Atomistic Modeling of DNA and Protein Structures

George C. Schatz

Introduction

In this lecture we introduce one of the most commonly used computational tools for studying protein and DNA structure, the use of empirical potential energy functions, and we describe their use in molecular mechanics and molecular dynamics calculations. As an application, we present results of a molecular mechanics study of the mechanical properties (mechanical pulling) of a small protein. We also describe modeling of DNA hairpin structures.

References:

Computer Simulation of Liquids, M. P. Allen and D. J. Tildesley, Clarendon Press, Oxford, 1987;

Introduction to Modern Statistical Mechanics, D. Chandler, Oxford, New York, 1987;

Understanding Molecular Simulations: from Algorithms to Applications, D. Frenkel and B. Smit, Academic Press; San Diego, 1996.

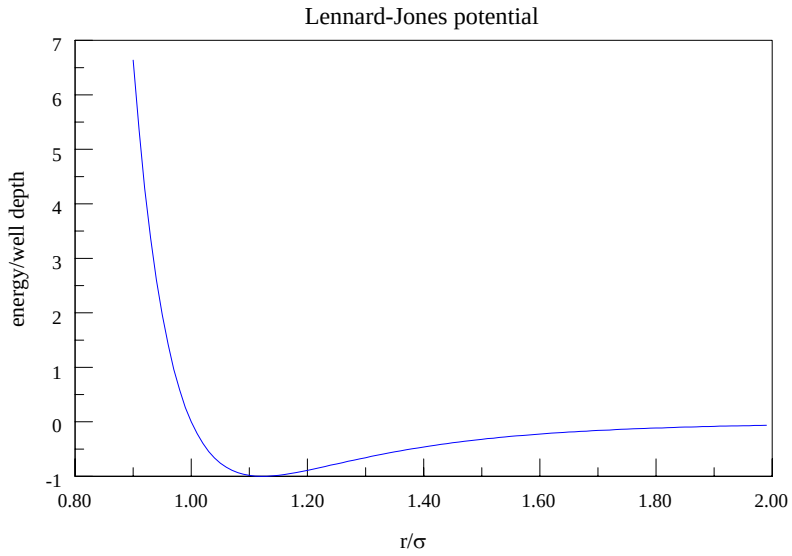
Molecular Modeling: Principles and Applications, (2nd ed) Andrew Leach, Prentice Hall, Englewood Cliffs, 2001.

What is a force field? This is the potential energy function V(R) that determines the interactions between the atoms in a molecule or solid.

The derivatives of this function (or more technically the gradients) give the forces that these atoms exert against each other. Thus Newton's equations say that F=ma, where the force F is related to V via F=-dV/dR, m is the mass and a is the acceleration.

A simple example of a force field is a Lennard-Jones, or 6-12 potential. This describes the interaction of two argon atoms pretty well, and it is often used to describe the interaction of two methane molecules or even, to a lesser degree of rigor, two water molecules.

$$V = 4\varepsilon\left(\left[\frac{\sigma}{R}\right]^{12} - \left[\frac{\sigma}{R}\right]^{6}\right)$$



For a liquid composed of argon, or methane, one can approximate that the overall potential is simply the sum of all pair-wise interactions:

$$V = \sum_{ij} 4\varepsilon_{ij} \left(\left[\frac{\sigma_{ij}}{R_{ij}} \right]^{12} - \left[\frac{\sigma_{ij}}{R_{ij}} \right]^{6} \right)$$

This neglects three-body and higher interactions, which works pretty well for argon, but is a problem for more complex materials.

Where do force fields come from?

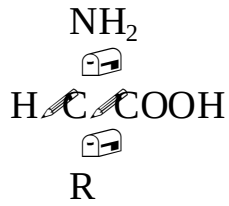
The correct force fields can be determined by **electronic structure calculations**, where one calculates the energies of the molecule or material as a function of the positions of the nuclei by solving the Schrödinger equation for the molecular orbitals. We will learn about this process in a later lecture, but suffice it for now to realize that this is a lot of work, and often it is not feasible to do.

As an alternative, one can guess a force field, such as the 6-12 potential that we just looked at. This has a little science in it, as the correct interatomic potential at long range looks like $-1/R^6$, however the $1/R^{12}$ part that describes the short range repulsion is just made up. The correct potential looks more like exp(-R), but it is numerically simpler to use a power law, and $1/R^{12}$ is mathematically convenient.

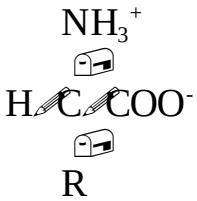
The 6-12 potential has two parameters, the well depth and the equilibrium distance. These are found by fitting results to experiment. You might think that for modeling liquid argon, ε and σ should be obtained by fitting the exact potential for the Ar dimer, but in fact it is better to fit properties of the liquid (such as density, boiling temperature, etc) if you want to obtain realistic estimates of the property of the liquid.

What is a protein?

Proteins are large molecules that are composed of peptide chains and sometimes other components (heme groups, sugars, nucleotides). The peptide is a polymer of amino acids. Amino acids have the general formula:



where R is one of 20 possible organic side chains that occur in biological systems. Actually in solution at neutral pH, all amino acids are substantially ionized into zwitterions:



however we'll ignore this here.

For the naturally occuring amino acids, there are 20 possibilities for side chains (the R group). Examples are (we'll use these later):

Name	abbr.	abbr.	formula
Alanine	Ala	A	CH_3
Aspartic acid	Asp	D	CH_2COOH
Arginine	Arg	R	$(CH_2)_3NHC(NH_2)_2$
Proline	Pro	P	$CH_2CH_2CH_2$-
Phenylalanine	Phe	F	CH_2-phenyl

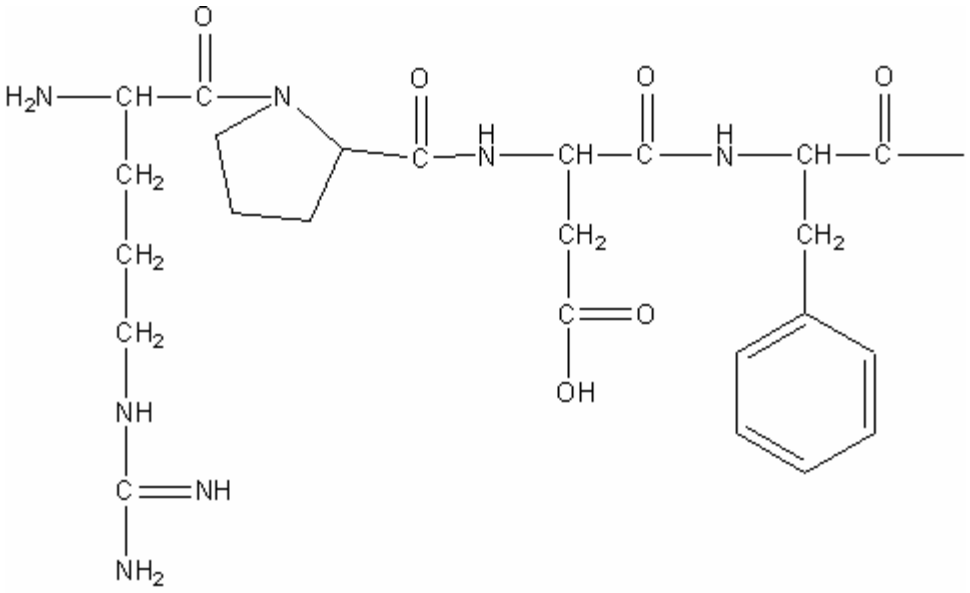
Peptides are formed by linking many amino acids together to form a polymer:

$$NH_3^+ - \underset{H}{\overset{R_1}{C}} - CO - N - \underset{H}{\overset{R_2}{C}} - CO - N - \underset{H}{\overset{R_3}{C}} - COO^-$$

A simple protein: BPTI (bovine pancreatic trypsin inhibitor)

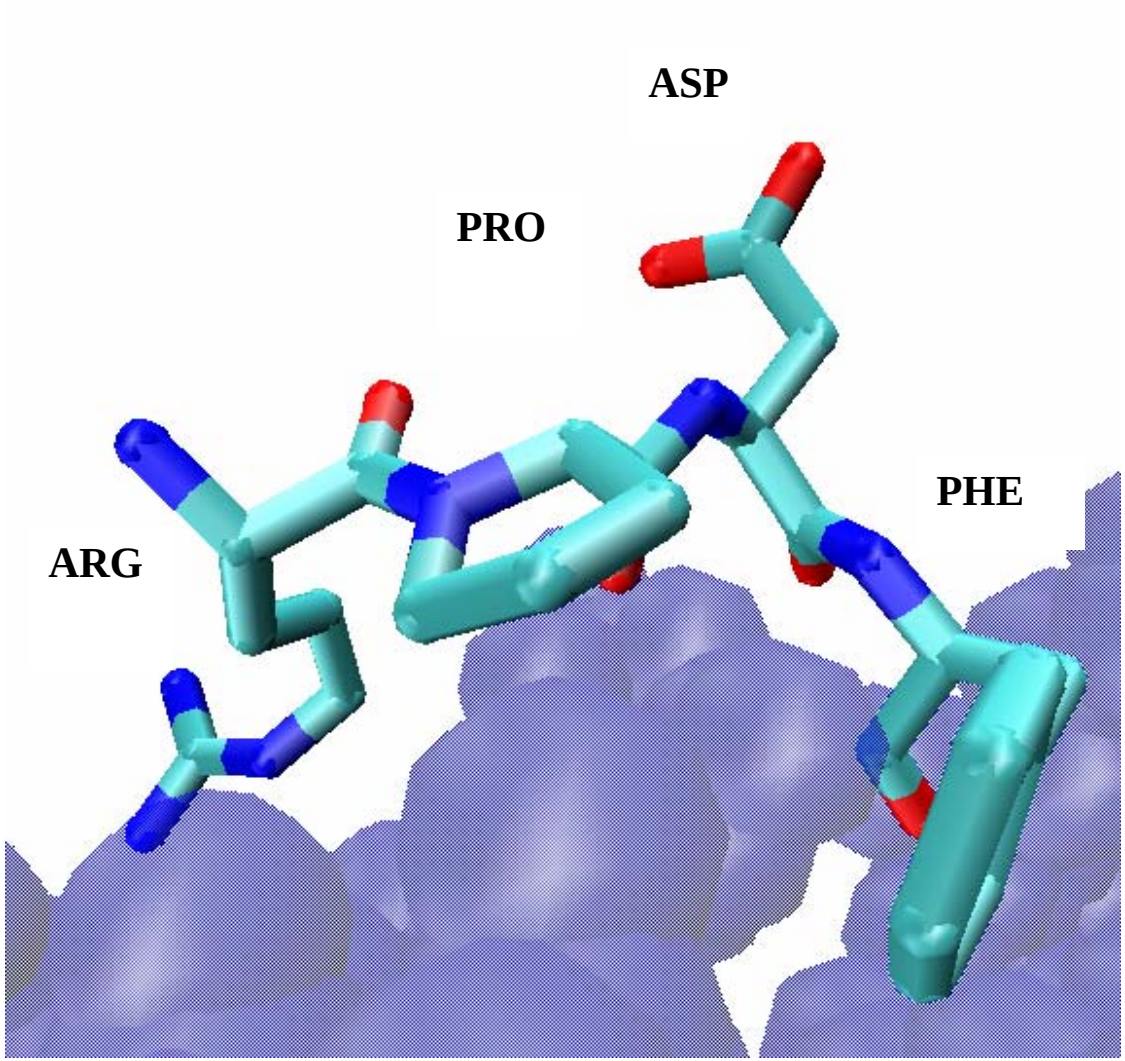
This consists of a single peptide having 58 amino acids. These are:

ARG PRO ASP PHE CYS LEU GLU PRO PRO TYR THR GLY PRO CYS LYS ALA ARG ILE ILE ARG TYR PHE TYR ASN ALA LYS ALA GLY LEU CYS GLN THR PHE VAL TYR GLY GLY CYS ARG ALA LYS ARG ASN ASN PHE LYS SER ALA GLU ASP CYS MET ARG THR CYS GLY GLY ALA

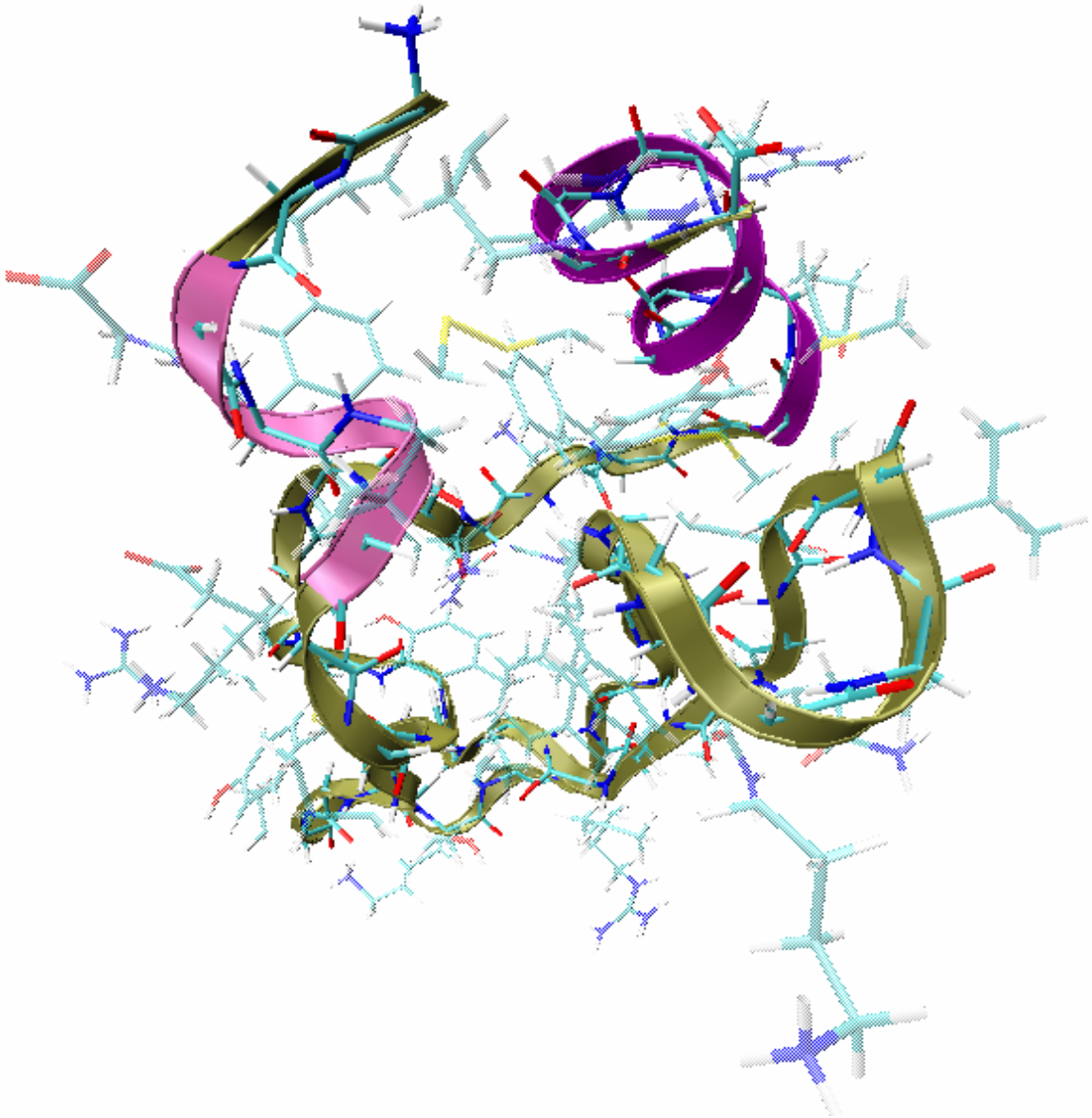


The structure of the first four residues (leaving out hydrogens) is:

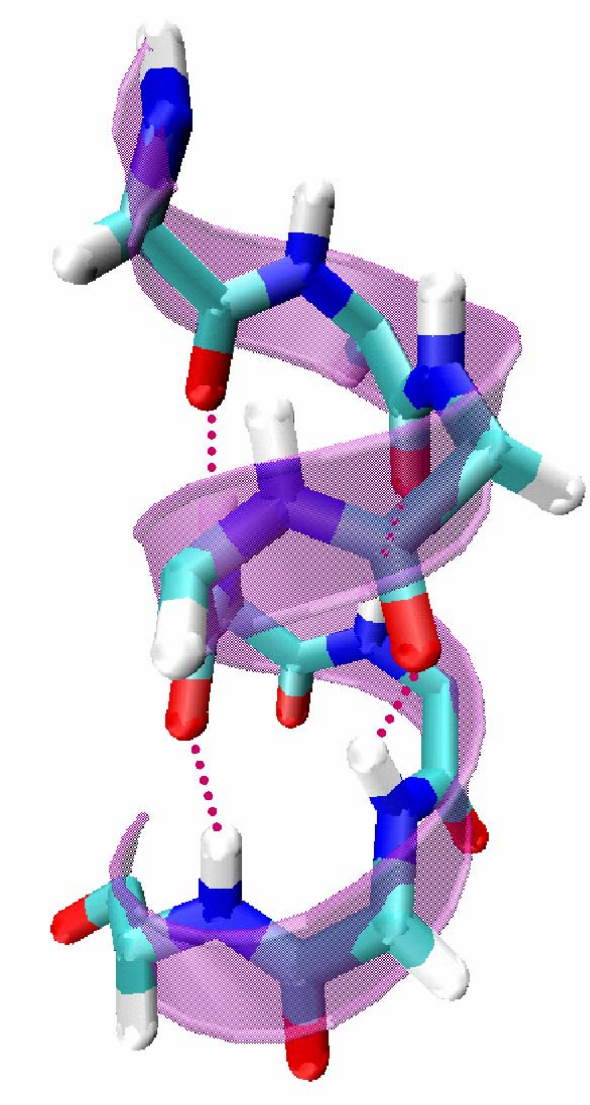
Color map: red – oxygen, blue – nitrogen, cyan – carbon.



The BPTI structure: Ribbon represents the backbone of protein.



Note that this structure consists of a jumble of several α-helices, where the general structure of each helix is:



How does one describe force fields for proteins?

There is no unique procedure, but the general idea is to write the potential as a sum of terms as follows:

1) Atoms that are directly bonded together are represented as a harmonic oscillator in terms of the interatomic separation

2) Three atoms that are bonded can have a bending potential (an oscillator function) as a function of the internal angle

3) Four atoms that form a bonded dihedral angle might have an oscillator function in terms of that angle.

4) In addition, there might be electrostatic interactions between partially charged atoms. The charges are determined by electronic structure methods (perhaps with empirical adjustments).

5) Some force fields include hydrogen bonds for O-H···O, O-H···N, etc.

6) Also, there is a 6-12 Lennard-Jones interaction between all atoms. This is especially important for describing non-bonded interactions.

CHARMM (Chemistry at HARvard using Molecular Mechanics:

``CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations",*J. Comp. Chem.* **198**3, 4, 187-217; A. D. MacKerell, Jr., et al. ``An All-Atom Empirical Energy Function for the Simulation of Nucleic Acids", *J. Am. Chem. Soc.* **199**5, *117*, 11946-11975)

$$V = \sum_{\text{bonds}} k_b \left(r - r_0\right)^2 + \sum_{\text{angles}} k_\theta \left(\theta - \theta_0\right)^2 + \sum_{1,3} K_{1,3}(S - S_0)^2$$

$$+ \sum_{\text{proper dihedrals}} \left|k_\phi\right| - k_\phi cos(n\phi) + \sum_{\text{improper dihedrals}} k_\omega (\omega - \omega_e)^2$$

$$+ \sum_{pairs, i \neq j} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{\varepsilon r_{ij}} \right]$$

Dirty Laundry: (as described in the literature)

Charges for the parameter sets were determined such that gas-phase molecule-water interaction energies and geometries were reproduced as well as dipole moments and heats of sublimation of the compounds.

Bond, angle, and dihedral force constants were set so as to match geometries and vibrational spectra for crystal structures, IR and raman intensities and 6-31G* gas-phase calculated properties.

A proper balance between solvent-solvent, solute-solvent and solute-solute interaction energies was sought with reference to the TIP3P water molecule.

Nucleic acid parameters were tested for their abilities to reproduce acid-base crystals with respect to lattice parameters, nonbonded parameters, and heats of sublimation.

vdW parameters were determined empirically

AMBER (Assisted Model Building with Energy Refinement):

S. J. Weiner, et al. ``A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins", *J. Am. Chem. Soc*., **198**2, *10*6, 765-784; S. J. Weiner, et al. ``An All Atom Force Field for Simulations of Proteins and Nucleic Acids", *J. Comp. Chem.* **198**6, 7, 230-252; W. D. Cornell, et al. ``A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules", *J. Am. Chem. Soc.* **199**5, *117*, 5179-5197)

$$\mathrm{V} = \sum_{\text{bonds}} K_r \left(r - r_{eq}\right)^2 + \sum_{\text{angles}} K_\theta \left(\theta - \theta_{eq}\right)^2 + \sum_{\text{dihedrals}} \frac{V_n}{2}\left[1 + \cos(n\phi - \gamma)\right]$$

$$+ \sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}}\right] + \sum_{\text{H-bonds}}\left[\frac{C_{ij}}{R_{ij}^{12}} - \frac{D_{ij}}{R_{ij}^{10}}\right]$$

What do you do with a protein force field?

1) If you know the structure (say from X-Ray measurements) you can "build" the protein and study its properties using molecular mechanics, molecular dynamics, and other methods. The standard source for protein structures is the Protein Data Bank (pdb). This is available at the web site: http://www.rcsb.org/pdb/

2) If you don't know the structure, but you know the sequence (say from genome data or from sequencing of an unknown protein), you can attempt to build the protein and determine its three-dimensional structure. This is the protein folding problem! Unfortunately this is an unsolved process and it is still an active area of research.

Classical Molecular Dynamics

If we have N particles, each with mass M, then the energy of the system is:

$$E = \sum_{k\alpha} \frac{1}{2} M v_{k\alpha}^2 + V$$

Here $v_{k\alpha}$ is the velocity of the αth component (α = x, y, z) of the kth particle, and V is the potential energy function.

Newton's equations of motion for this system are:

$$M \frac{d^2 X_{k\alpha}}{dt^2} = M \frac{dv_{k\alpha}}{dt} = -\frac{\partial V}{\partial X_{k\alpha}} = F_{k\alpha}$$

These define a set of 3N differential equations for the coordinates $X_{k\alpha}$. To solve these equations, we have to specify initial values of the coordinates and momenta of the particles, and then we numerically integrate the differential equations.

Integrating the classical equations of motion:

There are many approaches to numerically integrating the classical equations of motion. Since these are coupled nonlinear ordinary differential equations, stability can be a problem. One example of a numerical integration method is called the leap-frog algorithm. This algorithm (which is closely related to what is often called the velocity Verlet method) defines the velocities $v_{k\alpha}$ and coordinates $X_{k\alpha}$ as follows:

$$v_{k\alpha}(t+\frac{1}{2}\delta t) = v_{k\alpha}(t-\frac{1}{2}\delta t) + (\delta t)M^{-1}F_{k\alpha}(t)$$

$$X_{k\alpha}(t+\delta t) = X_{k\alpha}(t) + (\delta t)v_{k\alpha}(t+\frac{1}{2}\delta t)$$

In this algorithm, the energy at time t is evaluated using the following expression for the velocity at time t:

$$v_{k\alpha}(t) = \frac{1}{2}[v_{k\alpha}(t-\frac{1}{2}\delta t) + v_{k\alpha}(t+\frac{1}{2}\delta t)]$$

Molecular Mechanics: This is energy minimization, i.e., from a given starting structure, the goal is to locate the lowest possible energy structure, ideally the global minimum.

In most protein structures, it is hard to locate the true local minimum, so inevitably one finds a local minimum that one hopes is at least similar to the global minimum. Finding minima can be done in many ways. One way is to follow the gradient downhill. Since this will inevitably lead to local minima, one usually performs a series of MM and MD calculations in which the MD part is used to explore structures.

Software for doing MM and MD calculations for proteins based on empirical force fields

There are many programs available that enable one to study the properties of biomolecules. Many have the same names as the force fields that they were originally designed for: CHARMM, Amber, GROMOS. However the force field is usually a separate product from the code, and is mostly public domain. Most simulation codes are commercial although some are cheap to academic users.

One code that is entirely public domain is Tinker. Yes there is a Tinker force field as well, but this is not commonly used. However the Tinker simulation code is very popular. It has the feature that it can use all of the commonly available (public domain) force fields, including some, like MM3, that are more sophisticated than CHARMM or Amber (MM3 is usually not used for proteins as the sophistication makes applications run more slowly.)

Tinker was developed by Jay Ponder, at Washington University in St. Louis. It is available at: **http://dasher.wustl.edu/tinke**r

BRIEF OVERVIEW OF TINKER SIMULATION PACKAGE

TINKER is a system of programs and routines for molecular mechanics and dynamics as well as other energy-based and structural manipulation calculations. Rather than incorporating all the functionality in one monolithic program, TINKER provides a set of relatively small programs that interact to perform complex computations.

The series of major programs included in the distribution system perform the following core tasks:

(1) build protein and nucleic acid models from sequence
(2) energy minimization and structural optimization
(3) analysis of energy distribution within a structure
(4) molecular dynamics and stochastic dynamics
(5) simulated annealing with a choice of cooling schedules
(6) normal modes and vibrational frequencies
(7) conformational search and global optimization
(8) transition state location and conformational pathways
(9) fitting of energy parameters to crystal data
(10) Distances and geometries
(11) molecular volumes and surface areas
(12) free energy changes for structural mutations

BRIEF OVERVIEW OF AMBER SIMULATION PACKAGE

The Amber simulation package (not to be confused with the Amber force field) consists of around 50 programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of organic and biological molecules such as DNA and protein, and analysis of the structures, non-covalent interactions, and dynamic trajectories.

In order to speed up computing of large molecules, the Amber codes have been optimized for explicit solvent periodic calculations and parallel computing.

Amber was originally developed by Peter Kollman at UCSF, and is now administered by David Case at Scripps.

Web site: http://amber.scripps.edu

CAPABILITIES of AMBER PACKAGE

(1) Pre-defined libraries containing common bio-residues such as amino acids and nucleic acids
(2) Molecular graphic interface to facilitate building and visualizing molecules
(3) Energy minimization and structural optimization
(4) Molecular dynamics with parallel computing ability
(5) Particle-mesh Ewald (PME) procedure is used to handle long-range electrostatic interactions.
(6) Rectangular and truncated octahedron periodic boundaries simulations as well as non-periodic simulations.
(7) Explicit solvation and implicit solvation models
(8) Simulated annealing with a choice of cooling schedules
(9) A variety of constraints for NMR structure refinement calculations
(10) Free energy calculations using thermodynamic integration
(11) QM/MM calculations
(12) Normal modes and vibrational frequencies
(13) Tools for analyzing and processing trajectory or coordinate
(14) Analysis of energy distribution within a structure

Application of Amber force field to the determination of the mechanical unfolding of BPTI

As an application of the Amber force field (and the Amber program), we show how to simulate the mechanical unfolding of the BPTI protein. This mimics AFM experiments in which proteins are pulled apart by attaching them to AFM tips (see next slide). This also shows up in the "real" world, leading to the mechanical stability of abalone shells (see second slide), or the elasticity of muscle. BPTI is not related to any of these studies, but it is convenient for simulation.

References to mechanical unfolding experiments and theory include:

Stretching Single Protein Molecules: Titin is a Weird Spring, H. P. Erickson, Science, 276, 1090 (1997)

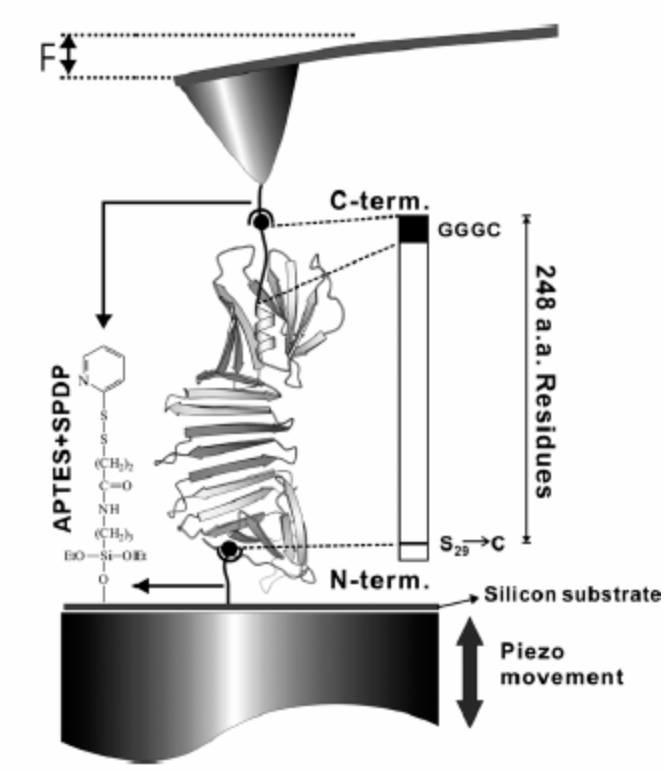
Molecular mechanistic origin of the toughness of natural adhesives, fibers and composites, Smith, B. L.; Schaffer, T. E.; Viani, M.; Thompson, J. B.; Frederick, N. A.; Kindt, J.; Belcher, A.; Stucky, G. D.; Morse, D. E.; Hansma, P. K. *Nature* **1999**, *399*, 761.

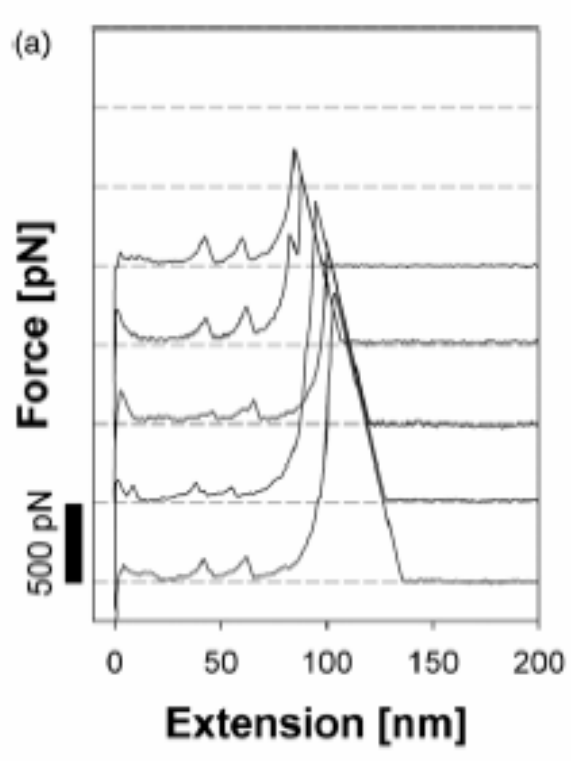
Ubiquitin-like Protein Domains Show High Resistance to Mechanical Unfolding Similar to That of the I27 Domain in Titin: Evidence from Simulations. Li, Pai-Chi; Makarov, Dmitrii E. Journal of Physical Chemistry B (2004), 108(2), 745-749.

Journal of Molecular Biology (2003), 333(5), 993-1002.

Unfolding Mechanics of Multiple OspA Substructures Investigated with Single Molecule Force Spectroscopy

Rukman Hertadi[1], Franz Gruswitz[2], Lin Silver[2], Akiko Koide[2,3]
Shohei Koide[2,3], Hideo Arakawa[1] and Atsushi Ikai[1*]





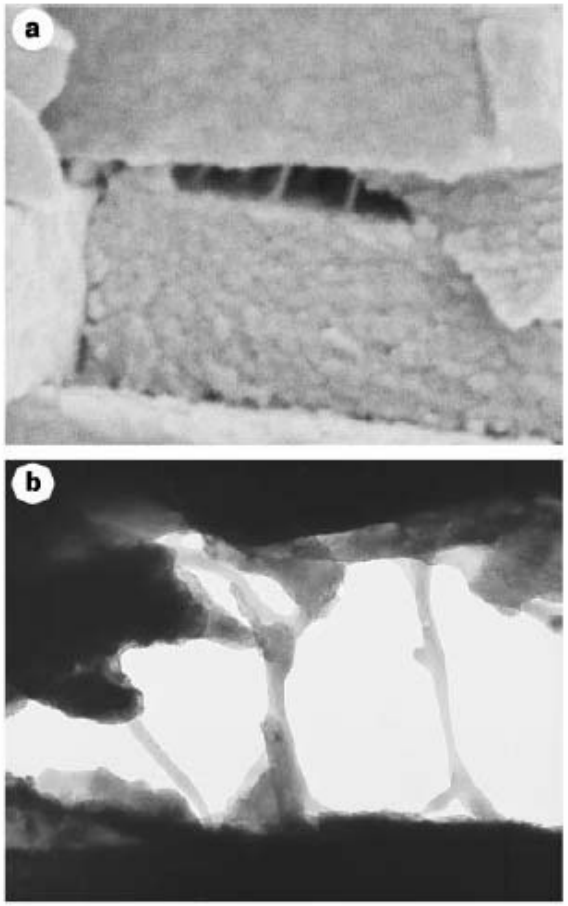


Figure 1 Scanning and transmission electron micrographs of a freshly cleaved abalone shell, showing adhesive ligaments formed between nacre tablets. **a**, Scanning electron micrograph of a freshly cleaved abalone shell showing adhesive ligaments formed between consecutive abalone nacre tablets on exertion of mechanical stress. The tablets are ~400 nm thick. **b**, Transmission electron micrograph of another cleaved abalone shell, showing the adhesive ligaments between nacre tablets. The space between the tablets is ~600 nm. Thus the ligaments can lengthen to many times the original spacing between the tablets, which is of the order of 30 nm.

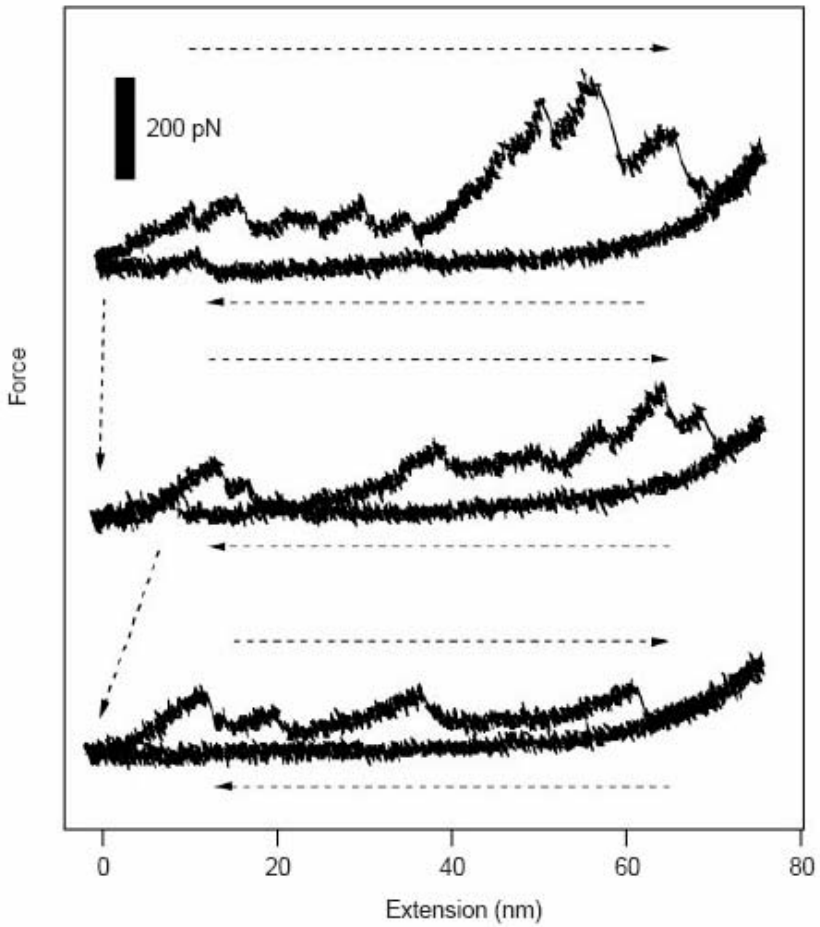


Figure 2 Consecutive force–extension curves, obtained using an atomic force microscope, from pulling on a freshly cleaved abalone nacre surface. Rupture events, with a sawtooth appearance, are visible in each of the curves. The surface was not touched between pulls, strong evidence that some refolding took place, possibly of domains in lustrin A. The approach and retract curves show hysteresis, indicating that the rupture events dissipate energy.

Smith, B. L.; Schaffer, T. E.; Viani, M.; Thompson, J. B.; Frederick, N. A.; Kindt, J.; Belcher, A.; Stucky, G. D.; Morse, D. E.; Hansma, P. K. *Nature* **1999**, *399*, 761.

Details of BPTI calculation

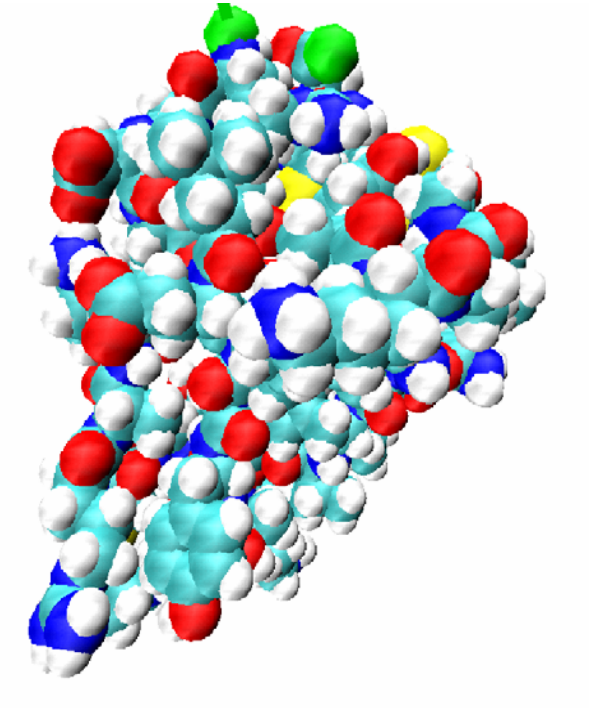
First, we download the 6PTI.pdb file of bovine pancreatic trypsin inhibitor (BPTI) from Protein Data Bank at following link:
http://www.rcsb.org/pdb/cgi/explore.cgi?pid=769111393047 9&pdbId=6PTI

The X-Ray structures usually have a few “problems” including missing or ill-defined residue structures. In the present case we had to define we had to define the conformation of residues 39 and 50, changed the cysteine residue name “CYS” to CYX, and deleted the CONECT records.

We also have to decide what to do about the solvent in which the structure is going to be immersed. “Explicit” solvent means that we surround the protein with lots of water molecules, while “Implicit” solvent means that we add additional terms to the protein force field that mimics the effect of interaction with the solvent. Here we chose to use the generalized Born implicit solvation model.

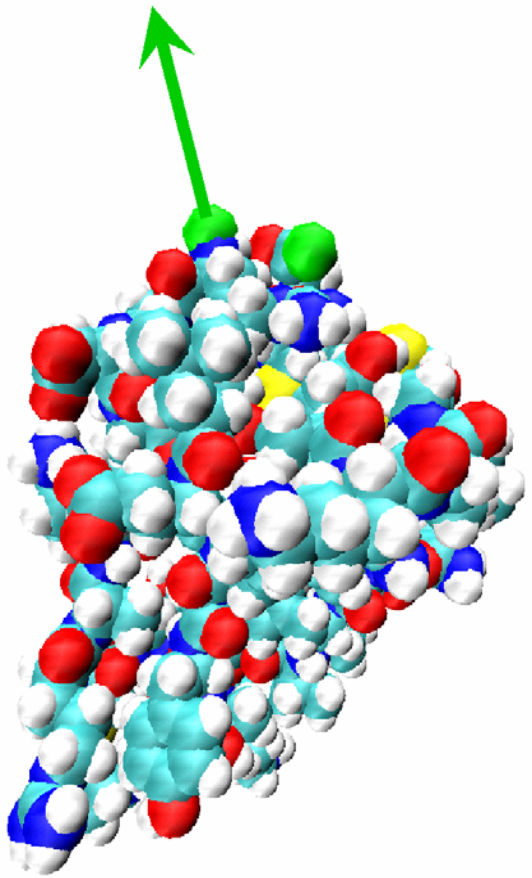
To study mechanical properties, it is necessary to exert force against portions of the molecule. Here we chose to pull on one end of the peptide chain while the other end is fixed. To do this we bonded two dummy atoms to the C- and N-terminal of the protein with a force constant of 100 kcal/mol Å^2 and a bond length of 1Å. All other force field parameters of the dummy atom such as angle, torsion, van der Waals and mass are all zero. (Note: Since we only did energy minimizations, a dummy atom with zero mass is acceptable. However, in a molecular dynamic simulation, the mass of dummy atoms must be some nonzero number, otherwise, simulation will crash.)

The BPTI protein with the dummy atoms is shown to the right.

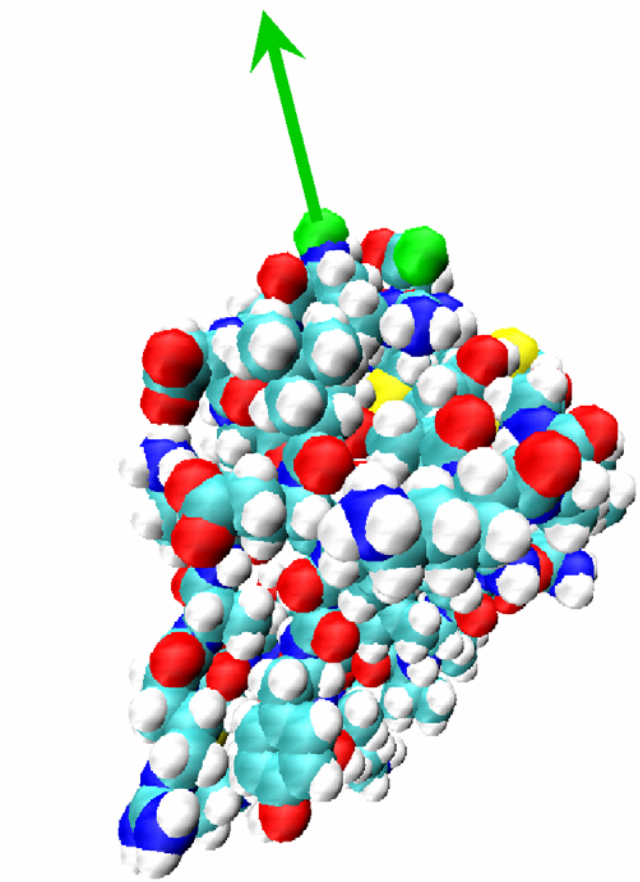


Color map: red – oxygen, white – hydrogen, blue – nitrogen, cyan – carbon, yellow – sulfur, green – dummy atom

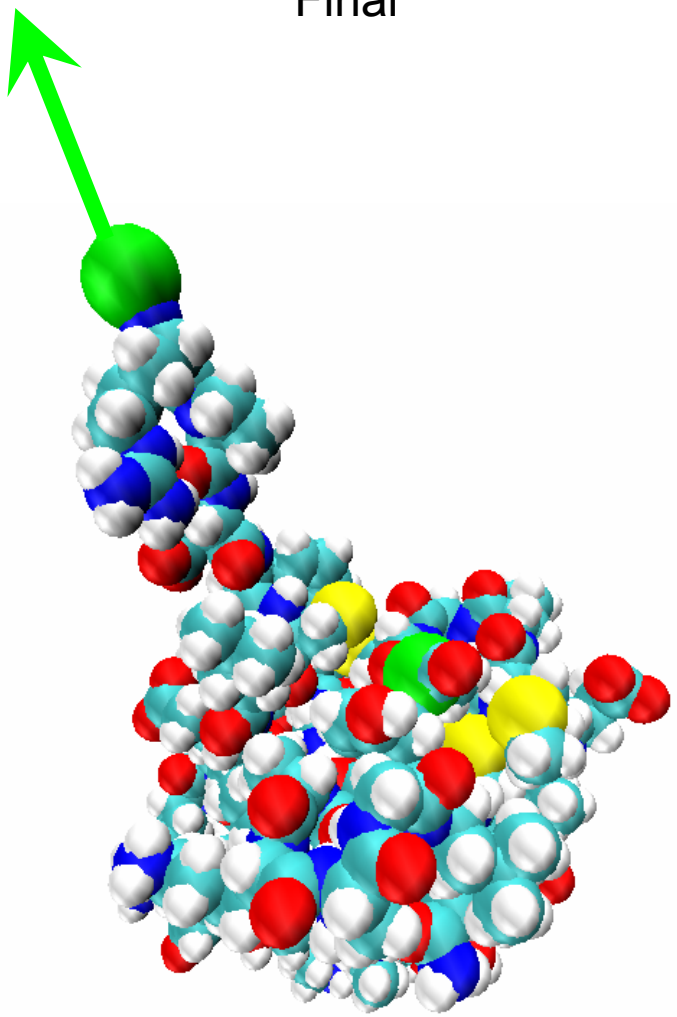
Next, we fixed the dummy atoms and performed a 25000-step energy minimization on this protein using steepest descent method and generalized Born solvation model. After the minimization, we manually modified the minimized structure to move the N-terminal dummy atom 0.5 Å in the direction from the nitrogen atom to the dummy atom. Then we did energy minimization again with both dummy atoms fixed. We repeated moving the N-terminal dummy atom 0.5 Å in the same direction, followed by minimization. This process was terminated when two residues of the N-terminal part of protein were pulled out.



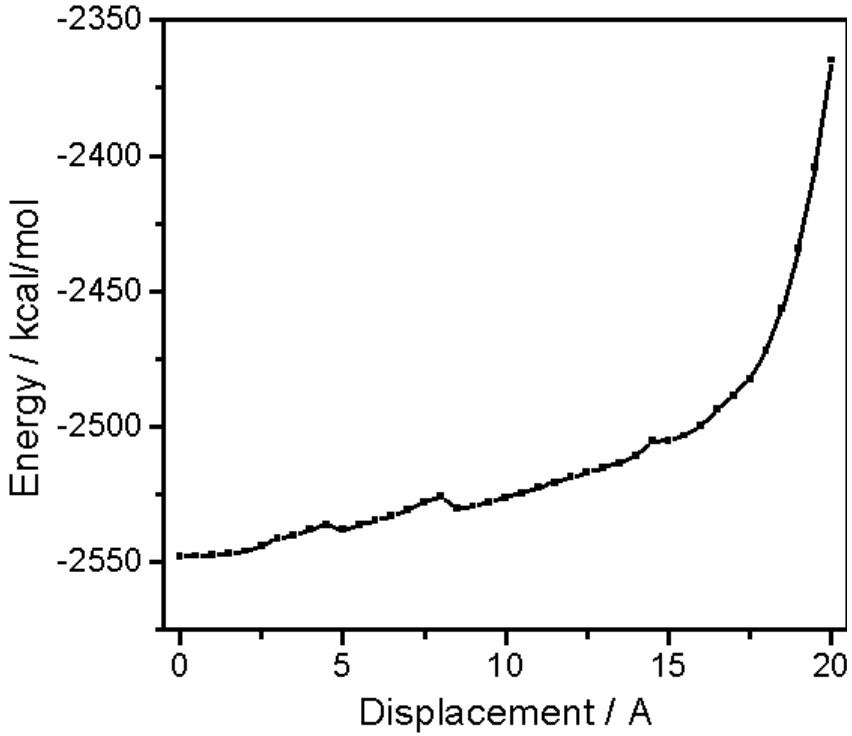
Initial

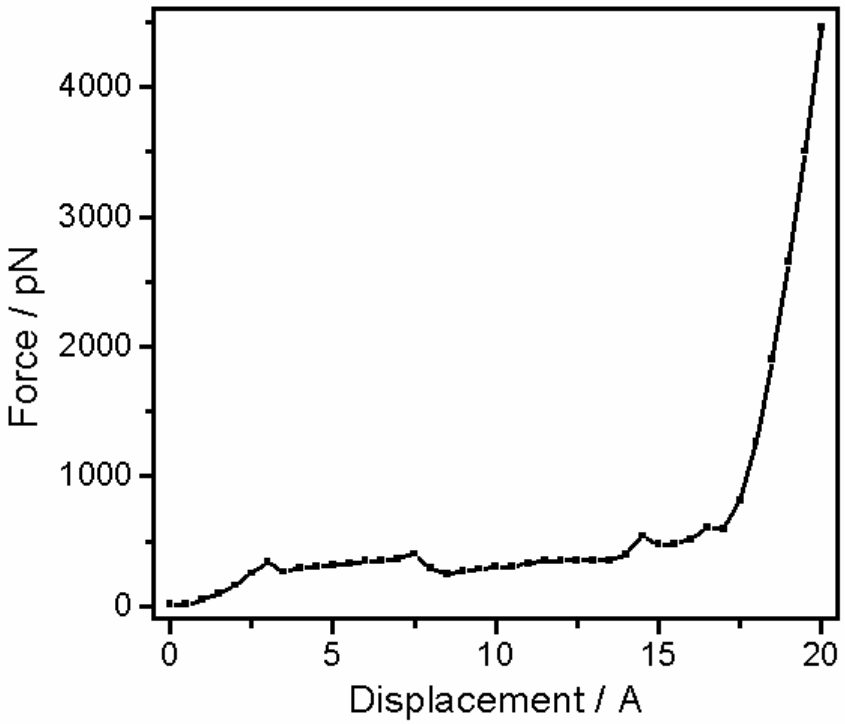


Final

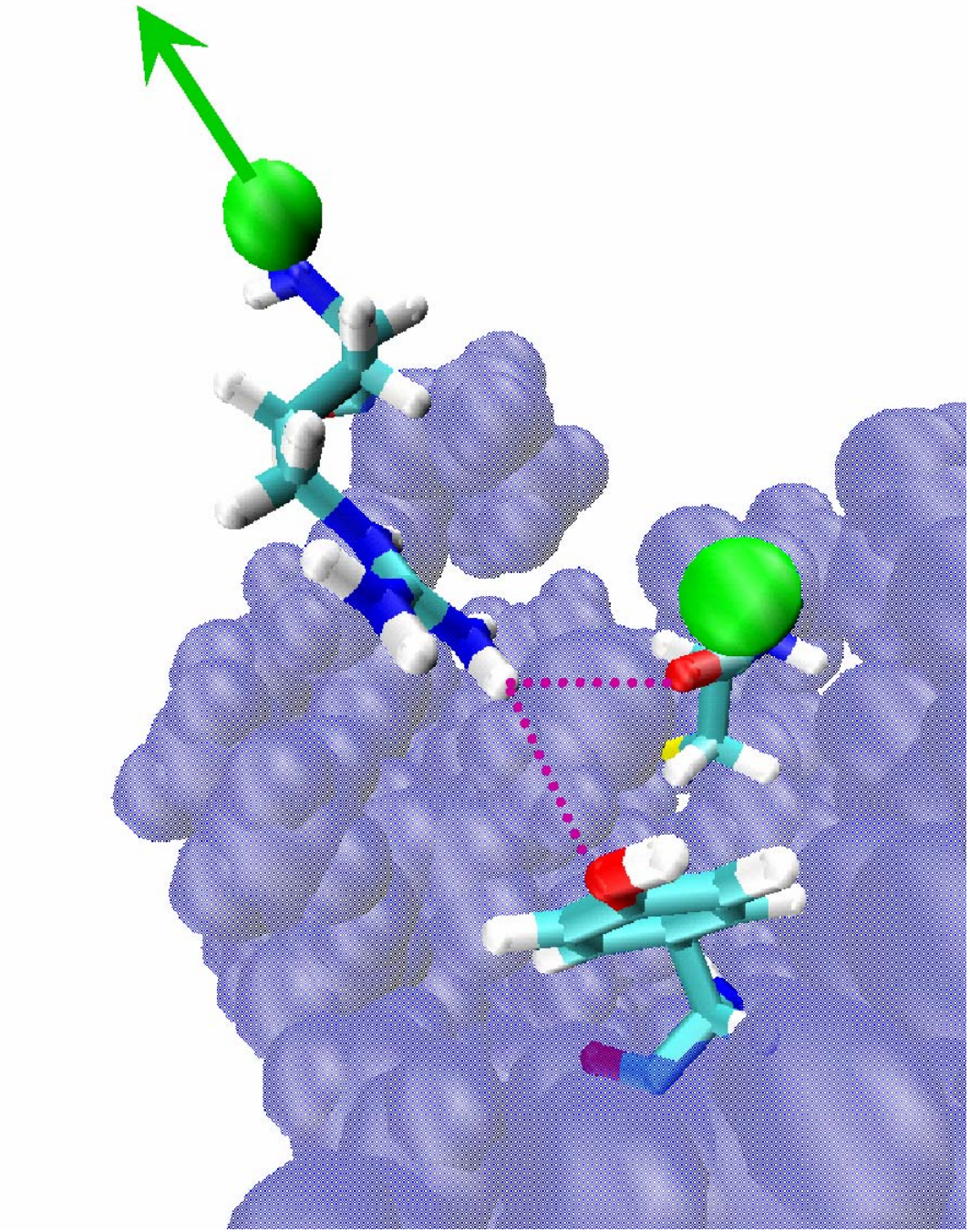


The energy and force versus dummy atom displacement for this process are shown below. There are small peaks on the curves, which occur when hydrogen bonds are broken. For example, the peak at 8A is from breaking hydrogen bonds between the N-terminal arginine residue and nearby residues (shown in next page).





The peak at 8A is from breaking hydrogen bonds between the N-terminal arginine residue and nearby residues

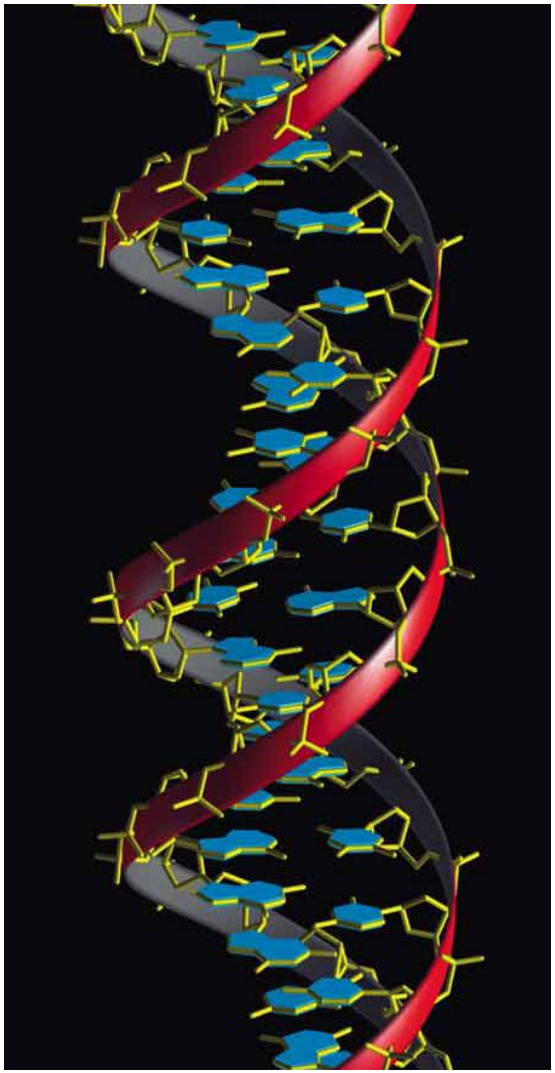


Molecular Dynamics Studies of DNA Structures

Hai Long and George C. Schatz

Northwestern University

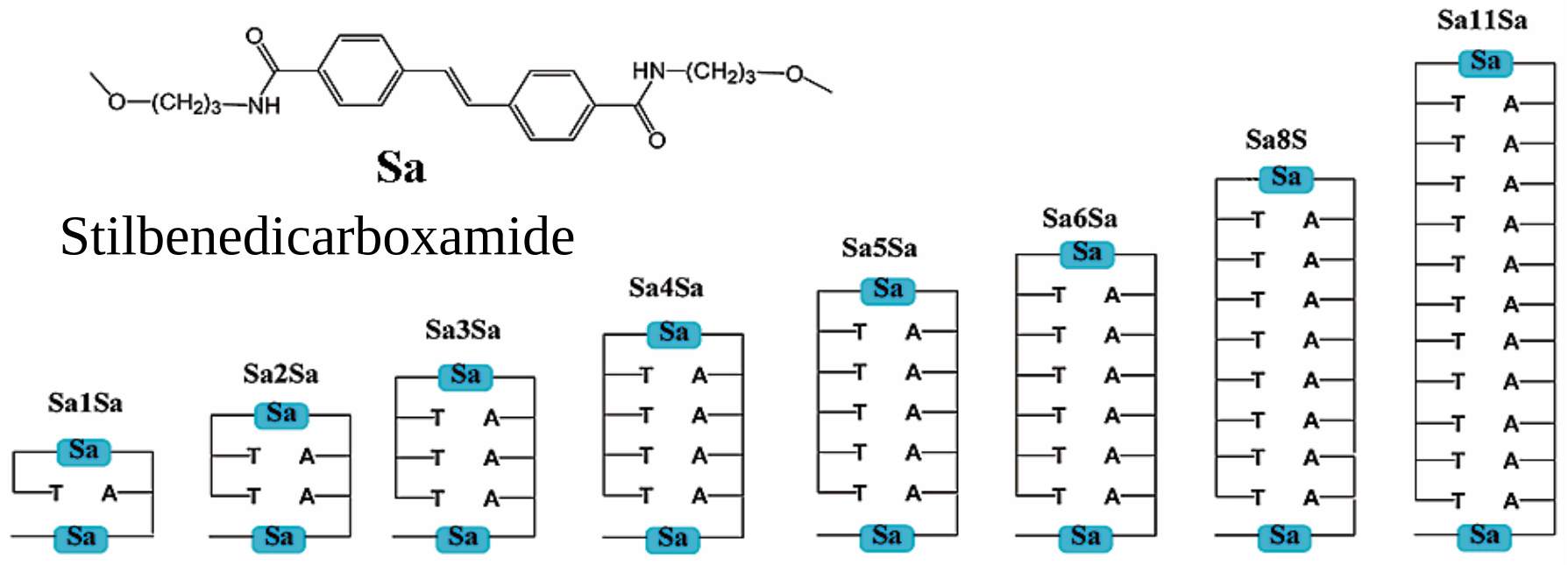
B-Type DNA Structure



- Right handed double helix
- DNA is highly charged
- 36º per base pair

  10 residues per turn
- 3.38 Å between base pairs or 33.8 Å between one turn
- Radius ~10 Å
- Contains a major and a minor groove

DNA Hairpins



Frederick D. Lewis, Xiaoyang Liu, Yansheng Wu, and Xiaobing Zuo, *J. Am. Chem. Soc.*, 125 (42), 12729 -12731, 2003.

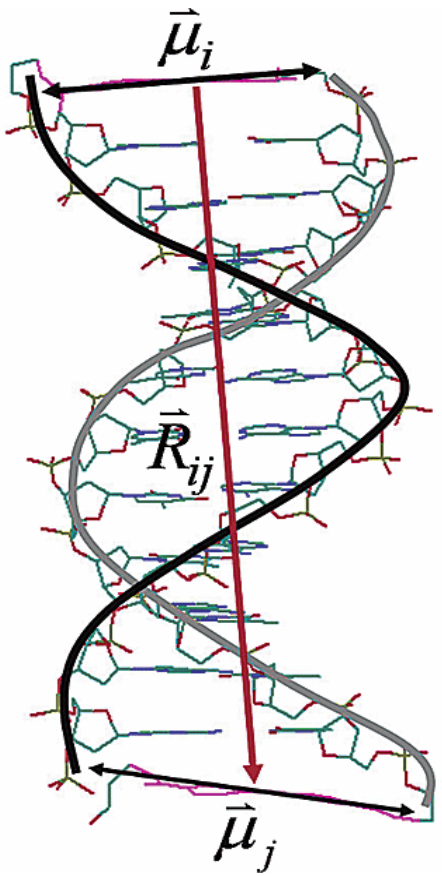
Properties of the Hairpin DNAs

- Stepwise evolution of circular dichroism (CD) spectra

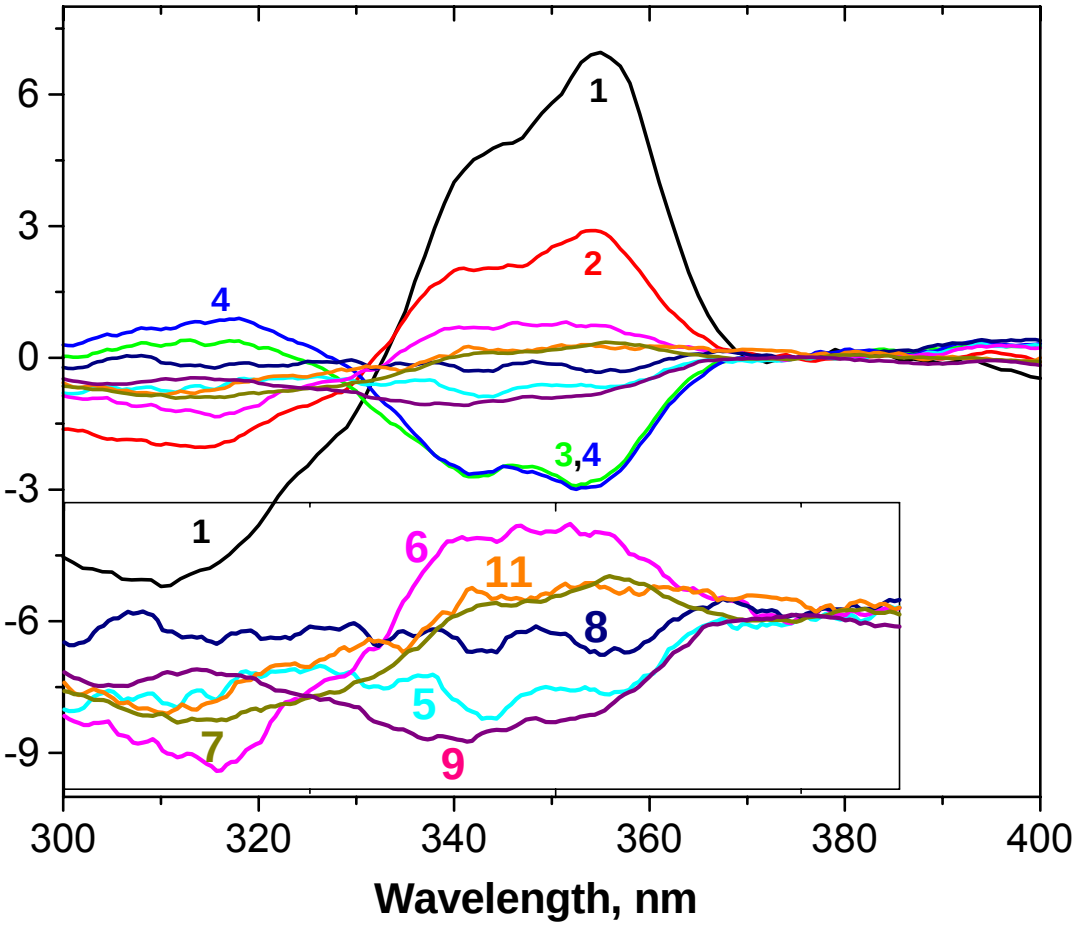
  CD Spectra depends on distance as well as the angle between stilbene chromophore

  3 vectors: μ_i, μ_j, and $\mathbf{R_{ij}}$

  Can we calculate **CD Spectra** by MD simulations?



Experimental CD
spectra of Sa(n)Sa.



Molecular Dynamic (MD) Simulation of DNA

- Force Field used in MD simulation

  CHARMM

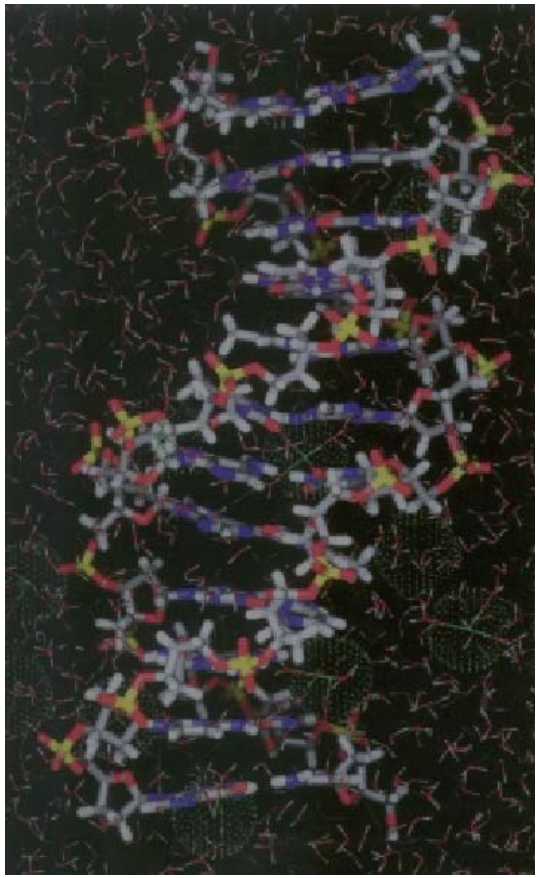
  GROMOS

- ✓ AMBER

$$E_{total} = \sum_{bonds} K_r (r - r_{eq})^2 + \sum_{angles} K_\theta (\theta - \theta_{eq})^2$$

$$+ \sum_{dihedrals} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j} [\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}}]$$

Non-bond: Van der Walls
Electrostatic

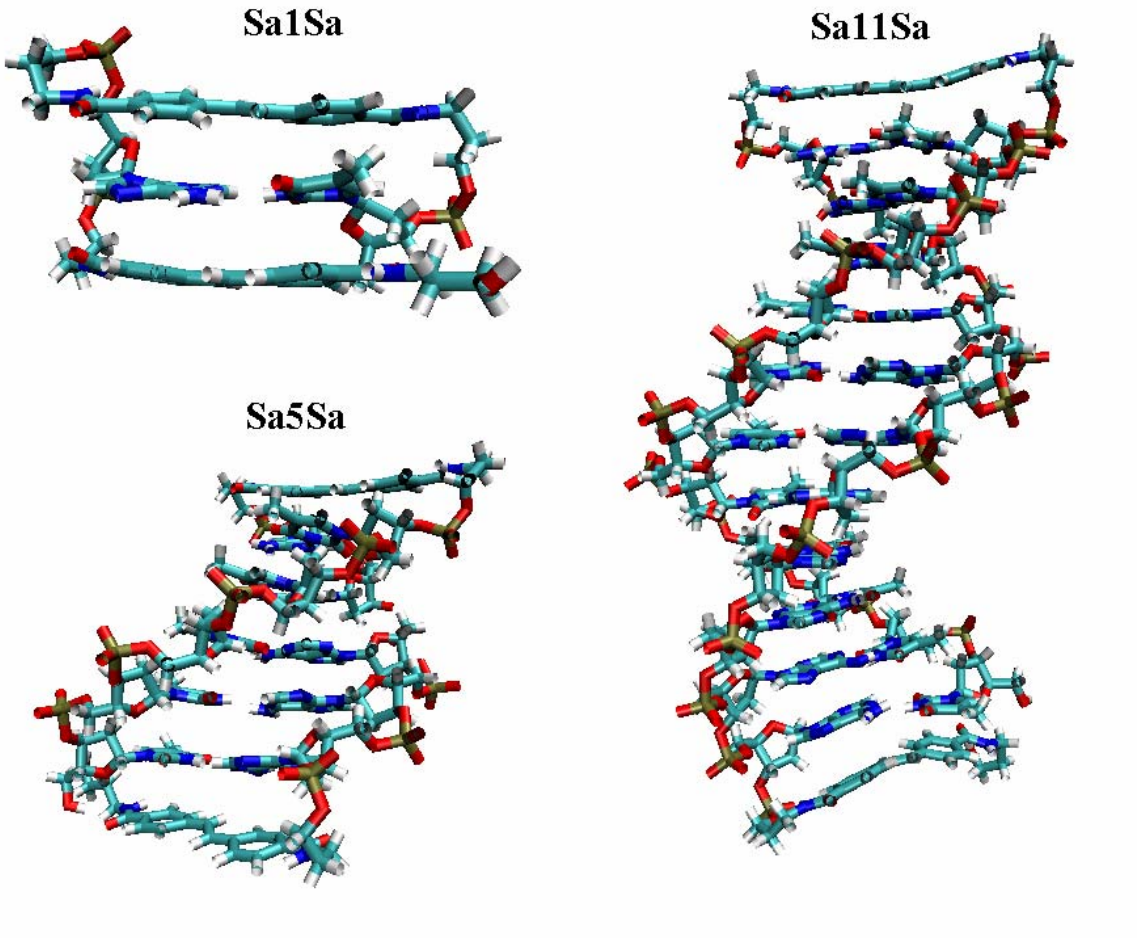


MD Simulation of Hairpin DNAs

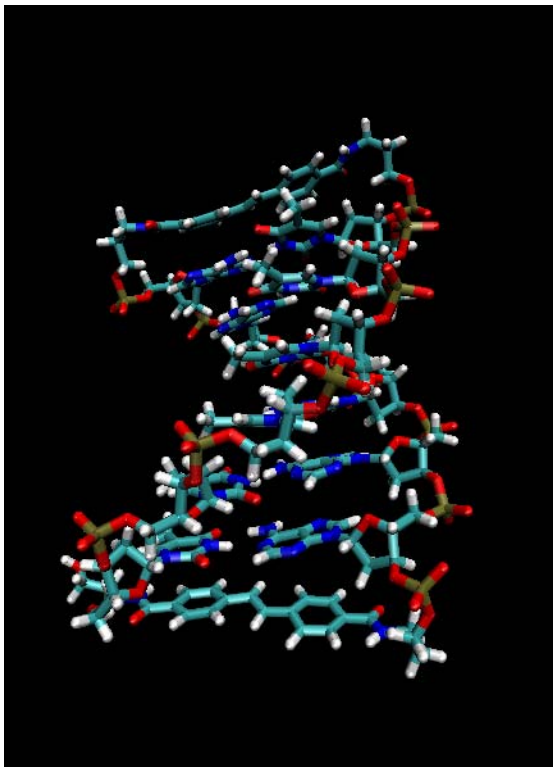
- Partial charge of atoms in the Sa residue:
  Calculated by GAMESS at 6-31G* level and fit by RESP method

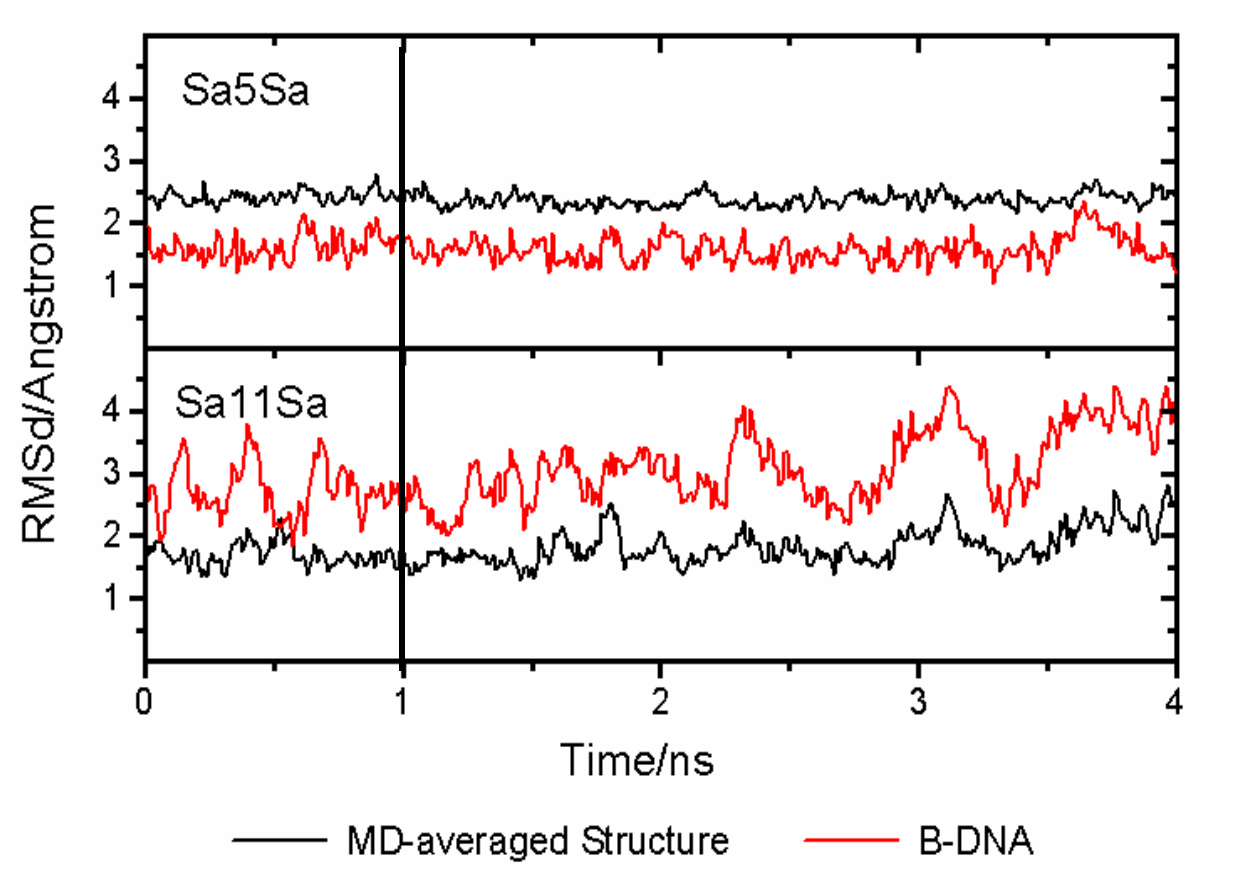
- Explicit water simulation
- After 1ns equilibrating run, begin sampling trajectories for another 3ns
- Family of the Hairpin DNAs:
  Sa1Sa, Sa2Sa, Sa3Sa, Sa4Sa, Sa4Sa, Sa5Sa, Sa6Sa, Sa8Sa, Sa11Sa

Snapshots of MD simulation of DNA Hairpins

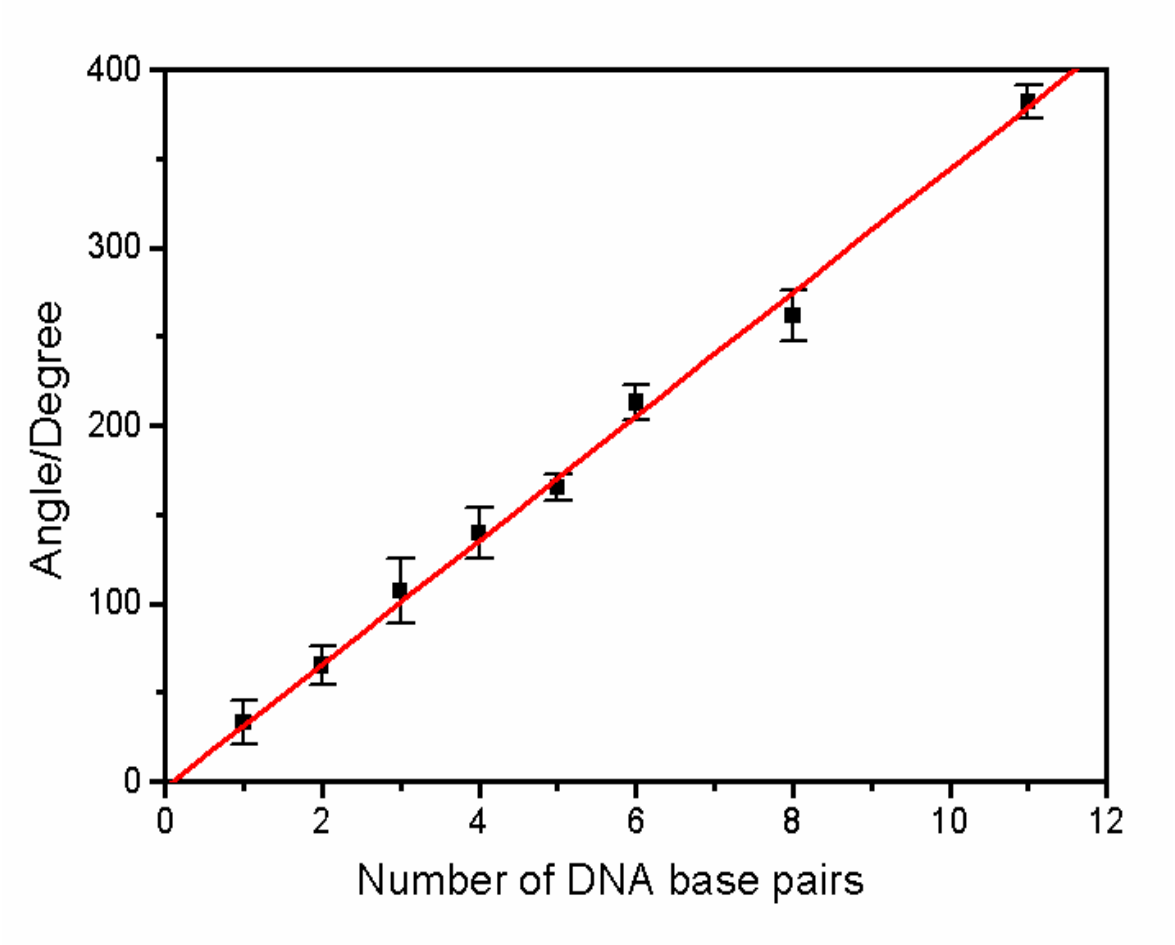


MD Simulation

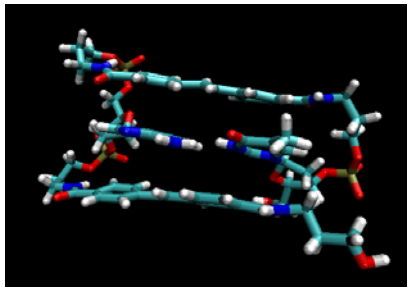




Angle Correlation

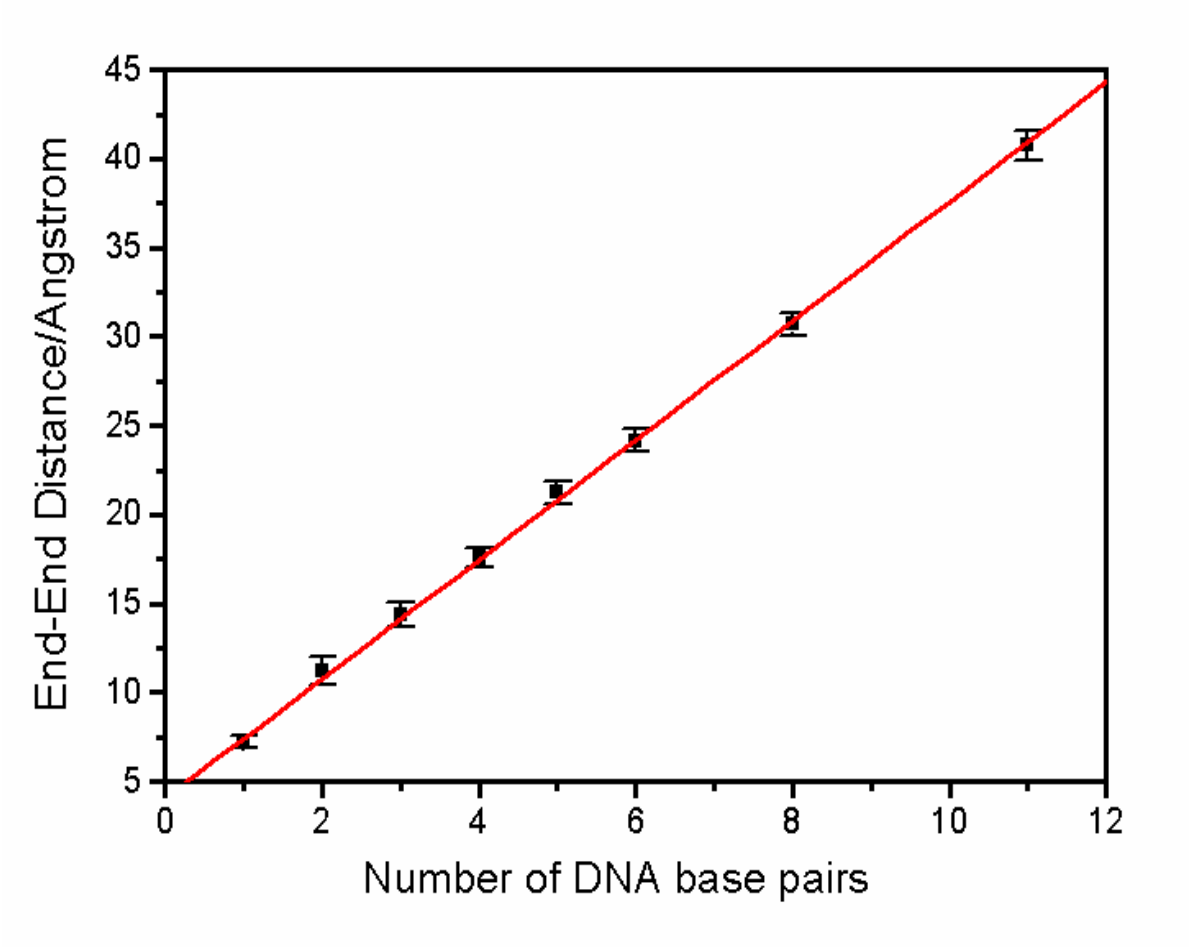


When X=1, Y=30˚ The average angle between one Sa and one DNA base pair is 15º

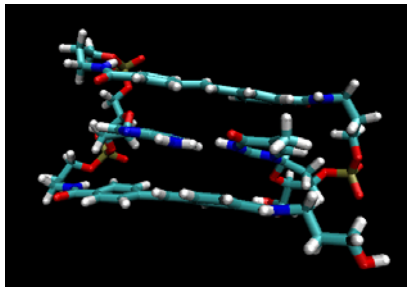


The angles between the Sa residues as a function of number of DNA base pairs. The linear fitting result is $y = (-4 \pm 5) + (34.8 \pm 0.8)x$ with R=0.9986.

Distance Correlation



When X=1, Y=7.4A
The average height between one Sa and one DNA base pair is 3.7A

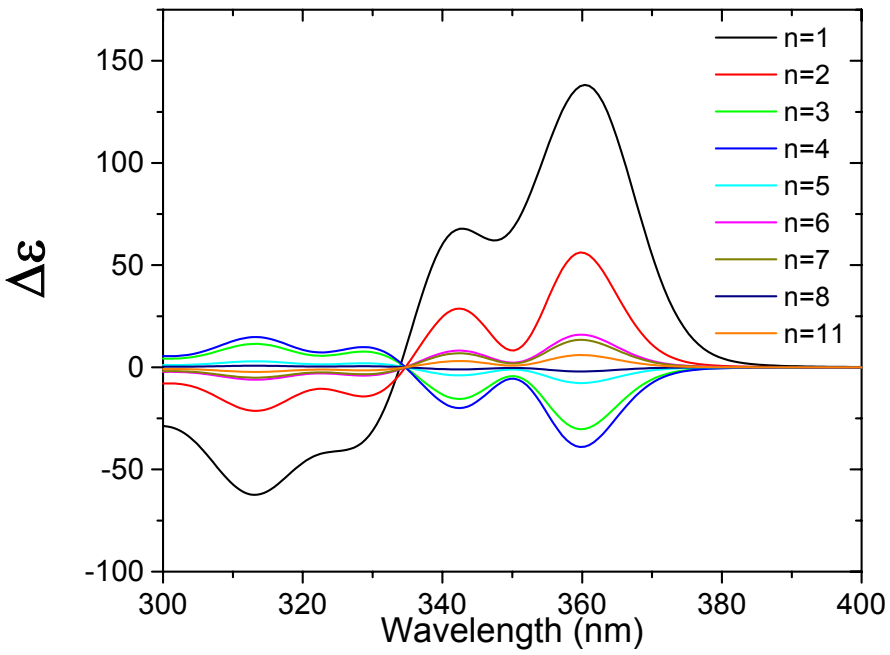


The distances between the Sa residues as a function of number of DNA base pairs. The linear fitting result is $y = (4.0 \pm 0.2) + (3.36 \pm 0.03)x$ with R=0.9997

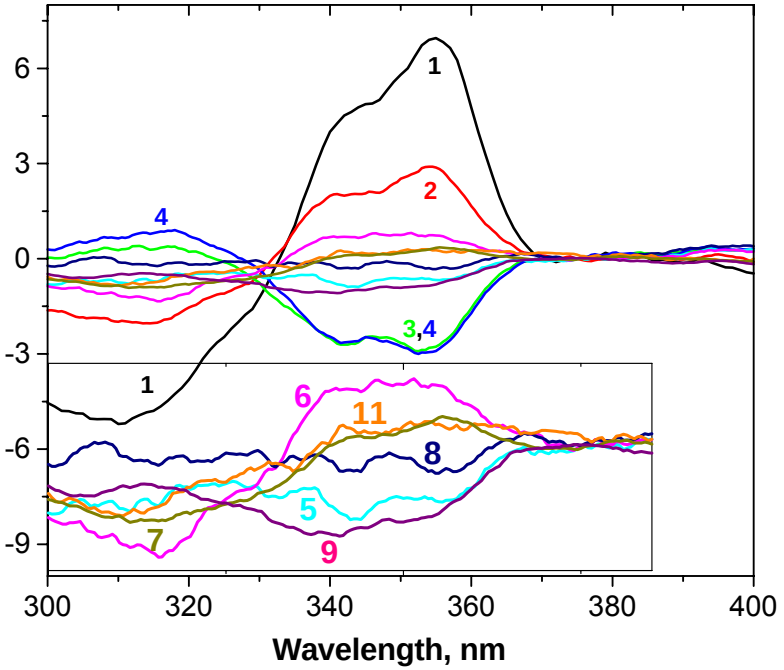
Calculation of CD Spectra from MD Simulation Results

Calculated by Xiaobing Zuo

Calculated CD Spectra using average angle and distance from simulations



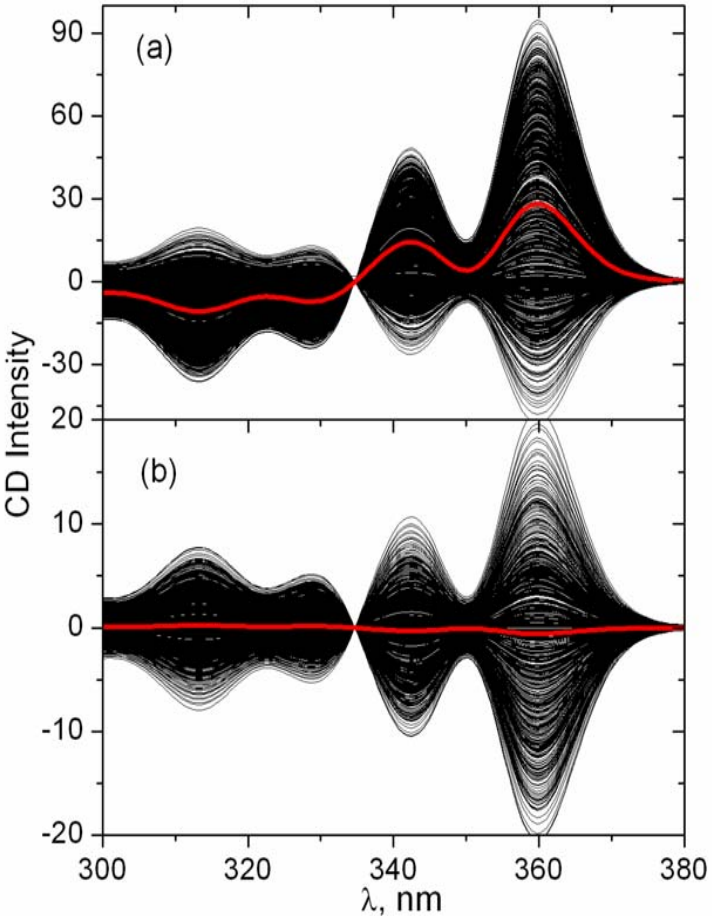
Experimental CD spectra of Sa(n)Sa.



Calculation of CD Spectra from MD Simulation Trajectories

CD spectra for conjugates 2 (a) and 5 (b) calculated from snapshots of MD simulated geometries

Calculated by Xiaobing Zuo



Experimental CD spectra of Sa(n)Sa.