Title
Structure and dynamics of nucleic acids and proteins by NMR

Author
Tonelli, Marco,

Publication Date
2003

Peer reviewed|Thesis/dissertation

Structure and Dynamics of Nucleic Acids and Proteins by NMR

by

Marco Tonelli

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

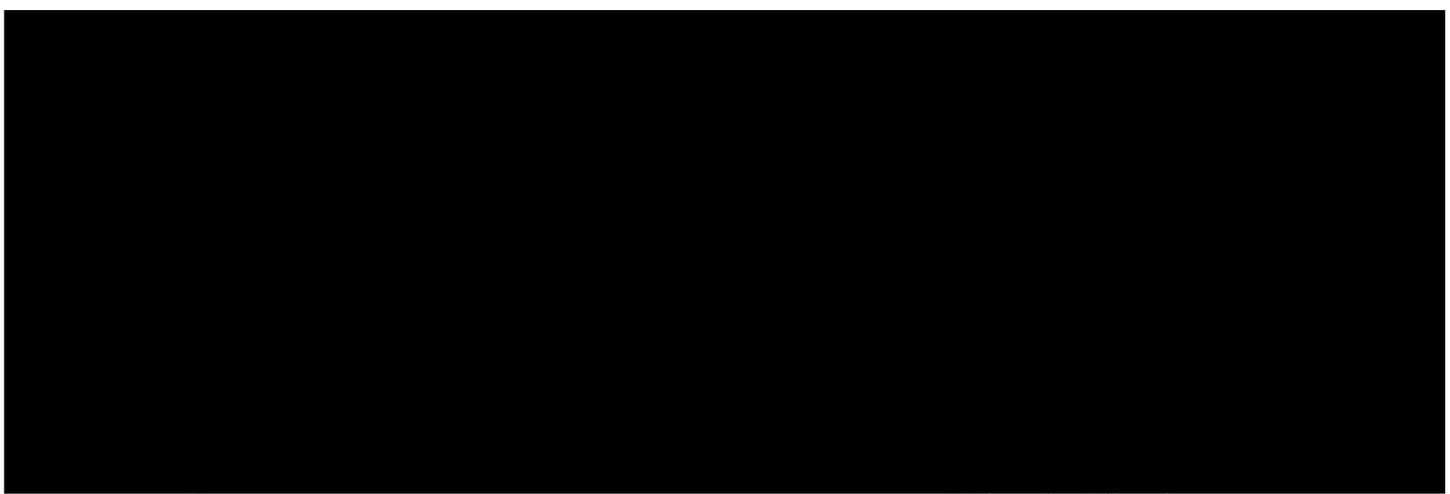
Pharmaceutical Chemistry

in the

GRADUATE DIVISION

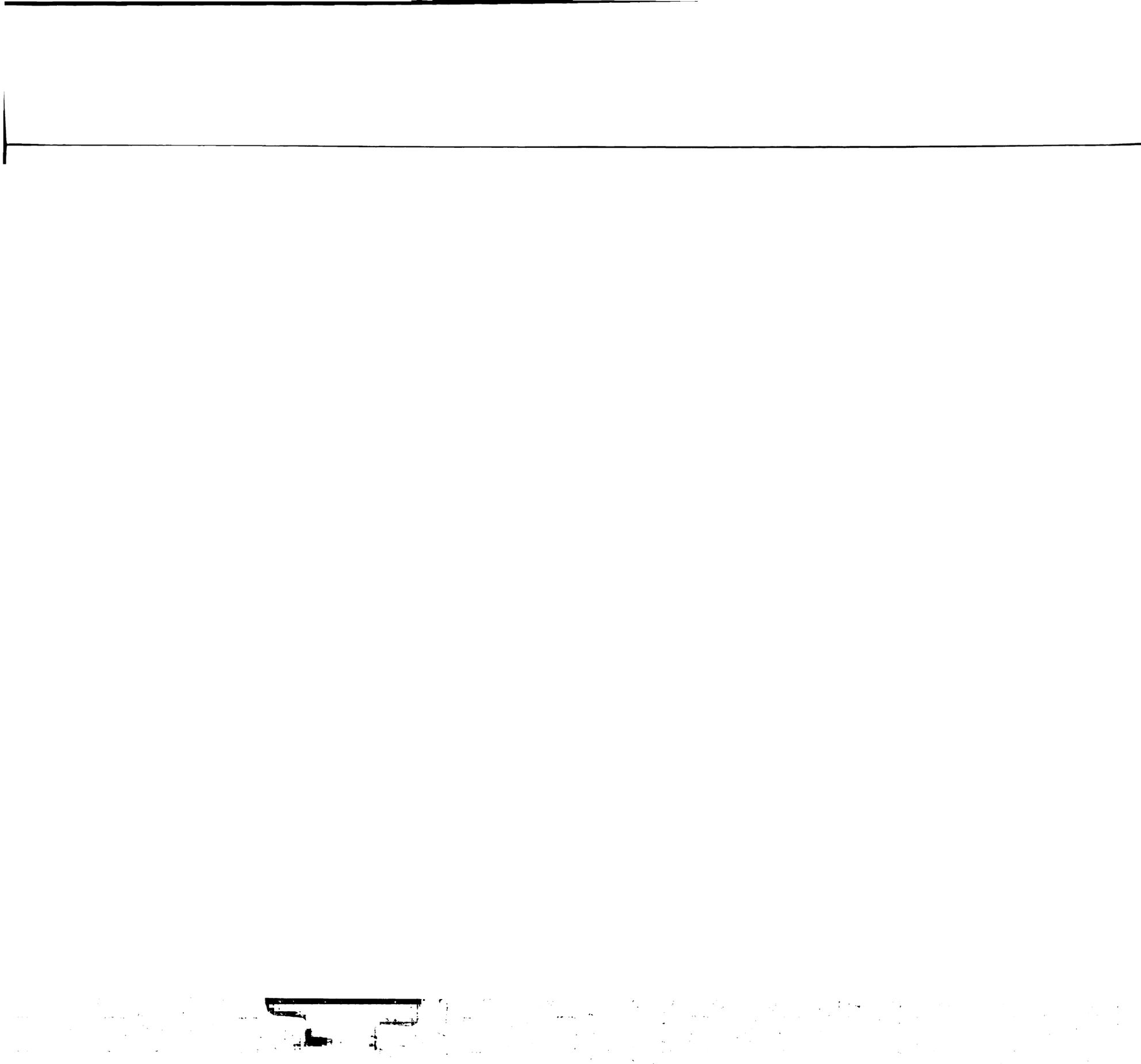
of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO



Date University Librarian

Degree Conferred:.....



To Tommaso and Matteo.

ACKNOWLEDGMENTS

Many people were instrumental to the completion of my doctorate. First of all, I wish to thank my advisor, Tom James, for giving me the opportunity to learn NMR while doing research in his laboratory and for always been there to guide my efforts in the right direction, Dave Agard, for being a second advisor to me and a great example to motivate me to work harder, Volker Doetsch, Dick Shafer, C.C. Wang and Peter Kollman, for being part of my thesis and/or orals committee. Special thanks to Vladimir Basus and my brother Davide for their friendship and technical help with the NMR machines and the computers. I am also very grateful to Anna Maria Bianucci, for bringing me to UCSF almost 10 years ago, and Reuben Peters, for preparing all the NMR samples of sTva47 that I ever needed to solve its structure.

Throughout the years I spent studying at UCSF, I've interacted with a lot of people from different laboratories. Many names come to mind, in no particular order I wish to thank for friendship, advise and help: Mark Grillo, Daina Avizonis, Chojiro Kojima, Anwer Mujeeb, Nick Ulyanov, Letitia Yao, Uli Schmitz, Alessandro Donati, He Liu, Shauna Farr-Jones, Carlos Gonzalez, Anton Filikov, Patrick Furrer, Tom Goddard, Yoko Shibata, Peter Lukawsky, Morten Sorensen, Don Kneller, Gary Wolfe, Hong Zhang, Moriz Mayer, Luis Comolli, Miriam Gochin.

Last but not least, I wish to thank my parents Giancarlo and Gabriella who have given me the opportunity to pursue my studies and have always been very supportive of the choices I made. And thanks most especially to my wife Grazia, who traveled half way around the world to be with me, and my two sons, Tommaso and Matteo, for giving a sense to my life.

Structure and Dynamics of Nucleic Acids and Proteins by NMR

Marco Tonelli

This thesis describes the use of nuclear magnetic resonance techniques to study the structure and dynamics of biological macromolecules in solution. Analysis of two-dimensional homonuclear spectra and complete relaxation matrix methods have been used to solve the high-resolution structure of an eleven base-pair DNA duplex that contains the consensus sequence for promoters recognized by σ^K RNA polymerase, and two modified ten base-pair DNA-RNA duplexes: [Rp]- and [Sp]-phosphorothioated hybrid duplexes, that have the proR or the proS oxygen atom of all the backbone phosphate groups in the DNA strand replaced by a sulfur, respectively. Studying the structure of phophorothioated hybrid duplexes is important to understand the mode of action of antisense inhibitors. On the other hand, multidimensional heteronuclear NMR techniques have provided the experimental information to refine the structure of a 47 amino acid fragment of the Tva receptor extracellular domain that is used by avian leucosis and sarcoma viruses to target host cells. Furthermore, every one of these molecules has clearly shown evidence of flexibility that I further investigated by taking advantage of the NMR capabilities. This is particularly important to provide a complete description of biological macromolecules in solution that is critical to fully understand their function.

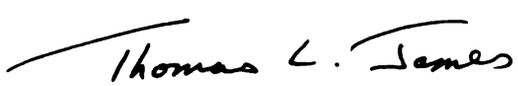


TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Appendix A

LIST OF TABLES

Chapter 2

Chapter 4

Chapter 5

Appendix A

1. Studying the Structu

Nuclear Magnetic R

This thesis describes the w
solid macromolecules in s
very attractive technique to
the complete description of t
that not only can NMR be
provides the means to invest
atomic resolution. Thus, even
X-ray crystallography studies.
behavior of the molecule i
axis of the growing numb
mode of a molecule is not s
facility needs to be taken i

On the other hand, NMR i
interaction of molecule in
nucleic acids and proteins, s
Subtle NMR spectra. Even
difficult or impossible to st
to make complicate things.

observations. 1D NMR, on
of two-dimensional nuclea
relation spectroscopy. TOC
applied restrained molec
the molecular dyn
in an

CHAPTER 1

1. Studying the Structure of Macromolecules in Solution by using Nuclear Magnetic Resonance Techniques

This thesis describes the work that I have done to investigate the structure of biological macromolecules in solution using nuclear magnetic resonance methods. NMR is a very attractive technique to study macromolecules that has the capacity of yielding a more complete description of the structure of a molecule than any other method available. In fact, not only can NMR be used to obtain a static picture of a molecule, but it also provides the means to investigate its dynamic behavior over a range of time scales with atomic resolution. Thus, even when a high resolution structure is already available from x-ray crystallography studies, NMR can still be useful to add a dynamic description of the behavior of the molecule in solution. This is of particular importance, since the analysis of the growing number of structures available has clearly indicated that a static picture of a molecule is not sufficient to fully account for function, and molecular flexibility needs to be taken into account.

On the other hand, NMR is a very insensitive technique that requires a high concentration of molecule in solution (high μM to mM). For large macromolecules, like nucleic acids and proteins, such high concentrations can lead to aggregation and, hence, unusable NMR spectra. Even when there is no aggregation, large molecules may be difficult or impossible to study due to the increasing number of signals that can overlap. To further complicate things, large molecules have also very efficient relaxation

---

Abbreviations. 1D NMR, one-dimensional NMR; 2D NMR, two-dimensional NMR; 2D NOE, two-dimensional nuclear Overhauser effect; 2QF-COSY, double-quantum-filtered correlation spectroscopy; TOCSY, total correlated spectroscopy; FID, free induction decay; rMD, restrained molecular dynamics; rMC, restrained Monte Carlo calculations; MDtar, molecular dynamics with time-averaged restraints; RMSD, root-mean-square deviation; R^x, sixth-root R-factor; RDC, residual dipolar coupling.

mechanisms that result in very broad peaks in NMR spectra. However, the use of stronger magnetic fields, more sensitive probes and the development of new pulse sequences and labeling strategies have been pushing NMR to study bigger and more complicated systems.

During the course of my PhD studies, I elucidated the structure of four biologically relevant macromolecules in solution. Every one of these molecules has clearly shown evidence of flexibility that I further investigated by taking advantage of the NMR capabilities. These structural studies are reported in this thesis. **Chapter 1** introduces the NMR methods used to study the structure of the macromolecules reported in later chapters. **Chapter 2** describes the structure of an eleven base-pair DNA duplex that contains the consensus sequence for promoters recognized by σ^K RNA polymerase. **Chapter 3** reports a discussion about internal flexibility of DNA duplexes through the analysis of inconsistencies in accurate NOE derived interproton distances. **Chapter 4** describes the high resolution structure of two modified ten base-pair DNA-RNA duplexes: [Rp]- and [Sp]-phosphorothio hybrid duplexs, that have the proR or the proS oxygen atom of all the backbone phosphate groups in the DNA strand replaced by a sulfur, respectively. Finally, **chapter 5** describes the use of heteronuclear multidimensional NMR to solve the solution structure of a 47 amino acid fragment of the Tva receptor extracellular domain. This receptor is used by avian leucosis and sarcoma viruses to enter host cells.

This chapter is not intended as a general overview of NMR methods but rather to introduce the techniques that I have used in my work of solving structures by NMR. The experimental details are reported in later chapters together with the results and discussion of each system.

1.1 Sample preparation

Once the macromolecule under investigation is available in purified form, a sample is prepared by dissolving it in the proper buffer at a certain pH and used to run NMR experiments at a chosen temperature. Changing the sample conditions will affect the appearance of the NMR spectra and we need to carefully select those conditions that will

...ch is the most exploitable
...ing the sample on the a
...nds the secondary structure
...spectra acquired in H_2O, sinc
...protons, the number and disp
...fingerprint of the structure of
...mple preparation. How ever,
...ucture calculation is interpr
...is also useful to run NOESY
...ch is the highest number o
...The main determinants of an N
...ed by the pH, salt and the temp
...of the macromolecule in the sa
...dimensional NMR spectr
...ay be too high as to get aggre
...ult in broadening of the peak
...physiological pH. How ever, th
...ranted in order to reduce the
...ment. Figure 1.1) (Creigh
...ludes the salt used for buffer
...ecules require the presence
...ecific calcium binding site, or
...ino acids. Finally, the temper
...macromolecule (e.g., for nuc
...ion. Within this limit, a hig
...ves better resolved spectra, by
...ucts (lower t_c), but it may al
...ctions in nucleic acids ar
...ange with bulk water.

result in the most exploitable spectra. This is best done directly by analyzing the effect of changing the sample on the appearance of certain NMR spectra. For example, in nucleic acids the secondary structure can be monitored by the presence of imino protons in 1D spectra acquired in H_2O, since H bonded imino protons can be observed. In ^{15}N labeled proteins, the number and dispersion of peaks in 2D ^{15}N HSQC spectra provides a fingerprint of the structure of the molecule making this an ideal experiment to use during sample preparation. However, since the most important information that will be used for structure calculation is interproton distances estimated from NOE cross-peak intensities, it is also useful to run NOESY experiments and choose those sample conditions that result in the highest number of cross-peaks in NOESY spectra.

The main determinants of an NMR sample are the concentration of the molecule under study, the pH, salt and the temperature chosen to acquire the spectra. The concentration of the macromolecule in the sample should be sufficiently high to be able to acquire multidimensional NMR spectra with good signal-to-noise ratio within a reasonable time, but not too high as to get aggregation of the macromolecule in solution (which would result in broadening of the peaks). Ideally, the pH should be as close as possible to the physiological pH. However, this is rarely possible for proteins, where we need the pH to be acidic in order to reduce the rate of exchange of amide protons with the bulk water to a minimum (Figure 1.1) (Creighton, 1993). Salt concentration is also important and includes the salt used for buffering the pH to the chosen value. Furthermore, some molecules require the presence of particular ions, like Ca^{2+} for Tva, since it contains a specific calcium binding site, or Mg^{2+} that is known to stabilize the structure of certain nucleic acids. Finally, the temperature should be well below the unfolding temperature of the macromolecule (e.g., for nucleic acid duplexes, below the melting temperature of the duplex). Within this limit, a higher temperature will result in sharper NMR peaks and, hence, better resolved spectra, by increasing the tumbling rate of the macromolecule in solution (lower τ_c), but it may also reduce the intensity of peaks involving imino and amino protons in nucleic acids and amide protons in proteins by increasing the rate of exchange with bulk water.

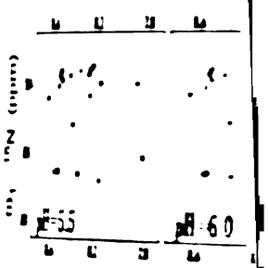


Figure 1.1. The same region
N-labeled sTva47 is sho
SQC peaks at higher pH va
with bulk water solvent.

1.2 Resonance Assign

Once the sample has been
assigning the resonances of th
assignments are used to extra
ically, in large molecules lik
ifying the number of proto
experiments are needed in ord
ecule in solution.

1.2.1 *Homonuclear 2D NMR*

Assignments for smaller pro
homonuclear NMR experiments
monly used are two-dimens
ally collected first with t
exchangeable proton resonance
changed to deuterium and, th
can be collected in H_2O to ass
2D NOESY spectra, we o
ace in space; the intensity of a
directly proportional to the dist
formation about the spatial arra

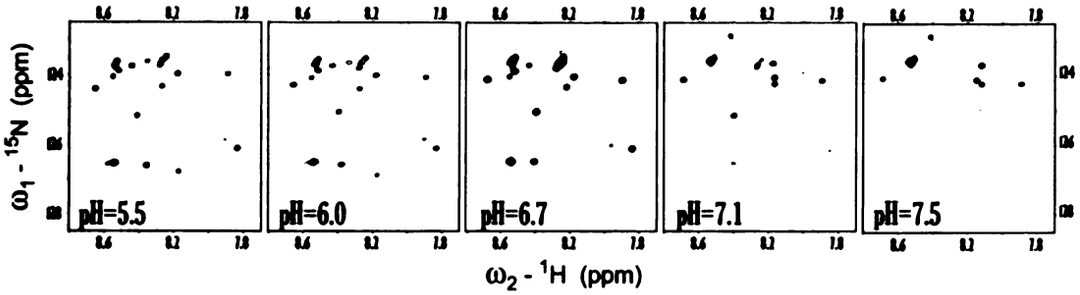


Figure 1.1. The same region of five ^{15}N-HSQC experiments acquired at increasing pH on ^{15}N-labeled sTva47 is shown. This series shows a decrease of the intensity of the HSQC peaks at higher pH values due to an increased rate of exchange of amide protons with bulk water solvent.

1.2 Resonance Assignments

Once the sample has been prepared, the next step is to acquire NMR experiments for assigning the resonances of the relevant nuclei in the molecule. Subsequently, assignments are used to extract useful structural information from other NMR spectra. Usually, in large molecules like the proteins and nucleic acids that we are interested in studying, the number of proton resonances is sufficiently high that multidimensional experiments are needed in order to extract useful information about the structure of the molecule in solution.

1.2.1 Homonuclear 2D NMR

Assignments for smaller proteins and nucleic acids can be accomplished by using 2D homonuclear NMR experiments with unlabeled samples. The experiments that are most commonly used are two-dimensional NOESY, TOCSY and COSY. These experiments are usually collected first with the sample dissolved in pure $^{2}H_2O$ to assign non-exchangeable proton resonances (while protons attached to nitrogen and oxygen are exchanged to deuterium and, thus, not observed in the spectra). Subsequently, spectra can also be collected in $^{1}H_2O$ to assign and extract information from exchangeable protons.

In 2D NOESY spectra, we observe off-diagonal peaks between protons that are nearby in space; the intensity of an NOE cross-peak observed between two protons is inversely proportional to the distance that separates them. Thus, NOESY spectra carry information about the spatial arrangement of the protons in the molecule and provide the

means for assigning their resonances by allowing one to perform sequential walks along the backbone of proteins or along the strands of nucleic acid duplexes. In TOCSY and COSY experiments, instead, the magnetization between nuclei is exchanged through scalar coupling, thus resulting in cross-peaks that correlate protons that are close within the chemical structure of the molecule. Thus, by combining the intraresidue and sequential connectivity provided by the spatial information extracted from NOESY spectra with the chemical bond connectivity provided by analyzing COSY and TOCSY experiments, most proton resonances in smaller proteins and nucleic acids can be assigned (Wüthrich, 1986).

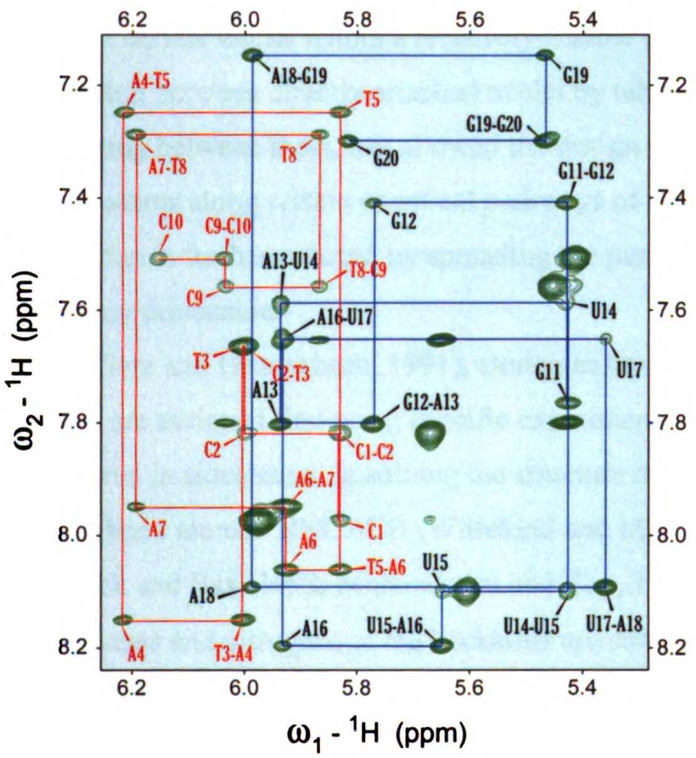


Figure 1.2. Portion of the 200ms NOESY spectrum for the [Sp]-hybrid, showing the sequential base-H1' fingerprint walks for the PSO (red) and RNA (blue) strands. For convenience, only cross-peaks along the sequential walks are labeled.

For example, I assigned the non-exchangeable protons of my DNA duplex and the phosphorothioated DNA-RNA hybrids duplexes, by identifying base-H1', base-H2'/H2" sequential walks in 2D NOESY spectra and using TOCSY and COSY spectra to identify H3', H4' and where possible H5'/H5" resonances. These spectra were recorded with the

samples dissolved in D_2O. Subsequently, I collected 2D NOESY spectra in H_2O to assign imino and amino protons in bases. Figure 1.2 shows the base-H1' sequential walks on the relevant portion of a 2D NOESY spectrum recorded with a 200ms mixing time on the [Sp]-PSO-RNA hybrid duplex sample.

1.2.2 Heteronuclear Multidimensional NMR

In larger molecules, severe resonance overlapping makes assignments very difficult or impossible even using two-dimensional spectra. Thus, to assign large molecules, other strategies have been developed that require acquiring multidimensional heteronuclear experiments with ^{15}N and/or ^{13}C labeled samples. In heteronuclear experiments, the ability to selectively excite certain nuclei within a relatively narrow chemical shift range and to transfer magnetization between directly attached nuclei by taking advantage of the large scalar coupling existing between them, has allowed the design of pulse sequences that specifically correlate atoms along certain chemical pathways of the molecule. Moreover, resonance overlap is further reduced by spreading the peaks along a third (or fourth) ^{13}C or ^{15}N frequency dimension.

For labeled proteins (Clore and Gronenborn, 1991), strategies have been developed such that backbone atoms are assigned first using specific experiments, then other spectra are collected to assign nuclei in sidechains. In solving the structure of Tva, I used two 3D spectra for assigning backbone atoms: HNCACB (Wittekind and Mueller, 1993) and CBCA(CO)NNH (Grzesiek and Bax, 1992; Muhandiram and Kay, 1994). In the first experiment, the amide protons and nitrogens of the backbone are correlated with the α and β carbons of the same residue and of the previous residue (i-1). The CBCA(CO)NNH pulse sequence, on the other hand, includes a selective magnetization transfer through the backbone carbonyl group, such that amide H and N are only correlated to the Cα and Cβ of the previous residue (Figure 1.3a). Thus, by combining the intraresidue and interresidue connectivity information obtained by analyzing these two spectra, I was able to assign the resonances of these nuclei while "walking" along the backbone of the protein (Figure 1.3b).

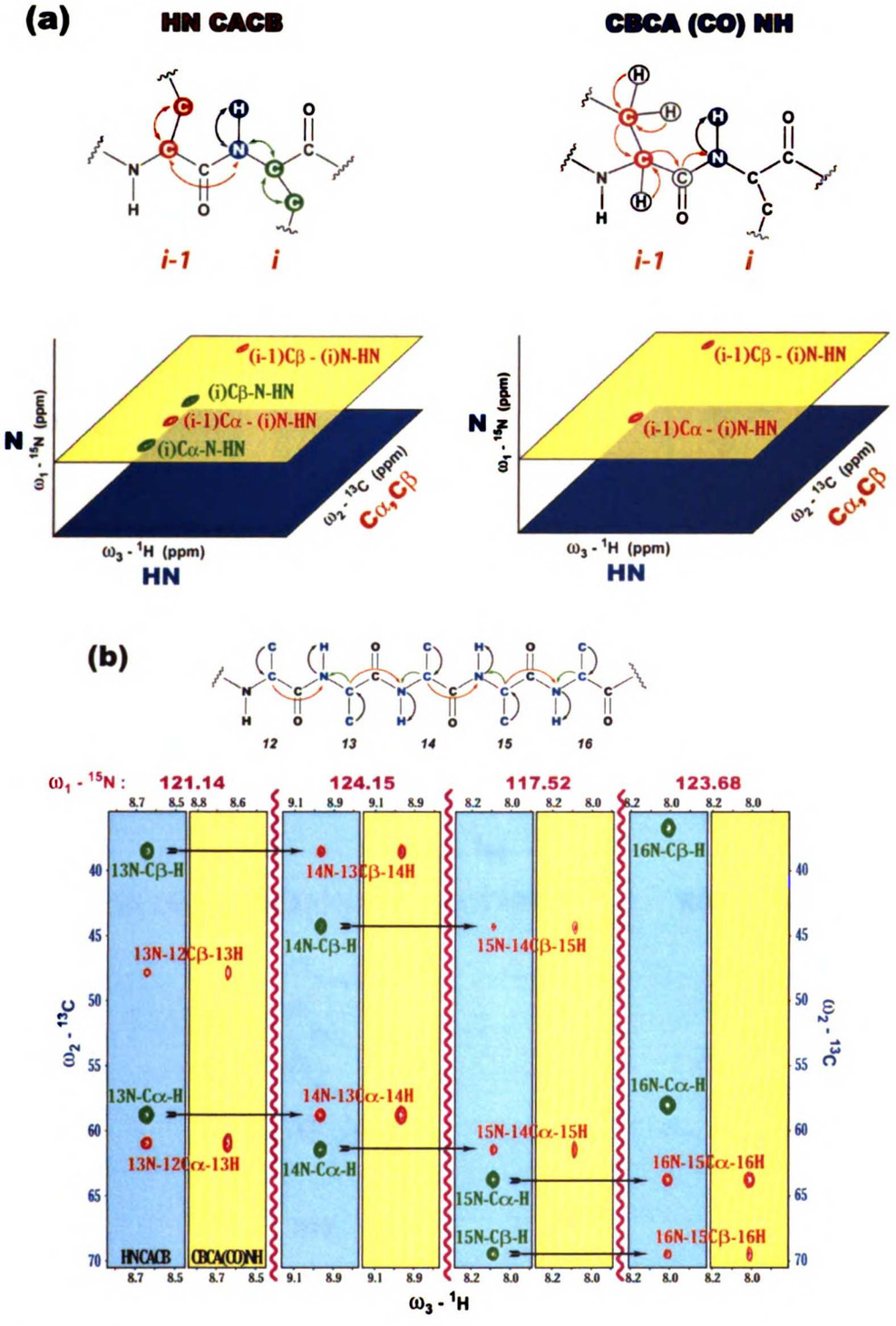


Figure 1.3. (a) Scheme showing the magnetization transfer pathway for the HNCACB and CBCA(CO)NNH experiments and an example of the intraresidue (green) and sequential (red) peaks we expect to find for each amide group in the protein. (b) Example of sequential walk from excerpts of HNCACB and CBCA(CO)NNH spectra acquired for sTva47.

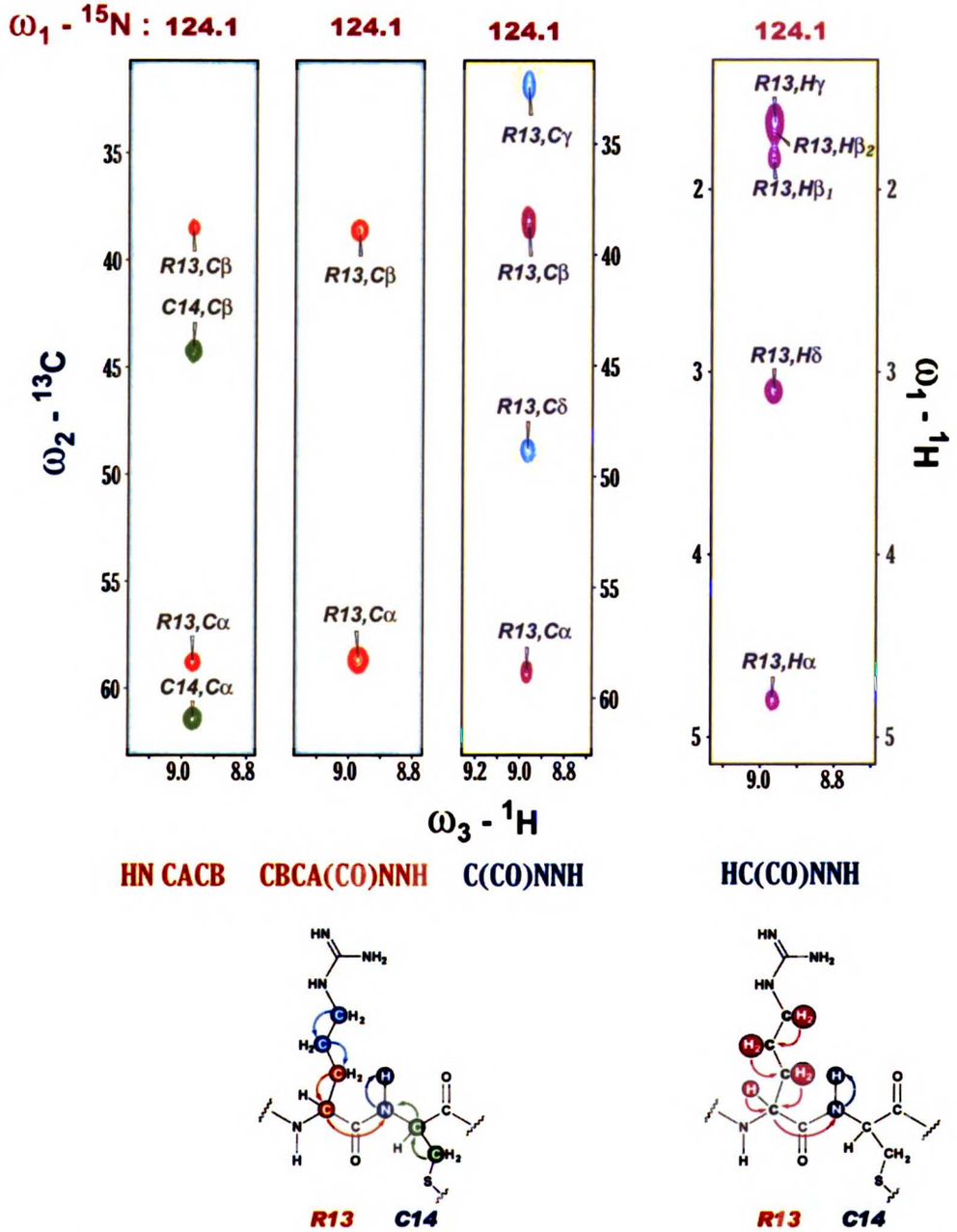


Figure 1.4. Excerpts of spectra acquired for ^{15}N,^{13}C-double-labeled sTva47 comparing peaks observed for the amide group of residue C14 in different 3D spectra: HNCACB, CBCA(CO)NNH, C(CO)NNH and HC(CO)NNH.

Once these backbone atoms have been assigned, other spectra are required to identify sidechain atoms and alpha nuclei. In my work on Tva, I started assigning sidechain

resonances in aliphatic residues by collecting two three-dimensional spectra using the C(CO)NNH and HC(CO)NNH pulse sequences (Grzesiek et al., 1993). In these experiments, the backbone amide protons and nitrogens are correlated with all the aliphatic sidechain ^{13}C of the previous residue in the C(CO)NNH spectrum and all aliphatic sidechain ^{1}H of the previous residue in the HC(CO)NNH one (Figure 1.4). Thus, if backbone amide ^{1}H and ^{15}N resonances are known, it is possible to identify all the aliphatic carbons and protons that belong to a particular residue. It is often not possible, however, to distinguish among all the protons and carbons within the same sidechain, especially for amino acids with longer sidechains.

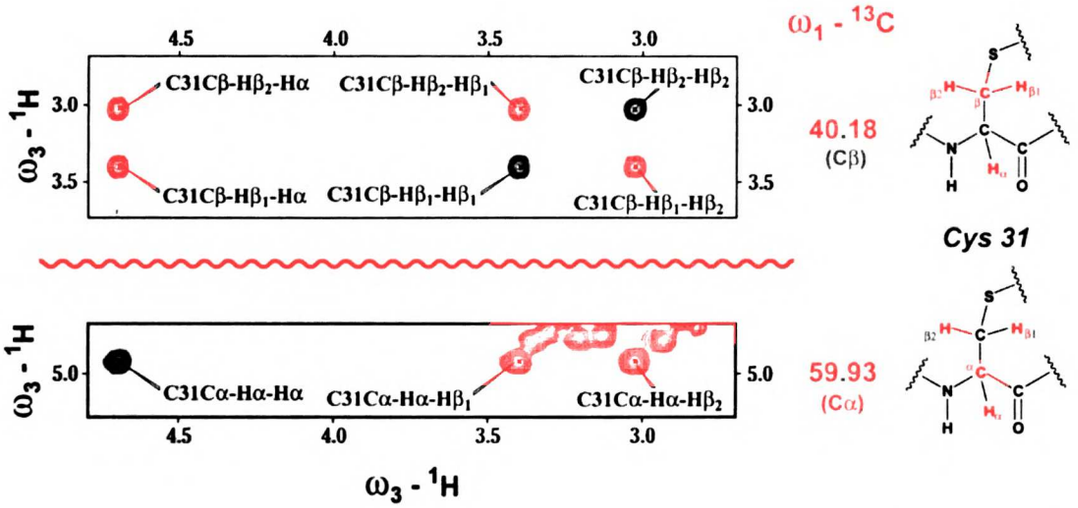


Figure 1.5. Excerpts from 3D HCCH-COSY spectrum of sTva47 showing the diagonal and cross-peaks observed for residue C14 on two different ^{13}C planes (Cα and Cβ).

The most common experiments used to distinguish between the various sidechain resonances are the three-dimensional spectra recorded using HCCH-TOCSY (Wang and Zuiderweg, 1995) and HCCH-COSY (Ikura et al., 1991) pulse sequences. These experiments allow dispersion of the 2D ^{1}H-^{1}H TOCSY and COSY spectra, respectively, into a third ^{13}C frequency dimension by utilizing three magnetization transfer steps: first from ^{1}H to its directly attached ^{13}C, then from the ^{13}C to neighboring ^{13}C nuclei and finally from ^{13}C back to the directly attached nuclei. The HCCH-COSY experiment has the advantage that only protons that are attached to adjacent carbons are correlated directly in the spectra, thus reducing overlap and simplifying identification of spins, but

is less sensitive than the TOCSY version and thus can't be acquired for all samples. My Tva sample was concentrated enough to allow me to collect a high quality three-dimensional HCCH-COSY spectrum, which I then used to identify all aliphatic carbon and proton resonances (Figure 1.5).

Finally, nuclei in aromatic sidechains can be assigned using specific pulse sequences and/or, once all other backbone and aliphatic sidechain resonances have been identified, by analyzing cross-peak connectivities in NOESY spectra. Thus, to identify aromatic protons in smaller proteins, two-dimensional NOESY spectra are recorded in D_2O samples; all the amide protons are exchanged to deuterium leaving only peaks from aromatic groups in the amide region. Alternatively, if a ^{13}C labeled sample is available, a three-dimensional ^{13}C HSQC-NOESY spectrum with the carbon frequency centered in the aromatic region can be used for assigning aromatic sidechains. In this spectrum, only NOE cross-peaks between protons attached to aromatic carbons and any other nearby proton are observed, thus, considerably reducing peak overlap and simplifying resonance identification. Furthermore, the peaks are also spread along a third aromatic ^{13}C frequency dimension. This strategy may be necessary for larger protein with crowded two-dimensional spectra.

1.3 Extracting structural information from NMR spectra

Resonance assignments are used to extract information about the structure of the macromolecule we are studying from the appropriate NMR experiments.

1.3.1 Interproton distances from NOE cross-peak intensities

Perhaps the most common and useful information that can be extracted from NMR spectra is interproton distances that are estimated from the cross-peak intensities in 2D NOESY spectra (Bull, 1987). During the mixing time period of the NOESY pulse sequence, the magnetization oriented along the Z-axis is transferred between nearby protons through dipole-dipole relaxation (Nuclear Overhauser effect) resulting in off-diagonal peaks in the final spectrum. The intensity of these NOESY cross-peaks is inversely proportional to the sixth-power of the distance between the protons that exchanged magnetization and can then be used to estimate such distance.

Depending on the mole
rected, mainly to help re
tification and integrati
eriment is used, the info
mension distances.

In smaller nucleic acids
spectra recorded with the
and H ^{1}e D_2O) is usually
ructure of the molecule. E
rotons that are attached to
ly exchanged with deu
om the spectra, leaving on
mplifying their identificatio
ntensities of H_2O in the
ually eliminating the large
rated in H_2O and requires

igure 1.6. 2D NOESY vs 3D
ecific overlap is greatly red
olving a third ^{15}N frequen

ote 1 -

Depending on the molecule under study, different types of NOESY spectra can be recorded, mainly to help reduce overlap in crowded systems, and to facilitate the identification and integration of NOE cross-peaks. However, no matter which type of experiment is used, the information extracted from the spectra remains the same: interproton distances.

In smaller nucleic acids and proteins, a combination of two-dimensional NOESY spectra, recorded with the sample dissolved in pure D_2O or in a mixture H_2O/ D_2O (with about 10 % D_2O) is usually sufficient to extract enough information to elucidate the structure of the molecule. By dissolving the sample in pure D_2O, all the exchangeable protons that are attached to nitrogen or oxygen and are exposed to the bulk solvent are rapidly exchanged with deuterium. Thus, the corresponding resonance will disappear from the spectra, leaving only cross-peaks from non-exchangeable protons and, thus, simplifying their identification and analysis. Furthermore, by drastically reducing the concentration of H_2O in the bulk solvent we increase the quality of the spectrum by virtually eliminating the large solvent peak that is otherwise observed in experiments recorded in H_2O and requires special pulsing techniques to suppress.

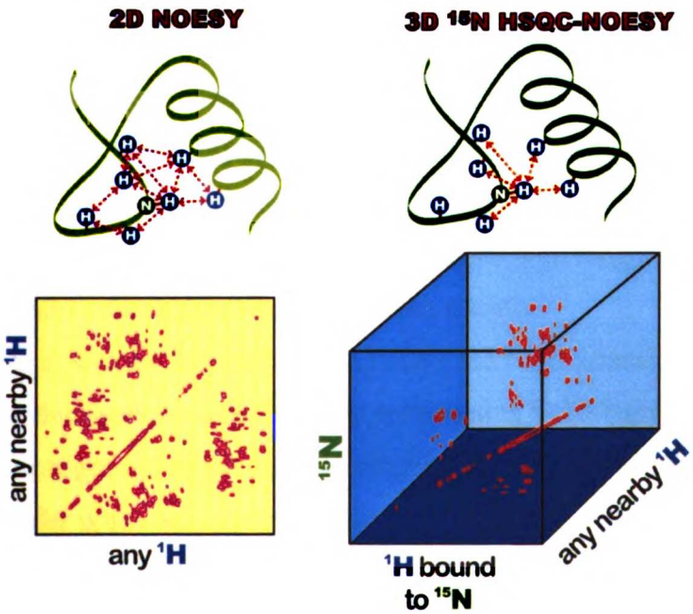


Figure 1.6. 2D NOESY *vs.* 3D ^{15}N HSQC-NOESY spectra. In the 3D spectrum the resonance overlap is greatly reduced because we observe only selected peaks that are also spread along a third ^{15}N frequency dimension.

In large molecules (Cl

NOESY spectra collected

labeled sample is available

include one or more selecti

^{15}N or ^{13}C prior and or afte

flows to select only for th

enriched to ^{15}N and or ^{13}C

are also spread along the a

There are different meth

distances (James, 1994a; Ja

approximation (Holak et al

from each other, so that ther

and interproton distances. Th

can be used to estimate unkn

proton pairs are not isolated,

magnetization during the mix

matrix phenomenon observe

al Berendsen, 1976; Olejnic

between two protons will be a

CSPA to analyze NOE data

will be more evident for weak

[9] and for NOE intensities

that longer times allow for m

On the other hand, we can

experiments run with shorter

using the initial linear portion

apply by using wider upper a

relation. In an extreme appr

Besides, NOE cross-peaks a

visual inspection of the peaks, a

which is commonly used for

In large molecules (Clore and Gronenborn, 1991), however, even two-dimensional NOESY spectra collected in D_2O can be too crowded to analyze. If an ^{15}N and/or ^{13}C labeled sample is available, three- or four-dimensional experiments can be recorded that include one or more selective magnetization transfers between directly attached ^{1}H and ^{15}N or ^{13}C prior and/or after the NOESY sequence. Depending on the pulse sequence, this allows to select only for those NOE cross-peaks in between nearby protons that are attached to ^{15}N and/or ^{13}C thus reducing overlap in the spectra. Furthermore, the peaks are also spread along the additional ^{15}N and/or ^{13}C frequency dimensions (Figure 1.6).

There are different methods of analyzing NOE cross-peaks to obtain interproton distances (James, 1994a; James, 1994b). The simpler one is called isolated spin pair approximation (Holak et al., 1987)(ISPA) and assumes that all proton pairs are isolated from each other, so that there is a straightforward relationship between NOE intensities and interproton distances. Thus, the NOE intensities that correspond to known distances can be used to estimate unknown distances from corresponding intensities. Obviously, proton pairs are not isolated, but rather they belong to an array of protons that exchange magnetization during the mixing time of NOESY experiments. This cross-polarization transfer phenomenon observed between multiple spins is known as spin-diffusion (Kalk and Berendsen, 1976; Olejniczak et al., 1986). Hence, the NOE intensity measured between two protons will be affected by the presence of other nearby protons and the use of ISPA to analyze NOE data will result in an error in the calculated distances. This error will be more evident for weaker intensities that correspond to longer distances (up to 80%) and for NOE intensities extracted from experiments with longer mixing times, since longer times allow for more cross-polarization transfer.

On the other hand, we can limit this error by using the NOE intensities measured from experiments run with shorter mixing times, by constructing NOE built up curves and using the initial linear portion of the curve to estimate the corresponding distances, or simply by using wider upper and lower bounds to enforce NOE distances during structure calculation. In an extreme approach to the problem of estimating distances from NOE intensities, NOE cross-peaks are classified into three or four groups, often only based on visual inspection of the peaks, and wide bounds are assigned to each group. While this approach is commonly used for protein structure calculation, it is usually not sufficient to

refine the structure of macromolecules with low proton density, like nucleic acids, at high resolution.

Thus, to take into account spin-diffusion and extract more accurate interproton distances, complete relaxation matrix methods have been developed (Borgias and James, 1989): a complete matrix of NOE intensities measured at a certain mixing time (τ_m) is built and subsequently diagonalized to generate a complete relaxation rate matrix according to equation:

$$\mathbf{R} = \frac{\ln\left[\frac{\mathbf{a}(\tau_m)}{\mathbf{a}(0)}\right]}{\tau_m}$$

where, **R** is the relaxation rate matrix, **a** is the intensity matrix and τ_m is the mixing time used to acquire the NOESY experiment. By diagonalizing this equation, we make sure to take into account of all relaxation pathways among the complete array of protons are taken into account. Individual distances are then calculated from the corresponding cross-relaxation rates. In order to build a complete intensity matrix, those intensities that could not be measured experimentally, because of overlap in the spectrum or limitations in the signal to noise ratio, are calculated from an initial model of the molecule. Surprisingly, the accuracy of the final distances is not very dependent on the initial model used, assuming that this model has good quality. MARDIGRAS (Borgias and James, 1990), a program developed in our laboratory that uses a complete relaxation matrix method, has been successfully used to estimate accurate interproton distances for elucidating the structure of nucleic acids and proteins. In fact, I used MARDIGRAS to calculate upper and lower bounds for my DNA duplex and the modified DNA-RNA hybrid duplexes.

While this approach has been widely used to extract distances from NOE intensities measured from two-dimensional spectra, intensities that are measured from heteronuclear three-dimensional NOESY experiments require an additional correction to take into account of the differences in magnetization transfer efficiency between protons and the attached heteronuclei for the different spins. In proteins however, the number of NOE cross-peak intensities extracted from three-dimensional spectra is usually so high that upper and lower bounds estimated using an ISPA-based approach are sufficient to refine

the structure of the molec

used the CALIBA routine

calculation (Guntert et al.

1.3.2 *Torsion angle info*

Torsion angles can be c

empirical relationship desc

are optimized for each sys

structure calculation metho

structure refinement progra

For smaller nucleic acid

corresponding homonuclear

analysis of cross-peak patte

solving the structure of my

homonuclear proton-proton

collected on a D_2O sample.

of the COSY cross-peaks (s

J couplings), J couplings ca

programs SPHINX and LIN

cross-peaks while using a v

line widths and digital reso

experimental cross-peak pa

fit.

For labeled proteins, exp

torsion angles of the molec

designed to manipulate mag

with specific torsion angle i

analysis of peaks in the resu

torsion angles can be estima

constraints in three-dimensio

the structure of the molecule at high resolution. Thus, in solving the structure of Tva, I used the CALIBA routine from DYANA to estimate upper bounds for structure calculation (Guntert et al., 1997).

1.3.2 Torsion angle information

Torsion angles can be estimated from the relevant J coupling constants by using the empirical relationship described by Karplus equations (Karplus, 1959). These equations are optimized for each system and are generally incorporated in the most common structure calculation methods, so that J coupling constants can be input directly into the structure refinement program.

For smaller nucleic acids and proteins, torsion angles can be estimated from the corresponding homonuclear proton-proton coupling constants that are extracted from the analysis of cross-peak patterns in two-dimensional COSY spectra (Bax et al., 1994). In solving the structure of my DNA undecamer, I estimated the sugar torsion angles from homonuclear proton-proton coupling constants extracted from DQF-COSY spectra collected on a D_2O sample. Since there are a number of factors that can affect the pattern of the COSY cross-peaks (spectral processing parameters, inherent proton line widths and J couplings), J couplings can not be extracted by direct measurement. Hence, I used the programs SPHINX and LINSHA (Widmer and Wüthrich, 1987) to simulate DQF-COSY cross-peaks while using a variety of J coupling values and taking into account the real line widths and digital resolution of the spectrum. Then, by matching simulated and experimental cross-peak patterns, I was able to estimate J coupling values for my DNA duplex.

For labeled proteins, experiments have been developed to allow extracting only certain torsion angles of the molecule (Bax et al., 1994). Again pulse sequences have been designed to manipulate magnetization along certain chemical pathways while labeling it with specific torsion angle information. These angles can then be estimated from the analysis of peaks in the resulting multidimensional spectra. For example, Φ backbone torsion angles can be estimated from comparison of the diagonal and cross-peak intensities in three-dimensional HNHA experiments (Vuister and Bax, 1993) (Figure 1.7).

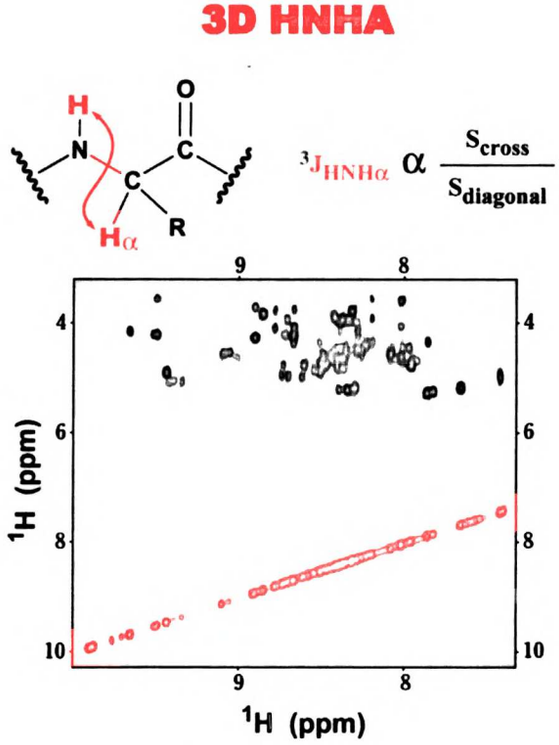


Figure 1.7. 2D projection of a 3D HNHA spectrum acquired on sTva47. **Scross**, intensity of off-diagonal peaks; **Sdiagonal**, intensity of diagonal peaks. The ratio between these intensities is proportional to the $J_{HNH\alpha}$ coupling constant and can used to estimate the ϕ torsion angle.

1.3.3 Residual dipolar couplings

Recently, strategies have been developed for extracting and utilizing the information given by residual dipolar couplings that are measured in samples with a very low degree of molecular alignment to the external magnetic field. This minimal alignment can be observed when molecules with non-zero magnetic susceptibility anisotropy are placed in strong magnetic fields. In most proteins and nucleic acids, however, the magnetic susceptibility anisotropy is too low to detect any degree of alignment even at high magnetic fields. Hence, to align the solute we need to render the solvent anisotropic. This is typically done by dissolving into the system large anisotropic particles like bicelles (Hansen et al., 1998; Tjandra and Bax, 1997) or phages (Hansen et al., 1998) that orient themselves in strong magnetic fields (liquid crystalline phase). The steric interaction

between such large oriented particles and our molecule of interest will result in the minimal alignment that we are looking for.

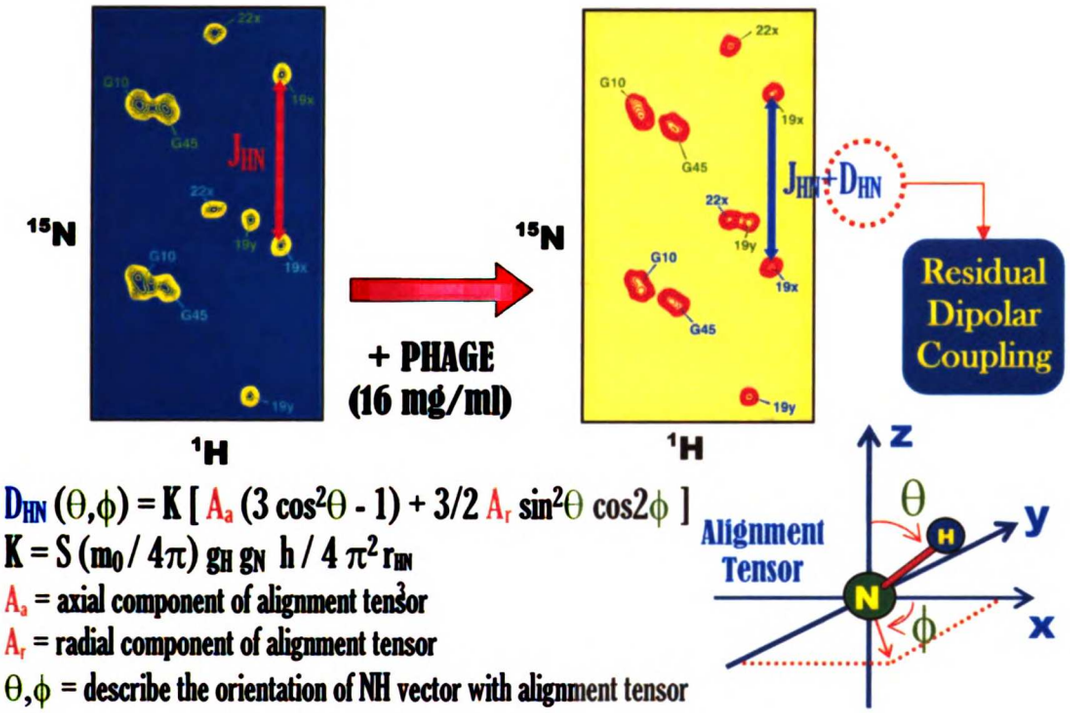


Figure 1.8. A section of two undecoupled ^{15}N HSQC spectra acquired on ^{15}N-labeled sTva47 before and after adding phage to the sample. The difference in peak splitting gives the RDC for a given HN group. The equation shows the relationship between RDC, the alignment tensor of the molecule in solution and the orientation of the HN vector relative to the alignment tensor of the molecule and, hence, the external magnetic field.

Typically, residual dipolar coupling values are extracted for certain bond vectors (like ^{1}H-^{15}N bond vectors in proteins) by measuring differences in peak splitting in non-decoupled spectra collected before and after adding the liquid crystalline medium to the sample (Figure 1.8) (Lerche et al., 1999). Residual dipolar couplings can then be used in structure calculation to optimize the orientation of these bond vectors with respect to the orientation of the external alignment or susceptibility tensor of the molecule (Tjandra and Bax, 1997). Given that all the bond vectors are oriented relative to the same reference axis system, residual dipolar couplings are fundamentally different from the strictly local NOE and J coupling restraints in that they can relate distant parts of the molecule. Thus, even the use of few RDCs can be very useful to improve the structure of a protein or nucleic acid. For example, RDCs can help in defining the relative orientation of separate

protein domains that can not otherwise be calculated very accurately by using only NOE and J coupling derived restraints. Similarly, RDCs have proven particularly useful in improving the quality of nucleic acid structures by correlating the orientation of base-pairs at both ends of duplex sections and/or by defining the relative orientation of duplex sections flanking bulge regions.

All the restraints derived from experiments are affected by error. This error can be introduced while extracting the data from NMR spectra or, later on, in the process of calculating the structural information from the experimental data. Thus, peak overlap, strong signals from solvent, low signal-to-noise and, in general, poor quality of NMR spectra, can lead to mistakes in the measurement of peak volumes or even in defining the exact peak position which will result in inaccurate NOE intensities, J coupling constants and RDCs. Then, error is also introduced in the process of calculating the structural parameters from the experimental data. For example, we may introduce error in NOE derived distances by neglecting spin diffusion, using the wrong correlation time to describe the overall tumbling of the molecule, etc. This uncertainty is taken into account during structure refinement by enforcing the experimental restraints as upper and lower bounds. A penalty will be imposed on the system only when the real value violates the upper or lower bounds. The width of the bounds, then, will have to be calculated to reflect the error in the NMR-derived data, with bigger errors resulting in wider bounds.

Finally, the quality of structure refined from NMR data will depend on the number, the accuracy and the distribution of the experimental restraints. Obviously, more restraints, more accurate and evenly distributed restraints will result in better defined structures.

1.4 Structure calculation

The information extracted from NMR spectra is then used to refine the structure of the macromolecule. Typically, for this purpose we use computer programs that are capable of generating structures by searching conformational space, while an empirical function is used to calculate the molecular energy (force field) or target function. To enforce the NMR information, the force field is modified to include terms that describe the

experimental data and impose a penalty on the system when the experimental restraints are violated. Thus, starting from an initial model of the molecule, that can also be randomly generated, an algorithm searches conformational space following a user-controlled protocol to generate structures with low target function or energy that are in agreement with the NMR restraints.

The most common of such methods implemented for NMR structure refinement are distance geometry (DG) and restrained molecular dynamic simulations (rMD) (Karplus and Petsko, 1990). In distance geometry, a variable target function algorithm is used to calculate macromolecular structures on the basis of NMR data, *e.g.*, using the program DIANA (Güntert et al., 1991). Instead, rMD searches conformational space by simulating molecular motion at a certain temperature by applying Newton's law. The forces acting on each atom are calculated from an empirical force field. While distance geometry has the ability of rapidly generating structures that satisfy the experimental restraints, simulation of molecular dynamics are time-consuming especially for large macromolecules, since Newton's law has to be applied on every atom of the system. Thus, DG is typically used to generate conformations when the fold of the macromolecule we are studying is not known. However, structures generated by DG usually have very poor energetics and need to be further refined by more sophisticated methods, like rMD, that make use of empirical force fields to describe the molecular energy. On the other hand, when a reasonable initial model is available, rMD simulations can be used directly to refine its structure, typically using an annealing protocol. In an annealing protocol, the temperature of the system is initially raised to overcome energy barriers, and then gradually cooled to lower the molecule into an energy minimum. A typical simulated annealing protocol is shown in figure 1.9.

Recently, novel rMD methods have been developed that are capable of rapidly generating structures by performing molecular dynamics in torsion angle space (Guntert et al., 1997). Using torsion angles to describe the structure of a molecule has the advantage of reducing the number of degrees of freedom and speeding up the calculations. Thus, these methods can replace DG to generate structures starting from random conformations when an initial model of the molecule is not available.

To refine the structure of the DNA duplex and phosphorothio hybrid duplexes, I used rMD since the global fold of the molecules is already known (right handed duplexes). Furthermore, to validate the structure of the DNA duplex, I independently refined it using miniCarlo, a locally developed program that uses helical parameters to describe nucleic acid duplexes and restrained MonteCarlo simulations to refine their structures. The final molecules generated by rMD and miniCarlo converged to virtually the same conformation, suggesting that the structure is defined by the NMR restraints and not the method used to refine it.

On the other hand, the structures of the phosphorothio hybrid duplexes were calculated using MD simulations with time-averaged restraints (MDtar). MDtar simulations are a particular type of rMD that can be used in flexible systems if we suspect that the NOE derived interproton distances carry inconsistencies due to conformational flexibility. In MDtar, the NOE restraints are enforced on the structures generated over a period of time, rather than at every step like for regular rMD. This allows the molecule to explore a wider range of conformational space and to simultaneously use an ensemble of structures to better satisfy NOE restraints affected by molecular flexibility. An ideal application for MDtar is to study sugar repuckering in nucleic acids.

Finally, to refine the structure of Tva, I used DYANA, a program that performs torsion angle dynamics (TAD) with a simulated annealing protocol, to generate initial conformations of the molecule starting from structures with random geometry. DYANA uses a simplified target function that only takes into account steric interactions and experimental restraints. The structures generated by DYANA with the lowest target function were further refined by rMD.

1.5 Assessment of the Quality of NMR Refined Structures

The main points that need to be analyzed to assess the quality of structures refined by NMR are accuracy, precision and overall quality (James, 1994a).

Precision or how well the structure is defined by the NMR restraints. The convergence of the structure to a common conformation when starting from different initial models (that can also be randomly generated structures) or by independently calculating it with

different refinement methods, gives an indication of how well the structure is defined by the NMR restraints. The RMSD values calculated among final ensembles of structures are used as a measure of the precision of the structure.

Accuracy or how well the refined structures represent the NMR data. This can be assessed by back-calculation of the NMR data from the final geometry and comparison with the original data extracted from experiments. For example, theoretical NOESY spectra can be simulated from refined models (e.g. using CORMA) and then compared to experimental ones (Keepers and James, 1984). The R factors that are calculated give an indication of the accuracy of the structure. Similarly, analysis of NMR restraints violation by the refined structures is also important, with more accurate structures resulting in less violation of experimental data.

Overall quality of refined structures. Finally, the energy of the refined molecule and the analysis of the parameters that are commonly used to describe its conformation (e.g. Ramachandran plots in proteins or helical parameters in nucleic acids) provide an indication of its quality.

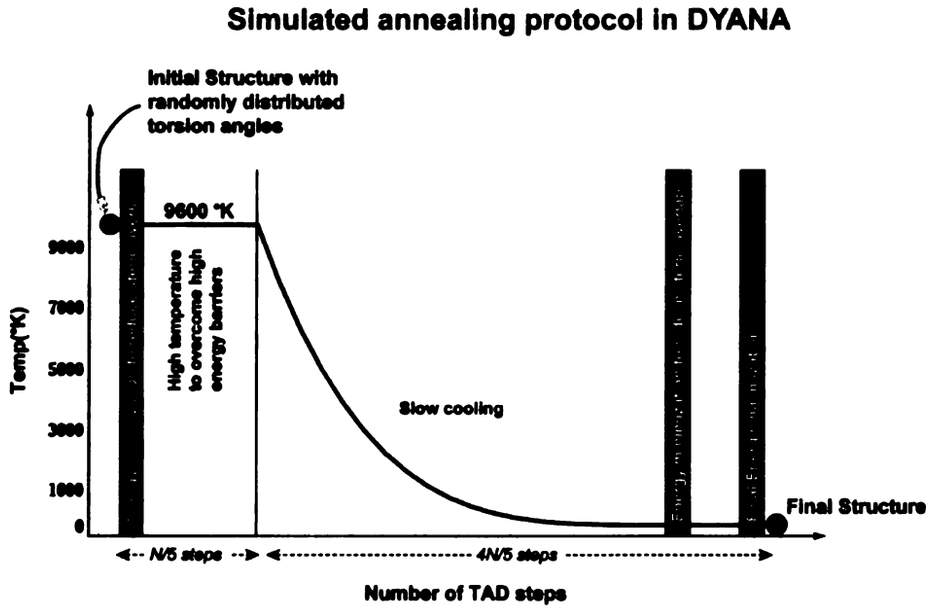


Figure 1.9. Typical simulated annealing protocol used by default in DYANA. Starting at very high temperature (9600°K) with a random structure, the system is gradually cooled to near 0°K to trap low energy conformations that are in agreement with the experimental restraints. TAD, torsion angle dynamics performed by DYANA to search conformational space. N, user-input number of TAD steps performed by DYANA to produce a final structure.

1.6 Studying structural flexibility by NMR

NMR allows studying molecules in solution and provides the means for investigating their flexibility (Cavanagh and Venters, 2001; Ishima and Torchia, 2000; Kay, 1998; Palmer et al., 2001). Furthermore, NMR has the unique capacity to investigate dynamic properties of molecules over a range of time scales with atomic resolution. Different kinds of motions will occur at different time-scales, affecting the spectral appearance and requiring specific strategies to be investigated.

Fast backbone and side chain fluctuations that occur in the sub-nanosecond range and are, thus, faster than the overall tumbling rate of the molecule, will not have much of an effect on spectral appearance. Still these motions can be studied in labeled molecules using ^{15}N and/or ^{13}C relaxation techniques which monitor the reorientation of ^{15}N-^{1}H and/or ^{13}C-^{1}H vectors by model-free analysis (Lipari and Szabo, 1982a; Lipari and Szabo, 1982b). This analysis is named model-free because it derives parameters without the need to invoke a specific model for internal motion. The parameters it provides are the order parameter (S^2) that gives a measure of the amplitude of the internal motion, the effective rate constant for internal dynamics (τ_e) as well as the overall molecular tumbling time of the molecule (τ_c). This analysis is useful to get an indication of the amplitude of structural variations or configurational entropy of the polypeptide chain.

If the dynamic process involves equilibrium between two or more conformers, the interconversion rate and population will dictate the appearance of the NMR spectra. If the exchange rate is faster than the chemical shift difference between the conformers, only one set of peaks is found in the spectra. This is also true for slightly slower exchange rates when one conformer has a dominant population compared to the others. Yet, if the motion involves a big change in chemical shift, the observed NMR peaks will be exchange-broadened due to an increase in the transverse relaxation rate, R_2. The contribution to R_2 from chemical exchange can then be identified by measuring R_2 as a function of the effective radiofrequency field strength using spin-lock or CPMG experiments (Akke and Palmer III, 1996; Loria et al., 1999). Thus, these experiments can be used to identify conformational exchange in the μs-to-ms time-scale range and can

provide the rate of conformational exchange, population and even entropy and enthalpy information about the system.

If the exchange rate between the conformers is slower than the chemical shift difference between them, multiple peaks will be observed in NMR spectra. Such motions within the ms-to-10s range can be monitored by measuring rates of magnetization transfer among protons with different chemical shifts (ZZ exchange experiments) (Wagner et al., 1985). Even slower motions can still be investigated with hydrogen/deuterium exchange experiments, *e.g.*, studying protein folding intermediates and pathways by hydrogen exchange (Englander, 2000).

Finally, we need to take into account that, even when a single set of peaks is observed in NMR spectra, the presence of dynamic processes will obviously have an effect on the NMR restraints and, hence, on the refined structure. Thus, the interproton distances derived from NOE intensities are typically averaged by motion and represent a weighted average of the distances sampled by a flexible molecule (LeMaster et al., 1988). Moreover, because of the inverse sixth-root relationship existing between NOE intensities and distances, this average value will be biased towards the conformer with the shortest distance, often resulting in inconsistencies between different classes of interproton distances.

However, we can turn this complication to our advantage by analyzing these inconsistencies to gain some insights into the dynamic processes that cause them. For example, while solving the structure of my DNA duplex, we concluded that a two-state equilibrium of the sugar moieties between the N and S puckers, combined with a rocking movement of the base about the χ torsion angle centered on the average value for the anti conformation, could explain the inconsistencies that we had observed between different classes of base-sugar intraresidue distances (see chapter 3). Sugar repuckering of the sugar rings must also be used to explain the homonuclear J couplings extracted from analysis of DQF-COSY cross-peak patterns (Gonzàlez et al., 1994; Schmitz et al., 1990). Finally, sugar repuckering can also be observed during MD simulations with time-averaged restraints, further suggesting that the NOE distances are biased by molecular flexibility (Schmitz et al., 1993; Torda et al., 1991).

On the other hand, the flexibility resulting from *cis-trans* isomerization shown by Tva is very slow, yielding NMR spectra with multiple sets of peaks. Furthermore, the rate of interconversion is so slow that no magnetization transfer can be identified between the conformers and thus each one can be treated as an independent entity in solution and its structure solved. In this case, NMR doesn't offer a time scale for the exchange process (except for lower limit), but it provides the means for solving the structure of each conformer at high resolution.

1.7 REFERENCES

Akke, M., and Palmer III, A. G. (1996). Monitoring Macromolecular Motions on Microsecond to Millisecond Time Scales by $R_{1\rho}$-R_1 Constant Relaxation Time NMR Spectroscopy. J Am Chem Soc *118*, 911-912.

Bax, A., Vuister, G. W., Grzesiek, S., Delaglio, F., Wang, A. C., Tschudin, R., and Zhu, G. (1994). Measurement of homo- and heteronuclear J couplings from quantitative J correlation. Methods Enzymol *239*, 79-105.

Borgias, B. A., and James, T. L. (1989). Two-Dimensional Nuclear Overhauser Effect: Complete Relaxation Matrix Analysis. In Methods in Enzymology, Nuclear Magnetic Resonance, Part A: Spectral Techniques and Dynamics, N. J. Oppenheimer, and T. L. James, eds. (New York, Academic Press), pp. 169-183.

Borgias, B. A., and James, T. L. (1990). MARDIGRAS--Procedure for Matrix Analysis of Relaxation for DIscerning Geometry of an Aqueous Structure. J Magn Reson *87*, 475-487.

Bull, T. E. (1987). Cross-Correlation and 2D NOE Spectra. J Magn Reson *72*, 397-413.

Cavanagh, J., and Venters, R. A. (2001). Protein dynamic studies move to a new time slot. Nat Struct Biol *8*, 912-914.

Clore, G. M., and Gronenborn, A. M. (1991). Structures of larger proteins in solution: three- and four-dimensional heteronuclear NMR spectroscopy. Science *252*, 1390-1399.

Creighton, T. E. (1993). Proteins : structures and molecular properties, 2nd edn (New York, W.H. Freeman).

Englander, S. W. (2000). Protein folding intermediates and pathways studied by hydrogen exchange. Annu Rev Biophys Biomol Struct *29*, 213-238.

Gonzàlez, C., Stec, W., Kobylanska, A., Hogrefe, R., Reynolds, M., and James, T. L. (1994). Structural Study of a DNA-RNA Hybrid Duplex with a Chiral Phosphorothioate Moiety by NMR: Extraction of Distance and Torsion Angle Constraints and Imino Proton Exchange Rates. Biochemistry *33*, 11062-11072.

Grzesiek, S., Anglister, J., and Bax, A. (1993). Correlation of Backbone Amide and Aliphatic Side-Chain Resonances in 13C/15N-Enriched Proteins by Isotropic Mixing of 13C Magnetization. Journal of Magnetic Resonance *B101*, 114-119.

Grzesiek, S., and Bax, A. (1992). Correlating backbone amide and side chain resonances in larger proteins by multiple relayed triple resonance NMR. Journal of the American Chemical Society *114*, 6291-6293.

Güntert, P., Braun, W., and Wüthrich, K. (1991). Efficient Computation of Three-dimensional Protein Structures in Solution from Nuclear Magnetic Resonance Data Using the Program DIANA and the Supporting Programs CALIBA, HABAS and GLOMSA. J Mol Biol *217*, 517-530.

Guntert, P., Mumenthaler, C., and Wuthrich, K. (1997). Torsion angle dynamics for NMR structure calculation with the new program DYANA. J Mol Biol *273*, 283-298.

Hansen, M. R., Mueller, L., and Pardi, A. (1998). Tunable alignment of macromolecules by filamentous phage yields dipolar coupling interactions. Nat Struct Biol *5*, 1065-1074.

Holak, T. A., Scarsdale, J. N., and Prestegard, J. H. (1987). A Simple Method for Quantitative Evaluation of Cross-Peak Intensities in Two-Dimensional NOE Spectra. J Magn Reson *74*, 546-549.

Ikura, M., Kay, L. E., and Bax, A. (1991). Improved three-dimensional 1H-13C-1H correlation spectroscopy of a 13C-labeled protein using constant-time evolution. J Biomol NMR *1*, 299-304.

Ishima, R., and Torchia, D. A. (2000). Protein dynamics from NMR. Nat Struct Biol *7*, 740-743.

James, T. L. (1994a). Assessment of the Quality of Derived Macromolecular Structures. Meth Enzymol *239*, 416-439.

James, T. L. (1994b). Computational Strategies Pertinent to NMR Solution Structure Determination. Curr Opin Struct Biol *4*, in press.

Kalk, A., and Berendsen, H. J. C. (1976). Proton Magnetic Relaxation and Spin Diffusion in Proteins. J Magn Reson *24*, 343-366.

Karplus, M. (1959). J Chem Phys *30*, 11-31.

Karplus, M., and Petsko, G. A. (1990). Molecular dynamics simulations in biology. Nature *347*, 631-639.

Kay, L. E. (1998). Protein dynamics from NMR. Nat Struct Biol *5 Suppl*, 513-517.

Keepers, J. W., and James, T. L. (1984). A Theoretical Study of Distance Determinations from NMR. Two-Dimensional Nuclear Overhauser Effect Spectra. J Magn Reson *57*, 404-426.

LeMaster, D. M., Kay, L. E., Brünger, A. T., and Prestegard, J. H. (1988). Protein dynamics and distance determination by NOE measurements. FEBS Lett *236*, 71-76.

Lerche, M. H., Meissner, A., Poulsen, F. M., and Sorensen, O. W. (1999). Pulse sequences for measurement of one-bond (15)N-(1)H coupling constants in the protein backbone. J Magn Reson *140*, 259-263.

Lipari, G., and Szabo, A. (1982a). Model-Free Approach to the Interpretation of Nuclear Magnetic Resonance Relaxation in Macromolecules. 1.Theory and Range of Validity. J Am Chem Soc *104*, 4546-4559.

Lipari, G., and Szabo, A. (1982b). Model-Free Approach to the Interpretation of Nuclear Magnetic Resonance Relaxation in Macromolecules. 2. Analysis of Experimental Results. J Am Chem Soc *104*, 4559-4570.

Loria, J. P., Rance, M., and Palmer, A. G. (1999). A relaxation-compensated Carr-Purcell-Meiboom-Gill sequence for characterizing chemical exchange by NMR spectroscopy. J Amer Chem Soc *121*, 2331-2332.

Muhandiram, D. R., and Kay, L. E. (1994). Gradient-Enhanced Triple-Resonance Three-Dimensional NMR Experiments with Improved Sensitivity. Journal of Magnetic Resonance *B103*, 203-216.

Olejniczak, E. T., Gampe\0Jr, R. T., and Fesik, S. W. (1986). Accounting for Spin Diffusion in the Analysis of 2D NOE Data. J Magn Reson *67*, 28-41.

Palmer, A. G., 3rd, Kroenke, C. D., and Loria, J. P. (2001). Nuclear magnetic resonance methods for quantifying microsecond-to-millisecond motions in biological macromolecules. Methods Enzymol *339*, 204-238.

Schmitz, U., Ulyanov, N. B., Kumar, A., and James, T. L. (1993). Molecular Dynamics with Weighted Time-averaged Restraints of a DNA Octamer: Dynamic Interpretation of NMR Data. J Mol Biol *234*, 373-389.

Schmitz, U., Zon, G., and James, T. L. (1990). Deoxyribose Conformation In $[d(GTATATAC)]_2$: Evaluation of Sugar Pucker by Simulation of Double-Quantum-Filtered COSY Cross-Peaks. Biochemistry *29*, 2357-2368.

Tjandra, N., and Bax, A. (1997). Direct Measurement of Distances and Angles in Biomolecules by NMR in a Dilute Liquid Crystalline Medium. Science *278*, 1111-1114.

Torda, A. E., Scheek, R. M., and van Gunsteren, W. F. (1991). Time Averaged Distance Restraints in NMR Based Structural Refinement. In Computational Aspects of the Study of Biological Macromolecules by Nuclear Magnetic Resonance Spectroscopy, J. C. H. e. al., ed. (New York, Plenum Press), pp. 219-225.

Vuister, G. W., and Bax, A. (1993). Quantitative J correlation: a new approach for measuring homonuclear three-bond J(HNH.alpha.) coupling constants in 15N-enriched proteins. Journal of the American Chemical Society *115*, 7772-7777.

Wagner, G., Bodenhausen, G., Mueller, N., Rance, M., Soerensen, O. W., Ernst, R. R., and Wuthrich, K. (1985). Exchange of two-spin order in nuclear magnetic resonance:

separation of exchange and cross-relaxation processes. Journal of the American Chemical Society *107*, 6440-6446.

Wang, H., and Zuiderweg, E. R. (1995). HCCH-TOCSY spectroscopy of 13C-labeled proteins in H2O using heteronuclear cross-polarization and pulsed-field gradients. J Biomol NMR *5*, 207-211.

Widmer, H., and Wüthrich, K. (1987). Simulated Two-Dimensional NMR Cross-Peak Fine Structures for ^{1}H Spin Systems in Polypeptides and Polydeoxynucleotides. J Magn Reson *74*, 316-336.

Wittekind, M., and Mueller, L. (1993). HNCACB, a High-Sensitivity 3D NMR Experiment to Correlate Amide-Proton and Nitrogen Resonances with the Alpha- and Beta-Carbon Resonances in Proteins. Journal of Magnetic Resonance *B101*, 201-205.

Wüthrich, K. (1986). NMR of Proteins and Nucleic Acids (New York, Wiley).

CHAPTER 2

2. NMR Structure of d(GCATATGATAG)·d(CTATCATATGC): A Consensus Sequence for Promoters Recognized by σ^K RNA Polymerase

The three-dimensional structure of d(GCATATGATAG)·d(CTATCATATGC), from the promoter region of a gene regulating sporulation in *Bacillus subtilis* mother cells, was determined utilizing two-dimensional nuclear Overhauser effect (2D NOE) and double quantum-filtered COSY (2QF-COSY) spectra. To minimize the effect of methods used to obtain restraints and refine structure, several variables were studied. Interproton distance bounds were calculated very conservatively by running the complete relaxation matrix program MARDIGRAS hundreds of times using 2D NOE spectra for exchangeable and for nonexchangeable protons at different mixing times, assuming different overall correlation times and different starting structures. The 435 distance restraints were used with two different structural refinement methods: restrained molecular dynamics (rMD) and restrained Monte Carlo calculations (rMC). Refinement using different procedures and starting structures resulted in essentially the same structure (<0.8 Å RMSD), indicating structure is defined by experimental restraints and not refinement method or variables used. R factors indicate the structures fit the experimental NOE data very well. Some helical parameters, notably large negative X-displacement, are characteristic of A-DNA but others are characteristic of B-DNA. As with TG/CA steps in other duplex DNA sequences studied in our laboratory, the two TG/CA steps have a positive roll, with T6-

---

Abbreviations. 1D NMR, one-dimensional NMR; 2D NMR, two-dimensional NMR; 2D NOE, two-dimensional nuclear Overhauser effect; 2QF-COSY, double-quantum-filtered correlation spectroscopy; TOCSY, total correlated spectroscopy; EDTA, ethylenediaminetetraacetate; FID, free induction decay; rMD, restrained molecular dynamics; rMC, restrained Monte Carlo calculations; MD-tar, molecular dynamics with time-averaged restraints; RMSD, root-mean-square deviation; R^x, sixth-root R-factor; P, pseudorotation phase angle.

G7 exhibiting the largest, and consequently a bent helix axis. The converged structure represents a time-averaged structure. However, multiple conformations, especially in deoxyriboses, were evident from vicinal coupling constants obtained from quantitative simulations of 2QF-COSY cross-peaks and from persistent inconsistencies in experimental distances due to nonlinear conformational averaging.

2.1 INTRODUCTION

Many important biological or pharmacological processes are effected by specific interactions of DNA with other molecules, e.g., proteins, RNA, mutagens or drugs. It has been observed that such interactions can be significantly affected by changes in the sequence of DNA. Indeed, effects are manifest even when mutated residues are not directly involved in contacting the protein, but the mutation induces conformational distortions in the DNA molecule (Koudelka et al., 1987). In particular, the three-dimensional arrangement of hydrogen-bonding sites, bases and helix geometry are structural features of DNA molecules specifically recognized by proteins. For example, specific promoter sequences exhibit 1000-10000 times stronger affinity for transcription factors than do random sequences (von Hippel et al., 1984). In duplex DNA, major and minor grooves provide the topology and there are sequence-dependent variations in the patterns of hydrogen bond acceptors and donors. There is little distinction in the hydrogen bond pattern in the minor groove, but there is significant variation in the major groove. The major groove also exhibits nonpolar patches in the C5 region of pyrimidines with an obvious difference even between C and T. The strength and kinetics of specific binding of a ligand to a double-stranded DNA depends on the ligand matching these and other contacts in a "direct readout" of sequence (Harrison and Aggarwal, 1990). However, the sequence-dependent conformation of DNA modifies the three-dimensional pattern of contacts to be made as an "indirect readout" of sequence. These features range from relatively gross changes in bending, to effects of groove width, and finally to local variations in disposition of contact points. Knowledge about sequence-dependent variations in DNA duplex geometry is essential to understand fully DNA function in biological processes. There may be some thermodynamic (and possibly kinetic)

advantages for a particular DNA free in solution to be in a conformation similar to that it will possess when bound. With binding, there is undoubtedly a favorable entropy change upon displacement of counterions, but some of that entropic advantage could be lost by the gain in entropy of the DNA itself upon complexation. If the free DNA structure closely resembles that of the bound state, there would be little entropic price to pay upon binding. In addition, conformational flexibility of the double helix at specific sites may play an important role in recognition by proteins, perhaps by providing loci with a propensity for conformational change which is then manifest upon protein binding.

While the overall folding geometry of proteins and RNA is often the goal of structure studies, when studying DNA duplexes the question to address is how the conformation varies with sequence. Since these sequence-dependent structural variations are very subtle, the structure must be determined with high resolution and high accuracy. A better understanding of the rules governing sequence-dependent structural heterogeneity in DNA could some day contribute to "rational design" of new drugs targeted to interact directly with a gene (DNA) rather than a gene product (protein).

Recently the quality of DNA structures determined via NMR has improved sufficiently to provide an alternative. Methodology for determination of structures with sufficient resolution has not been easily achieved, but the ability to determine an accurate, high-precision structure of nearly any DNA duplex of length $\leq$15 base pairs is now possible with sufficient care and effort (Schmitz and James, 1995). Especially useful NMR experiments are two-dimensional nuclear Overhauser enhancement (2D NOE) spectra and, to a lesser extent, double-quantum filtered correlated (2QF-COSY) spectra which are used for extracting proton-proton distances and sugar torsion angles, respectively. These NMR-derived parameters are subsequently used as structural restraints in refinement methods such as restrained molecular dynamics (rMD) calculations, which search conformational space to find structures consistent with the experimental data. Other refinement methods are also available: in particular, distance geometry algorithms are commonly used for protein structural refinement, and restrained Monte Carlo (rMC) calculations have occasionally been used.

Improvements in structure determination have also been achieved by developing methods to increase the accuracy and the number of structural restraints, such as,

extraction of interproton distances from 2D NOE intensities via a "complete relaxation matrix approach" (James, 1991), and sugar torsional angles determination from proton-proton coupling constants by applying optimized Karplus equations (Celda et al., 1989; Gonzàlez et al., 1994; Schmitz et al., 1990). MARDIGRAS (Matrix Analysis of Relaxation for DIscerning GeometRy of Aqueous Structures), is the algorithm we use to calculate reliable distance bounds from NOESY spectra (Liu et al., 1995). On the other hand, accurate coupling constants can be extracted by simulation of COSY cross-peak patterns with the programs SPHINX and LINSHA (Gonzàlez et al., 1994; Widmer and Wüthrich, 1986).

The DNA sequence d(GCATATGATAG)·d(CTATCATATGC) studied in this paper is found in the promoter region of the gene *gerE* which is expressed in *Bacillus subtilis* mother cell during sporulation (Zheng et al., 1992). Gene expression in the *B. subtilis* is a hierarchical cascade of at least four temporally controlled gene sets. The first three each contain a regulatory gene that governs expression of the next gene set in the pathway. The σ^K RNA polymerase (carrying transcription factor σ^K) has been shown to transcribe *in vitro* from *gerE* and at least three other genes involved in sporulation. The promoter regions of these genes each contains sequences similar to CATA– – – TA at about position -10 relative to their transcription start sites, suggesting that this particular sequence is important for recognition of the promoters by σ^K RNA polymerase. A related sequence is also found in the promoter region of genes transcribed by σ^E RNA polymerase. Thus, this undecamer DNA molecule contains the consensus sequence for promoters recognized by RNA polymerases carrying the sporulation σ factors.

This study also continues our investigation of the structural qualities of alternating purine-pyrimidine sequences, especially A+T-rich sequences. Previous studies of DNA duplexes in our laboratory have shown clear sequence-dependent variations in the geometry which has been statistically analyzed (Ulyanov and James, 1994). Interestingly, two of the sequences studied exhibit a noticeable bend of the global helix axis at each of three TG·CA steps. However, in each of the three cases, the T was preceded by a pyrimidine. The DNA molecule investigated in this paper presents a TG·CA step, with a preceding A, in the middle of the sequence interrupting an alternating series of AT steps.

Certainly the high-resolution structure of this molecule will prove valuable for elucidating conformational variations related to DNA sequence.

2.2 MATERIALS AND METHODS

2.2.1 NMR experiments

Sample Preparation. The two DNA undecamers, d(GCATATGATAG) and d(CTATCATATGC), were synthesized on an Applied Biosystems Inc. model 380B synthesizer and purified via reverse-phase HPLC as described previously (Stolarski et al., 1992). The DNA duplex was prepared for NMR by titration one undecamer in phosphate buffer solution, pH 7.0, with a similar solution of the complementary strand, monitored via 1D NMR. The resulting solution (20 mM sodium phosphate, 0.1 M NaCl, 0.1 mM EDTA, pH 6.9) was about 2.5 mM in duplex. For experiments with exchangeable protons, the sample was lyophilized and the dried product dissolved in 90% H_2O-10% ^2H_2O.

NMR Spectroscopy. ^1H NMR experiments were run at 600 MHz on a Bruker AMX600 spectrometer. All 2D NMR measurements were carried out at 25°C except for 2D NOE experiments in H_2O which were also run at 5 and 10°C.

For quantitative analysis, four 2D NOE spectra in ^2H_2O were recorded at mixing times of 70, 130, 200 and 270 ms. The experiments were carried out using exactly the same conditions without removing the sample from the magnet. Quadrature detection in t1 was achieved by TPPI (Marion and Wüthrich, 1983) of the pulse immediately preceding the increment period t1 in 90° steps along with t1. Low-power continuous-wave irradiation was applied to the residual water signal during recycle and mixing delays. A spectral width of 6024 Hz was used with the carrier frequency set to the HDO resonance frequency. A total of 1024 FIDs of 4K real data points were collected in t1 with 32 scans at each t1 value. The delay between scans was 2.5 sec.

2D NOE spectra in H_2O were acquired with a 1-1 echo excitation pulse centered on H_2O. Data were collected for a spectral width of 12500 Hz and the excitation maximum set to the imino-proton region at 13.2 ppm. Spectra were recorded with a 158 ms mixing

time at 5, 10 and 25°C. 512 FIDs of 2K real points were collected in t1 with 64 scans at each t1 value.

2D NOE spectra were processed using TRIAD, the NMR software within the Sybyl software package of TRIPOS, Inc. (St. Louis, MO). A Gaussian window function was used for resolution enhancement in both dimensions. Prior to Fourier transformation, the FIDs were zero-filled to give a final 2K x 4K data set with a digital resolution of 1.5 Hz/point in ω_2 and 2.9 Hz/point in ω_1 in the case of the spectra recorded in ^2H_2O. With the larger spectral width, spectra acquired in H_2O had a lower final resolution: 3.1 Hz/point in ω_2 and 6.1 Hz/point in ω_1.

Pure absorption 2QF-COSY data were recorded using the TPPI scheme, with 1024 FIDs of 4K points collected in t1 and 48 scans for each t1 delay. Acquisition times were 340 ms and 85 ms in t2 and t1, respectively, with a repetition delay between scans of 2.5 seconds. The spectral width was set to 6024 Hz. The averaged FIDs were apodized using a squared sine bell function, shifted by 30°, in both dimensions and zero-filled to a final data size of 1K × 4K points. Digital resolution of the processed spectrum was 1.47 Hz/point in ω_2 and 5.88 Hz/point in ω_1.

The total correlated spectroscopy (TOCSY) data were acquired by standard methods (Bax and Davies, 1985; Braunschweiler and Ernst, 1983). A 4K × 1K spectrum was acquired with 32 transients at each t1 value. Spectral width was set to 6024 Hz in both dimensions. This spectrum was processed to yield a 2K × 1K spectrum using a sine-square apodization function shifted by 70° applied in both dimensions prior to Fourier transform.

Measurement of ^1H spin-lattice relaxation times T1 was performed by the inversion-recovery method. Spin-spin relaxation times T2 were determined using the Hahn spin-echo method. Imino proton exchange rates with bulk water at 5, 10 and 25°C were determined by analyzing the exponential decay of imino protons while a spin-lock field was applied selectively to the water signal (Adams and Lerner, 1992; Liu et al., 1993).

^1H-detected ^{13}C NMR spectra were acquired at 25°C on a Varian Unityplus 600 MHz spectrometer equipped with a triple-resonance probe and actively-shielded gradient coils. These spectra were all processed and analyzed in the usual way with the locally written

programs STRIKER and SPARKY (Day and Kneller, UCSF, 1995). The ^{1}H-detected heteronuclear multiple-quantum coherence (HMQC) spectrum (Bax and Subramanian, 1986) was recorded using an unlabeled sample and the following parameters: 256 FIDs in the t1 dimension, 64 repetitions for each FID, 2048 points in the t2 dimension, and a spectral width of 8000 Hz in each dimension, resulting in about 13 ppm and 53 ppm spanned in the proton and carbon dimensions, respectively.

^{13}C T1 values were measured using a pulse sequence that includes a double DEPT transfer of magnetization from ^{1}H to ^{13}C and then back again to ^{1}H, after being labeled by T1 of ^{13}C, for detection (Nirmala and Wagner, 1988). This pulse sequence had been modified to use gradients instead of homospoil pulses. Due to the low sensitivity of this method on an unlabeled sample like ours, a large number of repetitions was required to obtain a useful signal, precluding two-dimensional experiments. Five 1D spectra, each with 8K data points and 8000 Hz spectral width, were acquired with relaxation delays of 2, 10, 100, 200 and 500 msec, and were used to estimate T1 values. To minimize experimental error and increase signal-to-noise ratio, each spectrum was acquired twice and the sum of corresponding spectra used to determine T1.

$^{13}C\{^{1}H\}$ NOE experiments were acquired in a similar fashion, with the pulse sequence modified to measure NOE instead of T1 values. Again, due to low sensitivity of the method on an unlabeled sample, 1D spectra were obtained rather than 2D spectra. Thus, 1D spectra of 8K data points and 8000 Hz width were acquired with and without NOE transfer being allowed. Each experiment was repeated twice and the sum of corresponding spectra used. The $^{13}C\{^{1}H\}$ NOE was determined by the difference of corresponding peak intensities acquired with and without allowing NOE transfer between atoms.

2.2.2 Analysis of NMR Spectra

Model Structures. Several structures were generated to be used as starting models for running MARDIGRAS, CORMA, molecular dynamics simulations and Monte Carlo calculations. Three different molecular modeling programs were used to generate all-atom models with all hydrogen atoms explicitly included: the module NUCGEN of AMBER Version 4.0 (Pearlman et al., 1990), Sybyl 6.0 (TRIPOS), and the program

DNAminiCarlo (Gorin et al., 1990) which allows generation of DNA duplex models by specifying generalized helical parameters. Standard A and B form DNA models (Arnott and Hukins, 1972; Arnott and Hukins, 1973) were created with each of the programs mentioned above. All structures were then energy-minimized by the three programs using their different empirical force-fields.

Extraction of proton-proton distance bounds from 2D NOE spectra. 2D NOE intensities were measured using the locally written NMR-processing program package SPARKY. Intensities were determined by line fitting the cross-peaks to a Gaussian function and subsequent integration of the theoretical curve (Weisz et al., 1992). With the 2K × 4K spectra, good fitting was achieved even for weak cross-peaks, allowing extraction of all intensities solely by line-fitting. Final intensities were obtained by averaging corresponding cross-peaks on both sides of the diagonal. In the few cases where the two cross-peaks intensities were significantly different, the most reliable was taken. Cross-peaks involving exchangeable protons were integrated from 2D NOE spectra recorded in H_2O and scaled accounting for the profile of the 1-1 echo excitation sequence.

Proton-proton distances were calculated from 2D NOE cross-peak intensities using the complete relaxation matrix approach implemented in the program MARDIGRAS (Borgias and James, 1990). A single correlation time for the whole molecule (isotropic motion) was assumed with no internal motion. For methyl groups, a three-state jump model was used (Liu et al., 1992). The overall isotropic correlation time τ_c was estimated by three independent methods (*vide infra*). To cover conservatively all uncertainty in the results, three sets of MARDIGRAS runs, with three different correlation time values (2.0, 3.0 and 4.0 ns) were performed. Experimental intensities involving nonexchangeable protons from four 2D NOE spectra acquired at 70, 130, 200 and 270 ms mixing times were used. Prior to the calculation of the distances, the experimental intensities were normalized with the theoretical ones by considering only intensities involving protons at fixed distances, i.e., H5-H6 in cytosines, methyl-H6 in thymines and H2'-H2" proton-proton distances. A total of eleven models were used as starting structures to run MARDIGRAS (standard and energy-minimized A- and B-DNA generated by AMBER,

Sybyl and DNAminiCarlo; energy-minimized A-DNA by Sybyl was not used – refer to model structure section above) for a total of 132 runs.

To calculate distances involving exchangeable protons, a single set of intensities acquired at 158 ms mixing time, was used, together with two starting models and three estimates of correlation time. To account for experimental errors, the RANDMARDI algorithm implemented in the MARDIGRAS software was applied (Liu et al., 1995).

For each proton pair, average values for distances, upper and lower bounds were calculated as well as standard deviations. The corresponding standard deviations were added and subtracted from the average distances, and the resulting values were compared to the average upper and lower bounds. Then, the two values that resulted in the larger width, were chosen as upper and lower bounds for structure refinement.

In the same fashion, the program CORMA (Borgias and James, 1988) was run to simulate NOE intensities from a model structure and to compare those intensities to the corresponding experimental intensities by calculating conventional R factors (R) and sixth-root R factors (R^x) (James, 1991). A 3.0 ns overall correlation time value was used for all models.

2QF-COSY simulation. 2QF-COSY cross-peak patterns for deoxyribose protons were simulated using the programs SPHINX and LINSHA (Widmer and Wüthrich, 1986). A six nuclei spin-system including H1', H2', H2", H3', H4' protons as well as ^{31}P bound to the 3'-O was considered. Weak coupling was assumed between all sugar protons except for the strong coupling between H2' and H2". Deoxyribose coupling constants were extracted by matching experimental and simulated cross-peak patterns along the better resolved ω2 dimension (Weisz et al., 1992). Then the conformation of the sugar rings was estimated by matching our experimental coupling constants to values calculated assuming a two-state model with rapid N-S interconversion for the sugar pucker. A graphical method described elsewhere (Gonzàlez et al., 1994) was used to find the best match. The proton-proton coupling constants determined by Rinkel and Altona for specific deoxyribose conformers were used (Rinkel and Altona, 1987).

2.2.3 Structure Refinement

Two different methods were used for structure refinement: simulation of molecular dynamics with experimental restraints (rMD) and restrained Monte Carlo calculations (rMC). The coordinates have been deposited at the Brookhaven Protein Data Bank; PDB Accession Number 1SKP.

Generation of restraints for structural refinement. The restraints used during molecular dynamics simulations and Monte Carlo calculations included interproton distances derived from experimental NOE data and Watson-Crick hydrogen bond distances and angles. Sugar torsional angles, from 2QF-COSY analysis, were not used, since the sugar conformation appeared to be well-defined by interproton distance restraints alone. The structural insights derived from scalar coupling data can be used as an independent monitor of the structural results.

The restraint term introduced into the potential energy function to enforce the experimental restraints takes the quintipartite form of a flat-well potential with a flat, parabolic and linear parts. The exact form of the restraint term used by AMBER has been previously described (Weisz et al., 1994). Upper and lower bounds from the MARDIGRAS analysis delimit the flat region of the potential. In the flat region, no energy penalty is applied. The force constant for the experimental distance restraints term, k_{NOE}, was set to 20 kcal/mol*Å^2 at 300 °K. This value was increased when the temperature was raised during rMD and rMC calculations.

Restraints for hydrogen bonds were added for Watson-Crick base pairs. These include distance restraints and flat angles between the three atoms forming the hydrogen bond. The values used are in agreement with crystallographic data (Saenger, 1984) and have been previously used in NMR structure refinement (Weisz et al., 1994). Use of Watson-Crick hydrogen bond restraints is especially important to prevent the two strands of the duplex from flying apart when the temperature of the system is raised during simulations. The force constant for Watson-Crick hydrogen bond distance restraints was always lower than for experimental restraints throughout the simulations.

Restrained Molecular Dynamics Simulations. Restrained molecular dynamics calculations were performed using the module SANDER of AMBER version 4.0

(Pearlman et al., 1990) on a cluster of Hewlett Packard 735 computers. Molecular dynamics simulations were performed using energy-minimized A- and B-DNA as starting models. These structures were generated by the NUCGEN module of AMBER and energy-minimized (*vide supra*). Prior to minimization, large (hydrated) sodium ions (radius = 5.0 Å) were added along the DNA backbone, 5 Å from the phosphorus atoms, to neutralize the phosphate negative charges; they were subsequently free to move. All MD simulations were run *in vacuo* with optimized parameters previously used to refine duplex structures in our lab (Gonzàlez et al., 1995; Mujeeb et al., 1993; Weisz et al., 1994).

Starting from B DNA, rMD were run for 50 ps at a constant temperature of 300 °K. No restraint forces were applied during the first 200 steps of the run. From step 201 to 1000, the restraint force constant was gradually increased to its final value (20 kcal/mol*Å^2) which was then maintained through the rest of the simulation. When A-DNA was used as starting structure for MD runs, an annealing protocol was used to overcome local energy barriers. The temperature of the system was gradually raised to 900 °K during the first 2000 steps of the run. The system was then kept at this temperature for the next 10000 steps and gradually cooled to 300 °K. The temperature was not changed for the last 10000 steps of the simulation. The restraint force constant was increased to 100 kcal/mol*Å^2 when the temperature was raised to 900 °K and then decreased to its final value of 20 kcal/mol*Å^2 when the temperature was lowered to 300 °K.

Restrained Monte Carlo Calculations. To assess dependence of the final geometry on the refinement method used, we also calculated the DNA undecamer structure by restrained Monte Carlo methods with the protocol we previously reported (Ulyanov et al., 1993). The DNAminiCarlo program, run on the Cray C90 at the Pittsburgh Supercomputing Center, uses generalized helical parameters as independent variables to generate DNA structures. For our undecamer, we had ten subsets of six variables defining each of the ten steps, plus eleven subsets of ten base-pair variables, with the sugar puckers defined by the phase angle of pseudorotation P (assuming pucker amplitude of ~35°).

Two DNA structures were generated by DNAminiCarlo with helical parameters typical of standard A- and B-DNA and energy-minimized by the same program. The same Monte Carlo simulation protocol (Markov chain) was used starting from both A- and B-DNA, consisting of a single cycle of simulated annealing. Starting at 100 °K, the temperature was gradually raised to 600 °K in six steps 100 °K apart. The NOE force constant, k_{NOE}, was increased from 5 kcal/mol*Å^2 to a maximum value of 80 kcal/mol*Å^2 at 500 °K and this value maintained when the temperature reached 600 °K. Temperature and k_{NOE} were kept constant for 1000 iterations before proceeding to the next fragment of the Markov chain. Each iteration was defined when all independent variables defining the DNA undecamer were changed. The last structure generated during each fragment was used as the starting point for the following one. The molecules were kept at 600 °K for 5000 iterations to allow crossing of conformational energy barriers. The system was then gradually cooled to 300 °K in three steps of 1000 iterations, each step being 100 °K apart. Simultaneously, the force constant was decreased to its final value of 20.0 kcal/mol*Å^2. At last, the temperature was maintained at 300 °K for 10000 iterations. To ensure proper base pairing, the force constant for the Watson-Crick hydrogen bond term was maintained at half the value used for k_{NOE}, except for the last 10000 iterations where a value of 5 kcal/mol*Å^2 was used.

All DNA independent variables from the last Markov chain fragment at 300 °K were averaged, and the mean values were used to construct the final structure. This was then restrained-minimized with the k_{NOE} set to 20 kcal/mol*Å^2.

Structure analysis and display. Values of the atomic root-mean-square deviation (RMSD) between model structures were calculated by the program CARNAL (written by Wilson S. Ross). Sugar puckers, torsion angles and helical parameters were calculated with the programs CURVES (Lavery and Sklenar, 1990) and Dials & Windows (Ravishanker et al., 1989). Values calculated by the two programs were quite similar. All structures were displayed using Midasplus (Gallo et al., 1989) and/or Sybyl (Tripos).

2.3 RESULTS

2.3.1 Proton Resonance Assignments

For the residues of the undecamer duplex, the following numbering scheme was used:

	1	2	3	4	5	6	7	8	9	10	11	
5'-	G	C	A	T	A	T	G	A	T	A	G	-3'
3'-	C	G	T	A	T	A	C	T	A	T	C	-5'
	22	21	20	19	18	17	16	15	14	13	12	

Nonexchangeable protons were assigned by following intranucleotide and sequential NOE connectivities typical for right-handed DNA (Scheek et al., 1983; Feigon et al., 1983) in 2D NOE spectra recorded from ^2H_2O. 2QF-COSY, TOCSY and natural abundance ^1H-^{13}C HMQC experiments aided with assignments. Due to severe peak overlap, few H4' and most H5'/H5" resonances were not assigned.

Exchangeable protons could be easily assigned by following well-established strategies (Boelens et al., 1985). Unfortunately, chemical shift degeneration between H3 in T9 and T15 and in T4 and T18 prevents assignment of most amino protons for A5 , A19 , A8 and A14. Moreover, the guanine amino proton resonances were not observed under conditions employed.

2.3.2 Sugar Conformation from Analysis of 2QF-COSY Spectra

The proton homonuclear $^3J_{1'2'}$, $^3J_{1'2"}$, $^3J_{2'3'}$, $^3J_{2"3'}$, $^3J_{2'2"}$ coupling constants were extracted by simulation of 2QF-COSY cross-peak patterns with SPHINX and LINSHA programs (Liu et al., 1995; Schmitz et al., 1990). Unfortunately, due to severe peak overlap, a significant number of residues were not analyzed: A3/A14, A5/A8/A17/A19 and T4/T18/C12 give fully or partially overlapped cross-peaks. Moreover, H2' and H2" of C12 are isochronous.

To obtain pseudorotational parameters, which describe the geometry of the sugar moiety, experimental J-coupling constants were compared with theoretical values using a graphical method (Gonzàlez et al., 1994). To match experimental values, we needed to assume a two-state model with rapid N-S interconversion, with S being the major conformer. Table 2.1 shows the fraction of the S major conformer f_S, the pseudorotation

angle P_S, and the amplitude Φ_S that gave the best match for each analyzed residue. For the minor N conformer, P_N was kept fixed to 9° and sugar pucker amplitudes were assumed to be the same for both S and N conformers ($\Phi_N = \Phi_S$). Except for the C12 terminal sugar, the fraction of minor N conformer was less than 25%. As expected, terminal residues have an increased population of N conformer, undoubtedly due to fraying of the duplex at its ends. Only for T13 was it possible to match the experimental J-coupling values by using only the S conformer (f_S=1.0). Also, for some residues, matches were found with different amplitudes Φ_S.

In conclusion, our analysis of 2QF-COSY cross-peaks suggests that sugars in DNA duplexes in solution are flexible, undergoing fast exchange between N and S conformers, with S being the favored conformer. Only well-resolved residues have been studied. However, 2QF-COSY pattern similarities suggest that overlapped sugars behave similar to the well-resolved ones.

Table 2.1. Phase angle of pseudorotation, P_S, pucker amplitude, Φ_S, and fraction, f_S, of theS-type conformer in d(GCATATGATAG)·d(CTATCATATGC). P_N was assumed to be 9° and Φ_N=Φ_S in the analysis.

Residue	f_S	P_S (deg)	Φ_S (deg)
G1	0.85-0.94	175-194	40
T6	0.85-0.9	145-171	30
G7	0.86-0.91	161-179	35
	0.85-0.95	179-194	40
T9	0.85-0.91	139-173	30
	0.79-0.88	130-149	35
A10	0.9-0.95	165-186	40
G11	0.77-0.8	139-163	35
	0.76-0.8	130-145	40
C12	0.45-0.56	163-198	35
	0.5-0.6	169-212	40
T13	0.9-1.0	157-184	35
T15	0.75-0.8	122-133	40
C16	0.8-0.9	155-183	30
	0.65-0.83	139-169	35
T20	0.77-0.85	132-151	35
	0.75-0.8	122-130	40
G21	0.80-0.95	155-175	40

2.3.3 MARDIGRAS Calculations

To refine the structure of our DNA undecamer at high resolution, we need to calculate accurate and, preferably, precise distance restraints. This is accomplished in our laboratory by using the program MARDIGRAS which solves the complete relaxation matrix, thus incorporating all network relaxation and multispin effects (Borgias and James, 1990). To construct a complete intensity matrix, MARDIGRAS first combines experimental intensities with intensities calculated from a starting model by the CORMA algorithm (Borgias and James, 1988; Keepers and James, 1984). The experimental intensities must be normalized prior to construction of the hybrid matrix.

Previous studies have shown that the initial model used to construct the hybrid matrix has little effect on the distances calculated by MARDIGRAS. However, in our system we have noticed that some distances depend on starting geometry more than others. Such dependency may reflect ambiguity in calculating the distances under the conditions used, so we decided to run MARDIGRAS using several starting models and to generate distance bounds by combining the results from all MARDIGRAS runs. Then, any ambiguity in distances calculated should be reflected in wider distance bounds. Eleven structures were used as initial models for MARDIGRAS. These structures were constructed using different molecular modeling programs to represent A- and B-DNA conformations; some differences were found between corresponding conformations generated by separate programs, so all models were used to run MARDIGRAS.

Estimation of Overall Correlation Time τ_C. Another important parameter in calculating proton-proton distances is the molecular motional model assumed, since motion can affect the value of the NOE intensities. For a hydrated undecamer duplex, the overall molecular tumbling can be considered isotropic and, hence, is well-approximated by a single correlation time, τ_C. The value of τ_C was estimated using various methods.

The simplest method to derive τ_C is from the proton spin-lattice relaxation time (T1) and spin-spin relaxation time (T2):

$$\tau_c = \frac{2}{\omega_H} * \sqrt{\frac{3 * T2}{T1}}$$

where ω_H, the proton Larmor frequency, is given by $2*\pi*\nu_H$, ν_H being the precession frequency in Hz and assuming that overall tumbling at 600 MHz is in the slow motion limit, $(\omega_H*\tau_c)^2 \gg 1$. T1 and T2 values were measured for resolved adenine H2 and H6/H8 protons at 25°C. The calculated τ_c values are reported in Table 2.2 together with the corresponding T1 and T2 values.

τ_c was also estimated from natural abundance ^{13}C T1 and $^{13}C\{^1H\}$ NOE values which were calculated from the height of broad overlapped peaks in the H1' region of 1D spectra at 25°C. τ_c values were extracted using equations describing the variation of ^{13}C T1 and $^{13}C\{^1H\}$ NOE with τ_c (Abragam, 1961), with all constants set to match experimental conditions. The τ_c values, 3.2-4.7 ns estimated from ^{13}C T1 values and 0.65-1.4 ns from $^{13}C\{^1H\}$ NOE, are not strictly in accord. They both, however, lay in the same range found by the proton T1/T2 method described above. Possible contributions from internal motion were ignored in all calculations.

Finally, it is also possible to utilize the iterative relaxation matrix calculations for an estimation of the correlation time with the assumption that the best fit for the experimental intensities is obtained for the "correct" correlation time. Unfortunately, this endeavor is not straightforward since the necessary normalization for experimental and model intensities obscures that effect. An independent normalization factor is required that must be kept constant when τ_c is varied. This type of normalization is feasible by adding the total intensities for 2D NOE slices of isolated protons, including the contribution from the diagonal peak, and determining the total intensity for a proton accordingly. Then, normalized theoretical intensities for these isolated proton slices can be compiled with MARDIGRAS using different correlation times. For the undecamer, results for diagonal, H6-methyl, H2'-H2" intensities fell into a range between 2 and 4 ns, which is quite similar to τ_c values derived from T1 and T2 relaxation parameters. However, inconsistent results were obtained for cytosine H5-H6 cross-peaks, where only the intensities obtained at short mixing times could be reproduced with reasonable correlation times, which were still larger (3-6 ns) than those obtained for the other protons.

Table 2.2. T1, T2 and τ_c values for most H6/H8 and some H2 protons.

Atom	T1 (sec)	T2(sec)	τ_c (ns)
A14 H8	2.70	0.0432	2.42
A3 H8	2.21	0.0254	2.86
A17 H8	1.92	0.0208	2.95
A5 H8,A19 H8	2.09	0.0190	3.21
A8 H8,A10 H8	2.22	0.0476	2.09
A3 H2	3.60	0.0615	2.34
C2 H6	1.22	0.0833	1.17
C22 H6	1.67	0.0258	2.46
C16 H6	2.23	0.0185	3.36
G7 H8	1.93	0.0465	1.97
T6 H6	1.75	0.0234	2.64
T9 H6	2.69	0.0303	2.89
T13 H6	1.25	0.0308	1.95
T20 H6	1.56	0.0272	2.32

In conclusion, the uncertainty in evaluating τ_c by different methods induced us to run MARDIGRAS calculations using three different values of the overall isotropic correlation time: 2, 3 and 4 ns.

Calculation of Nonexchangeable Proton-Proton Distances. MARDIGRAS calculations were performed on four sets of experimental intensities recorded at four mixing times (70, 130, 200 and 270 ms), using eleven different structures as starting models and with three different correlation time values (2.0 ns, 3.0 ns and 4.0 ns) for a total of 132 runs. Combining results from all these different runs, a total of 447 unique proton-proton distances were obtained. Table 2.3 shows the number of distances calculated from intensities at each mixing time, specifying how many intraresidue, interresidue and interstrand distances were found. Due to uncertainty in stereospecific assignment of H5' and H5" protons, all 41 cross-peaks involving these atoms were not used for structure refinement, as well as five cross-peaks implying proton n – methyl (n+2) interactions, due to spin diffusion. Ultimately, a total of 399 distances from nonexchangeable protons, i.e., about 18 per residue, were used for structure determination.

Table 2.3. Total number of nonexchangeable proton-proton distances calculated by MARDIGRAS for the four different mixing times at which the NOE spectra were acquired. The number of intraresidue, interresidue and interstrand distances are also shown. † – distances between nonexchangeable protons only. * – distances involving exchangeable protons extracted from 2D NOE in H_2O.

τ_m (ms)	Distances			
	Total	Intraresidue	Interresidue	Interstrand
70	328	187	134	7
130	409	225	162	14
200	438	237	174	16
270	445	236	185	15
all τ_m†	447	239	192	16
158*	36	8	8	20

Overlapped Peaks in 2D NOE Spectra. All corresponding nonexchangeable protons in residues A5 and A19 were found to be overlapped as well as all corresponding sugar protons in residues T4 and T18. As a result, no isolated intraresidue cross-peaks for A5 and A19 and only few T4-A5 and T18-A19 interresidue cross-peaks could be found. This overlap can be justified by noting that residues T4-A5 and T18-A19 are located in the center of a self-complementary segment of the DNA sequence; residues 2-7 and residues 16-21 have the sequence CATATG. Consequently, these residues must have very similar chemical environments, suggesting nearly identical three-dimensional structures. We can safely divide by two the volume of overlapped cross-peaks and assign each half to the degenerate proton pairs. By doing this, we were able to obtain 24 experimental restraints, which are included in the above-mentioned 399 nonexchangeable bounds. These restraints are very important (Figure 2.1), since they belong to a region of the molecule with otherwise few or no restraints at all. For example, A5 and A19 have no intraresidue restraints if we ignore the overlapped cross-peaks.

With overlapped cross-peaks included, the experimental restraints are well distributed along our DNA undecamer sequence with 15-20 distance restraints for each residue (Figure 2.1).

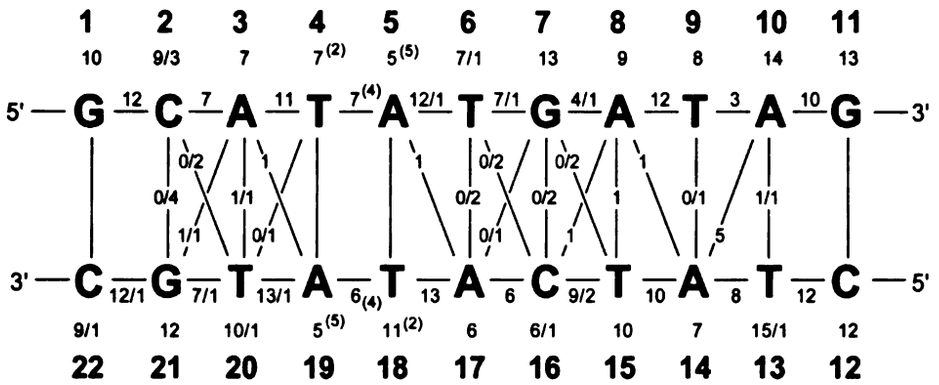


Figure 2.1. Distribution of NOE-derived distance restraints for the DNA undecamer duplex. Bold numerals indicate the sequence order. Plain numerals indicate number of intraresidue, sequential, and interstrand restraints with placement implying attribution. For interstrand restraints, the first number gives the number involving only nonexchangeable protons and the second gives the number involving at least one exchangeable proton. Due to peak overlap, some cross-peaks result in restraints being assigned to two similar residues (see text for method of attribution); the number within parenthesis: two, four, or five, are overlapped restraints residue pairs T4/T18 and A5/A19.

Estimation of distance error bounds for molecular dynamic simulations. With the conservative approach employed (*vide supra*), the average width of our calculated bounds is 0.77 Å for 447 total distances. For most proton pairs, the bound width increases with the absolute value of the distance, i.e., longer distances have wider bounds. However, the exact width depends on the nature of the two protons involved. The width of our bounds is larger than the average width of bounds previously used to refine DNA structures in our laboratory. The effect of having such wide bounds on structure determination will be further investigated later in this paper.

Problems with Short Repetition Delay. All NOE spectra in ^2H_2O were acquired using a relatively short repetition delay (2.5 sec) that does not allow for complete relaxation of all proton longitudinal magnetization between scans. The major effect is that cross-peak intensities will depend differentially upon the delay time, varying most for peaks involving protons with very different T1 values, such as H2 protons in DNA molecules. The effect of short repetition delay on NOE intensities of DNA duplexes has been previously discussed (Liu et al., 1996). To test for such effect, we corrected our intensities with the program SYMM (Liu et al., 1996). The distances calculated from

such intensities were in all cases within the original bounds, indicating that our conservative approach already accounted for any error deriving from incomplete relaxation.

MARDIGRAS calculations entailing exchangeable protons. NOE intensities from 2D spectra recorded in H_2O were corrected to account for the excitation profile of the 1-1 echo water suppression sequence used.

MARDIGRAS calculations were performed with experimental NOE intensities acquired at 10°C with 158 ms mixing time, using three different estimates of correlation time (2.0, 3.0 and 4.0 ns) and two different initial models (energy-minimized A- and B-DNA structures generated by AMBER). Nonexchangeable distances were fixed to the values previously calculated from ^2H_2O data: the MARDIGRAS complete hybrid intensity matrix was constructed by combining the normalized experimental exchangeable proton intensities with intensities calculated from initial model distances and "known" nonexchangeable distances. The lower bound of distances involving exchangeable protons were corrected to account for the lost of magnetization through exchange with the bulk water (Adams and Lerner, 1992; Gonzàlez et al., 1995; Liu et al., 1993) . To be conservative, the largest determined exchange rate (6.0 sec^{-1}) was used for all amino and imino protons. Finally, to account for the contribution to distance error from noise in the spectra and integration errors, the RANDMARDI algorithm was applied (Liu et al., 1995). Upper and lower bounds were chosen from the wide set of distances yielded by the calculations.

A total of 36 new proton-proton distance restraints involving imino and amino protons were calculated (Table 2.3). These bounds are especially important because they include 20 interstrand restraints (Figure 2.1).

Comparing distances calculated by MARDIGRAS using different overall correlation times. Due to some discrepancy in correlation time values estimated using different methods, three different sets of MARDIGRAS runs were performed assuming τ_c = 2.0, 3.0 and 4.0 ns. Since this range is somewhat larger than previously used in our laboratory, we investigated the effect of correlation time on distances calculated. To summarize, we found the deviation between distances calculated with the different τ_c

values is always <10%, usually much less. The MARDIGRAS normalization scheme eliminates most, but not all, of the error introduced by uncertainty in τ_c.

2.3.4 Structure Refinement Results

The coordinates have been deposited at the Brookhaven Protein Data Bank; PDB Accession Number 1SKP.

Restrained Molecular Dynamics Calculations. Ideally, the structure of our DNA undecamer duplex should be fully characterized by the experimental restraints. However, the method chosen for structural refinement may play an important role in defining conformation, especially for regions of molecules which have sparse or conflicting experimental restraints.

We have already discussed how we extracted structural restraints from the experimental data, allowing for experimental error and testing those variables that could affect the final distances. Our final list of experimental restraints consists of 435 sets of proton-proton upper and lower distance bounds (Figure 2.1). Furthermore, with the large number of MARDIGRAS runs we performed (*vide supra*), our bounds are, on average, wider than corresponding bounds previously used in our laboratory to refine DNA duplex structures. The effect of wider bounds on structure calculation will be further investigated here.

Since we already know the general folding of the DNA in solution, i.e., right-handed duplex, we can directly use structure refinement methods that require a "good" starting model. One popular method utilizes molecular dynamics with experimental restraints to search conformational space for structures which are compatible with both the experimental data and a chemical force field. For duplex DNA, we run restrained MD simulations starting from two different models: energy-minimized A- and B-DNA. *Amd1* and *Bmd* are the final structures, shown in Figure 2.2, obtained from starting A and B models, respectively. It is important to run multiple simulations starting from different models to verify the quality of the experimental restraints; weak or inconsistent bounds may not be able to drive different initial models to the same final structure.

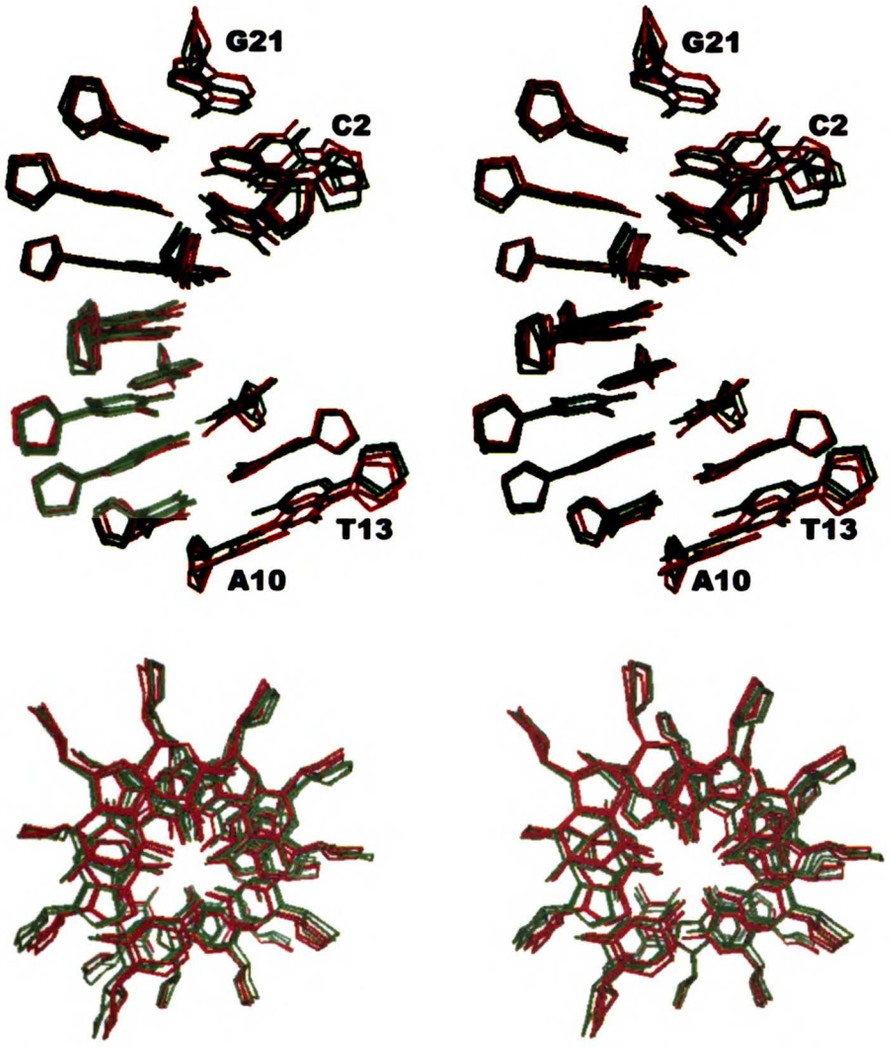


Figure 2.2. Stereo view of four superimposed final structures for the duplex: two models refined by molecular dynamic simulations starting from B and A form DNA, *Bmd* and *Amd1*, respectively (green), and two models refined by Monte Carlo calculation starting from B and A models, *Bmc* and *Amc* (red). The top figure displays only base and sugar heavy atoms of the 20 internal residues. The bottom view, by showing the superimposed structures along the helical axis, illustrates the A-like hole in the center of the duplex.

To assess differences between the various structures, we calculated the atomic root-mean-square- deviation (RMSD). Three RMSD values were determined (Table 2.4): all heavy atoms, all heavy atoms except those in terminal residues, and all heavy atoms except terminal and backbone atoms. The refined structures, *Amd1* and *Bmd*, have a much lower RMSD than the corresponding initial models, A- and B-DNA, and indicate convergence to essentially the same structure. A closer look at the RMSD values reveals that convergence is worse for terminal residues than for the internal ones (compare first

and second values in Table 2.4), suggesting that experimental bounds acting on terminal residues are not consistent with a single structure, probably due to fraying at the duplex ends. In contrast, the low deviation shown by internal residues, suggests that they are well-defined by the experimental data. Also, the low RMSD values calculated between *Amd1* and *Bmd* with or without taking into account backbone atoms (second and third RMSD values) suggest that these structures have very similar backbone geometry. Apparently, even though no experimental data directly restrain backbone atoms (except protons belonging to the deoxyribose ring) and no empirical torsion angle restraints were used, the backbone conformation, surprisingly, does not depend on initial geometry. Evidently the geometry of internal bases and sugars, being well-defined by NMR bounds, dictates the backbone conformation.

Depending on initial model, the protocol followed during the MD run may be essential to achieve convergence (Weisz et al., 1994). Thus, rMD starting from the A model was run with a cycle of simulated annealing. During the trajectory, the temperature reached a maximum of 900 °K to allow conformational changes to occur. Starting from B-DNA, the temperature was maintained at 300 °K throughout the simulation.

The final geometry also depends on the restraint force constant value, k_{NOE}. This value is empirically chosen to balance the driving force produced by the experimental restraints and the force field. Our goal is to satisfy the experimental data without forcing the molecule into high energy conformations. To select a k_{NOE} value, we ran several rMD simulations starting from A and B models with increasing k_{NOE}. The lowest force constant that allowed a good convergence between final structures, 20 kcal/mol*Å^2, was chosen. This value was used to calculate *Amd1* and *Bmd* structures.Energies for the converged structures, deducting the restraint energy penalty, are comparable to those of the starting models (Table 2.5). For each molecule, two values of distance violation are reported. These are calculated to the center position of the experimental bounds range, and to the closest bound, Δ_{avg} and Δ_{closer}, respectively. *Amd1* and *Bmd* give sensibly smaller distance violations than the initial models, especially the A model (Table 2.5). The large width of our experimental bounds is reflected in the large difference between Δ_{avg} and Δ_{closer} values.

mA*	0										
	6.1										
mB*	4.7	0									
	4.3										
	2.1	5.0									
Amd1	1.7	3.9	0								
	1.5	3.6									
	2.1	5.0	0.7								
Amd2	1.7	3.9	0.4	0							
	1.5	3.6	0.3								
	2.4	4.7	1.0	1.1							
Bmd	1.8	3.8	0.5	0.5	0						
	1.6	3.5	0.3	0.4							
	3.3	3.7	2.6	2.4	2.2						
Bmd-f	2.6	2.9	2.0	2.0	1.8	0					
	2.4	2.6	1.8	1.8	1.7						
	2.2	4.6	1.2	1.1	0.8	2.1					
Bmd-n	1.6	3.6	0.6	0.6	0.5	1.7	0				
	1.4	3.4	0.5	0.5	0.4	1.6					
	3.8	3.2	3.7	3.6	3.6	2.5	3.3				
mA¶	2.8	2.4	3.0	2.9	2.9	2.1	2.7	0			
	2.7	2.2	2.9	2.8	2.8	1.9	2.6				
	5.1	1.3	4.2	4.1	3.9	2.9	3.8	2.9			
mB¶	3.9	1.1	3.3	3.2	3.1	2.3	3.0	2.2	0		
	3.5	1.0	3.0	3.0	2.9	2.1	2.8	1.9			
	2.3	4.4	1.3	1.3	0.9	1.9	0.8	3.2	3.6		
Amc	1.7	3.6	0.7	0.8	0.6	1.6	0.6	2.6	2.8	0	
	1.4	3.3	0.7	0.7	0.5	1.5	0.5	2.6	2.6		
	2.3	4.3	1.3	1.2	0.8	1.8	0.8	3.1	3.4	0.5	
Bmc	1.7	3.5	0.8	0.8	0.6	1.6	0.6	2.6	2.7	0.3	0
	1.5	3.2	0.7	0.7	0.5	1.4	0.5	2.5	2.5	0.2	
	mA*	**mB***	**Amd1**	**Amd2**	**Bmd**	**Bmd-f**	**Bmd-n**	**mA¶**	**mB¶**	**Amc**	**Bmc**

Table 2.4. RMSD values calculated between initial and refined models. For each pair of structures, three RMSD values were determined taking into account all heavy atoms, all heavy atoms except terminal residues, and all heavy atoms except terminal residues and backbone atoms. *mA* and *mB* are energy-minimized A and B form DNA models generated by AMBER (*) and DNAminiCarlo (¶). Corresponding models generated by the two programs were found to be different (as it can be seen by their RMSD values). *Amd1* and *Amd2* are final models refined by AMBER from *mA** using two different seeds for random number generation (*vide infra*); *Bmd*, *Bmd-f* and *Bmd-n* are final structures calculated by AMBER starting from the *mB** model: *Bmd* was obtained by enforcing our experimental NOE restraints, while *Bmd-f* without applying experimental restraints (free MD), and *Bmd-n* was calculated using empirically narrowed bounds. *Amc* and *Bmc* are final structures calculated by DNAminiCarlo starting from *mA¶* and *mB¶* minimized models, respectively.

Even when rMD runs start from the same model, velocities assigned to the atoms at the beginning of the simulation may affect the conformational search. These velocities are randomly generated from a user-selected seed number input to AMBER. To evaluate the effect of using different seed numbers, we calculated a new structure, *Amd2*, running rMD starting from A form with a different random seed number. This structure is essentially identical to *Amd1*, as indicated by low RMSD values, as well as comparable energies and distance violations (Tables 2.4 and 2.5).

Table 2.5. Potential energies, E_{pot}, constraint energies, E_{cnst}, and distance violations values for initial and final structures. Two kinds of distance violation values are calculated: to the center position of the bounds ranges, Δ_{avg}, and to the closest bound, Δ_{closer}. Energy values are calculated using different forcefields in AMBER (rMD) and DNAminiCarlo (rMC) programs. Hence, the values are not comparable. Refer to the caption of Table 2.4 for a legend of the structure names.

Structures	E_{pot}-E_{cnst} (kcal/mol)	E_{cnst} (kcal/mol)	Δ_{avg} (Å)	Δ_{closer} (Å)
mA*	-1035.471	9464.204	1.086	0.744
mB*	-1196.288	1347.482	0.500	0.218
Amd1	-1096.467	147.740	0.345	0.075
Amd2	-1067.156	140.544	0.339	0.073
Bmd	-1142.620	158.007	0.345	0.079
Bmd-f	-1306.004	3211.108	0.604	0.321
Bmd-n	-1083.511	431.977	0.316	0.081
mA¶	-226.034	-	0.774	0.481
mB¶	-179.325	-	0.515	0.234
Amc	-47.427	704.450	0.435	0.155
Bmc	-56.071	587.310	0.421	0.144

We need to consider that our wide bounds may not have enough restraining power to drive the structure towards its target conformation, letting the chemical force field dominate. To test for this effect, we ran two independent MD simulations: the first without experimental restraints and the second applying bounds with reduced width. Both simulations started from B-DNA models using the same protocol as with wide bounds. We observed that free MD yields a final structure (*Bmd-f*) that, judging from RMSD values, energies and distance violations, is intermediate between A and B form but distant from *Amd1* and *Bmd*. On the contrary, the structure calculated with tighter bounds

(*Bmd-n*) has features similar to the structures calculated with wider bounds (Tables 2.4 and 2.5).

Perhaps the most important criterion for evaluating quality of NMR structures is to estimate their ability to reproduce the experimental data by calculating residual factors, R and R^x. These are determined comparing the NOE intensities calculated for a particular structure at a given mixing time with the corresponding experimental intensities. This is done with the program CORMA (Borgias and James, 1988; Keepers and James, 1984). For each structure we calculated R and R^x values individually for NOE spectra acquired at four mixing times (Table 2.6). R factors sensibly improved after rMD refinement. *Bmd*, *Amd1* and *Amd2* all have comparable R factors, lower than those of the starting geometries, especially A form. As expected, with narrow bounds a slight decrease in R factors resulted, while free MD yielded a structure (*Bmd-f*) with R factors higher than B form. This indicates that the experimental bounds are driving the structure during rMD and not the empirical force field. Consistent with our observation that NOE spectra recorded at intermediate mixing times (130 and 200 ms) gave spectra of better quality, they also had lower R factors.

Restrained Monte Carlo Calculations. For final corroboration of the structures calculated by rMD, we used our experimental bounds as restraints with a different structure refinement method: restrained Monte Carlo calculations (rMC). The Monte Carlo subroutines of the DNAminiCarlo program were used for this purpose. With DNAminiCarlo, not only is the conformational search method different than MD, but also the force field and the definition of molecular structure are different from those used in AMBER. In fact, DNAminiCarlo uses helical parameters to describe the structure of DNA duplexes, rather than atomic coordinates.

Again, to test dependence on initial model, we ran rMC starting from energy-minimized A- and B-DNA models (both generated by DNAminiCarlo). However, in contrast to our rMD simulations, the same protocol was used, consisting of a single cycle of simulated annealing starting from both A and B models. The same value of restraint force constant as in rMD simulations, 20 kcal/mol*$Å^2$, was chosen. Two final structures, *Amc* and *Bmc*, were obtained starting from energy-minimized A and B forms, respectively. RMSD values calculated between *Amc* and *Bmc* are well below 1.0 Å, even

when heavy atoms of terminal residues are considered (Table 2.4). Thus, starting from A and B models, rMC achieved even better convergence than rMD. It is possible that the fewer degrees of freedom allowed during rMC simulations result in better convergence. It is most reassuring that RMSD values calculated against rMD structures indicate that Monte Carlo calculations find a similar geometry as rMD, especially with B-DNA as starting model. *Amc* and *Bmc* structures are shown in Figure 2.2 superimposed on the final models refined by rMD (*Amd1* and *Bmd*).

Energies of structures calculated with different force fields cannot be compared directly. Instead, their energies can be compared to those of A and B forms minimized with the corresponding force fields. Thus, rMC final structures have energies that are about 100 kcal/mol higher than both A and B form models. As in rMD, distance violations decrease also during Monte Carlo refinement; however, the rMC final structures have values that are about 0.1 Å higher than those of rMD structures (Table 2.5).

Analyzing R factors for rMC structures (Table 2.6), we noticed the final structures and the initial B model have similar values, both being considerably smaller than the initial A model R factors. A closer look, reveals that intraresidue NOE intensities are better satisfied by the initial B model, while the interresidue intensities give smaller R factors with the final structures. In addition, some starting models may well satisfy intraresidue intensities, since these depend only on sugar pucker and glycosidic torsional angle. Interresidue bounds, instead, are more complex and can be better accounted for only after restrained conformational search. Because of inconsistencies in NOE intensities (due to experimental error, molecular flexibility, etc.), improved fitting of interresidue restraints may dictate disruption of local geometry, worsening intraresidue R factors.

Energies, distance violations and R factors of structures determined by Monte Carlo methods could be improved by performing consecutive cycles of simulated annealing, as we previously described (Ulyanov et al., 1993). However, our goal was to refine DNA models by rMC methods to conformations similar to those found by rMD. Even though the structure refinement method, the force field and the initial models are different from those used in rMD refinement, we found essentially the same structure. This strongly indicates that the experimental data are responsible for the final geometry and that the

structure we calculated represents a good time-averaged representation of our DNA undecamer in solution.

Table 2.6. Sixth-Root R factors (R^x) calculated by CORMA for initial and refined structures using experimental 2D NOE intensities acquired at four mixing times. Refer to the caption of Table 2.4 for a legend of the structure names.

	270 ms mixing time			200 ms mixing time			130 ms mixing time			70 ms mixing time		
structure	intra	inter	global	intra	inter	global	intra	inter	global	intra	inter	global
*mA**	0.160	0.211	0.182	0.183	0.265	0.217	0.206	0.299	0.244	0.272	0.278	0.274
*mB**	0.062	0.093	0.075	0.063	0.100	0.079	0.071	0.122	0.093	0.087	0.118	0.100
Amd1	0.050	0.060	0.054	0.051	0.062	0.056	0.052	0.063	0.057	0.054	0.067	0.060
Amd2	0.051	0.060	0.055	0.053	0.061	0.056	0.053	0.062	0.056	0.055	0.065	0.059
Bmd	0.050	0.060	0.055	0.051	0.063	0.056	0.053	0.064	0.058	0.055	0.068	0.060
Bmd-f	0.075	0.110	0.090	0.074	0.126	0.096	0.079	0.161	0.113	0.104	0.180	0.135
Bmd-n	0.048	0.058	0.052	0.046	0.063	0.053	0.046	0.069	0.056	0.050	0.073	0.059
mA¶	0.118	0.142	0.128	0.122	0.168	0.141	0.130	0.207	0.162	0.139	0.248	0.184
mB¶	0.062	0.097	0.077	0.062	0.110	0.082	0.071	0.136	0.098	0.111	0.147	0.126
Amc	0.068	0.076	0.071	0.068	0.081	0.073	0.068	0.085	0.075	0.116	0.106	0.111
Bmc	0.067	0.073	0.069	0.067	0.076	0.070	0.068	0.081	0.073	0.112	0.100	0.107

2.4 DISCUSSION

Analysis of Refined Structures. Perhaps the most striking feature of our final structure is the presence of a hole in the center of the DNA duplex when viewed along the helix axis (Figure 2.2 bottom). This quality is observed in canonical A-DNA, but not in B form DNA, and is reflected in the large negative X-displacement of the bases (Figure 2.3). This finding is in accord with a pattern that has been observed in our laboratory for other DNA sequences. In fact, with improvement of the NMR technique, it has become evident that DNA duplexes in solution deviate somewhat from canonical B-DNA with some features tending to resemble the A-DNA conformation. The X-displacement and inclination of the bases are among those helical parameters that deviate most towards A-form. The refined model of our undecamer, however, shows features that are closer to A-form than any DNA duplex previously solved in our laboratory. Since our experimental restraints are much wider than those used to solve other duplexes, we considered the possibility that this large bound width could be the reason for the strong A character we found for our undecamer. To investigate further the determinants of our final

conformation, we ran two independent MD simulations: one using experimental bounds that were empirically narrowed and the second without experimental restraints. The final structure, *Bmd-n*, obtained with narrowed restraints, has X-displacement and inclination of the bases closer to canonical B-DNA, which indeed suggests that our wide bounds are responsible for the A-like features exhibited by our undecamer. Based on this, one could further argue that the empirical force field, permitted by wide experimental bounds, is driving the structure towards A-form. This hypothesis, however, is not supported by free MD simulations, which show that the force field actually drives the molecule towards a structure, *Bmd-f*, with more B-like helical parameters. Moreover, very similar structures, that bear some A-DNA characteristics, were also refined by Monte Carlo methods (*Amc* and *Bmc* models), which calculate molecular energies using a force field different than that in AMBER (Gorin et al., 1990). Consequently, we cannot positively identify any single factor responsible for the strong A character shown by our final models. Perhaps, the large width of our bounds causes different types of restraints to become important in defining the DNA structure. At the same time, we cannot rule out the dependence of the final geometry on the particular sequence of our duplex, since we do not know what the effect of wider bounds on other sequences would be.

One important consequence of X-displacement and inclination of the bases being closer to A type is that the global shape of the molecule resembles that of A form. This may explain why our final structure, surprisingly, gives lower RMSD values when compared to A- than to B-DNA initial models (see Table 2.4).

Helical parameters differences between refined rMD and rMC structures. To investigate how precisely the helical parameters are defined by the experimental bounds, we ran 100 ps of rMD simulations at a temperature of 300 K starting from the structure refined from B-DNA (*Bmd*). The same value of k_{NOE}, 20.0 kcal/mol*Å^2, was used throughout the simulation. At the end, for each helical parameter, we calculated its average and standard deviation values. A first look at the standard deviation of several parameters shows that terminal residues have a broader distribution suggesting, as expected, that the structure of the duplex is less defined at its ends. A similar procedure was also used to obtain final rMC structures. In Monte Carlo simulations, in fact, the independent helical parameters of all the structures generated during the last Markov

chain fragment at 300 K were averaged and the mean values were used to construct the final model. A selection of helical parameters from the 100 ps rMD simulation and for two final rMC structures, *Bmc* and *Amc*, are plotted together in Figure 2.3.

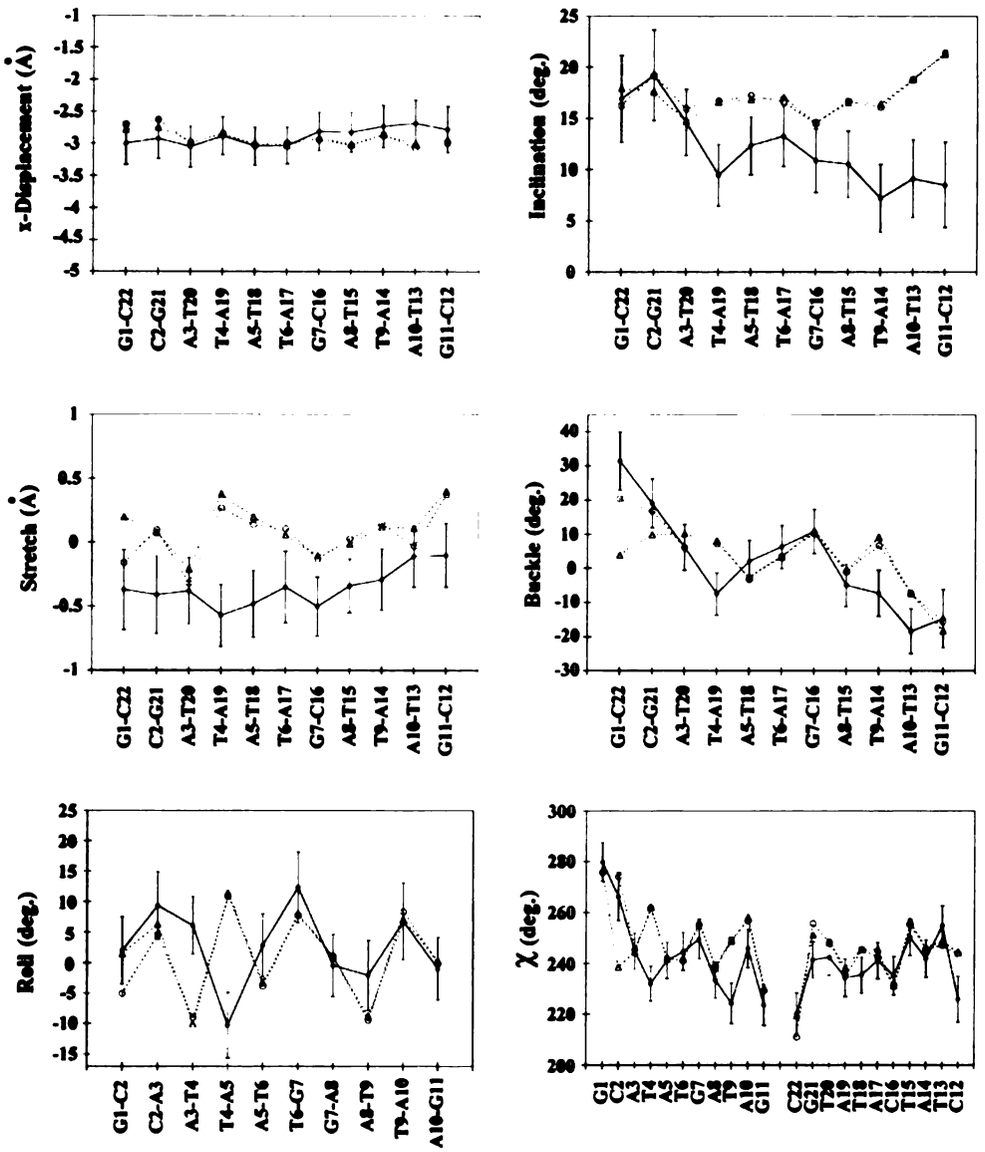


Figure 2.3. Selected helical parameters and glycosidic torsion angle calculated from 100 ps of rMD simulation at 300 K starting from the *Bmd* model (solid curve; average and standard deviation shown) and from the final *Bmc* (o) and *Amc* (Δ) models .

Considering the low RMSD values we calculated between these structures, especially for internal residues, we would expect helical parameters to be very similar. Surprisingly, while the two rMC models are essentially identical, significant changes are found with

rMD helical parameters. In particular, base-pair parameters, e.g., opening and stretch, and base parameters, such as inclination and Y-displacement, manifest the biggest differences. Other rMC helical parameter differences are mainly contained within the standard deviation of the corresponding rMD parameters. Moreover, parameters that define the geometry of the base-pair and steps involving residue T4 show the biggest differences between rMD and rMC structures, while T6, G7 and A8 are among the best-defined residues. To understand this deviation in helical parameters, we need to consider the differences in how the DNA duplex structure and experimental data are handled during rMD and rMC refinements.

First of all, DNAminiCarlo and AMBER, the programs used to run rMC and rMD simulations, respectively, describe DNA duplexes in two different ways. In DNAminiCarlo, the independent helical variables are used to describe duplex geometry, with a single parameter, the pseudorotation phase angle P, defining sugar puckers. Given a value of P, the sugar conformation with minimum energy is taken. Consequently, base and sugar rings during Monte Carlo simulations are locked into optimal conformations. In contrast, with AMBER, the molecules are defined by the cartesian coordinates of their atoms. Thus, AMBER allows more degrees of freedom for the structure to satisfy experimental restraints, which explains why models refined by rMD have lower R-factors and distance violations than those refined by rMC (Tables 2.5 and 2.6).

Second, from analysis of our experimental data, we found that some classes of restraints are systematically inconsistent with others and, hence, all restraints cannot be satisfied simultaneously by any single conformation. A more detailed analysis of NOESY and COSY data induced us to believe that, this lack of consistency, arises from flexibility of the sugar moieties and about the glycosidic bonds in DNA duplexes (see Chapter 3). Differences in how these incongruities are "assimilated" into the DNA structure, may lead to the observed deviations between rMD and rMC refined models. Thus, during molecular dynamics simulations, incompatibilities in the experimental bounds can be partially or fully satisfied by forcing base and sugar rings into slightly distorted conformations. On the other hand, in Monte Carlo calculations, the geometry of base and sugar rings cannot be altered to accommodate experimental inconsistencies, since only

optimal conformations are permitted. Hence, during rMC runs, locally inconsistent restraints perturb the helical geometry.

Inconsistencies between experimental bounds and empirical hydrogen bond restraints, can also contribute or be responsible for observed deviations of base-pair parameters (e.g., opening or stretch). In structures refined by rMD, the bases are often twisted to satisfy experimental bounds and flat Watson-Crick hydrogen bonds, simultaneously. Instead, during Monte Carlo calculations, since bases can not be bent, hydrogen bonds are somewhat disrupted to enforce experimental restraints. As expected, inspection of rMC and rMD final models reveals that the bases have nonflat geometry in structures refined by rMD, but not in the rMC ones. Also, the pucker amplitude of sugar rings, Φ, reaches unrealistic values during simulations of molecular dynamics, while it is contained within a narrow range (35-40°) in Monte Carlo runs (see Φ in Figure 2.4).

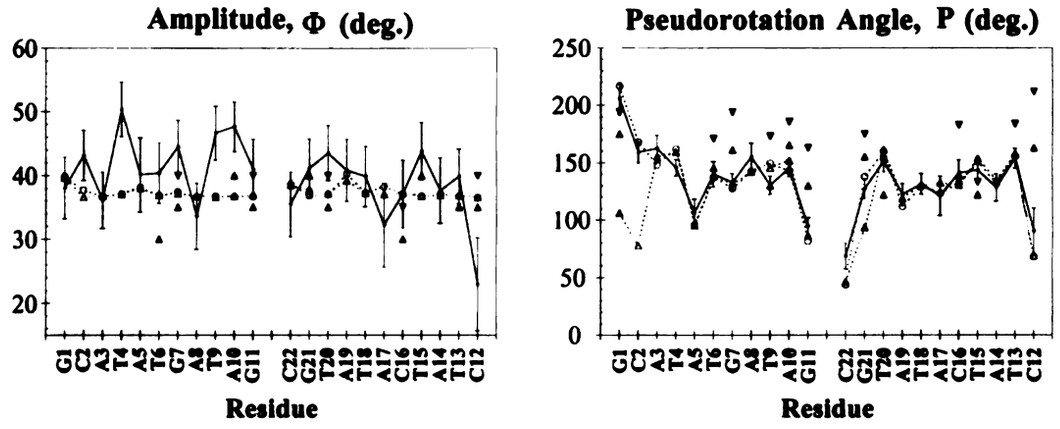


Figure 2.4. Phase angle of pseudorotation P and amplitude Φ calculated for sugar moieties of structures generated during 100 ps of rMD simulation at 300 K starting from the *Bmd* model (solid curve; average and standard deviation shown), for the final *Bmc* (o) and *Amc* (Δ) models refined by rMC calculations, and extracted from 2QF-COSY cross-peak analysis (π).

A third factor leading to differences between rMC and rMD structures is that the helical parameters extracted for rMD structures may not be properly calculated due to deformed base ring geometry. Artifacts in measuring helical parameters may also cause the observed discrepancy between these parameters and RMSD values, i.e., how structures can give RMSD values <1 Å yet show differences in helical parameters. For

example, helical parameters of T4-A19 base-pairs and T4-A5 steps are quite different between rMD and rMC final models (Figure 2.3). However, the bases of T4-A5 steps from rMD and rMC overlap with an RMSD value of 0.25 Å (Figure 2.5). Visually, the only significant difference between these steps is in the geometry of T4, which is rather distorted in rMD structures.

We conclude that helical parameter differences between rMD and rMC structures mainly derive from internally inconsistent experimental bounds. These inconsistencies distort base rings and promote unusual sugar pucker amplitudes in rMD structures, and local helical parameter perturbations in rMC structures refined with DNAminiCarlo. Moreover, helical parameters in rMD final structures may be ambiguously calculated because of nonflat base geometry.

In spite of all this, rMD and rMC final models overlap with RMSD values well below 1 Å (considering only internal residues, Table 2.4), indicating that the structure of our DNA undecamer duplex depends little on the refinement method used and is essentially defined by experimental bounds. The relatively large number of accurate distance bounds per base (15-20 restraints/base), probably compensates for the smaller restraining power of the wider bounds. In other words, a large number of accurate bounds, defined with lower precision, will lead to a high-quality structure. By superimposing corresponding base-pair steps of rMC and rMD structures, we found that T6-G7, G7-A8, A8-T9 steps are better defined and superimpose with lower RMSD values than the overall structures. The conformation of these residues is defined by more experimental bounds than other regions of the duplex (Figure 2.1), consistent as well with the notion that the experimental data and not the refinement method determines the structure.

Visually analyzing final models. Since we cannot fully rely on analyzing our final structures by their helical parameters, we visually inspected the final models. First, the geometry of terminal base-pairs appears to be distorted, probably due to fraying of the duplex at its ends. In particular, terminal and penultimate residues show a high positive buckle at the 5'-end of the duplex and a high negative buckle at the 3'-end (Figure 2.3). This is also associated with the sugar pucker of terminal residues being shifted towards the N conformation (Figure 2.4) and some unusual backbone torsional angles.

Consequently, the geometry of these residues should not to be used for structural analysis.

Looking at internal residues in structures refined by AMBER, we noticed that all thymidine residues have nonplanar base geometry with the methyl group displaced from the base plane towards the 3' end. The thymine base ring is also oriented with its C5-C6 edge towards the 3' end as if being pulled by the methyl group. Also in adenosine residues, the H8 atom is shifted from the base plane towards the 3' end, but not as much as the methyl group in thymidines. Moreover, the distortion caused by H8 does not seem to change the orientation of the base ring. In rMC structures, thymine and adenine bases are oriented similarly. As expected, the thymine methyl groups and the adenine H8 atoms remain in the base plane, since DNAminiCarlo does not allow deformation of base geometry. As a result of the orientation of thymine and adenine bases, T-A and A-T base-pairs should have negative propeller twist, TA steps positive roll and AT steps close to zero or negative roll.

Guanine bases behave like adenine bases, while cytidines resemble thymidines, with the H5 atom shifted towards the 3' end. However, the G7-C16 base-pair appears flatter than A-T base-pairs. Other GC base-pairs in our sequence are either terminal or penultimate, so their structural features are subject to terminal fraying.

TG steps have positive roll, with T6-G7 manifesting the largest roll. This observation is in accord with the TG geometry in other sequences studied in our laboratory. A big positive roll causes the helix axis to be bent. This feature may be typical for pyrimidine-purine steps, such as TG and TA.

Sugar and backbone conformations. Sugar puckers in the undecamer have P values characteristic of the C2'-*endo* pucker, except for terminal residues, with rMD and rMC structures in good agreement (Figure 2.4). However, sugar pucker amplitudes, Φ, oscillate within a restricted range in Monte Carlo final models (35-40°), but reach unrealistically high values in the rMD structures (up to 50°, Figure 2.4). As already mentioned, this is probably a compromise to accommodate inconsistencies in experimental restraints which is allowed during rMD runs by the AMBER program, but not by DNAminiCarlo in rMC calculations.

Figure 2.4 also shows the sugar conformational parameters, P and Φ, that were determined from deoxyribose vicinal coupling constants (Table 2.1). This information was not used for structural refinement but could be used to cross-validate the final structures. Except for terminal residues, values for P_S, the pseudorotation angle for the major conformer, determined from 2QF-COSY analysis are within 10° of the corresponding values in the refined structure, with the COSY-derived values being systematically larger. The pucker amplitude ranges between 30 and 40°. To calculate the sugar conformation from proton-proton coupling constants, we had to assume a two-state model with the sugar ring undergoing fast S/N conformational exchange. However, rMD and rMC procedures assume a conformational search for only a single conformation with no internal motion. By ignoring conformational exchange, the sugar rings in the refined structures have lower P values.

The backbone conformation of rMC- and rMD-refined structures exhibit similar torsion angles typical of B-DNA. Again, deviations are found in terminal residues and, surprisingly, also for residue T4. This is just another indication of the peculiar nature of this residue.

Thymidine 4 is a special residue. Further inspection of T4 in rMD structures reveals that the methyl group is out of the base plane in the 3'-direction, pulling the C5 atom of the ring in the same direction. At the same time, the base is still engaged in flat hydrogen bonding with A19. The result is that the T4 base has a rather distorted geometry in the final rMD models (Figure 2.5). In rMC structures, T4 has a much flatter base with the methyl group in-plane, but the hydrogen bonds to A19 are perturbed (Figure 2.5). This explains the observed differences between rMD- and rMC-refined structures in most T4-A19 base-pair and some T4-A5 step parameters, and indicates that some restraints acting on T4 are not consistent. Whether this is the result of internal flexibility or experimental error, we do not know.

Analysis of distance restraint violations in the final structures. An indication of internal inconsistencies in our experimental data comes from analysis of bound violations in final structures. For this purpose, we consider models refined by rMD and rMC separately. With the larger number of degrees of freedom permitted for the structure by AMBER, rMD-refined models give lower distance violations than rMC structures. Also,

we noticed that some types of restraints involving sugar protons are consistently violated. In most of these cases, the bounds are shorter than the distances in the final structures. Only a few distances violate the experimental restraint upper bound. Internal flexibility, such as N-S repuckering of the sugar ring, may be responsible for such systematic bound violations. In fact, NOE intensities are averaged by rapid exchange between two local conformations. Moreover, because of the inverse sixth root relationship between distances and intensities, estimated distances will be biased towards the conformer with the shortest distance. Even presence of a small amount (e.g., 10-20%) of a second conformer may significantly affect those experimental bounds with shorter distances in such a minor conformer. Fast N-S sugar repuckering was also suggested by COSY cross-peak analysis.

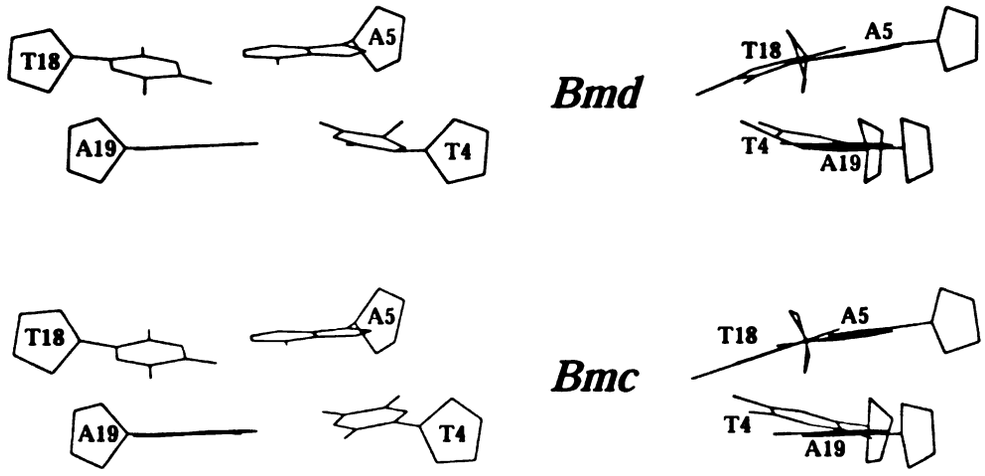


Figure 2.5. Side-by-side views of the T4-A5·T18-A19 base-pair step of the *Bmd* (green) and *Bmc* (red) final structures. While the RMSD between these structures is only 0.25 Å (calculated using only the heavy atoms of the bases), they exhibit significant differences in helical parameters (see Figure 2.3).

However, other experimental bounds, such as H1'-base intraresidue restraints, suggest that sugar repuckering by itself is not enough to fully explain bound violations. H1'-base intraresidue distances in refined models are consistently longer than the corresponding experimental bounds, but this cannot be accounted for by means of sugar flexibility, since H1'-base intraresidue distances do not directly depend on sugar conformation. A more

detailed analysis of experimental distance bounds is being conducted in our laboratory and should provide insights into conformational flexibility which will be the subject of a future paper.

2.5 CONCLUSIONS AND FUTURE DIRECTIONS

Determining solution structures involves several steps ranging from extraction of structural information from NMR spectra to refinement of the structure using such restraints. During each step, variables must be selected and assumptions made. The resulting geometry could be affected by the methods and variables chosen, especially if the experimental data poorly define the structure. In our study, we have tried to minimize the effect of method on the final outcome and, in the process, we have examined the influence of some of the choices on the resulting structure. Interproton distances and bounds were calculated very conservatively from 2D NOE intensities by running MARDIGRAS hundreds of times using 2D NOE spectra for exchangeable protons and for nonexchangeable protons recorded at different mixing times, assuming different overall correlation times and several different starting structures (*vide supra*). As a result, the distance bounds we obtained are wider than for any previous studies conducted in our laboratory, raising the question whether these bounds would have enough restraining power to yield the correct geometry. For this reason, we decided to refine the structure of our molecule using two different methods: restrained molecular dynamics (rMD) and restrained Monte Carlo calculations (rMC). These methods have different conformational search procedures, different empirical force fields, and even a different way of describing the DNA duplex structure (atomic coordinates vs. helical parameters). Nonetheless, rMC and rMD yielded essentially the same structures, with atomic RMSD well below 1 Å between final structures. Hence, we are confident that our final structure is defined by experimental restraints and does not depend significantly on the refinement method or variables used. While convergence to the same final conformation, even using different refinement methods, indicates that the experimental restraints are defining the structure of our duplex, it does not tell us how well this structure describes the real conformation of the molecule in solution. Comparison of the experimental 2D NOE intensities and

intensities calculated for the refined structures via conventional R factors and sixth-root R factors (R^x) is one good criterion indicating that the converged structures fit the experimental NOE data very well.

Surprisingly, refined rMD and rMC structures exhibiting low RMSD revealed significant differences in helical parameter values. Analysis suggests that these discrepancies depend on the effect of bound inconsistencies on structure calculation by the two different methods. In particular, the larger degrees of freedom permitted during rMD causes sugar and base rings to be distorted in accommodating inconsistencies in experimental data. In contrast, this is not possible in the rMC refinement. It was also interesting that helical parameters would often suggest that two structures differ more than what appears from simple visual inspection. It turns out that base distortions, such as those occurring during MD calculations, may impair extraction of helical parameters. Because of this, our paper does not rely much upon helical parameter analysis.

As noted in previously solved NMR structures of DNA duplexes, some helical parameters deviate from those of canonical B-DNA, taking on some characteristics of A-DNA. This quality is even more pronounced in the refined duplex structure reported here, being most obvious in the large negative X-displacement (Figure 2.2). While the current study utilized larger experimental bounds widths than previously employed, calculations using empirically narrowed bounds and no bounds in MD calculations could not conclusively demonstrate that the wider bounds were the source of the stronger A character; we cannot exclude the possibility that our particular duplex sequence plays the key role in defining final overall geometry.

As with the three TG steps in other duplex DNA sequences studied in our laboratory (Weisz et al., 1994; Mujeeb et al., 1993), the two TG steps in the undecamer have a positive roll, with T6-G7 exhibiting the largest roll. A big positive roll causes the helix axis to be bent. This is in accord with the results of a normal vector analysis of DNA duplex structures deposited in the Brookhaven Protein Data Bank (Dickerson, 1998). Dickerson found that roll bending occurs almost exclusively at pyrimidine-purine steps. DNA bending at TG steps has also been held responsible for the unusual electrophoretic mobility shown by duplexes containing such steps (Beutel and Gold, 1992; Nagaich et al., 1994).

In each of the three TG·CA steps we have previously encountered in solution structures determined in our lab (Mujeeb et al., 1993; Weisz et al., 1994), bending was observed, but the T was preceded by a pyrimidine. In the σ^K promoter sequence studied here, we also find roll bending although the TG·CA step has a preceding A and is in the middle of the sequence interrupting an alternating series of AT steps. We might speculate that bending in the free solution structure may confer some advantage in binding of the promoter to the σ^K transcription factor.

In conclusion, we successfully determined a unique structure for our DNA duplex. We believe that our model is an accurate representation of the time-averaged structure of the duplex in solution. Further studies are necessary to elucidate its dynamic properties, which are clearly suggested by our experimental data.

2.6 ACKNOWLEDGMENTS

This work has been published on Biochemistry, Vol. 37, No. 34, 1998. I wish to thank Enzio Ragg for acquiring the NMR spectra, Krystyna Leslak for synthesizing and purifying the DNA strands, Anna Maria Bianucci and Carlos González for help during the early stages of this work, He Liu, Uli Schmitz and Nicolai Ulyanov for useful discussions, and Vladimir Basus for help with running the heteronuclear NMR experiments. M.T. was supported in part by a fellowship from the University of Pisa. Use of the facilities of the Pittsburgh Supercomputing Center for some calculations and the UCSF Computer Graphics Laboratory, supported by NIH Grant RR01081, are gratefully acknowledged.

2.7 REFERENCES

Abragam, A. (1961). Principles of Nuclear Magnetism (Oxford, Oxford University Press).

Adams, B., and Lerner, L. (1992). A Simple One-Dimensional Method for Measuring Proton Exchange Rates in Water. J Magn Reson *96*, 604-607.

Arnott, S., and Hukins, D. W. L. (1972). Optimized Parameters for A-DNA and B-DNA. Biochem Biophys Res Commun *47*, 1504-1509.

Arnott, S., and Hukins, D. W. L. (1973). B-DNA and Implications for the Analysis of X-ray Diffraction Data from Fibers of Biopolymers. J Mol Biol *81*, 93-105.

Bax, A., and Davies, D. G. (1985). MLEV-17-Based Two-Dimensional Homonuclear Magnetization Transfer Spectroscopy. J Magn Reson *65*, 355-360.

Bax, A., and Subramanian, S. (1986). Sensitivity-Enhanced Two-Dimensional Heteronuclear Shift Correlation NMR Spectroscopy. J Magn Reson *67*, 565-569.

Beutel, A. B., and Gold, L. (1992). In Vitro Evolution of Intrinsically Bent DNA. J Mol Biol *228*, 803-812.

Boelens, R., Scheek, R. M., Dijkstra, K., and Kaptein, R. (1985). Sequential Assignment of Imino- and Amino-Proton Resonances in ^{1}H NMR Spectra of Oligonucleotides by Two-Dimensional NMR Spectroscopy Application to a *lac* Operator Fragment. J Magn Reson *62*, 378-386.

Borgias, B. A., and James, T. L. (1988). COMATOSE: A Method for Constrained Refinement of Macromolecular Structure Based on Two-Dimensional Nuclear Overhauser Effect Spectra. J Magn Reson *79*, 493-512.

Borgias, B. A., and James, T. L. (1990). MARDIGRAS--Procedure for Matrix Analysis of Relaxation for DIscerning Geometry of an Aqueous Structure. J Magn Reson *87*, 475-487.

Braunschweiler, L., and Ernst, R. R. (1983). Coherence Transfer by Isotropic Mixing: Application to Proton Correlation Spectroscopy. J Magn Reson *53*, 521-528.

Celda, B., Widmer, H., Leupin, W., Chazin, W. J., Denny, W. A., and Wüthrich, K. (1989). Conformational Studies of d-(AAAAATTTTT)$_2$ Using Constraints from Nuclear Overhauser Effects and from Quantitative Analysis of the Cross-Peak Fine Structures in Two-Dimensional ^{1}H Nuclear Magnetic Resonance Spectra. Biochemistry *28*, 1462-1470.

Dickerson, R. E. (1998). DNA Bending: The Prevalence of Kinkiness and the Virtues of Normality. Nucl Acid Res *26*, 1906-1926.

Feigon, J., Denny, W. A., Leupin, W., and Kearns, D. R. (1983). Proton Nuclear Magnetic Resonance Investigation of the Conformation and Dynamics in the Synthetis Deoxyribonucleic Acid Decamers d(ATATCGATAT) and d(ATATGCATAT). Biochemistry *22*, 5930-5942.

Gallo, K., Huang, C., Ferrin, T. E., and Langridge, R. (1989). Molecular Interactive Display and Simulation (MIDASplus) (San Francisco, University of California, San Francisco).

Gonzàlez, C., Stec, W., Kobylanska, A., Hogrefe, R., Reynolds, M., and James, T. L. (1994). Structural Study of a DNA-RNA Hybrid Duplex with a Chiral Phosphorothioate Moiety by NMR: Extraction of Distance and Torsion Angle Constraints and Imino Proton Exchange Rates. Biochemistry *33*, 11062-11072.

Gonzàlez, C., Stec, W., Reynolds, M., and James, T. L. (1995). Structure and Dynamics of a DNA-RNA Hybrid Duplex with a Chiral Phosphorothioate Moiety. NMR and Molecular Dynamics with Conventional and Time-averaged Restraints. Biochemistry *34*, 4969-4982.

Gorin, A. A., Ulyanov, N. B., and Zhurkin, V. B. (1990). S-N transition of the sugar ring in B-form DNA. Molek Biol (Eng trans) *24*, 1036-1047.

Harrison, S. C., and Aggarwal, A. K. (1990). DNA Recognition by Proteins with the Helix-Turn-Helix Motif. Ann Rev Biochem *59*, 933-969.

James, T. L. (1991). Relaxation Matrix Analysis of 2D NOE Spectra to Obtain Accurate Biomolecular Structural Constraints and to Assess Structural Quality. Curr Opin Struct Biol *1*, 1042-1053.

Keepers, J. W., and James, T. L. (1984). A Theoretical Study of Distance Determinations from NMR. Two-Dimensional Nuclear Overhauser Effect Spectra. J Magn Reson *57*, 404-426.

Koudelka, G. B., Harrison, S. C., and Ptashne, M. (1987). Effect of non-contacted bases on the affinity of 434 operator for 434 repressor and Cro. Nature *326*, 886-889.

Lavery, R., and Sklenar, H. (1990). CURVES 3.0. Helical Analysis of Irregular Nucleic Acids (Paris, Laboratory for Theoretical Biology, CNRS).

Liu, H., Kumar, A., Weisz, K., Schmitz, U., Bishop, K. D., and James, T. L. (1993). Extracting Accurate Distances and Bounds from 2D NOE Exchangeable Proton Peaks. J Am Chem Soc *115*, 1590-1591.

Liu, H., Spielmann, H. P., Ulyanov, N. B., Wemmer, D. E., and James, T. L. (1995). Interproton Distance Bounds from 2D-NOE Intensities: Effect of Experimental Noise and Peak Integration Errors. J Biomolec NMR *6*, 390-402.

Liu, H., Thomas, P. D., and James, T. L. (1992). Averaging of Cross-Relaxation Rates and Distances for Methyl, Methylene and Aromatic Ring Protons due to Motion or Overlap: Extraction of Accurate Distances Iteratively via Relaxation Matrix Analysis of 2D NOE Spectra. J Magn Reson *98*, 163-175.

Liu, H., Tonelli, M., and T.L.James (1996). Correcting NOESY Cross-Peak Intensities for Partial Relaxation Effects Enabling Accurate Distance Determination. J Magn Reson *111, Series B*, 85-89.

Marion, D., and Wüthrich, K. (1983). Application of Phase Sensitive Two-Dimensional Correlated Spectroscopy (COSY) for Measurements of ^{1}H-^{1}H Spin-Spin Coupling Constants in Proteins. Biochem Biophys Res Commun *113*, 967-974.

Mujeeb, A., Kerwin, S. M., Kenyon, G. L., and James, T. L. (1993). Solution Structure of a Conserved DNA Sequence from the HIV-1 Genome:Restrained Molecular Dynamics Simulation with Distance and Torsion Angle Restraints Derived from 2D NMR Spectra. Biochemistry *32*, 13419-13431.

Nagaich, A. K., Bhattacharyya, D., Brahmachari, S. K., and Bansal, M. (1994). CA/TG Sequence at the 5' End of Oligo(A)-tracts Strongly Modulates DNA Curvature. J Biol Chem *269*, 7824-7833.

Nirmala, N. R., and Wagner, G. (1988). Measurement of ^{13}C Relaxation Times in Proteins by Two-Dimensional Heteronuclear ^{1}H-^{13}C Correlation Spectroscopy. J Am Chem Soc *110*, 7557-7558.

Pearlman, D. A., Case, D. A., Caldwell, J. C., Seibel, G. L., Singh, U. C., Weiner, P., and Kollman, P. A. (1990). AMBER, version 4.0 (San Francisco, University of California, San Francisco).

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(CGCGAATTCGCG) double helix: "curves", dials and windows. J Biomolec Struct Dyn *6*, 669-699.

Rinkel, L. J., and Altona, C. (1987). Conformational Analysis of the Deoxyribofuranose Ring in DNA by means of Sums of Proton-proton Coupling Constants: A Graphical Analysis. J Biomol Struct Dyn *4*, 621-649.

Saenger, W. (1984). Principles of Nucleic Acid Structure (New York, Springer).

Scheek, R. M., Russo, N., Boelens, R., Kaptein, R., and Boom, J. H. v. (1983). Sequential Resonance Assignments in DNA ^{1}H NMR Spectra by Two-Dimensional NOE Spectroscopy. J Am Chem Soc *105*, 2914-2916.

Schmitz, U., and James, T. L. (1995). How to Generate Accurate Solution Structures of Double-Helical Nucleic Acid Fragments Using Nuclear Magnetic Resonance and Restrained Molecular Dynamics. In Methods in Enzymology, Nuclear Magnetic Resonance and Nucleic Acids, T. L. James, ed. (New York, Academic Press), pp. 3-44.

Schmitz, U., Zon, G., and James, T. L. (1990). Deoxyribose Conformation In $[d(GTATATAC)]_2$: Evaluation of Sugar Pucker by Simulation of Double-Quantum-Filtered COSY Cross-Peaks. Biochemistry *29*, 2357-2368.

Stolarski, R., Egan, W., and James, T. L. (1992). Solution Structure of the *Eco RI* DNA Octamer Containing 5-Fluorouracil via Restrained Molecular Dynamics Using Distance and Torsion Angle Constraints Extracted from NMR Spectral Simulations. Biochemistry *31*, 7027-7042.

Ulyanov, N. B., and James, T. L. (1994). Statistical Analysis of DNA Duplex Structures in Solution Derived by High Resolution NMR. Appl Magn Reson *7*, 21-42.

Ulyanov, N. B., Schmitz, U., and James, T. L. (1993). Metropolis Monte Carlo Calculations of DNA Structure using Internal Coordinates and NMR Distance Restraints: An Alternative Method for Generating High-Resolution Solution Structure. J Biomolec NMR *3*, 547-568.

von Hippel, P. H., Bear, D. G., Morgan, W. D., and McSwiggen, J. A. (1984). Protein-nucleic acid interactions in transcription: a molecular analysis. Ann Rev Biochem *53*, 389-419.

Weisz, K., Shafer, R. H., Egan, W., and James, T. L. (1992). The Octamer Motif in Immunoglobulin Genes: Extraction of Structural Constraints from Two-Dimensional NMR Studies. Biochemistry *31*, 7477-7487.

Weisz, K., Shafer, R. H., Egan, W., and James, T. L. (1994). Solution Structure of the Octamer Motif in Immunoglobulin Genes via Restrained Molecular Dynamics Calculations. Biochemistry *33*, 354-366.

Widmer, H., and Wüthrich, K. (1986). Simulation of Two-Dimensional NMR Experiments Using Numerical Density Matrix Calculations. J Magn Reson *70*, 270-279.

Zheng, L., Halberg, R., Roels, S., Ichikawa, H., Kroos, L., and Losick, R. (1992). Sporulation regulatory protein GerE from Bacillus subtilis binds to and can activate or repress transcription from promoters for mother-cell-specific genes. J Mol Biol *226*, 1037-1050.

2.8 SUPPLEMENTARY MATERIAL

Supplement table 2.7. Non-exchangeable proton chemical shifts (ppm) of d(GCATATGATAG) ·d(CTATCATATGC) at 25°C. [a]Not assigned.

residue	H8/6	H5	CH_3	H2	H1'	H2'	H2"	H3'	H4'	H5'	H5"
G1	7.96				5.99	2.63	2.80	4.86	4.25	3.73	3.73
C2	7.48	5.44			5.70	2.19	2.51	4.91	*a*	*a*	*a*
A3	8.36			7.66	6.31	2.74	3.00	5.06	4.47	4.23	4.14
T4	7.19		1.51		5.70	2.11	2.50	4.90	*a*	*a*	*a*
A5	8.25			7.16	6.21	2.61	2.91	5.02	4.43	*a*	*a*
T6	7.06		1.35		5.70	1.95	2.38	4.87	*a*	*a*	*a*
G7	7.82				5.60	2.65	2.76	5.01	4.35	4.14	4.07
A8	8.17			7.70	6.20	2.60	2.89	5.00	4.46	*a*	*a*
T9	7.14		1.41		5.58	1.93	2.29	*a*	4.12	*a*	*a*
A10	8.16			7.44	6.07	2.67	2.86	5.03	4.40	4.14	4.07
G11	7.69				6.02	2.43	2.28	4.64	4.19	4.14	4.27
C12	7.85	5.93			5.89	2.20	2.58	4.68	4.11	3.81	3.81
T13	7.60		1.73		5.79	2.31	2.61	4.94	4.24	4.10	4.05
A14	8.40			7.47	6.33	2.74	3.00	5.06	4.48	4.24	4.17
T15	7.18		1.41		5.90	2.06	2.47	4.87	4.22	*a*	*a*
C16	7.52	5.62			5.62	2.14	2.45	4.86	4.16	*a*	*a*
A17	8.29			7.47	6.21	2.66	2.94	5.01	4.42	4.18	4.12
T18	7.17		1.45		5.67	2.11	2.49	4.88	4.20	*a*	*a*
A19	8.24			7.19	6.22	2.62	2.92	5.01	4.42	*a*	*a*
T20	7.10		1.37		5.76	1.98	2.38	4.87	*a*	*a*	*a*
G21	7.86				5.92	2.60	2.70	4.98	4.37	*a*	*a*
C22	7.46	5.41			6.20	2.19	2.19	4.51	4.07	*a*	*a*

Supplement table 2.8. Exchangeable proton chemical shifts (ppm) of d(GCATATGATAG) ·d(CTATCATATGC) at 10°C. [a] Not assigned.[b] $NH_2(1)$ and $NH_2(2)$ denote non-hydrogen and hydrogen bonded amino protons.

Residue	NH[a]	$NH_2(1)$[b]	$NH_2(2)$[b]
G1	12.89	*a*	*a*
C2		8.26	6.41
A3		*a*	6.16
T4	13.08		
A5		*a*	*a*
T6	13.24		
G7	12.09	7.05	*a*
A8		*a*	5.79
T9	13.32		
A10		7.69	6.26
G11	12.99	*a*	*a*
C12		7.72	6.95
T13	13.58		
A14		*a*	6.35
T15	13.34		
C16		8.19	6.62
A17		7.48	6.12
T18	13.06		
A19		*a*	*a*
T20	13.45		
G21	12.59	*a*	*a*
C22		8.02	6.49

Supplement table 2.9. Vicinal proton coupling constants (in Hz) of d(GCATATGATAG)·d(CTATCATATGC) from quantitative simulations of DQF-COSY cross-peaks. $^3J_{2'2''}$ was assumed to be -14.0 ± 0.5 Hz. [a]Uncertainty ±0.3 Hz. [b]Uncertainty ±0.5 Hz.

residue	$^3J_{1'2'}$[a]	$^3J_{1'2''}$[a]	$^3J_{2'3'}$[b]	$^3J_{2''3'}$[b]
G1	9.1	5.7	4.9	2.4
T6	8.9	6.6	7	2.3
G7	9.4	6	5	2.4
T9	8.7	6.6	7	2.6
A10	9.5	5.4	4.9	2.3
G11	8.5	6.2	6.5	3.1
C12	5.4	6.8	6.6	5.5
T13	9.8	5.8	5.7	2.1
T15	8.8	5.9	7.1	3.4
C16	8.4	6.6	6.8	2.3
T20	8.5	5.9	7	3.4
G21	9.2	5.6	5.2	2.5

CHAPTER 3

3. Insights into the Dynamic Nature of DNA Duplex Structure Via analysis of Nuclear Overhauser Effect Intensities

Sequence-dependent structures of DNA duplexes in solution can be reliably determined using NMR data if care is taken to determine restraint bounds accurately. This entails use of complete relaxation matrix methods to analyze multidimensional nuclear Overhauser effect (NOE) spectroscopic cross-peak intensities yielding accurate distance restraints. Studies of various DNA duplexes by NMR have suggested that there may be some limited internal motions. First, it is typically not possible to reconcile all vicinal proton coupling constants in the deoxyribose ring with a single conformer. In addition, with the increased accuracy of interproton distance measurements afforded by the complete relaxation matrix algorithm MARDIGRAS, we find certain inconsistencies in distances which can most readily be ascribed to limited conformational flexibility, since conformational averaging is nonlinear. As base-sugar interproton distances depend on both sugar pucker and glycosidic torsion angle χ, any motion involving these structural variables should also be reflected by the experimental data. Possible motional models have been considered to account for all of the data for three DNA duplexes. Analysis of intraresidue base-sugar interproton NOE bounds patterns suggests a motional model with individual sugars in equilibrium between S (2'-*endo*) and N (3'-*endo*) conformations, with S being the preferred conformer. As sugar repuckering is correlated with changes in glycosidic torsion angle χ, different sugar conformers imply different values for χ, but this is insufficient to account for all data. While a possible two-state jump between *anti*

Abbreviations. 1D NMR, one-dimensional NMR; 2D NMR, two-dimensional NMR; 2D NOE, two-dimensional nuclear Overhauser effect; 2QF-COSY, double-quantum-filtered correlation spectroscopy; FID, free induction decay; rMD, restrained molecular dynamics; rMC, restrained Monte Carlo calculations; MD-tar, molecular dynamics with time-averaged restraints; RMSD, root-mean-square deviation; R^x, sixth-root R-factor; P, pseudorotation phase angle; S^2, order parameter.

and *syn* glycosidic conformers was considered, it was found to be incapable of accounting for all data. However, a model with restricted diffusion (rocking) about the glycosidic bond in addition to sugar repuckering was capable of accommodating all experimental data. This motional model is in qualitative agreement with experimental ^{13}C relaxation-derived order parameter values in a DNA duplex.

3.1 INTRODUCTION

The nuclear Overhauser effect (NOE) has been widely used to investigate the three-dimensional (3D) structure of biologically relevant macromolecules in solution. In a typical two-dimensional NOE (2D NOE) experiment, the dipole moments of two hydrogen atoms within a distance of *ca.* 5 Å will couple, yielding an off-diagonal peak in the spectrum. To a first approximation, the intensity of this cross-peak is inversely proportional to the sixth power of the distance between the two atoms involved. NOE intensities are often used to calculate distances between hundreds to thousands of pairs of hydrogen atoms; these distances are then used to restrain the structure during structural refinement procedures.

NOE intensities may also be affected by conformational fluctuations that occur during the time of the NMR experiment. In the case of exchange faster than the largest chemical shift difference of a particular proton in the various exchanging conformers (typically >1000 Hz), only one set of NOE peaks is observed in the spectrum. The intensity of these peaks is a weighted average of the intensities given by the atoms in each of the conformers. Complicating the situation, the manner in which NOE intensities are averaged depends (a) on the relative population of conformers, (b) on the rate of exchange between conformers, which can be alternatively viewed as internal motion, and (c) on the angular fluctuations if the exchange is extremely fast.

For purposes of structure determination, a single rigid conformation is assumed, so a single distance corresponding to each NOE intensity is naturally derived. The averaging in any case is obviously not a simple geometric average of distances in the different conformers. When the internal motion is faster than the overall correlation time for molecular tumbling, the measured distance is subject to $\langle r^{-3} \rangle$ averaging and depends as

well on angular dispersion; in this case, the apparent distance between a pair of nuclei is generally slightly weighted toward the shortest of the distances in the various conformers if the internal motion (conformational exchange) is not isotropic (Keepers and James, 1984), but the apparent distance is identically equal to the geometric mean distance, weighted by conformer populations, if by chance the conformational exchange entails isotropic jumps (LeMaster et al., 1988) . However, if jumps between the individual conformers are slow compared to the reciprocal of the overall correlation time but fast compared to the relaxation time, the apparent distance is subject to $\langle r^{-6} \rangle$ averaging over individual proton positions; so the apparent distance may be strongly distorted from the geometric average position.

It will be recognized that a Karplus relationship between vicinal scalar coupling constants and torsion angles is also not linear, so torsion angles are not geometrically averaged by conformational fluctuations. Even with perfect data and with careful analysis, with such nongeometric averaging, the individual distance and torsion angle restraints will be internally inconsistent. However, inconsistencies in NOE-derived distances encode information about the on-going conformational equilibrium and, in principle, could be used to study it.

To infer anything from sets of inconsistent distances derived from NOE intensities, it is essential that they be determined with high accuracy. Our distances are calculated by use of the program MARDIGRAS (Borgias and James, 1990; Liu et al., 1993; Liu et al., 1995), which is based on complete relaxation matrix analysis. Consequently, MARDIGRAS is able to account for network relaxation and multispin effects and, hence, to calculate proton-proton distances with greater accuracy than the commonly employed two-spin or isolated spin-pair approximation.

Previous NMR studies conducted on DNA duplexes show that certain classes of distances are inconsistent with others (Ulyanov et al., 1995). Indeed, some are inconsistent with the duplexes possessing structure in either the B or A family of structures, i.e., those which one would imagine are energetically most reasonable. We have also observed that a single conformer will rarely account for all coupling constants measured via detailed simulation and fitting, but reasonable fits are obtained with a two-

state model representing a rapid interconversion between S- and N-type sugar puckers (Rinkel and Altona, 1987; Schmitz and James, 1995).

The current investigation examines possible causes for these inconsistencies, focusing on the likelihood that internal motions are responsible. In our studies, we will consider inconsistencies involving intraresidue base-to-sugar distances. Consequently, we will be concerned with flexibility involving: (a) sugar pucker, which is described by the phase angle **P** and the amplitude Φ of pseudorotation, and (b) rotation about the glycosyl bond C1'-N (described by the torsion angle χ) within the same residue (Figure 3.1). This entails a relatively simple subset of conformational space, whose members can be almost fully described by two variables, **P** and χ angles, with the sugar pucker amplitude Φ being allowed to vary only within a narrow range. Furthermore, from an NMR point of view, it is a well-defined system, with 6-9 hydrogen atoms depending on which residue we are studying. Interactions among these atoms give NOE cross-peaks that can be used for our structural studies.

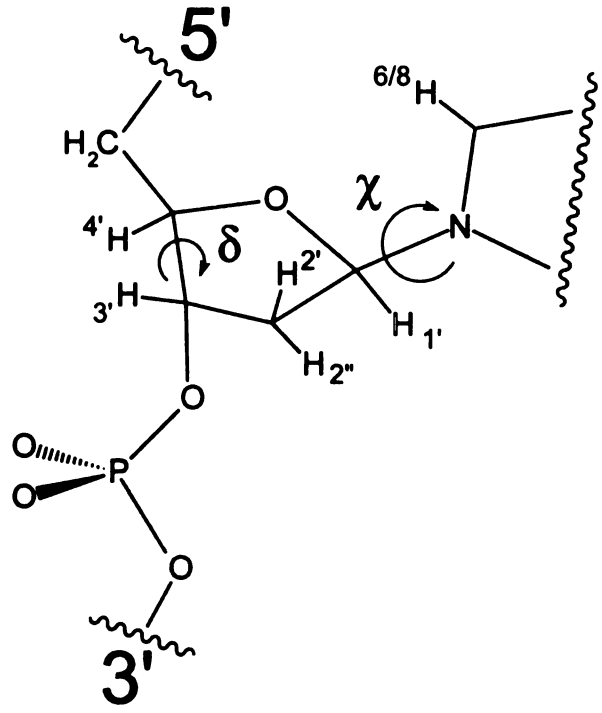


Figure 3.1. A portion of a deoxyribose nucleotide with labeling of the base and sugar hydrogen atoms and defining torsion angles δ and χ.

Two principal low-energy sugar conformations have been described in nucleotide structures: C3'-*endo* with $0° < \mathbf{P} < 36°$ (in the "north" of the **P** circle, or N) and C2'-*endo* with $144° < \mathbf{P} < 190°$ ("south", or S). As for the χ angle, crystallographic (Saenger, 1984) and NMR data (Wüthrich, 1986) indicate that the base can adopt two orientations relative to the sugar moiety, called *syn* (χ angle values 190-270°) and *anti* (40-90°). A good correlation between the glycosidic angle χ and the torsional angle δ, which is strongly correlated with the sugar pucker, has also been demonstrated (Fratini et al., 1982).

Since base-sugar interproton distances depend on both sugar pucker and χ torsion angle, any motion involving these structural variables should also be reflected by the experimental data. Thus, to justify base-sugar interproton distances fully, we need to consider the effect of sugar repuckering and glycosyl bond flexibility simultaneously. This approach offers new insights for DNA structure analysis, since so far, sugar repuckering and χ flexibility have mostly been considered independently.

3.2 MATERIALS AND METHODS

3.2.1 Interproton Distances Derived from 2D NOE Spectra.

NOE-derived distance bounds from three independent DNA duplexes are considered: (I) the Pribnow box octamer with sequence d(GTATAATG)·d(CATTATAC) (Schmitz et al., 1992); (II) the octamer motif found in both the promoter and enhancer regions in immunoglobulin genes which is contained in a DNA decamer with sequence d(CATTTGCATC)·d(GATGCAAATG) (Weisz et al., 1994); and (III) the σ^K consensus sequence undecamer d(GCATATGATAG)·d(CTATCATATGC) (Tonelli et al., 1998).

All proton-proton distance bounds for these three DNA duplexes were calculated from 2D NOE experiments by MARDIGRAS (Borgias and James, 1990; Liu et al., 1993; Liu et al., 1995) and were previously reported; the protocols used to determine these distances were also described. It will be noted that the bounds for the σ^K consensus sequence are broader. This is due to a very conservative approach used in determination

of bounds; indeed, to account for integration error and spectral noise, the RANDMARDI version of the software has been utilized (Liu et al., 1995).

Of all the distances that were calculated and used for structure refinement, only base-to-sugar intraresidue bounds are analyzed in the present paper. In particular, we focus our attention on: H6/8-H1', H6/8-H2', H6/8-H2", H6/8-H3' and H6/8-H4' intraresidue distances. The H5-sugar proton distances of cytosines and methyl-sugar proton distances in thymine residues manifest a similar behavior as H6/8-sugar, but will not be discussed further in this paper.

3.2.2 Molecular Models.

Distances corresponding to those measured from 2D NOE data were obtained from various molecular models, including canonical DNA structures (A- and B-DNA). These canonical models were generated by AMBER (Cornell et al., 1995). The canonical models were also used as starting models for MARDIGRAS calculations and structure refinement as described in the original papers.

To generate a larger pool of conformers capable of accommodating most of the experimental NMR data, we utilized molecular dynamics with weighted time-averaged restraints (MD-tar). MD-tar permits a broader search of conformational space in a restrained molecular dynamics (rMD) simulation, as the calculation does not demand that the measured value of the distance be satisfied exactly at every single step of the simulation as in standard rMD calculations (Torda et al., 1989). MD-tar simulations of the octamer pribnow box have been described already (Schmitz et al., 1993). The MD-tar simulations result in a very large number of structures each of which may fit a different subset of the experimental interproton distance restraints. Monte Carlo simulations were also carried out on a DNA dimer with sequence d(CA)·d(TG) with the program DNAminiCarlo (Ulyanov et al., 1992; Zhurkin et al., 1982). No restraints were used in the Monte Carlo simulations in order to produce some structures broadly covering conformational space. The expanded pool of conformers are used in the present study to explore the conformational space accessible to duplex DNA – to examine whether the reproducibly inconsistent experimental distances can be reconciled by conformational exchange.

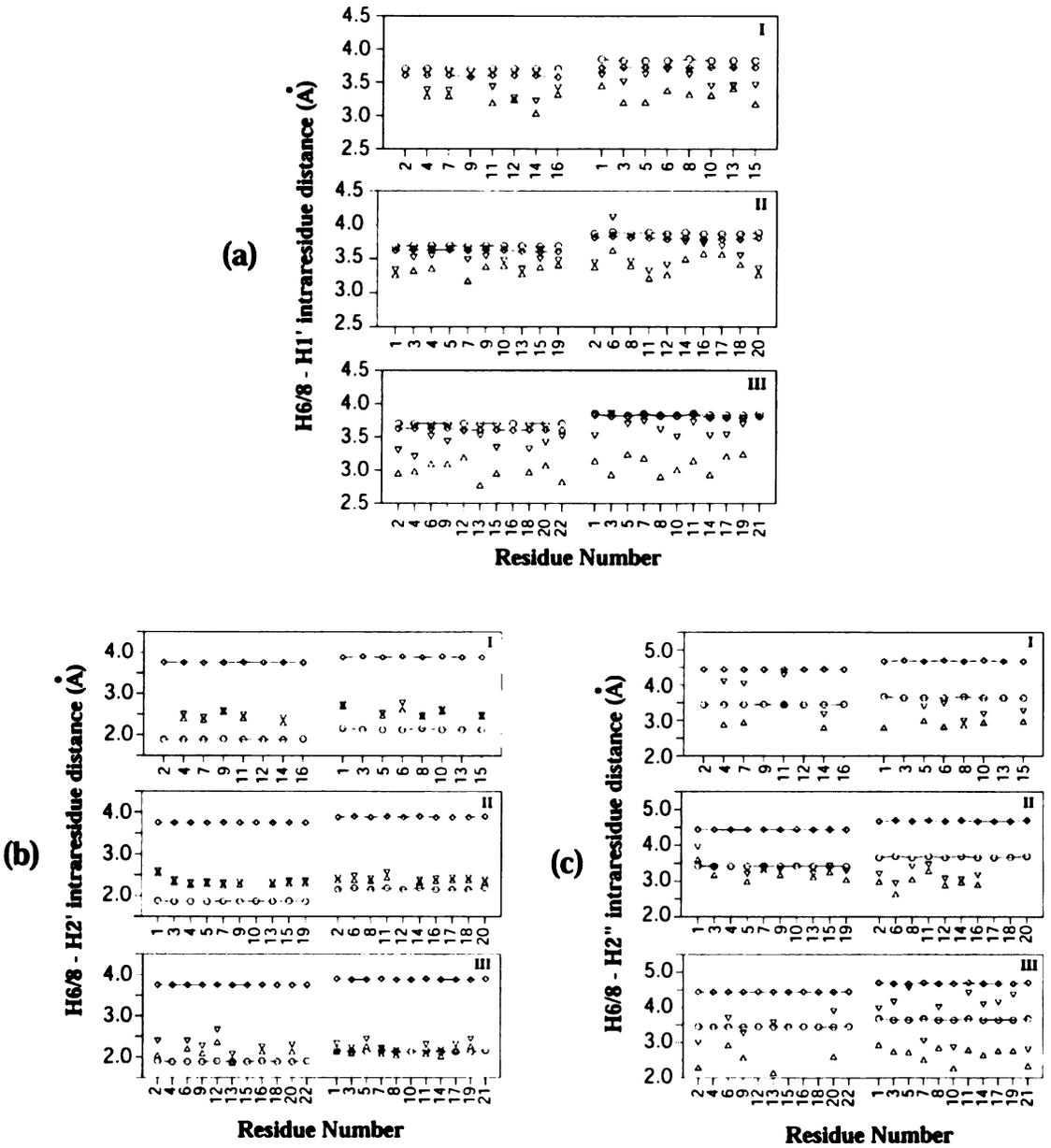
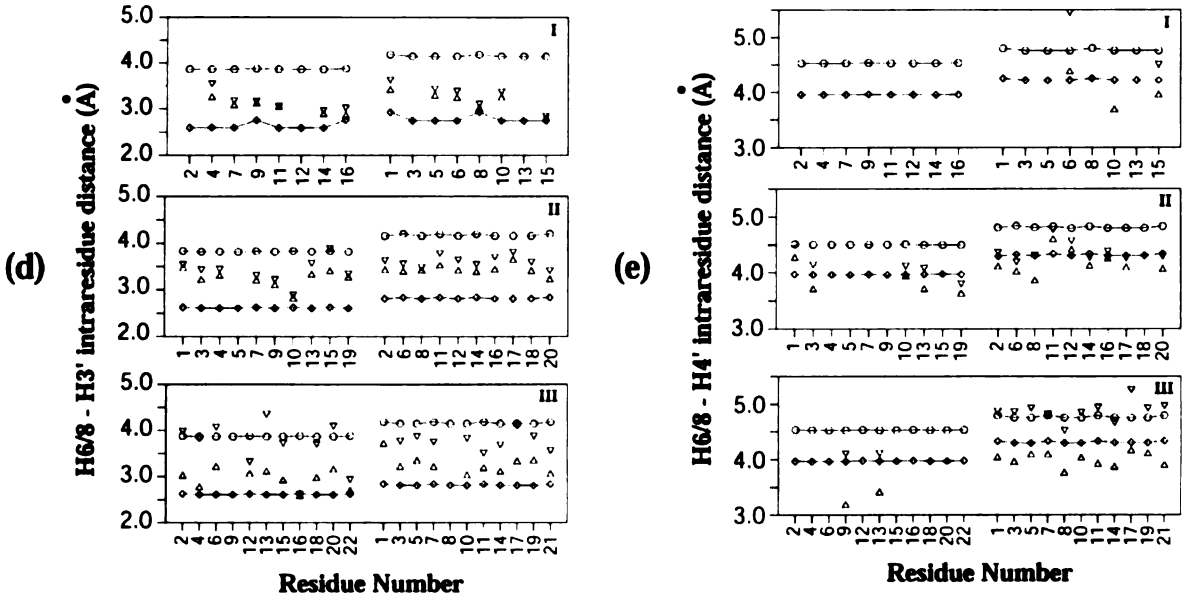


Figure 3.2. NOE distance bounds (lower Δ, upper □) calculated by MARDIGRAS and the corresponding distances from standard A-DNA (-◊-) and B-DNA (-o-) for the pribnow box octamer (I) d(GTATAATG)· d(CATTATAC), immunoglobulin gene decamer (II) d(CATTTGCATC)·d(GATGCAAATG), and σ^K consensus sequence undecamer (III) d(GCATATGATAG)·d(CTATCATATGC). H6/8-H1' (a), H6/8-H2' (b), H6/8-H2" (c), H6/8-H3' (d) and H6/8-H4' (e) intraresidue bounds and standard distances are shown (next page).

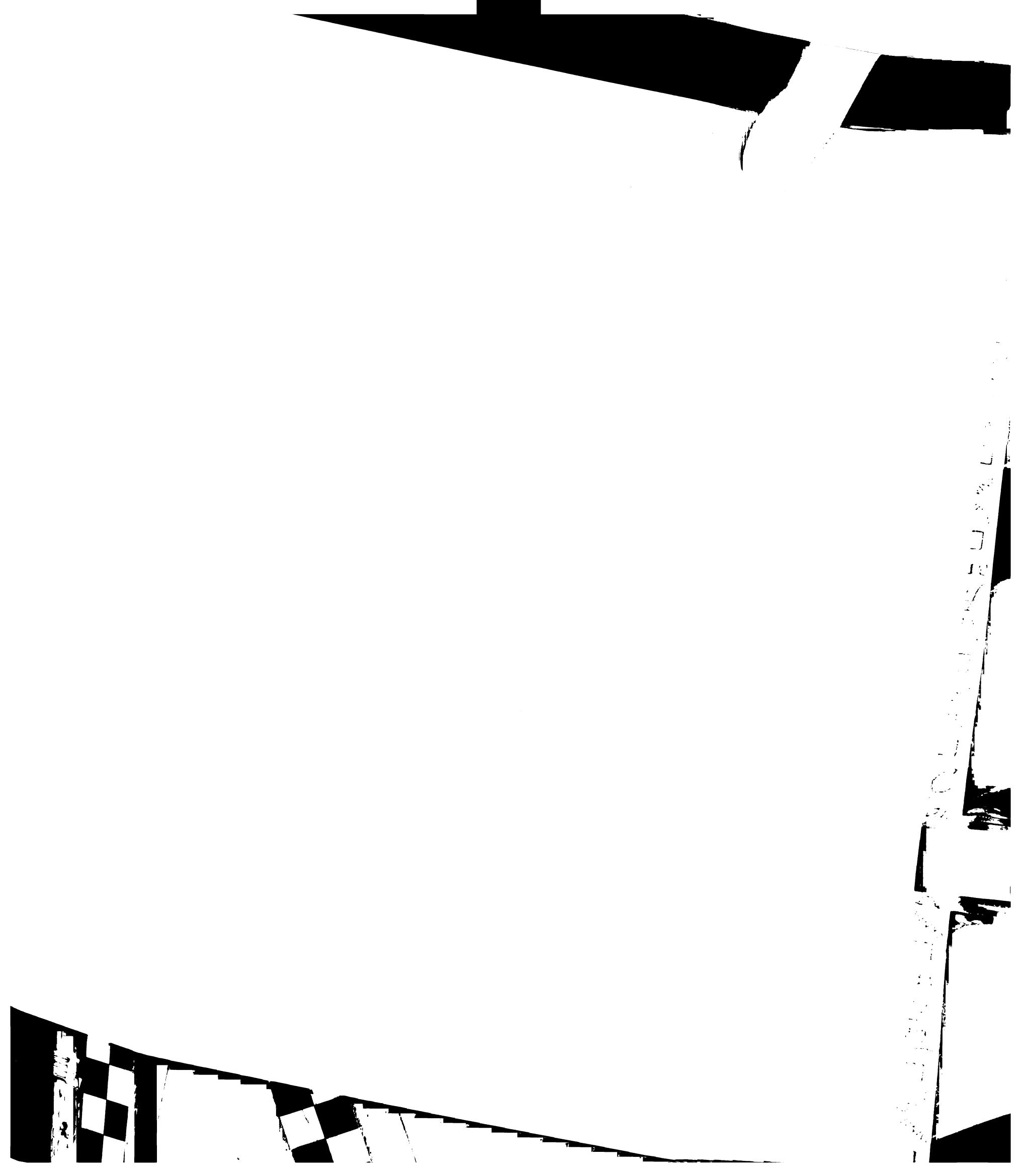


3.3 RESULTS AND DISCUSSION

In this section we analyze the interproton distance bounds for three different DNA duplexes previously investigated in our lab. It is evident in Figure 3.2 that the bounds for the σ^K consensus sequence undecamer (duplex III) were determined in an extremely conservative fashion (Tonelli et al., 1998), taking into account numerous potential errors ranging from uncertainties in overall rotational correlation time, to peak integration and spectral noise contributions (Liu et al., 1995). We will formulate two plausible motional models entailing motions about the glycosyl bond to explain the data. Further examination of the NOE bounds and MD-tar simulation data will lead us to favor one of these models. This inclination will be further strengthened by analysis of the pool of structures generated by Monte Carlo calculations.

3.3.1 Base-to-H2' and base-to-H3' distances derived from 2D NOE data indicate sugar repuckering motions.

First, let us consider H6/8-H2' and H6/8-H3' NOE distance bounds. These distances are dramatically different in A- and B-DNA, since they are strongly correlated with sugar pucker: (a) H6/8-H2' is about 1.7 Å shorter in B-DNA, and (b) H6/8-H3' is about 1.4 Å shorter in A-DNA.



By comparing the experimental NOE data to the canonical A- and B-DNA models, it is evident that these distances contradict each other: base-H2' bounds are consistently somewhat longer than standard B-DNA but not too close to A-DNA (Figure 3.2b), while base-H3' distance bounds are consistently shorter than found in B-DNA and closer to A-DNA (Figure 3.2d) than are base-H2' bounds. This apparent contradiction can be justified by the presence of an on-going conformational equilibrium with fast exchange between N and S sugar puckers. NOE data also indicate that S is the favored conformation. As noted earlier, the presence of a S/N sugar repuckering equilibrium is also suggested by scalar-coupling based analysis (Lane, 1993; Schmitz et al., 1990). To understand how repuckering can lead to the apparent inconsistencies in the NOE distance bounds, we need to consider two factors affecting the NOE bounds, i.e., position of the equilibrium and actual distances in each conformer with a bias towards the conformers with shorter values. Thus, base-H2' bounds are longer and closer to standard B-DNA distances (S sugar pucker) because the major conformer has S pucker and a shorter base-H2' distance; base-H3' bounds, on the other hand, are shorter than B-DNA and further shifted towards A-DNA, because the N pucker has the shortest base-H3' distance.

We have been able to reproduce base-H2' and base-H3' experimental data by back-calculation of the NOE intensities using a mixture of A- and B-DNA model structures (Donati and James, unpublished results). This was done with the algorithm used in the program CORMA (Keepers and James, 1984) that can generate NOE intensities for interconverting mixtures of up to 5000 different conformations. We note also that we have developed another method, termed PARSE, that generates an ensemble of conformers with an assessment of the probability of each conformer, using cross-relaxation rates simulated by CORMA, by consideration of a much wider pool of conformers which includes structures resembling A- and B-DNA (Ulyanov et al., 1995). We have found in simulated data sets that PARSE is able to select the few correct interconverting conformers from a much wider pool (hundreds) of conformers (Ulyanov et al., 1998).

Similar considerations can be made for base-H2" and base-H4' NOE bounds. However, these distances seem to be affected by other factors as well, so we will discuss them later in the paper.

3.3.2 Base-to-H1' distances are shorter than in either A- or B-DNA: two possible motional models.

Since the intraresidue H6/8-H1' distance is a direct function of the glycosidic torsion angle, we expect the value of this distance bound to reflect any motion about the glycosidic bond. Canonical DNA structures, A and B, have similar H6/H8-H1' distances (3.4 Å for pyrimidines, 3.7 Å for purines in both A- and B-DNA), even with χ torsion angles differing by about 60°. Surprisingly, NOE bounds are consistently shorter than either A- and B- DNA distances (Figure 3.2a). To understand this observation, we must consider the dependence of this distance on the torsion angle χ (Wüthrich, 1986). The H6/H8-H1' distance varies sinusoidally as the torsion angle χ is changed, such that the H6/H8-H1' distance in canonical A- and B-DNA lie on either side of the top of a sinusoidal curve. Thus, changing the χ angle to a value between that of standard A- and B-DNA results in a shorter H6/H8-H1' distance.

On this basis, two plausible internal motion models can be proposed to explain the experimental data: (1) the jump diffusion model, which describes the effect of a jump between two stable conformations (*syn* and *anti*); (2) the restricted diffusion model, which allows restricted rotation about the glycosidic bond between the angles $\chi \pm \Delta\chi$, where χ is a value in the *anti* conformational region. In the first model, the *syn* conformer has a much shorter base-H1' distance (about 1.3 Å shorter) than the *anti*. In the restricted diffusion model, rocking about the glycosyl bond produces conformers with shorter base-H1' distance. In either case, because of the distance-weighting of NOE intensities, the NOE-derived distances will be shorter than either A- or B-DNA in accord with experimental data.

Moreover, neither of these two models is physically unreasonable. Experimental data from x-ray crystallography and NMR indicate that both *anti* and *syn* conformers can be present in oligonucleotides. On the other hand, Drew et al. (Drew et al., 1981) reported that in many cases the sugar ring appears to be rocking about the glycosidic bond. Establishing whether one or the other of these motions is occurring in our DNA duplexes in solution is our task in this paper.

Assuming that distances are averaged as the sixth-root by fast motion (*vide supra*): (1) for the two-state jump diffusion model, we calculated the fraction of the *syn* conformer that needs to be present at equilibrium in order to explain the experimental data; and (2) for the restricted diffusion model, assuming a normal distribution of the χ torsion angle with an average value of 255° for purines and 235° for pyrimidines (typical of B-DNA), we calculated how broad this distribution should be in order to shorten base-H1' distances by the observed amount. We found that the experimental NOE data can be satisfied by: (1) less than 10% of *syn* conformer at equilibrium for either purines or pyrimidines (Figure 3.3a), or (2) a normal distribution with a standard deviation of 30-40° for purines and even higher for pyrimidines (50-60°) (Figure 3.3b).

While the percent of *syn* conformer calculated is in agreement with what has been previously estimated (Gaudin et al., 1995), the standard deviation values we found are a bit larger than expected. This may be the result of over-simplification in our calculations. We reserve discussion of this issue in more detail later in the paper. At this point let us just conclude that analysis of the H6/8-H1' NOE distances alone do not allow us to favor unequivocally one model over the other.

3.3.3 Consideration of all intraresidue base-sugar distances.

So far, we have only analyzed H6/H8-H2', H6/H8-H3' distances and H6/H8-H1' distances independently. Now, let us consider all the base-to-sugar distances, including distances to H2" and to H4' that we have not discussed yet. While a base-H1' distance depends directly only on the glycosidic torsion angle, the other base-sugar proton-proton distances are determined by both the χ angle and sugar pucker.

Intraresidue distances to H2", like those to H2', are shorter in B-DNA than in A-DNA. We might therefore expect the corresponding NOE bounds to behave alike. However, if we look carefully at the experimental values, we notice that, while base-H2' bounds are mostly in agreement or longer than in B-DNA (Figure 3.2b), base-H2" bounds are consistently shorter than in standard B-DNA (Figure 3.2c). In a similar way, comparing base-H3' NOE bounds with base-H4' bounds, we note that they are both longer in B-DNA

than in A-DNA, but the bounds for base-H4' are systematically shorter and closer to A-DNA than are base-H3' bounds (Figures 3.2d and 3.2e).

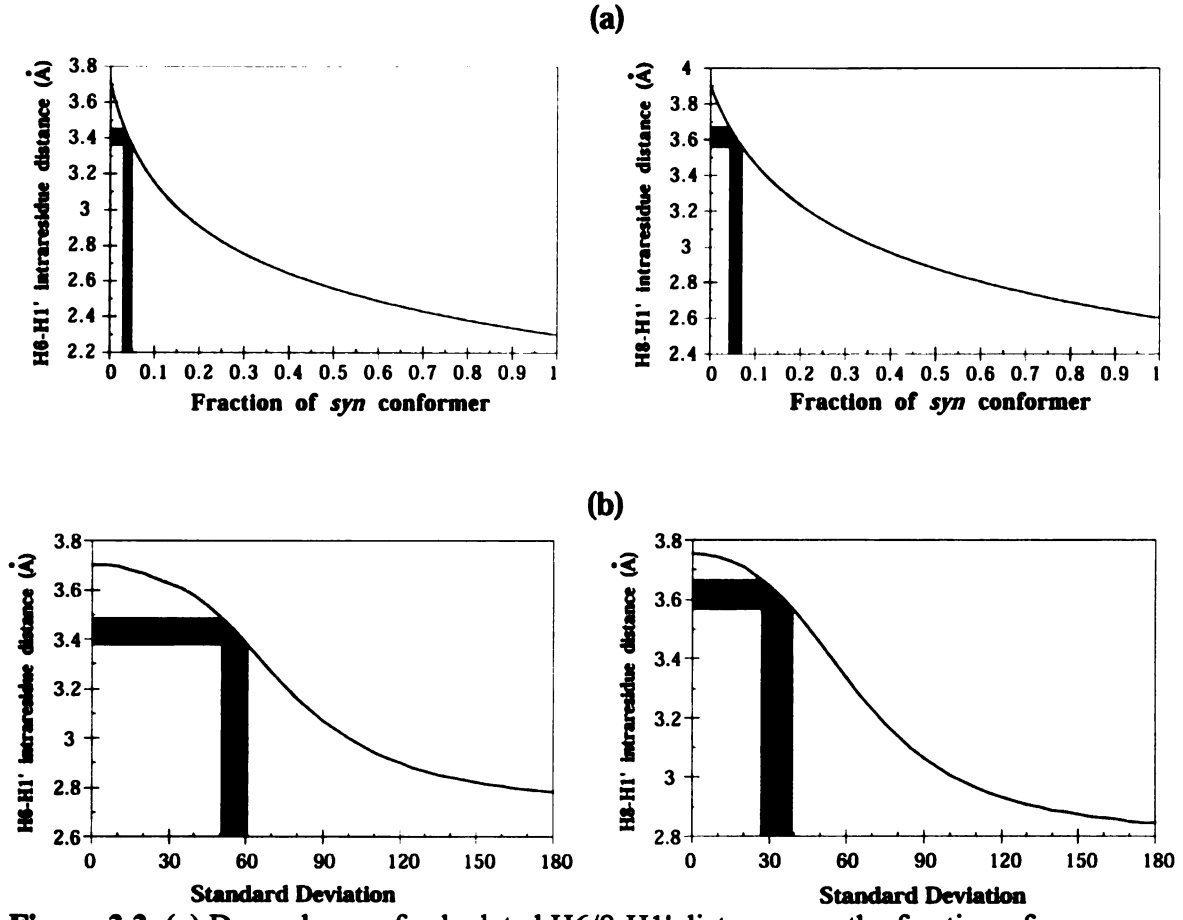


Figure 3.3. (a) Dependence of calculated H6/8-H1' distances on the fraction of *syn* conformer present in equilibrium with the *anti* conformer. The distance value read at 0 fraction of *syn* conformer represents the actual H6/8-H1' distance of the *anti* conformer and, vice versa, the value read at 1 fraction of *syn* is the actual distance of the *syn* conformer (no *syn/anti* equilibrium occurring). (b) Dependence of calculated H6/8-H1' distances on the standard deviation assuming a normal distribution of the χ torsion angle about an average value of 235° for pyrimidines and 255° for purines. The distance value read at 0 standard deviation represents the H6/8-H1' distance when no restricted rotational motion of χ is occurring. In both (a) and (b), distances were calculated assuming sixth root averaging of distances for a single conformer. The shaded cyan areas represent the ranges of fraction of *syn* conformer (a) and standard deviation (b) that are required to match the experimental base-H1' NOE upper bounds of all non-terminal residues in all three duplexes.

Sugar repuckering, which accounted for base-H2' and base-H3' bounds observations, cannot fully explain what we see with bounds to H2" and to H4'. There must be some other variable that is affecting base-sugar distances differently. This causes bounds to

H2" and H4' to be shorter than those to H2' and H3', respectively. Since the only other structural variable directly affecting these distances is the torsion angle χ, motion about this angle must be responsible for these observations.

We conclude that any motion about the glycosyl bond that is responsible for short base-H1' bounds is also shortening base-H2" and base-H4' distances, while it is not affecting base-H2' and base-H3' distances. Of the two previously proposed motional models involving the χ angle, we will find out that only one can explain this behavior.

3.3.4 Feasible internal motional models: analysis of MD-tar trajectories.

To help explain the observed NOE bounds, we analyzed structures generated during MD-tar simulations. The trajectories we used were run on the pribnow box octamer duplex (Schmitz et al., 1993). In MD-tar simulations, the experimental restraints and the associated penalty function is monitored as a running average with exponential weighting to emphasize more recent steps during an rMD simulation. As the interproton distance restraints are enforced as a running average rather than at every snapshot as in the usual rMD simulations, MD-tar has proved to explore conformational space more efficiently. This allows MD-tar to account better for any molecular flexibility encoded in the experimental data which cannot be satisfied by a single conformation.

First, we examined how the χ torsion angle and the pseudorotation angle P change during the simulation. In agreement with the original paper (Schmitz et al., 1993), we observed that: (1) torsion angle χ appears to be more dispersed compared to what has been seen in regular r-MD; (2) the *syn* conformer at the glycosidic bond is only found in terminal residues and in one sub-terminal residue (adenine 15); (3) the sugar rings are jumping between N and S conformations, with the S conformer dominant; (4) the change in P is associated with a smaller, but consistent, change in the χ torsion angle value. This is in agreement with the previously observed correlation between the glycosidic torsion angle and the backbone/sugar torsion angle δ (Fratini et al., 1982), since δ is strongly related to the sugar pucker. Recapitulating, for individual residues during MD-tar simulations two interchanging conformations with different P and χ values exist. These

are characterized by sugar puckers in the N and S regions. The *syn* conformer is populated only for the terminal residues of the duplex.

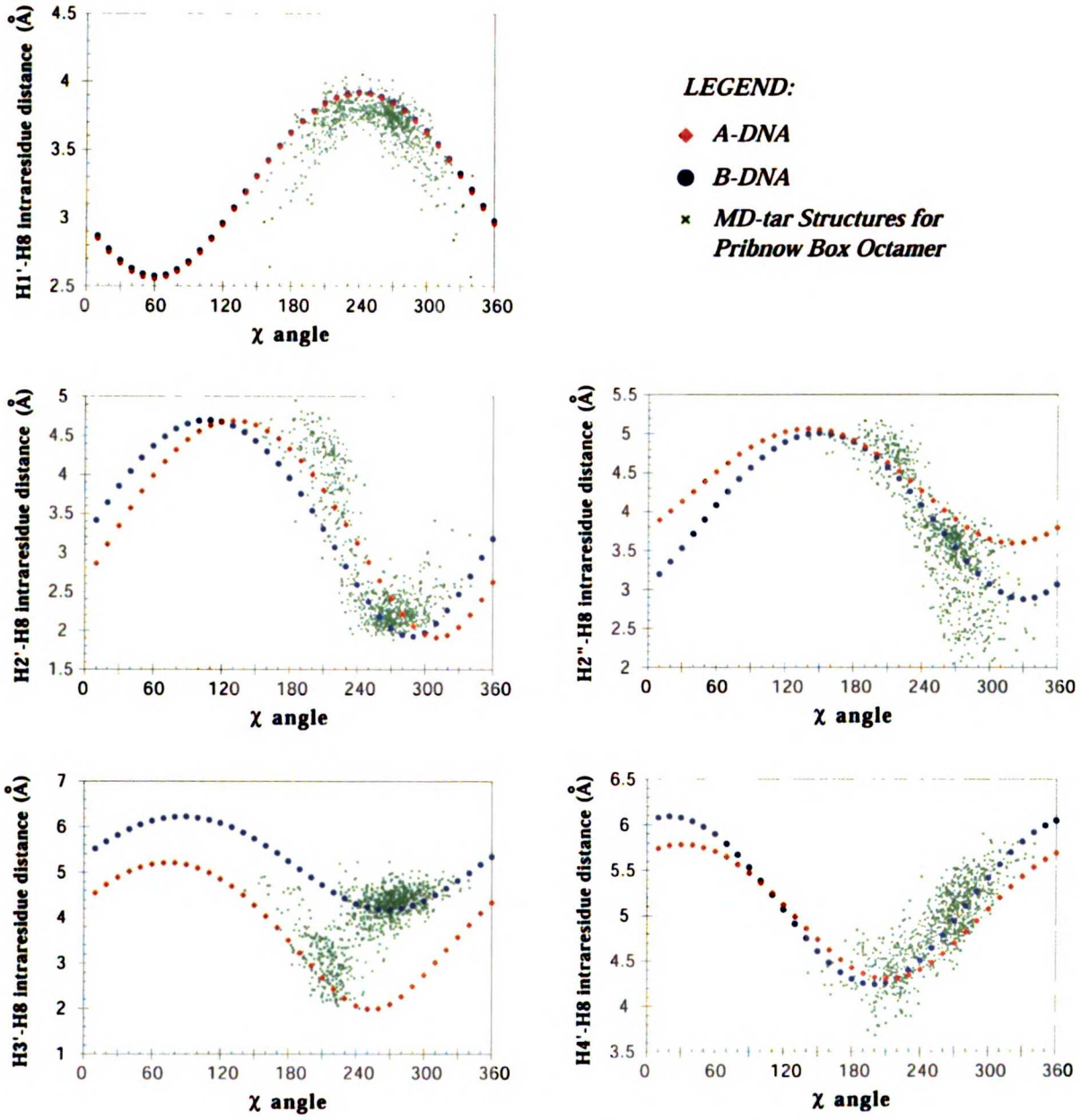


Figure 3.4. Correlation between H8-sugar proton distances and glycosidic torsion angle for adenosine 5 in the pribnow box octamer duplex d(GTATAATG)·d(CATTATAC) obtained from MD-tar simulations (~900 structures, green). The plots also show how H8-sugar proton distances change with χ angle in adenosine residues with either S (blue) and N (red) sugar pucker. These curves were obtained simply by measuring the base H8-sugar proton distances while manually changing χ starting with two standard A and B adenosine residues.

Finally, in our aim to explain the observed NOE bounds, we plotted base-sugar interproton distances against the glycosidic torsion angle value for each snapshot of the trajectory. This allows us to see directly any dependence of the distances on variations of χ. Figure 3.4 shows all base-sugar distances vs. χ angle plots for adenine 5. It is evident for those distances that change significantly between N and S sugar puckers (base H8-H2', H8-H2" and H8-H3' in particular), two interchanging populations corresponding to conformers with S and N sugar pucker. The two curves shown in the plots represent how the distances change with the χ angle in standard A and B adenine residues. These curves were obtained simply by measuring the H8-sugar distances while "manually" changing χ starting with two standard A and B adenine residues. This entire range is, of course, physically not very realistic; however, it helps in understanding the plots.

Now, let us reconsider the two proposed motional models of motion about the χ angle: the jump diffusion model and the restricted diffusion model. Using the plots in Figure 3.4, we will try to justify the NOE patterns observed for all H8-sugar distances using first one model and then the other. To simplify our analysis when trying to explain the effect of χ flexibility, we will focus our attention on the major population with S sugar pucker. The minor N conformer is used to analyze the effect of sugar repuckering.

The jump diffusion model describes the effect of a jump between *syn* and *anti* conformers. These structures have preferred χ regions from 190 to 270° (*anti*) and from 40 to 90° (*syn*) (Wüthrich, 1986). The MD-tar structures (Figure 3.4) reflect the experimental H8-H1' NOE bounds (Figure 3.2a) which are shorter than in A- or B-DNA, both being in the *anti* region. The *syn* conformer has H8-H1' distances about 1.3 Å shorter than the *anti* conformer. Thus, the presence of small amounts of a *syn* conformer in equilibrium with the *anti* conformer, can explain the observed short values. Base H8-H2', H8-H2", H8-H3' and H8-H4' bounds are directly affected by sugar pucker as well as χ angle. Sugar repuckering with S/N equilibrium can partially explain the experimental NOE values. However, we noticed that H8 bounds to H2" and to H4' are relatively shorter when compared to both A- and B-DNA than are bounds to H2' and to H3', respectively (see Figure 3.2). We attributed this difference to χ flexibility (*vide supra*). In the *syn* conformer, all these distances have about the same value as in the *anti* conformer

or, in some cases, they are longer. Thus, the jump diffusion model does not explain shortening of H8 NOE bounds to H2" and H4' relative to those to H2' and H3', respectively.

The restricted diffusion model allows restricted diffusion about the glycosidic bond between the angles $\chi \pm \Delta\chi$, with an average χ value of about 255° for adenines with S sugar pucker. Since deviations of χ from the average value decrease the distance to H1', this model can explain the observed base H8-H1' short bounds. Base H8-H2' and H8-H3' distances fall in a valley of the sinusoidal curve, while H8-H2" and H8-H4' distances lie on a slope of the curve. Consequently, experimental distances of H8 to H2' and to H3' will increase when χ oscillates around the center position. Instead, distances to H2" and to H4', will increase when χ changes upward and decrease when χ moves downward the slope. Since NOE distance bounds are biased towards short values, we conclude that rocking about the glycosidic bond can justify short H8-H2" and H8-H4' bounds, while bounds to H2' and H3' are little affected. Obviously, the same arguments hold for H6 protons in pyrimidines as for purine H8.

Thus, of the two proposed models for χ flexibility, only the restricted diffusion model, combined with S/N sugar repuckering, is able to justify the observed base-sugar bound patterns.

3.3.5 Analysis of a pool of conformers generated by unrestrained Monte Carlo calculations

So far, we have analyzed structures being generated by MD-tar simulations. One could argue that biases affecting this method may also prejudice our conclusions. Consequently, we decided to analyze a pool of conformers generated by the Monte Carlo module of the DNAminiCarlo program. Unrestrained energy minimization calculations were previously carried out on a two base-pair DNA fragment with sequence d(CA)·d(TG). Improved sampling of conformational space was achieved by applying the "scanning procedure" described elsewhere (Ulyanov et al., 1992; Zhurkin et al., 1982).

This pool of conformers was analyzed in a similar fashion as the MD-tar trajectories described in the previous section. Figure 3.5 shows the dependence of intraresidue base-

sugar distances on the χ angle value for the adenine residue. As for MD-tar, it is easy to notice the presence of two populations corresponding to conformers with S and N sugar pucker.

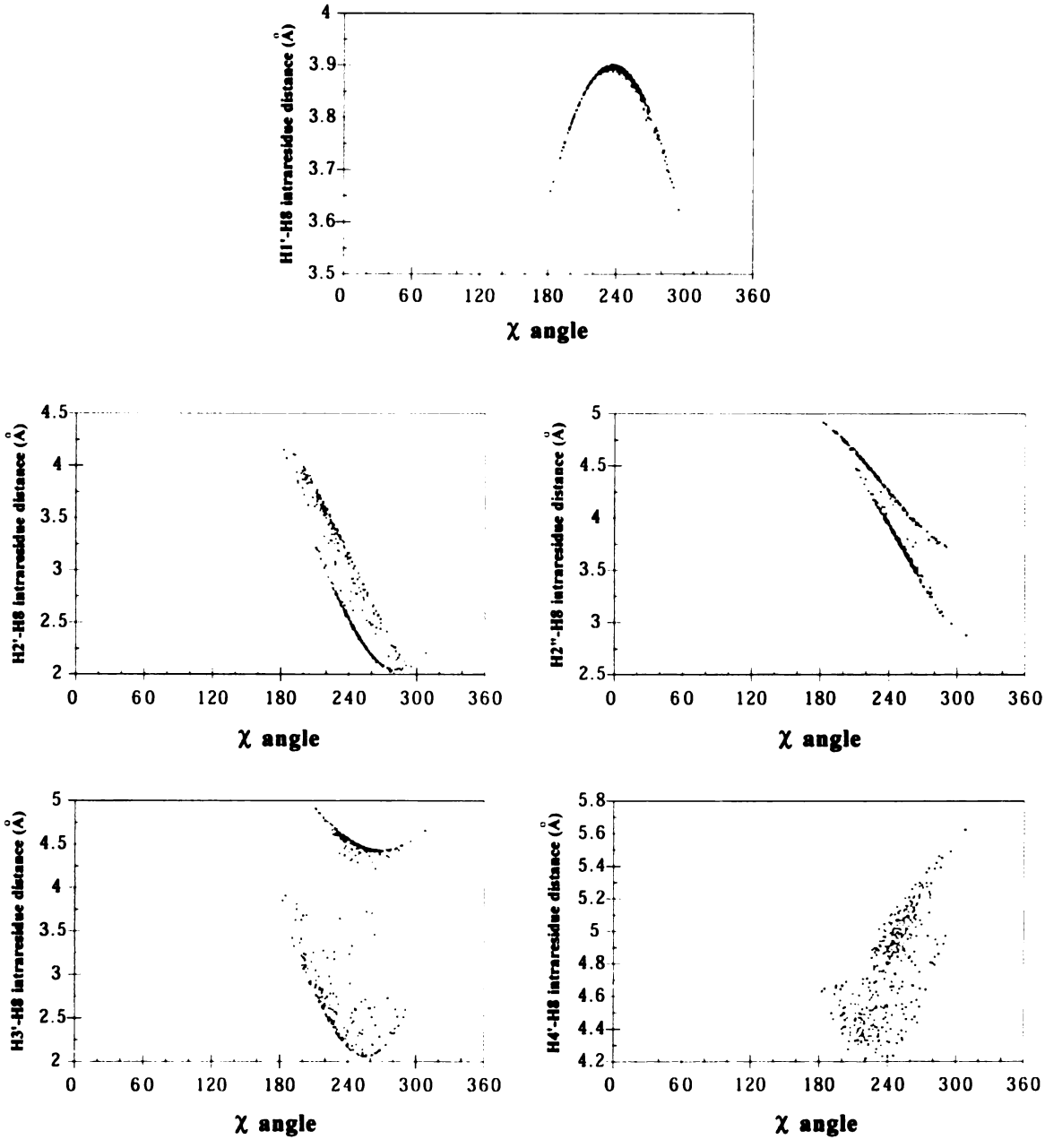


Figure 3.5. Correlation between H8-sugar proton distances and glycosidic torsion angle for adenosine 2 in the dimer duplex d(CA)·d(TG) obtained from unrestrained energy minimization performed by the Monte Carlo modules of the DNAminiCarlo program. Covering of conformational space was achieved by applying the "scanning procedure" that has been described elsewhere (Ulyanov et al., 1992, Zhurkin et al., 1982).

Two main differences can be pointed out: (1) the torsion angle χ is shifted towards smaller values, and (2) there is a narrower range of distances in the pool of conformers generated by the Monte Carlo method. The shift in the χ angle value probably results from the NOE restraints being applied during MD-tar but not during DNAminiCarlo calculations. In particular, short H8-H2' and H8-H2" bounds are pulling the χ torsion angle to higher values during MD-tar simulations in order to satisfy the restraints. Secondly, because of the narrower ranges of distances, the two populations corresponding to N and S sugar conformers are more clearly defined in the DNAminiCarlo plots. This may appear somewhat surprising if we consider that the Monte Carlo simulations were run free of experimental restraints. However, it can be easily understood since, in DNAminiCarlo, sugar pucker amplitudes are allowed to change only within a small range, while there are no restrictions in MD-tar simulations. In fact, in the MD trajectories, it is easy to find conformations with unrealistic sugar pucker amplitudes. These are often the result of inconsistencies in the NOE data enforced during the simulation.

With the differences being taken into account, the DNAminiCarlo plots support the same type of conclusions we draw from analysis of MD-tar simulations. Specifically, the existence of a S/N sugar pucker equilibrium in tandem with restricted rotation about the glycosidic bond best explains the NOE patterns.

3.4 CONCLUSIONS AND FUTURE DIRECTIONS

3.4.1 A motional model for duplex DNA: sugar repuckering and glycosidic torsion angle flexibility.

Analysis of intraresidue base-sugar interproton NOE bounds patterns suggests a motional model with S/N sugar repuckering, with S being the preferred conformer. Associated with repuckering, there is a small, but consistent, change in the glycosidic torsion angle χ. In addition to this, the χ torsion angle fluctuates around an average value, as described by a restricted diffusion model. Thus, sugar repuckering accounts for (a) base H8/H6-H2' and H8/H6-H2" bounds which are longer than in typical B-form DNA

conformation, and (b) H8/H6-H3' and H8/H6-H4' bounds which are shorter than in standard B-DNA and closer to A-DNA. On the other hand, restricted rotation about the glycosidic torsion angle justifies (a) H8/H6-H1' bounds shorter than in both A- and B-DNA, and (b) H8/H6-H2" and H8/H6-H4' bounds relatively shorter than for H8/H6-H2' and H8/H6-H3' bounds, respectively, when compared to standard A- and B-DNA values.

Even if we cannot exclude the presence of small amounts of *syn* conformer at equilibrium with the *anti* conformer in our analysis, it is clear that the jump diffusion model cannot fully explain the experimental data. Hence, this model is not required.

We pointed out earlier that the standard deviation estimated for the restricted diffusion model is broader than what has being previously estimated. However, our calculations were performed assuming a single gaussian distribution centered about the χ angle value typical of S sugar pucker. Our study also suggests the presence of a smaller percent of N conformer with shifted χ torsion angle. The reasoning for variations in H8/H6-H1' distances with restricted χ angle diffusion on the S conformer also applies to the N conformer. Moreover, both MD-tar and DNAminiCarlo simulations suggest a more scattered distribution of the χ angle in the N conformer. Thus, the presence of a small percent of N conformer with a broader χ angle distribution should further diminish the measured distance. Of course, this implies a narrower distribution for the χ angle in the S conformer. By repeating our calculations considering two χ angle distributions corresponding to the N and S conformers, we could estimate the size of the distributions, but this would entail nearly as many parameters as observables.

Finally, we need to consider the effect that random rotational molecular motion has on NOE intensities. In the model-free approach of Lipari and Szabo (Lipari and Szabo, 1982a; Lipari and Szabo, 1982b) the information on internal motions is specified by two model-independent quantities: an order parameter S^2 which is a measure of the spatial restriction of the motion and an effective correlation time τ_e which is a measure of the rate of the random internal motion. Kojima and James (unpublished results), using the model-free analysis on six relaxation parameters measured on a ^{13}C-labeled DNA duplex d(CATTTGCATC)·d(GATGCAAATG) estimated internal motions in the picosecond

time scale (10-40 ps), with an order parameter $S^2 \approx 0.8$ for C-H vectors in both bases and sugars. Using a simple wobble-in-a-cone model (London, 1980) gives some physical perspective to an order parameter; a value of 0.8 means that a C-H vector can rapidly reorient up to an angle of 44°.

Existence of such a fast internal motion, faster than the overall tumbling time of the duplex, will reduce cross-relaxation rates and, consequently, NOE intensities. Model-free analysis indicates that motional averaging of distances will change from a $\langle r^{-6} \rangle$ dependence to a ($S^2 \times \langle r^{-6} \rangle$). Thus, distances calculated from NOE intensities will be overestimated if we do not take into account the contribution from fast internal motion. However, our intensities, calculated by MARDIGRAS, were normalized against fixed distances (cytosines H5-H6 and thymines methyl-H6 distances). Since all C-H vectors in DNA were found to have order parameters around 0.8, we expect all proton-proton pairs to have similar order parameter values as well. Normalization of the NOE intensities should minimize any contribution from fast internal motion as long as the internal motions are comparable for all proton pairs (Kumar et al., 1992).

Studying molecular flexibility through analysis of proton-proton distance values offers an advantage, compared with the heteronuclear relaxation parameter studies alone, of yielding a realistic picture of the type of motion in the context of known structural information. Our model describes two types of motion: restricted rotation about an average position of the glycosidic torsion angle χ and N/S sugar repuckering associated with a shift of the average χ value by about 60°. It is reasonable to argue that these two motions occur on a different time scale, with sugar repuckering being slower, since it involves transitions between different energy minima. If this is the case, then the value of τ_e, found by the model-free analysis of relaxation parameters, should reflect the time scale of the fastest motion, i.e., χ angle fluctuations. While the time scale for sugar repuckering is not truly known, it is likely to be in the nanosecond range and, consequently, to be difficult to distinguish from overall molecular tumbling but to have nonnegligible effects on cross-relaxation rates.

In recent years, methodology development has permitted us to determine quite well the time-average, sequence-dependent structure of double helical nucleic acids. Key to

this has been the use of interproton distances determined from two-dimensional nuclear Overhauser effect spectroscopy using the complete relaxation matrix approach embodied in the software MARDIGRAS. However, with the increasing number of studies conducted in our laboratory, it has become clear that certain classes of distances are inconsistent with others in each of the structures studied. Indeed, some are inconsistent with the duplexes possessing structure in either the B or A family of structures, i.e., those which one would imagine are energetically most reasonable.

The current investigation has examined possible causes for these inconsistencies, focusing on the likelihood that internal motions are responsible. It has been demonstrated that repuckering of the deoxyribose ring, which scalar coupling-based spectroscopy results suggest occurs, cannot reconcile all the NOE data. However, restricted rotational diffusion about the glycosidic bond, together with sugar repuckering, does account for the NOE data, indicating that this is a general phenomenon as it has been observed in each of the DNA duplexes studied in our laboratory. A more complete description of the dynamic nature of the duplex accounting for repuckering and glycosidic bond diffusion will require an analysis using more sophisticated modeling and additional data.

3.5 ACKNOWLEDGMENT

This work has been published on Biochemistry, Vol. 37, No. 33, 1998. I wish to thank Dr. Chojiro Kojima for helpful discussions. Use of the facilities of the Pittsburgh Supercomputing Center for some calculations and the UCSF Computer Graphics Laboratory, supported by NIH Grant RR01081, are gratefully acknowledged.

3.6 REFERENCES

Borgias, B. A., and James, T. L. (1990). MARDIGRAS--Procedure for Matrix Analysis of Relaxation for DIscerning Geometry of an Aqueous Structure. J Magn Reson *87*, 475-487.

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc *117*, 5179-5197.

Drew, H. R., Wing, R. M., Takano, T., Broka, C., Tanaka, S., Itakura, K., and Dickerson, R. E. (1981). Structure of a B-DNA Dodecamer: Conformation and Dynamics. Proc Nat Acad Sci (USA) *78*, 2179-2183.

Fratini, A. V., Kopka, M. L., Drew, H. R., and Dickerson, R. E. (1982). Reversible bending and helix geometry in a B-DNA dodecamer: CGCGAATTBrCGCG. J Biol Chem *257*, 14686-14707.

Gaudin, F., Paquet, F., Chanteloup, L., Beau, J. M., Nguyen, T. T., and Lancelot, G. (1995). Selectively ^{13}C-enriched DNA: dynamics of the C1'-H1' vector in d(CGCAAATTTGCG)$_2$. J Biomolec NMR *5*, 49-58.

Keepers, J. W., and James, T. L. (1984). A theoretical study of distance determinations from NMR two-dimensional nuclear Overhauser effect spectra. J Magn Reson *57*, 404-426.

Kumar, A., James, T. L., and Levy, G. C. (1992). Macromolecule Structure Refinement by Relaxation Matrix Analysis of 2D NOE Spectra: Effect of Differential Internal Motion on Internuclear Distance Determination in Proteins. Israel J Chem *32*, 257-261.

Lane, A. N. (1993). NMR Studies of Dynamics in Nucleic Acids. Prog NMR Spectr *25*, 481-505.

LeMaster, D. M., Kay, L. E., Brunger, A. T., and Prestegard, J. H. (1988). Protein dynamics and distance determination by NOE measurements. FEBS Lett *236*, 71-76.

Lipari, G., and Szabo, A. (1982a). Model-Free Approach to the Interpretation of Nuclear Magnetic Resonance Relaxation in Macromolecules. 1.Theory and Range of Validity. J Am Chem Soc *104*, 4546-4559.

Lipari, G., and Szabo, A. (1982b). Model-Free Approach to the Interpretation of Nuclear Magnetic Resonance Relaxation in Macromolecules. 2. Analysis of Experimental Results. J Am Chem Soc *104*, 4559-4570.

Liu, H., Kumar, A., Weisz, K., Schmitz, U., Bishop, K. D., and James, T. L. (1993). Extracting Accurate Distances and Bounds from 2D NOE Exchangeable Proton Peaks. J Am Chem Soc *115*, 1590-1591.

Liu, H., Spielmann, H. P., Ulyanov, N. B., Wemmer, D. E., and James, T. L. (1995). Interproton Distance Bounds from 2D-NOE Intensities: Effect of Experimental Noise and Peak Integration Errors. J Biomolec NMR *6*, 390-402.

London, R. E. (1980). Intramolecular Dynamics of Proteins and Peptides as Monitored by Nuclear Magnetic Relaxation Measurements. In Magnetic Resonance in Biology, J. S. Cohen, ed. (New York, John Wiley & Sons), pp. 1-69.

Rinkel, L. J., and Altona, C. (1987). Conformational Analysis of the Deoxyribofuranose Ring in DNA by means of Sums of Proton-proton Coupling Constants: A Graphical Analysis. J Biomol Struct Dyn *4*, 621-649.

Schmitz, U., and James, T. L. (1995). How to Generate Accurate Solution Structures of Double-Helical Nucleic Acid Fragments Using Nuclear Magnetic Resonance and Restrained Molecular Dynamics. In Methods in Enzymology, Nuclear Magnetic Resonance and Nucleic Acids, T. L. James, ed. (New York, Academic Press), pp. 3-44.

Schmitz, U., Sethson, I., Egan, W., and James, T. L. (1992). Solution Structure of a DNA Octamer Containing the Pribnow Box via Restrained Molecular Dynamics Simulation with Distance and Torsion Angle Constraints Derived from 2D NMR Spectral Fitting. J Mol Biol *227*, 510-531.

Schmitz, U., Ulyanov, N. B., Kumar, A., and James, T. L. (1993). Molecular Dynamics with Weighted Time-averaged Restraints of a DNA Octamer: Dynamic Interpretation of NMR Data. J Mol Biol *234*, 373-389.

Schmitz, U., Zon, G., and James, T. L. (1990). Deoxyribose Conformation In [d(GTATATAC)]$_2$: Evaluation of Sugar Pucker by Simulation of Double-Quantum-Filtered COSY Cross-Peaks. Biochemistry *29*, 2357-2368.

Tonelli, M., Ragg, E., Bianucci, A. M., Lesiak, K., and James, T. L. (1998). Nuclear Magnetic Resonance Structure of d(GCATATGATAAG): A Consensus Sequence for Promoters Recognized by σ^K RNA Polymerase. Biochemistry *37*, 11745-11761.

Torda, A. E., Scheek, R. M., and van Gunsteren, W. F. (1989). Chem Phys Lett *157*, 289-294.

Ulyanov, N., Gorin, A. A., Zhurkin, V. B., Chen, B.-C., Sarma, M. H., and Sarma, R. H. (1992). Systematic Study of Nuclear Overhauser Effects vis-a-vis Local Helical Parameters, Sugar Puckers, and Glycosidic Torsions in B DNA: Insensitivity of NOE to Local Transitions in B DNA Oligonucleotides Due to Internal Structure Compensations. Biochemistry *31*, 3918-3930.

Ulyanov, N. B., Mujeeb, A., Donati, A., Furrer, P., Liu, H., Farr-Jones, S., Konerding, D. E., Schmitz, U., and James, T. L. (1998). Determination of Structural Ensembles from NMR Data: Conformational Sampling and Probability Assessment. In Molecular Modeling of Nucleic Acids, N. B. Leontis, and J. J. SantaLucia, eds. (Washington, D.C., American Chemical Society), pp. 181-194.

Ulyanov, N. B., Schmitz, U., Kumar, A., and James, T. L. (1995). Probability Assessment of Conformational Ensembles: Sugar Repuckering in a DNA Duplex in Solution. Biophys J *68*, 13-24

Weisz, K., Shafer, R. H., Egan, W., and James, T. L. (1994). Solution Structure of the Octamer Motif in Immunoglobulin Genes via Restrained Molecular Dynamics Calculations. Biochemistry *33*, 354-366.

Wüthrich, K. (1986). NMR of Proteins and Nucleic Acids (New York, Wiley).

Zhurkin, V. B., Lysov, Y. P., Florentiev, V. L., and Ivanov, V. I. (1982). Torsional flexibility of B-DNA as revealed by conformational analysis. Nucl Acids Res *11*, 1811-1830.

CHAPTER 4

4. Dynamic NMR Structures of [Rp]- and [Sp]-phosphorothioated DNA-RNA hybrids: is flexibility required for RNase H recognition ?

Chemically modified DNA oligonucleotides have been crucial to the development of antisense therapeutics. High-resolution structural studies of pharmaceutically relevant derivatives have been limited to only a few molecules. We have used NMR to elucidate the structure in solution of two DNA-RNA hybrids with the sequence d(CCTATAATCC)·r(GGAUUAUAGG). The two hybrids contain an unmodified RNA strand, whereas the DNA strand contains one of two different sugar-phosphate backbone linkages at each nucleotide: (1) [Rp]-phosphorothioate or (2) [Sp]-phosphorothioate. Homonuclear two-dimensional spectroscopy afforded nearly complete nonlabile proton assignments. Distance bounds, calculated from the NOE cross-peak intensities *via* a complete relaxation matrix approach with the program MARDIGRAS, were used to restrain the structure of the two hybrids during simulations of molecular dynamics. Analysis of MD trajectories suggest that both hybrids are highly flexible, requiring the use of time-averaged restraints (MDtar) to generate ensembles of structures capable of satisfying the NMR data. In particular, the deoxyribose sugars of the DNA strand show strong evidence of repuckering. Furthermore, deoxyribose sugar repuckering is accompanied by increased flexibility of overall helical geometry. These observations,

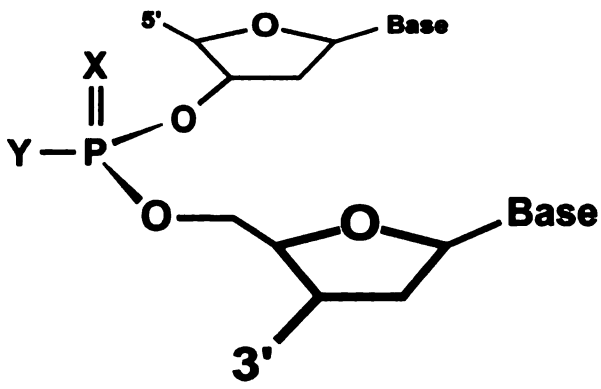
---

Abbreviations. PSO, phosphorothioated oligodeoxyribonucleotide; [Rp]-hybrid, [Rp]-PSO-RNA hybrid duplex; [Sp]-hybrid, [Sp]-PSO-RNA hybrid duplex; RNase H, ribonuclease H; 1D, one-dimensional; 2D, two-dimensional; 2D NOE, two-dimensional nuclear Overhauser effect; DQF-COSY, double-quantum-filtered correlation spectroscopy; TOCSY, total correlated spectroscopy; FID, free induction decay; rMD, restrained molecular dynamics; MD-tar, molecular dynamics with time-averaged restraints; RMSD, root-mean-square deviation; R^x, sixth-root R-factor; P, pseudorotation phase angle.

together with the analysis of the crystal structure of a hybrid duplex in complex with RNase H, suggested that this flexibility may be required for recognition by RNase H.

4.1 INTRODUCTION

Antisense oligonucleotide inhibitors are designed to disrupt function of specific RNA targets by forming hybrid duplexes with complementary RNA sequences. Mode of action of the first generation antisense oligonucleotides utilized ribonuclease H (RNase H), a ubiquitous sequence-nonspecific endonuclease that cleaves phosphodiester bonds in the RNA strand of RNA/DNA duplexes but leaves the DNA strand intact. Such a convenient property of RNase H allowed recycling the oligodeoxyribonucleotides after the RNA strand was hydrolyzed. Antisense oligonucleotides often include phosphorothioate modifications of the backbone, which confer resistance to single-strand-specific nucleases. At the same time, RNA hybrids with phosphorothioated oligodeoxyribonucleotides (PSO) are still recognized by RNase H (Agrawal and Iyer, 1997; Crooke and Bennett, 1996; DeLong et al., 1997; Temsamani and Guinot, 1997). One serious problem with this strategy is a relatively low specificity of both DNA-RNA recognition and recognition of hybrid duplexes by RNase H. Indeed, it has been shown that mismatched hybrid duplexes are still recognized by RNase H, which leads to the cleavage of unintended RNA sequences. Antisense oligonucleotides of newer generations are designed to form very stable hybrid duplexes with complementary RNA due to various base, sugar or backbone modifications. This comes most often, if not always, at the expense of the hybrid ceasing to be a substrate for RNase H. The mode of action of new antisense drugs is steric block; formation of very stable hybrid duplexes interferes with normal function of target RNA (Altmann et al., 1997; Freier and Altmann, 1997; Marquis and Grindel, 2000; Monia, 1997). Nevertheless, recycling of antisense oligonucleotides is a property of the RNase H mode of action too attractive to lose; a promising approach includes ligation of new-generation oligonucleotides with DNA or PSO, thus combining the stability of the hybrid duplex with a window for the RNase H action (Lima and Crooke, 1997; McKay et al., 1999).



	X	Y
phosphate	O	O
[Rp]-phosphorothioate	O	S
[Sp]-phosphorothioate	S	O
phosphorodithioate	S	S

Figure 4.1. Scheme showing definitions of [Rp] and [Sp] oxygens.

Many details of the enzymatic catalysis by RNase H have been elucidated in the past several years (reviewed by Krakowiak et al (Krakowiak et al., 2002)). An important development in the structural studies includes a crystal structure of a complex between HIV-1 reverse transcriptase (which contains an RNase H domain) and an RNA-DNA hybrid duplex (Sarafianos et al., 2001). However, many questions remain unanswered, one of the most puzzling being, why RNase H cleaves the RNA strand in RNA-DNA or RNA-PSO but not in RNA-RNA duplexes and not in hybrid duplexes of RNA with new antisense drugs. In an attempt to answer these questions, we have been studying a series of complexes of antisense oligonucleotides with a prototype RNA target, which includes the Pribnow box sequence. Previously, we described high-resolution NMR studies in solution of hybrid duplexes of RNA with complementary DNA, [Rp]-phosphorothioate, and phosphorodithioate (Furrer et al., 1999; Gonzàlez et al., 1994; Gonzàlez et al., 1995). Here we present NMR data for the [Sp]-PSO-RNA hybrid with the same sequence and determine solution structures for both [Sp]-PSO-RNA and [Rp]-PSO-RNA hybrids. The main conformational feature of these structures, common for both hybrids, is unusually high flexibility of deoxyribose rings in PSO strands. A similar behavior has been observed before for the DNA strand in a DNA-RNA hybrid (Gonzàlez et al., 1995; Gyi et al., 1998). Such flexibility required a special approach for structure determination, because no single structure can adequately describe observed NMR data. Here we used a combination of MDtar (Torda et al., 1990) and PDQPRO (Ulyanov et al., 1995) methods to determine an ensemble of conformations consistent with experimental data. Flexibility of deoxyribose sugar rings is accompanied by increased flexibility of overall helical

geometry, including the inclination and X-displacement parameters. Comparing our results with the crystal structure of a complex between RNase H and a hybrid duplex, we put forward a hypothesis that this flexibility may be required for the RNase H recognition.

4.2 MATERIAL AND METHODS

4.2.1 Sample preparation and NMR spectroscopy

The hybrid duplex was prepared for NMR by titration of a solution of the modified [Sp]-phosphorothioate DNA strand with a similar solution of the RNA strand. The 1:1 stoichiometry was ensured by monitoring the disappearance of the DNA single strand peaks in a 1D NMR spectrum. After each addition of RNA strand, the sample was re-annealed directly into the NMR tube to ensure proper forming of the duplex. The final concentration of hybrid in the NMR sample is ~1 mM dissolved in 0.1 mM EDTA, 100 mM NaCl and 20 mM sodium phosphate (pH 6.5) buffer. The sample was then lyophilized twice in order to exchange water for ^2H_2O and finally placed in a 5 mm Shigemi tube.

Two 2D NOESY, a 2D TOCSY and a 2D DQF-COSY spectra were collected on the [Sp]-phosphorothioate hybrid sample in ^2H_2O at 600 MHz on a Varian Inova spectrometer. All measurements were taken at 293°K using a spectral width of 6600 Hz in both dimensions and 4096 x 2048 data points along ω_2 and ω_1 dimensions, respectively. These spectra were then processed using the program NMRPipe (Delaglio et al., 1995) to a yield a final 2K x 2K data set; a combination of a Gaussian and sine-bell window functions were applied for resolution enhancement in each dimension. Subsequently, the processed spectra were imported into our locally developed NMR data analysis program SPARKY (Goddard and Kneller, 1998) for assignment and integration. The two 2D NOESY spectra were recorded in hyper-complex mode with mixing times of 100 and 200 ms to allow for build up of NOE cross-peaks. These spectra were acquired with 32 scans for each t_1 fid and a delay of 8 second between scans. The pure absorption 2D DQF-COSY was also acquired with 32 scans for each t_1 fid, but a repetition delay of 3 second between scans. A somewhat stronger apodization function was applied during

processing to reduce cancellation of the anti-phase components of the DQF-COSY cross-peaks. Finally, the 2D TOCSY spectrum was collected using a mixing period of 50 ms, and only 8 scans for each t_1 fid were sufficient. 2D NOESY spectra were also acquired with the sample dissolved in H_2O, but were only used for evaluating the structure of the duplex and assignments of exchangeable resonances and not to extract information for structure refinement.

4.2.2 Assignments

The 2D NOESY spectra acquired in ^2H_2O were used for assigning non-exchangeable protons by identifying base-base, base-H1', base-H2'/H2" (for the DNA strand) and base-H3' sequential walks (Wüthrich, 1986). Intraresidue connectivities provided by 2D TOCSY and COSY spectra were helpful to distinguish among overlapped resonances. Proton assignments for the [Sp]-hybrid are given in Table 4.1.

Table 4.1. Non-exchangeable proton chemical shifts (ppm) of [Sp]-phophorothioate hybrid at 20°C. [a]Not assigned. [b]Degenerate H2' and H2" resonances.

	H1'	H2'	H2"	H3'	H4'	H5'	H5"	H6/8	H2	H5	M7
C1	5.831	2.594[b]		4.768	4.119	3.942	3.866	7.973		5.977	
C2	6.002	2.474	2.692	4.883	4.319	4.195	4.121	7.819		5.670	
T3	6.003	2.516	2.705	5.026	4.326	4.204	*a*	7.663			1.634
A4	6.215	2.595	2.872	5.021	4.448	4.224	*a*	8.150	7.046		
T5	5.830	2.237	2.577	4.982	4.260	4.322	*a*	7.248			1.341
A6	5.925	2.704[b]		5.088	4.503	4.090		8.060	6.467		
A7	6.194	2.566	2.814	4.932	4.448	4.382	4.290	7.948	7.652		
T8	5.867	2.280	2.519	4.885	4.262	*a*	*a*	7.287			1.219
C9	6.035	2.296	2.579	4.830	4.280	*a*	*a*	7.556		5.452	
C10	6.155	2.099	2.296	4.477	4.112	4.168	4.249	7.507		5.416	
G11	5.430	4.726		4.421	4.168	3.886	3.793	7.766			
G12	5.771	4.695		4.623	4.482	4.121	*a*	7.410			
A13	5.936	4.601		4.530	*a*	4.131	*a*	7.802	7.800		
U14	5.427	4.407		4.368	4.443	4.080	*a*	7.588		5.031	
U15	5.649	4.654		4.629	4.476	*a*	*a*	8.097		5.605	
A16	5.932	4.442		4.625	4.555	4.203	*a*	8.196	6.963		
U17	5.360	4.398		4.492	*a*	4.091	*a*	7.652		5.088	
A18	5.988	4.636		4.727	4.472	*a*	*a*	8.092	6.882		
G19	5.468	4.388		4.428	*a*	4.065	*a*	7.146			
G20	5.817	4.047		4.207	*a*	*a*	*a*	7.300			

4.2.3 Extracting interproton distance restraints

In each of the two 2D NOESY spectra, the cross-peaks intensities were measured with SPARKY by line-fitting of the peaks to a Gaussian function and integration of the theoretical curve. These intensities were then used to calculate accurate distance restraints by using the complete relaxation matrix approach implemented in MARDIGRAS (Borgias and James, 1990). Two of the three starting models that were used for MARDIGRAS calculations were generated by the miniCarlo program (Zhurkin et al., 1991) using standard helical parameters for A-form and B-form nucleic acid duplexes. The third starting model, H-form, was also generated by miniCarlo; it has C3'-*endo* sugar puckers for the RNA strand and C2'-*endo* for the PSO strand. All three structures were energy-minimized by the same program. Several values of correlation time were used to represent the isotropic motion of the molecule as a whole but, in the end, only distances calculated using 5, 6 and 7 ns were used to estimate the bounds for structure calculation. These values of τ_c were chosen by analyzing their ability to accurately reproduce certain intra-sugar distances that are known to change only within a narrow range, independently of the sugar conformation. MARDIGRAS was run with the RANDMARDI option set to 200 (Liu et al., 1995). Thus, for each of the two sets experimental NOE intensities, 200 new sets were generated by adding noise, randomly calculated within user-chosen limits, directly to the intensities before calculation of the distances. In conclusion, three starting models and three values of isotropic τ_c were used to run MARDIGRAS on the two experimental NOE intensities sets measured at 100 and 200 ms mixing times, for a total of 18 MARDIGRAS calculations. Each one of these calculations was then repeated 200 times on randomly modified intensity sets generated by the RANDMARDI option selected in MARDIGRAS. All these distances, generated by MARDIGRAS, were combined to calculate average values and standard deviations. Then, for each proton pair, upper and lower limits to be used for structure calculation were chosen between the wider of the bounds calculated using the average distance value added and subtracted of the standard deviation and the bounds generated by building a distribution of the distances and eliminating 10% from each end of the distribution.

Sample preparation and NMR spectroscopy for the [Rp]- hybrid were performed following a strategy similar to that described here for the [Sp]-hybrid and have already

been reported elsewhere (Furrer et al., 1999). However, even though the distances for the [Rp]-hybrid were calculated previously, we repeated the procedure for both hybrids, in order to use compatible computational schemes for both molecules. The main difference between the two systems was in the concentration of hybrid in the NMR sample, being higher for the [Rp]-hybrid. However, given that the all spectra for the [Rp]-hybrid were also collected at a higher temperature (30°C *vs* 20°C), the values of τ_c used in MARDIGRAS calculations were the same for both hybrids (5,6,7 ns); see Table 4.2 for a summary of MARDIGRAS-calculated distance restraints for both [Rp]- and [Sp]-hybrids.

Table 4.2. Number and average width for distance bounds calculated from NOE intensities with MARDIGRAS from the NOE intensities for the [Rp]- and [Sp]-hybrids.

	[Sp]-hybrid		[Rp]-hybrid	
	# distances	average width	# distances	average width
Total	208	1.508	196	1.661
Intraresidue	77	1.294	65	1.742
Interresidue	118	1.664	117	1.697
Cross-strand	13	1.362	14	0.981
DNA strand	118	1.533	100	2.168
Intraresidue	50	1.390	39	2.168
Interresidue	68	1.638	61	2.168
RNA strand	77	1.494	82	1.158
Intraresidue	27	1.115	26	1.103
Interresidue	50	1.698	56	1.184

Finally, together with the NOE-derived distance restraints, some empirical restraints were also included for structure refinement. These include typical hydrogen bond distance and flat angle restraints to ensure proper pairing of the bases. Also, loose distance restraints (3.0 Å - 8.0 Å upper and lower bounds) between each sodium counter ion and two phosphorous atoms were imposed to maintain a sodium atom in the space in between two phosphate groups. These restraints account for the position of 18 ions, 9 for each strand. The remaining two ions, required to neutralize the system, were let free to move in the bulk solvent. Furthermore, later in our attempt to refine the hybrids, non-NOE cross-strand distance restraints (4.5 Å lower bound) were added between H1' protons: H1'(i) - H1'(j+2), where (j) is the nucleotide base-paired to (i), yielding a total of 8 distance restraints. These bounds were needed to keep the two strands from coming

too close in the minor groove of the duplex during the MD simulations. A summary of experimental and empirical restraints used in structure refinement is reported in Table 4.3.

4.2.4 Structure calculations

Upper and lower bounds generated by MARDIGRAS were used to refine the structure of both hybrids by restrained molecular dynamics simulations (rMD) or by MD with time averaging (MDtar). The program AMBER 6 was used (Case et al., 2000).

The initial models used at the start of AMBER calculations for the two phophorothioate hybrids, were generated with miniCarlo by restrained minimization of a typical A-form hybrid duplex. These initial structures were further energy-minimized with AMBER *in vacuo* and, then, placed in periodic solvent boxes containing 10 Å of TIP3P water molecules surrounding the hybrid in each direction. Subsequently the system was neutralized by adding sodium counter ions. The solvated systems were then equilibrated with standard protocols. Briefly, after an initial minimization, 50 ps of MD simulations were run at constant volume, followed by 150 ps of MD simulations at constant pressure, for a total of 200 ps of equilibration. The water molecules were let free to equilibrate from the beginning of the simulation, while the hybrid duplex was kept fixed by positional restraints that were gradually reduced towards the end of the equilibration. The sodium counter ions were also initially fixed by positional restraints, but let free to equilibrate after 40 ps of MD simulations at constant pressure, their position being then restrained in the neighborhood of phosphate groups by the empirical bounds described above. In fact, the empirical restraints, as well as the NOE derived distance bounds, were gradually built up during the equilibration period.

Production runs for regular restrained MD and MDtar calculations were started independently from the equilibrated system. Each production calculation was run for a period of 1ns. Throughout the equilibration and production runs, the particle mesh Ewald (PME) method (Essmann et al., 1995) was used to treat electrostatic interactions (default in AMBER 6 for solvated systems), and the temperature was kept constant at 300°K by using the Berendsen coupling algorithm (Berendsen et al., 1984) with a time constant of 1 ps. Other parameters were set to typical values for MD: a 9 Å cut-off for non-bonded

interactions, SHAKE on bonds involving hydrogen atoms (Ryckaert et al., 1977), a 1 fs (0.001ps) time step. The atomic coordinates were saved every 1000 steps (1ps) for a total of 1000 structures for each rMD or MDtar run. All trajectories were stripped of the water molecules and analyzed by visual inspection using the program MOIL-View (Simmerling et al., 1995). In the end, four simulations were selected for further analysis: two rMD and two MDtar runs, one of each for the two [Sp]- and [Rp]-phosphorothioate hybrids. 1000 pdb files were generated for each trajectory and the program Curves (Lavery and Sklenar, 1996) was run on each saved duplex structure to calculate the helical parameters. The programs MIDAS (Ferrin et al., 1988), Chimera (Huang et al., 1996) and Molmol (Koradi et al., 1996) were used to visualize and further analyze the coordinates. Representative structural ensembles for each 1 ns production run were generated using the locally developed PDQPRO algorithm (Ulyanov et al., 1995). CORMA (Keepers and James, 1984) was then run for each simulation on the whole ensemble of 1000 structures (with equal probability for each structure), on the ensemble of PDQPRO selected structures (using PDQPRO probabilities for each model) and separately on each pdb file, to back-calculate theoretical NOESY spectra. R factors were measured by comparison of the simulated NOE intensities with the corresponding experimental ones.

The complete list of experimental and empirical restraints used for structure refinement and the force constant used to enforce them is reported in Table 4.3. During MDtar runs, only the NOE-derived distance bounds were subjected to the exponentially weighted third-root time-averaging with an exponential time constant of 20 ps and the 'pseudo-force' option (Pearlman and Kollman, 1991; Torda et al., 1990); all other restraints were applied continuously.

Table 4.3. Restraints used in the rMD equilibration and rMD and MDtar production runs for the [Rp]- and [Sp]-hybrids. The force constant used to enforce these bounds during restrained MD simulations, k, is given in units of kcal/(mol · Å^2).

	Bounds (Å)		[Sp]-hybrid		[Rp]-hybrid	
	lower	upper	# bounds	k	# bounds	k
NOE distance restraints			208	2.5	196	5.0
H bond distance restraints			24	3.75	24	7.5
H bond flat angle restraints			24	5.0	24	10.0
Na^+ - P distance restraints	3.0	8.0	32	0.5	32	1.0
H1'(i) - H1'(j+2) restraints	4.5	40.0	8	25.0	8	50.0

4.2.5 PDQPRO calculations

Representative structural ensembles were selected from the entire 1-ns MDtar trajectories using the PDQPRO algorithm (Ulyanov et al., 1995). The program finds a subset of structures and their probabilities so that the resulting ensemble produces the best fit of observed NMR parameters. Experimental homonuclear cross-relaxation rates were derived from the NOE data using MARDIGRAS simultaneously with calculating distance restraints (see above).

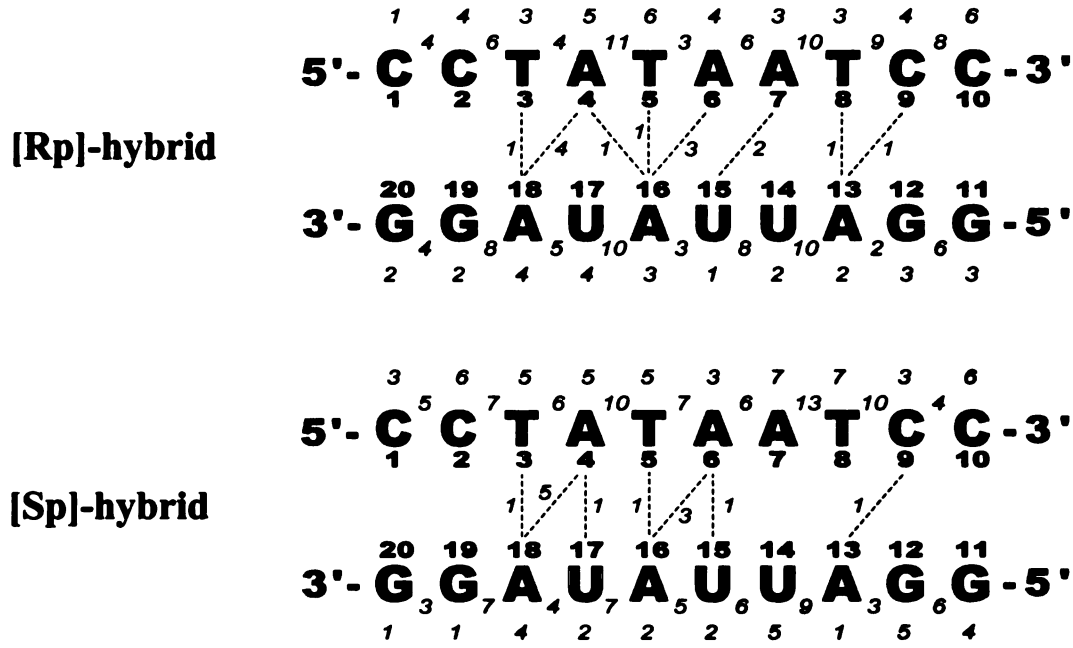


Figure 4.2. Sequence numbering (bold) and distance restraint distributions (intraresidue, interresidue and cross-strand) for the [Rp]- and [Sp]-hybrids.

4.3 RESULTS

4.3.1 NMR spectroscopy and interproton distance restraints

Proton assignments were performed following the procedure indicated earlier. We were able to identify all base protons and most sugar protons, except for H4' of residues A13, U17, G19 and G20. Some H5'/H5" protons were also tentatively assigned based on their NOE and TOCSY contacts, but not used for structure refinement. Not surprisingly, resonances for the RNA strand are similar to those already assigned for the other hybrid

duplexes with the same sequence (Furrer et al., 1999), while resonances for the modified DNA duplex strand are somewhat different. DNA and RNA strand sequential walks for the fingerprint region are shown in Figure 4.3a, and the proton assignments for the [Sp]-hybrid are given in Table 4.1. Interproton distance restraints were calculated from experimental NOE data as described in Materials and Methods, using a complete relaxation matrix analysis (MARDIGRAS) with the random error perturbation procedure (RANDMARDI). Even though the distances have been calculated for the [Rp]-hybrid previously (Furrer et al., 1999), we repeated the procedure for both hybrids, in order to use compatible computational schemes for both molecules. The statistics of distance restraints is similar for both hybrids, with ca. 10 restraints per residue and an average flat-well width of 1.7 and 1.5 Å for the [Rp]- and [Sp]-hybrids, respectively (Table 4.2). These numbers do not include interproton distances with low variation, such as H1'-H2', which do not carry any structural information. The distribution of the calculated distance restraints is shown in Figure 4.2. Distance bounds for the RNA residues are largely consistent with the helical A conformation and with C3'-*endo* sugar pucker of riboses. In contrast, distances involving deoxyribose protons in [Sp]- or [Rp]-PSO strands are not consistent with either C3'-*endo* or C2'-*endo* sugar conformations. Intraresidue H2'-H6/H8 distances are consistent with the B-like sugar conformations (C2'-*endo* or C1'-*exo*), but they are too short for C3'-*endo* conformations. At the same time, intraresidue H3'-H6/H8 distances are consistent with the A-like C3'-*endo* conformations, but they tend to be too short for B-like conformations (Figure 4.4). A similar behavior has been observed previously for the DNA residues in regular DNA-RNA hybrids (Gonzàlez et al., 1994; Gyi et al., 1998), and to a lesser extent for DNA duplexes (Mujeeb et al., 1992; Rinkel and Altona, 1987; Schmitz et al., 1992b; Schmitz et al., 1990; Tonelli and James, 1998; Weisz et al., 1992). This is qualitatively confirmed by the DQF-COSY spectrum of the [Sp]-hybrid (Figure 4.3b): the presence of H2'-H1' peaks rules out pure C3'-*endo* sugar pucker for the PSO residues. In principle, this situation may arise either due to some non-standard conformations of sugar rings in PSO residues, or by flexibility of these residues when both C2'-*endo* and C3'-*endo* conformers are present.

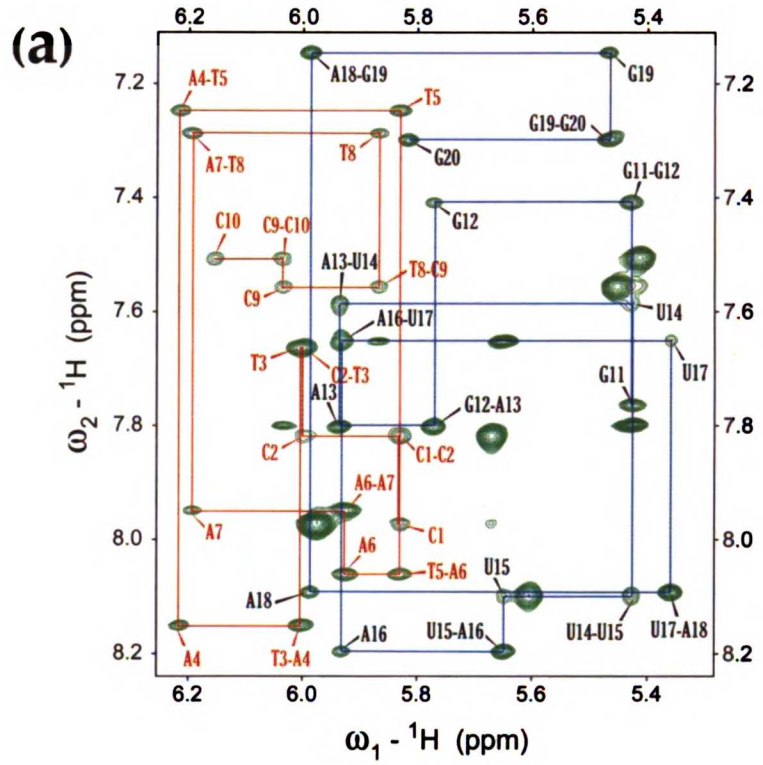


Figure 4.3a. Portion of the 200ms NOESY spectrum for the [Sp]-hybrid, showing the sequential base-H1' fingerprint walks for the PSO (red) and RNA (blue) strands. For convenience, only cross-peaks along the sequential walks are labeled.

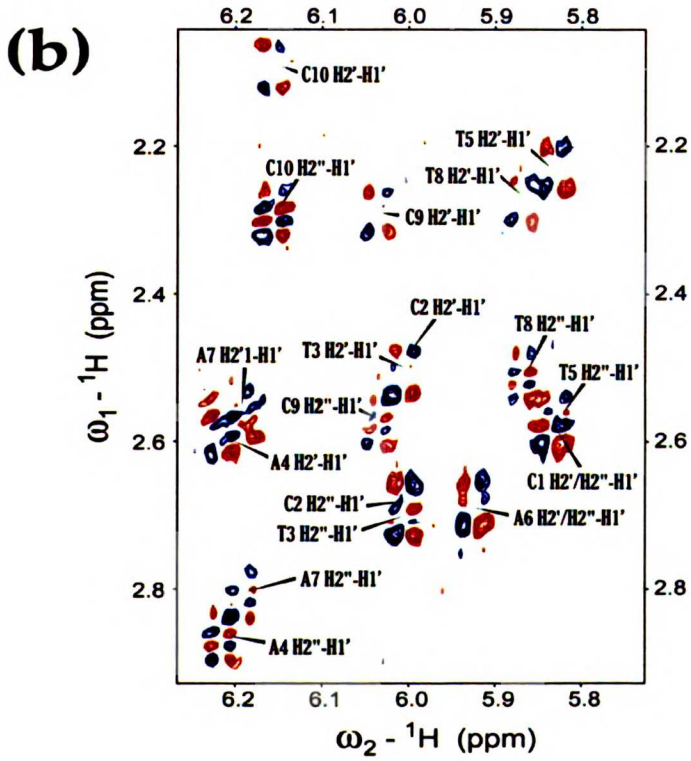


Figure 4.3b. Portion of the DQF-COSY spectrum for the [Sp]-hybrid, showing H1'-H2' and H1'-H2" cross-peaks in the PSO strand.

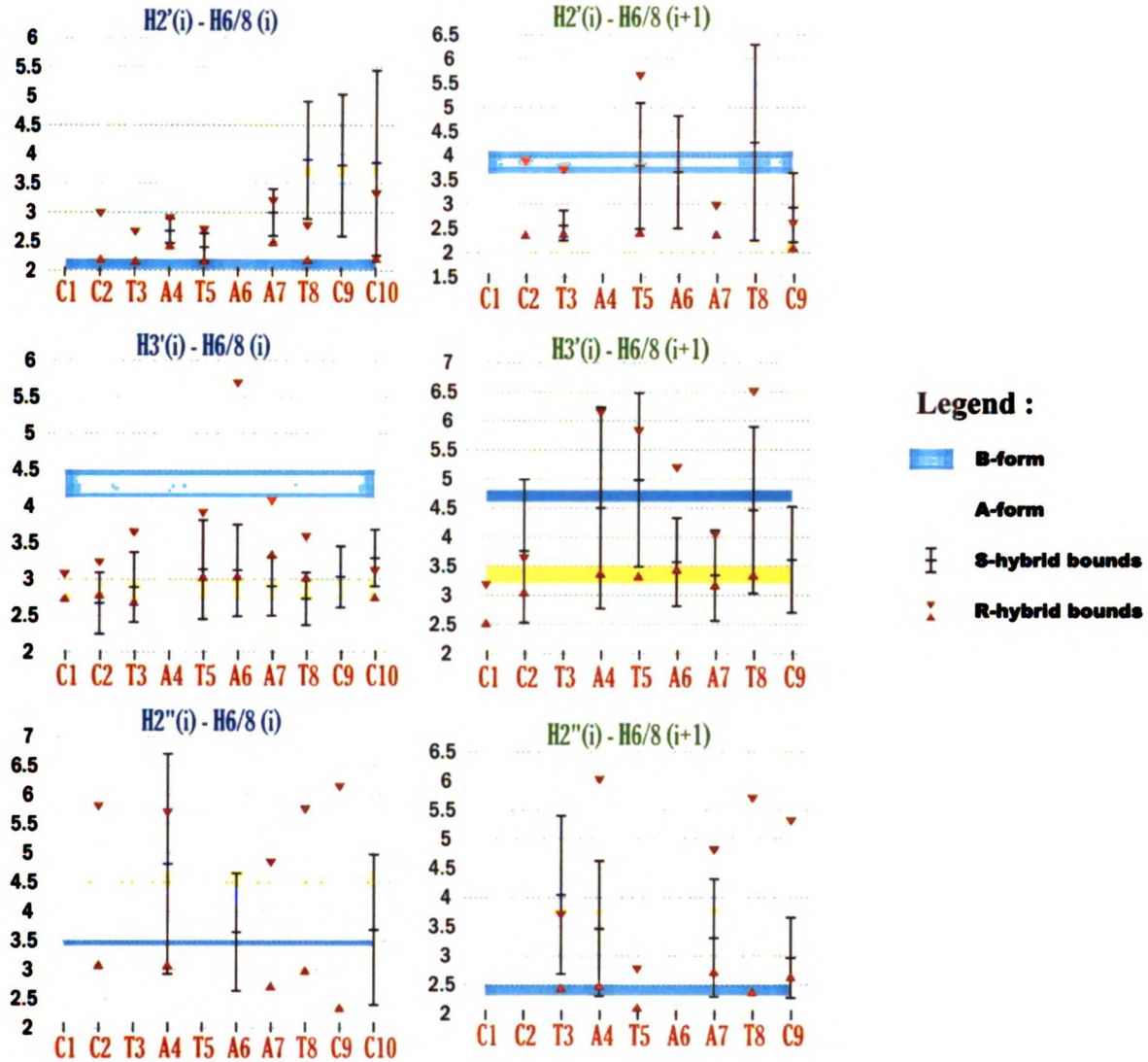


Figure 4.4. Intraresidue and sequential distance restraints H2'-H6/H8, H3'-H6/H8 and H2"-H6/H8 for [Rp]-PSO (red triangles) and [Sp]-PSO (black error bars) strands. The colored areas indicate distance ranges typical for B-form (yellow) and A-form (cyan) duplexes.

4.3.2 Structure calculation for the [Sp]-hybrid

Initially, we attempted to refine the [Sp]-hybrid by rMD in *vacuo* using a distance-dependent dielectric constant and big hydrated sodium ions to neutralize the charges in the phosphate groups. The structure appears to be stable at the beginning of the simulation, when no experimental restraints are enforced. However, as soon as we start ramping up the force constant for the NOE restraints, the sodium ions fly away and the

duplex conformation rapidly deteriorates and keeps fluctuating among bad geometries, indicating that, at these conditions, the restraints we apply are not able to define any reasonable structure.

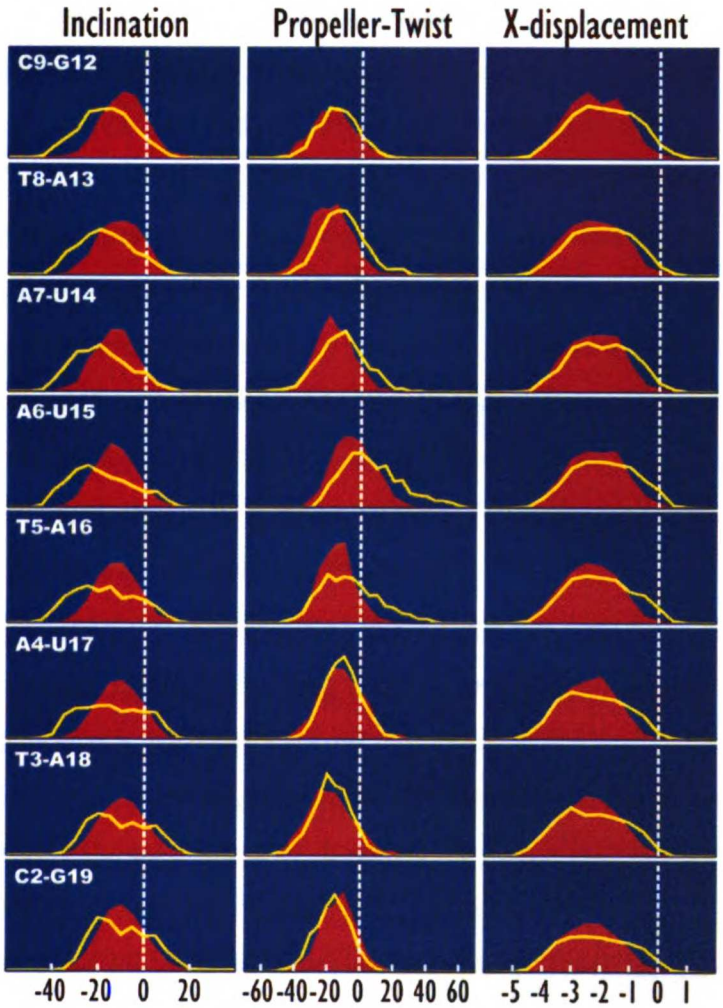


Figure 4.5. Relative population of Inclination, Propeller Twist and X-displacement for internal base pairs of the [Sp]-hybrid calculated from MDtar (red area) and rMD (yellow line) ensembles using CURVES.

rMD refinement of fully solvated [Sp]-hybrid with particle mesh Ewald (PME) model for calculation of electrostatic interactions failed to converge; average atomic RMSD deviation for snapshots taken each picosecond is 2.4 Å (Table 4.5). The structure continues drifting during 1 ns of simulation towards conformations with negative inclination (as low as -40.0 for T5-A16, A6-U15 and A7-U14 base pairs), positive propeller twist (with peaks higher than 50.0 for A6-U15 and 40.0 for T5-A16) and zero

x-displacement (Figure 4.5). In the course of the trajectory, the same base pairs are weakened, despite the use of hydrogen bond restraints. This is accompanied by sugar repuckering in some residues (e.g., T3, blue lines in the left panels of Figure 4.6a). At the end of the 1-ns simulation, the duplex appears elongated with very narrow minor groove and short cross-strand contacts (e.g., H1'(i) to H1'(j+2) close to 3 Å, where j is the nucleotides base-paired to i). Surprisingly, this conformational change is accompanied by an increase in the distance restraint violation and, hence, of the constraint energy (from 40 to 80 Kcal/mole A). All our attempts to stabilize the rMD simulations by changing the conditions, such as force constant for the NMR restraints, adding and modifying empirical restraints, have only been partially successful in slowing down the drifting of the molecule. In the end, the molecule is irreversibly driven towards structures with bad geometry. The average distance restraint deviation <Rdev> is quite high for these structures (0.14 Å). This is clearly a failed trajectory; we show some data for this trajectory only for the sake of comparison (Table 4.4).

MDtar trajectory of the fully solvated [Sp]-hybrid was run with PME electrostatics as described in Materials and Methods, starting with independent equilibration process. In contrast to rMD calculations, base pairs stayed intact during the course of the trajectory. Unlike rMD calculations, it is not required in MDtar simulations that all restraints are satisfied simultaneously for each snapshot. Instead, the distances are third-root averaged, with more recent snapshots weighted exponentially more heavily. Effectively, it is required that restraints are satisfied for the distances averaged during each fragment of trajectory equal in length to the exponential time constant for distance averaging (20 ps for current simulations). Correspondingly, to assess the quality of such a trajectory, the third-root averaged distances must be compared with the experimental restraints. Indeed, the corresponding distance deviation index, dev$\langle r^{-3} \rangle$ is very low for the trajectory (0.08 Å), while the distance deviations are very high for individual structures, with the average <Rdev> of 0.16 Å (Table 4.4). Another index assessing the quality of the refinement is sixth-root weighted R^x-factor, which directly compares calculated and experimentally measured NOE intensities. R^x-factors were calculated using CORMA for the two NOESY data sets, and their average values and standard deviations are reported in Table 4.4. These R^x-factors characterize how well each individual structure fits experimental

data. In addition, ensemble R^x-factors (Schmitz et al., 1992a) (ensR^x in Table 4.4) were calculated for each ensemble of structures, assuming fast exchange between individual conformers. Ensemble R^x-factor assesses quality of an ensemble as a whole, rather than each individual member of the ensemble. Based on these indices, we consider this refinement successful, and the MDtar trajectory a representation of [Sp]-hybrid in solution.

Table 4.4. Structural parameters for the rMD-refined, MDtar-refined and PDQPRO-selected-MDtar-refined structures for the [Rp]- and [Sp]-hybrids. Definition of structural parameters: $<E_1>$, average AMBER energy (kcal/mol) calculated on the whole solvated system with water molecules and counter ions; $<E_2>$, average AMBER energy (kcal/mol) calculated by stripping the water molecules and counter ions, adding big hydrated sodium ions to neutralize the hybrid and optimizing their position by energy minimization while keeping the hybrid fixed; Rdev, average deviation of distances from the corresponding NOE bounds; dev$<r^{-3}>$, average third-root deviation of distances from the corresponding NOE bounds; Rx, average of sixth-root CORMA R-factor calculated independently on each structure of the ensemble; ensRx, sixth-root CORMA ensemble R-factor. Where relevant, the standard deviation from the average is written within parentheses below the corresponding average value.

Ensemble	# pdb	$<E_1>$	$<E_2>$	Rdev	R^x (x100)		dev$<r^{-3}>$	ensR^x (x100)	
					(1)	(2)		(1)	(2)
[Sp]-hybrid:									
rMD	1000	-42196 (60)	-1147 (20)	0.141 (0.02)	8.83 (0.42)	8.49 (0.40)	0.095	7.31	7.04
MDtar	1000	-42199 (57)	-1153 (12)	0.161 (0.02)	8.99 (0.39)	8.64 (0.35)	0.081	6.84	6.72
PDQPRO	26[a]	-42208 (58)	-1153 (10)				0.075	6.36	6.38
[Rp]-hybrid:									
rMD	1000	-42248 (58)	-1168 (12)	0.095 (0.01)	8.34 (0.34)	7.03 (0.26)	0.076	7.42	6.20
MDtar	1000	-42238 (59)	-1151 (16)	0.116 (0.01)	8.64 (0.39)	7.43 (0.32)	0.063	7.05	5.91
PDQPRO	15[b]	-42233 (44)	-1148 (18)				0.065	6.79	5.80

[a] Sorted probabilities of 13 PDPQRO structures with probabilities above 3%: 10.8, 9.0, 8.7, 8.7, 6.7, 6.2, 5.3, 5.1, 4.1, 3.8, 3.6, 3.2, 3.1%. Thirteen more structures have probabilities between 0.1 and 2.7%.

[b] Sorted probabilities of 15 PDQPRO structures: 19.0, 15.0, 12.9, 10.7, 7.2, 6.5, 5.5, 5.2, 4.5, 4.3, 4.2, 3.0, 1.0, 0.6

Similarly to rMD, in MDtar the molecule drifts towards geometries with negative inclination, but to a lesser extent (-20.0 at most) and not irreversibly. In fact, during MDtar simulations, the molecule fluctuates between structures with negative and near-zero inclination, for an average value of -8.7 for the whole trajectory (versus -13.1 for rMD; Table 4.6). Furthermore, propeller twist becomes consistently positive only for A6-U15 base pair (up to ~20.0), while x-displacement remains negative (Figure 4.5). While for the RNA residues, sugars stayed in the C3'-*endo* conformations (Figure 4.6a, right panels), sugars for the PSO residues adopted multiple conformations (Figure 4.6a, red trace in the left panels and red histograms in the middle panels). Figure 4.7a shows distribution of Twist, Slide and Roll helical parameters calculated with the CURVES program; the distributions are mostly mono-modal for the [Sp]-hybrid.

One thousand saved pdb files from the MDtar trajectory were subjected to the PDQPRO analysis. PDPQRO finds a subset of structures and their probabilities so that the resulting ensemble produces the best fit of experimental NMR parameters, NOE-derived cross-relaxation rates. PDQPRO selected 26 structures out of 1000 with the probabilities ranging from 0.1 to 10.8%. Amber energies and various figures of merit are reported for the PDQPRO ensemble in Table 4.4; average values and standard deviations were calculated by taking into account the probability of each structure. Members of PDQPRO ensemble with probabilities greater than 3% are shown in Figure 4.8.

Table 4.5. Root mean square deviation within the ensemble of all rMD, MDtar or PDQPRO-selected structures for the [Sp]- and [Rp]-hybrids. The values calculated using all heavy atoms except those in terminal base pairs (all) and heavy atoms only in the DNA or RNA strands except the terminal base-pairs (labeled DNA and RNA, respectively) are reported in the table.

Ensemble	[Sp]-hybrid			[Rp]-hybrid		
	all	DNA	RNA	all	DNA	RNA
rMD	2.4 (1.3)	1.8 (0.9)	1.9 (1)	0.9 (0.2)	0.7 (0.2)	0.7 (0.2)
MDtar	1.4 (0.4)	1.2 (0.3)	1.1 (0.4)	1.5 (0.5)	1.2 (0.4)	1.0 (0.4)
PDQPRO	1.3 (0.4)	1.2 (0.3)	1.1 (0.4)	1.5 (0.5)	1.2 (0.4)	1.0 (0.4)

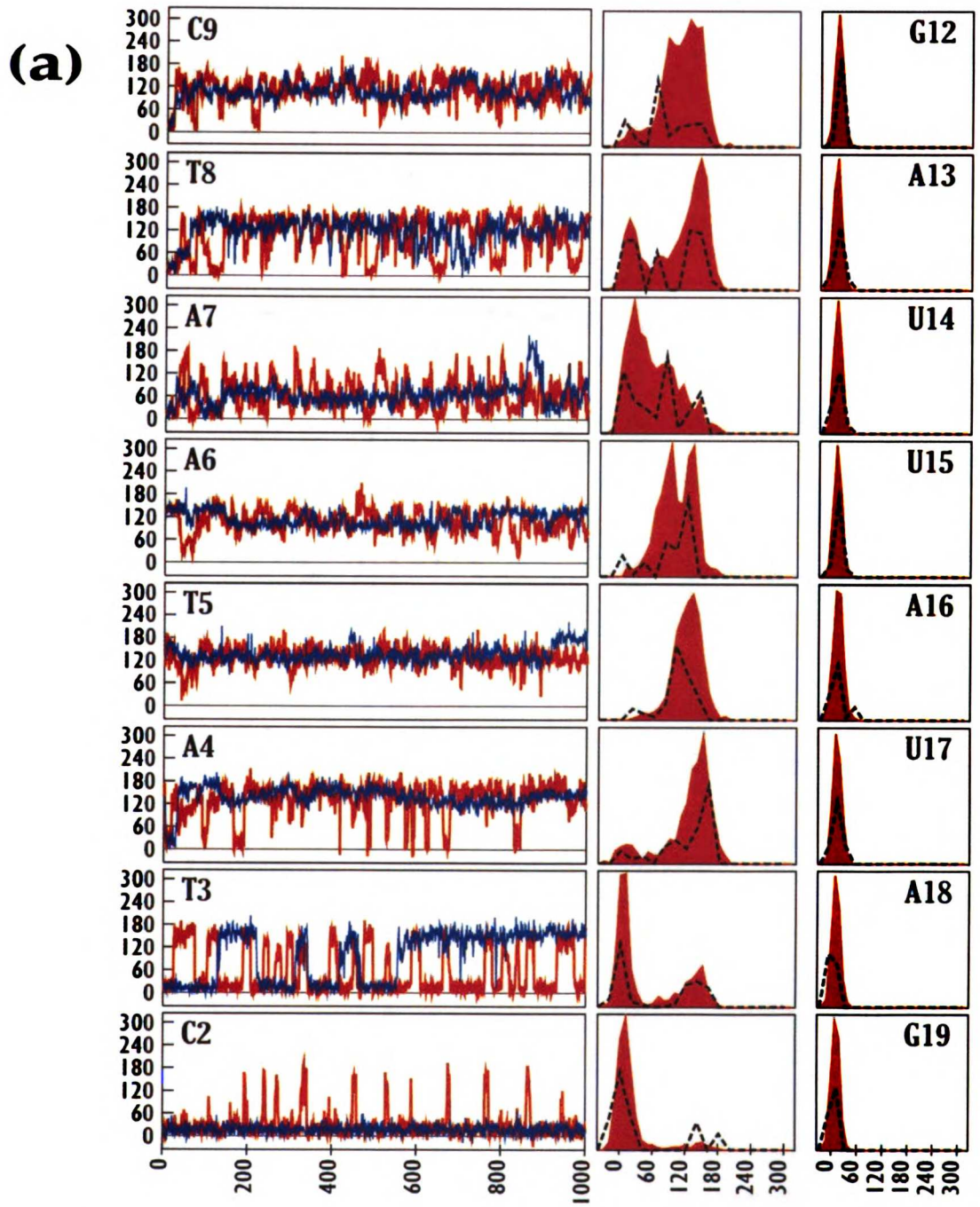
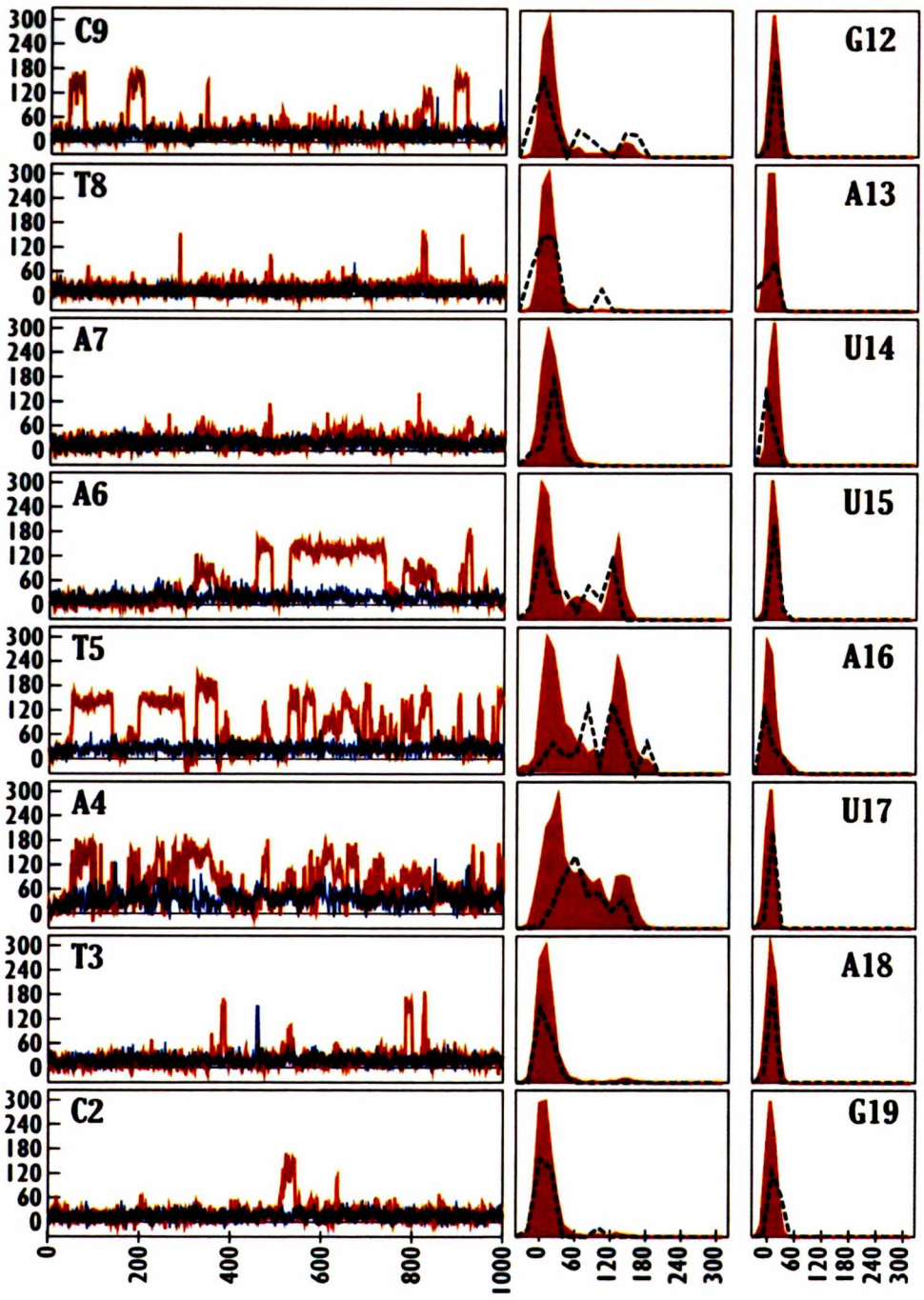


Figure 4.6. Sugar pucker analysis: (a) [Sp]-hybrid; (b) [Rp]-hybrid (next page). Left panels: time courses of the pseudorotation phase angle (**P**) for internal nucleotides of the PSO strand measured from the rMD (blue lines) and MDtar (red lines) production runs. Center panels: relative population of P values for internal nucleotides of the PSO strand measured from the MDtar ensemble (red histograms) and structures selected by PDQPRO from the MDtar ensemble (black dotted lines). The PDQPRO distributions are weighted based on PDQPRO probabilities. Right panel: similar relative population of P as shown in the center panels but measured for internal nucleotides of the RNA strand, instead.

hy-

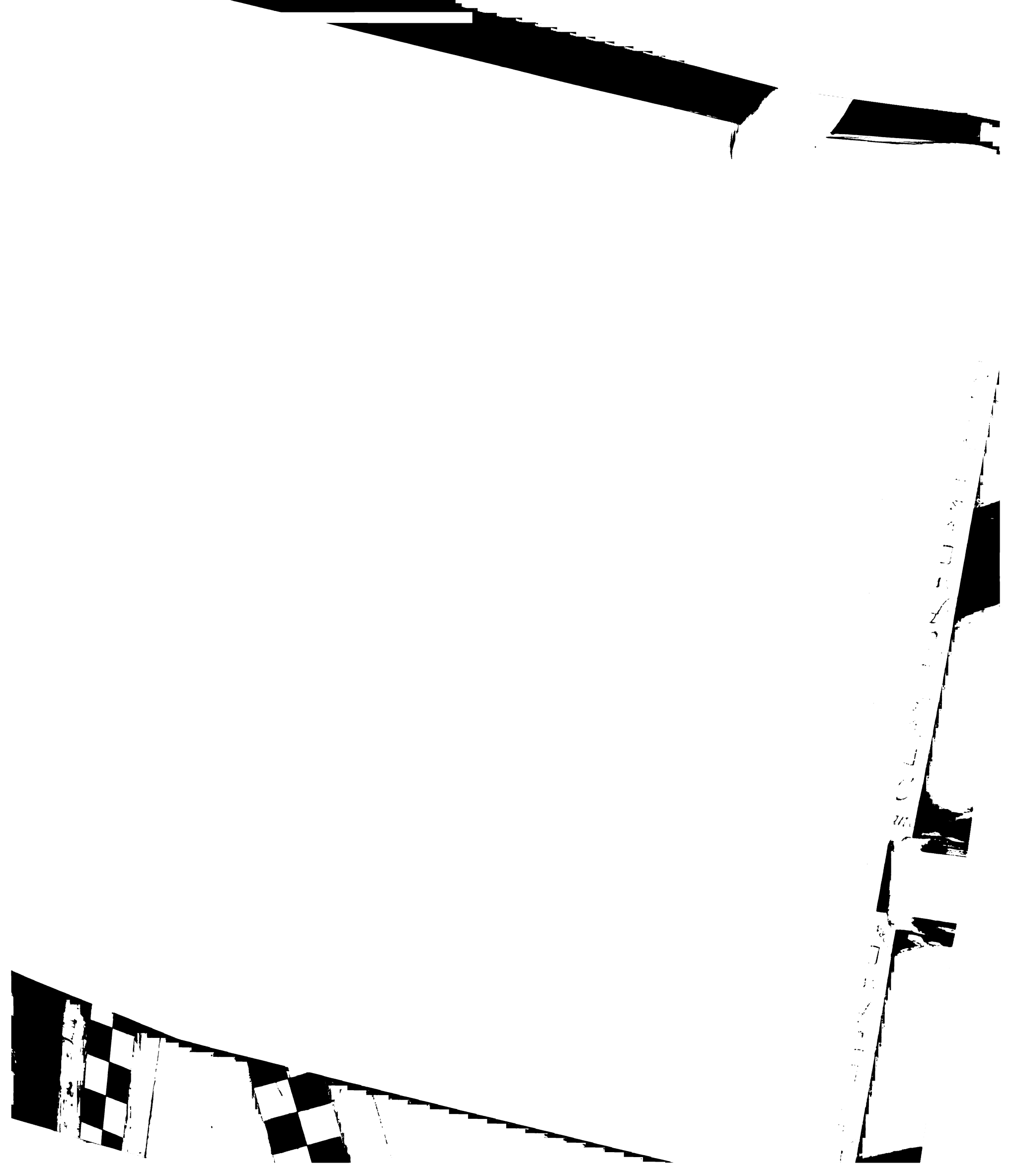
sim

(b)



4.3.3 Structure calculation for the [Rp]-hybrid

Calculations for this molecule were carried out similarly to the [Sp]-hybrid. The [Rp]-hybrid seems to behave better during PME rMD and MDtar production runs. Neither simulation yielded unreasonable duplex geometries, allowing us to use a stronger force



constant for NMR restraints (Table 4.3). Consequently, the refined structure for the [Rp]-hybrid features lower distance deviation values and lower R factors compared to the [Sp]-hybrid.

The <Rdev> index for rMD on the [Rp]-hybrid is 0.10 Å, which could be considered acceptable. Still, the MDtar ensemble description of the molecules offers a significant improvement in figures of merit. Index dev$<r^{-3}>$ for MDtar is 34% lower than <Rdev> for rMD, and ensemble R^x-factor for MDtar is 16-18% lower than average R^x-factor for rMD (the range is given for the two data sets, at 150 and 300 ms). Thus, we consider MDtar trajectory a more accurate description of the [Rp]-hybrid in solution than a conventional rMD description. PDQPRO calculations selected 15 conformers out of the MDtar trajectory; probabilities of individual conformers varied from below 1 to 19%. Parameters of the PDQPRO ensemble are reported in Table 4.4. Figures 4.6b and 4.7b show distributions of sugar conformations and helical parameters for the MDtar and PDQPRO ensembles for the [Rp]-hybrid.

Table 4.6. Average and standard deviation (within parenthesis) values for selected helical parameters calculated using CURVES (Lavery and Sklenar, 1996) from the ensemble of all rMD, MDtar or PDQPRO-selected structures for the [Sp]- and [Rp]-hybrids.

	Roll	Slide	Twist	Rise	Inclin	Propeller	X-displ
[Sp]-hybrid:							
rMD	0.8 (11.6)	-2.04 (0.48)	28.6 (4.7)	3.41 (0.38)	-13.1 (13.0)	-7.8 (15.6)	-1.83 (1.10)
MDtar	3.8 (9.8)	-1.86 (0.40)	29.6 (4.2)	3.34 (0.37)	-8.7 (9.1)	-12.0 (12.8)	-2.06 (0.91)
PDQPRO	4.9 (9.7)	-1.75 (0.43)	29.8 (4.6)	3.29 (0.38)	-5.1 (7.9)	-14.8 (12.1)	-2.35 (0.73)
[Rp]-hybrid:							
rMD	4.0 (6.7)	-1.88 (0.31)	31.0 (3.4)	3.22 (0.33)	-2.5 (6.0)	-20.9 (9.8)	-2.94 (0.62)
MDtar	3.9 (8.5)	-1.86 (0.40)	30.5 (4.5)	3.27 (0.36)	-7.8 (8.9)	-16.5 (11.9)	-2.20 (0.94)
PDQPRO	3.8 (8.6)	-1.88 (0.41)	30.0 (4.7)	3.26 (0.35)	-11.8 (7.5)	-15.9 (12.1)	-1.75 (0.75)

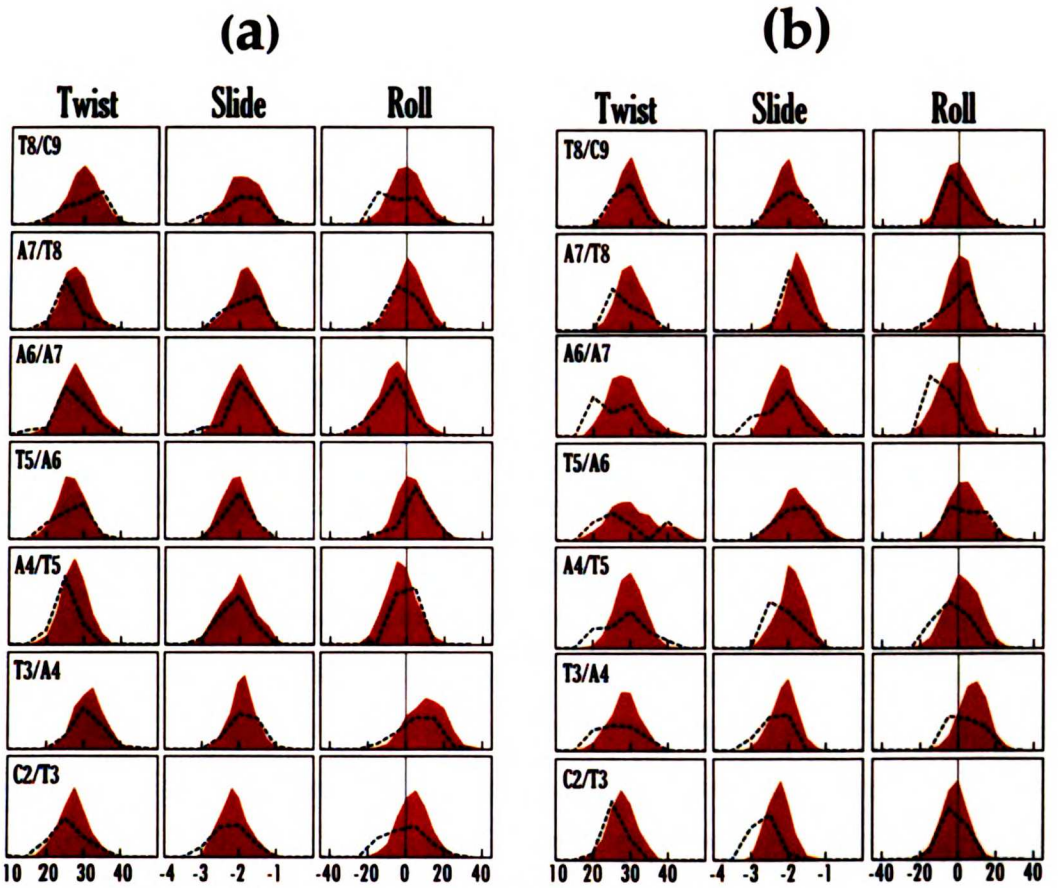


Figure 4.7. Relative population of Twist, Slide and Roll for internal steps calculated from MDtar (red area) and PDQPRO (black dotted line) ensembles: (a) [Sp]-hybrid; (b) [Rp]-hybrid. The PDQPRO distributions are weighted based on PDQPRO probabilities.

4.4 DISCUSSION

4.4.1 rMD versus MDtar calculations

Molecular dynamics with distance restraints is a robust refinement method of NMR structures for nucleic acids when a molecule possesses a well-defined structure in solution (Mujeeb et al., 1993; Schmitz et al., 1992b; Tonelli et al., 1998; Weisz et al., 1994). A successful rMD refinement leads to a set of similar structures; quality of refinement can be assessed by calculating residual violation of distance restraints (Rdev) and by comparing experimental and predicted NOE intensities (using, *e.g.*, sixth-root weighted R^x-factor). Often it is necessary to find an optimal balance between empirical

Chapt

T

bas

exp

notev

force field and experimental restraints, so that satisfying restraints does not compromise conformational energy. Other factors contributing to successful refinement include accurate experimental restraints and accurate empirical force field. To define a nucleic acid conformation, especially in the absence of residual dipolar couplings, distance bounds must be as tight as warranted by experimental data; such bounds can be determined by using full relaxation matrix analysis of NOE data, e.g., with MARDIGRAS. From the perspective of the empirical force field, it may be critical to use explicit solvent with the PME scheme, especially for RNA molecules, where hydroxyl groups of riboses may form spurious hydrogen bonds during *in vacuo* calculations (K-S Ryu, unpublished data). On the other hand, when rapidly exchanging multiple conformations contribute to the observed time-averaged NMR signal, the derived experimental restraints correspond to some virtual structure, which may or may not be significantly populated in solution. Practically, enforcing such restraints during conventional rMD refinement may lead to a number of different scenarios. If the virtual structure has in fact a relatively low energy, *i.e.*, represents one of the solution conformers, or if the empiric force field is not accurate enough to reflect the differences in the populations of solution conformers, then the refinement may produce a well-defined structure, and we will just miss the fact of the existence of multiple conformers. If the experimental restraints are not precise enough (we are assuming that they are accurate though), *e.g.*, if distance bounds are not very tight, then the resulting structure will not be very well defined, or it will be defined more by the empiric force field than by experimental data. The only scenario when we can infer the existence of multiple conformers based on conventional refinement is when the experimental restraints are sufficiently tight and the virtual structure has high conformational energy. Practically, this will be manifest by high conformational energy, high restraint energy, or both; increasing the weight of restraints would only deteriorate the quality of the structure. By this token, NMR data for the [Sp]-hybrid clearly cannot be fit by a single conformation. The rMD structures have high residual distance deviation Rdev and relatively high NOE-based R^x-factors for both data sets at 100 and 200 ms (Table 4.4). At the same time, as explained in Results, rMD produces structures of poor quality for the [Sp]-hybrid. It is noteworthy though that it is difficult to correlate the quality of the structure as assessed

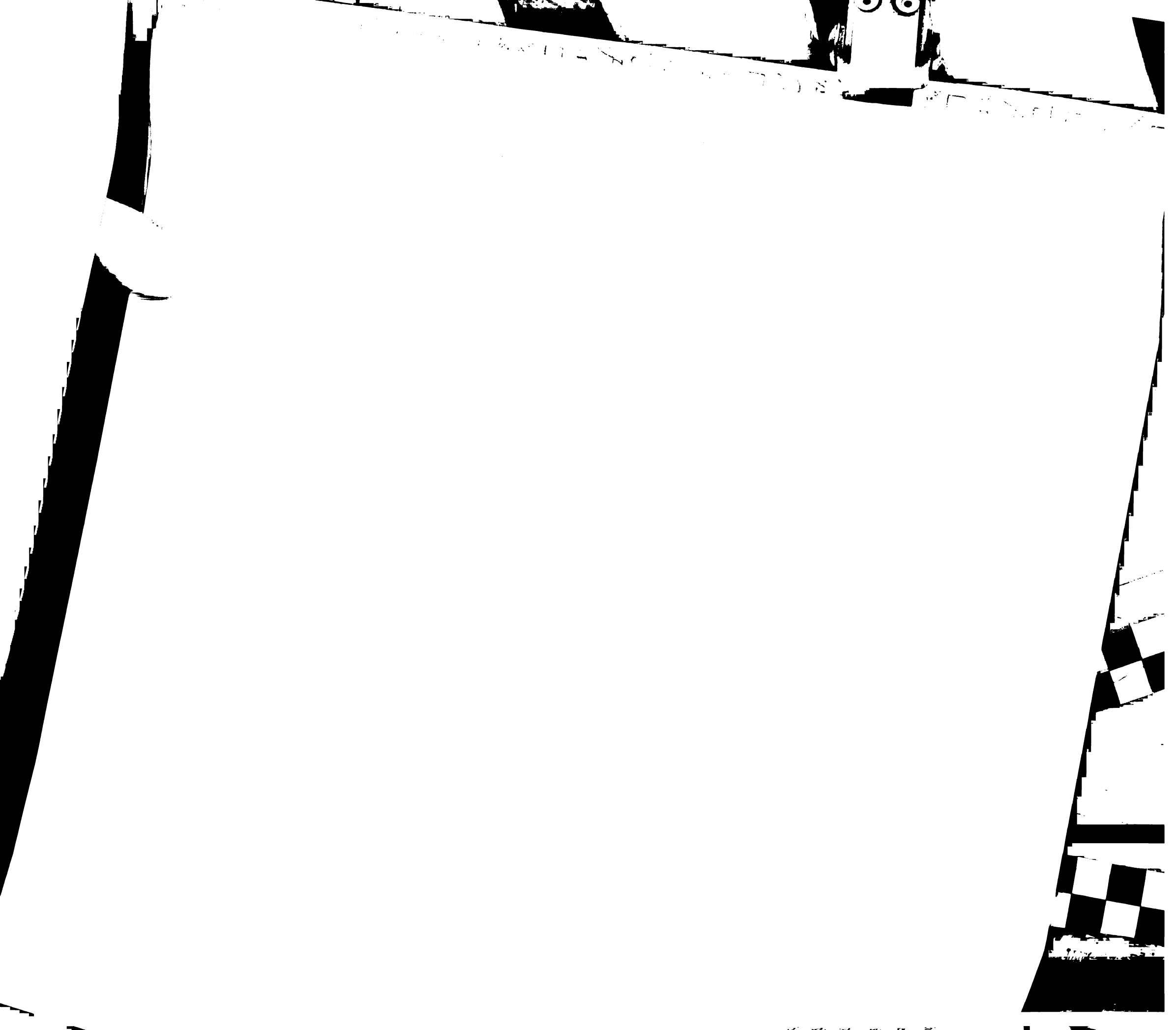
by visual inspection and analysis of helical parameters with the AMBER conformational energy (Table 4.4). Indeed, while the energy difference between feasible and unfeasible conformations may be of the order of 10^0 kcal/mole, the total AMBER energy is of the order of 10^3-10^4 kcal/mole.

There are several approaches proposed for NMR refinement of systems with multiple conformers (Bonvin and Brünger, 1995; Fennen et al., 1995; Görler et al., 2000; Gyi et al., 1998; Kemmink and Scheek, 1995; Pearlman, 1996; Torda et al., 1990; Ulyanov et al., 1995). We have successfully used MDtar method of Torda *et al.* in the past (Gonzàlez et al., 1995; Schmitz et al., 1993; Yao et al., 1997). Instead of enforcing all restraints simultaneously for each snapshot, it is only required in MDtar simulations that the restraints are satisfied over the course of trajectory for the appropriately averaged parameters, third-root averaged distances in this case. Individual MDtar snapshots have very high residual distance deviations for the [Sp] hybrid (average Rdev is 0.16 Å, Table 4.4), even higher than individual snapshots of rMD trajectory. However the distances third-root averaged over the whole MDtar trajectory have a low residual deviation of 0.08 Å (dev$<r^{-3}>$ in Table 4.4). Comparing the MDtar trajectory to individual snapshots of rMD with the average Rdev of 0.14 Å (rMD assumes that there is a single conformer), this is a 42% improvement. Ultimately, the quality of refinement must be judged by how well the refinement structures predict experimentally measured parameters, NOE intensities in this case. Ensemble R^x-factors, ensR^x, show a better than 20% improvement for the MDtar trajectory compared to average R^x-factors of rMD (Table 4.4). Interestingly, ensemble figures of merit, dev$<r^{-3}>$ and ensR^x, calculated for the rMD trajectory, also show some improvement compared to average values of Rdev and R^x for the same trajectory, although the restraints were imposed on each individual time frame and not on the distances time-averaged over the rMD trajectory. Apparently, this happened because even rMD simulations captured some degree of flexibility of [Sp]-hybrid, such as sugar flexibility in residues T3, A7, T8 (Figure 4.6a). However, the improvement is not as dramatic as in the case of MDtar, and the overall quality of structures was poor (Results).

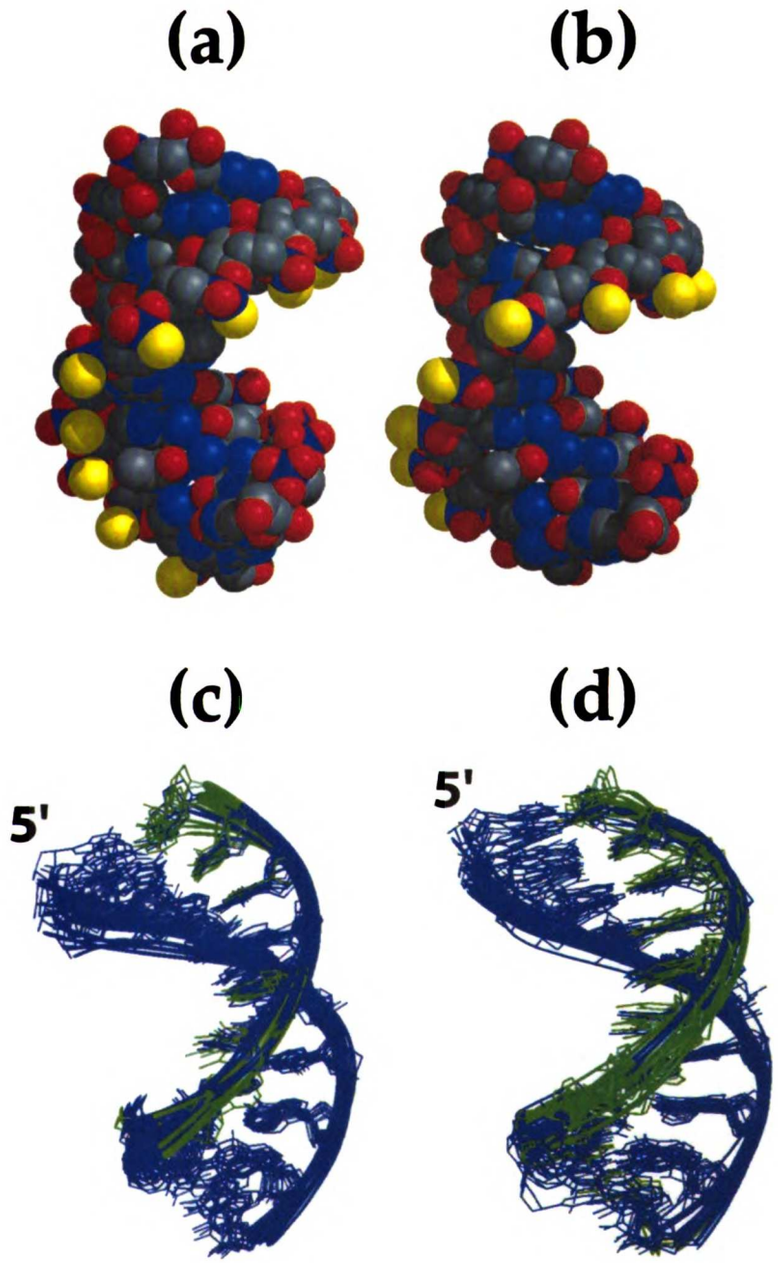
The situation is somewhat less clear in the case of the [Rp]-hybrid. To begin with, snapshots of the rMD trajectory do not have such high residual distance violation, 0.095

Å on average, and the overall quality of structures is better than in the case of the [Sp]-hybrid (Results). This could be because [Rp]-hybrid is indeed somewhat less flexible, or because the MARDIGRAS-calculated distance restraints are less tight (average flat-well width of 1.66 Å compared to 1.51 Å for the [Sp]-hybrid). Nevertheless, the MDtar trajectory shows a significant amount of flexibility for the [Rp]-hybrid as well (Figures 4.6b, 4.7b). Most importantly, ensemble figures of merit are still significantly improved for the MDtar compared to average individual figures of merit for the rMD: distance deviation by 34%, and NOE-based R^x-factors by 15% (Table 4.4).

It must be clear of course, that we do not have enough of experimental data to define each conformer to a high resolution, especially for very flexible molecules like the [Sp]-hybrid or [Rp]-hybrid, where much of flexibility occurs at nucleotide level with sugar repuckering of deoxyribose rings (see below). Instead, MDtar trajectory must be considered as a representative ensemble of structures capturing elements of flexibility necessary to fit the observed data. In this case, it was necessary to assume sugar flexibility in the PSO strands in order to explain certain experimental distance restraints, most notably intraresidue H2'-H6/H8 and H3'-H6/H8 (Results). Such representative ensemble is not unique; it can be selected in different ways; the ensemble of 1000 MDtar structures is likely to be redundant, in the sense that a smaller set of structures should be able to explain the experimental data. Apart from these considerations, 1000 structures are not very convenient to deal with when analyzing them. Previously, we have used PDQPRO in combination with MDtar to select a small ensemble capable of fitting the NMR data (Aramini et al., 2000; Schmitz et al., 1998). Given a pool of potential conformations, PDQPRO finds a subset of structures and their probabilities with the best fit to the experimental data. Even though there is no intrinsic mechanism in PDQPRO to reduce the size of the ensemble, in our experience, finding the best fit does reduce the potential pool significantly, most probably, due to the elimination of redundant structures. Out of 1000 MDtar structures, PDQPRO calculations selected 26 conformers for the [Sp]-hybrid and 15 for [Rp]-hybrid. During the selection process, PDQPRO optimizes a quadratic function of NOE-derived cross-relaxation rates, but not directly distance deviation or NOE R^x-factor. Because of that, it cannot be expected that ensemble figures of merit, dev$\langle r^{-3} \rangle$ and ensR^x will further improve; however, they did increase a little for



both hybrids compared to full MDtar trajectories (Table 4.4). Importantly, PDQPRO ensembles appear to be good representatives of the full MDtar trajectories: they have similar average and standard deviations of AMBER energy, almost identical average pair-wise atomic RMSD (Tables 4.4, 4.5). Even the distributions of sugar pseudorotation phase angles and helical parameters (Figures 4.6, 4.7) have many common features between MDtar and PDQPRO ensembles.



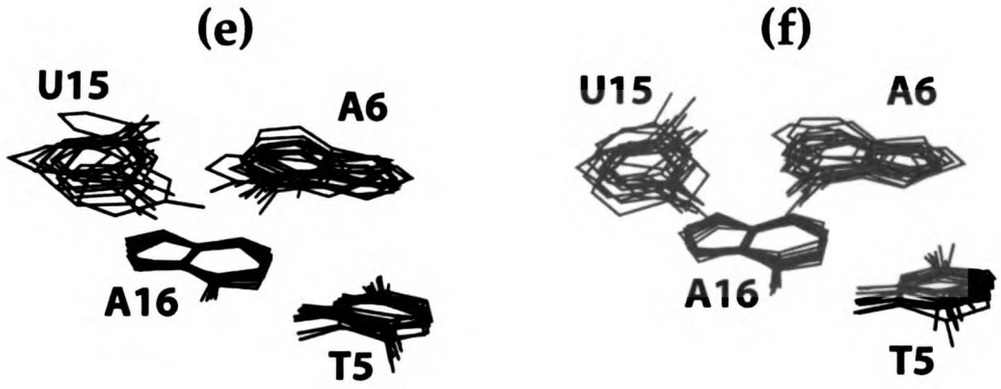


Figure 4.8. Molecular graphics showing (a) most populated PDQPRO conformer for the [Rp]-hybrid (19%); (b) most populated PDQPRO conformer for the [Sp]-hybrid (11%); (c and e) superposition of eleven PDQPRO conformers with populations greater than 3% for the [Rp]-hybrid; (d and f) superposition of thirteen PDQPRO conformers with populations greater than 3% for the [Sp]-hybrid. In (a) and (b) atoms are colored according to the atom type. In (c-f) residues with C3'-*endo* sugar pucker are colored blue, and residues with C2'-*endo* sugar pucker are colored green. In (e) and (f), bases T5 and A16 are superimposed.

4.4.2 *Sugar conformations*

The most striking feature of the solution ensembles for both hybrids is rigid C3'-*endo* conformations of riboses in RNA strands and very flexible deoxyriboses in PSO strands (Figures 4.6). This feature of PSO-RNA hybrid is common with regular DNA-RNA hybrids (Gonzàlez et al., 1995; Gyi et al., 1998). It has been known that deoxyriboses are also flexible in DNA-DNA duplexes (Mujeeb et al., 1992; Rinkel and Altona, 1987; Schmitz et al., 1992b; Schmitz et al., 1990; Tonelli and James, 1998; Tonelli et al., 1998; Weisz et al., 1992), however, the population of N-conformations is relatively small and often can be ignored during refinement without significant deterioration of quality of structures. In contrast, populations of both N- and S-puckers are high for deoxyriboses in both DNA-RNA and PSO-RNA duplexes. Ignoring sugar flexibility in this case, could either lead to completely unphysical structures (as in the case of rMD refinement of the [Sp]-hybrid), or lead to a sub-optimal virtual structure (as in the case of the [Rp]-hybrid); this has been also noted by others (Gyi et al., 1998). Flexibility of deoxyriboses appears to be more pronounced for the [Sp]-hybrid and to depend on sequence (Figure 4.6a). However, it is not clear to what degree this result depends on what types of distance

restraints were measured for each residue; some of the NOE cross-peaks could not be integrated because of spectral overlap. For regular DNA-RNA hybrids, sequence dependence of sugar flexibility has been reported before; pyrimidine-rich DNA appears to be more flexible than purine-rich DNA strands (Gyi et al., 1998).

4.4.3 *4.4.3 Helical parameters*

The overall geometry of [Sp]- and [Rp]-hybrids is shown in Figures 4.7a and 4.7b. The conformations of the two hybrids are close to each other, with the atomic RMSD between the most probable PDQPRO conformers of 1.3 Å. The sulfur atoms are pointing toward solution in the [Sp]-hybrid; they are directed more toward the major groove in the [Rp]-hybrid, but they are far from making any specific interactions with atoms from the other strand or from the walls of the groove. The global conformational parameters of the hybrids are clearly intermediate between those typical for classical B- and A-conformations (Table 4.6). X-displacement is ca. –2 Å; it still creates a characteristic central hole in the duplex, but not as large as in the canonical A-conformation (a typical X-displacement for A-forms is ca. –4 Å). Inclination of base pairs relative to the global axis of the duplex is negative, which is typical of B-conformations; however, the variation of this parameter between individual conformers is very high for both hybrids, with the standard deviation of 8-9°. Local helical parameters are sequence-dependent and exhibit a high degree of variation between conformers; some of them appear to have bimodal distributions (Figure 4.7). Of note is, *e.g.*, the T5-A6 step, which has low-twist and high-twist conformers. For the [Rp]-hybrid, twist varies between 22 and 40° for this step; for the [Sp]-hybrid, the range is 16—34°. This step precedes a stretch of two adenines; in DNA, TA steps preceding a stretch of adenines have been shown to undergo conformational averaging (Kennedy et al., 1993; McAteer et al., 1995; McAteer and Kennedy, 2000; Schmitz et al., 1992b), although the structural detail of this dynamic process are not known.

4.4.4 *Implications for RNase H recognition*

DNA-RNA and PSO-RNA hybrid duplexes are recognized by RNase H with the subsequent cleavage of RNA residues. All these types of hybrids are extremely flexible in

solution, with deoxyriboses of DNA or PSO strands undergoing conformational switches between C3'-*endo* and C2'-*endo* puckers. On the other hand, RNA-RNA duplexes with rigid C3'-*endo* riboses are not substrates for RNase H. Furthermore, introducing 2'-substituents that lock deoxyriboses in C3'-*endo* conformations and result in rigid DNA-RNA structures, also result in the disappearance of RNase H recognition [Ryu, unpublished data]. We propose a hypothesis that it is the flexibility of the hybrid duplex required for the RNase H recognition. This hypothesis is consistent with the recent crystal structure of a complex of HIV-1 reverse transcriptase and a DNA-RNA hybrid duplex (Sarafianos et al., 2001). As expected, all RNA residues in this hybrid have C3'-*endo* sugar puckers. On the other hand, the DNA residues clearly exhibit conformational flexibility: part of the hybrid duplex interacting with the polymerase domain has deoxyriboses in C3'-*endo* conformation, while DNA residues interacting with the RNase H domain have C2'-*endo* sugar puckers. Base pairs X-displacement for the portion of the hybrid interacting with RNase is between –2.3 and –2.0 Å and the inclination is ca. –4° (Sarafianos et al., 2001), not unlike the values determined in this work for both PSO-RNA hybrid duplexes. Furthermore, the irregularities in the base pairing of the hybrid DNA-RNA duplex in the vicinity of the RNase H active site (Sarafianos et al., 2001) suggest, that the helical conformation of the hybrid is also very flexible, similar or even higher than observed for the PSO-RNA hybrids in this work.

4.5 CONCLUSIONS AND FUTURE DIRECTIONS

We have used homonuclear 2D NMR to solve the structure of an [Sp]-phophorothioated DNA-RNA hybrid duplex and the corresponding [Rp]-isomer with the same sequence. Our results clearly indicate that these hybrids are very flexible. Indeed, the [Sp]-hybrid required the use of time-averaged restraints during MD simulations in order to get an ensemble of structures capable of describing the experimental data. The [Rp]-hybrid appears to be less flexible, so we were able to produce stable trajectories even with regular rMD simulations. It did, however, show a significant amount of flexibility when MDtar was used, yielding an ensemble of structures that fit the NMR data better (see Results).

Furthermore, the flexibility that we observe for the hybrids is not only localized at the sugars of the DNA strands, that undergo repuckering, but also involves the overall helical geometry as indicated by the high variability of helical parameters in our final ensembles of structures. Indeed, some helical parameters even exhibited a bimodal distribution (see Results and Discussion).

It is, of course, possible that this conformational variability is not real, but rather is the indirect result of the paucity of our NMR restraints that are relatively few in number and have quite large flat-well widths (the average widths are 1.51 Å and 1.66 Å in the [Sp]- and [Rp]-hybrids, respectively). In this regard, it is interesting to notice that the sugar rings in the RNA strand of both hybrids have C3'-*endo* conformations and don't show any sign of repuckering, even though, they are restrained by fewer distance bounds that the DNA deoxyribose sugars (35% and 18% less distance restraints for the RNA strand in the [Sp]- and [Rp]-hybrid, respectively). Moreover, the crystal structure of a DNA-RNA hybrid bound to HIV-1 reverse transriptase (Sarafianos et al., 2001) shows a great deal of conformational heterogeneity, not only limited to the pucker of the DNA sugars but also involving the helical geometry, that resembles the structures of our hydrids (see Discussion).

Hence, we conclude that our PSO-RNA hybrids must indeed be flexible both at the sugar ring level and global helical geometry. Furthermore, also taking into account that RNA-RNA duplexes with rigid riboses are not substrates for RNase H, we hypothesize that sugar and helical flexibility must be critical for the RNase H to recognize duplexes.

4.6 ACKNOWLEDGEMENTS

I gratefully thank Dr. Nick Ulyanov for essential help in producing and analyzing the data, Drs. Boleslaw Karwowski and Wojciech J. Stech for providing the purified [Rp]- and [Sp]-PSO strands and Dr. Vladimir Basus for help running the NMR machines.

4.7 REFERENCES

Agrawal, S., and Iyer, R. P. (1997). Perspectives in antisense therapeutics. Pharmacol Ther *76*, 151-160.

Altmann, K.-H., Fabbrot, D., Dean, N. M., Geigert, T., Monia, B. P., Müller, M., and Nicklin, P. (1997). Second-Generation Antisense Oligonucleotides: Structure-Activity Relationships and the Design of Improved Signal-Transduction Inhibitors. Biochemical Society Transactions *24*, 630-637.

Aramini, J. M., Mujeeb, A., Ulyanov, N. B., and Germann, M. W. (2000). Conformational dynamics in mixed alpha/beta-oligonucleotides containing polarity reversals: A molecular dynamics study using time-averaged restraints. J Biomol NMR *18*, 287-302.

Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., Di Nola, A., and Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. J Chem Phys *81*, 3684-3690.

Bonvin, A. M. J. J., and Brünger, A. T. (1995). Conformational Variability of Solution Nuclear Magnetic Resonance Structures. Journal of Molecular Biology *250*, 80-93.

Borgias, B. A., and James, T. L. (1990). MARDIGRAS--Procedure for Matrix Analysis of Relaxation for DIscerning Geometry of an Aqueous Structure. J Magn Reson *87*, 475-487.

Case, D. A., Pearlman, D. A., Caldwell, J. W., Cheatham, I. T. E., Ross, W. S., Simmerling, C., Darden, T., Merz, K. M., Stanton, R. V., Chen, A., *et al.* (2000). Amber 6.0, University of California, San Francisco. Amber 60, University of California, San Francisco.

Crooke, S. T., and Bennett, C. F. (1996). Progress in antisense oligonucleotide therapeutics. Annu Rev Pharmacol Toxicol *36*, 107-129.

Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J., and Bax, A. (1995). NMRPipe: A Multidimensional Spectral Processing System Based on UNIX Pipes. J Biomolec NMR *6*, 277-293.

DeLong, R. K., Nolting, A., Fisher, M., Chen, Q., Wickstrom, E., Kligshteyn, M., Demirdji, S., Caruthers, M., and Juliano, R. L. (1997). Comparative pharmacokinetics, tissue distribution, and tumor accumulation of phosphorothioate, phosphorodithioate, and methylphosphonate oligonucleotides in nude mice. Antisense Nucleic Acid Drug Dev *7*, 71-77.

Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995). A smooth particle mesh Ewald method. J Chem Phys *103*, 8577-8593.

Fennen, J., Torda, A. E., and Gunsteren, W. F. v. (1995). Structure Refinement with Molecular Dynamics and a Boltzmann-Weighted Ensemble. J Biomolec NMR *6*, 163-170.

Ferrin, T. E., Huang, C. C., Jarvis, L. E., and Langridge, R. (1988). The MIDAS display system. Mol Graphics *6*, 13-27.

Freier, S. M., and Altmann, K. H. (1997). The ups and downs of nucleic acid duplex stability: structure-stability studies on chemically-modified DNA:RNA duplexes. Nucleic Acids Research *25*, 4429-4443.

Furrer, P., Billeci, T. M., Donati, A., Kojima, C., Karwowski, B., Sierzchala, A., Stec, W., and James, T. L. (1999). Structural Effect of Complete R_P Phosphorothioate and Phosphorodithioate Substitutions in the DNA Strand of a Model Antisense Inhibitor-Target RNA Complex. J Mol Biol *285*, 1609-1622.

Goddard, T. D., and Kneller, D. G. (1998). SPARKY (University of California, San Francisco).

Gonzàlez, C., Stec, W., Kobylanska, A., Hogrefe, R., Reynolds, M., and James, T. L. (1994). Structural Study of a DNA-RNA Hybrid Duplex with a Chiral Phosphorothioate Moiety by NMR: Extraction of Distance and Torsion Angle Constraints and Imino Proton Exchange Rates. Biochemistry *33*, 11062-11072.

Gonzàlez, C., Stec, W., Reynolds, M., and James, T. L. (1995). Structure and Dynamics of a DNA-RNA Hybrid Duplex with a Chiral Phosphorothioate Moiety. NMR and Molecular Dynamics with Conventional and Time-averaged Restraints. Biochemistry *34*, 4969-4982.

Görler, A., Ulyanov, N. B., and James, T. L. (2000). Determination of the Populations and Structures of Multiple Conformers in an Ensemble from NMR Data: Multiple-Copy Refinement of Nucleic Acid Structures using Floating Weights. J Biomol NMR *16*, 147-164.

Gyi, J. I., Lane, A. N., Conn, G. L., and Brown, T. (1998). Solution Structures of DNA□RNA Hybrids with Purine-Rich and Pyrimidine-Rich Strands: Comparison with the Homologous DNA and RNA Duplexes. Biochem *37*, 73-80.

Huang, C. C., Couch, G. S., Pettersen, E. F., and Ferrin, T. E. (1996). Chimera: An Extensible Molecular Modeling Application Constructed Using Standard Components. Pacific Symposium on Biocomputing *1*, 724.

Keepers, J. W., and James, T. L. (1984). A Theoretical Study of Distance Determinations from NMR. Two-Dimensional Nuclear Overhauser Effect Spectra. J Magn Reson *57*, 404-426.

Kemmink, J., and Scheek, R. M. (1995). Dynamic Modelling of a Helical Peptide in Solution Using NMR Data: Multiple Conformations and Multi-Spin Effects. J Biomolec NMR *5*, 33-40.

Kennedy, M. A., Nuutero, S. T., Davis, J. T., Drobny, G., and Reid, B. R. (1993). Mobility at the TpA Cleavage Site in the T_3A_3-Containing *Aha*III and *Pme* Restriction Sequences. Biochemistry *32*, 8022-8035.

Koradi, R., Billeter, M., and Wüthrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. J Mol GraphJournal of Molecular Graphics *14*, 51-55, 29-32.

Krakowiak, A., Owczarek, A., Koziolkiewicz, M., and Stec, W. J. (2002). Stereochemical Course of Escherichia coli RNase H. Chembiochem *3*, 1242-1250.

Lavery, R., and Sklenar, H. (1996). CURVES 5.1. Helical Analysis of Irregular Nucleic Acids (Paris, Laboratoire de Biochimie Theoretique, CNRS).

Lima, W. F., and Crooke, S. T. (1997). Binding Affinity and Specificity of *Escherichia coli* RNase H1: Impact on the Kinetics of Catalysis of Antisense Oligonucleotide-RNA Hybrids. Biochemistry *36*, 390-398.

Liu, H., Spielmann, H. P., Ulyanov, N. B., Wemmer, D. E., and James, T. L. (1995). Interproton Distance Bounds from 2D-NOE Intensities: Effect of Experimental Noise and Peak Integration Errors. J Biomolec NMR *6*, 390-402.

Marquis, J. K., and Grindel, J. M. (2000). Toxicological evaluation of oligonucleotide therapeutics. Curr Opin Mol Ther *2*, 258-263.

McAteer, K., Ellis, P. D., and Kennedy, M. A. (1995). The effects of sequence context on base dynamics at TpA steps in DNA studied by NMR. Nucleic Acids Res *23*, 3962-3966.

McAteer, K., and Kennedy, M. A. (2000). NMR evidence for base dynamics at all TpA steps in DNA. J Biomol Struct Dyn *17*, 1001-1009.

McKay, R. A., Miraglia, L. J., Cummins, L. L., Owens, S. R., Sasmor, H., and Dean, N. M. (1999). Characterization of a potent and specific class of antisense oligonucleotide inhibitor of human protein kinase C-alpha expression. J Biol Chem *274*, 1715-1722.

Monia, B. P. (1997). First- and Second-Generation Antisense Inhibitors Targeted to Human c-*raf* Kinase: *in vitro* and *in vivo* Studies. Anti-Cancer Drug Design *12*, 327-339.

Mujeeb, A., Kerwin, S. M., Egan, W., Kenyon, G. L., and James, T. L. (1992). A Potential Gene Target in HIV-1: Rationale, Selection of a Conserved Sequence, and Determination of NMR Distance and Torsion Angle Constraints. Biochemistry *31*, 9325-9338.

Mujeeb, A., Kerwin, S. M., Kenyon, G. L., and James, T. L. (1993). Solution Structure of a Conserved DNA Sequence from the HIV-1 Genome:Restrained Molecular Dynamics Simulation with Distance and Torsion Angle Restraints Derived from 2D NMR Spectra. Biochemistry *32*, 13419-13431.

Pearlman, D. A. (1996). FINGAR: A New Genetic Algorithm-Based Method for Fitting NMR Data. J Biomolec NMR *8*, 49-66.

Pearlman, D. A., and Kollman, P. A. (1991). Are Time-Averaged Restraints Necessary for NMR Refinement? A Model Study for DNA. J Mol Biol *220*, 457-479.

Rinkel, L. J., and Altona, C. (1987). Conformational Analysis of the Deoxyribofuranose Ring in DNA by means of Sums of Proton-proton Coupling Constants: A Graphical Analysis. J Biomol Struct Dyn *4*, 621-649.

Ryckaert, J. P., Cicotti, G., and Berendsen, H. J. C. (1977). Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Comput Phys *23*, 327-341.

Sarafianos, S. G., Das, K., Tantillo, C., Clark, A. D., Jr., Ding, J., Whitcomb, J. M., Boyer, P. L., Hughes, S. H., and Arnold, E. (2001). Crystal structure of HIV-1 reverse transcriptase in complex with a polypurine tract RNA:DNA. Embo J *20*, 1449-1461.

Schmitz, U., Donati, A., James, T. L., Ulyanov, N. B., and Yao, L. (1998). Small structural ensembles for a 17-nucleotide mimic of the tRNA T psi C-loop via fitting dipolar relaxation rates with the quadratic programming algorithm. Biopolymers *46*, 329-342.

Schmitz, U., Kumar, A., and James, T. L. (1992a). Dynamic Interpretation of NMR Structural Data: Molecular Dynamics with Weighted Time-averaged Restraints and Ensemble R-Factor. J Am Chem Soc *114*, 10654-10656.

Schmitz, U., Sethson, I., Egan, W., and James, T. L. (1992b). Solution Structure of a DNA Octamer Containing the Pribnow Box via Restrained Molecular Dynamics Simulation with Distance and Torsion Angle Constraints Derived from 2D NMR Spectral Fitting. J Mol Biol *227*, 510-531.

Schmitz, U., Ulyanov, N. B., Kumar, A., and James, T. L. (1993). Molecular Dynamics with Weighted Time-averaged Restraints of a DNA Octamer: Dynamic Interpretation of NMR Data. J Mol Biol *234*, 373-389.

Schmitz, U., Zon, G., and James, T. L. (1990). Deoxyribose Conformation In $[d(GTATATAC)]_2$: Evaluation of Sugar Pucker by Simulation of Double-Quantum-Filtered COSY Cross-Peaks. Biochemistry *29*, 2357-2368.

Simmerling, C., Elber, R., and Zhang, J. (1995). MOIL-View- a program for visualization of structure and dynamics of biomolecules and STO - a program for

computing stochastic paths. In Modeling of biomolecular structures and mechanisms, A. Pullman, ed. (Kluwer, Netherlands, Academic Press), pp. 241-265.

Temsamani, J., and Guinot, P. (1997). Antisense oligonucleotides: a new therapeutic approach. Biotechnol Appl Biochem *26*, 65-71.

Tonelli, M., and James, T. L. (1998). Insights into the Dynamic Nature of DNA Duplex Structure Via Analysis of Nuclear Overhauser Effect Intensities. Biochemistry *37*, 11478-11487.

Tonelli, M., Ragg, E., Bianucci, A. M., Lesiak, K., and James, T. L. (1998). Nuclear Magnetic Resonance Structure of d(GCATATGATAAG): A Consensus Sequence for Promoters Recognized by σ^K RNA Polymerase. Biochemistry *37*, 11745-11761.

Torda, A. E., Scheek, R. M., and van Gunsteren, W. F. (1990). Time-average Nuclear Overhauser Effect Distance Restraints Applied to Tendamistat. J Mol Biol *214*, 223-235.

Ulyanov, N. B., Schmitz, U., Kumar, A., and James, T. L. (1995). Probability Assessment of Conformational Ensembles: Sugar Repuckering in a DNA Duplex in Solution. Biophys J *68*, 13-24

Weisz, K., Shafer, R. H., Egan, W., and James, T. L. (1992). The Octamer Motif in Immunoglobulin Genes: Extraction of Structural Constraints from Two-Dimensional NMR Studies. Biochemistry *31*, 7477-7487.

Weisz, K., Shafer, R. H., Egan, W., and James, T. L. (1994). Solution Structure of the Octamer Motif in Immunoglobulin Genes via Restrained Molecular Dynamics Calculations. Biochemistry *33*, 354-366.

Wüthrich, K. (1986). NMR of Proteins and Nucleic Acids (New York, Wiley).

Yao, L. J., James, T. L., Kealey, J. T., Santi, D. V., and Schmitz, U. (1997). The Dynamic NMR Structure of the TΨC-Loop: Implications for the Specificity of tRNA Methylation. J Biomol NMR *9*, 229-244.

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 Nat Acad Sci USA *88*, 7046-7050.

CHAPTER 5

5. The Solution Structure of the Viral-Binding Domain of Tva, the Cellular Receptor for Subgroup A Avian Leukosis and Sarcoma Virus

The cellular receptor for subgroup A Avian Leukosis and Sarcoma Virus (ALSV-A) is Tva, which contains a motif related to repeats in the Low Density Lipoprotein Receptor (LDLR) ligand-binding domain (LBr) and which is necessary for viral entry. As observed with LBr repeats of LDLR, the 47 residue LBr domain of Tva (sTva47) requires calcium during oxidative folding to form the correct disulfide bonds. Furthermore, calcium enhances the structure of correctly oxidized sTva47, as well as its ability to bind the viral Envelope protein. However, solution NMR studies indicate that, even in the presence of excess calcium, sTva47 exists in an ensemble of conformations. Nonetheless, as reported here, the structure of the predominant sTva47 solution conformer closely resembles that of other LBr repeats, with identical SS binding topology and octahedral calcium coordination. The location of W48 and other critical residues on the surface suggests a region of the molecule necessary for Env binding and to mediate post-binding events important for ALSV-A cell entry.

Abbreviations. ALSV-A, subgroup A Avian Leukosis and Sarcoma Virus; Tva, receptor for ALSV-A; aa, amino acid; sTva, soluble Tva ectodomain; sTva47, 47aa soluble Tva LBr; Env, viral envelope fusion glycoprotein; SU, surface subunits of Env; TM, transmembrane subunits of Env; SU(A)-rIgG (SU(A)-immunoadhesin), soluble fusion of ALSV-A Env SU to constant region of immunoglobulin G; API, soluble trimeric ALSV-A Env; sTva-immunoadhesin, fusion of sTva to the constant region of immunoglobulin G; LDLR, low density lipoprotein receptor; LBr, ligand-binding repeat of LDLR; MBP, maltose binding protein; MBP-sTva47, fusion of sTva47 to MBP; HBS, HEPES buffered saline; EDTA, ethylenediaminetetraacetate; NMR, nuclear magnetic resonance; 2D, two-dimensional; 3D, three-dimensional; HSQC, heteronuclear single quantum coherence spectra; NOE, nuclear Overhauser enhancement; NOESY, NOE spectroscopy; COSY, correlation spectroscopy; rmsd, root mean square deviation.

5.1 INTRODUCTION

Enveloped viruses must fuse their viral membrane with the membrane of the host cell to initiate a productive infection. In retroviruses this membrane fusion is carried out by the viral envelope fusion glycoprotein (Env[1]), which is composed of surface (SU) and transmembrane (TM) subunits that form a trimer of SU/TM heterodimers. Env binds to cellular receptors through the SU subunit, mediating viral attachment, and generally defining target specificity. In many cases, following the initial interaction between Env and a cellular receptor, internalization from the cell surface is triggered by interaction with another cellular receptor (e.g. the Human Imunodeficiency Virus) or through endosomal internalization. A two-step conformational change mechanism is generally involved in converting Env to the fusogenic state. Initially, the surface subunits are partially disassembled (either by receptor binding or reduction in pH in an endosomal compartment), which acts to prime Env to undergo further conformational change. Final conversion to the fusogenic state is mediated by interaction with an additional receptor or a (further) drop in pH (White, 1992).

By contrast with the influenza A virus, which, upon endosomal internalization, uses the decrease in pH to trigger viral entry (Bullough et al., 1994; Carr et al., 1997), most retroviruses are believed to use a pH-independent mechanism where receptor binding directly triggers viral entry. Because of its apparent simplicity (i.e., utilization of a single receptor), the Avian Leukosis Sarcoma Virus (ALSV) has become a model system to study pH-independent viral cell entry. Surprisingly, recent data suggest that, for ALSV cell entry, receptor binding must be followed by a decrease in pH (Mothes et al., 2000). In any case, for ALSV, receptor binding serves to drive conformational changes necessary for viral cell entry.

The ALSV family is comprised of five major subgroups (A – E), that differ in host range, receptor binding, and interference patterns (Vogt, 1977). The receptor for subgroup A ALSV (Tva) has been cloned from both quail (Bates et al., 1993) and chicken (Young et al., 1993) and shown to be necessary for ALSV-A cell entry (Young et al., 1993). Tva encodes a small (83aa) ectodomain that contains a single copy of a cysteine-rich motif first identified as seven repeats in the ligand-binding domain of the

Low Density Lipoprotein Receptor (LDLR) (Bates et al., 1993). This ~40aa LDLR ligand-binding repeat (LBr) is sufficient to reconstitute ALSV-A cell entry (Rong and Bates, 1995). Mutational analysis of the cysteines in the Tva LDLR LBr suggested a first-third, second-fifth, and fourth-sixth disulfide bonding pattern (Belanger et al., 1995), which matches that later determined for the LBr of LDLR (Bieri et al., 1995a; Bieri et al., 1995b; Blacklow and Kim, 1996). Intriguingly, further mutational analysis has suggested that tryptophan 48 $(W48)^2$ serves a critical role in post-binding events required for the viral fusion protein to assume a liposome binding conformation (Hernandez et al., 1997) and for viral cell entry (Zingler and Young, 1996).

The binding of low density lipoprotein to the LDLR is dependent on the presence of calcium (Goldstein and Brown, 1974). Further, the LBr of LDLR receptor family require calcium during oxidative folding to selectively form the correct disulfide bonds (Atkins et al., 1998; Bieri et al., 1995b; Dolmer et al., 1998; North and Blacklow, 1999) and calcium stabilizes their structures (Bieri et al., 1998; Fass et al., 1997; Huang et al., 1999; North and Blacklow, 2000). We have shown a similar requirement for calcium in the bacterially-expressed soluble Tva LBr (sTva47) from the quail sequence. Calcium was necessary for the selective formation of the correct disulfide bonds during oxidative folding and enhanced its structural stability, as also observed by Wang et al (Wang et al., 2001). We have also measured the binding of sTva47 to an ALSV-A SU-immunoadhesin to be enhanced ~30-fold by the presence of calcium.

In this communication we present the three-dimensional structure of the major solution conformation assumed by sTva47 with excess calcium, solved using heteronuclear NMR techniques. Not surprisingly, we found that it closely resembles that of LDLR LBr domains, especially in the C-terminal half that shows the most sequence similarity and contains a calcium binding site. In fact, although calcium was not explicitly included during structural refinement, the four conserved acidic side chains and two backbone carbonyl oxygen atoms were restricted by the NMR restraints to a correct position for octahedral coordination. The structure reveals that residues known to be important for binding, comprise and cluster around a hydrophobic patch on the receptor surface

5.2 MATERIALS AND METHODS

5.2.1 Construction of sTva expression vectors.

The quail Tva LDLR LBr (Tva residues 11-51, Figure 1) was subcloned into pBluescript (Stratagene) using PCR to introduce an in-frame 5' BamHI and 3' stop codon and EcoRI site. These restriction sites were further used to subclone into pET12a (Stratagene) or MBP-fusion pMAL (New England Biolabs) vectors. All constructs were verified by sequencing.

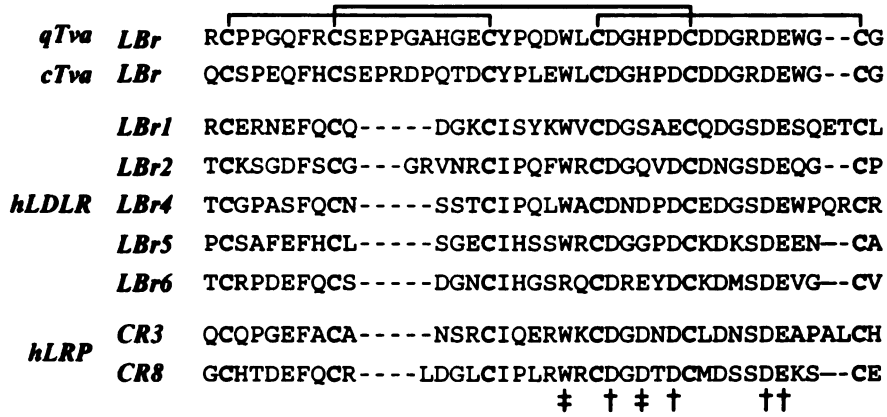


Figure 5.1. Tva contains an LBr. Alignment of Tva LBr (quail and chicken, qTva and cTva, respectively) with previously characterized LBr from human LDLR (hLDLR) and human LDLR related protein (hLRP). Conserved cysteines characteristic of LBr are in bold with the disulfide bond pairings connected above the alignment. The daggers below the alignment indicate the residues involved in calcium chelation (‡, coordination by backbone carbonyl; †, coordination by side chain carboxyl) as suggested by the crystal structure of hLDLR LBr5 (Fass et al., 1997).

5.2.2 Expression & Purification of MBP-sTva47.

The cytoplasmic MBP-sTva47 fusion yielded the best expression (~40% of total soluble protein) and was subsequently utilized, although the subcloning protocol leaves six residues (ISEFGS) on the N-terminus of the 41aa Tva LBr after Factor Xa release from MBP. Expression and purification was performed essentially following the recommended protocol. However, it was necessary to incorporate an anion exchange step prior to affinity purification. Clarified extract was loaded onto a HiTrap Q column (Pharmacia), washed and eluted with a 0-1 M NaCl gradient (MBP-sTva47 elutes in a broad peak from 150-450 mM). Because approximately half of the total expressed MBP-sTva47 is found in the flow through and wash from the amylose (affinity) column, these

were combined, diluted 1:1 with buffer, and rerun over the HiTrap Q and amylose columns, as above.

5.2.3 Oxidation and proteolytic release of sTva47.

Oxidation of sTva47 was routinely performed in the context of the fusion protein. Purified MBP-sTva47 was diluted to $A_{280} = 0.5$ with buffer and oxidized by the addition of 3 mM reduced glutathione, 0.3 mM oxidized glutathione, and 5 mM $CaCl_2$, followed by incubation at 4°C for at least 12 hours (Bieri et al., 1995a; Bieri et al., 1995b; Blacklow and Kim, 1996). Oxidized sTva47 could be released from MBP with Factor Xa (New England Biolabs) cleavage and separated by ultrafiltration through a 10 kDa molecular weight cutoff membrane (Amicon). The released sTva47 in the filtrate was then polished on a preparative scale C18 reversed phase column (Vydac) using a Pharmacia AKTA HPLC. Fractions containing sTva47 (>99% purity as judged by electrospray mass spectrometry) were pooled and lyophilized. This lyophilized protein was brought up into the appropriate buffers and concentrations determined using the calculated extinction coefficient, $\varepsilon = 12{,}900\ M^{-1}\ cm^{-1}$.

5.2.4 Surface plasmon resonance binding studies.

Binding studies were carried out using a BIAcore (BIAcore AB) with the MBP-sTva47 fusion immobilized (the relatively much larger MBP provided a means to leave sTva47 largely free of potentially interfering surface effects). Oxidized MBP-sTva47 was immobilized following the recommended protocol (Pharmacia BioSensor). All binding experiments were performed using degassed and 0.2 μ filtered HEPES Buffered Saline (HBS; 10 mM HEPES, 150 mM NaCl) and either 2.5 mM EDTA or 5 mM $CaCl_2$, with freshly added 0.0005% P20 surfactant, as the running buffer. ALSV-A SU-immunoadhesin (SU(A)-rIgG) fusion protein was dialyzed into HBS+2.5 mM EDTA and diluted 5-fold into either the same or HBS + 5 mM $CaCl_2$ and the concentration determined by A_{280} using the calculated extinction coefficient, $\varepsilon = 99{,}920\ M^{-1}\ cm^{-1}$ (P20 surfactant was added to 0.0005% prior to use). Injections of 10 mM HCl were used to regenerate the binding surfaces before and between injections of SU(A)-rIgG.

5.2.5 NMR structural determination of sTva47.

Labeled sTva47 was obtained from expression in appropriately labeled minimal media (M9) culture. Expression, purification, and oxidation were carried out as described above. An NMR sample was then prepared with ~2 mM ^{15}N,^{13}C labeled sTva47 in 50 mM d_3-NaAcetate, 5 mM $CaCl_2$ and 10% D_2O at pH 5.5. Preliminary NMR experiments were carried out a Bruker DMX 600 MHz spectrometer, while standard 2D and 3D heteronuclear experiments for structure elucidation were acquired at 15°C on a Varian Unity 600 MHz spectrometer using sequences from the ProteinPack package available online from the Varian user library. All NMR experiments were processed using nmrPipe (Delaglio et al., 1995) and visually analyzed and manipulated with the locally developed program Sparky 3 (Tom Goddard and Don Kneller, University of California, San Francisco).

2D-^{15}N-HSQC, 3D-HNCACB, 3D-CBCA(CO)NH experiments were used to assign all amide ^{1}H, amide ^{15}N, $^{13}C\alpha$ and $^{13}C\beta$ resonances following standard techniques for backbone assignments (Clore and Gronenborn, 1991; Clore and Gronenborn, 1998). 2D-^{13}C-HSQC with the carrier frequency centered in the ^{13}C aliphatic region, 3D-C(CO)NH, 3D-HC(CO)NH and 3D-HCCH-COSY spectra were then used to assign the remaining aliphatic side chain ^{1}H and ^{13}C resonances. Aromatic side chain resonances were also assigned by using a 2D-^{13}C-HSQC spectrum centered in the ^{13}C aromatic region and by inspecting cross-peak connectivities in 2D- and 3D-NOESY experiments.

Restraints for structural calculation were extracted from three 3D-NOESY experiments: a NOESY-^{15}N-HSQC spectrum and two NOESY-^{13}C-HSQC spectra, one centered in the ^{13}C aliphatic region and the other in the ^{13}C aromatic region. All NOESY experiments were acquired using a 200 ms mixing time to allow NOE intensities to build up. The program Sparky was used to measure the volumes of NOE cross-peaks by integration. A total of 481, 880 and 128 NOE cross-peak volumes were extracted from the NOESY-^{15}N-HSQC, NOESY-^{13}C-aliphatic-HSQC and NOESY-^{13}C-aromatic-HSQC spectra, respectively. The CALIBA macro of DYANA 1.5 (Guntert et al., 1991) was then used to convert the integrated volumes to upper distance limits. Pseudoatom corrections were introduced where stereospecific assignments were not available

(Wüthrich et al., 1983). Lower bounds were also added to help define the structure and were conservatively set to 2.0 Å below the corresponding upper bounds. Furthermore, during the structural refinement process, negative restraints (lower bounds set to 4.0 Å for proton pairs that show no NOE cross-peaks in the spectra) were also introduced to eliminate unrealistic close contacts found in preliminary structures (Schmitz et al., 1998; Yao et al., 1997).

These upper and lower bounds were used in DYANA-1.5 (Guntert et al., 1997) and AMBER-6 (Cornell et al., 1995) to refine the structure of sTva47. One hundred structures were first calculated with DYANA using a simulated annealing cycle of 10,000 torsion angle dynamic steps to yield each structure. The twenty structures with the lowest target function were imported into AMBER and further refined by running several consecutive cycles of simulated annealing. The first six simulated annealing cycles were 60 ps long with 12 ps high temperature periods at 1000°K to help overcome local energy barriers. These were followed by nine consecutive 100 ps long annealing cycles with 20 ps spent at 1000°K, for a total of 15 annealing cycles and 1260 ps of AMBER molecular dynamic simulations. At the end of each cycle, the experimental distance restraints were analyzed against the calculated structures to identify and correct errors in assignments, integration and underestimated upper bounds due to spin diffusion. The last frame of the annealing cycle was then used to restart a new cycle using the newly refined set of distance bounds. Details of the final set of restraints are shown in Table 5.1 and Figure 5.2D. The final structures were obtained by averaging the coordinates saved during the last 5 ps of equilibration at 300°K of the last annealing cycle and restrained energy-minimized. The force constant used in AMBER for NOE distance restraints was 20 kcal/mole; this value was increased to 100 kcal/mole during high temperature dynamics at 1000°K. During AMBER calculations, a generalized Born implicit solvation model (Tsui and Case, 2000) was used to account for the effects of water solvation during molecular dynamic simulations, with the salt concentration set to the actual ion strength of our sample.

NOE cross-peak connectivities and preliminary structural calculations suggested a unique disulfide bonding pattern in agreement with the topology suggested by mutagenic studies (Belanger et al., 1995) and comparison to other LDLR LBrs. Hence, the three

disulfide bonds were defined for further structural calculations with DYANA and AMBER. On the contrary, calcium was not explicitly included as well as no dihedral angle restraints were used during our simulations. Moreover, the first 5 residues of sTva47 (I5-G9), that show only intraresidue and short-range NOE contacts, were poorly defined in early calculated structures and were not included in the calculations that yielded our final structures. The twenty lowest energy structures of the predominant conformer of sTva47 have been deposited in the Protein Data Bank, along with the NMR constraints and chemical shift assignments (PDB ID code 1K7B).

Table 5.1. Constraint and Structural Statistics.

No. of NOE distance restraints	
intraresidue *(i = 0)*	120
short-range *(i = 1)*	346
medium-range *(1 < i < 4)*	268
long-range *(i > 4)*	282
Total	1016
pairwise RMSD for overlays (Å) (backbone, heavy)[a]	
6-47	0.25, 0.75
7-46	0.20, 0.73
7-25	0.23, 0.83
26-46	0.05, 0.59
Ramachandran plot region[b]	
in favoured (%)	63.9
in additional allowed (%)	35.0
in generously allowed (%)	1.1
restraint satisfaction	
rms violation (Å)	0.041 ± 0.0001
maximum NOE violation	0.286
average no. of NOE violations	23

[a] MOLMOL (Koradi et al., 1996) was used to determine RMSD values for overlays.
[b] Procheck-NMR (Laskowski et al., 1996) was used to generate Ramachandran plots to assess the quality of the structures.

5.3 RESULTS AND DISCUSSION

5.3.1 sTva47 requires calcium for folding, structure and function

Purified sTva47 required oxidative folding to form the correct disulfide bonds. As was found for oxidation of LBr from the LDLR family of receptors (Bieri et al., 1998; Blacklow and Kim, 1996; Dolmer et al., 1998; North and Blacklow, 1999), calcium was required to selectively form a single disulfide bonding pattern, which we assumed matched that of the native Tva and the LDLR LBr as previously suggested (Belanger et al., 1995) (Wang et al have reported similar findings (Wang et al., 2001)). After suitable conditions were defined, oxidation was carried out using purified MBP-sTva47 in order to use correctly oxidized sTva47 attached to MBP for surface plasmon resonance binding experiments (see below).

Inspection of ^{15}N-HSQC spectra, revealed that the addition of excess (5 mM) calcium dramatically reduces the number (from >120 to 84) and increases the dispersion of cross-peaks in ^{15}N-edited HSQC spectra (data not shown). This indicates that, even when correctly oxidized, the Tva LBr requires bound calcium to stabilize its tertiary structure. Similar results were also found by Wang et at for Tva (Wang et al., 2001) and for LBr of the LDLR family of receptors (Bieri et al., 1998; Blacklow and Kim, 1996; Dolmer et al., 1998; North and Blacklow, 1999).

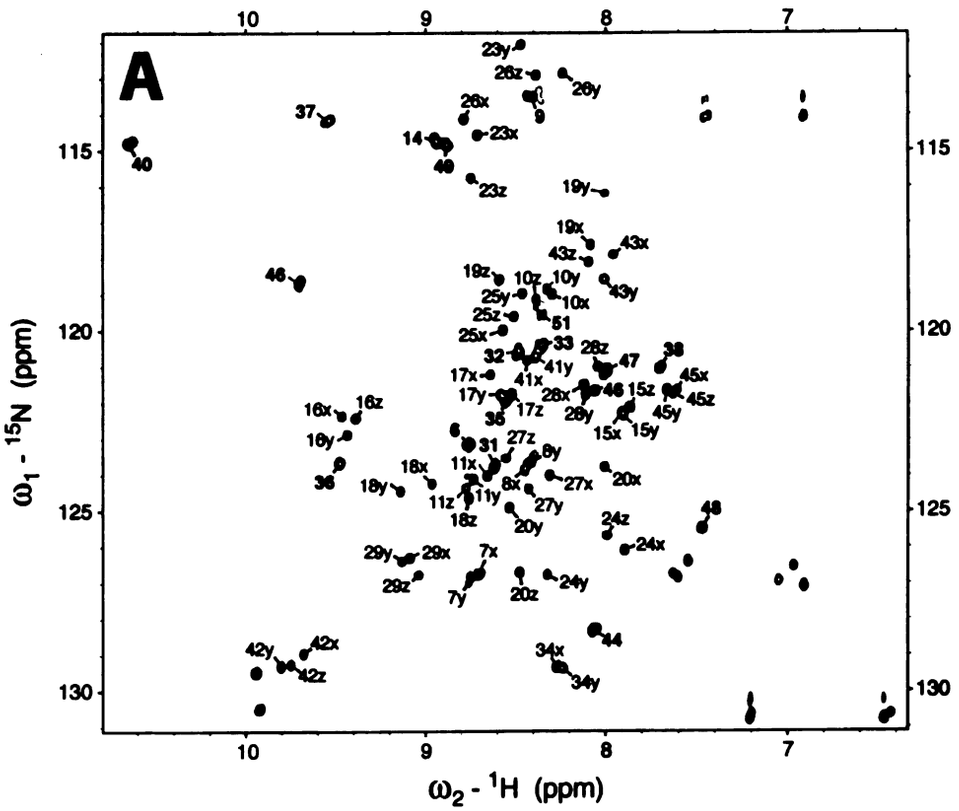
ALSV binds to its cellular receptor through the surface subunit (SU) of its Envelope glycoprotein (Env). To quantitatively address the ability of sTva47 to bind ALSV-A Env, we performed surface plasmon resonance binding experiments using a previously characterized ALSV-A SU-immunoadhesin (SU(A)-rIgG) fusion (Zingler and Young, 1996). In the absence of calcium (i.e., 2.5 mM EDTA) the binding constant was 5.5 nM. Notably, the presence of 5 mM calcium increased the affinity of this interaction 27.5-fold (K_d = 0.2 nM) (data not shown). This range of disassociation constant is similar to those previously reported, where the presence of calcium was not quantified (Balliet et al., 1999; Zingler and Young, 1996). These results demonstrate that calcium increases the affinity of sTva47 for the SU of ALSV-A, presumably by stabilizing its tertiary structure, as detailed below. This requirement for calcium for oxidative folding, structural stability, and receptor binding is consistent with the results of studies of LBr from the LDLR

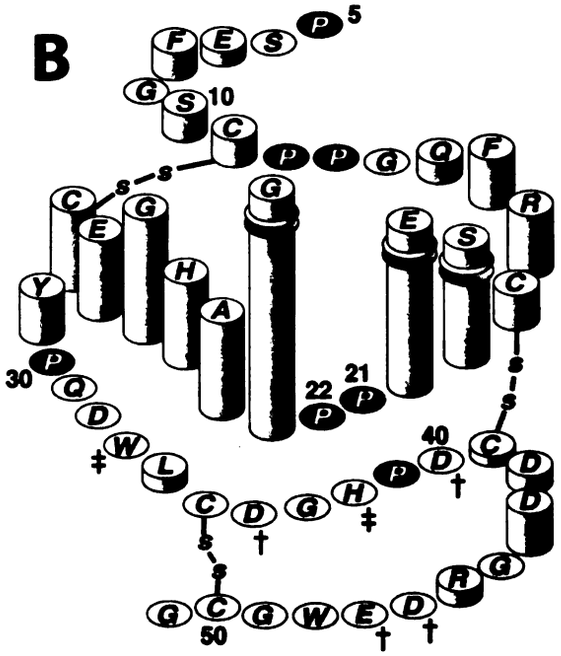
family of receptors (Atkins et al., 1998; Bieri et al., 1998; Blacklow and Kim, 1996; Dolmer et al., 1998; Fass et al., 1997; Huang et al., 1999; North and Blacklow, 1999).

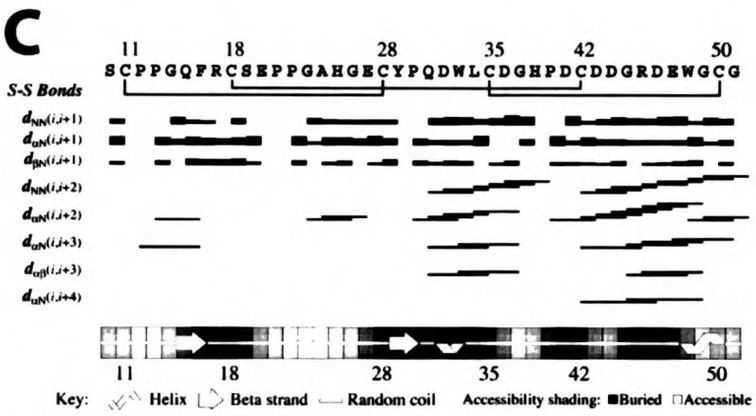
5.3.2 Structural characterization of sTva47

Surprisingly, the ^{15}N-HSQC spectrum of properly folded sTva47 acquired in the presence of excess calcium shows a larger number of cross-peaks than expected (Figure 5.2A), indicating the presence of multiple conformations which do not interconvert within the NMR time scale. Changes in the pH, protein concentration, salt concentration (including calcium concentration) and temperature (15-35°C range) of the sample did not seem to favor one set of cross-peaks over others.

Resonance assignments revealed three sets of cross-peaks for some residues indicating the presence of three different conformers (Figure 5.2A). Plotting the chemical shift difference between the three conformers for each residue (Figure 5.2B) showed that this conformational flexibility is localized to the loop region between C18 and C28 with the largest deviations found around two adjacent proline residues: P21 and P22. This observation suggested that a *cis-trans* isomerization equilibrium about the peptide bonds of P21 and P22 may be responsible for the multiple conformers. This hypothesis was demonstrated and the configuration of the two peptide bonds in the three conformers determined by analyzing Hα(i)-Hα(i+1) and Hα(i)-Hδ(i+1) NOE connectivities between residues E20-P21 and P21-P22. The conformers were found to have *trans-trans*, *trans-cis* and *cis-trans* configuration about the P21 and P22 peptide bonds, respectively, with the *trans-trans* conformer being the most abundant one, based on the relative intensities of ^{15}N-HSQC cross-peaks. In this manuscript we report the structure of the predominant *trans-trans* conformer. We note that this conformational flexibility seems to be restricted to the C18-C28 loop in the N-terminal half of Tva, far away from the putative binding site of the viral envelope glycoprotein. Moreover, the two proline residues undergoing *cis-trans* isomerization have been mutated without impairing the ability of Tva to function as a receptor for ALSV-A (Rong et al., 1998a; Rong et al., 1998b; Zingler et al., 1995). Thus, we conclude that while this prolyl isomerization process is a significant structural property of the quail Tva, it does not play an important role in viral receptor functionality.


A
10
9
8
7
115
120
125
130
ω_1 - ^{15}N (ppm)
ω_2 - ^{1}H (ppm)


B
5
10
21
22
30
40
50



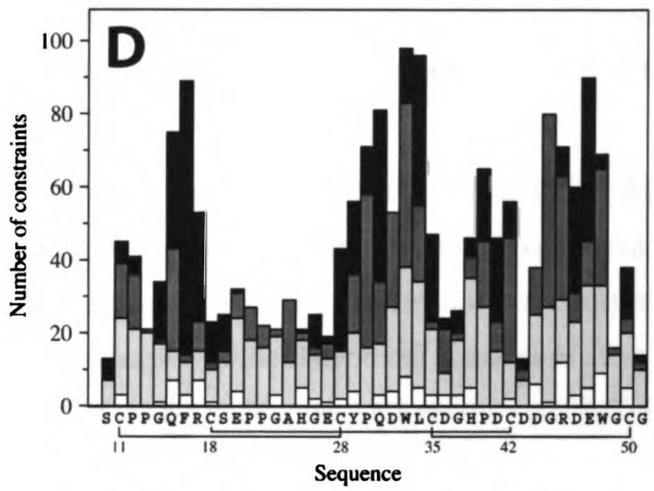


Figure 5.2. (A) ^{1}H-^{15}N HSQC spectrum of sTva47 with assignments. The three different conformers observed for some residues are labeled x, y and z, where x refers to the dominant *trans-trans* conformer whose structure was solved. Cross-peaks from side-chain groups were also assigned and are explicitly labeled in the figure. Folded peaks are shown with black labels. (B) Cartoon representation of sTva47 molecule with disulfide bonds and daggers to indicate residues that are involved in calcium coordination (‡, coordination by backbone carbonyl; †, coordination by side chain carboxyl). The height of the amino acid cylinders is proportional to the chemical shift difference between the three x, y, z conformers. (C) Summary of local constraints used in deriving the structure of sTva47 including disulfide bonds and short- to medium-range NOE-derived backbone constraints. Secondary structure and average estimated accessibility are shown at the bottom as well. (D) Plot of intraresidue, short-range (i=1), medium-range (1<i<4) and long-range (i>4) NOE distance restraints per residue.

In general, the RMSD values, energies, geometry and Ramachandran statistics are good for all the calculated structures of sTva47 (Table 5.1). The structure appears to be formed from N- and C-terminal structural domains. The N-terminal domain is folded

around the side chain of R17 (that has NOE contacts to 14 other residues) and is dominated by the C18-C28 loop with a short two strand parallel β-sheet at its bottom, also described in other LDLR LBr structures. This loop, which undergoes the isomerization process described above, is highly exposed to solvent and less well defined by NMR restraints. The C-terminal domain, on the other hand, appears to be better defined with a larger number of distance restraints, probably due to the stabilizing effect of calcium coordination (Table 5.1 and Figures 5.2 and 5.3). The two domains are linked together by the C18-C41 disulfide bond at one end and the backbone of residues Q31-L34, that form a short 3_{10}-helix (also found in other LDLR LBr), at the other end. In between these is the hydrophobic core of the protein comprised of the side chains of residues F16, Y29, P30 and W33. Another short 3_{10}-helix (residues W48 to C50) is found at the C-terminus.

The structure of sTva47 was compared to LR5 (Fass et al., 1997), the fifth repeat of LDLR, chosen to represent the LBr fold. Not surprisingly, given the identical S-S topology and the putative calcium binding site, the structures are very similar with the backbone of the C-terminal and part of the N-terminal domains of sTva47 (residues C28-C50) overlapping to 1.0 Å of the corresponding LR5 residues (C17-C39) (Figure 5.4A). Moreover, the RMSD value decreases to 0.6 Å if only the C-terminal domain is considered (L34-G49 of sTva47 overlapped to R23-N38 of LR5). The main differences between the structures are in the N-terminal domain, mainly because of the 5 amino acid longer C18-C28 loop found in sTva47 compared to LR5. This is also the region of Tva that undergoes prolyl isomerization.

The acidic side chains of D36, D40, D46 and E47, and the two backbone carbonyl oxygen atoms from residues L34 and H38, are correctly positioned by the NMR restraints to coordinate calcium. Together they form an octahedral cage, closely resembling the calcium binding site found in LR5 (Figure 5.4B). The distances between opposing oxygen atoms are somewhat longer than expected, ~7 Å compared to ~5 Å, probably because of electrostatic repulsion between their negative charges in the absence of an explicit calcium atom during structural calculations. Given this finding and the important role that calcium plays in viral receptor function described earlier, we conclude that the calcium binding site found in LDLR LBr is conserved in Tva.

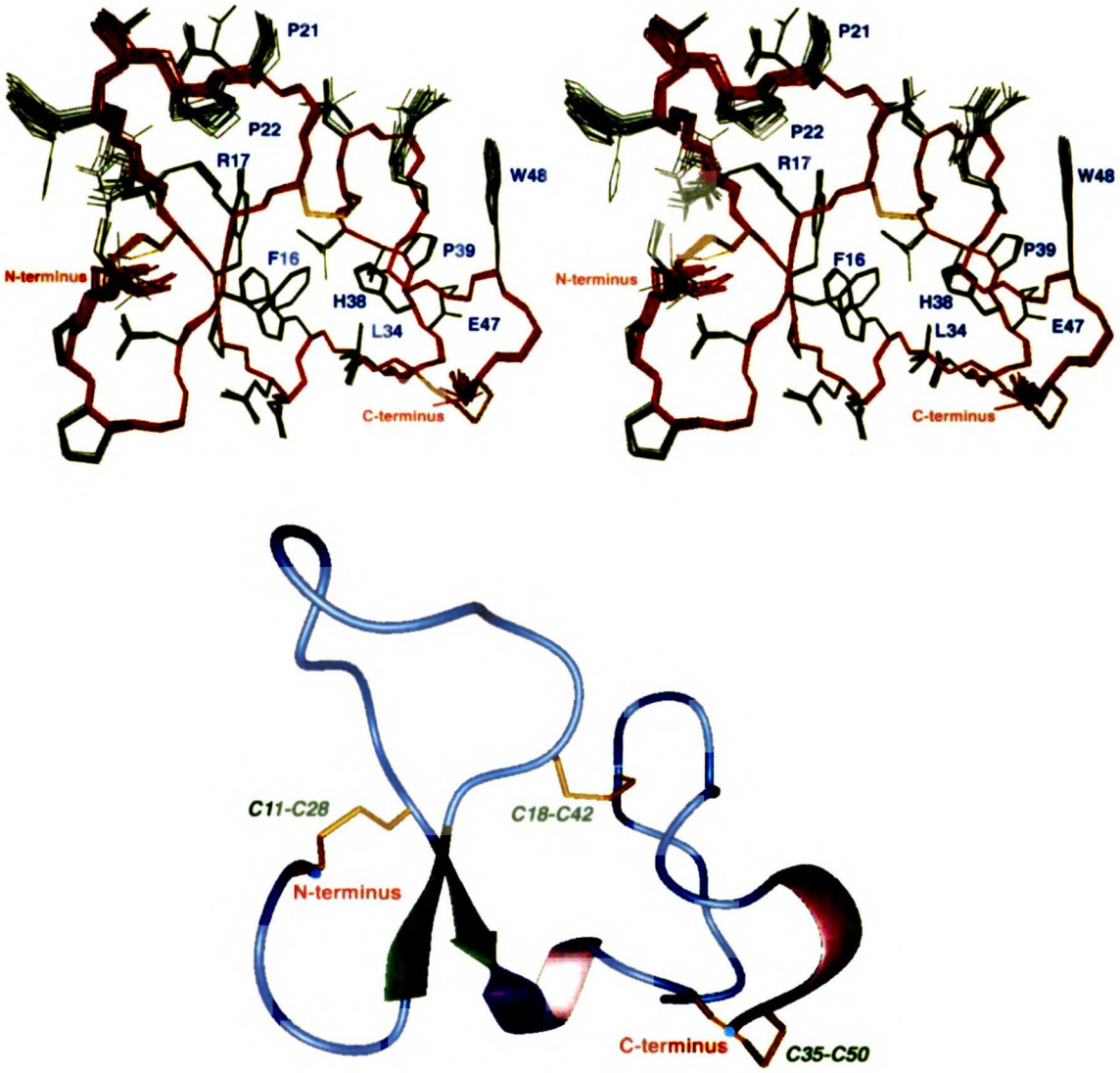


Figure 5.3. Top panel: stereo view of the 20 lowest energy structures of sTva47 superimposed. Backbone atoms (red) are illustrated along with side chains (green). Disulfide bonds are shown in yellow. Bottom panel: lowest energy structure of sTva47 with secondary structural elements. This figure was prepared with the program MOLMOL (Koradi et al., 1996).

5.3.3 Implications for viral receptor function

Gain and loss of function studies have pointed to residues that are important or critical for Tva to act as an ALSV-A receptor. Most of these residues are located in the C-terminal portion of the protein suggesting that this is where the bulk of the viral binding site is located. Among them are highly conserved residues that are critical to ensure proper folding of the LBr module of Tva: D46 and E47, whose acidic side chains

coordinate calcium (Rong et al., 1998b; Wang et al., 2001), and C35 and C50, that form a disulfide bridge (Belanger et al., 1995). Mutations of these residues most likely exert their effect by disrupting the Tva structure at the viral binding site.

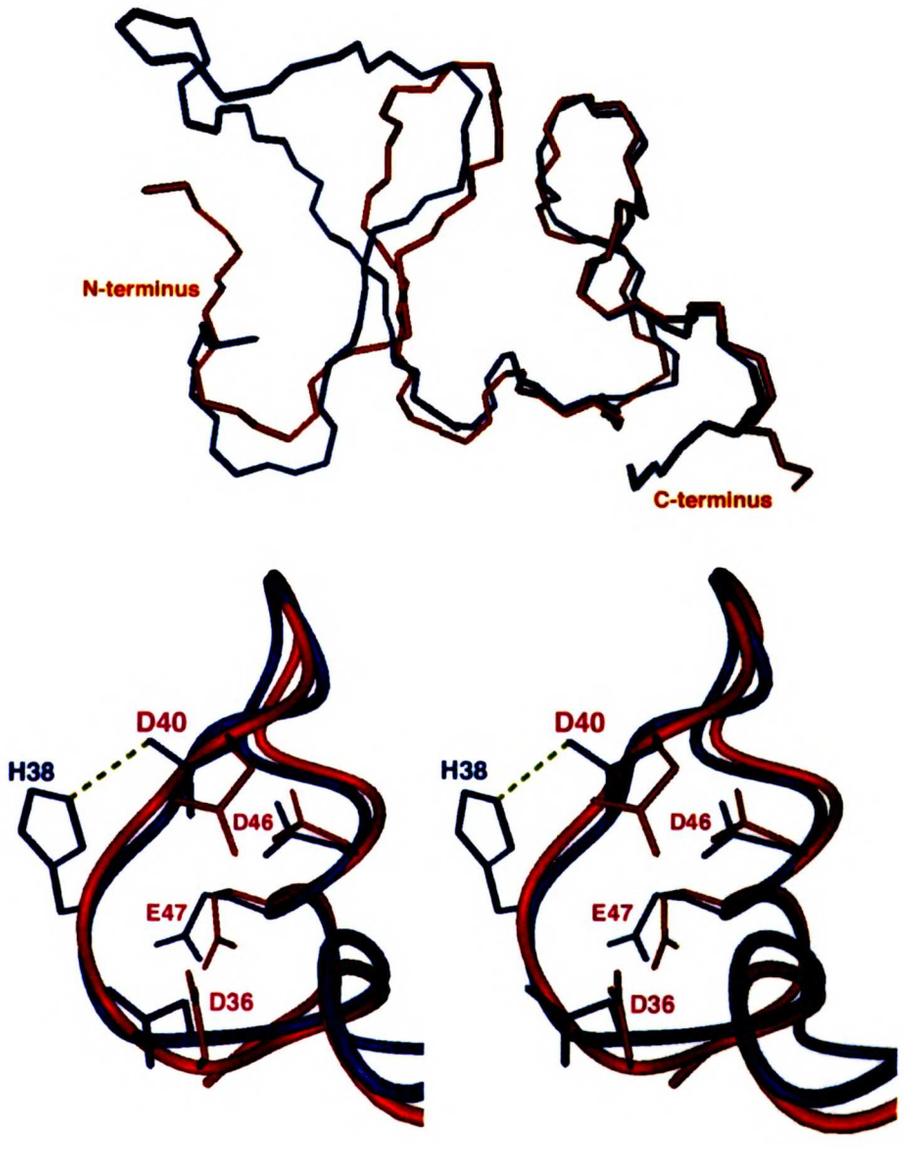


Figure 5.4. Top panel: comparison of the backbone atom coordinates of sTva47 (light blue) and LR5 (red), the fifth repeat of LDLR that was chosen as representative of the LBr domain fold (residues 24-46 of sTva47 were superimposed to residues 17-39 of LR5). Bottom panel: comparison of the calcium binding sites of sTva47 and LR5. The H-bond formed between H38 and D40 of Tva is also shown in green. The coordinates of LR5 were obtained from the Protein Data Bank (1ajj). Structural comparisons were performed using the program MOLMOL (Koradi et al., 1996).

Other important amino acids (W48, G49, L34 and H38) are not conserved within the LBr family and may play a more direct role in viral receptor function. Perhaps one of the most interesting of these residues is W48 (Hernandez et al., 1997; Rong et al., 1998b; Zingler et al., 1995; Zingler and Young, 1996). Our sTva47 structure shows the aromatic side chain of W48 at the surface of the protein, laying flat against the backbone and side chain of R45 (whose $^{1}H\alpha$ resonance is shifted upfield to 2.1 ppm due to its proximity to the aromatic ring) and to a lesser extent residues D43 and G44. Its position is well defined by NMR restraints. Given its exposed location, it is unlikely that W48 is critical for protein folding and, hence, its functional importance is most likely to derive from a direct interaction with the viral envelope glycoprotein. Furthermore, the presence of a glycine residue adjacent to W48 was also found to be important for function. While it was hypothesized that a bulkier side chain might interfere with the correct placement of the W48 aromatic side chain (Rong et al., 1998a), our structure shows that position 49 could easily host a bulkier side chain without altering the conformation of W48. An alternative explanation is that G49 forms part of the binding surface and its mutation sterically interferes with Tva contacting the envelope glycoprotein. This is confirmed by the fact that the insertion of two amino acid residues after G49 (as found in LDLR LB1 and LB4) also disrupts viral receptor function (Rong et al., 1998a), as if altering the binding surface.

L34, which some studies found to be critical for Tva receptor function (Rong et al., 1998b; Zingler et al., 1995), seems to play an important role for protein folding, since, as expected (Wang et al., 2001), it is mostly buried in the hydrophobic core of the protein. However, together with the highly conserved residue F16, L34 also forms a hydrophobic patch on the surface of Tva that is adjacent to W48 and may contact Env-A directly.

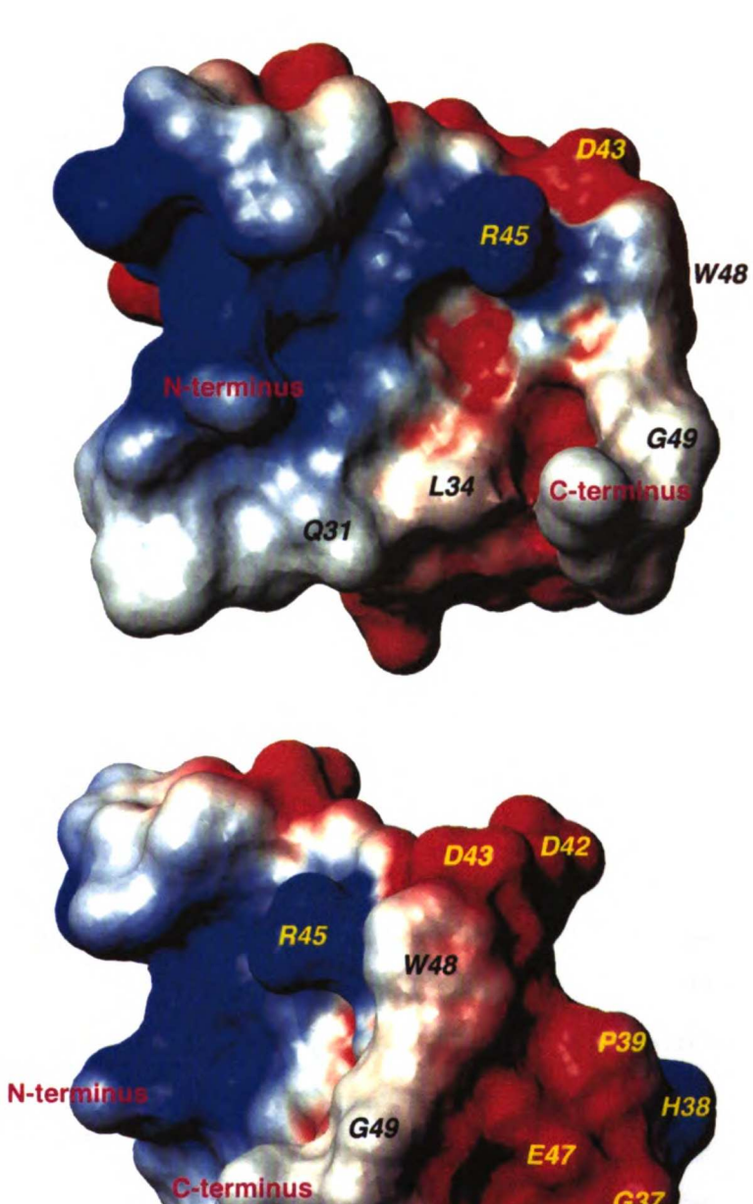
Finally, gain of function experiments using a bone fide LDLR repeat also pointed to H38 playing a role in viral receptor function (Rong et al., 1998a). This residue is not conserved among LBrs and in our structure is at the surface of Tva suggesting that it interacts directly with the envelope glycoprotein. H38, however, also seems to be important for calcium coordination, as it provides one of the two backbone carbonyl oxygen atoms that bind calcium and its aromatic side chain forms a H-bond to the acidic side chain of D40, which also coordinates the calcium ion (Figure 5.4B).

Although most of the work done so far points to the viral binding site being near the C-terminus of the Tva LBr domain, there have been indications that residues from the N-terminal domain of the molecule may also play a minor role in viral receptor function (Rong et al., 1998b). In a recent report, three mutations in the hr1 region of the SU subunit of Env-A, E149K, Y142N and Y142N/E149K, were found to reduce the binding affinity for quail but not chicken sTva (Holmen et al., 2001). Since the C-terminal half of chicken and quail Tva are identical, these results point to a positively charged residue present in the quail but not chicken N-terminal domain of Tva that contacts the hr1 region of Env-A (Figure 5.1). A possible candidate is H25, which is Gln in the chicken sequence. Our structure, however, shows that the H25 side chain is located on the opposite side of the molecule far away from the putative SU binding surface. The only other positively charged residue found in the quail but not chicken sequence is R10, surprisingly just outside the LBr domain (Gln in the chicken sequence). Unfortunately our sequence differs from both quail and chicken in that has a Ser at this position. Furthermore, in our structure this Ser is poorly defined both because its exposure at the surface and its proximity to the disordered N-terminus. However, it is quite plausible that an Arg residue at this position could contact Env-A directly. While this interaction may not contribute significantly to Tva-SU binding affinity, Env mutations like the ones described above could readily introduce steric or electrostatic clashes capable of disrupting binding of quail Tva to Env-A.

5.3.4 Analysis of Tva surface

Inspection of the molecular surface of Tva reveals a hydrophobic patch at the C-terminus, composed of residues W48 and G49, that extends towards the N-terminal domain with L34, Q31 and F16 (Figure 5.5A). The other side of the C-terminal domain is highly polar (Figures 5.5B-C) with the positively charged H38 residue surrounded by negatively charged acidic side chains. The viral binding site is most likely centered around residues W48-G49, extending on one side towards H38 and on the other towards the hydrophobic patch formed by L34 and F16. In between W48, G49 and H38 we find the acidic side chain of E47 which also coordinates calcium. Being located in such a hot spot at the Tva surface suggests that E47 may also play a more direct role in viral receptor function,

perhaps by interacting with those basic residues on the hr2 region of Env-A that mutagenic studies found to be important for Tva binding (Rong et al., 1997).



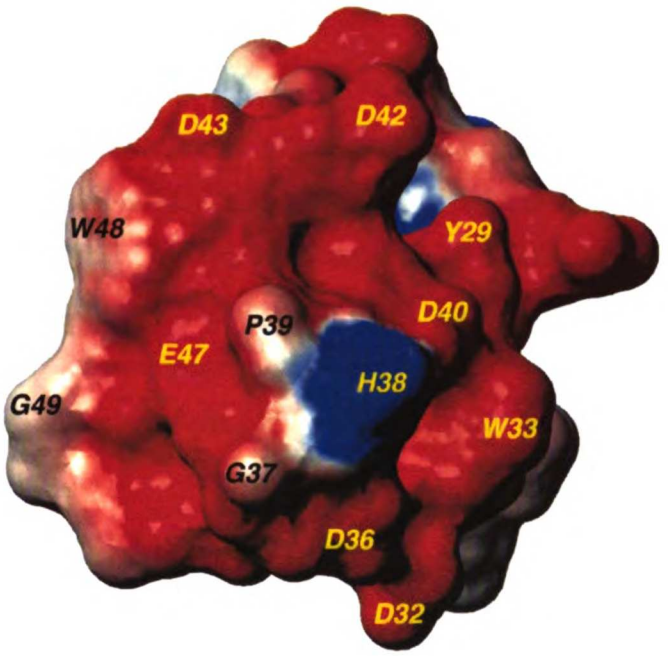


Figure 5.5. Previous page and above: molecular surface of the sTva47 structure colored by electrostatic potential. In bottom panels, the modules are rotated ~ 90° around the vertical axis. This figure was prepared with the program MOLMOL (Koradi et al., 1996).

5.4 CONCLUSIONS

We have functionally and structurally characterized sTva47, a 47 amino acid peptide that contains the LBr domain of Tva that is sufficient and necessary to interact with ALSV-A and trigger a change in conformation of Env. The three-dimensional structure of sTva47 is remarkably similar to that of other LBr domains, with identical S-S topology and a conserved calcium binding site near the C-terminus. Furthermore, all the residues that are known to be important for viral receptor function are clustered in the C-terminal half of the molecule. These findings explain our observations that calcium is required for oxidative folding, structural stability and also its ability to increase the affinity of Tva for Env-A. Finally, surface analysis points to a putative binding site in the C-terminal domain of Tva where important and critical residues (W48, G49, H38, L34 and E47) are exposed at the surface and likely to interact directly with the viral envelope glycoprotein.

5.5 ACKNOWLEDGMENTS

This work has been published on FEBS Letters 509 (2001), 161-168. We wish to thank Dr. Reuben Peters for preparing all the NMR samples and getting this project startet. We thank Dr. A. Weiss (UCSF/HHMI) for use of a BiaCore, Dr. C. Turck (UCSF/HHMI) for use of a LCQ (Finnigan) mass spectrometer, Dr. J.A.T. Young (Harvard Med) for providing SU(A)-rIgG, Dr. T.M. Handel for use of a Bruker DMX 600 MHz spectrometer for preliminary NMR analysis, Dr. V. Basus for help acquiring NMR experiments on a Varian Unity 600 MHz spectrometer and Dr. C. Kojima for helpful discussions.

5.6 REFERENCES

Atkins, A. R., Brereton, I. M., Kroon, P. A., Lee, H. T., and Smith, R. (1998). Calcium Is Essential for the Structural Integrity of the Cysteine-Rich, Ligand-Binding Repeat of the Low-Density Lipoprotein Receptor. Biochemistry *37*, 1662-1670.

Balliet, J. W., Berson, J., D'Cruz, C. M., Huang, J., Crane, J., Gilbert, J. M., and Bates, P. (1999). Production and Characterization of a Soluble, Active Form of Tva, the Subgroup A Avian Sarcoma and Leukosis Virus Receptor. J Virol *73*, 3054-3061.

Bates, P., Young, J. A. T., and Varmus, H. E. (1993). A Receptor for Subgroup A Rous Sarcoma Virus is Related to the Low Density Lipoprotein Receptor. Cell *74*, 1043-1051.

Belanger, C., Zingler, K., and Young, J. A. T. (1995). Importance of Cysteines in the LDLR-Related Domain of the Subgroup A Avian Leukosis and Sarcoma Virus Receptor for Viral Entry. J Virol *69*, 1019-1024.

Bieri, S., Atkins, A. R., Lee, H. T., Winzor, D. J., Smith, R., and Kroon, P. A. (1998). Folding, Calcium Binding, and Structural Characterization of a Concatemer of the First and Second Ligand-Binding Modules of the Low-Density Lipoprotein Receptor. Biochemistry *37*, 10994-11002.

Bieri, S., Djordjevic, J. T., Daly, N. L., Smith, R., and Kroon, P. A. (1995a). Disulfide Bridges of a Cysteine-Rich Repeat of the LDL Receptor Ligand-Binding Domain. Biochemistry *34*, 13059-13065.

Bieri, S., Djordjevic, J. T., Jamshidi, N., Smith, R., and Kroon, P. A. (1995b). Expression and disulfide-bond connectivity of the second repeat of the human LDL receptor. FEBS Lett *371*, 341-344.

Blacklow, S. C., and Kim, P. S. (1996). Protein folding and calcium binding defects arising from familial hypercholesterolemia mutations of the LDL receptor. Nature Struct Biol *3*, 758-761.

Bullough, P. A., Hughson, F. M., Skehel, J. J., and Wiley, D. C. (1994). Structure of influenza hemagglutinin at the pH of membrane fusion. Nature *371*, 37-43.

Carr, C. M., Chaudhry, C., and Kim, P. S. (1997). Influenza hemagglutinin is spring-loaded by a metastable native conformation. Proc Natl Acad Sci USA *94*, 14306-14313.

Clore, G. M., and Gronenborn, A. M. (1991). Structures of larger proteins in solution: three- and four-dimensional heteronuclear NMR spectroscopy. Science *252*, 1390-1399.

Clore, G. M., and Gronenborn, A. M. (1998). NMR structure determination of proteins and protein complexes larger than 20 kDa. Curr Opin Chem Biol *2*, 564-570.

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. (1995). A Second Generation Force Field For the Simulation of Proteins, Nucleic Acids, and Organic Molecules. J Am Chem Soc *117*, 5179-5197.

Delaglio, F., Grzesiek, S., Buister, G. W., Zhu, G., Pfeifer, J., and Bax, A. (1995). NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J Biomol NMR *6*, 277-293.

Dolmer, K., Huang, W., and Gettins, P. G. W. (1998). Characterization of the Calcium Site in Two Complement-like Domains from the Low-Density Lipoprotein Receptor-

Related Protein (LRP) and Comparison with a Repeat from the Low-Density Lipoprotein Receptor. Biochemistry *37*, 17016-17023.

Fass, D., Blacklow, S., Kim, P. S., and Berger, J. M. (1997). Molecular basis of familial hypercholesterolemia from structure of LDL receptor module. Nature *388*, 691-693.

Goldstein, J. L., and Brown, M. S. (1974). Binding and Degradation of Low Density Lipoproteins by Cultured Human Fibroblasts. J Biol Chem *249*, 5153-5162.

Guntert, P., Braun, W., and Wuthrich, K. (1991). Efficient Computation of 3-Dimensional Protein Structures in Solution From Nuclear Magnetic Resonance Data Using the Program Diana and the Supporting Programs Caliba, Habas and Glomsa. J Mol Biol *217*, 517-530.

Guntert, P., Mumenthaler, C., and Wuthrich, K. (1997). Torsion angle dynamics for NMR structure calculation with the new program DYANA. J Mol Biol *273*, 283-298.

Hernandez, L. D., Peters, R. J., Delos, S. E., Young, J. A. T., Agard, D. A., and White, J. M. (1997). Activation of a Retroviral Membrane Fusion Protein: Soluble Receptor-induced Liposome Binding of the ALSV Envelope Glycoprotein. J Cell Biol *139*, 1455-1464.

Holmen, S. L., Melder, D. C., and Federspiel, M. J. (2001). Identification of key residues in subgroup A avian leukosis virus envelope determining receptor binding affinity and infectivity of cells expressing chicken or quail Tva receptor. J Virol *75*, 726-737.

Huang, W., Dolmer, K., and Gettins, P. G. W. (1999). NMR Solution Structure of Complement-like Repeat CR8 from the Low Density Lipoprotein Receptor-related Protein. J Biol Chem *274*, 14130-14136.

Koradi, R., Billeter, M., and Wuthrich, K. (1996). Molmol - a Program For Display and Analysis of Macromolecular Structures. J Mol Graphics *14*, 51-55.

Laskowski, R. A., Rullmannn, J. A., MacArthur, M. W., Kaptein, R., and Thornton, J. M. (1996). AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol Nmr *8*, 477-486.

Mothes, W., Boerger, A. L., Narayan, S., Cunningham, J. M., and Young, J. A. T. (2000). Retroviral Entry Mediated by Receptor Priming and Low pH Triggering of an Envelope Glycoprotein. Cell *103*, 679-689.

North, C. L., and Blacklow, S. C. (1999). Structural Independence of Ligand-Binding Modules Five and Six of the LDL Receptor. Biochemistry *38*, 3926-3935.

North, C. L., and Blacklow, S. C. (2000). Solution Structure of the sixth LDL-A module of the LDL receptor. Biochemistry *39*, 2564-2571.

Rong, L., and Bates, P. (1995). Analysis of the Subgroup A Avian Sarcoma and Leukosis Virus Receptor: the 40-Residue, Cysteine-Rich, Low-Density Lipoprotein Receptor Repeat Motif of Tva is Sufficient To Mediate Viral Entry. J Virol *69*, 4847-4853.

Rong, L., Edinger, A., and Bates, P. (1997). Role of basic residues in the subgroup-determining region of the subgroup A avian sarcoma and leukosis virus envelope in receptor binding and infection. J Virol *71*, 3458-3465.

Rong, L., Gendron, K., and Bates, P. (1998a). Conversion of a human low-density lipoprotein receptor ligand-binding repeat to a virus receptor: Identification of residues important for ligand specificity. Proc Natl Acad Sci USA *95*, 8467-8472.

Rong, L., Gendron, K., Strohl, B., Shenoy, R., Wool-Lewis, R. J., and Bates, P. (1998b). Characterization of Determinants for Envelope Binding and Infection in Tva, the Subgroup A Avian Sarcoma and Leukosis Virus Receptor. J Virol *72*, 4552-4559.

Schmitz, U., Donati, A., James, T. L., Ulyanov, N. B., and Yao, L. (1998). Small structural ensembles for a 17-nucleotide mimic of the tRNA T Psi C-Loop via fitting dipolar relaxation rates with the quadratic programming algorithm. Biopolymers *46*, 329-342.

Tsui, V., and Case, D. A. (2000). Molecular dynamics simulations of nucleic acids with a generalized born solvation model. J Am Chem Soc *122*, 2489-2498.

Vogt, P. K. (1977). Genetics of RNA tumor viruses. In Comphrensive Virology, H. Fraenkel-Conrat, and R. Wagner, eds. (New York, Plenum), pp. 341-455.

Wang, Q.-Y., Dolmer, K., Huang, W., Gettins, P. G. W., and Rong, L. (2001). Role of Calcium in Protein Folding and Function of Tva, the Receptor of Subgroup A Avian Sarcoma and Leukosis Virus. J Virol *75*, 2051-2058.

White, J. M. (1992). Membrane Fusion. Science *258*, 917-924.

Wüthrich, K., Billeter, M., and Braun, W. (1983). Pseudo-structures for the 20 common amino acids for use in studies of protein conformations by measurements of intramolecular proton-proton distance constraints with nuclear magnetic resonance. J Mol Biol *169*, 949-961.

Yao, L. J., James, T. L., Kealey, J. T., Santi, D. V., and Schmitz, U. (1997). The dynamic NMR structure of the T Psi C-loop: Implications for the specificity of tRNA methylation. Journal of Biomolecular Nmr *9*, 229-244.

Young, J. A. T., Bates, P., and Varmus, H. E. (1993). Isolation of a Chicken Gene That Confers Susceptibility to Infection by Subgroup A Avian Leukosis and Sarcoma Viruses. J Virol *67*, 1811-1816.

Zingler, K., Belanger, C., Peters, R. J., Agard, D. A., and Young, J. A. T. (1995). Identification and Characterization of the Viral Interaction Determinant of the Subgroup A Avian Leukosis Virus Receptor. J Virol *69*, 4261-4266.

Zingler, K., and Young, J. A. T. (1996). Residue Trp-48 of Tva is Critical for Viral Entry but Not for High-Affinity Binding to the SU Glycoprotein of Subgroup A Avian Leukosis and Sarcoma Viruses. J Virol *70*, 7510-7516.

APPENDIX A

A. Preliminary Structures for the *trans-cis* and *cis-trans* conformers of sTva47

As already described earlier in chapter 5, sTva47 exists in solution as a mixture of three slow-exchanging conformers that arise from *cis-trans* isomerization about the peptide bonds of residues P21 and P22. Analysis of NOE cross-peak intensities suggested that the three conformers have configuration: *trans-trans*, *trans-cis* and *cis-trans* of the P21, P22 peptide bonds. As shown in figure 5.2B, this dynamic process mainly involves the loop region comprised between C18 and C28 that exhibits three different sets of resonances, one for each conformer. On the other hand, resonances from other parts of the molecule are almost completely overlapped. From these observations we predicted the structure for the three conformers to be very similar, except for the C18-C28 loop region that assumes different geometry depending on the configuration of the peptide bonds P21 and P22. Furthermore, from the relative intensity of ^{15}N-HSQC peaks we were able to estimate the relative population to be 2:1:1 for the *trans-trans*, *trans-cis* and *cis-trans* conformers, respectively.

In chapter 5, we described the structure of the *trans-trans* conformer that we refined using interproton distances extracted from 3D NOESY spectra to restrain the molecule during MD simulations. Initially, we ran several cycles of refinement with the program DYANA (Guntert et al., 1997) that uses torsion angle dynamics to rapidly generate structures that comply with the experimental restraints. At each cycle, these structures were analyzed in order to improve the quality and quantity of the NOE distance restraints. Once we reached a good convergence among the pool of DYANA-generated structures (low RMSD values), we used the program AMBER (Case et al., 2000) for a

---

Abbreviations. Tva, receptor for ALSV-A; aa, amino acid; sTva, soluble Tva ectodomain; sTva47, 47aa soluble Tva LBr; Env, viral envelope fusion glycoprotein; NMR, nuclear magnetic resonance; 2D, two-dimensional; 3D, three-dimensional; HSQC, heteronuclear single quantum coherence spectra; NOE, nuclear Overhauser enhancement; NOESY, NOE spectroscopy; rmsd, root mean square deviation.

more sophisticated refinement of the selected molecules that takes into account of electrostatic, as well as, steric interactions.

We are currently refining the structure for the other two conformers of sTva47, *trans-cis* and *cis-trans,* using a procedure similar to the one used for solving the structure of the *trans-trans* conformer. Given that the population for the *trans-cis* and *cis-trans* conformers is only half that of the *trans-trans* conformer (see above), it is not surprising that, for these molecules, we were able to extract a fewer number of NOE distance restraints (see Table A.1). At the present time, we have, for the *trans-cis* conformer, a converged DYANA-generated pool of structures that needs to be further optimized with AMBER (see Table A.2 for RMSD values). For the *cis-trans* conformer, on the other hand, we need to run more cycles of DYANA refinement in order to improve its distance restraint set (in particular to improve the number of interproton distance restraints for the flexible loop region where we only have few of them; see Figures A.1 and A.2).

Table A.1. NOE-derived distance restraints for the three conformers of sTva47.

# NOE distance bounds	*trans-trans*	*trans-cis*	*cis-trans*
intraresidue *(i = 0)*	120	111	101
short-range *(i = 1)*	346	319	303
medium-range *(1 < i < 4)*	268	257	256
long-range *(i > 4)*	282	256	245
Total	1016	943	905

In figure A.1, we show the DYANA-generated pools of structures for the *trans-cis* (Figure A.1a) and *cis-trans* (Figure A.1b) conformers of sTva47 at the current state of refinement. It can be noticed that the C18-C28 loop (with the flexible residued P21 and P22) is the least well defined portion of the molecule, while the C-terminal domain is extremely well defined (see also RMSD valued in Table A.2). This observation does not come as a surprise, since these are also the regions of the molecule with the lowest and highest number distance restraints, respectively (see Figure A.2 for distance restraint distribution for the three conformers). The low number of NOE-derived interproton distances for the flexible loop is probably a consequence of the fact that, as already shown by the structure of the *trans-trans* conformer, this loop is mostly exposed to the

bulk solvent and its residues only make few short-range contacts. On the other hand, the C-terminal domain contains a specific calcium binding coordination site that makes it very stable, with many long-range contacts.

Figure A.3 shows a superimposition of the current structures for the three conformers of sTva47. As expected, the C18-C28 loop, because of the different configuration of the P21 and P22 peptide bonds, is so different that it makes no sense to calculate an RMDS value among them(Figure A.3b). The rest of the Tva molecule overlaps fairly well (Figure A.3a), with RMSD values of 0.58 Å for the backbone and 0.81 Å for all heavy atoms of residues 11-18 and 28-51. Finally, the C-terminal domains (residues 31-47) are virtually identical (RMSD values of 0.23 and 0.63 for the backbone and all heavy atoms, respectively).

Table A.2. RMSD values calculated among the pools of refined structures for the three conformers of sTva47. RMSD values were calculated using the program MOLMOL (Koradi et al., 1996).

resdidues	*trans-trans*		*trans-cis*		*cis-trans*	
11-46	0.16	0.51	0.49	0.91	0.63	0.95
27-46	0.03	0.42	0.17	0.78	0.09	0.55
18-28	0.12	0.60	0.48	0.93	0.98	1.57

In conclusion, we are solving the structure for the other two conformers of sTva47 in solution: *trans-cis* and *cis-trans*. Preliminary results show that, as expected, the conformers differ only for the geometry of the C18-C28 loop that undergoes isomerization at the P21 and P22 peptide bonds. In particular, the C-terminal domain, that contains the specific calcium coordination site and where the putative binding site for the viral envelope glycoprotein is located (see chapter 5), is virtually identical. Hence, we conclude that while this prolyl isomerization process is a significant structural property of Tva, it does not play an important role in viral receptor functionality. This is also supported by the fact that P21 and P22 of Tva can be mutated to other amino acids without loosing viral receptor function (Rong et al., 1998a; Rong et al., 1998b; Zingler et al., 1995).

A.1 REFERENCES

Case, D. A., Pearlman, D. A., Caldwell, J. W., Cheatham, I. T. E., Ross, W. S., Simmerling, C., Darden, T., Merz, K. M., Stanton, R. V., Chen, A., *et al.* (2000). Amber 6.0, University of California, San Francisco. Amber 60, University of California, San Francisco.

Guntert, P., Mumenthaler, C., and Wuthrich, K. (1997). Torsion angle dynamics for NMR structure calculation with the new program DYANA. J Mol Biol *273*, 283-298.

Koradi, R., Billeter, M., and Wuthrich, K. (1996). Molmol - a Program For Display and Analysis of Macromolecular Structures. J Mol Graphics *14*, 51-55.

Rong, L., Gendron, K., and Bates, P. (1998a). Conversion of a human low-density lipoprotein receptor ligand-binding repeat to a virus receptor: Identification of residues important for ligand specificity. Proc Natl Acad Sci USA *95*, 8467-8472.

Rong, L., Gendron, K., Strohl, B., Shenoy, R., Wool-Lewis, R. J., and Bates, P. (1998b). Characterization of Determinants for Envelope Binding and Infection in Tva, the Subgroup A Avian Sarcoma and Leukosis Virus Receptor. J Virol *72*, 4552-4559.

Zingler, K., Belanger, C., Peters, R. J., Agard, D. A., and Young, J. A. T. (1995). Identification and Characterization of the Viral Interaction Determinant of the Subgroup A Avian Leukosis Virus Receptor. J Virol *69*, 4261-4266.

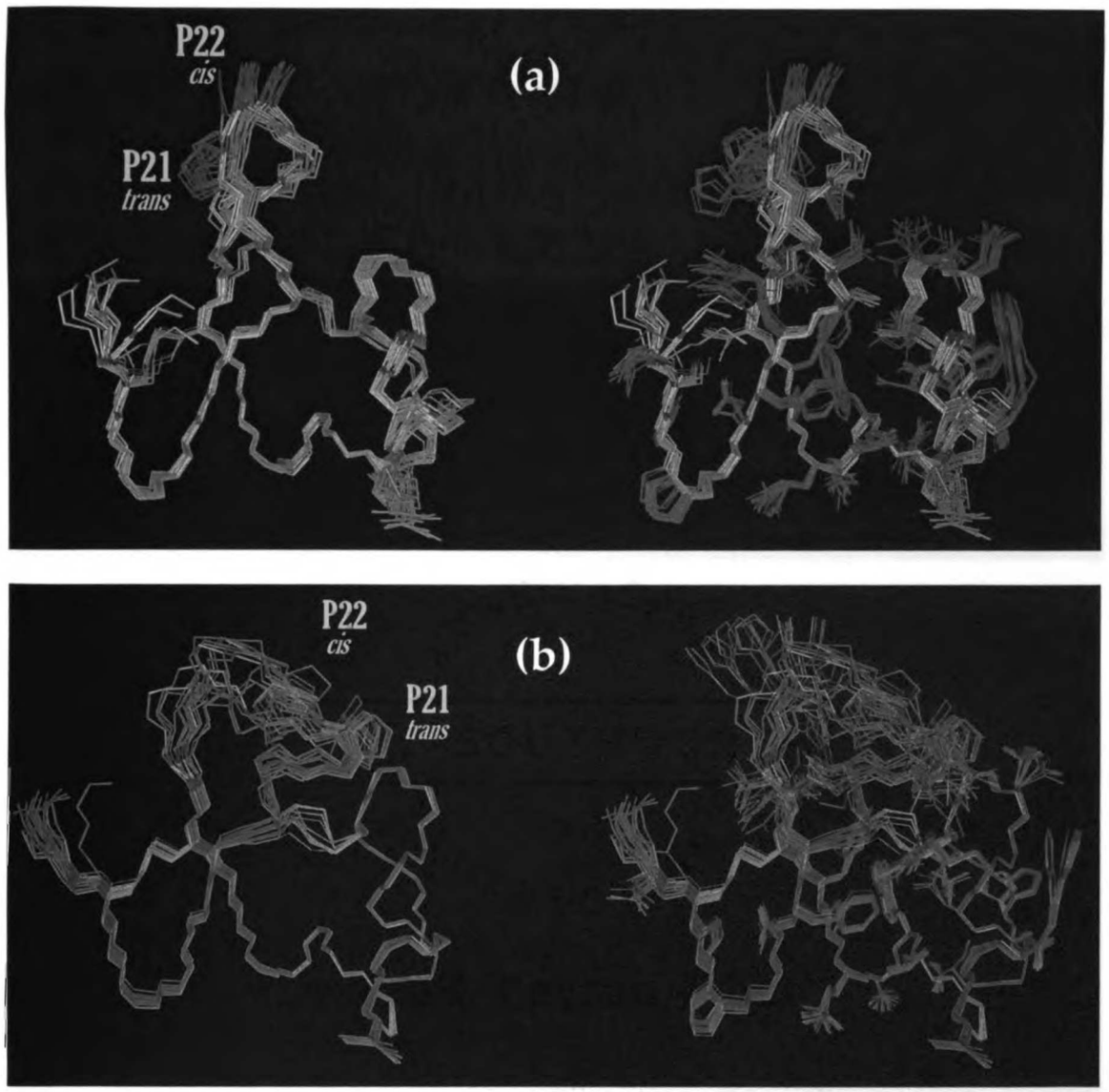


Figure A.1. View of the 20 lowest energy structures generated by DYANA for the *trans-cis* (a) and *cis-trans* (b) conformers.

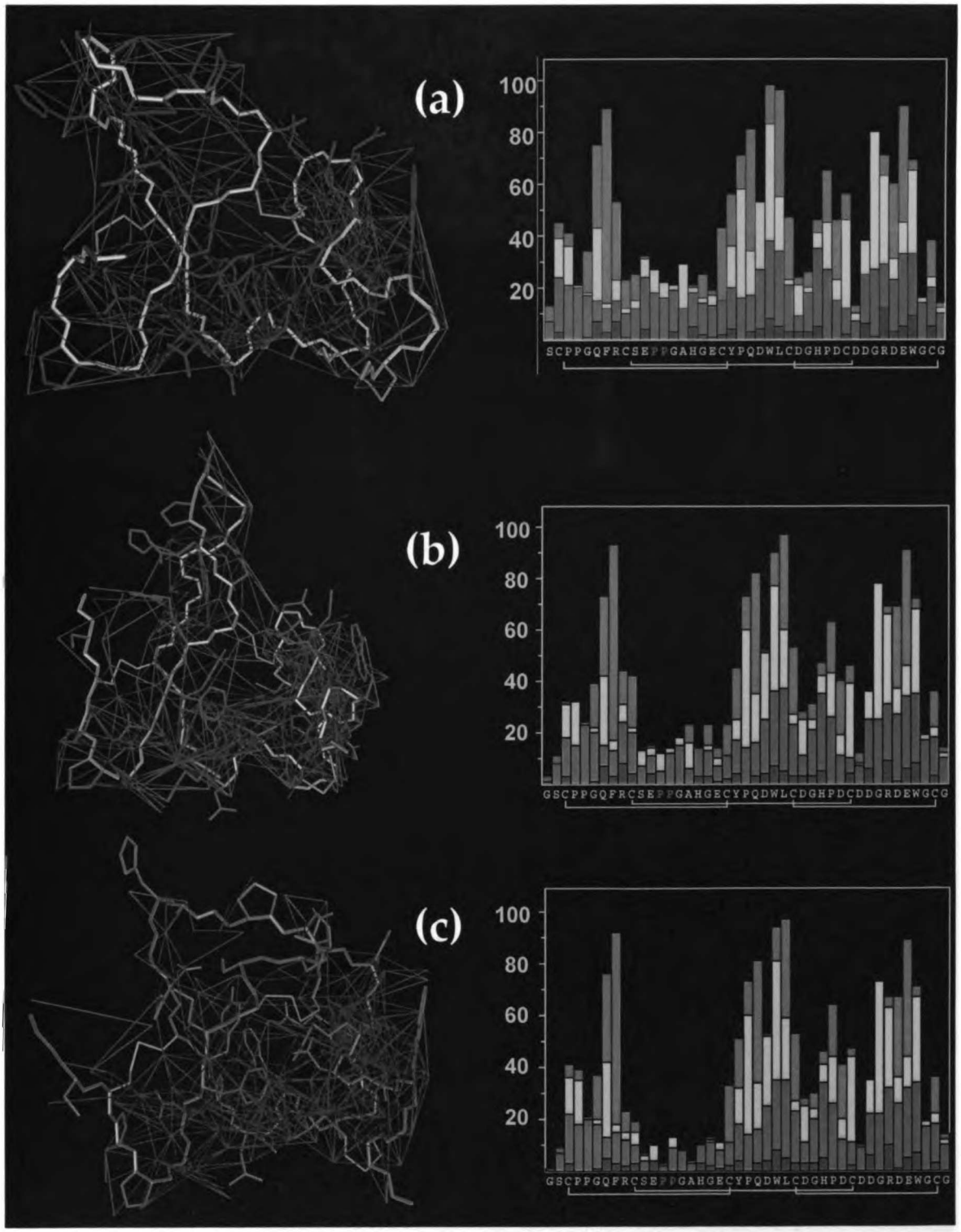


Figure A.2. Distribution of NOE distance restraints for the three conformers of sTva47: (a) *trans-trans*, (b) *trans-cis* and (c) *cis-trans*. Right panels, the distance restraints are shown as brown lines directly on the best structure of the refined pool for each conformer. Left panels, distribution of distance restraints along the sTva47 sequence: intraresidue (green), short-range (cyan), medium-range (yellow) and long-range (magenta) distances are shown.

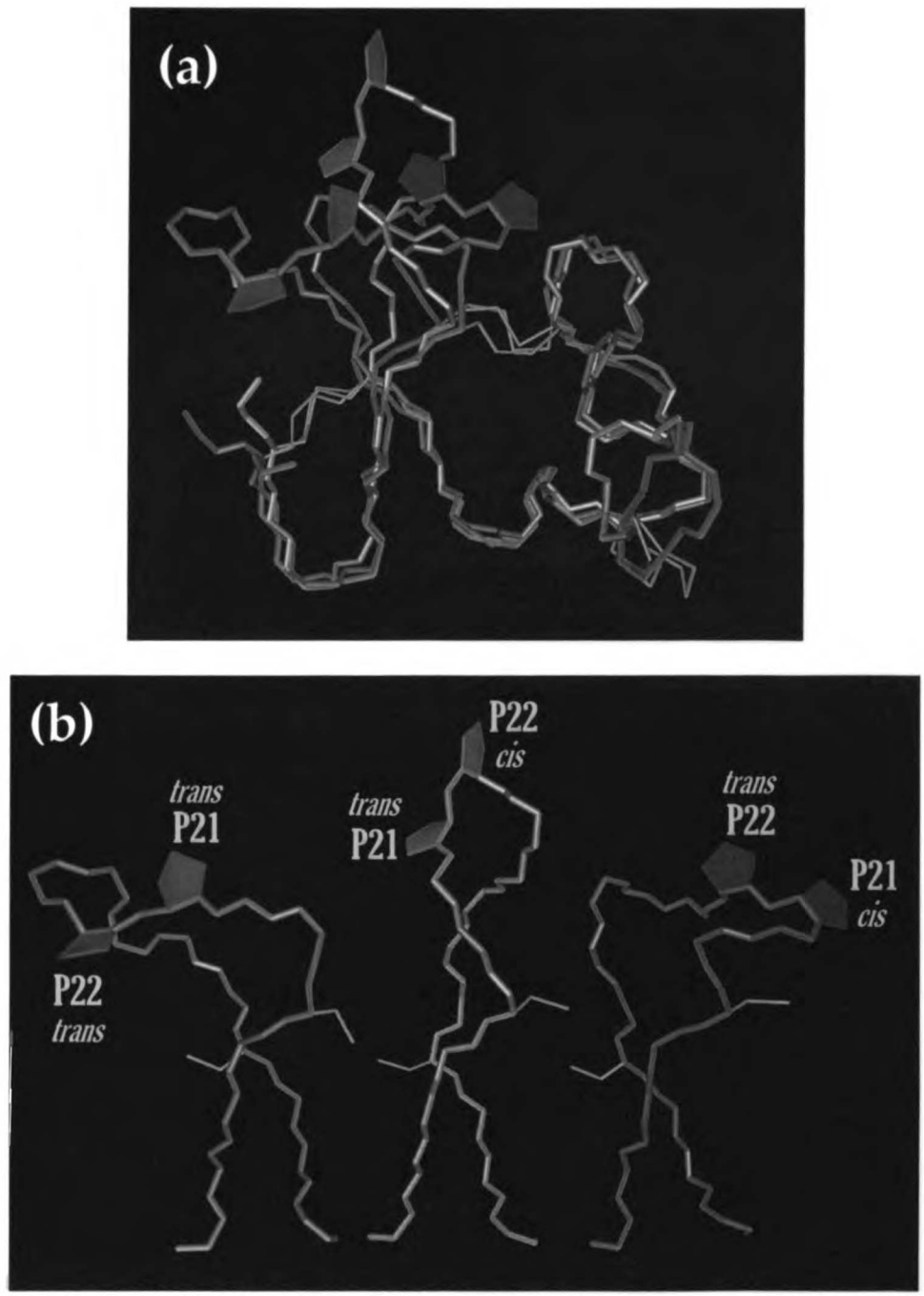
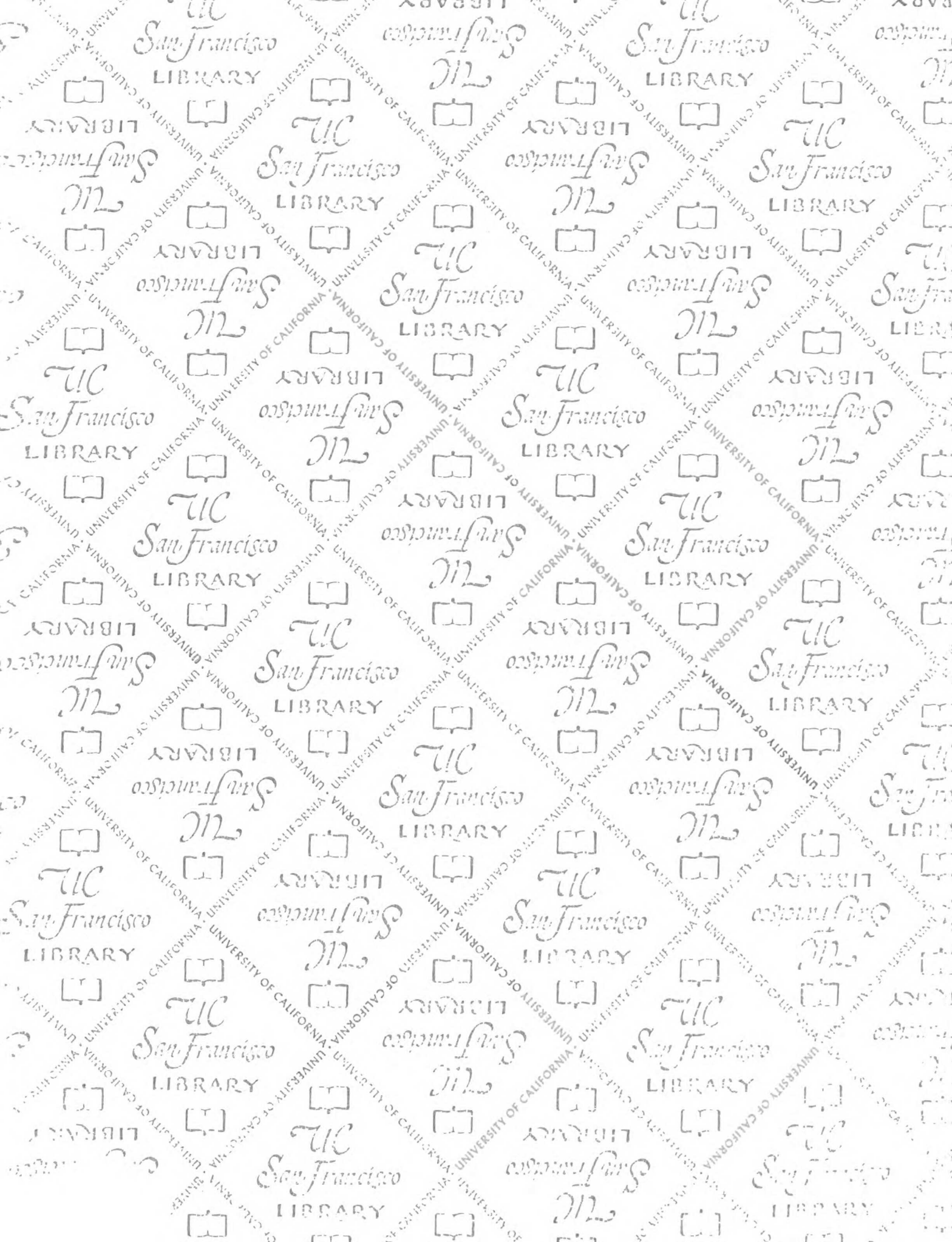


Figure A.3. Comparison of *trans-trans* (blue), *trans-cis* (pink) and *cis-trans* (green) structures of sTva47. In (a) the three conformers are overlapped using backbone and all heavy atoms of residues 11-18 and 28-51 (see text for RMSD values). In (b) the C18-C28 loop for the three conformers is shown. The P21 and P22 residues that undergo *cis-trans* isomerization are labeled.


UC
San Francisco
LIBRARY
UNIVERSITY OF CALIFORNIA