displacements for each of the intervals (a minimum of 5 and a maximum of 20) were averaged for each time point, and a line was fitted to the average, using only the data points between 5 and 30 ps. From the Einstein-Smoluchowski equation (Atkins 1990), the diffusion coefficient in three dimensions is the limiting slope of this line, divided by six:

$$\langle r^2 \rangle = 6Dt$$

where $\langle r^2 \rangle$ is the mean square displacement, D is the diffusion coefficient and t is time.

2.5.3 Radial distribution functions

For the isotropic pure water and **sod3** systems, all the possible pairwise radial distribution functions were calculated up to a radius of 12Å. The histogram of distances was collected in CHARMM23 (Brooks 1983), and the g(r) was calculated in Kaleidagraph (Abelbeck Software, 1993):

$$g(r) = (v / n_{total}) * (n(r) / 4\pi r^2 dr)$$

where v = volume of a 12Å sphere, n_{total} = total number of pairs found inside a 12Å radius, n(r) = number of pairs found between distances r and r + dr.

The simulations including macromolecules are very anisotropic systems, where the standard definition of radial distribution functions breaks down because of the excluded volume and the lack of spherical symmetry. Mehrotra and Beveridge (Mehrotra 1980) developed a method for analizing the structure and energetics of solvent around complex solutes. This analysis is based on the proximity criterion, which classifies each solvent molecule according to its nearest solute atom. The volume belonging to each atom is alloted by drawing bisectors to each atom pair of the solute, and is estimated numerically by a Monte Carlo method. This method was used recently to characterize the solvation of DNA (Guamieri 1996). The code implemented by Dr. Mihaly Mezei was used to generate radial distribution functions up to a radius of 12Å for macromolecule-sodium and macromolecule-water.

For the macromolecule-sodium analysis, the normalization density was derived from the actual sodium density in the simualtion cell. The analysis was done in the absence of water (a reduced dynamics trajectory file was generated for each system omitting the water molecules). In order to account in an average manner for the effect of the water in screening the interaction, the interaction energies between the sodium and the macromolecules (in Table 6.1) were

calculated assuming a dielectric constant of 80. These energies were not used as absolute numbers, and their ranking is insensitive to the scaling factor included as a dielectric constant.

Experimental data suggest that sodium ions bind to DNA fully hydrated (Black 1994; Collins 1997). Sodium-water radial distribution functions were calculated (in the DNA and **athdna** simulations) and compared to the ones from **sod3** to study the changes in sodium hydration upon condensation to DNA and complex formation.

For the macromolecule-water and water-water analysis, the experimental water density was used for normalization, thus allowing for direct comparison between the different simulations. The interaction energies between water and the solutes (in Table 7.1) were calculated assuming a dielectric constant of 1.

In order to study the perturbation of water structure by the macromolecules, radial distribution functions were calculated for water molecules in a 4Å shell from the center of the atoms most exposed to the surface of the macromolecules and sodium ions. The radial distribution function for the angle formed between the water dipole and the vector connecting the solute atoms to the water oxygen was also calculated to monitor the different effects of the solutes in the organization of water around them.

3 Validation of the Simulations

3.1 Structural Stability

There are no experimentally determined structures against which to compare the free DNA simulations. Hence, structural stability of the DNA molecules was judged from 2D rms plots in which each DNA structure of the simulation (separated 1.5 ps from each other) was superimposed and compared to all the others. Figure 3.1 shows these plots for the nine free DNA simulations and for the DNA molecule in **athdna**. The upper triangle of Figure 3.1.A displays the data for all the heavy atoms, and the lower triangle, the data for the heavy atoms of the eight basepairs constituting the TATA box of **mlp**. Given that the TATA box region follows the behavior of the whole dodecamer, as evidenced by the marked square in this figure, Figures 3.1.B-F contain the information for all the heavy atoms of the other simulated dodecamers; the upper and lower triangles correspond to different simulations, indicated in the axes. In all cases, there is a departure from the initial structure, and **mlp** shows an instance of revisiting a conformation 1 ns later (see the low rms region corresponding to the comparison of structures at 240 ps and 1240 ps). As another measure of stability, the simulations were viewed with the animation module of QUANTA (Molecular Simulations Inc. 1992), and no instances of fraying were observed. This was confirmed by measuring the N1 - N3 distances

for all the basepairs as a function of time (data not shown).

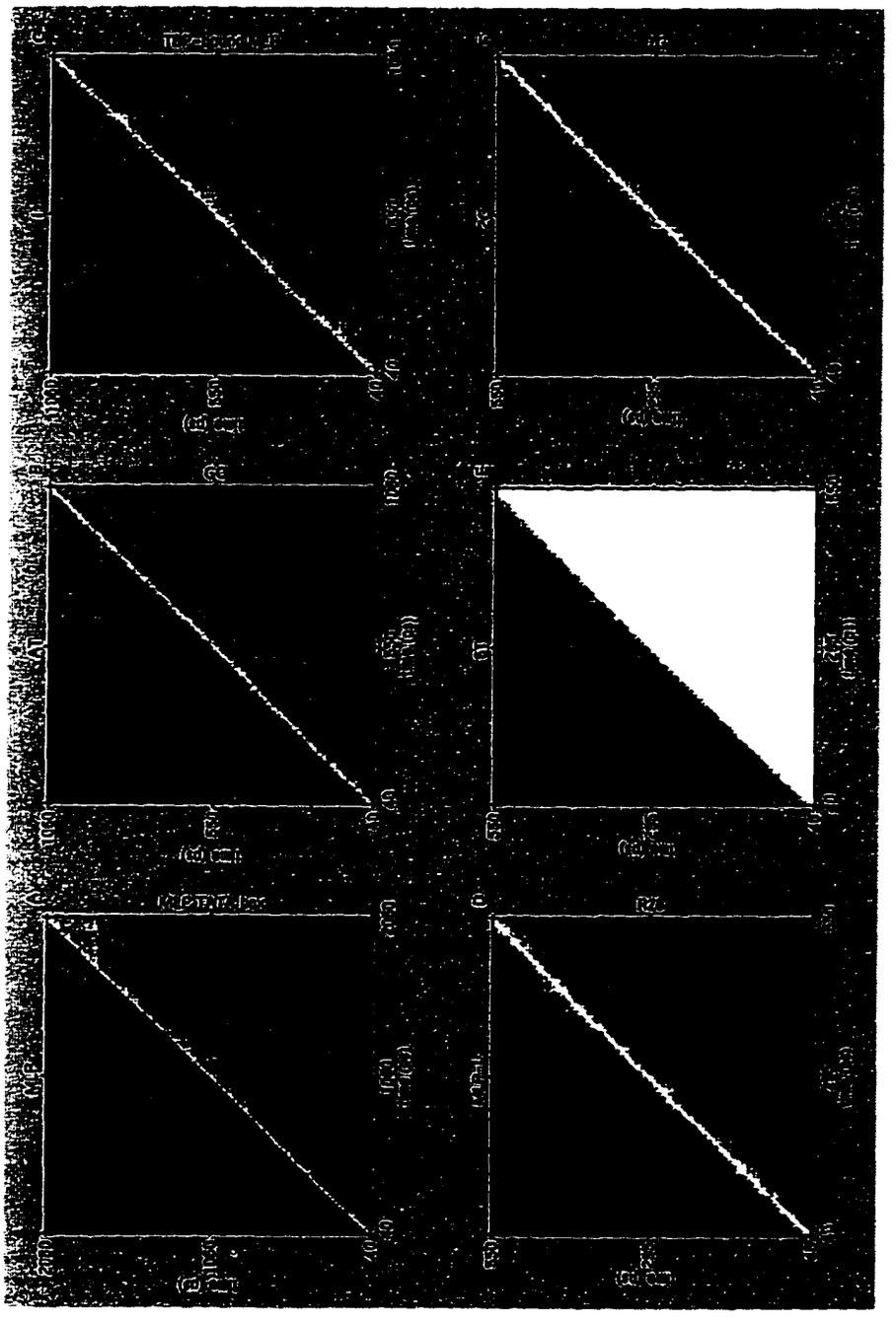

Figure 3.1 Two dimensional rmsd plots for DNA. A: **mlp** heavy atoms (top) and TATA box heavy atoms (bottom). The scale bar (in 0.1Å) is shown on Figure 3.3. B-F: heavy atom rmsd for the whole dodecamers, indicated in the axes.

In the case of the **ath** and **athdna** simulations, the starting points are actual structures. Figure 3.2 shows the rms for the Cα as a function of time for the ATH2/mlp complex and free ATH2, compared to the parent structures. The complex seems to require ~500 ps to equilibrate, and to do so as a whole, judging from the rms for the heavy atoms in the DNA; free TBP shows an oscillatory behavior. From this plot it seems that the conformations of free and bound TBP are within thermal fluctuations from each other. The 2D rms plots for these simulations are shown in Figure 3.3, confirming the information displayed in the 1D plots.

Miaskiewicz and Ornstein (Miaskiewicz 1996) reported simulations for ATH2, both free and complexed to mlp. Contrary to what they find, there is no evidence of a collapse of the free TBP structure in the **ath** simulation presented here. Figure 3.4 shows the time evolution of the distance between the Cα of R38 and L129, located at the tips of the stirrups, for free and bound TBP. In **ath**, this distance is seen to oscillate, never going below 20Å. In **athdna** this distance is constrained by the DNA. Miaskiewicz and Ornstein (Miaskiewicz 1996) did their simulations with the AMBER potential (Pearlman 1995), in a sphere of water, starting from the bound conformation, and did not include the crystallographic internal waters. There are two waters that link together strands 1(1'), strands 5(5') and the N-termini of helices 2(2'), very close to the hinge point identified by them. Perhaps the collapse observed in the Miaskiewicz and Ornstein (Miaskiewicz 1996) simulation is due to the lack of the waters, and not to an inherent property of TBP.

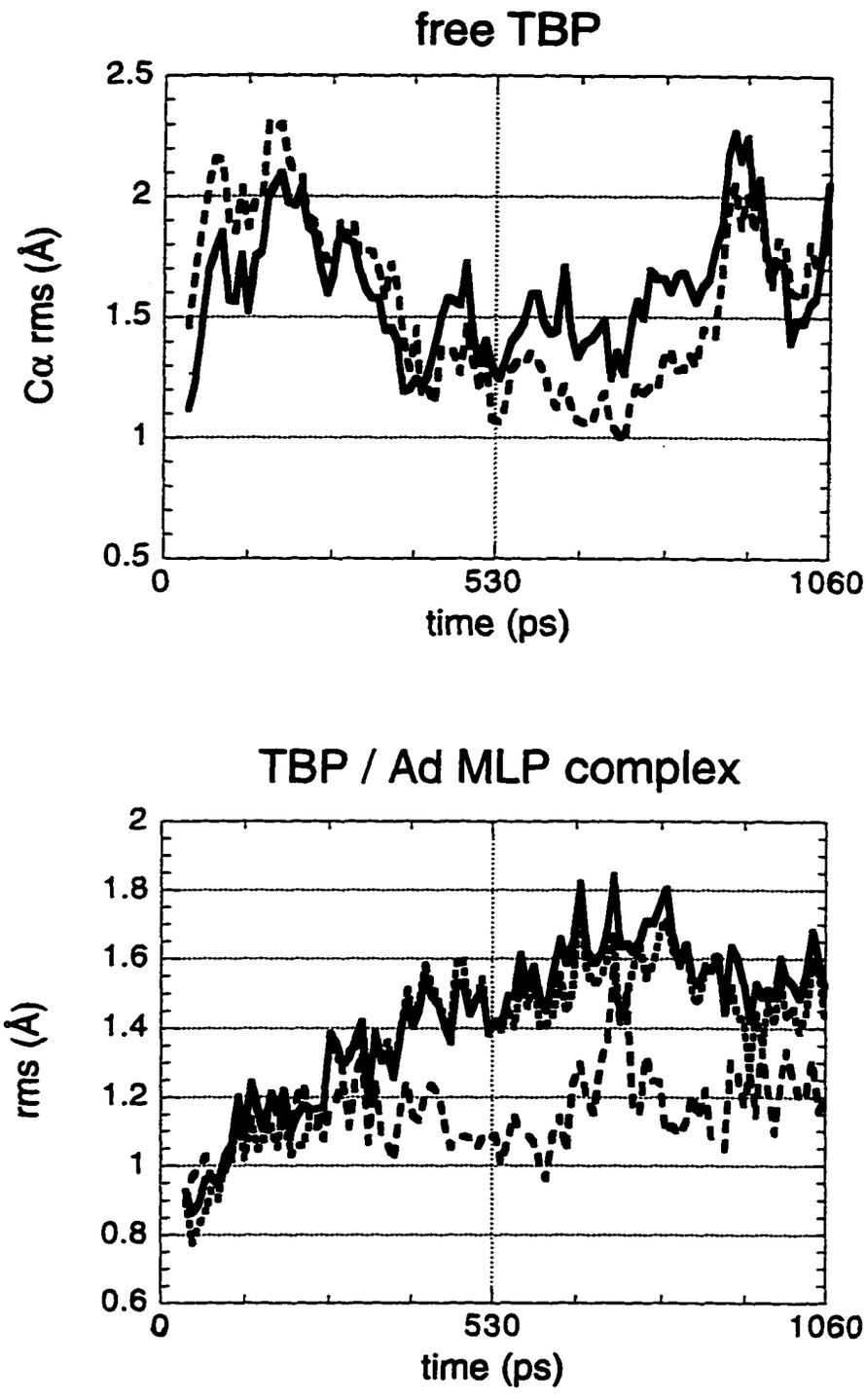

Figure 3.2 Time evolution of the rms to the starting structures for free TBP (top) and its complex with DNA (bottom). Top: solid line: rms to Cα of 1VOK (Nikolov 1994), the starting structure for the **ath** simulation; broken line: rms to Cα of PDT025 (Kim 1994), the starting structure for the **athdna** simulation. Bottom: solid line: rms to Cα and DNA heavy atoms of PDT025; broken line: rms to Cα of PDT025; dotted line: rms to DNA heavy atoms of PDT025.

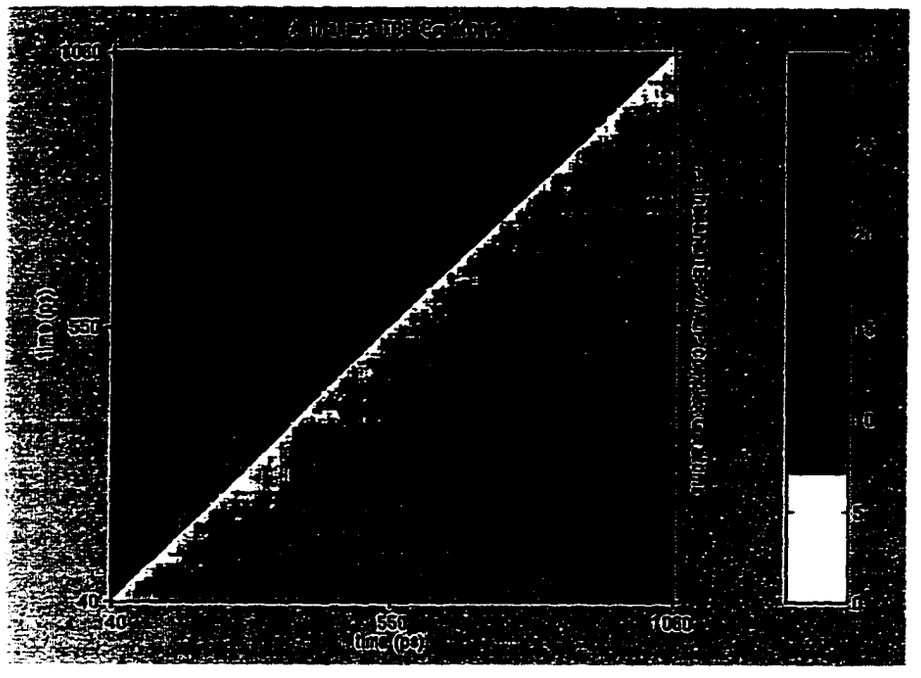

Figure 3.3 Two dimensional rmsd plots for the Cα of free TBP and TBP complexed with DNA. Top: free TBP; bottom, TBP complexed with mlp. The scale bar (in 0.1 Å) is shown on the right.

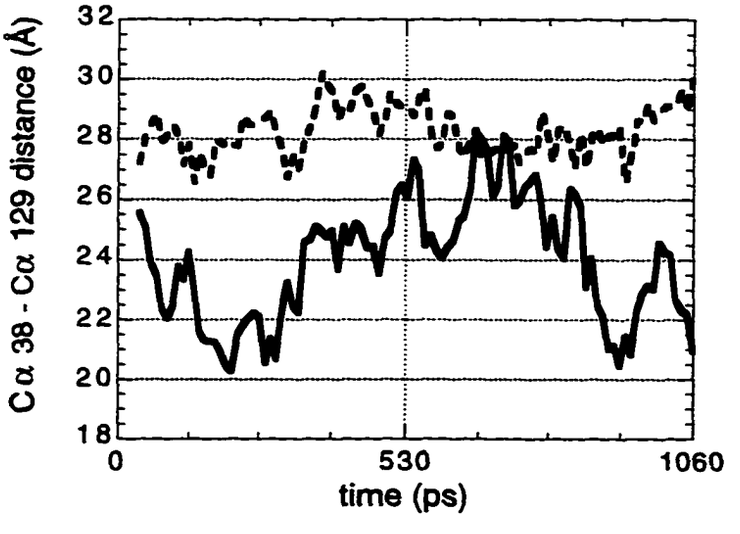

Figure 3.4 Time evolution of the distance between the tips of the TBP stirrups. Solid line: free TBP; broken line: TBP in complex with DNA.

3.2 Comparison to Available Experimentally Determined Structures

To validate the free DNA conformations, the structures generated by CHARMM were compared to those found in high resolution X-ray structures and NMR structures refined using a forcefield (listed in Table 2.2). The compared properties included consecutive phosphorus-phosphorus (P - P) distances, sugar puckers and glycosidic bond torsional angles, and basepair step geometric parameters. To validate the dynamic behavior of TBP and its complex with DNA, the atomic fluctuations of the Cα atoms were compared to the B factors of the Cα atoms of the parent crystal structures. The DNA geometry was compared to the 99% confidence intervals obtained for the available crystal structures (Table 1.3).

3.2.1 Backbone property: consecutive P - P distances

Figure 3.5 shows the distribution of P - P distances: the simulations of free DNA (labeled as CHARMM) produce a major peak at 6.8Å, very close to the center of the distributions for NMR structures and B-DNA. The secondary peak is intermediate between A-DNA and B-DNA characteristic values. In Figure 3.6 the P - P distance distributions for each of the simulations is shown. There are sequence dependent variations in these profiles, and there is no obvious correlation between the length of the simulation and the population of A-DNA like distances.

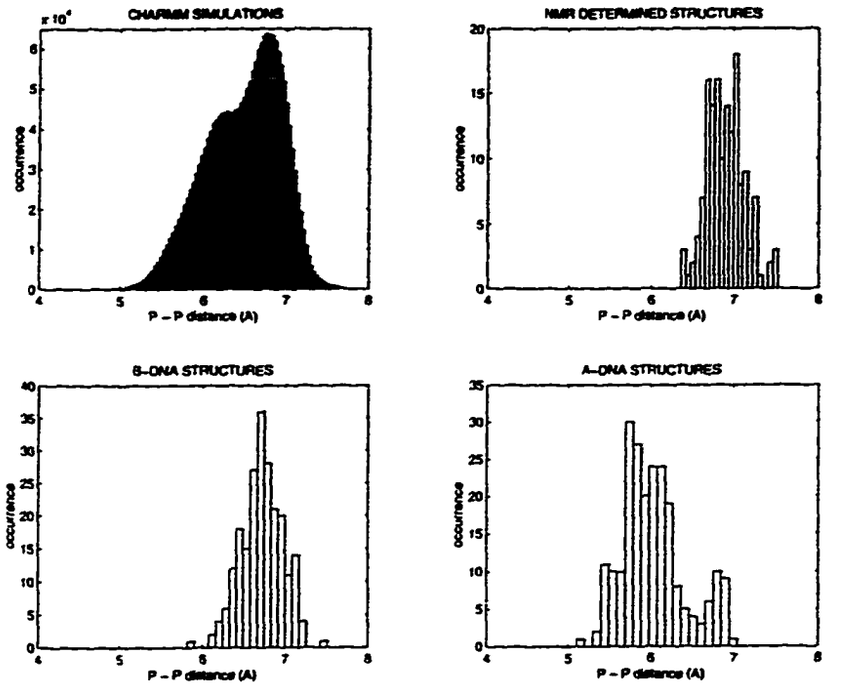

Figure 3.5 Comparison of consecutive P-P distance distributions between the free DNA simulations and experimental data.

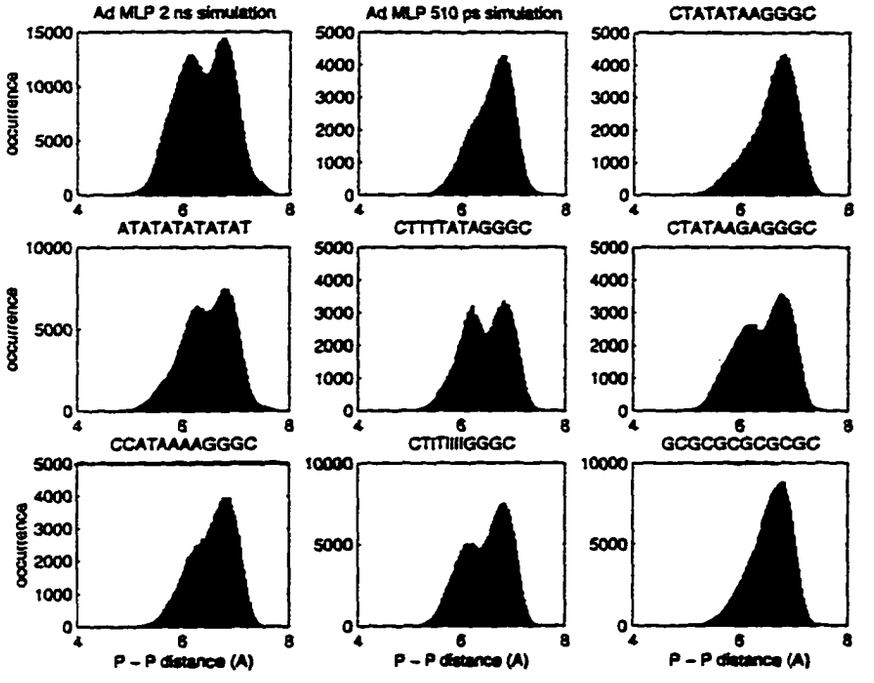

Figure 3.6 Consecutive P-P distance distributions for the free DNA simulations.

3.2.2 Nucleotide properties: sugar pucker and χ

The P - P distances are correlated with the sugar pucker (Saenger 1983). In Table 3.1 are shown the populations for the sugar puckers, classified according to the nucleotide to which they belong and binned in 36° intervals centered at the listed phase. Also listed, in Table 3.2, are the populations for the different simulations, and the values from crystal and NMR structures.

Table 3.1 % Population of the sugar pseudorotation cycle for the simulation data (sample size: 2040000)

Classified by the simulation to which they belong

PHASE	mlp	mlp-l	2c	6t	7g	r28	i	at	gc
$C_{3'}$ endo	40.6	19.3	21.3	20.0	35.6	38.7	30.4	38.0	16.9
$C_{4'}$ exo	3.1	4.7	3.9	2.6	3.3	2.0	2.2	3.0	3.8
$O_{4'}$ endo	9.1	15.5	14.7	12.7	9.8	5.4	8.2	7.9	8.4
$C_{1'}$ exo	24.4	34.6	32.4	35.2	27.5	21.9	27.9	24.7	40.4
$C_{2'}$ endo	14.6	21.3	25.0	26.1	19.6	25.8	24.9	22.5	29.2
$C_{3'}$ exo	0.2	0.3	0.2	0.4	0.2	0.4	0.3	0.6	0.2
$C_{4'}$ endo	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$O_{4'}$ exo	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$C_{1'}$ endo	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$C_{2'}$ exo	8.1	4.2	2.5	3.1	4.1	5.8	6.2	3.4	1.1

Classified by the nucleotide to which they belong

PHASE	A	C	G	I	T	ALL
$C_{3'}$ endo	42.35	13.06	43.70	55.64	22.90	31.19
$C_{4'}$ exo	4.35	1.56	3.88	3.23	2.63	3.13
$O_{4'}$ endo	11.95	6.06	9.63	9.45	9.82	9.56
$C_{1'}$ exo	19.72	42.74	17.97	9.85	37.15	28.99
$C_{2'}$ endo	14.88	34.91	14.69	6.62	25.57	21.94
$C_{3'}$ exo	0.42	0.34	0.27	0.10	0.18	0.30
$C_{4'}$ endo	0.00	0.00	0.00	0.00	0.00	0.00
$O_{4'}$ exo	0.00	0.00	0.00	0.00	0.00	0.00
$C_{1'}$ endo	0.00	0.00	0.00	0.00	0.00	0.00
$C_{2'}$ exo	6.32	1.32	9.85	15.11	1.75	4.88

Overall, the most populated puckers are C3' endo / C2' exo and C1' exo / C2' endo, with a non-negligible population of C4' exo / O4' endo. The O4' exo region of the pseudorotation cycle is practically never visited, in accordance with the energy profile for furanoses (Saenger 1983). These tables show that even for DNA dodecamers with the same composition (**2c** and **7g** for example), the population of the sugar puckers varies with sequence. Given that TBP imposes C3' endo sugars on the bound DNA, this sequence dependent behavior might be part of a preparation for binding (see Chapter 4).

Table 3.2 % Population of the sugar pseudorotation cycle for X-ray and NMR data (sample size: 770)

	nucleotide					structure class		
PHASE	A	C	G	I	T	ADNA	BDNA	NMR
$C_{3'}$ endo	15.22	40.49	42.45	0.0	18.12	83.68	3.26	0.00
$C_{4'}$ exo	0.72	5.67	0.82	0.0	0.00	3.13	1.45	1.94
$O_{4'}$ endo	2.17	5.26	2.45	0.0	5.80	1.04	6.52	4.37
$C_{1'}$ exo	27.54	14.98	10.61	0.0	47.10	1.74	32.61	34.47
$C_{2'}$ endo	43.48	29.96	33.88	100.0	27.54	1.39	47.46	59.22
$C_{3'}$ exo	7.97	2.02	3.27	0.0	0.72	0.69	8.33	0.00
$C_{4'}$ endo	0.00	0.00	0.41	0.0	0.00	0.00	0.36	0.00
$O_{4'}$ exo	0.72	0.00	0.00	0.0	0.00	0.35	0.00	0.00
$C_{1'}$ endo	0.00	0.00	0.82	0.0	0.72	1.04	0.00	0.00
$C_{2'}$ exo	2.17	1.62	5.31	0.0	0.00	6.94	0.00	0.00

The sugar pucker is also correlated to the glycosyl bond angle χ. The 2D histograms for sugar pucker and glycosyl bond torsional angles, in Figure 3.7, indicate both A-DNA and B-DNA like structures, shown by the black isopopulation contours. The gray area represents the wide range of conformations generated by CHARMM. The overlap with experimental data, indicated by the asterisks, is very satisfactory, as is the known correlation between these two parameters (Saenger 1983).

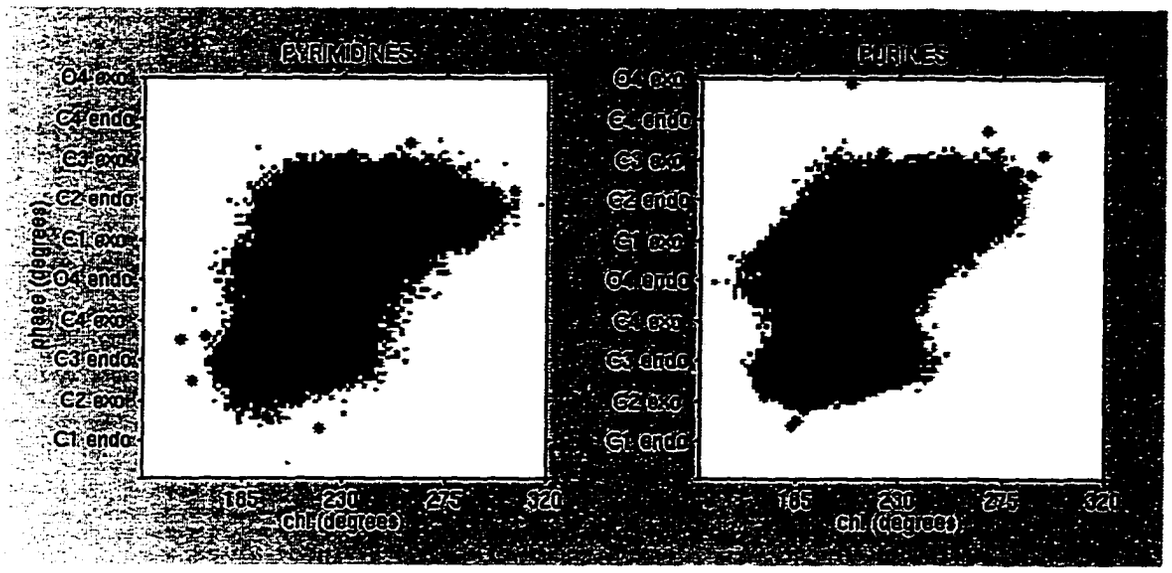

Figure 3.7 Two dimensional distributions for sugar pucker and glycosyl bond torsional angle for the free DNA simulations. Comparison to experimental data (asterisks). Dark gray: configurational area explored by the simulations; black contours: isopopulation contours for the distributions generated by the simulations.

3.2.3 Basepair step geometry

The distributions of the orientation-independent basepair step geometrical parameters are represented in the 1D histograms in Figure 3.8 for the three different kinds of basepair steps (purine-purine = RR, purine-pyrimidine = RY, and pyrimidine-purine = YR). The averages and standard deviations for the six parameters are listed in Table 3.3. In agreement with the

general trends observed experimentally, purine-purine (RR) steps are found to be more rigid than purine-pyrimidine (RY) or pyrimidine-purine (YR) steps (Hogan 1983), YR steps being the most flexible (Chen 1985; Hogan 1983).

Table 3.3 Basepair step geometrical parameters for free DNA dodecamers

	shift	*slide*	*rise*	*tilt*	*roll*	*twist*
all	0.1 ± 0.8	-1.4 ± 0.8	3.4 ± 0.5	0.5 ± 6.6	3.1 ± 12.8	32.0 ± 5.3
RR	0.1 ± 0.7	-1.8 ± 0.7	3.3 ± 0.4	1.0 ± 6.4	1.4 ± 11.5	31.9 ± 4.8
RY	0.0 ± 0.7	-1.0 ± 0.5	2.9 ± 0.4	1.3 ± 5.6	14.4 ± 11.3	29.2 ± 5.4
YR	0.2 ± 1.0	-0.9 ± 0.7	3.7 ± 0.5	1.1 ± 7.3	1.0 ± 13.0	33.6 ± 5.7

mean ± standard deviation

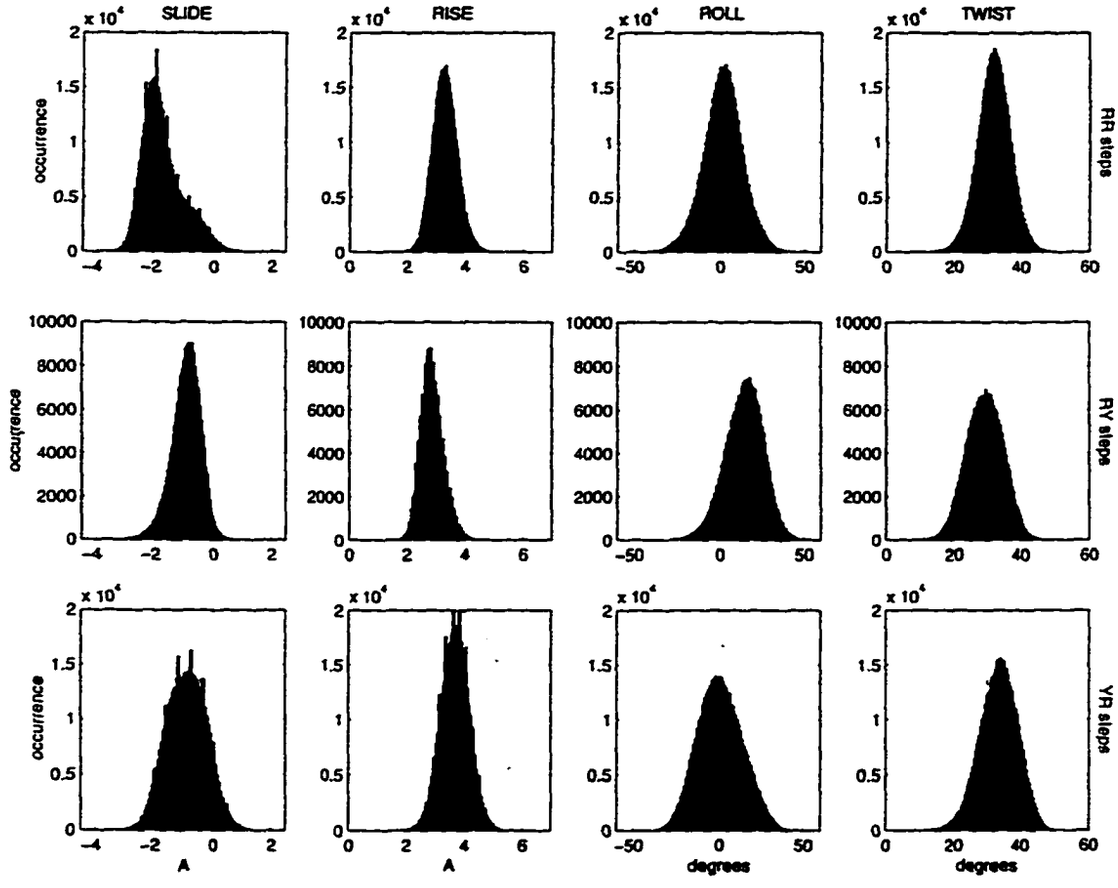

Figure 3.8 Selected sequence dependent basepair step geometrical parameter distributions for the free DNA simulations.

More interesting are the correlations between these parameters . The 2D histograms in Figure 3.9 reveal a satisfactory overlap between the experimental data (asterisks) and the calculated profiles represented by the isopopulation contours in black on the gray background showing the ranges of values from the conformations generated by CHARMM. All the outliers correspond to CA steps. The anisotropy of bending can be seen in Figure 3.9.C, both from the ranges of the *tilt* and *roll* axes, and from the shapes of the distributions. The anisotropy has been explained (Yanagi 1991; Zhurkin 1979) by the fact that *tilt*ing involves the relatively unfavorable stretching and compressing the sugar-phosphate backbone, while *roll*ing can be achieved without much perturbation of the backbone. There is also anisotropy in the *shift* and *slide* motions (figure 3.9.A), manifested in that most of the values explored by *shift* are mainly restricted to ±1 Å, while *slide* has greater excursions from the perfectly straight axis. This is because, in general, *slid*ing improves stacking, while *shift*ing decreases it (Calladine 1982; Suzuki 1996).

There is a clear asymmetry in bending: it is easier to bend by compressing the major groove of the DNA, as shown by the greater incidence of positive values of *roll*. This has been explained by the greater chance of a steric clash between the bases in the minor groove, as opposed to the major groove (Calladine 1982; Gorin 1995). There is also a greater tendency to unwind than to overwind, compared with B-DNA. Inverse correlations are evident between *roll* and *twist* (figure 3.9.D) and *rise* and *roll* (figure 3.9.H), as is the direct correlation between *rise* and *twist* (figure 3.9.I).

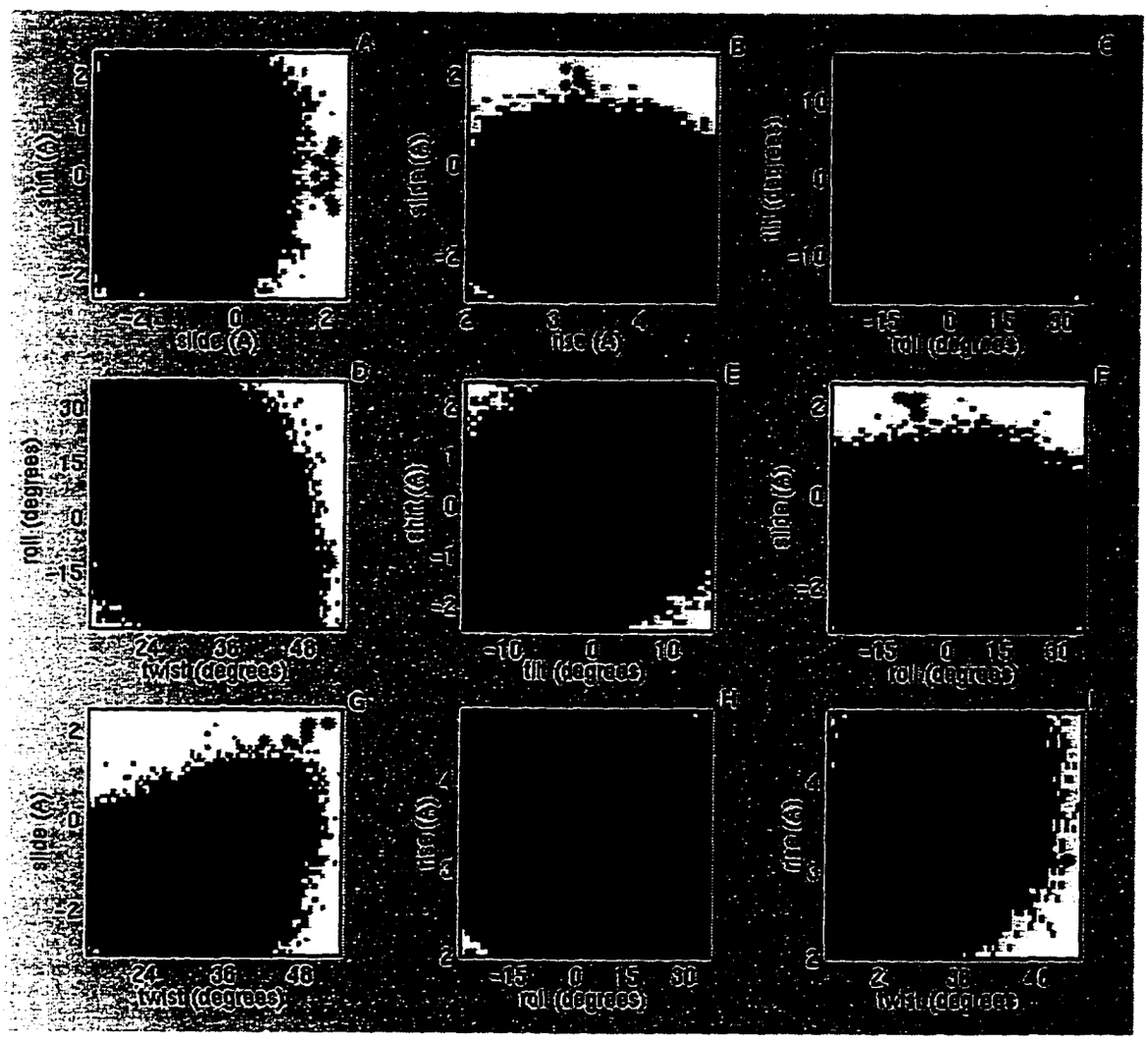

Figure 3.9 Two dimensional distributions for selected basepair step geometrical parameters for the free DNA simulations. Comparison to experimental data (asterisks).Dark gray: configurational area explored by the simulations; black contours: isopopulation contours for the distributions generated by the simulations.

Regarding the DNA molecule in **athdna**, Table 3.4 compiles the averages and standard deviations for the six geometrical parameters describing basepair steps. The most notable differences with the intervals derived from the crystal structures (Table 1.3) are the more normal *shift* at step 2, and more exaggerated *rise* and *roll* at the kink steps (1 and 7). Otherwise, the same trends are observed in the simulated and actual structures, including the return to average values closer to those of free DNA outside the TATA box (steps 8-9).

Table 3.4 Basepair step geometrical parameters for DNA simulated in complex with *A. thaliana* TBP:

step	*shift*	*slide*	*rise*	*tilt*	*roll*	*twist*
1	0.7 ± 0.9	-2.1 ± 0.6	6.1 ± 0.6	0.6 ± 6.2	52.2 ± 12.5	21.2 ± 5.3
2	-0.3 ± 0.6	-1.7 ± 0.4	3.2 ± 0.4	1.4 ± 5.1	13.6 ± 8.4	21.6 ± 4.3
3	0.3 ± 0.5	1.3 ± 0.8	3.5 ± 0.5	4.4 ± 5.2	10.3 ± 9.0	20.3 ± 3.9
4	0.1 ± 0.5	0.9 ± 0.4	3.6 ± 0.5	-4.8 ± 5.1	29.8 ± 8.5	13.9 ± 3.4
5	-0.3 ± 0.5	1.5 ± 0.5	3.1 ± 0.4	1.5 ± 6.2	17.4 ± 7.4	20.7 ± 3.9
6	0.1 ± 0.5	1.4 ± 0.5	3.2 ± 0.5	6.5 ± 5.2	26.0 ± 7.1	29.6 ± 4.2
7	0.7 ± 0.9	-0.1 ± 0.6	6.7 ± 0.6	5.0 ± 5.9	43.3 ± 10.5	22.7 ± 5.5
8	0.0 ± 0.7	-1.9 ± 0.5	2.9 ± 0.4	-1.4 ± 6.6	7.1 ± 10.6	34.1 ± 5.8
9	0.7 ± 0.5	-1.9 ± 0.3	3.2 ± 0.3	2.3 ± 5.9	-13.4 ± 11.0	32.0 ± 4.1

mean ± standard deviation. Note: steps are listed 5' to 3' along the TATA box

All these comparisons show that the CHARMM simulations reproduce both the average properties of free DNA and the observed correlations among these structural parameters. Furthermore, the entry labeled "all" in Table 3.3 defines the thermally accessible range of conformations for general sequence DNA, following the approach of Olson and coworkers (Olson 1995).

3.2.4 TBP Cα B-factors

Having determined that the simulations of TBP and its complex with mlp are stable structurally (section 3.1 above), the dynamic behavior of these simulations was assayed by comparing the Cα fluctuations calculated from structurally stable parts of the trajectories (400 - 910 ps for **ath** and 550 - 1060 ps for **athdna**) to the B-factors for the free and DNA-bound ATH2 TBPs reported in the PDB.

The top of figure 3.10 shows the comparison for free TBP. The agreement in the trends is remarkably good, except for the C-terminal stirrup. This is part of the dimerization interface, which is absent in the simulation, yielding a TBP that is both structurally and dynamically more symmetric. The symmetry observed in the simulation puts in question the hypothesis (Kim 1993) that TBP choses the orientation for binding because a more rigid C-terminal domain would be prefered to bind the more flexible TATA 5' half of the binding site, while the more flexible N-terminal domain would prefer to bind the more rigid AAAG 3' half of the box.

The comparison for DNA-bound TBP is depicted in the bottom of Figure 3.10, and here the trends match without exception. The two stirrups have less mobility, and there is an increase in the mobility of the loops connecting helix 1 and strand 2, and helix 2 and strand 1' .

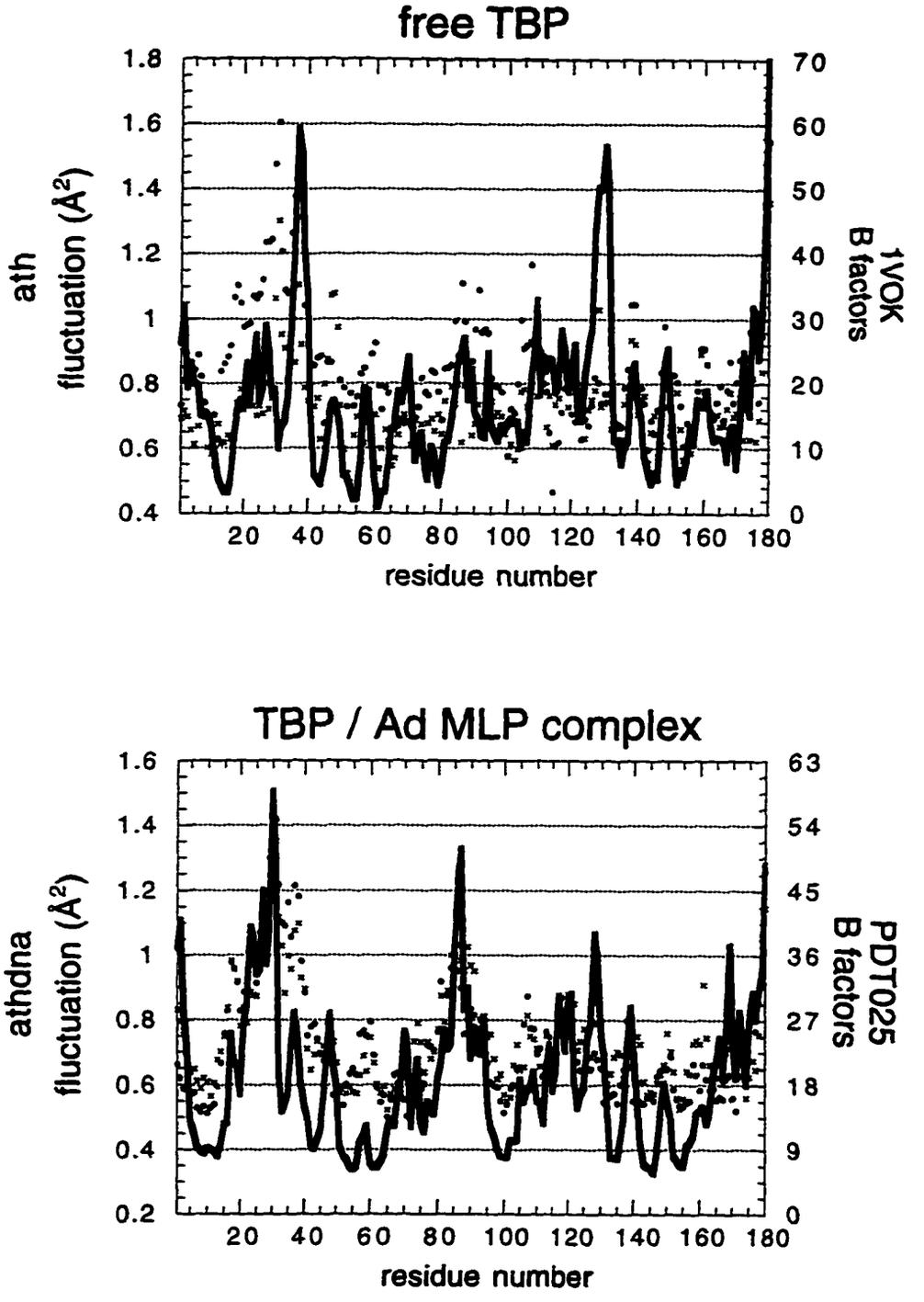

Figure 3.10 Cα fluctuations for free TBP and TBP in complex with DNA. Comparison to crystallographic B factors. Top: solid line: Cα fluctuations for **ath**, x and •: B factors for Cα of both copies of 1VOK (Nikolov 1994). Bottom: solid line: Cα fluctuations for **athdna**, x and •: B factors for Cα of both copies of PDT025 (Kim 1994).

3.3 Equilibration of the counterion distribution

A molecule with the high charge density of DNA requires special attention to the approximations used in the calculations. A particular concern is the effect that a finite cutoff might have on the ionic environment and on the DNA structure itself (Jayaram 1994; Jayaram 1996; Mirzabekov 1979; Strauss 1996).

Figure 3.11 presents the resulting radial distribution functions of the three pairs of charged species in the **mlp** system: sodium-sodium, sodium-phosphorus (as the center of the phosphate group), and phosphorus-phosphorus; these were calculated with CHARMM, and treated as if this were an isotropic system (the position of the peaks in these functions, which is the point of the calculation, is not affected by the normalization). The evidence for problems in handling electrostatics has been shown to be the artificial accumulation of ion pairs exactly at the cutoff distance (Neumann 1980). The shift function makes the interaction zero exactly at the distance of 12Å marked by the vertical line in the plots shown in Figure 3.11. It is clearly evident that there is no accumulation of pairs in any case, suggesting that this cutoff in the CHARMM simulations is sufficiently long so as not to distort the DNA or impose an artificial structure on the ion distribution. These results lead to the conclusion that the cutoff does not distort the DNA (from the P - P rdf), and it does not impose any artifactual structure on the ion distribution around DNA (from the Na^+-P rdf).

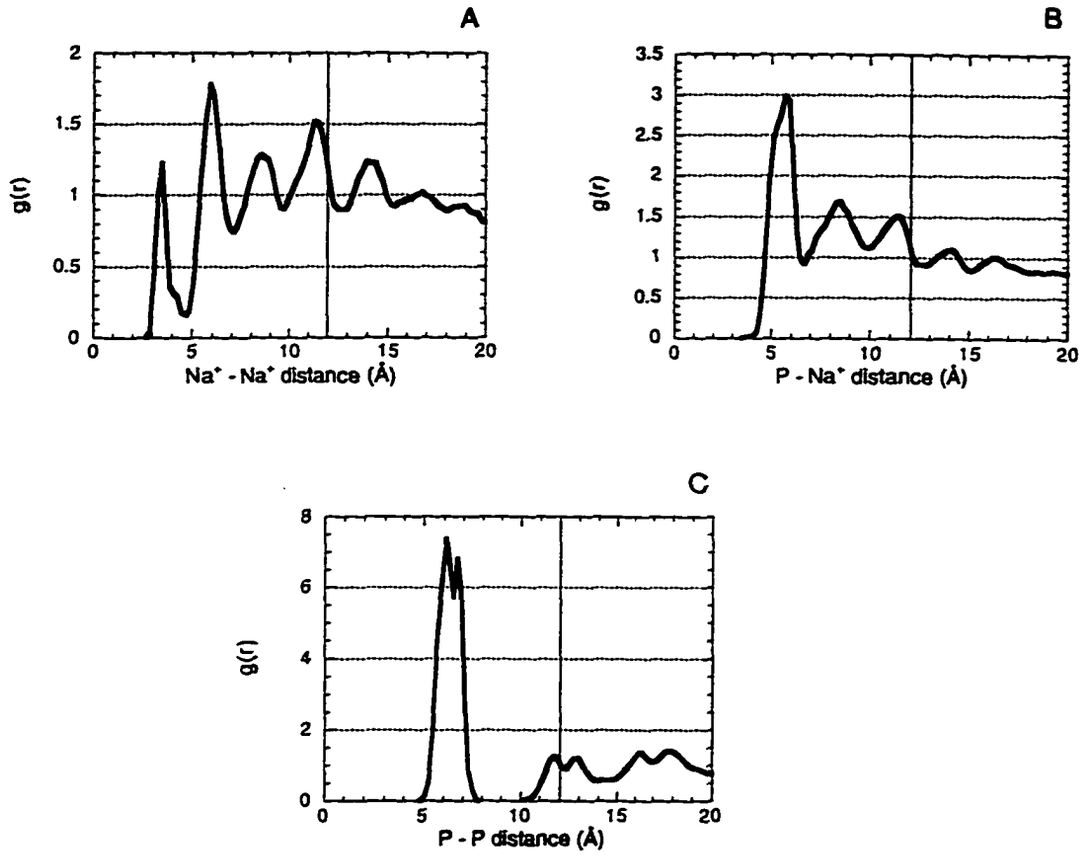

Figure 3.11 Radial distribution functions (rdf) for the charged species in **mlp**. A: Na+-Na+ rdf; B: Na+-P rdf; C: P-P rdf. The vertical line at 12Å indicates the place where the electrostatic interaction between these atoms becomes zero.

Another concern is the time required to equilibrate the counterion distribution around DNA, especially because the counterions could influence the conformation of the DNA (Mirzabekov 1979; Strauss 1996). **mlp** is the longest simulation (2 ns), so it was chosen to examine the relaxation behavior of the sodium ions. A qualitative picture of the equilibration of the sodium ions

can be found in Figure 3.12. This figure was generated by plotting the positions visited by sodium atoms throughout the 2 ns of production time. The "head on" view at the bottom indicates that radial equilibration has been achieved, since there is no empty space left by the ions (the hole in the center is where the DNA dodecamer lies). The side view (on top), on the other hand, has empty spaces at the ends of the simulation cell, which are places where no sodium atom has ventured.

A more quantitative approach is found in Figure 3.13, which shows the DNA-sodium radial distribution functions (calculated with the proximity criterion described in section 2.5) for the three intervals of the simulation where the DNA structure was stable (defined from figure 3.1.A). The last two intervals are very similar, suggesting that it takes at least 500 ps to equilibrate the ions. **mlp** turned out to be a poor choice to study equilibration, because one sodium ion associated with the O6 of a terminal guanine, and did not exchange in over 1.5 ns. That is the explanation for the peak at 2Å, which does not change appreciably with time. The replica simulation **mlp-I** did not show any long lived associations between sodium ions and the DNA, and neither did any of the other dodecamers with identical 5' end sequences (**6t** and **7g**) suggesting that this association is not sequence dependent. The only other simulation presenting long lived sodium-ketone oxygen associations is **athdna**, also shown in Figure 3.13, where one sodium is located at exactly the same position as in **mlp**, and another is coordinated by the O6 of guanines 10 and 11 of the sense strand.

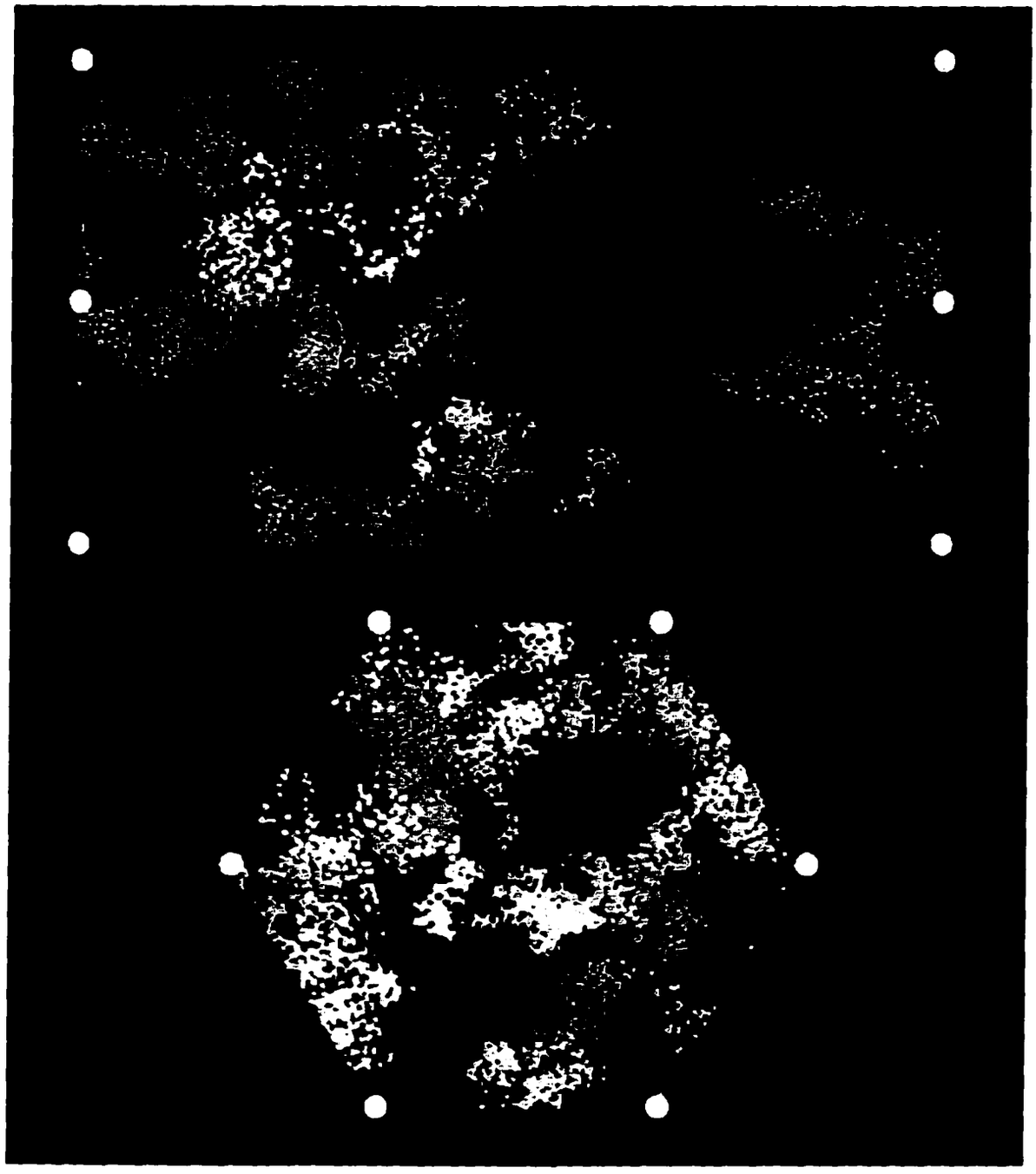

Figure 3.12 Space visited by sodium ions during 2 ns in **mlp**. Each sodium is colored differently. The larger white dots indicate the boundary of the simulation cell.

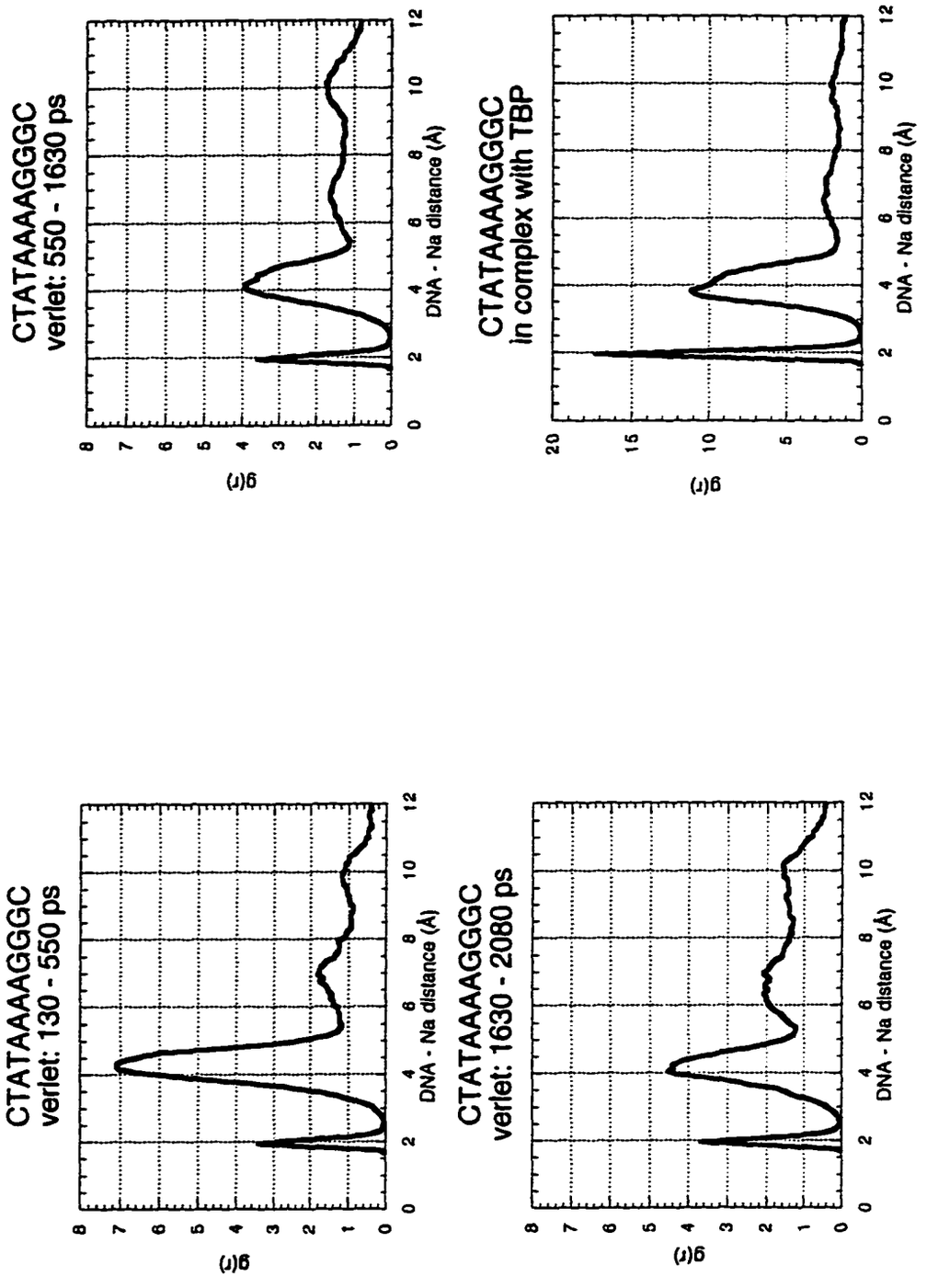

Figure 3.13 Time evolution of DNA-sodium radial distribution functions for **mlp** and comparison to **athdna**.

An independent simulation of another DNA oligomer (not described in this work) shows that these sodiums can exchange, indicating that this is not a problem of the interaction potential, but a sampling (and luck) problem.Similar long lived sodium-DNA associations in simulations carried out with the AMBER potential (Pearlman 1995) have been described, with the difference that those happen in the minor groove (Young 1997). In the simulations described in this thesis, there are no instances of ions coming into the minor groove of DNA.

The method of proximity analysis (see Section 2) allows to assign sodium ions to each base along the DNA. The coordination number up to the border of the simulation cell is plotted in Figure 3.14 for the sense and antisense strands of **mlp**, for the three time intervals analyzed above. While the radial distribution functions indicate the presence of a condensation shell (Manning 1978) (the peak around 4Å), a disappointing finding is that there seems to be no evidence of an increase in sodium condensation as one walks into the dodecamer, away from the ends, even after disregarding the guanine 12 in the antisense strand (this is the guanine bound to the sodium). The increase in sodium population towards the center of the DNA oligomer was expected from the simulations done by Olmsted et al. (Olmsted 1989), and its absence in **mlp** could be due to the lack of added salt. The actual number of sodium ions assigned to each base oscillates with time, a finding that could be interpreted as having achieved equilibrium, even by 550 ps. In conclusion, different measures give different pictures regarding equilibration. The strongest

argument for a lack of convergence in the sodium distribution is the appearance of the "non-exchanging" sodium ions; on the other hand, the oscillations in the coordination numbers for each base indicate that convergence has been achieved.

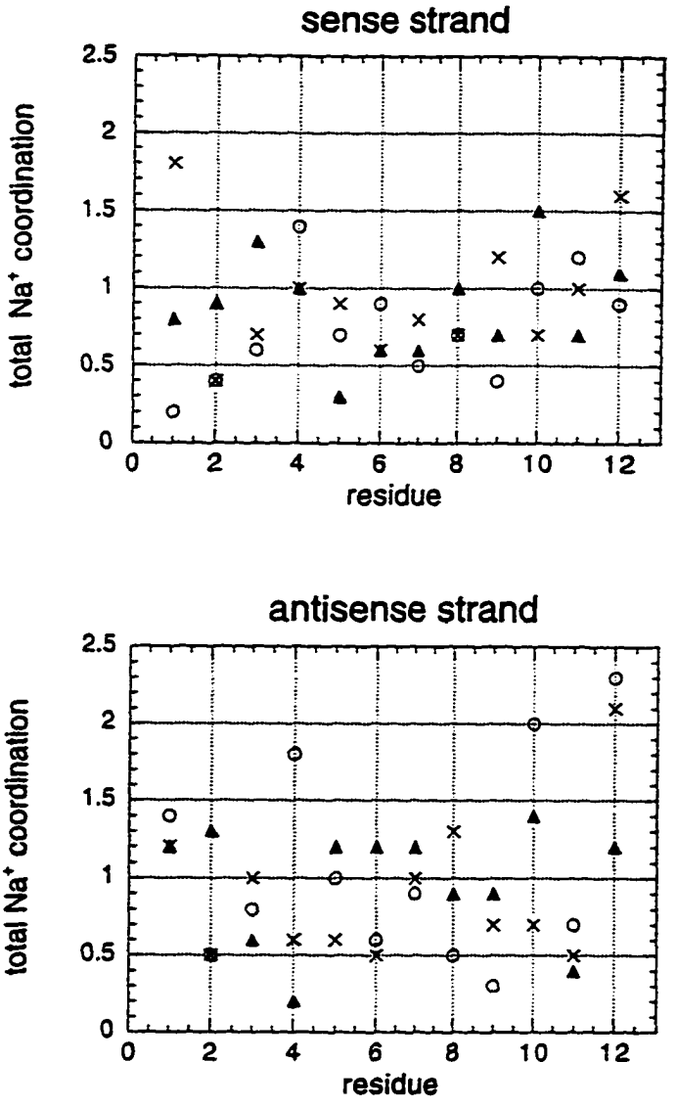

Figure 3.14 Time evolution of the sodium coordination number for each base in **mlp**. Both strands are read 5' to 3'. o: 130-550 ps; x: 550-1630 ps; triangles: 1630-2080 ps.

3.4 Solvent structure and dynamics

The simulations of pure water, and water with only 3 sodium atoms (**sod3**) were included as a control for characterizing the structure and dynamics of the pure solvent and the simplest sodium solution at a comparable concentration to that used in the simulations with macromolecules.

Figure 3.15 shows all the possible pairs of radial distribution functions (calculated with CHARMM) for these simple systems. Regarding the pure water simulation, the three g(r) have the major peaks at the expected distances for liquid water (Narten 1971); the oxygen-oxygen g(r) shows an exaggerated first peak, a feature of TIP3 water (Jorgensen 1983). The sodium-sodium g(r) indicates the existence of contact pairs and solvent separated pairs. The plot is "noisy" because there are only three particles to average, albeit for 2 ns. The sodium-oxygen and sodium-hydrogen g(r)s are consistent with a first hydration shell of 6 waters, as expected for a sodium solution. The water g(r)s for the **sod3** simulation are identical to those obtained for pure water.

To examine the effect that sodium ions have on the structure of water, the three g(r)s for water were calculated for those water molecules within 4Å of any sodium atom. The results are shown in Figure 3.16, and are again consistent with 6 water molecules forming the first hydration shell (Marchese 1984) and references therein).

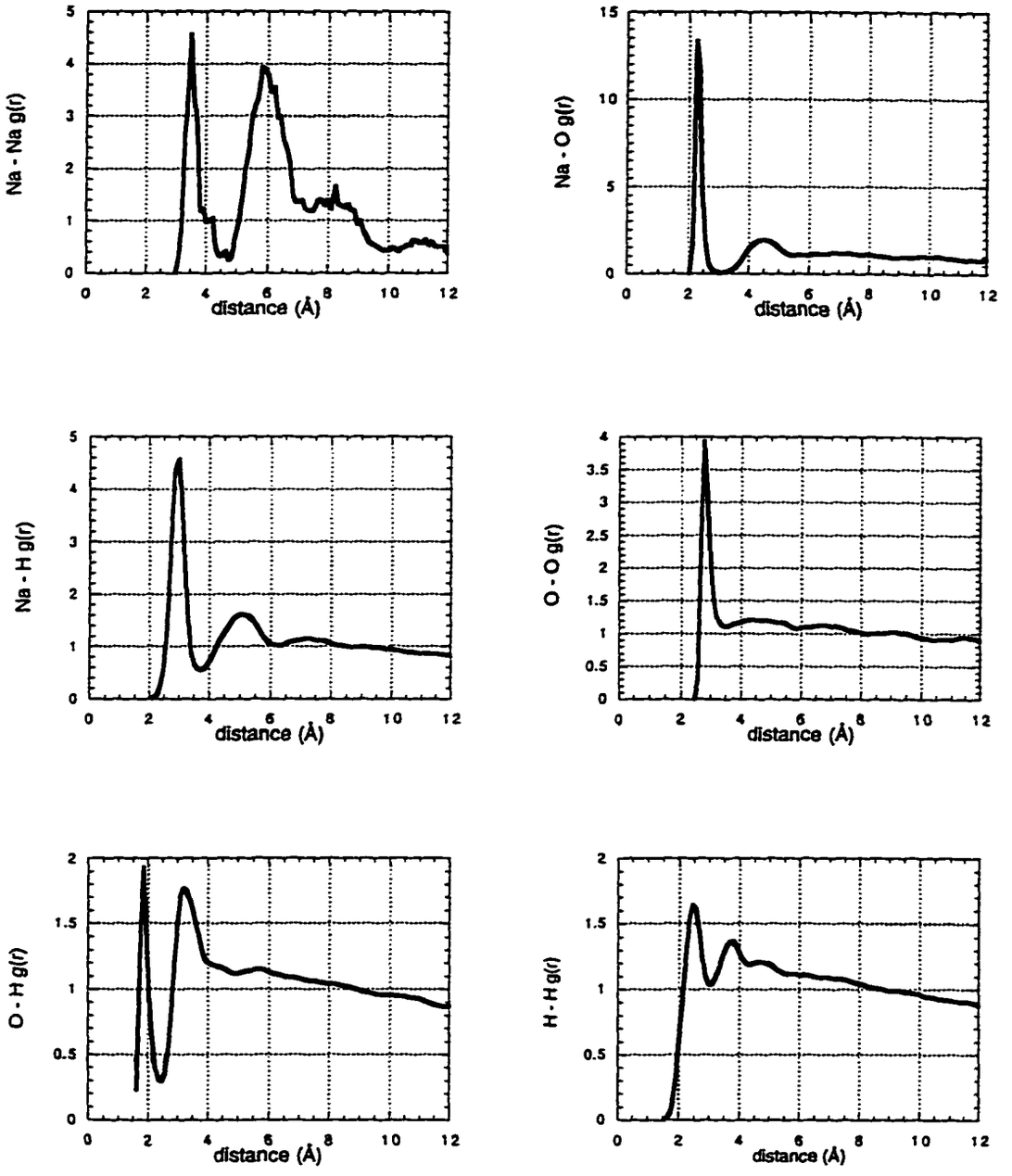

Figure 3.15 Pairwise radial distribution functions for the pure water and **sod3** simulations.

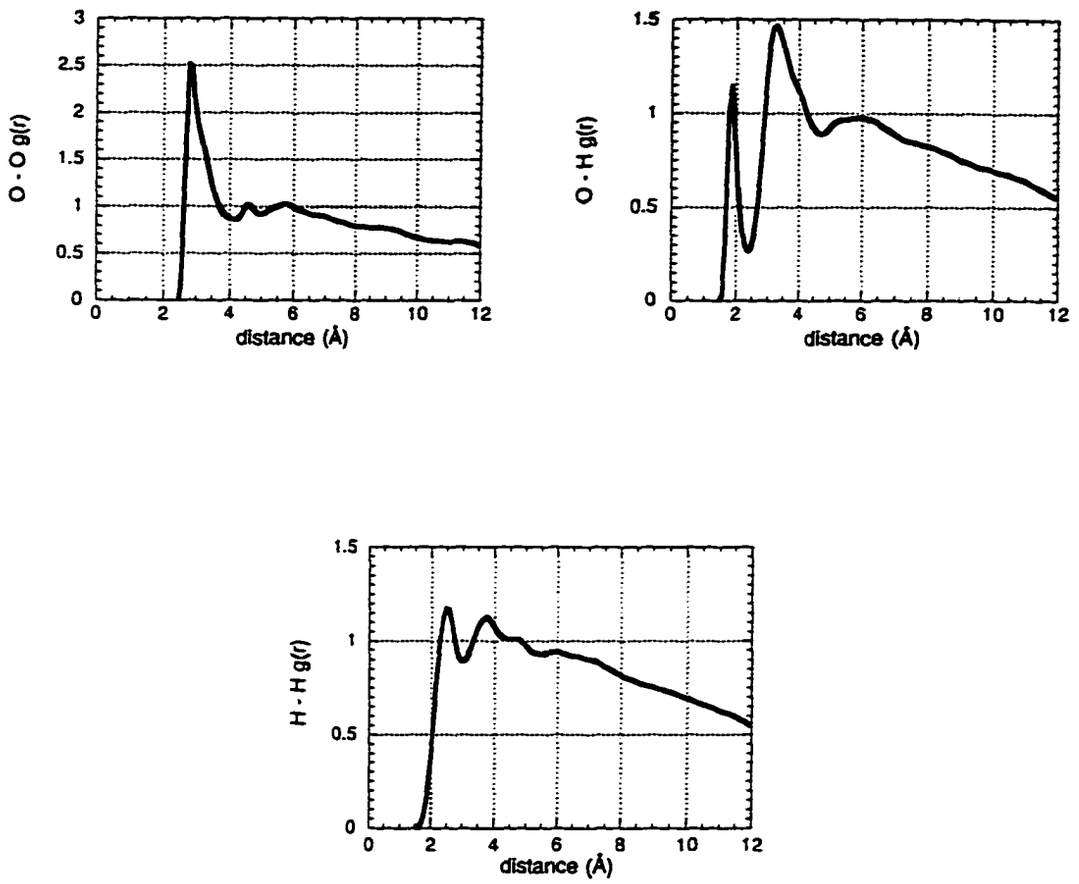

Figure 3.16 Pairwise radial distribution functions for the water molecules in the first hydration shell of sodium ions (**sod3** simulation).

Because one of the purposes of the simulations is to analyze the flexibility of DNA and TBP, it is necessary to characterize also dynamic properties of the solvent. To that effect, the diffusion coefficients for water (using the oxygen atom as the center of mass) and sodium ions were calculated for all the simulations, and the results are gathered in Table 3.5.

Table 3.5 Sodium and water diffusion coefficients (10^{-9} m^2/s)

simulation system	sodium	water
pure water	------	2.4
sod3	0.50	2.1
mlp	0.17	0.7
mlp-l	0.27	0.6
2c	0.27	0.7
6t	0.35	0.5
7g	0.30	0.6
r28	0.20	0.4
i	0.22	0.7
at	0.23	0.5
gc	0.26	0.5
ath	------	1.1
athdna	0.01	0.1

The value for diffusion of pure water agrees very well with the reported experimental value of 2.3 x 10^{-9} m^2/s (Mills 1973). Water diffusion is seen to slow down in the presence of DNA more than in the presence of TBP. This behavior agrees with NMR measurements of O^{17} labeled water in DNA solutions (van Dijk 1987).

The calculated diffusion coefficient for sodium is slower than the experimental value (1.33 x 10^{-9} m^2/s (Atkins 1990)), and the reasons could be as varied as a concentration effect, the lack of anions in the simulation or lack of convergence due to poor statistics (only three sodiums are present in the

simulation cell). Nonetheless, it is seen again that DNA slows down the motion of the sodium ions, consistent with the electrostatic attraction between the oligoanion and the sodium.

In conclusion from the validation section, the comparison to experimental data suggests that both the interaction potential used (CHARMM23, (MacKerell 1995)) and the simulation protocol are capable of generating realistic structures of the DNA and the protein systems, and provide insight into relevant elements of the dynamic behavior of the simulated macromolecules and their environment.

4 DNA Bendability: Preparation of DNA for Binding by TBP

4.1 Introduction

The idea that the sequence dependent local conformation and dynamics of DNA contributes to its recognition by ligands is not new (Calladine 1982; Hagerman 1990; Hagerman 1992; Harrington 1994; Travers 1991; Travers 1992; Zhurkin 1979). The relevant properties can be characterized using a variety of parameters, such as groove widths, global bending angles, or local geometrical parameters. A common approach is to rationalize preferential directions for DNA bending and flexibility based on the assumption that the structure and dynamics of a DNA molecule can be understood from the properties of its constituent basepair steps, where the basepairs are sometimes modeled as planes. Examples of such studies (Goodsell 1994; Gorin 1995; Suzuki 1995; Suzuki 1996; Ulyanov 1995; Yanagi 1991; Young 1995) rest on the analysis of structures obtained from X-ray crystallography and NMR that are available in the NDB (Berman 1992) or PDB (Bernstein 1977), and on inferences from a variety of solution techniques, such as anomalous migration in gel electrophoresis (Crothers 1992; Haran 1994), differential reactivity to agents that break DNA (Price 1993) and cyclization efficiency assays (Kahn 1994; Lyubchenko 1993). Computational studies of sequence-dependent DNA properties include Monte Carlo simulations at various levels of detail (Olson

1995; Sarai 1996; Sarai 1988; Sarai 1989; Srinivasan 1987; Zhurkin 1985; Zhurkin 1979; Zhurkin 1991), adiabatic mapping of the properties of basepair steps and DNA oligomers (Hunter 1993; Hunter 1997; Poncin 1992; Poncin 1992; Sanghani 1996; Zakrzewska 1992), and mechanical models of DNA (Calladine 1982; Calladine 1986; Calladine 1996; Calladine 1988; Goodsell 1994).

The extreme deformation imposed on DNA by TBP binding suggests that this induced conformation is not easily accessible to free DNA at room temperature, making the energy required to acquire this conformation a likely selectivity determinant (Kim 1994; Kim 1993; Kim 1993; Struhl 1994). DNA bound by TBP has been subject of detailed analyses by Juo et al. (Juo 1996), Suzuki et al. (Suzuki 1996), Guzikevich-Guerstein and Shakked (Guzikevich-Guerstein 1996) and Lebrun et al. (Lebrun 1997). These works are based on a static analysis of DNA structures, and on mechanical models of DNA. In summary, Juo et al. (1996) propose that TATA boxes are designed to achieve flexibility (in minor groove width and positive *roll*ing) and close contact between the protein and the floor of the minor groove (hence, no guanines are allowed because the exocyclic amino group would push the protein away from the floor of the groove); alternating AT is preferred over homopolymeric stretches because the latter tend to have narrower minor grooves and are also more rigid than the former. Suzuki et al. (Suzuki 1996) explain binding preferences as a matter of matching curvatures between the DNA and TBP. According to them, the issue is the ability to compress or expand residues across the β-sheet,

pointing into the protein core, and to avoid clashes between the C5-methyl groups of the thymines, which point towards the major groove of the DNA. Guzikevich-Guerstein and Shakked (Guzikevich-Guerstein 1996) propose that TATA boxes are selected on their propensity to achieve an A-DNA like conformation, given the remarkable similarity in backbone conformations between canonical A-DNA and TBP-bound DNA (called TA-DNA by them); the transition to A-DNA is assumed to start at the 5' end of the TATA box, because this sequence has been crystallized in A-DNA form (Shakked 1983). Lebrun et al. (Lebrun 1997) suggest that TBP binding is akin to DNA stretching, due to the similar structures generated by TBP-bound DNA and DNA stretched from the 3' ends of the central four basepairs of the TATA box. They find that the structural transition is energetically more feasible if particular phosphates are neutralized, and hypothesize that these neutralizations aid the B- to A-DNA transition of the center of the TATA box.

The analysis of the available crystal structures allows for the characterization of the final TBP-bound DNA geometry, but is limited in the understanding of which properties are necessary to get there, and which follow from these. The only mechanistic explanations proposed this far are Lebrun's (Lebrun 1997) (discussed above) and Elcock's (Elcock 1996), who can reproduce approximately the degree DNA bending produced by TBP by reducing the dielectric constant between the phosphates in the minor groove (neither of these studies focused on the sequence dependence of the deformations they find). Hence, an analysis of average, static properties might

be incomplete. If the free DNA molecules exhibit transiently some characteristics of the DNA geometry found in the complex with TBP, one could hypothesize that these characteristics are not only the trigger for binding, but also the selectivity determinants. These transient, extreme conformations, are very likely to be sequence dependent, and are inherently time-dependent phenomena, particularly amenable for study with molecular dynamics simulations.

The following sections describe an analysis of both average and transient DNA structural properties; this is an extension of the analysis presented in Pastor et al. (Pastor 1997). The characteristics that TBP uses to distinguish its binding site are unknown, so the first problem is deciding what to measure; the properties in question could manifested, for example, in the global bending of the binding site, in the phosphate-sugar backbone conformation and/or in the basepair stacking patterns. This chapter is a search for those properties that might be used by TBP to select its DNA binding partner.

4.2 Average structural properties

4.2.1 DNA Global Conformational Analysis

The extreme widening of the minor groove and C3'-endo sugar puckers displayed in DNA bound by TBP prompted the Burley (Kim 1993) and Shakked (Guzikevich-Guerstein 1996) labs to propose an intermediate in the binding

reaction that would look like A-DNA. As there are no experimentally derived structures for any of the sequences studied in this work, the simulation results were examined for evidence of such a conformational intermediate for the DNA alone, in the absence of TBP. The 2D root mean square (rms) distance plot for the 2 ns run of **mlp** indicated that at least three distinct structures were visited during the simulation trajectory (Figure 3.1.A). As a first characterization of these structures, the rms difference from A-DNA and B-DNA for the eight simulated dodecamers was calculated as a function of time. The plots for all of them (Figure 4.1) show a progressive departure from B-DNA and a concomitant approximation to A-DNA, converging after 550 ps towards an rms value around 4Å from either canonical structure. There are nevertheless some differences in the distance from A-DNA and B-DNA at which different sequences stabilize. For example, **i** stabilizes between 4 and 5Å from A-DNA, but **2c** keeps drifting towards A-DNA even after 550 ps. While these rms values are too large to classify the structures as either A-DNA or B-DNA, the results clearly differ from both these canonical structures.

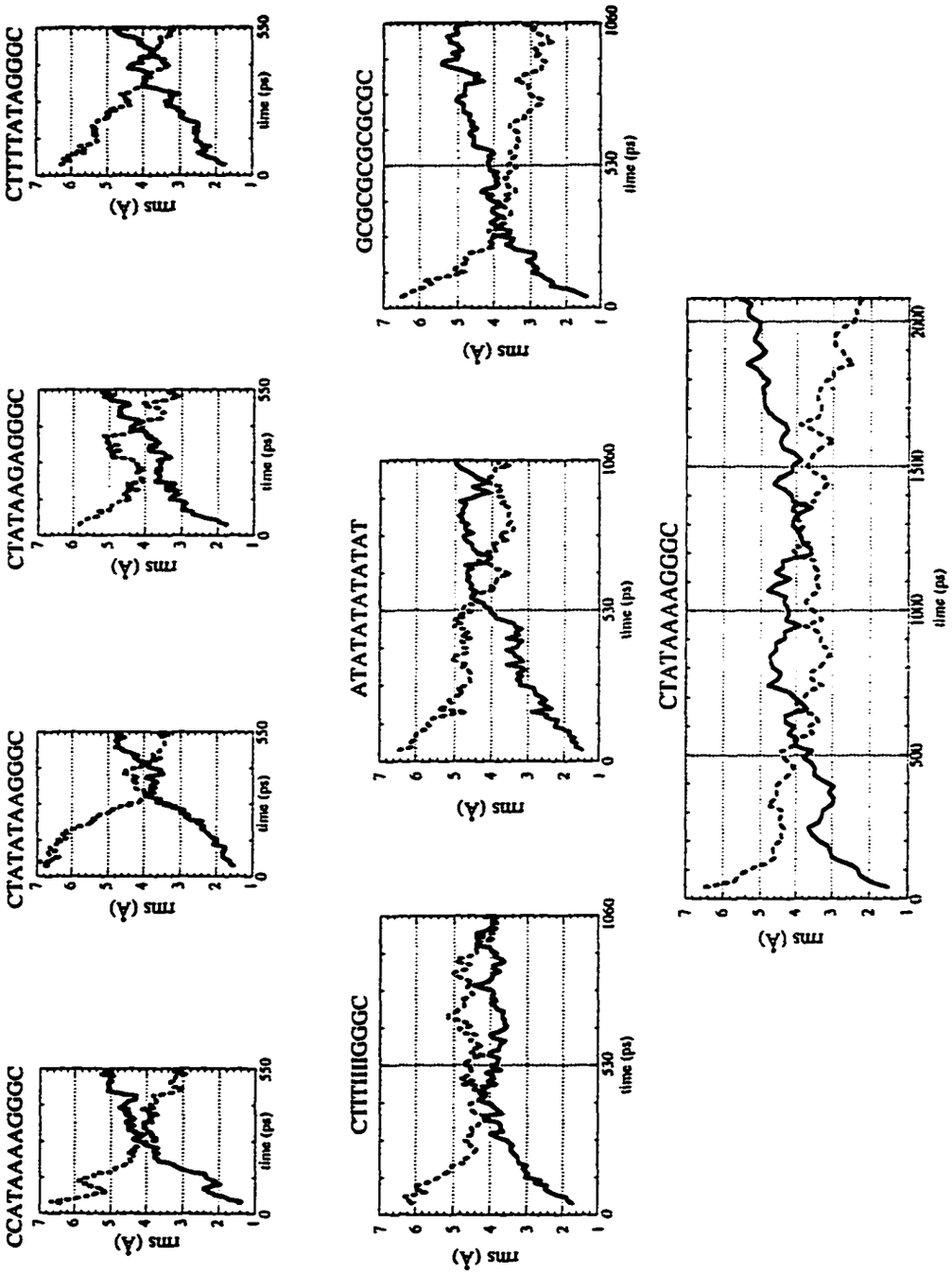

Figure 4.1 Time evolution of the rms to A- and B-DNA for the free DNA simulations. Solid line: rms to B-DNA (starting structure in the simulations); broken line: rms to A-DNA (Arnott 1972).

Another prominent characteristic of DNA bound by TBP is the ~90° bend. To explore the extent of bending, the DNA helix axes were calculated for the average structures obtained from the intervals in the simulation in which the structures showed little or no drift to another structure, reflected in their being within 1.8Å rms of each other (defined by the 2D rms plots in Figure 3.1). The superimposition of the helical axes for the average structures of the three known binding sites (**mlp**, **6t**, **at**) and the negative control (**gc**) revealed them to be indistinguishable from each other and to lack any significant bending (Figure 4.2).

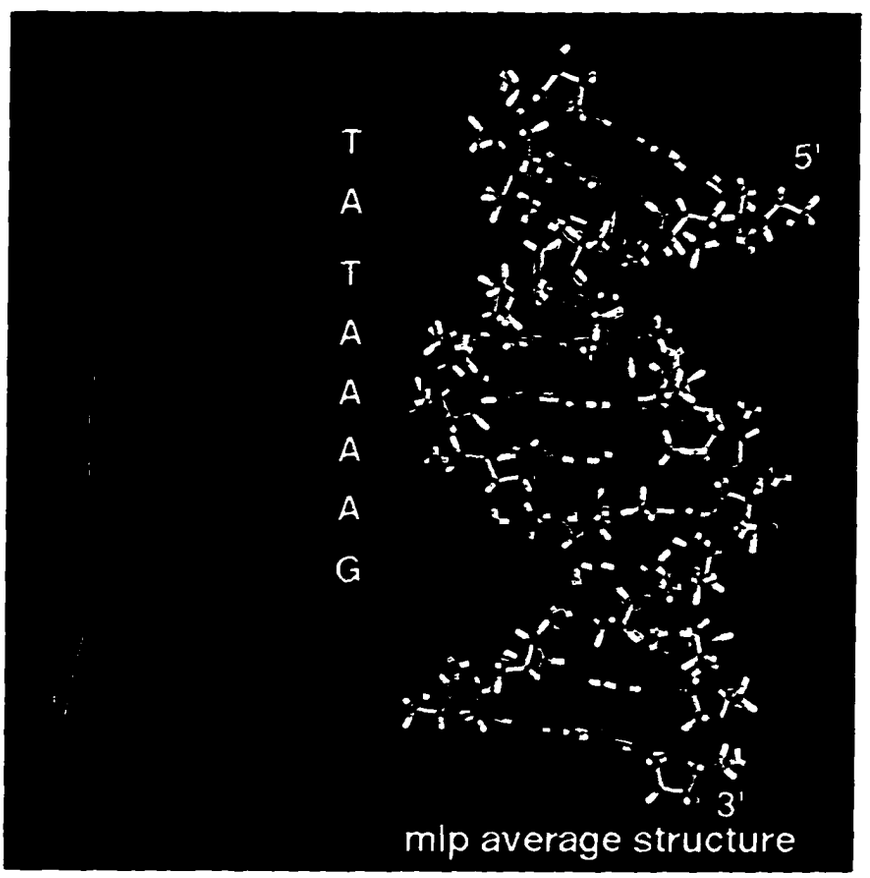

Figure 4.2 Comparison of the helical axes for the average structures of known TBP binding sites (**mlp** in purple, **6t** in blue, **at** in red) and a negative control (**gc** in green). The average structure for **mlp** is shown on the right, with the helix axis in purple, and the sequence of the TATA box on its left; the 5' and 3' ends of the sense strand are labeled. (Black and white version: blue = darkest gray; green = dark gray; red = gray; magenta = lightest gray).

The absence of a distinguishing feature among the calculated average structures of the various dodecamers moved the search for the selectivity determinants for TBP binding to the analysis of more local properties of the DNA oligomers, both bound to TBP and free.

4.2.1 P - P distances

As described by Guzikevich-Guerstein and Shakked (Guzikevich-Guerstein 1996), the DNA bound by TBP has adjacent interphosphate distances closer to those of A-DNA than to those of B-DNA. In line with this finding, a predisposition for shorter P - P distances could be a hallmark of good binding sites for TBP. Figure 4.3 shows the distribution of consecutive P - P distances for the TATA box region of the simulated dodecamers, both free and bound to TBP. While there are obvious differences in the distributions for the different dodecamers, there is no correlation between the ability to be bound by TBP and the population of distances around 6Å: the worst substrate (**gc**) has a very similar profile to one of the best binding sequences (**6t**). The split peaks found in some of these distributions can reflect either a structural drift to shorter P - P distances throughout the simulation, or inhomogeneity in the P - P distances along the DNA dodecamer.

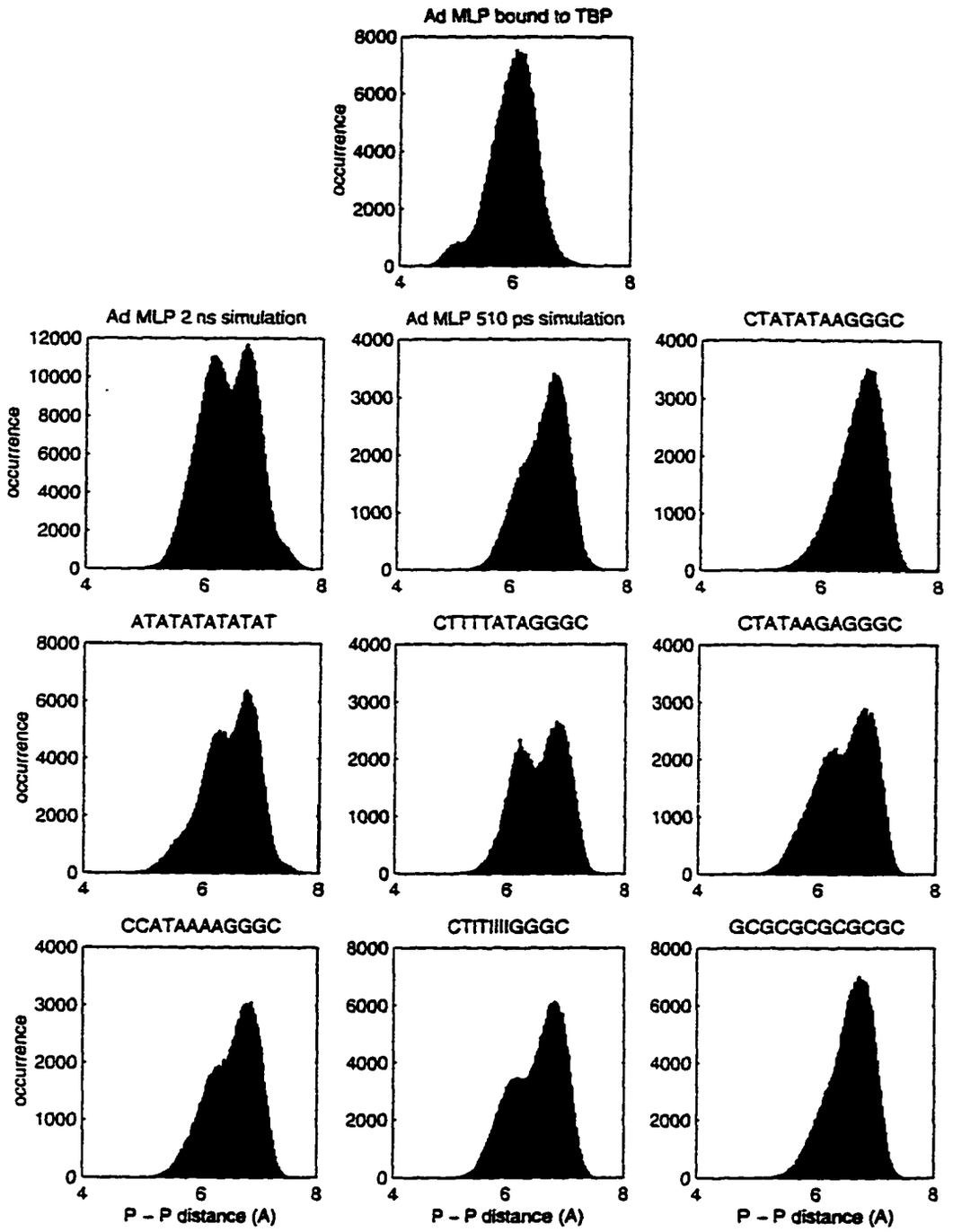

Figure 4.3 Consecutive P-P distance distributions for the TATA boxes of the free DNA simulations and **athdna.**

4.2.2 Basepair step sequence dependent geometry

Overall, TBP-bound DNA is found to exhibit low *twist* and positive *roll*; the first and last basepair steps have high *rise*, as a consequence of the insertion of phenylalanine residues from TBP (see Table 1.3). To look for the appearance of such specific local properties in the simulations, all the basepair step conformations generated in the simulations were scanned to identify those steps that have average values biased towards these characteristics. As a first approximation, the basepair steps were separated into three classes, depending on their sequence, effectively averaging the contributions from neighboring sequences. As shown in Table 3.3 and Figure 3.8, RY steps (R stands for purine and Y for pyrimidine) were found to share the properties of low *twist* and positive *roll*, while YR steps were found to have high *rise*.

These findings lead to a proposal for a canonical binding site that is based on the average properties of basepair steps and on the local requirements for TBP binding identified from the complexes with TATA box DNA. The canonical binding site that satisfies these requirements should have YR steps at the ends of the binding site, while RY steps would be needed throughout. Consequently, the combination that best accommodates the requirements is an alternating YR sequence. The classification of basepair steps into three broad categories does not permit a distinction between alternating sequences that contain GC basepairs from those that contain only AT basepairs.

The sequence dependent behavior found above justified a finer analysis.

Table 4.1 contains the averages and standard deviations for the fourteen types of steps found in the simulated dodecamers.

Table 4.1 Sequence dependent basepair step parameters

step	*shift*	*slide*	*rise*	*tilt*	*roll*	*twist*
AA	0.1 ± 0.6	-1.5 ± 0.7	3.2 ± 0.4	0.0 ± 6.4	3.3 ± 10.1	31.7 ± 4.8
AG	-0.1 ± 0.7	-1.5 ± 0.8	3.4 ± 0.4	-0.2 ± 6.4	1.7 ± 11.7	32.4 ± 4.5
GG	0.2 ± 0.7	-2.2 ± 0.4	3.3 ± 0.4	2.5 ± 6.4	-.3 ± 12.6	31.8 ± 4.9
GA	0.0 ± 1.1	-1.2 ± 0.5	3.0 ± 0.4	0.6 ± 6.7	7.5 ± 14.5	30.1 ± 7.6
II	0.3 ± 0.6	-2.0 ± 0.4	3.2 ± 0.4	1.2 ± 5.6	-1.1 ± 10.6	32.4 ± 4.7
IG	-0.4 ± 0.6	-2.4 ± 0.4	3.4 ± 0.4	2.9 ± 5.4	0.2 ± 8.0	31.2 ± 3.7
AT	0.1 ± 0.7	-1.1 ± 0.5	3.0 ± 0.4	2.0 ± 5.9	10.6 ± 10.6	30.4 ± 5.4
GC	0.0 ± 0.5	-0.8 ± 0.4	2.7 ± 0.3	0.2 ± 5.0	21.7 ± 9.7	27.2 ± 5.0
IT	0.4 ± 0.6	-0.7 ± 0.4	2.7 ± 0.3	-0.6 ± 4.6	18.1 ± 8.1	27.4 ± 4.8
TA	0.2 ± 1.0	-1.0 ± 0.7	3.6 ± 0.5	1.2 ± 7.5	3.2 ± 13.1	33.2 ± 5.9
CG	0.3 ± 0.9	-0.7 ± 0.7	4.0 ± 0.4	1.8 ± 6.4	-3.2 ± 12.1	33.9 ± 5.3
CA	-0.2 ± 1.2	-0.8 ± 0.5	3.6 ± 0.4	-0.5 ± 6.9	-4.7 ± 10.2	35.0 ± 4.3
TG	-1.1 ± 0.6	-1.2 ± 0.4	3.6 ± 0.4	-4.3 ± 5.0	-1.7 ± 7.8	32.7 ± 3.4
TI	0.5 ± 1.1	-0.7 ± 0.4	3.6 ± 0.4	1.7 ± 6.7	-5.4 ± 10.6	35.6 ± 4.4

According to this table, the fact that TBP binds poorly, if at all, to alternating GC sequences indicates that the poor binding is very probably due to the steric clash between the TBP sidechains and the exocyclic amino group of the G base, and not to the conformational preferences of the DNA, which seem to be even better suited for TBP binding than those of alternating AT.

4.2.3 TATA boxes and TITI boxes

TBP was defined biochemically as a minor groove ligand because of its ability to bind to sites containing IC basepairs but not GC basepairs (Lee 1991; Starr 1991). Sterically, IC and AT basepairs are identical in the minor groove side, explaining the absence of steric clashes between the side chains of TBP and the floor of the minor groove, and hence the ability to be bound by TBP. Because sequence dependent behavior was observed in the simulated DNA molecules, it became worthwhile to explore the equivalences and differences between **mlp** and an analogue where AT basepairs were substituted by IC basepairs (i).

Figures 4.4 and 4.5 show the distributions for the four basepair step parameters most relevant to the conformational change imposed by TBP, for an "mlp" sequence built from all the available basepair steps with the particular sequence of each step, and i in the region of the TATA box, read 5' to 3' from top to bottom of the page. The overall pattern of alternation of *rise*, *roll* and *twist* is shared by both sequences in the 5' half of the TATA box, although the actual position and widths of the distributions differ somewhat. For example, for *slide* in step 3, i is more prone to have less negative values (which is good for binding, see below), and the converse is seen in step 6. Another interesting difference is seen in *roll* in step 5, where "mlp" has a symmetrical distribution around 0 degrees, and i has a significant population with negative values, closing the minor groove.

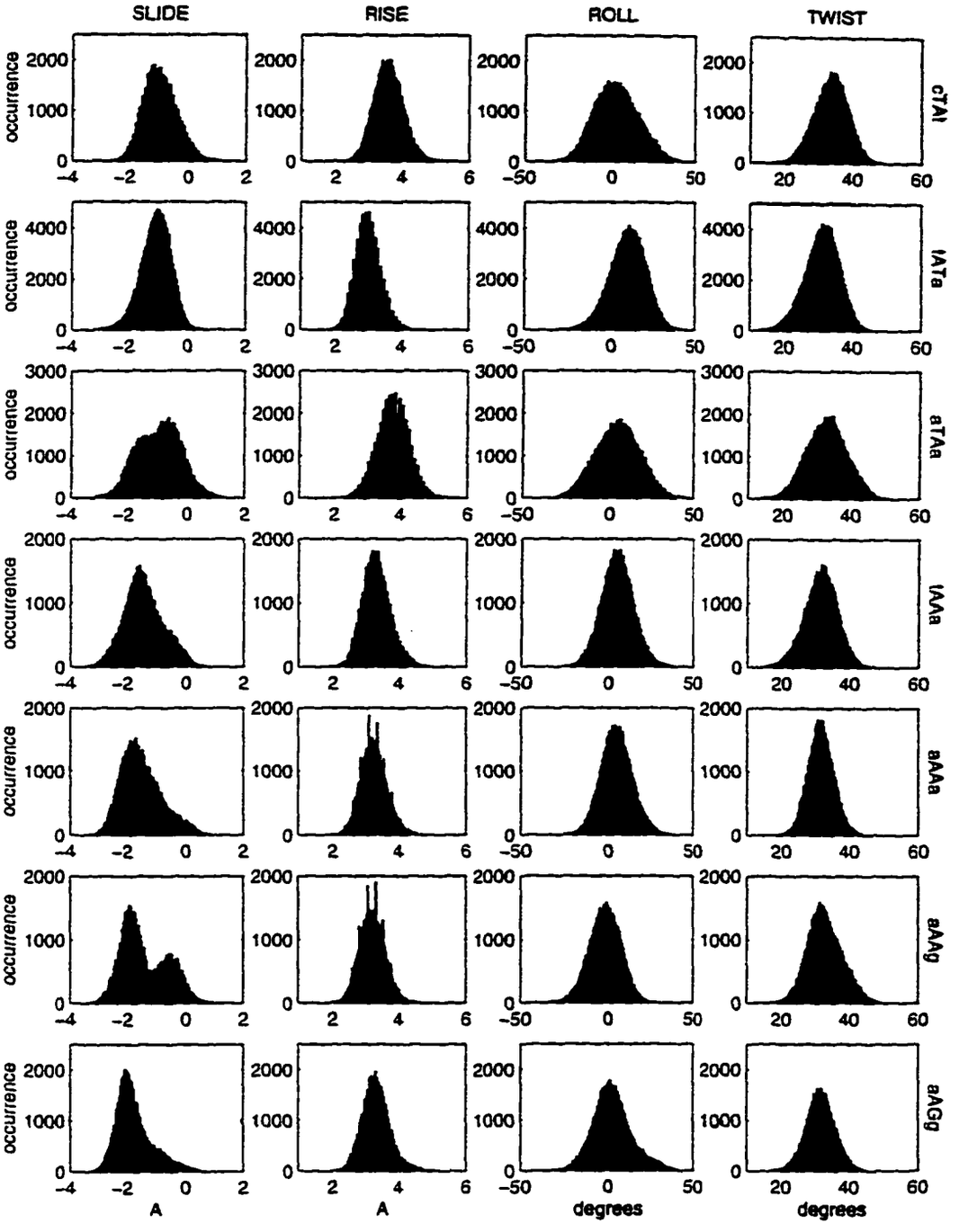

Figure 4.4 Selected basepair step geometrical parameters for a TATA box. The basepair steps are listed 5' to 3', and are indicated at the right of each row. The flanking nucleotides are shown in lowercase.

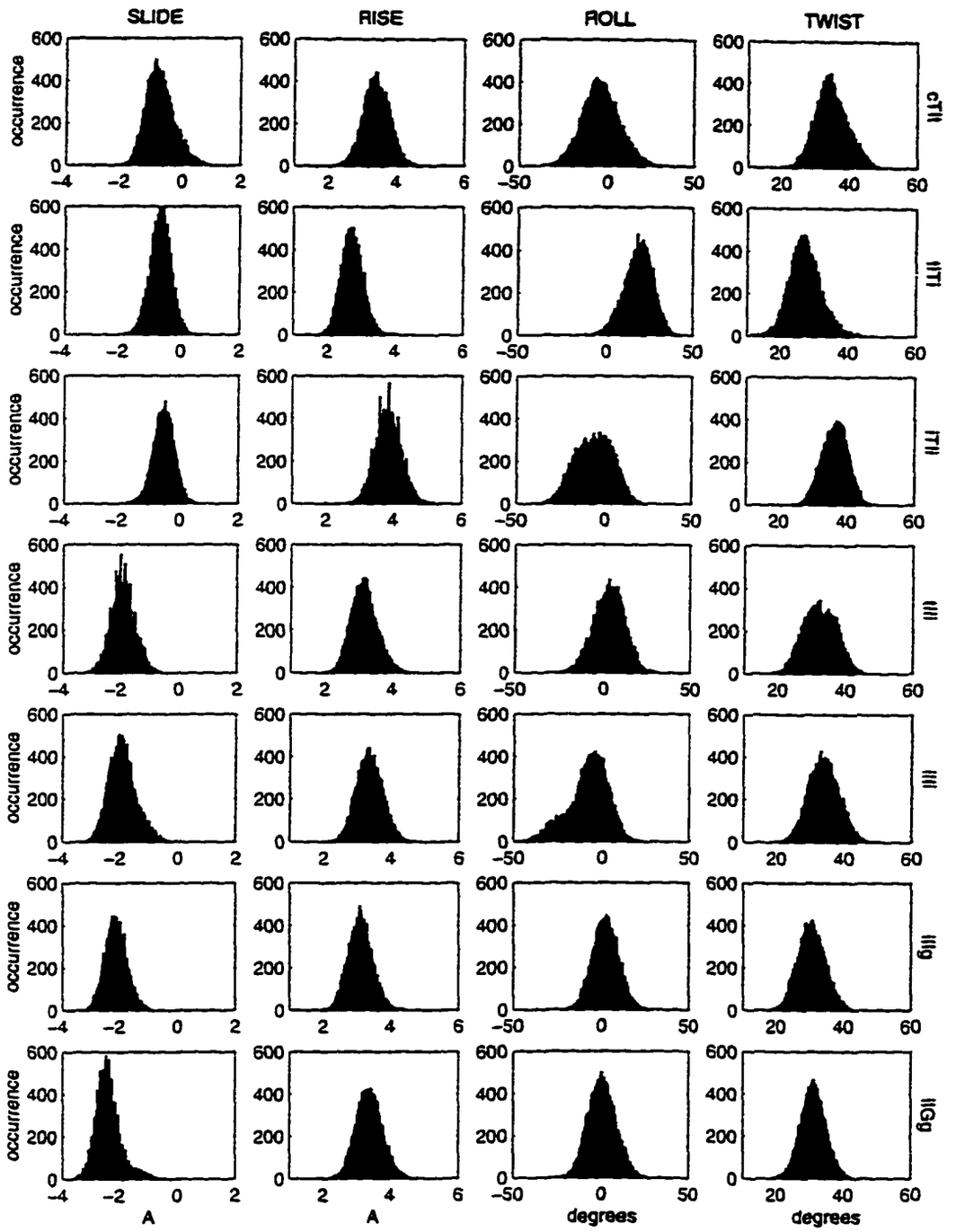

Figure 4.5 Selected basepair step geometrical parameters for a TITI box. The basepair steps are listed 5' to 3', and are indicated at the right of each row. The flanking nucleotides are shown in lowercase.

4.3 Transient structural properties: Achieving the conformation in the complex

The analysis presented thus far has shown the existence of preformed, incipient bending sites on the DNA which are in a direction that favors TBP binding. Now the dynamic component is explored, by searching for properties of the complex that are visited rarely during the simulations of free DNA, and which may be more populated for particular DNA sequences, making these preferred binding sites.

4.3.1 Sugar conformation

A trend found in the bound sites by TBP is that the sugars are in a conformation typical of A-DNA (Guzikevich-Guerstein 1996). This is true for most of the sugars in the TATA box in the **athdna** simulation, as depicted in the time traces of the dihedral angle δ, which is related to the sugar pucker (Figure 4.6). A value around 83° is typical of A-DNA, and 144° is the standard for B-DNA. The nucleotides constituting steps 1 and 7 show a flickery behavior, alternating between both conformations throughout the simulation, especially at the 5' ends of the strands. This is probably due to the asymmetric interaction of the inserting phenylalanine residues at the kink sites: these side chains stack against the base of the sense and antisense strand 5' ends, leaving the sugars free, while the corresponding partner stacks against the sugar of the opposite strand.

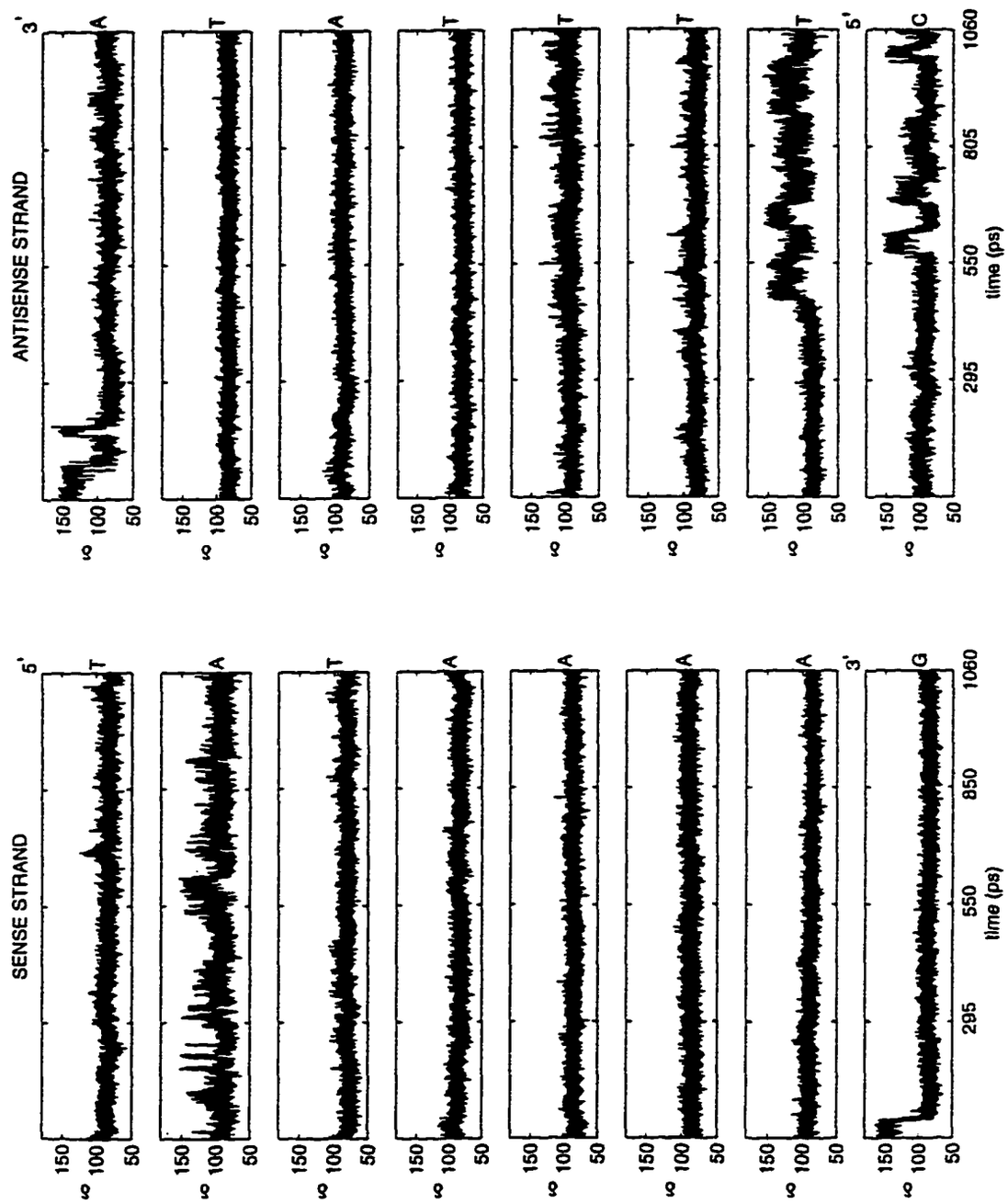

Figure 4.6 Time evolution of the DNA backbone torsion angle δ for the TATA box sugars in **athdna**. The nucleotide is indicated at the right of each plot; the sense strand reads 5' to 3' from top to bottom and the antisense strand reads 3' to 5'.

To examine whether a tendency to have the sugars in the A-DNA conformation existed in the different sequences, similar plots were generated for the sugars in the TATA boxes of the free DNA simulations. These are shown in Figure 4.7. Most of the sugars are seen to repucker extensively through the simulation, with the interesting exception of the adenine sugars in A-tracts (**mlp**, **mlp-l**, **2c**) and the I-tract (**i**); these sugars tend to lock in the A-DNA conformation, and this happens first for the 3'-most sugar of the tract. This trend is absent in **r28**, probably because the A-tract ends too close to the 3' end of the strand. It is also absent in **6t** and **at**, two of the best TBP binding sites.

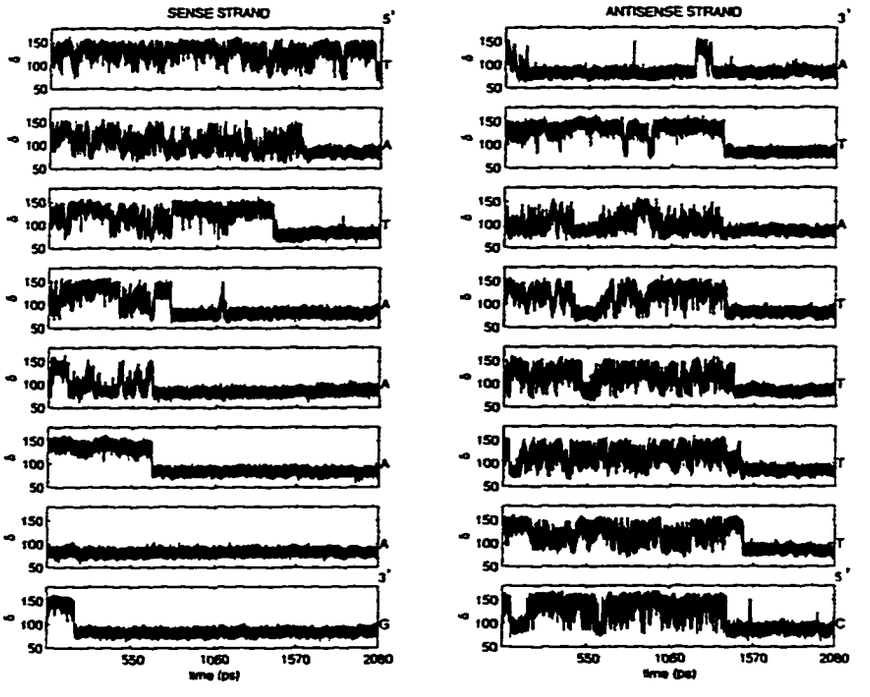

Figure 4.7 Time evolution of the DNA backbone torsion angle δ for the TATA box sugars in **mlp**.The sense and antisense strands (and their 5' and 3' ends, respectively) are labeled at the top of each column, and the nucleotide is indicated on the left of each plot.

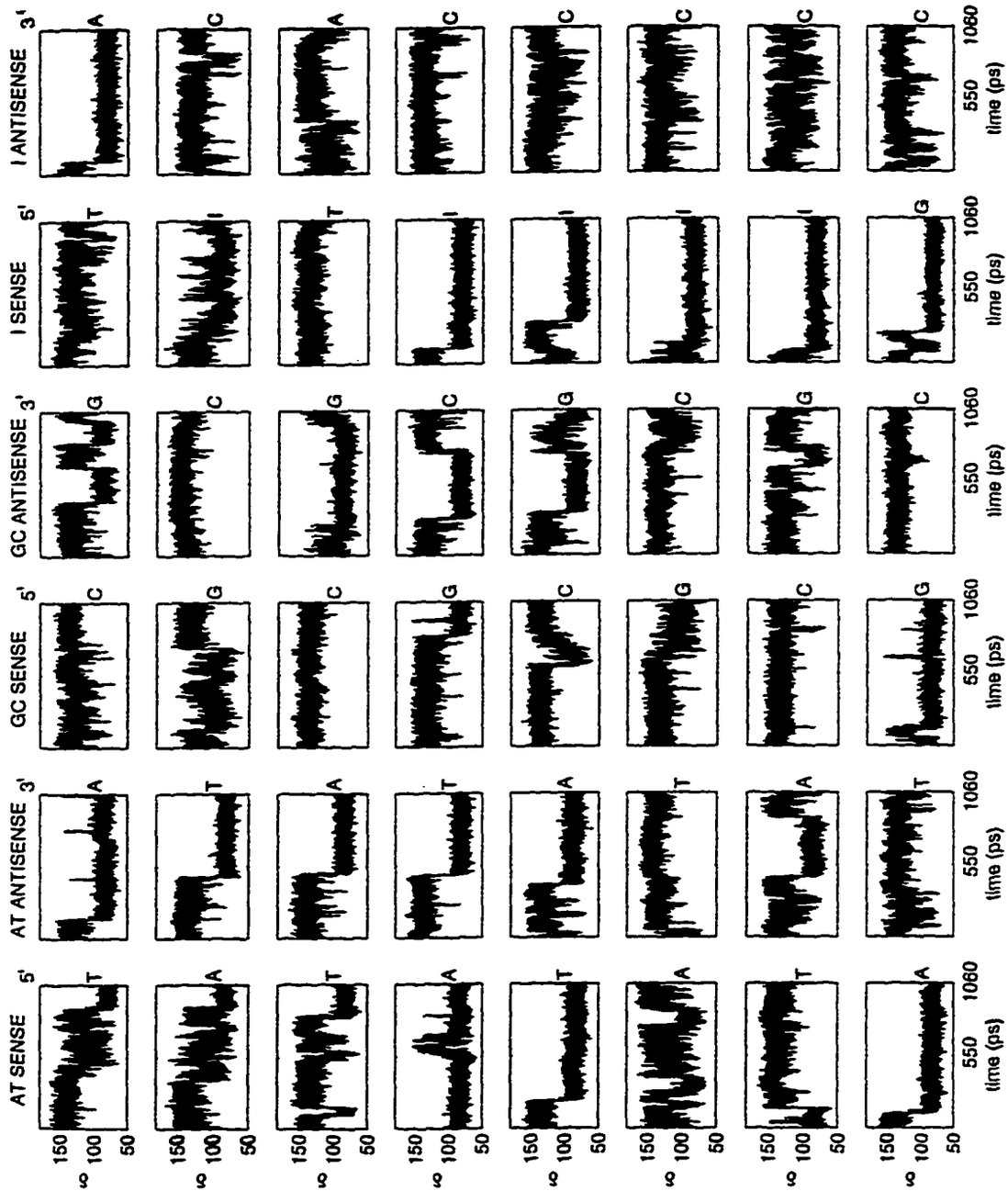

Figure 4.7 Time evolution of the DNA backbone torsion angle δ for the TATA box sugars in **at**, **gc** and **i**. The sense and antisense strands (and their 5' and 3' ends, respectively) are labeled at the top of each column, and the nucleotide is indicated on the left of each plot.

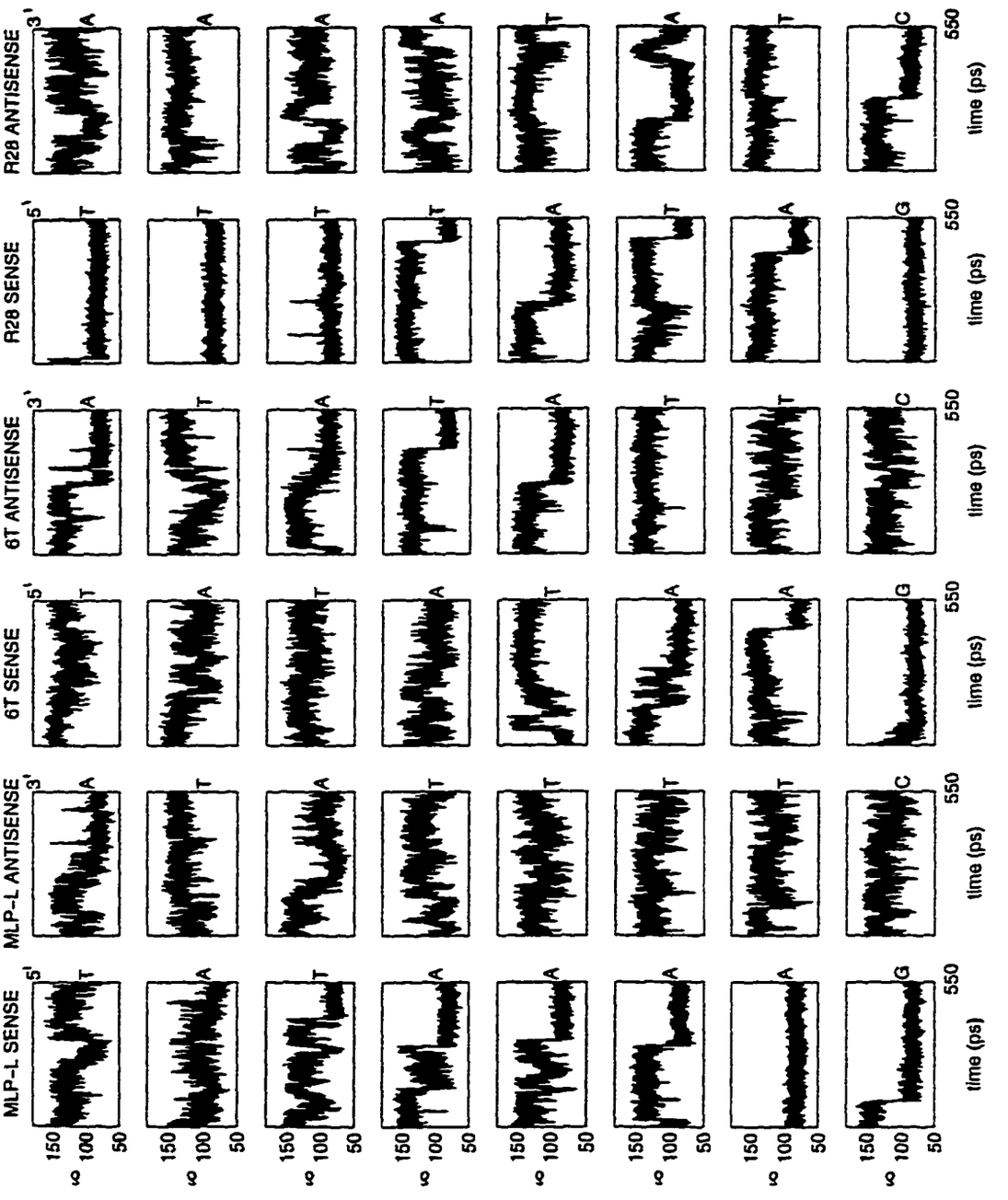

Figure 4.7 Time evolution of the DNA backbone torsion angle δ for the TATA box sugars in **mlp-l**, **6t** and **r28**. The sense and antisense strands (and their 5' and 3' ends, respectively) are labeled at the top of each column, and the nucleotide is indicated on the left of each plot.

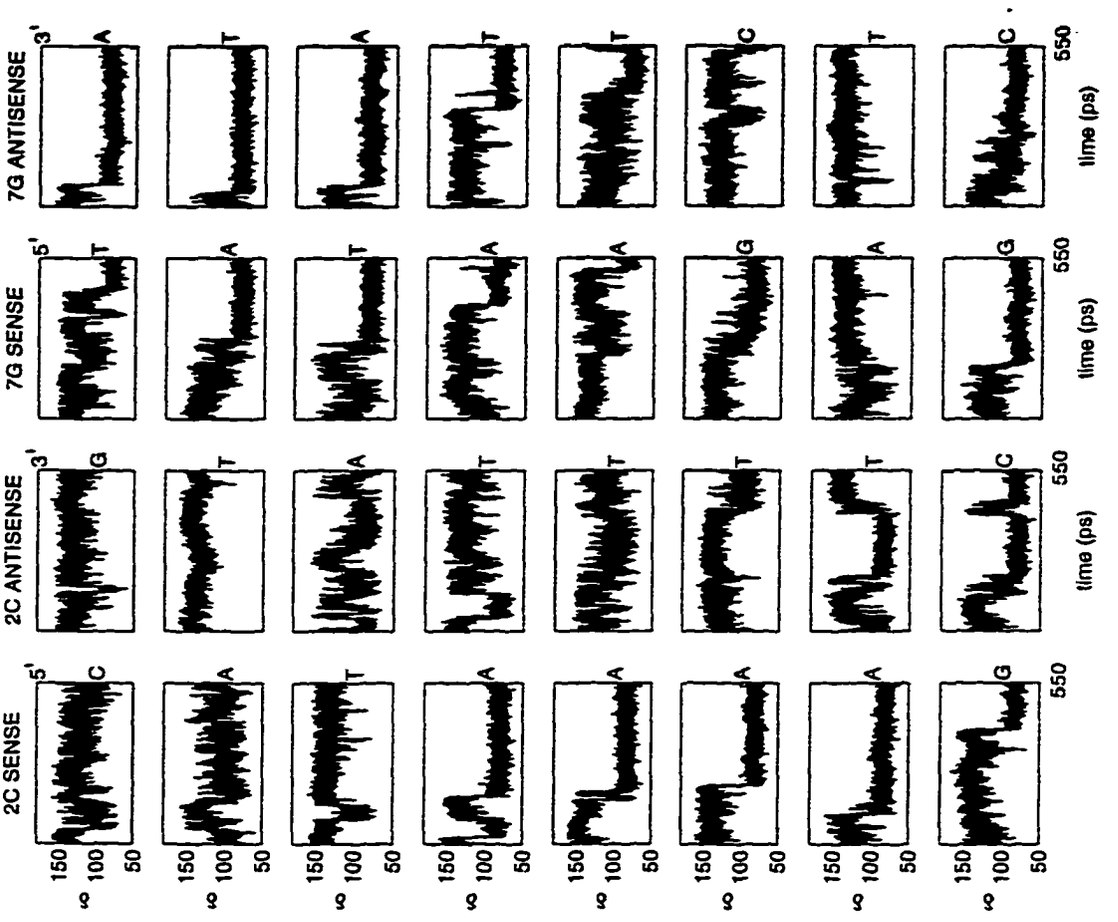

Figure 4.7 Time evolution of the DNA backbone torsion angle δ for the TATA box sugars in **2c** and **7g**. The sense and antisense strands (and their 5' and 3' ends, respectively) are labeled at the top of each column, and the nucleotide is indicated on the left of each plot.

The tendency to lock the sugars of the A-tract in the A-DNA conformation correlates with the ability to bind to TBP in the case of **mlp** and the two point mutations where the culprit for lack of binding is thought to be a steric clash (**2c**, **7g**). The absence of this feature in the two other proven binding sites (**at**, **6t**) suggests that different sequences may exploit different molecular strategies in order to prepare for TBP binding.

4.3.2 Basepair step geometry

The general consensus of the canonical TBP binding site defined above can be refined further to yield more specific sequence characteristics. This refinement is based on the premise that TBP binding would be facilitated if the DNA adopted even transiently a set of local conformations corresponding to the geometries observed in the complexes with TBP. Such sparsely populated conformations can be identified from the results of the simulations, which offer over 1 million conformations for the basepair steps that constitute the simulated dodecamers.

To identify specifically the targets for the search of conformational space, the basepair step parameters for the 8 DNA oligomers found in crystal complexes with TBP were calculated, and 99% confidence intervals were constructed from the distribution of the geometrical parameters. These intervals, shown in Table 1.3, define the range of conformational variability compatible with the formation of a complex with TBP. These ranges of values for the geometrical parameters were compared to the range contained between the mean ± 1 standard deviation of the values observed for all the basepair steps from the free DNA simulations. These comparisons identify parameters that are outside the thermally accessible interval for general sequence DNA (Olson 1995) (shown in the first line of Table 3.3). The same procedure was followed with the intervals obtained from the DNA molecule in **athdna**, producing a very similar set of relevant parameters, all of which are listed in Table 4.2. These are expected to be useful selectivity indicators because good TBP substrates will be

more likely than poor binding sequences to visit these conformational ranges. Nevertheless, at this point it is impossible to know which ones are necessary, and which are a consequence of the others.

These selected parameters and their conformational ranges were used as filters for the conformations generated in the simulations, and the number of times that each different basepair step visited each of these discriminant conformational ranges was counted; the basepair steps were classified according to their flanking sequences, resulting in the tetrads listed in Table 4.2. Those properties that were found to have as the most frequent visitor (rated by a χ^2 score) a step that binds to TBP, were considered as "selectivity determinants" which are probably used by TBP to identify productive binding sites. These are assumed to be necessary conditions for binding.

From both selection schemes (crystal structure and simulation derived), only two selectivity determinant sites emerged from the analysis: the positive *slide* that is apparently required at both basepair steps 3 and 5, which flank the central basepair step. This agrees with the finding by Suzuki and collaborators (Suzuki 1996), who point out that positive *slide* at these two positions is required to align the floor of the minor groove for recognition by the sidechains of TBP.

The sequence that can access most efficiently the particular range of positive *slide* required at these steps is TA. Because the occurrence of this geometry is extremely small, it is possible that the slow binding of TBP (Coleman 1995; Hoopes 1992; Parkhurst 1996; Perez-Howard 1995; Petri

1995; Starr 1995) may be explained by the low effective concentration of DNA competent for binding that is suggested by such a small population.

As a consequence of their being straight on average and rigid (see Figure 3.8 and Table 4.1), and hence not predisposed to bend, RR steps are absent from the best tetrad column. Notably, sequences that are not bound efficiently by TBP do appear as best tetrads in steps of the TATA box, especially GC and CG steps. In this case, the reason these tetrads are selected over all the others is because the distributions themselves are centered closer to the requirements imposed by TBP, not because of an increase in flexibility, as indicated by the standard deviations for all the basepair step parameters listed in Table 4.1. The fact that GC and CG steps display adequate conformations for being TBP substrates suggests that TBP selectivity is first based on the absence of steric clashes at the interface, and that it is then modulated by the average or transient geometry of the different basepair steps.

Table 4.2 Basepair step properties contributing to selectivity

		crystal structure derived		simulation derived	
step	parameter	best tetrad: % occurrence	worst tetrad: % occurrence	best tetrad: % occurrence	worst tetrad: % occurrence
	rise	gcgc: 9.90	tata: 0.03	ataa: 0.06	gggc: 0.00
1	roll	cgcg: 5.82	aaag: 0.00	cgcg: 2.53	tata: 0.05
	twist	cgcg:13.96	aggg: 0.95	cgcg:43.45	aggg: 7.48
2	shift	atag:30.33	atat: 2.51	-----	-----
	twist	cgcg: 1.93	aaaa: 0.02	cgcg:37.00	aggg: 5.66
3	**SLIDE**	aTAt: 0.13	gggc: 0.00	aTAt: 2.19	tata: 0.11
	twist	-----	-----	cgcg:25.85	aggg: 2.56
	slide	atat: 0.57	gggc: 0.00	atat: 2.20	tata: 0.11
4	roll	cgcg:24.43	aaag: 0.13	cgcg:49.42	aaag: 0.34
	twist	cgcg: 0.51	gcgc: 0.01	cgcg: 2.16	aaaa/g: 0.01
	SLIDE	aTAt: 0.00	tata: 0.00	aTAt: 0.36	tata: 0.01
5	roll	cgcg:32.61	aaag: 0.28	-----	-----
	twist	cgcg:32.99	gcgc: 5.09	cgcg:28.21	aggg: 3.16
	slide	atat: 0.88	tata: 0.04	atat: 0.61	gggc: 0.00
6	roll	cgcg:32.90	aaag: 0.53	cgcg:51.45	aaag: 0.86
	twist	cgcg:38.69	aggg: 6.97	-----	-----
	slide	gcgc:25.73	gggc: 0.21	-----	-----
7	rise	gcgc: 2.28	tata: 0.00	zero hits	zero hits
	roll	cgcg: 3.60	aggg: 0.03	cgcg:12.40	aaag: 0.00

steps are defined in Table 1.2; best tetrad: tetrad that visits the conformational range defined by the crystal structures or the simulation of TBP/DNA complexes with the highest frequency, scored by a χ^2 test; worst tetrad: tetrad that visits the conformational range defined by the crystal structures or the simulation of TBP/DNA complexes with the lowest frequency, scored by a χ^2 test (double entries shared the lowest frequency and χ^2 score). The parameters in **UPPERCASE** are those that selected as the best tetrad one that has actually been crystallized in a complex with TBP. ----- means that that parameter is not outside the thermally accesible range for general sequence DNA.

Step 4 in the TATA sequence, for which the search for conformations consistent with complex formation failed to pick out any tetrad found in the available crystallized complexes or in the simulation of the complex, is the only one where H bonds are formed between TBP and the minor groove of the DNA. This suggests that the H bonds might pay for the energy cost of unwinding, opening the minor groove and sliding the basepairs to induce the complex geometry. A rough estimate of the free energy penalty for this structural transition in the basepair step can be obtained from a statistical analysis of the simulation results, from

$$n_i / n_{ground} = \exp(\Delta G/RT)$$

where n_i is the number of conformations appropriate for this interaction, n_{ground} is the number of conformations inside the thermally accessible range, and ΔG is the free energy difference between the conformation found in the complexes with TBP and the conformation for the most populated state. The calculation for the two tetrads found in crystals with TBP indicate that tATa always has a lower penalty than tAAa (4.5 versus 5.1 kcal/mol for *slide*; 1.3 versus 2.0 kcal/mol for *roll*; 3.6 versus 4.2 kcal/mol for *twist*), a trend that matches the relative affinities of TBP for these sequences (Wong 1994). Furthermore, tATa can make six hydrogen bonds to TBP, while tAAa can only make five. Assuming additivity and a 1.5 kcal/mol contribution from each H bond (Jen Jacobson 1995), the formation of six H bonds would balance the nearly 9.5 kcal/mol energy penalty for the distortion required in the tATa steps. Notably, GC steps were unable to

achieve the *slide* requirements for the central basepair step, even after 1 ns of simulation; the reason for this is elusive, since modeling of this step at position 4 in the ATH2/mlp complex does not cause any clashes between the bases or sugars that would explain the reluctance to achieve that particular conformation.

4.4 The Ideal TBP Binding Site

Combining the findings from the analysis of the average properties of the basepair steps with those from the transient conformational properties, a canonical TBP binding site of the form: YRTATAYR is proposed. The considerations underlying this definition of the canonical binding site further lead to the prediction that the closer a DNA sequence is to this consensus, the better its TBP binding ability should be.

The prediction is probed by the relation between the adherence of the calculated sequences to the canonical TBP binding site, and the available data for binding affinity determined by Wong and Bateman (Wong 1994). As shown in Table 4.3, **at** and **6t** match the consensus almost completely, and are found to have the highest affinity for TBP. Because **mlp** is missing the second TA of the canonical sequence, it should bind less well; its affinity is found to be 3 to 4 times lower than that of **at** or **6t**. While **2c** and **7g** are assumed not to be bound by wild type TBP (apparently, **7g** can be bound by TBP, with a lower affinity (Brenowitz, personal communication)), the binding constants measured with the appropriate TBP mutants should be in the same range as that for **mlp**.

Table 4.3. Comparison to equilibrium binding constants (from (Wong 1994))

consensus sequence	**YRTATAYR**	$K_{eq}(10^{-9})$
sequences:		
mlp	**TATAAAAG**	3.7
2c	**CATAAAAG**	n.a.
7g	**TATAAGAG**	n.a.
6t	**TATATAAG**	1.1
at	**TATATATA**	1.4

the sequences indicate both **average** and transient specificity determinants; n.a. : not available

In conclusion, both average and transient conformational properties of DNA seem to be part of the selectivity determinants used by TBP to chose its binding sites, there being no steric clashes (**gc** is a good substrate as far as *rise*, *roll* and *twist* conformational properties are concerned (Tables 4.1 and 4.2), and also comes close to an A-DNA conformation (Figure 4.1), but it is a poor binding site). While **mlp** showed a tendency to become closer to an A-DNA like conformation after 2 ns of simulation, as hypothesized by Guzikevich-Guerstein and Shakked (Guzikevich-Guerstein 1996), there is no clear correlation between a tendency to become A-DNA like and promoter strength from the simulations carried out in this work (even superimposed on the "natural" tendency of the CHARMM potential to promote A-DNA conformations (Yang 1996)). In agreement with Juo et al. (Juo 1996), alternating TA sequences are selected because of the propensity for unstacking at TA steps;

on the other hand, the results shown here suggest that CG steps have an even greater tendency to unstack, contrary to their hypothesis. Also, AT steps here have the clear function of opening the minor groove as a consequence of the persistent positive *roll*, again in opposition to their conclusions. Crystal structure analysis suggests that RY steps will *roll* closing the minor groove, and this has been explained by a clash between the rims of the bases in the major groove and the sugar in the backbone (Suzuki 1996; Suzuki 1996). The simulations described here show a clear tendency for these RY steps to *roll* opening the minor groove, indicating that such a clash can be avoided efficiently.

The importance and sequence preferences of *slide* pointed out by Suzuki et al. (Suzuki 1996) also result from the present work, but from the analysis of transient properties; this could be the trigger for binding, and could also be related to the slow binding of TBP. The correlation between adherence to the proposed consensus binding site and the affinity constants (Table 4.3) further indicates that the properties identified are indeed related to sequence specificity.

An interesting corollary is that different sequences might exploit different subsets of the properties found here. For example, **mlp** satisfies the derived consensus sequence at the 5' half of the TATA box, and locks the sugars of the 3' half in a conformation similar to the one needed in the complex; conversely, **at** matches the derived consensus perfectly, but keeps the sugar flexibility. Still to be determined are sequence dependent hydration properties and interactions with counterions, which may further modulate the binding affinity.

5 TBP Dynamics and the Binding to DNA

5.1 Global structural comparison between free and bound TBP

The crystal structures of free and bound TBP showed that the protein conformation changes very little between these two states (Geiger 1996; Juo 1996; Kim 1994; Kim 1993; Nikolov 1996; Nikolov 1995; Tan 1996). This is also found in the **ath** and **athdna** simulations, as is evident from the oscillatory pattern in the rms plot *versus* time for the Cα atoms of **ath** compared to the structure in the crystal complex (Figure 3.2). Figure 5.1 shows the superposition of the protein Cα traces for the average structures of TBP in **ath** and **athdna**. This fit has an rms of 1.1Å over all the backbone atoms. There are small differences throughout the whole structure, and the stirrups are slightly more open in the complex than in the free protein (as expected from Figure 3.4).

While the structures are very similar, the simulations revealed differences in the dynamics of the protein. Figure 5.2 is a plot resulting from subtracting the Cα atom fluctuations of free TBP from those of TBP in complex with DNA (from Figure 3.10), as a function of residue number. The residues with changes greater than 0.4 $Å^2$ are labeled in the plot. As expected, the two stirrups (N35-A40 and P127-P131) are more rigid in the complex than in free TBP, due to the interaction with DNA. The C-terminus of helix 2' is also stabilized by binding to DNA. An unexpected finding is the increase in mobility of residues R30 an N31,

located in the loop between helix 1 and strand 2. Close examination of the B-factors from the crystal structures also reveals this change, lending credibility to the results obtained from the simulations.

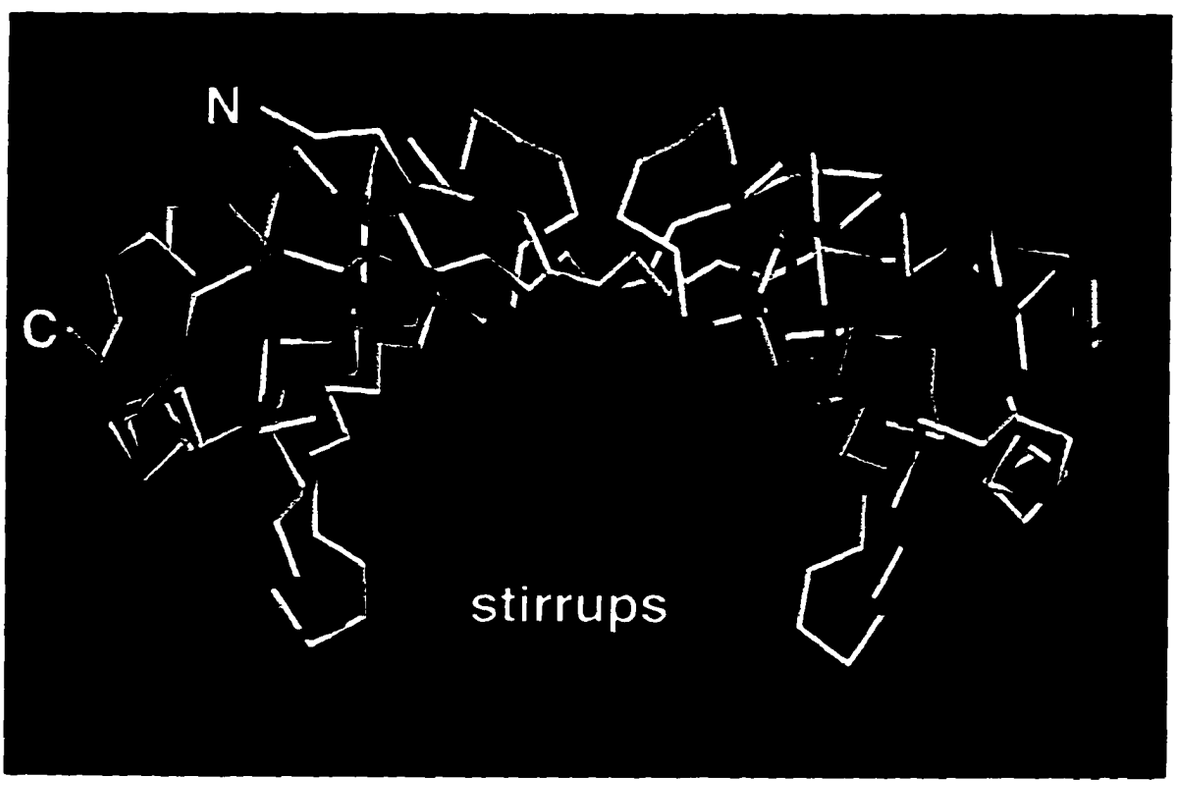

Figure 5.1 Comparison of the Cα traces for the average structures of free TBP (white) and TBP complexed to mlp (red / gray). The stirrups and the N and C-termini are labeled.

The symmetry related loop in the C-terminal half of TBP did not show such order-disorder transition, but instead has higher mobility in helix 2', close to P172, which is probably related to fixing the end of the helix. The loop connecting the two domains also displays increased mobility in the complex;

this could be related to the ease of accommodating insertions in this area of TBP (see the alignment in Table 1.1). The increase in mobility is interesting because it is a favorable entropic contribution to binding, and also because it could be related to the decrease in heat capacity upon binding (Sturtevant 1977). This is reminiscent of the unfolding of an α helix in BamHI when it binds to DNA (Newman 1995), although here there is no change in secondary structure, just an increase in the fluctuations.

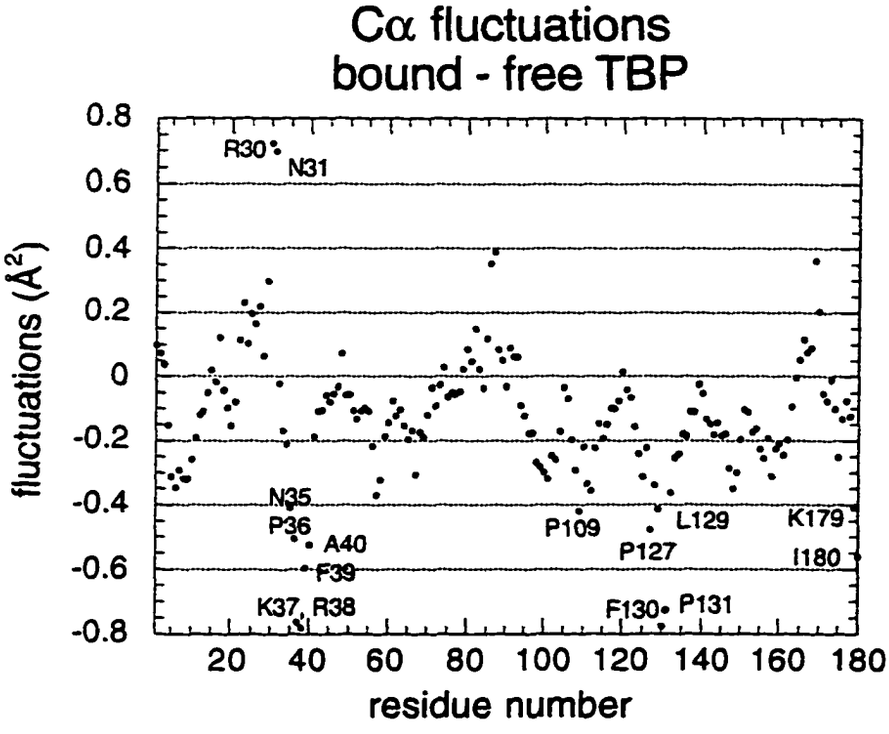

Figure 5.2 Comparison of Cα fluctuations between free TBP and TBP in complex with DNA.

The side chains of some TBP residues pointing towards the core of the protein show substantial motion, both when free and when bound to DNA, as evidenced by numerous rotamer changes of these side chains throughout the simulations. These residues are shown in Figure 5.3. They are clustered, and tend to map to the interfaces between helix 1(1'), helix 2(2') and the body of the β-sheet, especially strands 3(3') and 4(4'). There are two internal water molecules, and mobile residues (C63, L144) are located near them. The interior of proteins is assumed to be well packed (Branden 1991), so these side chain rotations must be correlated to each other and/or to protein backbone motions; this is probably related to their being localized to clusters at the interface between secondary structural elements. The residues that are capable of rotating in both **ath** and **athdna** simulations are I10 (at the interface between the two domains), M44, I55 (strands 3 and 4, close to helix 1 and the C-terminal end of helix 2), I110 (in helix 1'), L174 (in helix 2'), and I142, L144 and I154 (in strands 4' and 5', interacting with L174 and I110). Contrary to the idea of a rigid C-terminal domain derived from the comparison of B-factors of the free and bound TBPs (Kim 1993), the rotations of these side chains suggest that both domains of the protein are rather flexible.

The distribution of the most mobile residues in the two domains is not symmetric, especially at the interface between helix 1(1') and the rest of the protein (see Figure 5.3). The remarkable feature of these side chain rotations is that their distribution in the molecule changes from that in free TBP to a different one in the complex. The interface between helix 1' and the body of the protein is

more mobile in **ath** than in **athdna**, and the opposite happens at the interface between helix 1 and the body of the protein, where **ath** has no rotating side chains but **athdna** has rotational activity between L20, I25, I55 and M44 (see Figure 5.3). This finding extends the analysis presented by Suzuki et al. (Suzuki 1996), where the asymmetry between the domains was ascribed to the difference in side chain volume at the interface between the β-strand and helix 2(2'), which was assumed to make the C-terminal domain more rigid (larger side chains) than the N-terminal domain (smaller side chains). Incidentally, L144 is the center of the C-terminal domain, as defined by Suzuki et al. (Suzuki 1996), and it was found to jump many times between two different χ_1 and χ_2 rotamers. The center of the N-terminal domain is A53, which does not have distinguishable rotamers, but its neighbors are very active in rotational transitions (I55 and M44). This suggests that large side chain volume is not always correlated with rigid structures. In addition to its possible effect on the symmetry of binding, the flexibility resulting from these side chain rotations could also be pertinent to the ability of TBP to recognize different variants of TATA boxes, as documented for the *trp* repressor (Gryk 1996), allowing the shape of the protein to adjust to small perturbations in the minor groove width or DNA bending angle.

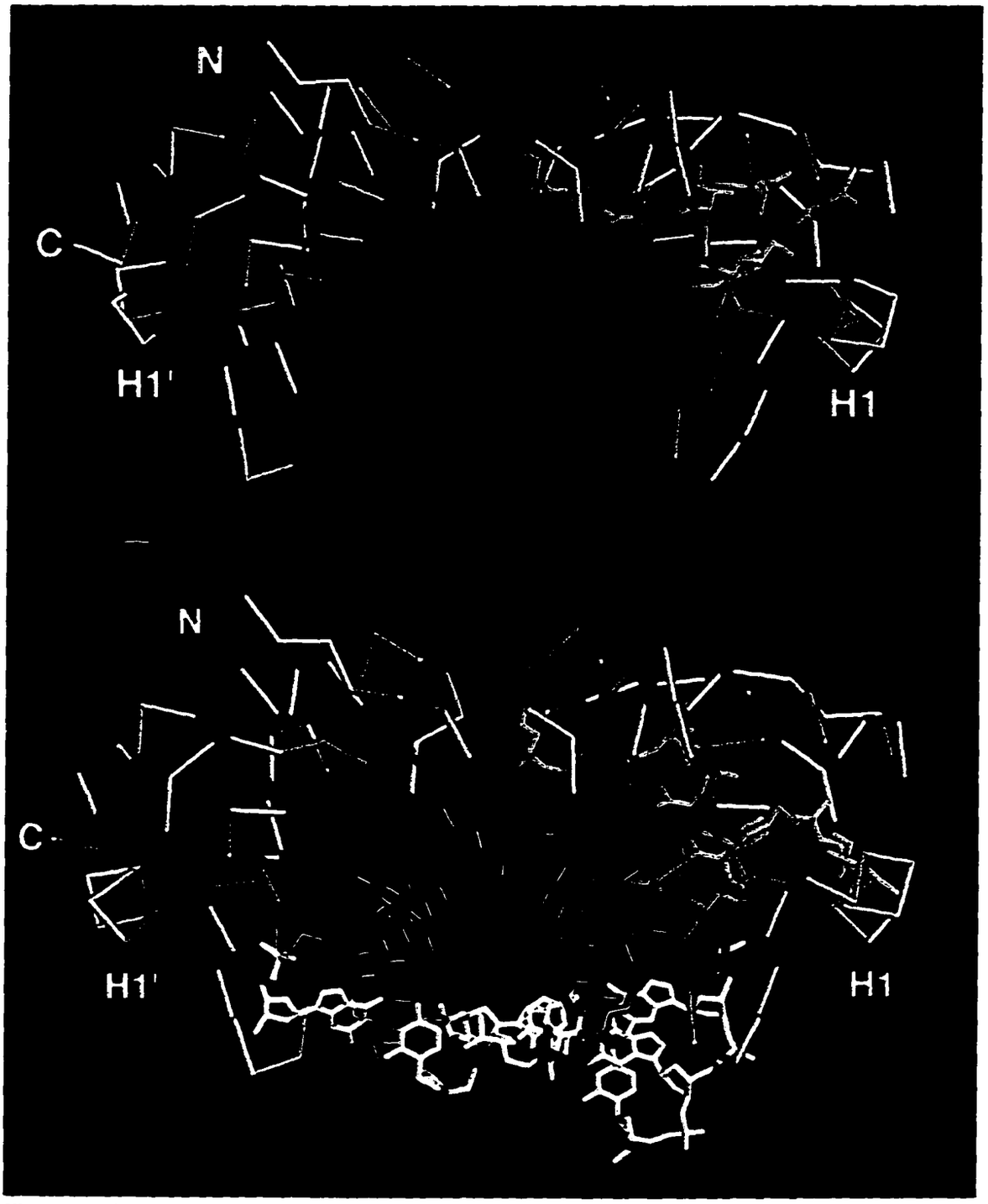

Figure 5.3 Residues that explore various side chain rotamers in the protein cores of **ath** (top) and **athdna** (bottom). Cα trace (green/light gray), DNA (blue/darkest gray (TATA box) and white), water (red/dark gray), N-terminal domain side chains (yellow/lightest gray (top clockwise: I10, C63, I46, L16, M44, I55, V83; bottom: I10, V14, M61, M44, I55, L20, I25 -- L54 and I43 interact with DNA)); C-terminal domain side chains (magenta (top clockwise: I110, L112, L115, L133, L174, I142, L144, I154, I100; bottom: I110, L174, I170, I152, I142, L144, I154 -- L145 interacts with DNA)). N-, C-termini, and helices 1(1') are labeled.

The significance for DNA binding of this change in the pattern of motions requires further analysis of the dynamics of the protein and the complex, but it clearly shows that the protein is indeed asymmetric insofar as its dynamic properties are concerned. The simulations carried out in this work include an asymmetric DNA binding site, which complicates the analysis of the dynamics of the protein when complexed to the DNA (it is not trivial to separate the influence of the dynamics of the DNA on the dynamics of the protein). Also, a clearer mechanistic picture should arise from the analysis of global motions of the protein, both free and bound to DNA; this could probably be achieved through a normal mode or quasiharmonic analysis over the stable parts of the trajectories (de Groot 1996; Hayward 1995; Hayward 1993; Steinbach 1996).

5.2 Analysis of TBP contacts with DNA

The interface between TBP and DNA is mostly hydrophobic and symmetric when comparing the 5' half to the 3' half of the TATA box (Juo 1996; Kim 1994; Kim 1993). A reduction in side chain mobility would be intuitively congruent with the side chains becoming trapped at the molecular interface (Ribas de Pouplana 1996), and would constitute a straightforward way to decrease the heat capacity of the system. However, an increase in side chain mobility upon complex formation has not been characterized before. An example of a mobile interface is provided by homeodomains (Billeter 1996; Hirsch 1995; Wilson 1995) and the steroid receptors (Gewirth 1995; Schwabe

1993; Schwabe 1995) bound to DNA, which show side chain flexibility at the interface of an α-helix with the hydrated major groove of the DNA, where one side chain makes contacts with up to three bases in the recognition site. In contrast, the TBP-minor groove interface is described as anhydrous and tightly packed (Kim 1994). Since side chain flexibility could be relevant for the understanding of the source of the very large change in heat capacity upon TBP binding to DNA (Petri 1995), the population of side chain rotamers for the residues in contact with DNA were compared between **ath** and **athdna**. The 5' to 3' symmetry was also examined, yielding novel insights into a possible mechanism for directional binding. The presentation of the results is ordered from the ends of the TATA box to the center, to facilitate the analysis of the symmetry of the interactions.

5.2.1 Hydrophobic contacts

The first symmetry break occurs at the ends of the recognition site (defined crystallographically as eight basepairs of the TATA box which are contacted by any protein side chain), where the C-terminal domain has P131, and the N-terminal domain has A40. P131 interacts with the adenine of basepair 1 of the TATA box (see Table 1.2), while A40 contacts the sugar of the guanine outside the TATA box (basepair 9), in agreement with chemical protection assays (Lee 1991). This difference was proposed by Juo et al. (Juo 1996) to be a determinant for the orientation of binding, but this has not been proven as yet. A40 is P40 in archaea, N40 in PFA and R40 in DROT, displaying

greater variability than position 131, where the only deviant is A131 in PFA. This substitution pattern indeed suggests an asymmetry in the TBP interaction requirements at the ends of the TATA box, and probably is related to the tolerance for any basepair at the 3' end of the TATA box (assuming, of course, unidirectional binding). In the exploration of a dynamic underpinning for the interactions, the two ends of the protein segment do not contribute, as alanine has no side chain degrees of freedom, and the proline does not repucker significantly; they will not be discussed further.

The kinks in the bound DNA are caused by the partial insertion of two pairs of phenylalanine residues. One of each pair (F39, F130) stacks against the base at the 3' of the kink, while the other partner (F56, F147) stacks against the sugar of the basepairing nucleotide. Both F39 and F130 are more disordered in **ath** than in **athdna** (Figure 5.4). F39 populates two rotamers in χ_1 (only one transition observed) and χ_2 (with multiple transitions), and F130 has a broad ψ distribution that is narrowed upon binding (probably related to the decrease in Cα fluctuations for this residue); also, F130 changes its χ_1 rotamer from +60° to 180° upon binding. F39 is the only inserting phenylalanine that is not 100% conserved through evolution, being I39 in PFA; furthermore, an F39L mutant is viable in SCE (Reddy 1991), suggesting that the van der Waals interactions between these aliphatic side chains and the base are comparable to the aromatic stacking. This contention is also supported by other examples of minor groove binding proteins, such as SRY (Werner 1995) and LEF-1 (Love

1995), which also use aliphatic side chains to stabilize the kinks between basepair steps.

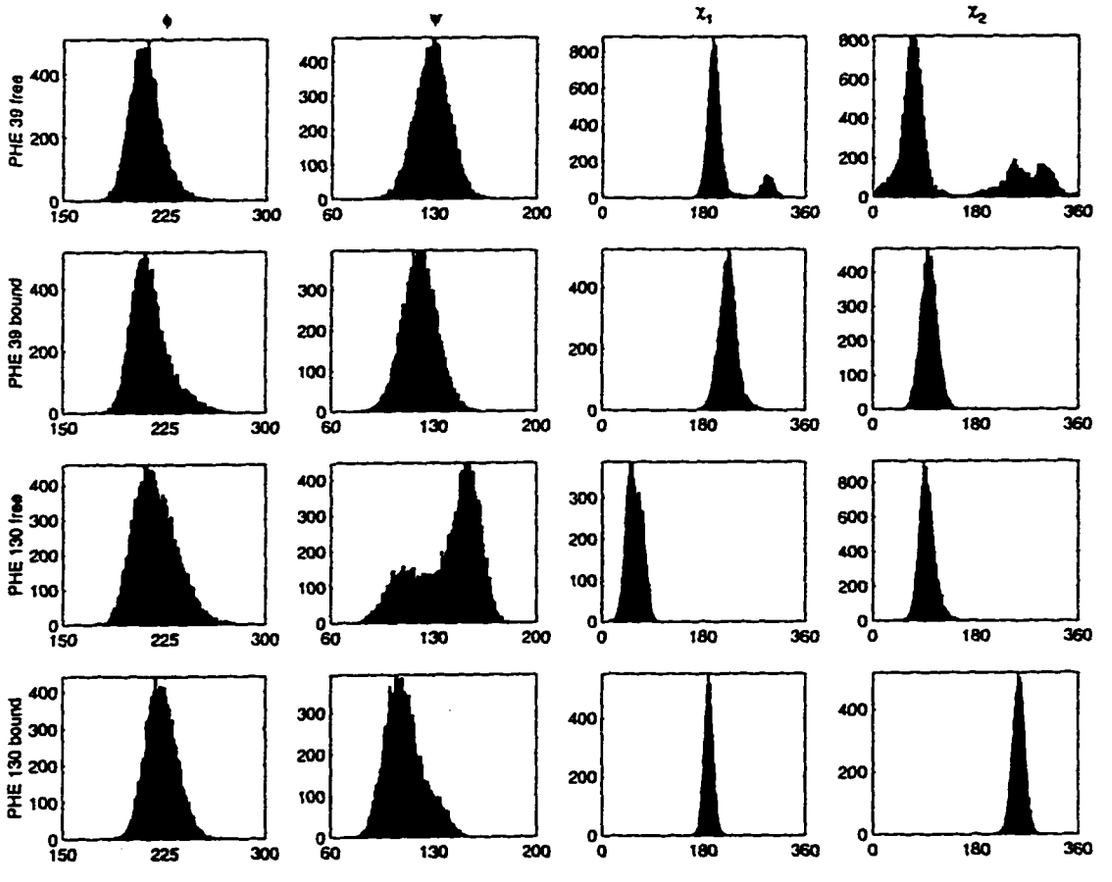

Figure 5.4 Comparison of rotamer populations for the inserting phenylalanines F39 and F130 in **ath** and **athdna**. The residue is indicated to the left of each row.

The next line of interactions involves L54 and L145, which read the C2 position of the adenines at basepairs 7 and 2, respectively, and are 100% conserved. In **ath**, L54 populates two rotamers for χ_1 and χ_2, but only one transition between the rotamers was observed. L145, on the other hand, rotates frequently over χ_2. The interface between TBP and DNA was described as tightly packed (Kim 1993), so it was very surprising to see that both L54 and L145 rotate over χ_2 even in the complex with DNA (Figure 5.5). The crystal complexes were examined for evidence of such a side chain rotation; for example, the two copies of SCE bound to the CYC-1 promoter (NDB structure PDT012 (Kim 1993)) have different rotamers for both L54 and L145. These structures were refined to a resolution of 1.8Å, which reduces the likelihood of this being an artifact of the refinement protocol. In the case of L54, while two χ_2 rotamers are sampled in both simulations, the frequency of the rotations is higher in the complex than in the free protein (where only one coupled χ_1 and χ_2 transition was observed). F39 appears to be hindering the rotation in **ath**; once F39 moves to insert into the DNA, L54 is free to rotate. This is both entropically favorable, and might contribute to the decrease in heat capacity. The flexibility observed for these side chains might also explain why in SCE bound to DNA, these residues are described as reading basepairs 3 and 6 (Arndt 1994; Kim 1993). L54 is one of the Arndt mutations which is capable of binding to TATAAGA, and this sequence can also be bound by wild type TBP, with ~1 kcal/mole less affinity (Brenowitz, personal communication). It is very

tempting to speculate that the exocyclic amino group of the guanine is hindering the rotation of this side chain, causing a decrease in entropy and hence destabilizing slightly the complex; also, this sequence displays an almost negligible change in heat capacity upon binding (Brenowitz, personal communication). It is possible that interactions at the interface, like the hypothesized restriction in rotation of the side chain of L54 because of steric hindrance caused by an exocyclic amino group, are responsible for the sequence dependent differences in the heat capacity change (Brenowitz, personal communication).

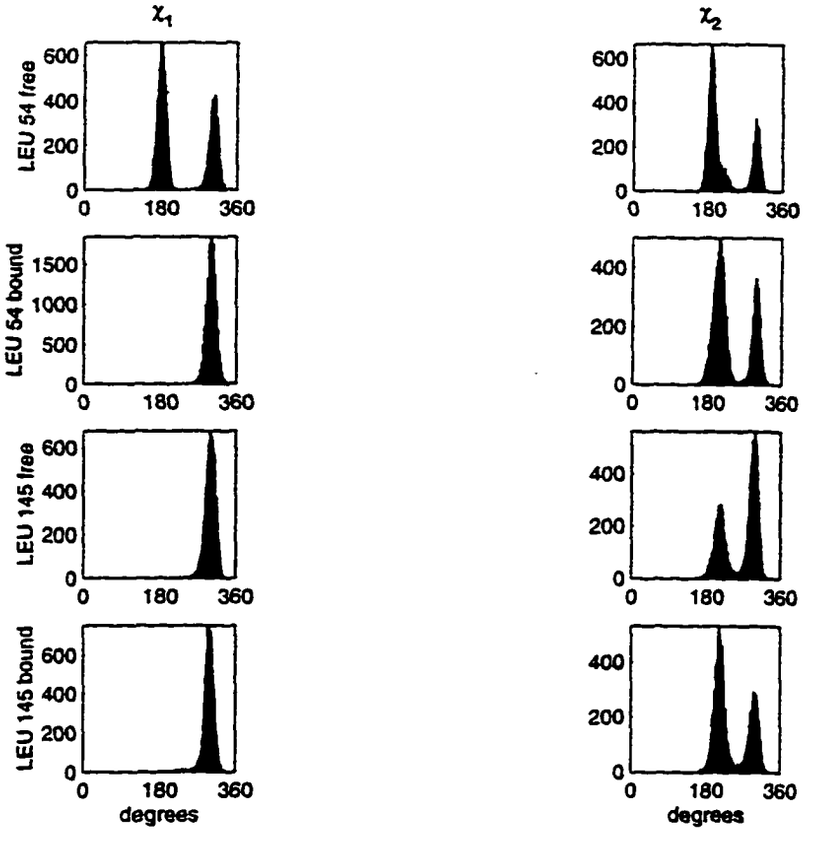

Figure 5.5 Comparison of rotamer populations for L54 and L145 in **ath** and **athdna**. The residue is indicated to the left of each row.

The four valines (V11, V62, V101, and V153) contacting the edges of the bases have symmetrical interactions, and their conformation does not change upon binding. They will not be discussed further.

TBP interacts extensively also with the backbone of the DNA. I43 and symmetry related I134 bind to the sugars of the antisense strand of basepair 7 and the sense strand of basepair 2, respectively. In both cases, binding to the DNA selects a particular rotamer of the manifold sampled when TBP is free (Figure 5.6). The residual flexibility in χ_2 can be understood from the substitution pattern for these residues: position 43 is a valine in HASA and MJA, and position 134 is a valine in EHI, PFA, HASA and MJA; this indicates that the interactions beyond Cγ are probably not necessary for stability.

Two symmetry related glutamines (Q8 and Q98) lie above the sugars for the antisense strand of basepair 3 and the sense strand of basepair 6. Q98 is substituted by glutamic acid in PFA; the pattern for Q8 is more complicated: valine in EHI and MJA, histidine in PFA, and glutamic acid in the rest of the archea. This site seems to use the whole side chain for interacting with the DNA, because all the dihedral angles are frozen into one conformer, compared to the practically free rotation in free TBP for χ_3 (Figure 5.7). Also contacting basepairs 3 and 6, but on the opposite strands to these glutamines, are T52 (opposite Q98) and V143, across from Q8. Their interactions appear to be symmetrical, and there is no change in rotamer populations upon binding to DNA.

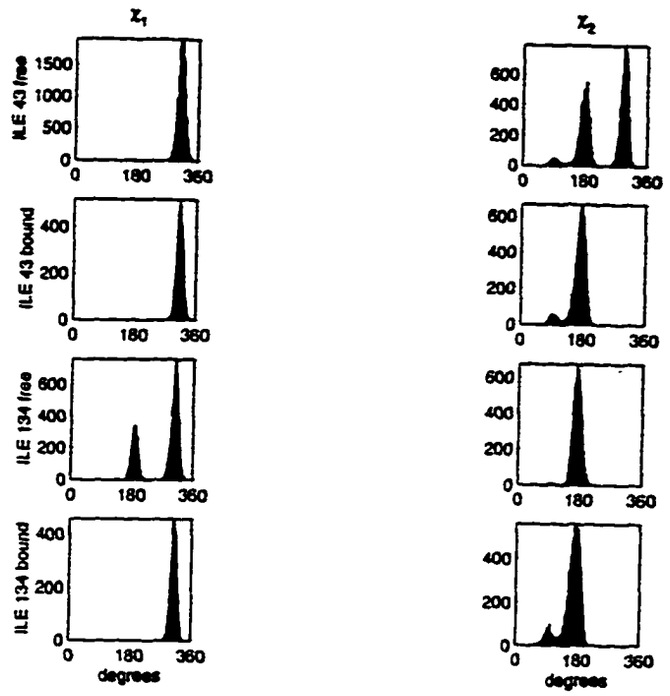

Figure 5.6 Comparison of rotamer populations for I43 and I134 in **ath** and **athdna**. The residue is indicated to the left of the rows.

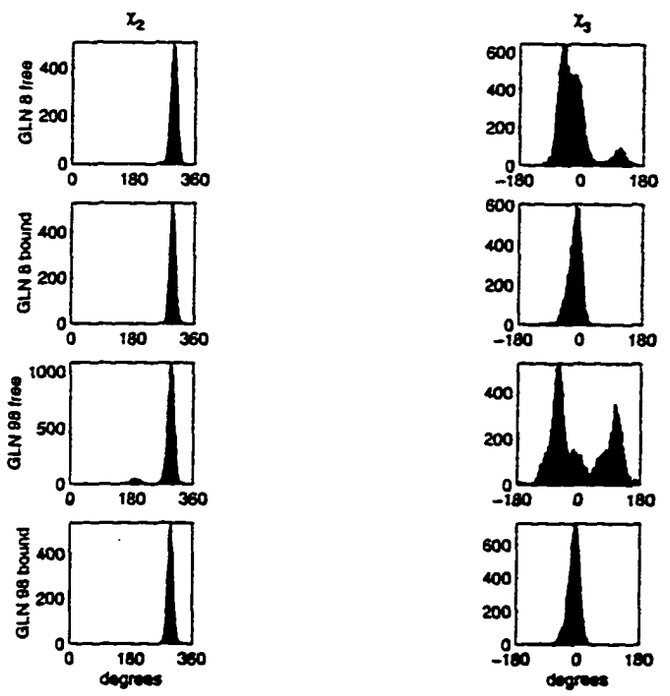

Figure 5.7 Comparison of rotamer populations for Q8 and Q98 in **ath** and **athdna**. The residue is indicated to the left of the rows.

5.2.2 H bonds between N and T residues and the DNA bases

The central basepair step in the TATA box (step 4, formed by basepairs 4 and 5) is the only location in the complex where asparagine (N9 and N99) and threonine (T64 and T155) side chains make H bonds to the edges of the DNA bases. Other threonines are found interacting with the sugars and the phosphates (for example, T13 and T52). In the complex between ATH2 and mlp, T64 does not engage in a H bond to the base, but rather interacts with the sugar of the next residue. This is probably due to the asymmetry of the mlp sequence, which has an AA step at this position; T64 will H bond to the base if presented with an AT step (Kim 1993). It should be recalled that T64 is substituted by a valine in MJA, indicating that the H-bonding ability at this position may not be essential for binding.

The two asparagines seem to be already in the right rotamer to interact with DNA, because there was no change in the populations of either χ_1 or χ_2. On the other hand, both threonines have different χ_1 rotamers in **ath** and **athdna**, (T64: t to g^-; T155: g^- to t) without any interconversion between the two.

N9 H bonds to the O2 of thymines in the antisense strand of step 4, while N99 interacts with the N3 of the basepaired adenines. The plots for H-bond distance and angle in Figure 5.8 reveal a marked asymmetry in the interactions between N9 and N99. The H bonds to the O2 have a much better geometry than those to the N3, both in distance and in angle. As expected, there is a small *tilt* component (-5.8°) for step 4 (see Table 3.4), reflecting the uneven compression of the basepair step. T155, on the other hand, makes a H bond to the adenine N3 in basepair 4, with acceptable distances and angles.

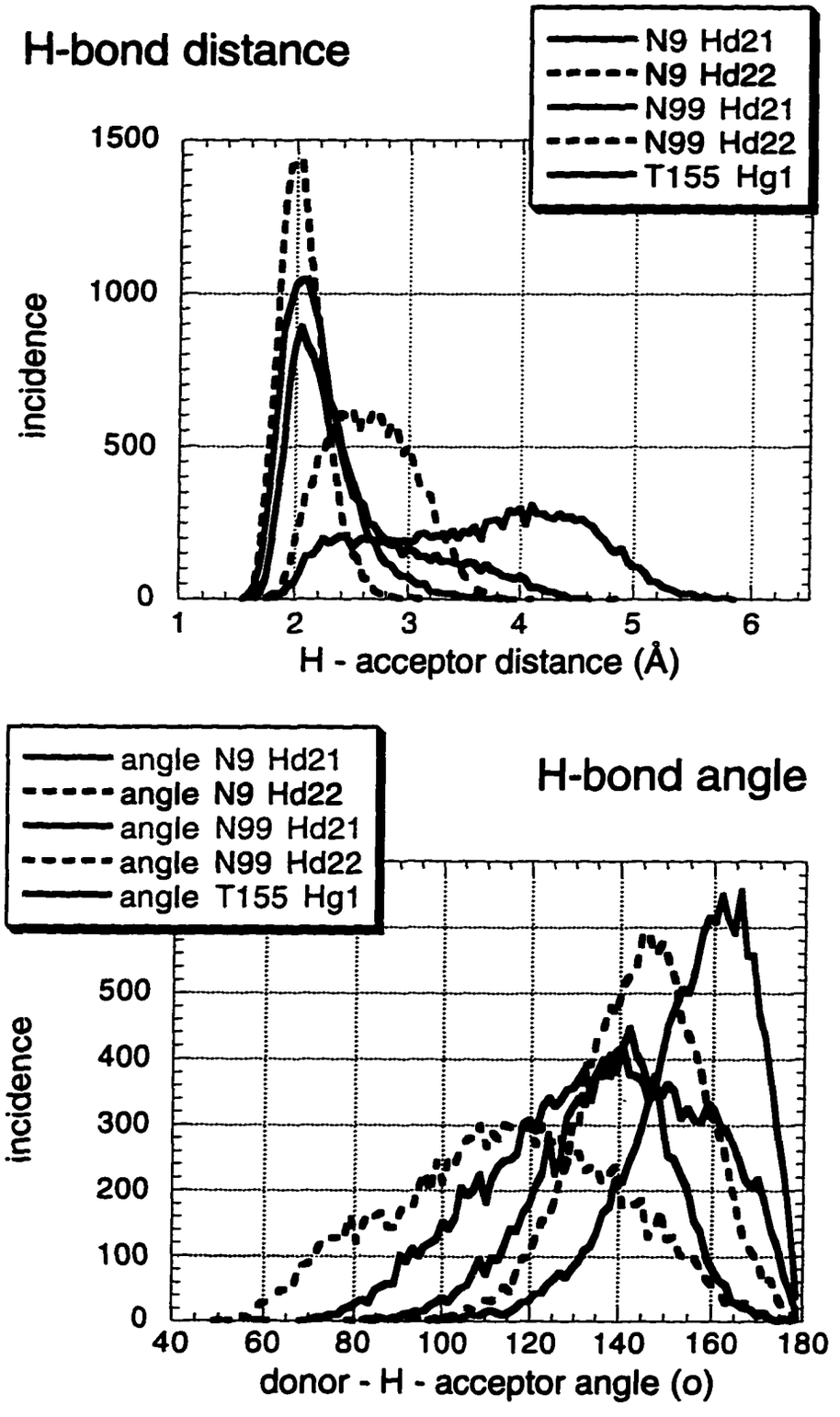

Figure 5.8 Distributions of H bond distances (top) and angles (bottom) beween N9, N99 and T155 and the central basepair step of the TATA box in **athdna**.

Examination of the dynamics trajectory revealed that Hδ22 of N99 seems to be making a H bond to the Oγ1 of T155, thus explaining the large spread in the distances of Hδ22 of N99 to the base. It appears that O-H-N H bonds are favored over N-H-N and O-H-O, especially if the structure between SCE and the CYC-1 promoter is included in the analysis (Kim 1993). In this case, there is an AT step instead of an AA step at the central position, and T64 makes a H bond to the N3 of the adenine in basepair 5 (that is an N-H-O bond); this same threonine prefers consistently not to interact with a thymine O2 (which would make an O-H-O bond) in the ATH2/mIp (Kim 1994) and HSA/mIp complexes (Nikolov 1996), turning to interact with a sugar.

5.2.3 Salt bridges and H bonds to the phosphates

Ethylation interference assays (Lee 1991) indicate that the following phosphates are protected by TBP (indicated in boldface):

phosphate #	1 2 3 4 5 6 7 8 9
sense strand	5'pTpApT**p**A**p**A**p**A**p**ApG**p**Gp3'
antisense strand	3'pApT**p**ApT**p**T**p**T**p**T**p**CpCp5'

The salt dependence of the binding constant is consistent with the release of 3.5 counterions from the surface of the DNA (Petri 1995), which is less than the ten protected phosphates suggested from the ethylation interference. The crystal structures of TBP/DNA complexes can explain the results from the ethylation interference, from the contacts of various lysine, arginine, serine, threonine and glycine residues located near the phosphates. A possible

explanation for the difference in number between the released counterions and the protected phosphates by the protein could be that the interactions between the side chains and the phosphates are short lived. In order to test this hypothesis, **athdna** was probed for TBP hydrogen atoms within 2.4Å of the phosphate oxygens. This is a generous upper bound for H bonding, and will also include van der Waals contacts between the side chains and the phosphate oxygens. This query resulted in a list of phosphates contacted by TBP side chains (Table 5.1), and in a list of the residues that interact with the phosphates. This list is shown in Table 5.2. The percentage of the time that such interactions exist is also included in these tables, and is defined as the number of trajectory frames for which a contact was observed, divided by the total number of frames analyzed. Hence, values over 100% mean that more than one residue was interacting with the particular site (Table 5.1), or that a particular side chain was making contact to more than one phosphate (Table 5.2).

The entries marked in boldface are those phosphates protected by TBP from ethylation; the patterns of protection do not match perfectly, indicated by the absence of an entry in the table for phosphates in boldface and the presence of entries for phosphates that are not found to be protected experimentally. This discrepancy could be due to the local electrostatic environment of the phosphates, which could attract or repel the ethylation reagent, and which is not taken into accoun in this steric analysis. Nonetheless, the simulation predicts the protection of ~11 phosphates, close to the

experimental finding. Neither pattern is perfectly symmetric, extending more towards the 3' side on the sense strand, probably reflecting the sequence asymmetry in TBP regarding the side chains in contact with these phosphates, especially those located at the stirrups.

Table 5.1 Phosphate oxygens interacting with TBP side chains

sense phosphate (5' - 3')	% of time in contact	% of time in contact	antisense phosphate (3' - 5')
pT		75.56	pA
pA	0.96	0.33	**pT**
pT	102.28		pA
pA	10.68	37.04	**pT**
pA		129.80	**pT**
pA	19.33	29.38	**pT**
pA		0.76	**pT**
pG	109.62		pC
pG			pC

Most of these interactions are with O2P, that is the oxygen lining the minor groove. O1P generally points away from the protein and into the compressed major groove. Four phosphates have very short lived interactions through O1P, and these belong to sense strand 4 and 9, and antisense strand 4 and 5 (the central basepair step and the 3' phenylalanine insertion site), suggesting an increase in the DNA backbone flexibility at these sites.

Table 5.2 TBP side chains as phosphate ligands

TBP residue (hydrogen)	DNA phosphate	% time in contact
R38 (Hδ1, Hη11, Hη12, Hη21, Hη22)	antisense 6 and 7	30.14
R45 (Hε, Hη21)	antisense 5	117.29
K50 (Hγ1, Hδ1, Hε1, Hζ1, Hζ2, Hζ3)	antisense 4	37.04
T52 (Hγ1, Hγ21, Hγ22, Hγ23)	antisense 5	12.51
S58 (Hγ1)	sense 8	44.13
K60 (Hδ1, Hε1, Hε2, Hζ1, Hζ2, Hζ3)	sense 8	65.49
L129 (Hδ11, Hδ12, Hδ13, Hδ22)	sense 2	0.96
I134 (Hγ22, Hδ1, Hδ2, Hδ3)	sense 3	0.15
R136 (Hε, Hη12, Hη21, Hη22)	sense 3 and 4	231.27
V143 (Hγ21, Hγ23)	sense 4	2.77
S149 (Hβ1, Hβ2, Hγ1)	antisense 1	75.37
K151 (Hγ2, Hε1, Hζ2, Hζ3)	antisense 1 and 2	0.52
K158 (Hζ1, Hζ3)	sense 6	19.33

The fact that various hydrogens from the same side chain contact a particular phosphate oxygen suggests that these interactions are formed and broken a number of times throughout the simulation, most likely because of the competition by water molecules for the formation of H bonds. The exchange with solvent also relates to side chain flexibility, because the side chain has to have rotated in order to interact with the phosphate through a different hydrogen. The rotamer populations in Figure 5.9 show that this is indeed the case for the four lysines found to interact with the phosphates (K50, K60, K151 and K158). Mutation of any of these lysines to leucine eliminates binding (Yamamoto 1992), and in particular the mutation of K151 causes a large

perturbation in the electrostatic potential of the protein (Pastor 1995). The pattern of rotamers for these lysines changes in all cases, retaining and even increasing the number of rotamers explored when TBP binds to DNA (see K151 in Figure 5.9). K151 explores only one rotamer in **ath** because it forms a salt link with D105; this interaction is effectively competed by the phosphates in DNA, and could be less permanent if small anions were to be included in the simulation. While this would probably increase the number of sampled rotamers for K151 in the free protein, hence reducing the calculated entropic contribution to binding, the flexibility observed in the complex will most likely remain.

A similar phenomenon occurs for the arginines (R38, R45 and R136). R136 shows only one conformer, in keeping with its being associated with two consecutive phosphates. R38 and R45 keep many rotamer populations upon binding, and they are linked by E33, the flexibility of which is increased upon complex formation (Figure 5.10).

The side chain flexibility displayed by these residues can help to understand the small salt dependence of the binding constant. While many phosphates interact with the protein (Table 5.1), only four are contacted close to 100% of the time, and this number correlates very nicely with the 3.5 ions released.

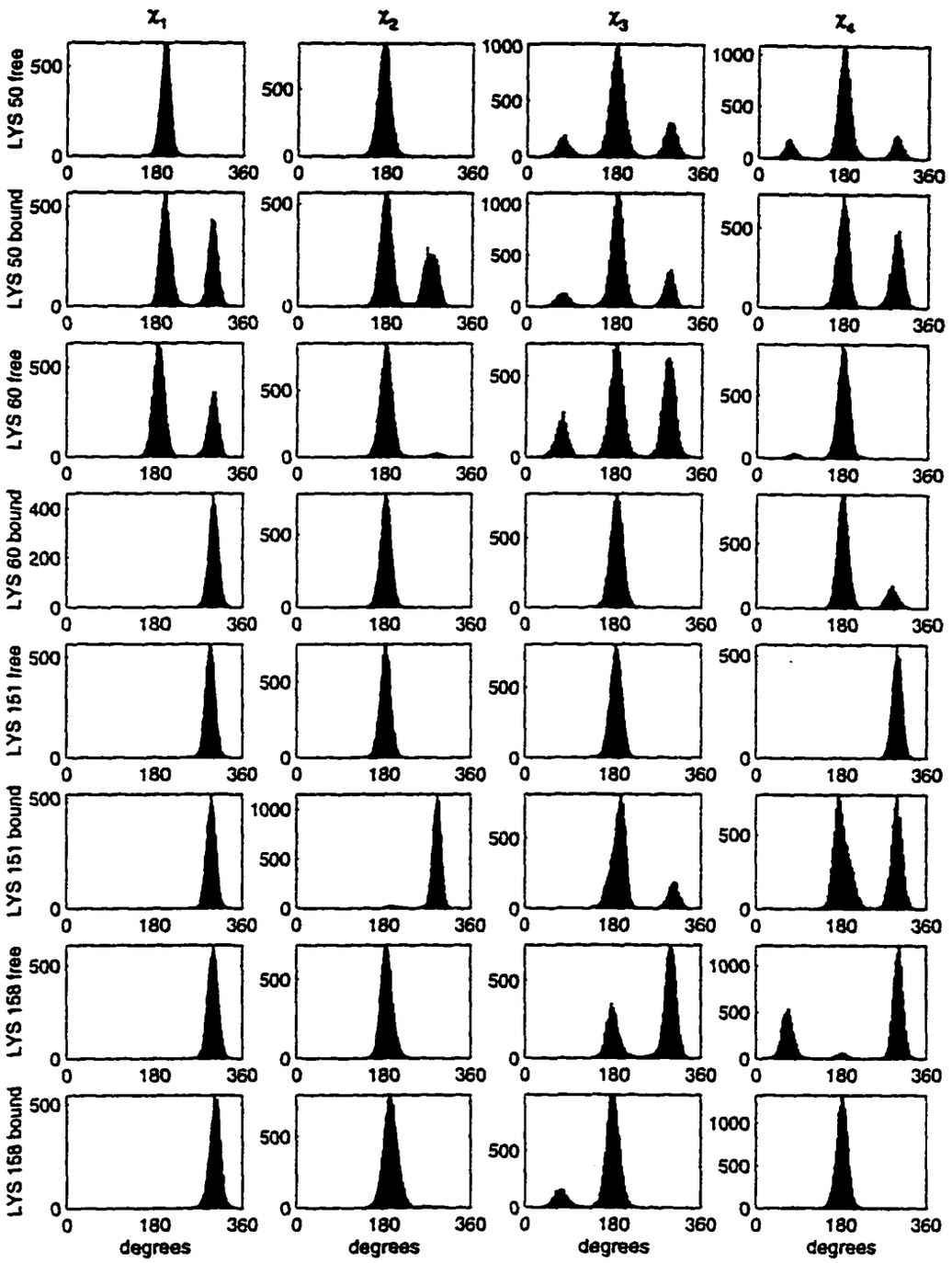

Figure 5.9 Comparison of rotamer populations for K50, K60, K151, and K158 in **ath** and **athdna**. The residues are indicated to the left of each row.

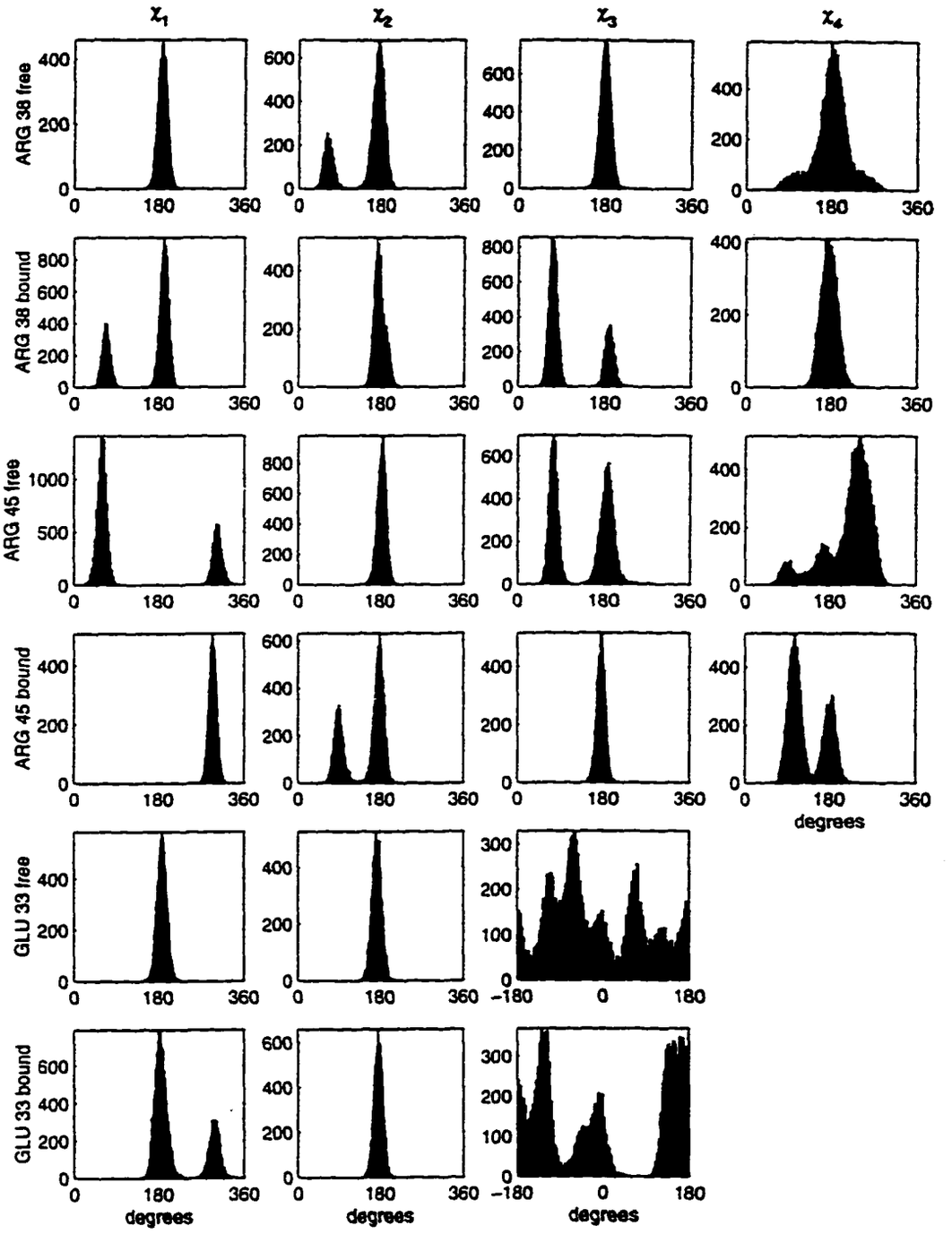

Figure 5.10 Comparison of rotamer populations for R38, R45, and E33 in **ath** and **athdna**. The residues are indicated to the left of each row.

It is noteworthy that the analysis of the rotamers of the side chain of residues in contact with DNA has provided some insights into the mechanism of binding specificity that are not possible to obtain from analyzing static structures. The H bonding to the central basepair step turned out to distinguish between the sense and antisense strands due to an apparent preference for O-H-N H bonds over the H bonds connecting two oxygens or two nitrogens, and this could be a source of directionality in the complex and of sequence specificity as well; the relative stability of these H bonds in a symmetric binding site remains to be tested. Most notable is the ability of L54 and L145 to rotate, even in the complex with DNA. Also remarkable is the induced flexibility in E33 by the neighboring arginines (R38 and R45), and the transient nature of the H bonds to the phosphates. These transient interactions caused by side chain flexibility would seem be important for binding, because they provide a favorable enthalpy (because the interaction exists on average), while also contributing favorably to the entropy of binding, and probably also to the increase in heat capacity.

6 Counterion Distribution and Release in the TBP/DNA Complex

DNA is a polyelectrolyte, and as such, it condenses cations near its surface (Manning 1978) and excludes anions from its vicinity; the amount of condensation depends on the axial charge density of the polyelectrolyte, and is relatively insensitive to the salt concentration of the solution, up to 1M. When a protein binds to DNA, it displaces the condensed cations, neutralizing the phosphate charges with its own positive side chains. Record and coworkers (Record 1978; Record 1976) have suggested that the release of cations from the surface of DNA is a stabilizing force for the formation of protein-DNA complexes, because of an increase in the entropy of these cations produced by the increase in the accessible volume in which they can roam when they are released into the bulk of the solution. The entropy difference decreases as the concentration of cations increases in the bulk of the solution, and this explains why the association constant for the protein with the DNA decreases as the salt concentration of the solution increases. This is a purely entropic effect, as it is assumed that the ions are condensed keeping their hydration shells, and hence the enthalpy remains the same.

The association constant of TBP to DNA shows a weak salt dependence, corresponding to the release of ~3.5 counterions (Petri 1995) according to the equation

$$d(\ln K) / d(\ln [salt]) = z\psi$$

where ψ is a Debye screening factor (0.88 for double stranded B-DNA (Record 1978)) and z is the actual number of counterions released from the DNA surface. At physiological salt concentrations (0.145M) and room temperature (300K), this translates to a free energy of ~3 kcal/mol, out of a total of ~13 kcal/mol for binding, which is clearly a non-negligible contribution.

The analysis of the ligands for the phosphates in the binding site discussed above has indicated that four phosphates are engaged in H bonds almost 100% of the time, approximately consistent with a measured release of 3.5 counterions (Petri 1995). As the simulations were carried out in the presence of sodium ions, it is possible to obtain a direct account of the nature of the counterion distribution and of the condensation layer upon which this interpretation of the salt dependence of binding rests, and to evaluate the diffference in number and binding energy of the counterions associated with the DNA when it is free or bound to TBP.

6.1 Sequence dependent DNA - Na^+ radial distribution functions

The radial distribution functions (g(r)) for the free and bound DNA were calculated with the proximity analysis (Mehrotra 1980) described in Chapter 2, defining the whole dodecamer as one molecule. The resulting g(r) is thus an average over all atom types at the surface of the molecule. The simulations were analyzed over the whole production phase, and the trajectory for **mlp**, being the longest one, was also broken down into three non-overlapping time

intervals for analysis (as described in Chapter 3).

The actual plots of the g(r) as a function of atomic centers distance, are shown in Figure 6.1, labeled on the top with the sequence of the dodecamer (see Figure 3.13 for the plot for **athdna**). In all cases there is a prominent peak centered around 4Å, which corresponds to the existence of a layer of water between the surface of the DNA and the sodium ions. This agrees with Manning's theory (Manning 1978), which describes the ions as associating to the polyelectrolyte without losing their hydration shells, and defines the width of the Manning layer to ~5Å. Only **mlp** and **athdna** show any direct associations between DNA and sodium (see Chapter 3). The height of the 4Å peak is different for each dodecamer, and this is related to the number of associated sodiums in the "condensation layer". This sequence dependence of the distribution could be relevant for sequence dependent binding, because it would affect the number of released counterions that provide part of the driving force for complex formation. The simulation runs for 1 or 2 ns show secondary and even tertiary peaks, approximately 3Å apart, hints of which can be seen in the shorter runs (**mlp-I**, **2c**, **6t**, **7g**, and **r28**); the correspondence between the incipient longer distance peaks observable in the shorter simulations, to those in the ns-long simulations is particularly clear for **mlp** and **mlp-I**. The coordination numbers and the binding energies for the ions in all the simulations are summarized in Table 6.1. The first shell radius is ~2.6 Å, and the second shell radius is ~5.2Å on average. Since this analysis was done in the absence of water (see Section 2.5), the interaction energies were calculated

assuming a dielectric constant of 80. The binding energy is the potential energy of interaction of the ions to the whole dodecamer (with a distance cutoff of 12Å), averaged over time and space. The pair energy (<pe>) is this average binding energy divided by the number of coordinated ions in the first shell (C) or over all distances (all C).

Table 6.1 Sodium coordination by the simulated DNA dodecamers

run	C	2C	<pe>	all C	<all pe>
mlp1	1.00	10.9	-5.34	22.0	-4.87
mlp2	1.01	7.3	-5.50	22.0	-4.58
mlp3	1.03	7.8	-6.14	22.0	-4.74
mlp	0.98	8.6	-5.62	22.0	-4.84
mlp-l	0.11	10.9	-6.40	22.0	-4.91
2c	0.03	12.1	-6.41	22.0	-4.96
6t	0.02	13.3	-5.98	22.0	-5.11
7g	0.03	12.7	-5.60	22.0	-5.08
r28	0.02	13.0	-6.42	22.0	-5.23
i	0.02	10.1	-6.41	22.0	-4.94
at	0.02	12.1	-5.63	22.0	-5.09
gc	0.03	10.3	-4.83	22.0	-4.96
athdna	1.79	6.3	-4.09	10.8	-3.16

mlp1: mlp(130-550 ps); mlp2: mlp(550-1630 ps); mlp3: mlp(1630-2080 ps)
C: coordination in the first shell; 2C: coordination in the first and second shells; all C: coordination up to the border of the simulation cell

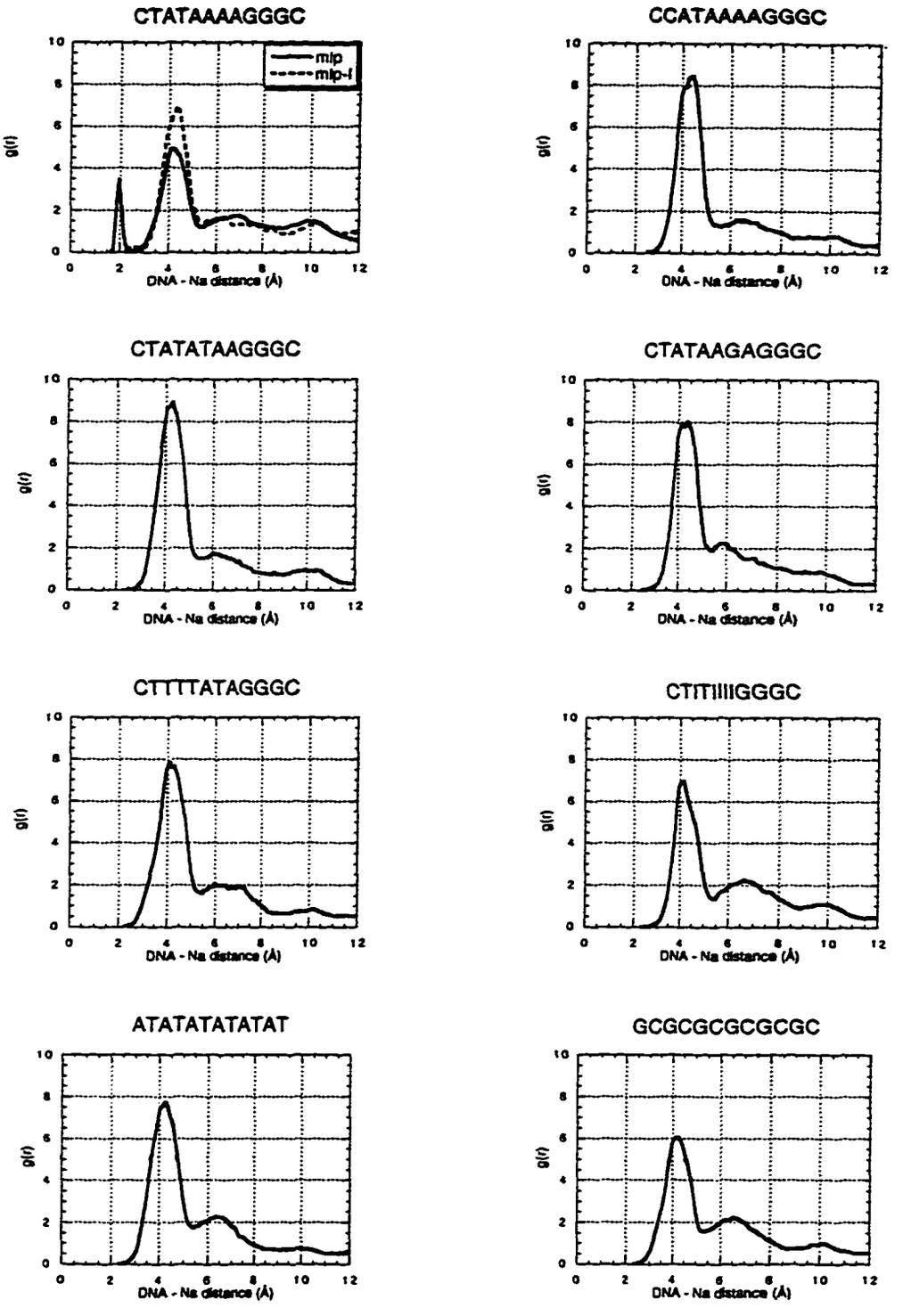

Figure 6.1 DNA-sodium radial distribution functions for the free DNA simulations. The sequence of each dodecamer is indicated at the top of the plots.

As discussed in the Validation chapter, **mlp** and **athdna** have ion-DNA contacts at guanine O6 positions, in the major groove. This is reflected in the coordination numbers for the first shell (column marked C). All the other simulations have negligible populations of sodiums in the first shell. The column marked 2C is the number of condensed counterions, that is, the number of sodium ions inside the Manning layer. The numerical values confirm the visual impression from the plots in Figures 6.1 and 3.13 that the different dodecamers condense different amounts of sodium. The differences observed from these simulations could be related to the different ^{23}Na quadrupolar relaxation rates measured for poly(dG-dC)$_2$ and poly(dA-dT)$_2$ (Nordenskiold 1984), from which the authors concluded that, since the axial charge density is the same for both molecules, DNA internal motion could contribute to the measured differences in ^{23}Na relaxation rates. The analysis of DNA conformation and dynamics presented in Chapter 4 indeed shows differences in the dynamic properties of alternating AT and alternating GC sequences, the latter being stiffer than the former (see Table 4.1).

As expected, the DNA in complex with TBP has less ions in the Manning layer, and sees less sodiums as a whole, due to the presence of the protein (10.8 ions compared to the 22 ions present in the simulation cell). The condensation fraction, calculated from the ratio of the number of condensed counterions to the total number of counterions in the simulation cell, ranges from 0.33 for the second simulation interval of **mlp**, to 0.60 for **6t**. The

corresponding value for a polyelectrolyte with the charge density of double stranded DNA is 0.76 (Manning 1978), and it is expected that oligoelectrolytes will condense less counterions due to the decrease in axial charge density produced by end effects (Olmsted 1989; Record 1978; Zhang 1996).

Besides there being a difference in the number of condensed ions, the DNA in the complex also binds the ions with less energy, about 1.5 kcal/mol less tightly than the average of the dodecamers (from the <all pe> column in Table 6.1). This is an enthalpic disadvantage, not considered by Record and collaborators in their analysis of the salt dependence of DNA ligand binding (Record 1978), and in agreement with the analysis of this kind of molecular interactions by Kim Sharp and his group (Misra 1994; Sharp 1995; Sharp 1995). The simulations were done in the absence of coions, so these numbers might still be affected by the addition of small anions to the simulated solution, but the comparison to the Manning - Record model holds, in that this model does not take the coions into account either.

6.2 Local counterion condensation and release

The *proximity* analysis (see section 2.5) assigns ion populations to each nucleotide of the dodecamers, making it possible to study local ion coordination numbers as a function of position in the dodecamers. Figure 6.2 displays the total coordination numbers for each DNA residue, in each strand. There appears to be a uniform sodium sheath over the dodecamers, as described for

mlp in the Validation section. Also, the distribution does not differ according to the length of the simulations, suggesting that the uniform distribution is not an artifact of poor equilibration. Rather, it seems to be due to the short length of the oligomers.

A comparison of the local counterion coordination between **mlp**, **mlp-l**, and **athdna** is shown in Figure 6.3. The residues with zero coordinated ions in **athdna** are contacted by R136 and S58 in the sense strand, and R45, K50 and S149 in the antisense strand, a very pleasant correspondence to the results of the analysis of TBP side chains contacting phosphates. The only position with zero coordinated ions that cannot be explained by contact to protein atoms is residue 2 of the antisense strand, which is outside the binding interface. A plausible reason is that this nucleotide is basepaired to one of the guanines coordinating a sodium ion through a long lived interaction with the O6 (see section 3.3).

The correspondence between the local ion concentration (see Figure 6.3) and the structural analysis of the TBP/DNA complex (Chapter 5) provides a solid mechanistic interpretation to the salt dependence of binding determined experimentally.

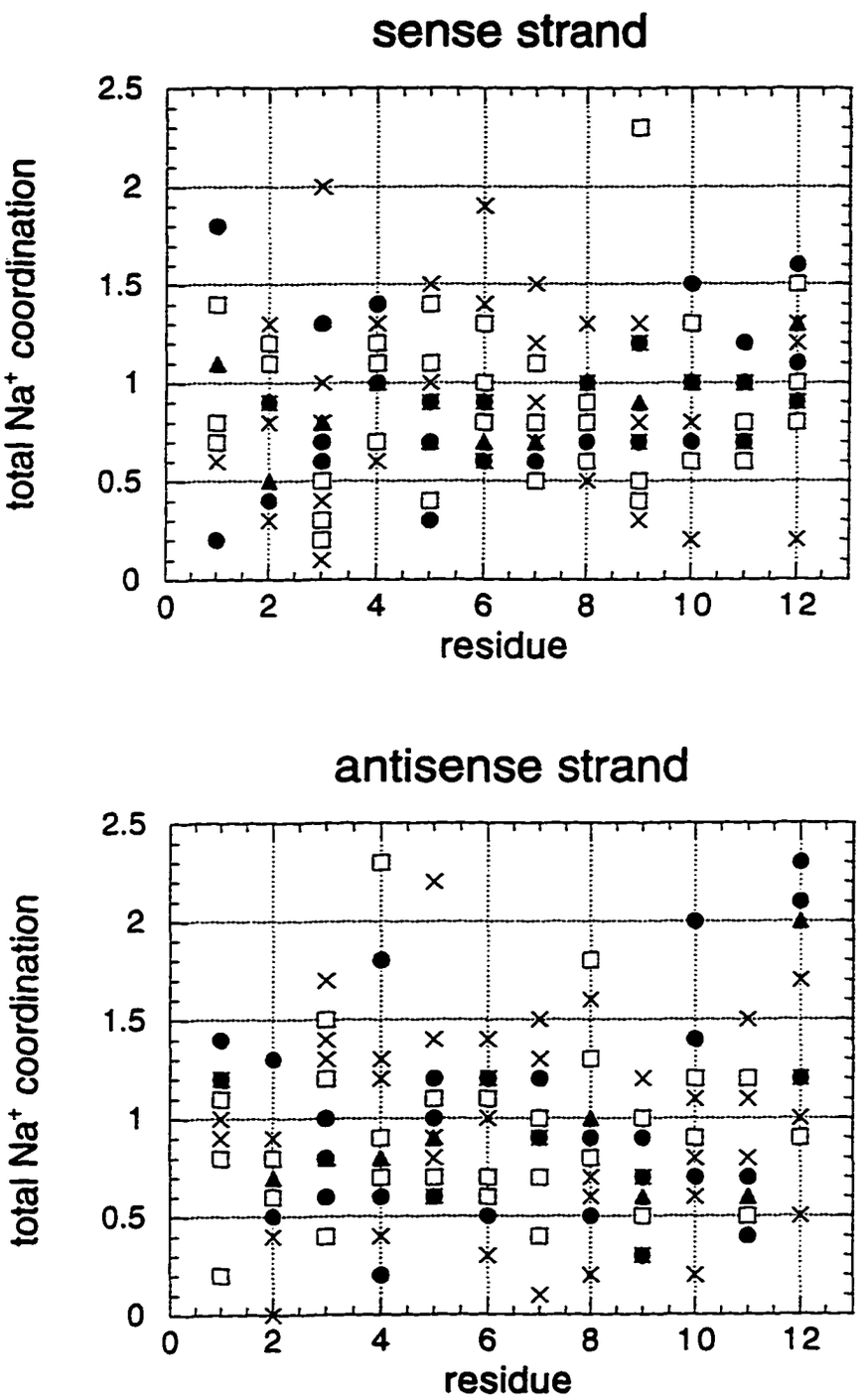

Figure 6.2 Sodium coordination number for each base of the free DNA simulations. Both strands are read 5' to 3'. Full circles: mlp1, mlp2 and mlp3; full triangles: **mlp**; x: **mlp-I**, **2c**, **6t**, **7g**, and **r28** (0.5 ns runs); squares: **i**, **at**, and **gc** (1 ns runs).

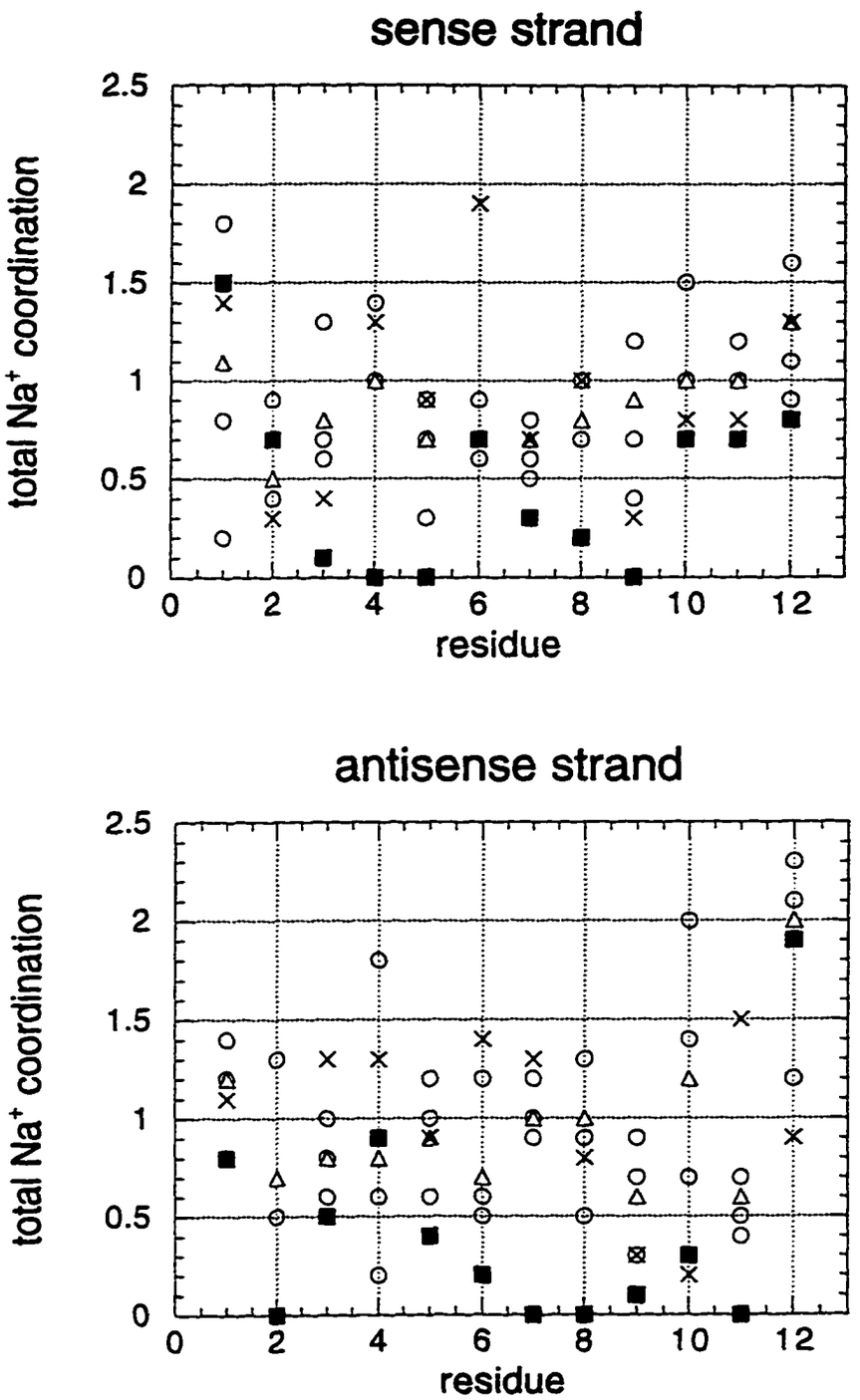

Figure 6.3 Comparison of sodium coordination number for free DNA and DNA in complex with TBP. Both strands are read 5' to 3'. o: mlp1, mlp2 and mlp3; triangles: **mlp**; x: **mlp-l**; filled squares: **athdna**.

Under the assumption that the interactions between TBP side chains and the phosphates will be independent of DNA sequence (as suggested by the identical salt dependence measured for the promoters E4 (TATATATA) and mlp (TATAAAAG) (Brenowitz, personal communication)), and that the more ions condensed, the more favorable the contribution to binding, the coordination numbers for the Manning layer listed in Table 6.1 can be used to rank the different dodecamers according to the polyelectrolyte contribution to the free energy of binding: **6t** > **r28** >**7g** > **at** = **2c** > **mlp-l** > **gc** > **i** > **mlp** (ranging from -7 to -2.3 kcal/mol at 300K and 0.15M [salt]). The ranking does not match the sequence preferences of TBP. Notwithstanding the failure to match the ranking of the experimentally determined binding affinities to the ranking of counterion condensation, it is clearly suggested by the results of the simulation that **mlp** and **mlp-l** release 2.3 and 4.6 counterions respectively, bracketing the experimental measure of 3.5 ions released (Petri 1995).

The first three ranked simulations are 0.5 ns long, and the last one listed is the longest simulation (2 ns); while there is no clear correlation between the length of the simulation run and the amount of condensed counterions (i.e., **at** and **2c** have the same amount of ions in the second shell, and the values for mlp2 and mlp3 indicate that this quantity is starting to oscillate with time), it would probably be better to compare simulations of the same duration, to eliminate the possibility of uneven equilibration of the counterion environment around the DNA. The amount of simulation time required to equilibrate the counterion distribution around DNA has not been determined, and the only

clear trend being that divalent ions take longer to equilibrate than monovalent ions (MacKerell 1997; Young 1997). For reviews on computational modeling of the ionic environment of DNA, the reader is refered to (Jayaram 1996). As discussed in the Validation section (3.3), different measures of the counterion distribution will indicate that either convergence has been achieved (see for example Figure 6.2, where the number of ions associated with each base is independent of the length of the simulation), or not (see the peak at 2Å for the DNA-sodium radial distribution functions for **mlp** and **athdna** in Figures 6.2 and 3.13, respectively, representing counterions that do not exchange with the solvent). The ^{23}Na relaxation rate studies carried out by Nordenskiöld et al. (Nordenskiold 1984) suggest the existence of at least three time scales for motion of ions around the DNA, going from picosecond rearrangement of the solvation shell, to axial diffusion in hundreds of picoseconds, and the longest one being on the order of hundreds of nanoseconds, assumed to be diffusion away from the DNA. The longest time range is not accessible to molecular dynamics simulations that include the solvent explicitly, because of the time of simulation and the size of the simulation cell required to simulate adequately radial diffusion (~100 Å (Reddy 1987)). The validity of the analysis presented in this chapter rests on the assumption that local, short ranged (and hence short lived) interactions between the sodium ions and the DNA are the main contributors to DNA sequence dependent condensation.

7 Water Release from TBP and DNA

The major driving force for macromolecular complex formation is assumed to be the hydrophobic effect (Spolar 1989; Spolar 1994), because the dehydration of the surfaces that become the interaction interface releases water molecules into the bulk solution, providing a large entropic gain. The amount of water released from the interface can be determined experimentally, for example, from the variation of the binding constant with osmolyte concentration (Parsegian 1995), but these experiments have not been carried out for TBP binding to DNA. An alternative to estimate the contribution from water release is to calculate the accessible surface area of reactants and products, and relate this area to the experimentally determined correlation between excluded area and free energy of solvation (Livingstone 1991; Spolar 1992; Spolar 1994). Assuming the contribution is entirely entropic:

$$\Delta S^0_{HE}(T) = 0.32\ \Delta A_{np}\ \ln(T / 386) = 1.35\ \Delta Cp^0\ \ln(T / 386)$$

where ΔA_{np} is the amount of hydrophobic surface occluded from the solvent upon complex formation, T is temperature, 386 is the temperature (in Kelvin) at which this entropic contribution becomes zero, and ΔCp is the change in heat capacity upon complexation.

The van't Hoff analysis of the binding of TBP to the E4 promoter (Petri 1995) revealed a decrease in heat capacity upon binding of unprecedented magnitude given the size of TBP: -3.5 kcal/mol•K. According to the analysis

presented by Spolar and Record (Spolar 1994), this ΔCp corresponds to a buried area of 12600 $Å^2$:

$$\Delta Cp = -0.32\ \Delta A_{np} + 0.14\ \Delta A_{p}$$

The solvent accessible surface area of TBP is ~11000 $Å^2$, so the estimated value for the buried area upon complex formation does not make structural sense for the particular case of TBP. Furthermore, the measured ΔCp is DNA sequence dependent, even for sequences that bind isosterically to TBP (Brenowitz, personal communication), and which would probably bury a similar amount of surface upon binding.

Other sources for the change in heat capacity have been proposed by Sturtevant (Sturtevant 1977). In particular, the rupture of H bonds (Madan 1996) and the stiffening of soft internal modes of the macromolecules or the water molecules (Guinto 1996) are proposed to contribute to a decrease in heat capacity. An increase in the number of isoenergetic conformations is proposed to be neutral for ΔCp.

The study of the temperature behavior of the internal modes of TBP and DNA requires either a normal mode or quasiharmonic analysis that is outside the scope of this work. On the other hand, an analysis of the hydration of TBP, DNA and sodium ions, and of the properties of the water molecules in the first hydration shells of these molecules is feasible, making it possible to discern the changes, if any, in H bonding or in water structure upon complex formation. This analysis also yields an approximate enthalpic cost for dehydrating the different

DNA sequences, which might also be a selectivity determinant used by TBP.

7.1 Solute - water radial distribution functions

The water distributions around the macromolecules were calculated with the proximity analysis (Mehrotra 1980) described in section 2.5. Individual radial distribution functions were obtained for the DNA dodecamers, TBP, and the sodiums. Similar to the results presented in Chapter 6, the radial distribution functions in Figure 7.1 represent an average over the different atom types present at the surface of the solutes thus defined. The water dipole distribution measures the orientation of the water molecules as a function of distance from the solute, where 180° corresponds to a linear arrangement between the line connecting the oxygen of the water to the solute atom, and the water dipole.

The three molecular species present in the simulations interact with water distinctly, as shown by the large differences in the g(r) and water dipole distributions in Figure 7.1. The differences observed in the water dipole orientation suggest the inequivalence of the local dielectric environment around these solutes. The dielectric properties of space around these molecules could probably be estimated from a molecular perspective from the orientation of the water dipoles (Allen 1987), as an alternative to a continuum electrostatics approach based on the Poisson-Boltzmann equation (Pack 1993). While there are differences between TBP, DNA and sodium, the absence of sensitivity to the presence of TBP in the DNA distributions and *viceversa* is very bewildering.

The water molecules organize differently around TBP alone (broken line in Figure 7.1) than around TBP in the complex, as seen from the corresponding g(r). The region of TBP responsible for this deviation remains to be determined, but it is tempting to assume that it is the hydrophobic surface (Pertsemlidis 1996) in contact with DNA that is responsible for the decrease in water density at longer distances. Incidentally, this long range effect could help understand why the attempts to explain all the change in heat capacity for this system from the properties of the first shell of hydration have failed so far.

A surprising finding is the insensitivity of water to the bound or free state of DNA. The plot in Figure 7.1 shows the superposition of the DNA-water g(r) for all the simulations, **athdna** represented with the bold line. The interesting feature of this distribution is a small peak at 2Å; this distance corresponds to an O-H H bond, and it is also present in the distributions for TBP.

The first hydration shell for sodium remains basically untouched by the presence of DNA and/or TBP, as shown by the superposition of all the g(r). This finding agrees with the condensation theory (Manning 1978) and experimental measures discussed in (Collins 1997). The divergence between the **sod3** simulation and the rest of the simulations occurs at the second shell, marked with an arrow, and beyond.

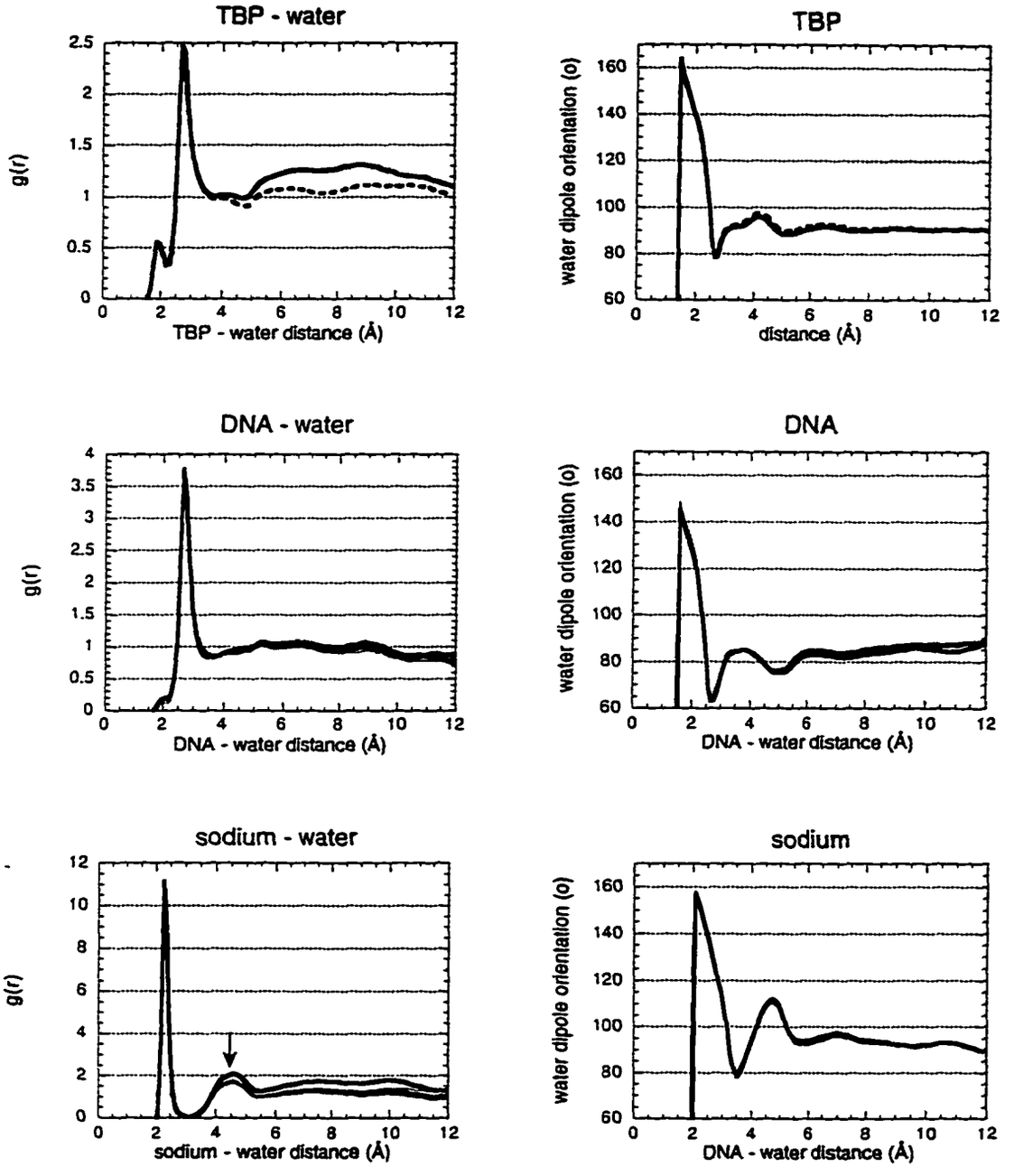

Figure 7.1 Radial distribution functions and water dipole orientation functions for TBP, DNA and sodium. Top: broken line: free TBP, continuous line: TBP bound to DNA. Bottom: the arrow marks the second hydration shell.

Table 7.1 gathers the coordination numbers for the first shell (C) and second shell (2C) of the solutes identified in the first column of the table. The pair energy (<pe>) is the average binding energy of a water molecule to the macromolecule, and $<pe_{ww}>$ is the pair energy of the waters in the first hydration shell. The hydration for each basepair can be calculated by taking the values in column C and dividing them by the number of basepairs in the DNA oligomers. The result is ~22 water molecules / basepair, in good agreement with experimental estimates (Chalikian 1994; Schneider 1995).

Assuming a binding reaction

$$DNA + 22\ Na^+ + TBP + 3\ water \rightarrow TBP/DNA + 22\ Na^+ + 3\ water,$$

(where the 3 water molecules are the crystallographic molecules) the amount of released water molecules upon complex formation can be estimated directly from the entries labeled **athdna**, **ath** and **mlp**. These are 60 water molecules; ~25 come from the protein (21 can be ascribed to the residues within 5Å of any DNA atom, most of which were described in Chapter 5), 3 come from the sodium ions, and the remaining 31 from the DNA. 60 water molecules correspond to an accessible surface area of ~600 $Å^2$ (Colombo 1992). This is approximately one fifth of the estimated buried area from the crystal structures of TBP and the TBP/DNA complex, and such a large discrepancy is currently not understood. It could be due to the definition of the hydration radii, which are derived from the water-solute radial distribution functions for a variety of small solutes (Mezei, personal communication). Assuming that the radii are

consistently underestimated, the second hydration shell numbers produce a new estimate of 442 released water molecules; these correspond to ~4420 $Å^2$ of surface buried , much closer to the 3100 $Å^2$ obtained from the structures. The radii for the protein and DNA residues are most likely underestimated and need to be reexamined.

Table 7.1 Water coordination by TBP and DNA

simulation	C	2C	<pe>	$<pe_{ww}>$
athdna	484.6	1996.7	-14.39	-3.25
ath	278.7	1330.1	-12.04	-3.66
mlp	265.0	1108.9	-16.19	-2.96
mlp-l	262.5	1057.3	-16.26	-2.98
2c	263.5	1068.4	-16.30	-2.96
6t	260.4	1032.2	-16.33	-2.96
7g	262.3	1035.4	-16.38	-2.96
r28	261.7	1004.5	-16.48	-2.93
i	269.6	1087.1	-16.18	-2.93
at	256.9	1037.2	-16.17	-2.88
gc	269.6	1091.7	-15.91	-2.90

C= coordination number in the first shell; 2C= coordination number in the second shell; <pe>= pairwise water-solute binding energy; $<pe_{ww}>$= pairwise water-water binding energy. athdna includes the contributions from TBP, DNA, sodium ions and crystallographic waters; ath includes also the contribution from crystallographic waters. All the free DNA simulations include the contribution from sodium ions.

There are small variations in the pair energies of the water molecules to the different DNA dodecamers, and these are probably not significant, though this remains to be tested. TBP does have a noticeably different interaction with

water, and so does the complex. There is an inverse correlation between the strength with which the macromolecule binds water and the strenght with which the water binds to itself (pe_{ww}), which holds if the data for sodium (Table 7.2) are included. The elucidation of the molecular reason for this correlation will require breaking down the macromolecules into residues and/or functional groups.

From Figure 7.1 it appeared that the sodium ions keep their first hydration shell regardless of the presence of DNA and/or TBP in the simulation cell. This is also shown in the coordination numbers summarized in Table 7.2 for all the sodium ions in the simulations.

Table 7.2 Water coordination by sodium ions in the simulations

simulation	C	2C	<pe>	<pe_{ww}>
athdna	5.6	21.3	-18.47	-1.50
mlp	5.7	22.9	-18.49	-1.64
mlp-l	5.8	21.4	-19.07	-1.62
2c	5.8	21.6	-19.12	-1.59
6t	5.7	20.3	-19.62	-1.55
7g	5.8	21.0	-18.94	-1.60
r28	5.8	20.0	-19.34	-1.57
i	5.8	22.1	-18.94	-1.60
at	5.8	21.1	-18.97	-1.53
gc	5.8	21.9	-18.69	-1.58
sod3	5.8	28.0	-17.50	-1.68

C= coordination number in the first shell; 2C= coordination number in the second shell; <pe>= pairwise water-solute binding energy; <pe_{ww}>= pairwise water-water binding energy

The only difference is a decrease in the interaction of sodium with the

water molecules in the first shell, by 1 - 1.5 kcal/mol. This is a bit surprising, since the g(r) look practically identical.

7.2 Local hydration analysis

The differences in hydration number in Table 7.1 for the various dodecamers prompted a more local analysis of the water molecules associated with each base. As the TBP/DNA interface is anhydrous, the enthalpic cost for stripping the coordinated waters from the surface of a particular DNA sequence, and the release of different numbers of water molecules, might also constitute sequence selectivity determinants.

Figures 7.2 shows the first shell coordination numbers for the free DNA molecules for the sense and antisense strands. The plots are not symmetric because the strands are not palindromic, and purines and pyrimidines coordinate different numbers of water molecules (Schneider 1995). The patterns displayed are diverse, with hints of periodic behavior for **at** and **gc**, reflecting the sequence and structure periodicity in these dodecamers. The strongest variation is seen for the sense strand of **i**, at the center of the dodecamer, with a sharp decrease at the IG step; the antisense strand for **i**, which is a long stretch of cytosines, behaves normally.

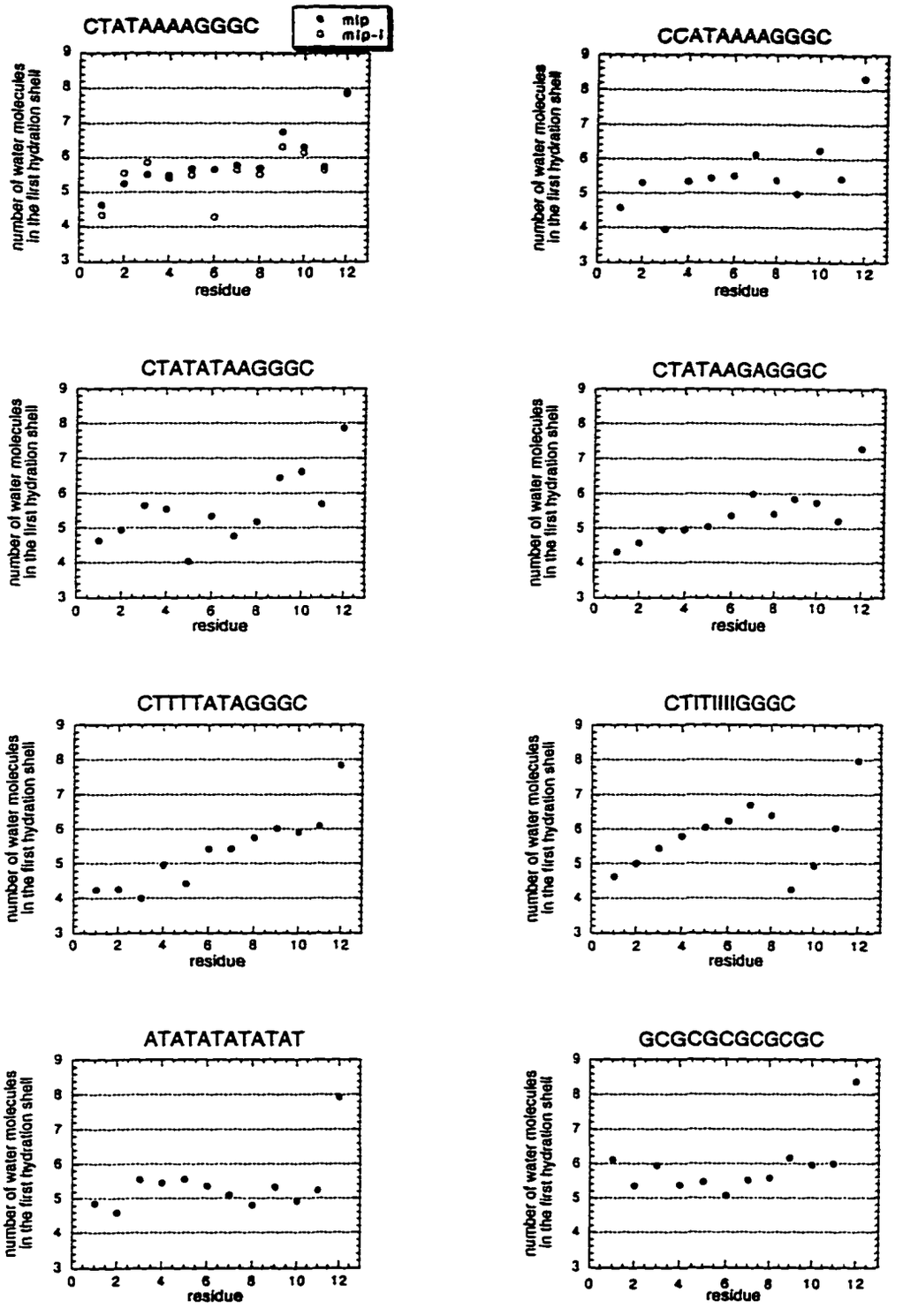

Figure 7.2 Water coordination numbers for the sense strand bases of the free DNA dodecamers, indicated on the top of each plot.

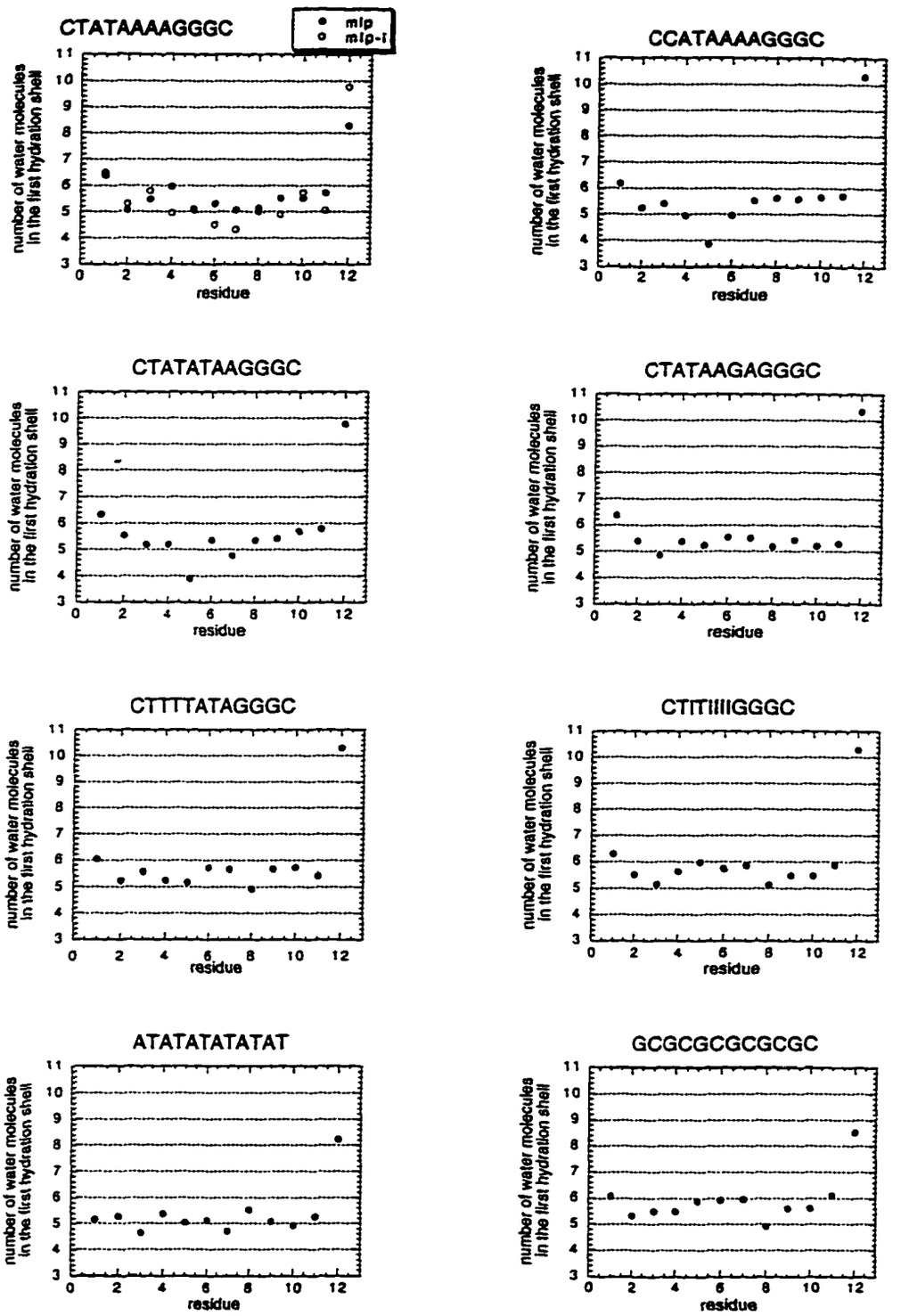

Figure 7.2 Water coordination numbers for the antisense strand bases of the free DNA dodecamers, indicated on the top of each plot. The strand reads 5' to 3' from left to right.

Similar, albeit less pronounced, dips are seen at base 5 of the antisense strands of **2c** and **6t**, that is the antisense side of an AG step, and is a place where TBP inserts phenylalanine residues into the basepair steps. Were this depletion in hydration real, it would probably facilitate insertion by reducing the enthalpic penalty from hydration loss at this site.

Figure 7.3 compares the hydration numbers for free mlp (**mlp** and **mlp-l**) and **athdna** for both strands of the DNA. As expected, there is a decrease in solvation upon complex formation, accounting for 31 water molecules. All bases in the complex are at least partially hydrated, because the analysis did not distinguish between the major groove and the minor groove. Also shown in this figure is the decrease in hydration for the TBP residues involved in interactions with DNA (defined as residues with atoms within 5Å of any DNA atom). These residues account for ~21 water molecules lost upon complexation to DNA. The numbers in this plot include the contributions from backbone atoms as well, resulting in non-zero hydration numbers for residues that would appear to be buried completely. The only residues that were not assigned any water molecules at all are: V11, V101, N9, T64, T155 and L145; the remarkable finding is that N99 and L54, the symmetry partners of N9 and L145, did get some secondary hydration assignments, despite being located at the core of the TBP/DNA interface (see Chapter 5). This could mean that the complex is not equally tight at all points on the surface of the DNA, a phenomenon which would create small openings for the buried atoms to which the analysis program would assign partial solvation.

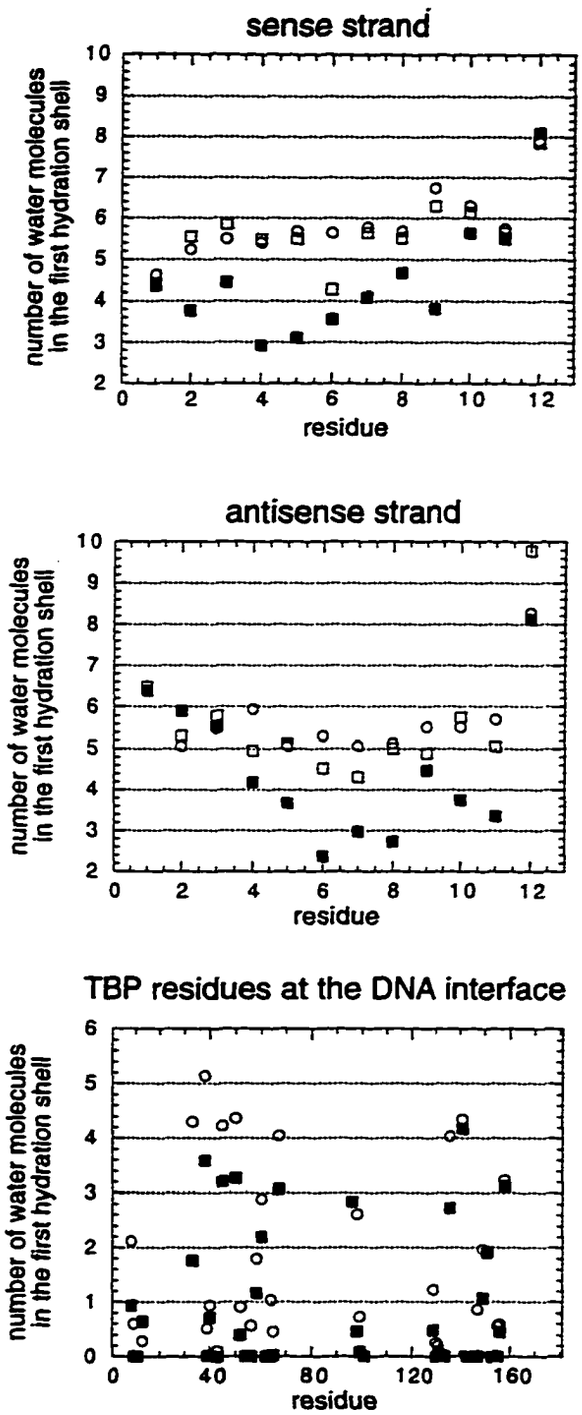

Figure 7.3 Comparison of water coordination numbers for free and bound mip. Filled squares: **athdna**; o: **mip** in the sense and antisense plots (both read 5' to 3' from left to right), **ath** in the bottom plot; open squares: **mip-l**.

Arginines, lysines and E33 are amongst the most solvated side chains, even in the complex, as expected by their having formal charges. This is also consistent with the flickery nature of the H bonds to the phosphates found in Chapter 5.

Further modulation of sequence specificity could be provided by differential hydration of the binding sites. To avoid end effects in addressing this aspect, the first shell hydration was studied only for the four central basepairs of the dodecamers. The results are summarized in Table 7.3, which shows that there can be differences of up to 9-10 water molecules between these DNA sequences (from comparing **i** with **6t**).

Table 7.3 First shell water coordination by the central four basepairs

simulation	sense C	sense E	antisense C	antisense E	total C	total E
athdna	15.43	-14.9	11.76	-16.8	27.19	-31.7
mlp	22.79	-15.6	20.56	-15.6	43.35	-31.2
mlp-I	20.94	-14.9	18.94	-15.3	39.88	-30.1
2c	22.42	-15.5	19.95	-16.0	42.37	-31.5
6t	19.32	-14.2	19.34	-14.5	38.66	-28.6
7g	21.80	-15.6	21.43	-15.8	43.23	-31.3
r28	21.01	-15.1	21.45	-14.7	42.46	-29.7
i	25.33	-17.1	22.71	-14.0	48.04	-31.1
at	20.81	-15.2	20.38	-14.7	41.19	-31.1
gc	21.63	-15.0	22.67	-14.5	44.30	-29.5

C= coordination number; E= pairwise binding energy (kcal/mol)

It is reassuring to note that when the strands are identical in sequence, the hydration numbers match much better (**at, gc, r28**) than when the strands are segregated into a purine strand and a pyrimidine strand (**mip, mip-i, 2c**), with the extreme being **i**.

A comparison of the binding energies between the bound DNA and the same sequence free in solution, shows that the water molecules are held more tightly in the complex, and that there are less water molecules bound. The effect of hydration has favorable enthalpic components (because the binding energy of the waters to the complex is more attractive than to the free DNA) and entropic components (from the release of water into bulk solution). In keeping with the fact that both enthalpic and entropic contributions from hydration seem to operate for TBP binding, a ranking of the oligomers that would reflect both of these factors was attempted. The ranking of the free DNA sequences by the pairwise DNA-water binding energy, starting from the one that binds to its hydration shell tighter produces the following:

2c > 7g > mip > i > mip-i > at > r28 > gc > 6t

If one assumes that there is an enthalpic penalty for dehydrating these basepairs, the best binding sequence would be **6t**, because it binds the hydration shell less tightly than any of the others.

A ranking related to the entropic gain from water release is produced for the number of bound water molecules:

i > gc > mip > 7g > r28 > 2c > at > mip-i > 6t

which makes **6t** the worst binding site if the entropic gain from releasing water

from the surface of DNA contributes to the free energy of binding. It should be stressed that the differences between all these sequences are subtle, and the fact that the two simulations for **mlp** appear at different places in the ranking also indicates the sensitivity of the water structure to slight changes in the underlying DNA structure (Chalikian 1994; Schneider 1995; Tippin 1997).

The two rankings performed above assume that the binding energy of the water to the DNA and the amount of released water molecules will be the same for all TBP/DNA complexes, regardless of sequence. A reason for the lack of agreement between the two rankings is that this underlying assumption might be wrong. Unfortunately, the ranking of the binding energy in the complex cannot be performed, as only one sequence (**mlp**) was simulated both free and bound to TBP.

7.3 Perturbation of the structure of water

One of the explanations proposed to account for changes in heat capacity is by the alteration of the structure of water by the dissolved solutes (Sturtevant 1977). Water structure is disrupted by ions and promoted by hydrophobic solutes (Madan 1996; Pertsemlidis 1996). The simulations on TBP and DNA have both ionic and hydrophobic surfaces, so the structure of water in a layer extending up to 4Å from any solute atom was calculated. Figures 7.4 and 7.5 contain the gOO(r) and gOH(r) for pure water, sodium, TBP and DNA.

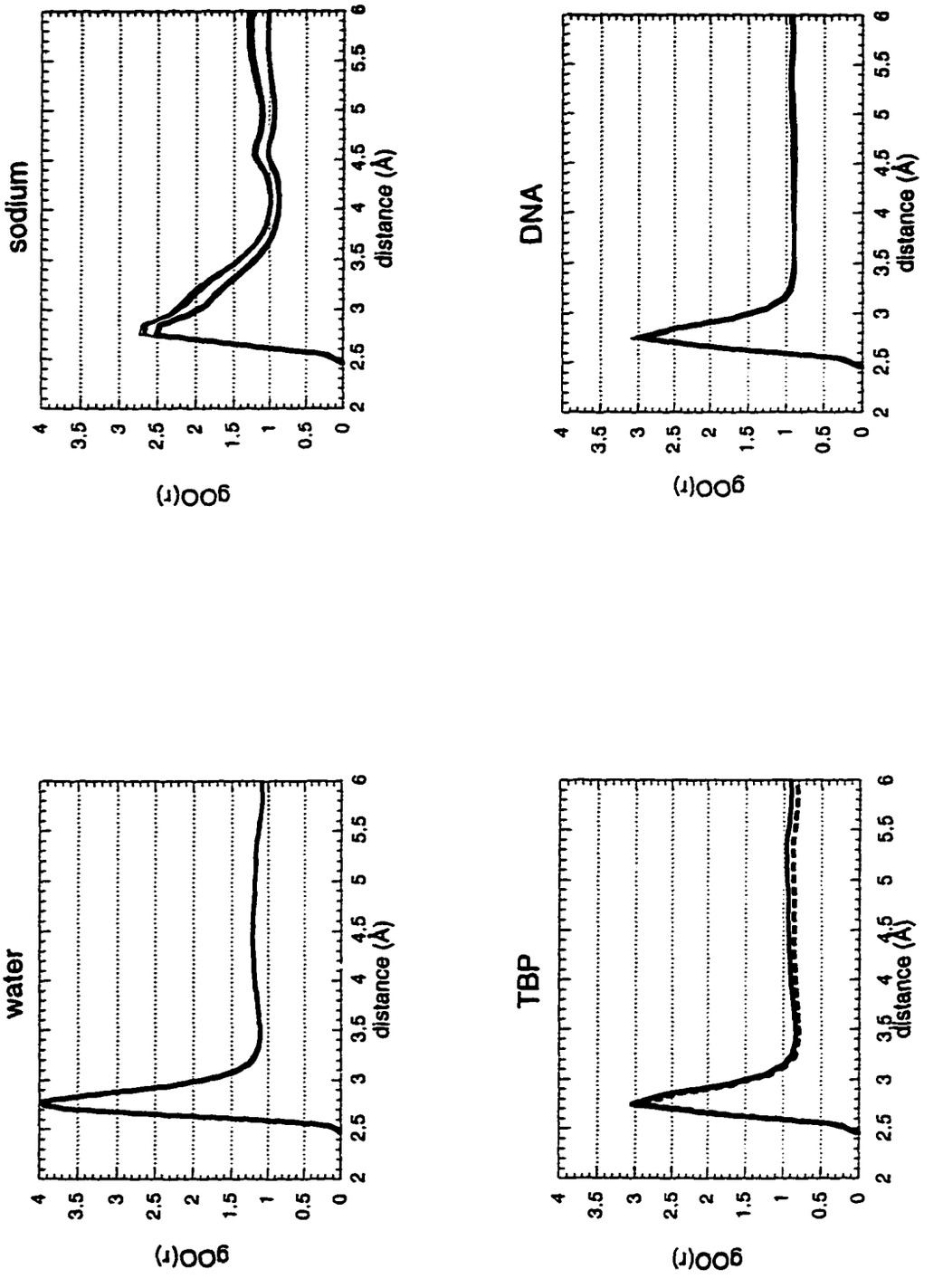

Figure 7.4 gOO(r) for pure water and the water in a 4Å layer around sodium, DNA and TBP. TBP plot: solid line: **athdna**; broken line: **ath**. Sodium plot: bold line **sod3**; thin lines: rest of the simulations

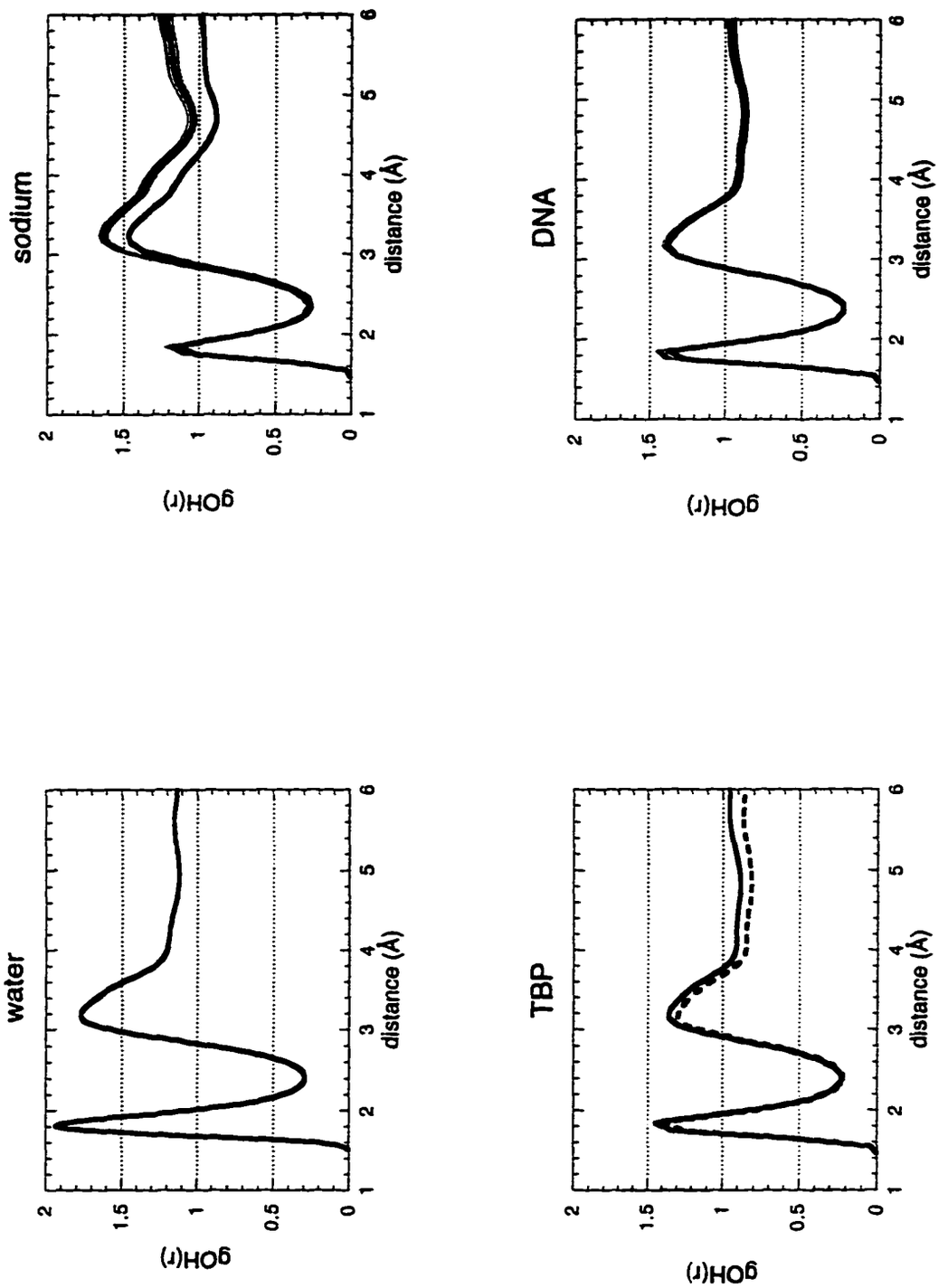

Figure 7.5 gOH(r) for pure water and the water in a 4Å layer around sodium, DNA and TBP. TBP plot: solid line: **athdna**; broken line: **ath**. Sodium plot: bold line **sod3**; thin lines: rest of the simulations

In all cases, there is a decrease in the height of the peak at 3Å for the gOO(r), when compared to the pure water simulation. This could be explained by the layer being thin enough for the presence of the solute to compete for space in the first hydration shells of these water molecules. While TBP and DNA produce basically the same distributions (probably a consequence of the heterogeneity of the surface), sodium ions produce a distinct effect on the gOO(r). This distribution is consistent with previous analysis of sodium hydration (Marchese 1984), and indicates a coordination of six water molecules, as expected from Table 7.2. The difference between the **sod3** simulation (bold line) and the rest of the simulations might be due to the absence of a complete second hydration shell in the latter, which displaces slightly the water molecules in the first hydration shell. This remains to be proven by looking at the actual spatial distribution of water centers around the sodium ions.

The gOH(r) in Figure 7.5 are the initial step towards quantitating H bond loss and formation as a consequence of TBP binding to DNA. The first peak corresponds to an O-H H bond, and again it is seen to decrease in all cases, most dramatically for sodium, where the relative heights of the first and second peak are inverted. In the case of TBP and DNA, it will be interesting to see if the decrease in the first peak for the gOH(r) is correlated to the existence of the peak around 2Å found for the TBP-water and DNA-water g(r) (Figure 7.1).

The analysis of the hydration of TBP and DNA presented in this section has shown that there are indeed differences in the structure of the water

solvating the macromolecules, sometimes extending far away from the molecular surface (see Figure 7.1 for the TBP-water g(r)). In the case of DNA, these differences can be ascribed to the sequence of the DNA strands, there being a clear distinction between hydration of purines and pyrimidines (Table 7.3 and Figure 7.2) and of inosine bases. The alteration of the water-water distributions by the presence of the solutes, noticeable both in the gOO(r) and in the first peak of the gOH(r) suggests that it might be worthwhile to carry out replicas of these simulations at higher and lower temperatures, in order to calculate the contribution of water structure breaking to the decrease in heat capacity.

8 Summary and Conclusions

The broad picture of DNA selectivity for TBP binding that emerges from the analysis presented in this work, indicates a mechanism that is determined by the confluence of a series of structural and dynamic properties. A major element in DNA selectivity determinants for TBP binding is the avoidance of steric clashes between the minor groove of the DNA and the side chains of TBP. Having satisfied this requirement, there is a delicate balance between the energetics of a series of steps including the appeareance in the DNA structure of preformed, incipient kinks, of fleeting excursions to the conformation found in the complex, and of changes in hydration and in counterion condensation. The superposition of the corresponding energy contributions is responsible for the relative ranking of the ability of TBP to recognize many different promoters with slightly different affinities (Wong 1994). It is to be expected, therefore, that the various attributes need not be displayed simultaneously by any one particular DNA sequence. Actually, it appears that some DNA binding sites are selected on the basis of their flexibility combined with a general predisposition to untwist and roll to open the minor groove, while others substitute the flexibility with a stronger tendency to adopt a conformation that resembles more the bound DNA. TBP is an active partner in complex formation, in spite of the lack of overall structural change upon complex formation. Rather, the analysis showed significant flexibility of the side chains of internal residues and of the residues at

the interface with the DNA. This flexibility (observed directly from the simulations) is redistributed as a consequence of DNA binding, and is also responsible for the interaction with the DNA through side chains that attach to and dissociate from DNA repeatedly. Thereby, the internal flexibility of TBP provides both a favorable enthalpy of interaction and a high configurational entropy.

The role of the water and counterions in the formation of the complex is pivotal. However, the analysis of the dynamics and distribution of the counterions and the water is subject to the uncertainties related to convergence issues. Nonetheless, the finding that a variety of properties computed from the simulation results are close to the experimentally measured values (e.g., the amount of condensation, the hydration / basepair, the diffusion coefficient for water, etc) suggests that the mechanistic and structural conclusions derived from the analysis of the ion and water distributions are likely to be at least qualitatively correct. In general, the strongest conclusions from this work come from the analysis of DNA and local sodium hydration properties, for which both time and ensemble averaging could be done, over the entire collection of independent simulations. In particular, the conclusions obtained for the conformational properties of DNA have remained practically unchanged over three rounds of analysis of increasing amounts of data.

While the results obtained thus far did not make it possible to evaluate the change in heat capacity upon binding, a number of findings, such as the alteration of water structure around the macromolecules and the rotation of TBP

side chains in contact with DNA, indicate that the actual calculation might be fruitful in this respect. To study the contribution from the macromolecules, probably the most straightforward way to extend the analysis presented here is to carry out a quasiharmonic analysis of the motions in the free protein and DNA molecules, and of the complex. Given that the ΔCp is sequence dependent, the simulation of TBP/DNA complexes bound to different DNA sequences also becomes interesting. It is noteworthy for further study that the hydration properties of the different basepairs varies sufficiently to make attractive a more in depth analysis, separating the contributions from the two grooves. Another example worth a more detailed classification of the hydration surfaces is the difference in hydration found for free and bound TBP. This is manifested in the divergent TBP-water distributions at long distances, and is currently lacking an explanation.

References

Abu Daya, A., Brown, P. M., and Fox, K. R. (1995). "DNA sequence preferences of several AT-selective minor groove binding ligands". Nucleic Acids Res. **23:** 3385-92.

Allen, M. P., and Tildesley, D. J. (1987). "Computer simulation of liquids", (New York: Oxford University Press).

Arndt, K. M., Ricupero, S. L., Eisenmann, D. M., and Winston, F. (1992). "Biochemical and genetic characterization of a yeast TFIID mutant that alters transcription in vivo and DNA binding in vitro". Mol. Cell Biol. **12:** 2372-82.

Arndt, K. M., Wobbe, C. R., Ricupero, H. S., Struhl, K., and Winston, F. (1994). "Equivalent mutations in the two repeats of yeast TATA-binding protein confer distinct TATA recognition specificities". Mol. Cell Biol. **14:** 3719-28.

Arnott, S., and Hukins, D. W. (1972). "Optimised parameters for A-DNA and B-DNA". Biochem. Biophys. Res. Commun. **47:** 1504-9.

Atkins, P. W. (1990). "Physical Chemistry", 4th edition Edition (New York: W. H. Freeman & Co).

Auffinger, P., Louise-May, S., and Westhof, E. (1995). "Multiple molecular dynamics simulations of the anticodon loop of tRNA-Asp in aqueous solution with counterions". J. Am. Chem. Soc. **117:** 6720-6.

Auffinger, P., and Westhof, E. (1996). "H-bond stability in the tRNA Asp anticodon hairpin: 3 ns of multiple molecular dynamics simulations". Biophys. J. **71:** 940-54.

Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S.-H., Srinivasan, A. R., and Schneider, B. (1992). "The nucleic acid database: a comprehensive relational database of three-dimensional structures of nucleic acids". Biophys. J. **63:** 751-759.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. E., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). "The Protein Data Bank: a computer-based archival file for macromolecular structures". J. Mol. Biol. **112:** 535-42.

Bernues, J., Carrera, P., and Azorin, F. (1996). "TBP binds the transcriptionally inactive TA5 sequence but the resulting complex is not efficiently recognised by

TFIIB and TFIIA". Nucleic Acids Res. **24:** 2950-8.

Billeter, M., Guntert, P., Luginbuhl, P., and Wuthrich, K. (1996). "Hydration and DNA recognition by homeodomains". Cell **85:** 1057-65.

Black, C. B., and Cowan, J. A. (1994). "Quantitative evaluation of electrostatic and hydrogen-bonding contributions to metal cofactor binding to nucleic acids". J. Am. Chem. Soc. **116:** 1174-8.

Branden, C., and Tooze, J. (1991). "Introduction to protein structure", (New York: Garland Publishing, Inc.).

Breathnach, R., and Chambon, P. (1981). "Organization and expression of eucaryotic split genes coding for proteins". Annu. Rev. Biochem. **50:** 349-83.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. (1983). "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations". J. Comp. Chem. **4:** 187-217.

Bryant, G. O., Martel, L. S., Burley, S. K., and Berk, A. J. (1996). "Radical mutations reveal TATA-box binding protein surfaces required for activated transcription in vivo". Genes Dev. **10:** 2491-504.

Burley, S. K., and Roeder, R. G. (1996). "Biochemistry and structural biology of transcription factor IID (TFIID)". Annu. Rev. Biochem. **65:** 769-99.

Calladine, C. R. (1982). "Mechanics of sequence-dependent stacking of bases in B-DNA". J. Mol. Biol. **161:** 343-352.

Calladine, C. R., and Drew, H. R. (1986). "Principles of sequence-dependent flexure of DNA". J. Mol. Biol. **192:** 907-18.

Calladine, C. R., and Drew, H. R. (1996). "A useful role for "static" models in elucidating the behaviour of DNA in solution". J. Mol. Biol. **257:** 479-85.

Calladine, C. R., Drew, H. R., and McCall, M. J. (1988). "The intrinsic curvature of DNA in solution". J. Mol. Biol. **201:** 127-37.

Chalikian, T. V., Sarvazyan, A. P., Plum, G. E., and Breslauer, K. J. (1994). "Influence of base composition, base sequence, and duplex structure on DNA hydration: apparent molar volumes and apparent molar adiabatic compressibilities of synthetic and natural DNA duplexes at 25 degrees C". Biochemistry **33:** 2394-401.

Chasman, D. I., Flaherty, K. M., Sharp, P. A., and Kornberg, R. D. (1993). "Crystal structure of yeast TATA-binding protein and model for interaction with DNA". Proc. Natl. Acad. Sci. USA **90:** 8174-8.

Chen, H. H., Rau, D. C., and Charney, E. (1985). "The flexibility of alternating dA-dT sequences". J. Biomol. Struct. Dyn. **2:** 709-19.

Chen, W., and Struhl, K. (1988). "Saturation mutagenesis of a yeast his3 "TATA element": Genetic evidence for a specific TATA-binding protein". Proc. Natl. Acad. Sci. USA **85:** 2691-2695.

Chiang, S. Y., Welch, J., Rauscher, F. r., and Beerman, T. A. (1994). "Effects of minor groove binding drugs on the interaction of TATA box binding protein and TFIIA with DNA". Biochemistry **33:** 7033-40.

Coleman, R. A., and Pugh, B. F. (1995). "Evidence for functional binding and stable sliding of the TATA binding protein on nonspecific DNA". J. Biol. Chem. **270:** 13850-9.

Coleman, R. A., Taggart, A. K., Benjamin, L. R., and Pugh, B. F. (1995). "Dimerization of the TATA binding protein". J. Biol. Chem. **270:** 13842-9.

Collins, K. D. (1997). "Charge density-dependent strength of hydration and biological structure". Biophys. J. **72:** 65-76.

Colombo, M. F., Rau, D. C., and Parsegian, V. A. (1992). "Protein solvation in allosteric regulation: a water effect on hemoglobin". Science **256:** 655-9.

Cormack, B. P., and Struhl, K. (1992). "The TATA-binding protein is required for transcription by all three nuclear RNA polymerases in yeast cells". Cell **69:** 685-96.

Coulombe, B., Li, J., and Greenblatt, J. (1994). "Topological localization of the human transcription factors IIA, IIB, TATA box-binding protein, and RNA polymerase II-associated protein 30 on a class II promoter". J. Biol. Chem. **269:** 19962-7.

Crothers, D. M., and Drak, J. (1992). "Global features of DNA structure by comparative gel electrophoresis". Methods Enzymol. **212:** 46-71.

de Groot, B. L., van Aalten, D. M. F., Amadei, A., and Berendsen, H. J. C. (1996). "The consistency of large concerted motions in proteins in molecular dynamics simulations". Biophys. J. **71:** 1707-13.

DeDecker, B. S., O'Brien, R., Fleming, P. J., Geiger, J. H., Jackson, S. P., and Sigler, P. B. (1996). "The crystal structure of a hyperthermophilic archaeal TATA-box binding protein". J. Mol. Biol. **264:** 1072-84.

Drapper, D. E. (1993). "Protein-DNA complexes: the cost of recognition". Proc. Natl. Acad. Sci. USA **90:** 7429-7430.

Elcock, A. H., and McCammon, J. A. (1996). "The low dielectric interior of proteins is sufficient to cause major structural changes in DNA on association". J. Am. Chem. Soc. **118:** 3787-3788.

Geierstanger, B. H., and Wemmer, D. E. (1995). "Complexes of the minor groove of DNA". Annu. Rev. Biophys. Biomol. Struct. **24:** 463-93.

Geiger, J. H., Hahn, S., Lee, S., and Sigler, P. B. (1996). "Crystal structure of the yeast TFIIA/TBP/DNA complex". Science **272:** 830-6.

Gewirth, D. T., and Sigler, P. B. (1995). "The basis for half-site specificity explored through a non-cognate steroid receptor-DNA complex". Nature Struct. Biol. **2:** 386-94.

Goodsell, D. S., and Dickerson, R. E. (1994). "Bending and curvature calculations in B-DNA". Nucleic Acids Res. **22:** 5497-5503.

Gorin, A. A., Zhurkin, V. B., and Olson, W. K. (1995). "B-DNA twisting correlates with base-pair morphology". J. Mol. Biol. **247:** 34-48.

Griffith, J. D., Makhov, A., Zawel, L., and Reinberg, D. (1995). "Visualization of TBP oligomers binding and bending the HIV-1 and adeno promoters". J. Mol. Biol. **246:** 576-84.

Gryk, M. R., Jardetzky, O., Klig, L. S., and Yanofsky, C. (1996). "Flexibility of DNA binding domain of trp repressor required for recognition of different operator sequences". Protein Sci. **5:** 1195-7.

Guarnieri, F., and Mezei, M. (1996). "Simulated annealing of chemical potential: a general procedure for locating bound waters. Application to the study of the differential hydration propensities of the major and minor grooves of DNA". J. Am. Chem. Soc. **118:** 8493-4.

Guinto, E. R., and Di Cera, E. (1996). "Large heat capacity change in a protein-monovalent cation interaction". Biochemistry **35:** 8800-4.

Guzikevich-Guerstein, G., and Shakked, Z. (1996). "A novel form of the DNA

double helix imposed on the TATA-box by the TATA-binding protein". Nature Struct. Biol. **3:** 32-7.

Hagerman, P. J. (1990). "Sequence-directed curvature of DNA". Annu. Rev. Biochem. **59:** 755-81.

Hagerman, P. J. (1992). "Straightening out the bends in curved DNA". Biochim. Biophys. Acta **1131:** 125-32.

Hahn, S., Buratowski, S., Sharp, P. A., and Guarente, L. (1989). "Yeast TATA-binding protein TFIID binds to TATA elements with both consensus and nonconsensus DNA sequences". Proc. Natl. Acad. Sci. USA **86:** 5718-5722.

Haran, T. E., Kahn, J. D., and Crothers, D. M. (1994). "Sequence elements responsible for DNA curvature". J. Mol. Biol. **244:** 135-143.

Harrington, R. E., and Winicov, I. (1994). "New concepts in protein-DNA recognition: sequence-directed DNA bending and flexibility". Prog. Nucleic Acid Res. Mol. Biol. **47:** 195-270.

Hayward, S., Kitao, A., and Go, N. (1995). "Harmonicity and anharmonicity in protein dynamics: a normal mode analysis and principal component analysis". Proteins **23:** 177-86.

Hayward, S., Kitao, A., Hirata, F., and Go, N. (1993). "Effect of solvent on collective motions in globular protein". J. Mol. Biol. **234:** 1207-17.

Hellman, J., and Chamberlin, M. (1988). "Structure and function of bacterial sigma factors". Annu. Rev. Biochem. **57:** 839-72.

Hernandez, N. (1993). "TBP, a universal eukaryotic transcription factor?". Genes Dev. **7:** 1291-308.

Hirsch, J. A., and Aggarwal, A. K. (1995). "Structure of the even-skipped homeodomain complexed to AT-rich DNA: new perspectives on homeodomain specificity". EMBO J. **14:** 6280-91.

Hoffmann, A., Sinn, E., Yamamoto, T., Wang, J., Roy, A., Horikoshi, M., and Roeder, R. G. (1990). "Highly conserved core domain and unique N terminus with presumptive regulatory motifs in a human TATA factor (TFIID)". Nature **346:** 387-390.

Hogan, M., LeGrange, J., and Austin, B. (1983). "Dependence of DNA helix flexibility on base composition". Nature **304:** 752-4.

Hoopes, B. C., LeBlanc, J. F., and Hawley, D. K. (1992). "Kinetic analysis of yeast TFIID-TATA box complex formation suggests a multi-step pathway". J. Biol. Chem. **267:** 11539-47.

Horikoshi, M., Bertuccioli, C., Takada, R., Wang, J., Yamamoto, T., and Roeder, R. G. (1992). "Transcription factor TFIID induces DNA bending upon binding to the TATA element". Proc. Natl. Acad. Sci. USA **89:** 1060-4.

Hunter, C. A. (1993). "Sequence-dependent DNA Structure. The Role of Base Stacking Interactions". J. Mol. Biol. **230:** 1025-1054.

Hunter, C. A., and Lu, X.-J. (1997). "DNA base stacking interactions: a comparison of theoretical calculations with oligonucleotide X-ray crystal structures". J. Mol. Biol. **265:** 603-19.

Jayaram, B., Aneja, N., Rajasekaran, E., Arora, V., Das, A., Ranganathan, V., and Gupta, V. (1994). "Modelling DNA in aqueous solutions". J. Sci. Ind. Res. **53:** 88-105.

Jayaram, B., and Beveridge, D. L. (1996). "Modeling DNA in aqueous solutions: Theoretical and computer simulation studies of the ion atmosphere of DNA". Annu. Rev. Biophys. Biomol. Struct. **25:** 367-94.

Jen Jacobson, L. (1995). "Structural-perturbation approaches to thermodynamics of site-specific protein-DNA interactions". Methods Enzymol. **259:** 305-44.

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983). "Comparison of simple potential functions for simulating liquid water". J. Chem. Phys. **79:** 926-35.

Juo, Z. S., Chiu, T. K., Leiberman, P. M., Baikalov, I., Berk, A. J., and Dickerson, R. E. (1996). "How proteins recognize the TATA box". J. Mol. Biol. **261:** 239-54.

Kahn, J. D., Yun, E., and Crothers, D. M. (1994). "Detection of localized DNA flexibility". Nature **368:** 163-6.

Kao, C. C., Lieberman, P. M., Schmidt, M. C., Zhou, Q., Pei, R., and Berk, A. J. (1990). "Cloning of a transcriptionally active human TATA binding factor". Science **248:** 1646-50.

Kim, J.-S., Kim, J., Cepek, K. L., Sharp, P. A., and Pabo, C. O. (1997). "Design of TATA box-binding protein/zinc finger fusions for targeted regulation of gene

expression". Proc. Natl. Acad. Sci. USA **94:** 3616-20.

Kim, J. L., and Burley, S. K. (1994). "1.9 A resolution refined structure of TBP recognizing the minor groove of TATAAAAG". Nature Struct. Biol. **1:** 638-53.

Kim, J. L., Nikolov, D. B., and Burley, S. K. (1993). "Co-crystal structure of TBP recognizing the minor groove of a TATA element". Nature **365:** 520-7.

Kim, Y., Geiger, J. H., Hahn, S., and Sigler, P. B. (1993). "Crystal structure of a yeast TBP/TATA-box complex". Nature **365:** 512-20.

Lavery, R., and Sklenar, H. (1989). "Defining the structure of irregular nucleic acids: conventions and principles". J. Biomol. Struct. Dyn. **6:** 655-667.

Lavery, R., and Sklenar, H. (1988). "The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids". J. Biomol. Struct. Dyn. **6:** 63-91.

Lebrun, A., Shakked, Z., and Lavery, R. (1997). "Local DNA stretching mimics the distortion caused by the TATA box-binding protein". Proc. Natl. Acad. Sci. USA **94:** 2993-8.

Lee, D. K., Horikoshi, M., and Roeder, R. G. (1991). "Interaction of TFIID in the minor groove of the TATA element". Cell **67:** 1241-50.

Leuther, K. K., Bushnell, D. A., and Kornberg, R. D. (1996). "Two-dimensional crystallography of TFIIB- and IIE-RNA polymerase II complexes: implications for start site selection and initiation complex formation". Cell **85:** 773-9.

Lewis, M., Chang, G., Horton, N. C., Kercher, M. A., Pace, H. C., Schumacher, M. A., Brennan, R. G., and Lu, P. (1996). "Crystal structure of the lactose operon repressor and its complexes with DNA and inducer". Science **271:** 1247-54.

Li, J. J., Kim, R. H., and Sodek, J. (1995). "An inverted TATA box directs downstream transcription of the bone sialoprotein gene". Biochem. J. **310:** 33-40.

Livingstone, J. R., Spolar, R. S., and Record, M. T. J. (1991). "Contribution to the thermodynamics of protein folding from the reduction in water-accessible nonpolar surface area". Biochemistry **30:** 4237-44.

Lorch, Y., and Kornberg, R. D. (1993). "Near-zero linking difference upon transcription factor IID binding to promoter DNA". Mol. Cell Biol. **13:** 1872-5.

Love, J. J., Li, X., Case, D. A., Giese, K., Grosschedl, R., and Wright, P. E. (1995). "Structural basis for DNA bending by the architectural transcription factor LEF-1". Nature **376:** 791-5.

Lyubchenko, Y. L., Shlyakhtenko, L. S., Appella, E., and Harrington, R. E. (1993). "CA runs increase DNA flexibility in the complex of lambda Cro protein with the OR3 site". Biochemistry **32:** 4121-7.

MacKerell Jr, A. D. (1997). "Influence of magnesium ions on duplex DNA structural, dynamic, and solvation properties". J. Phys. Chem. B **101:** 646-50.

MacKerell Jr, A. D., Wiorkiewicz-Kuczera, J., and Karplus, M. (1995). "An all-atom empirical energy function for the simulation of nucleic acids". J. Am. Chem. Soc. **117:** 11946-11975.

Madan, B., and Sharp, K. (1996). "Heat capacity changes accompanying hydrophobic and ionic solvation: a Monte Carlo and random network model study". J. Phys. Chem. **100:** 7713-21.

Manning, G. S. (1978). "The molecular theory of polyelectrolyte solutions with applications to the electrostatic properties of polynucleotides". Quart. Rev. Biophys. **11:** 179-246.

Marchese, F. T., and Beveridge, D. L. (1984). "Pattern recognition approach to the analysis of geometrical features of solvation: application to the aqueous hydration of Li+, Na+, K+, F-, and Cl-". J. Am. Chem. Soc. **106:** 3713-20.

Marsh, T. L., Reich, C. I., Whitelock, R. B., and Olsen, G. J. (1994). "Transcription factor IID in the Archaea: sequences in the Thermococcus celer genome would encode a product closely related to the TATA-binding protein of eukaryotes". Proc. Natl. Acad. Sci. USA **91:** 4180-4.

Mehrotra, P. K., and Beveridge, D. L. (1980). "Structural analysis of molecular solutions based on quasi-component distribution functions. Application to [H2CO]aq at 25 oC". J. Am. Chem. Soc. **102:** 4287-94.

Mezei, M. (1997). "Optimal position of solute for simulations". J. Comp. Chem. **18:** 812-5.

Miaskiewicz, K., and Ornstein, R. L. (1996). "DNA binding by TATA-box binding protein (TBP): a molecular dynamics computational study". J. Biomol. Struct. Dyn. **13:** 593-600.

Mills, R. (1973). "Self-diffusion in normal and heavy water in the range 1 - 45

degrees". J. Phys. Chem. **77:** 685-8.

Mirzabekov, A. D., and Rich, A. (1979). "Asymmetric lateral distribution of unshielded phosphate groups in nucleosomal DNA and its role in DNA bending". Proc. Natl. Acad. Sci. USA **76:** 1118-21.

Misra, V. K., Hecht, J. L., Sharp, K. A., Friedman, R. A., and Honig, B. (1994). "Salt effects on protein-DNA interactions". J. Mol. Biol. **238:** 264-280.

Narten, A. H., and Levy, H. A. (1971). "Liquid water: molecular correlation functions from X-ray diffraction". J. Chem. Phys. **55:** 2263-9.

Neumann, N., and Steinhauser, O. (1980). "The influence of boundary conditions used in machine simulations on the structure of polar systems". Mol. Phys. **39:** 437-54.

Newman, M., Strzelecka, T., Dorner, L. F., Schildkraut, I., and Aggarwal, A. K. (1995). "Structure of Bam HI endonuclease bound to DNA: partial folding and unfolding on DNA binding". Science **269:** 656-63.

Nikolov, D. B., and Burley, S. K. (1994). "2.1 A resolution refined structure of a TATA box-binding protein (TBP)". Nature Struct. Biol. **1:** 621-37.

Nikolov, D. B., Chen, H., Halay, E. D., Hoffman, A., Roeder, R. G., and Burley, S. K. (1996). "Crystal structure of a human TATA box-binding protein/TATA element complex". Proc. Natl. Acad. Sci. USA **93:** 4862-7.

Nikolov, D. B., Chen, H., Halay, E. D., Usheva, A. A., Hisatake, K., Lee, D. K., Roeder, R. G., and Burley, S. K. (1995). "Crystal structure of a TFIIB-TBP-TATA-element ternary complex". Nature **377:** 119-28.

Nikolov, D. B., Hu, S. H., Lin, J., Gasch, A., Hoffmann, A., Horikoshi, M., Chua, N. H., Roeder, R. G., and Burley, S. K. (1992). "Crystal structure of TFIID TATA-box binding protein". Nature **360:** 40-6.

Nordenskiold, L., Chang, D. K., Anderson, C. F., and Record Jr, M. T. (1984). "23Na NMR relaxation study of the effects of conformation and base composition on the interactions of counterions with double-helical DNA". Biochemistry **23:** 4309-4317.

Olmsted, M. C., Anderson, C. F., and Record Jr, M. T. (1989). "Monte Carlo description of oligoelectrolyte properties of DNA oligomers: range of the end effect and the approach of molecular and thermodynamic properties to the polyelectrolyte limits". Proc. Natl. Acad. Sci. U S A **86:** 7766-7770.

Olson, W. K., Babcock, M. S., Gorin, A., Liu, G., Marky, N. L., Martino, J. A., Pedersen, S. C., Srinivasan, A. R., Tobias, I., Westcott, T. P., and et al. (1995). "Flexing and folding double helical DNA". Biophys. Chem. **55:** 7-29.

Orphanides, G., Lagrange, T., and Reinberg, D. (1996). "The general transcription factors of RNA polymerase II". Genes Dev. **10:** 2657-83.

Pack, G. R., Garrett, G. A., Wong, L., and Lamm, G. (1993). "The effect of a variable dielectric coefficient and finite ion size on Poisson-Boltzmann calculations of DNA-electrolyte systems". Biophys. J. **65:** 1363-70.

Parkhurst, K. M., Brenowitz, M., and Parkhurst, L. J. (1996). "Simultaneous binding and bending of promoter DNA by the TATA binding protein: real time kinetic measurements". Biochemistry **35:** 7459-65.

Parsegian, V. A., Rand, R. P., and Rau, D. C. (1995). "Macromolecules and water: probing with osmotic stress". Methods Enzymol. **259:** 43-94.

Parvin, J. D., McCormick, R. J., Sharp, P. A., and Fisher, D. E. (1995). "Pre-bending of a promoter sequence enhances affinity for the TATA-binding factor". Nature **373:** 724-27.

Pastor, N., Pardo, L., and Weinstein, H. (1997). "Does TATA Matter? A structural exploration of the selectivity determinants in its complexes with TBP". Biophys. J. **73:** in press.

Pastor, N., and Weinstein, H. (1995). "Electrostatic analysis of DNA binding properties in lysine to leucine mutants of TATA-box binding proteins". Protein Eng. **8:** 543-9.

Pearlman, D. A., Case, D. A., Caldwell, J. W., Ross, W. S., Cheatham III, T. E., Ferguson, D. M., Seibel, G. L., Singh, U. C., Weiner, P. K., and Kollman, P. A. (1995). "AMBER 4.1". University of California, San Francisco.

Perez-Howard, G. M., Weil, P. A., and Beechem, J. M. (1995). "Yeast TATA binding protein interaction with DNA: fluorescence determination of oligomeric state, equilibrium binding, on-rate, and dissociation kinetics". Biochemistry **34:** 8005-17.

Pertsemlidis, A., Saxena, A. M., Soper, A. K., Head Gordon, T., and Glaeser, R. M. (1996). "Direct evidence for modified solvent structure within the hydration shell of a hydrophobic amino acid". Proc. Natl. Acad. Sci. USA **93:** 10769-74.

Petri, V., Hsieh, M., and Brenowitz, M. (1995). "Thermodynamic and kinetic characterization of the binding of the TATA binding protein to the adenovirus E4 promoter". Biochemistry **34:** 9977-84.

Poncin, M., Hartmann, B., and Lavery, R. (1992). "Conformational sub-states in B-DNA". J. Mol. Biol. **226:** 775-94.

Poncin, M., Piazzola, D., and Lavery, R. (1992). "DNA flexibility as a function of allomorphic conformation and of base sequence". Biopolymers **32:** 1077-103.

Poon, D., Knittle, R. A., Sabelko, K. A., Yamamoto, T., Horikoshi, M., Roeder, R. G., and Weil, P. A. (1993). "Genetic and biochemical analyses of yeast TATA-binding protein mutants". J. Biol. Chem. **268:** 5005-13.

Price, M. A., and Tullius, T. D. (1993). "How the structure of an adenine tract depends on sequence context: a new model for the structure of TnAn DNA sequences". Biochemistry **32:** 127-36.

Pugh, B. F., and Tjian, R. (1991). "Transcription from a TATA-less promoter requires a multisubunit TFIID complex". Genes Dev. **5:** 1935-45.

Ravishanker, G., Swaminathan, S., Beveridge, D. L., Lavery, R., and Sklenar, H. (1989). "Conformational and helicoidal analysis of 30 ps of molecular dynamics on the d(CGCGAATTCCGCG) double helix: "Curves", Dials and Windows". J. Biomol. Struct. Dyn. **6:** 669-699.

Record Jr, M. T., Anderson, C. F., and Lohman, T. M. (1978). "Thermodynamic analysis of ion effects on the binding and conformational equilibria of proteins and nucleic acids: the roles of ion association or release, screening, and ion effects on water activity". Quart. Rev. Biophys. **11:** 103-78.

Record Jr, M. T., Ha, J.-H., and Fisher, M. A. (1991). "Analysis of equilibrium and kinetic measurements to determine thermodynamic origins of stability and specificity and mechanism of formation of site-specific complexes between proteins and helical DNA". Methods Enzymol. **208:** 291-343.

Record Jr, M. T., and Lohman, T. M. (1978). "A semiempirical extension of polyelectrolyte theory to the treatment of oligoelectrolytes: application to oligonucleotide helix-coil transitions". Biopolymers **17:** 159-166.

Record Jr, M. T., Lohman, T. M., and de Haseth, P. (1976). "Ion effects on ligand-nucleic acid interactions". J. Mol. Biol. **107:** 145-158.

Reddy, M. R., Rossky, P. J., and Murthy, C. S. (1987). "Counterion spin

relaxation in DNA solutions: a stochastic dynamics simulation study". J. Phys. Chem. **91:** 4923-4933.

Reddy, P., and Hahn, S. (1991). "Dominant negative mutations in yeast TFIID define a bipartite DNA-binding region". Cell **65:** 349-57.

Rhodes, D., Schwabe, J. W. R., Chapman, L., and Fairall, L. (1996). "Towards and understanding of protein-DNA recognition". Phil. Trans. R. Soc. Lond. B **351:** 501-9.

Ribas de Pouplana, L., Auld, D. S., Kim, S., and Schimmel, P. (1996). "A mechanism for reducing entropic cost of induced fit in protein--RNA recognition". Biochemistry **35:** 8095-102.

Rice, P. A., Yang, S.-w., Mizuuchi, K., and Nash, H. A. (1996). "Crystal structure of an IHF-DNA complex: a protein-induced DNA U-turn". Cell **87:** 1295-306.

Robert, F., Forget, D., Li, J., Greenblatt, J., and Coulombe, B. (1996). "Localization of subunits of transcription factors IIE and IIF immediately upstream of the transcriptional initiation site of the adenovirus major late promoter". J. Biol. Chem. **271:** 8517-20.

Roeder, R. G. (1996). "The role of general initiation factors by RNA polymerase II". Trends Biochem. Sci. **21:** 327-35.

Rowlands, T., Baumann, P., and Jackson, S. P. (1994). "The TATA-binding protein: a general transcription factor in eukaryotes and archaebacteria". Science **264:** 1326-9.

Saenger, W. (1983). "Principles of nucleic acid structure", (New York: Springer-Verlag).

Sanghani, S. R., Zakrzewska, K., Harvey, S. C., and Lavery, R. (1996). "Molecular modelling of (A4T4NN)n and (T4A4NN)n: sequence elements responsible for curvature". Nucleic Acids Res. **24:** 1632-7.

Sarai, A., Jernigan, R. L., and Mazur, J. (1996). "Interdependence of conformational variables in double-helical DNA". Biophys. J. **71:** 1507-18.

Sarai, A., Mazur, J., Nussinov, R., and Jernigan, R. L. (1988). "Origin of DNA helical structure and its sequence dependence". Biochemistry **27:** 8498-8502.

Sarai, A., Mazur, J., Nussinov, R., and Jernigan, R. L. (1989). "Sequence

dependence of DNA conformational flexibility". Biochemistry **28:** 7842-7849.

Schneider, B., and Berman, H. M. (1995). "Hydration of the DNA bases is local". Biophys. J. **69:** 2661-9.

Schultz, M. C., Reeder, R. H., and Hahn, S. (1992). "Variants of the TATA-binding protein can distinguish subsets of RNA polymerase I, II, and III promoters". Cell **69:** 697-702.

Schumacher, M. A., Choi, K. Y., Zalkin, H., and Brennan, R. G. (1994). "Crystal structure of LacI member, PurR, bound to DNA: minor groove binding by alpha helices". Science **266:** 763-770.

Schwabe, J. W., Fairall, L., Chapman, L., Finch, J. T., Dutnall, R. N., and Rhodes, D. (1993). "The cocrystal structures of two zinc-stabilized DNA-binding domains illustrate different ways of achieving sequence-specific DNA recognition". Cold Spring Harb. Symp. Quant. Biol. **58:** 141-7.

Schwabe, J. W. R., Chapman, L., and Rhodes, D. (1995). "The oestrogen receptor recognizes an imperfectly palindromic response element through an alternative side-chain conformation". Structure **3:** 201-213.

Seeman, N. C., Rosenberg, J. M., and Rich, A. (1976). "Sequence-specific recognition of double helical nucleic acids by proteins". Proc. Natl. Acad. Sci. USA **73:** 804-8.

Shakked, Z., Rabinovich, D., Kennard, O., Cruse, W. B., Salisbury, S. A., and Viswamitra, M. A. (1983). "Sequence-dependent conformation of an A-DNA double helix. The crystal structure of the octamer d(G-G-T-A-T-A-C-C)". J. Mol. Biol. **166:** 183-201.

Sharp, K. A. (1995). "Polyelectrolyte electrostatics: salt dependence, entropic, and enthalpic contributions to free energy in the nonlinear Poisson-Boltzmann model". Biopolymers **36:** 227-243.

Sharp, K. A., Friedman, R. A., Misra, V., Hecht, J., and Honig, B. (1995). "Salt effects on polyelectrolyte-ligand binding: comparison of Poisson-Boltzmann, and limiting law/counterion binding models". Biopolymers **36:** 245-62.

Sklenar, H., Etchebest, C., and Lavery, R. (1989). "Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis". Proteins **6:** 46-60.

Spolar, R. S., Ha, J. H., and Record, M. T. J. (1989). "Hydrophobic effect in

protein folding and other noncovalent processes involving proteins". Proc. Natl. Acad. Sci. USA **86:** 8382-5.

Spolar, R. S., Livingstone, J. R., and Record Jr, M. T. (1992). "Use of liquid hydrocarbon and amide transfer data to estimate contributions to thermodynamic functions of protein folding from the removal of nonpolar and polar surface from water". Biochemistry **31:** 3947-3955.

Spolar, R. S., and Record, M. T. J. (1994). "Coupling of local folding to site-specific binding of proteins to DNA". Science **263:** 777-84.

Srinivasan, A. R., Torres, R., Clark, W., and Olson, W. K. (1987). "Base sequence effects in double helical DNA. I. Potential energy estimates of local base morphology". J. Biomol. Struct. Dyn. **5:** 459-96.

Starr, D. B., and Hawley, D. K. (1991). "TFIID binds in the minor groove of the TATA box". Cell **67:** 1231-40.

Starr, D. B., Hoopes, B. C., and Hawley, D. K. (1995). "DNA bending is an important component of site-specific recognition by the TATA binding protein". J. Mol. Biol. **250:** 434-46.

Steinbach, P. J., and Brooks, B. R. (1996). "Hydrated myoglobin's anharmonic fluctuations are not primarily due to dihedral transitions". Proc. Natl. Acad. Sci. USA **93:** 55-9.

Strauss, J. K., Prakash, T. P., Roberts, C., Switzer, C., and Maher, I., L.J. (1996). "DNA bending by a phantom protein". Chemistry & Biology **3:** 671-8.

Strubin, M., and Struhl, K. (1992). "Yeast and human TFIID with altered DNA-binding specificity for TATA elements". Cell **68:** 721-30.

Struhl, K. (1994). "Duality of TBP, the universal transcription factor". Science **263:** 1103-4.

Sturtevant, J. M. (1977). "Heat capacity and entropy changes in processes involving proteins". Proc. Natl. Acad. Sci. USA **74:** 2236-40.

Suck, D., Lahm, A., and Oefner, C. (1988). "Structure refined to 2A of a nicked DNA octanucleotide complex with DNase I". Nature **332:** 464-468.

Suzuki, M., Allen, M. D., Yagi, N., and Finch, J. T. (1996). "Analysis of co-crystal structures to identify the stereochemical determinants of the orientation of TBP on the TATA box". Nucleic Acids Res. **24:** 2767-73.

Suzuki, M., and Yagi, N. (1995). "Stereochemical basis of DNA bending by transcription factors". Nucleic Acids Res. **23:** 2083-91.

Suzuki, M., Yagi, N., and Finch, J. T. (1996). "Role of base-backbone and base-base interactions in alternating DNA conformations". FEBS Lett. **379:** 148-52.

Tan, S., Hunziker, Y., Sargent, D. F., and Richmond, T. J. (1996). "Crystal structure of a yeast TFIIA/TBP/DNA complex". Nature **381:** 127-51.

Tang, H., Sun, X., Reinberg, D., and Ebright, R. H. (1996). "Protein-protein interactions in eukaryotic transcription initiation: structure of the preinitiation complex". Proc. Natl. Acad. Sci. USA **93:** 1119-24.

Tippin, D. B., and Sundaralingam, M. (1997). "Comparison of major groove hydration in isomorphous A-DNA octamers and dependence on base sequence and local helix geometry". Biochemistry **36:** 536-43.

Travers, A. A. (1991). "DNA bending and kinking - sequence dependence and function". Curr. Opin. Struct. Biol. **1:** 114-122.

Travers, A. A. (1992). "DNA conformation and configuration in protein-DNA complexes". Curr. Opin. Struct. Biol. **2:** 71-77.

Ulyanov, N. B., and James, T. L. (1995). "Statistical analysis of DNA duplex structural features". Methods Enzymol. **261:** 90-120.

van Dijk, L., Gruwel, M. L. H., Jesse, W., de Bleijser, J., and Leyte, J. C. (1987). "Sodium ion and solvent nuclear relaxation results in aqueous solutions of DNA". Biopolymers **26:** 261-284.

von Hippel, P. H., and Berg, O. G. (1986). "On the specificity of DNA-protein interactions". Proc. Natl. Acad. Sci. USA **83:** 1608-1612.

Wang, Y., Jensen, R. C., and Stumph, W. E. (1996). "Role of TATA box sequence and orientation in determining RNA polymerase II/III transcription specificity". Nucleic Acids Res. **24:** 3100-6.

Wang, Y., and Stumph, W. E. (1995). "RNA polymerase II/III transcription specificity determined by TATA box orientation". Proc. Natl. Acad. Sci. USA **92:** 8606-10.

Wemmer, D. E., and Dervan, P. B. (1997). "Targeting the minor groove of DNA".

Curr. Opin. Struct. Biol. **7:** 355-61.

Werner, M. H., and Burley, S. K. (1997). "Architectural transcription factors: proteins that remodel DNA". Cell **88:** 733-6.

Werner, M. H., Gronenborn, A. M., and Clore, G. M. (1996). "Intercalation, DNA kinking, and the control of transcription". Science **271:** 778-84.

Werner, M. H., Huth, J. R., Gronenborn, A. M., and Clore, G. M. (1995). "Molecular basis of human 46X,Y sex reversal revealed from the three-dimensional solution structure of the human SRY-DNA complex". Cell **81:** 705-14.

White, S., Baird, E. E., and Dervan, P. B. (1996). "Effects of the A.T/T.A degeneracy of pyrrole-imidazole polyamide recognition in the minor groove of DNA". Biochemistry **1996:** 12532-7.

Whitehall, S. K., Kassavetis, G. A., and Geiduschek, E. P. (1995). "The symmetry of the yeast U6 RNA gene's TATA box and the orientation of the TATA-binding protein in yeast TFIIIB". Genes Dev. **9:** 2974-85.

Wilson, D. S., Guenther, B., Desplan, C., and Kuriyan, J. (1995). "High resolution crystal structure of a paired (Pax) class cooperative homeodomain dimer on DNA". Cell **82:** 709-19.

Wobbe, C. R., and Struhl, K. (1990). "Yeast and human TATA-binding proteins have nearly identical DNA sequence requirements for transcription in vitro". Mol. Cell Biol. **10:** 3859-67.

Wong, J. M., and Bateman, E. (1994). "TBP-DNA interactions in the minor groove discriminate between A:T and T:A base pairs". Nucleic Acids Res. **22:** 1890-6.

Xu, L. C., Thali, M., and Schaffner, W. (1991). "Upstream box/TATA box order is the major determinant of the direction of transcription". Nucleic Acids Res. **19:** 6699-704.

Yamamoto, T., Horikoshi, M., Wang, J., Hasegawa, S., Weil, P. A., and Roeder, R. G. (1992). "A bipartite DNA binding domain composed of direct repeats in the TATA box binding factor TFIID". Proc. Natl. Acad. Sci. USA **89:** 2844-8.

Yanagi, K., Prive, G. G., and Dickerson, R. E. (1991). "Analysis of local helix geometry in three B-DNA decamers and eight dodecamers". J. Mol. Biol. **217:** 201-14.

Yang, L., and Pettitt, B. M. (1996). "B to A transition of DNA on the nanosecond time scale". J. Phys. Chem. **100:** 2564-2566.

Young, M. A., Jayaram, B., and Beveridge, D. L. (1997). "Intrusion of counterions into the spine of hydration in the minor groove of B-DNA: fractional occupancy of electronegative pockets". J. Am. Chem. Soc. **117:** 59-69.

Young, M. A., Ravishanker, G., and Beveridge, D. L. (1995). "Analysis of local helix bending in crystal structures of DNA oligonucleotides and DNA-protein complexes". Biophys. J. **68:** 2454-68.

Zakrzewska, K. (1992). "Static and dynamic conformational properties of AT sequences in B-DNA". J. Biomol. Struct. and Dynam. **9:** 681-93.

Zhang, W., Bond, J. P., Anderson, C. F., Lohman, T. M., and Record, M. T., Jr. (1996). "Large electrostatic differences in the binding thermodynamics of a cationic peptide to oligomeric and polymeric DNA". Proc. Natl. Acad. Sci. USA **93:** 2511-6.

Zhurkin, V. B. (1985). "Sequence-dependent bending of DNA and phasing of nucleosomes". J. Biomol. Struct. Dyn. **2:** 785-804.

Zhurkin, V. B., Lysov, Y. P., and Ivanov, V. I. (1979). "Anisotropic flexibility of DNA and the nucleosomal structure". Nucleic Acids Res. **6:** 1081-96.

Zhurkin, V. B., Ulyanov, N. B., Gorin, A. A., and Jernigan, R. L. (1991). "Static and statistical bending of DNA evaluated by Monte Carlo simulations". Proc. Natl. Acad. Sci. USA **88:** 7046-50.