Multi-Scale Modelling of DNA:

Parametrizations of Coarse-Grain Sequence-Dependent Models of DNA Mechanics

John H. Maddocks

Institut de Mathématiques B

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

IPAM, September, 2005

Good general principle:

The answer should depend on the question...

Thus to identify the appropriate level of coarse-graining need to understand the scale of the experimental data that is to be modelled.

Today I'll only discuss naked, Crick-Watson, B-form double helical DNA at various length scales.

The WLC/TWLC/Vanilla Elastic rod or Euler Elastica proven to be a very useful model at scales of a thousand or so base pairs on up, i.e. at scales of many persistence lengths where effects are largely entropy dominated and sequence-dependent effects not so crucial.

WLC/TWLC/Elastica has the Hamiltonian (in stat mech language) or Lagrangian (in mechanics language) or just energy

$$U(\kappa(s), u_3(s); B, C, \hat{u}_3) := \frac{1}{2}\int_0^L B\kappa^2 + C(u_3 - \hat{u}_3)^2 \quad ds$$

or a simple finite difference approximation thereof.

Only three constant, model parameters A=B, C, $\hat{u}_3$, so can fit directly to experiment, which is the inverse problem in math speak, and off you go to simulate other similar systems, or various direct problems, to get interesting information.

Note that the TWLC/Elastica is a) isotropic (equal bending responses), and b) uniform (constant coefficient).

Modifications available for various things, for example self-avoidance, but sequence-dependence is at most rather weak.

But in some circumstances this basic model is not appropriate:

- It is believed that length scales of 10's–100's of bps very pertinent to biology. DNA wrapped on nucleosomes, packed in phage heads, looping of repressors and promoters, TATA boxes, A-tracts, ...

- For 10–200 bp i.e. at or below a persistence length, sequence-dependent effects of elasticity and shape seem to be important. The details of AT pairing, versus CG pairing, and different stackings between purines and pyrimidines seems to modulate both the intrinsic shape and the stiffnesses of the idealized sequence-independent Crick-Watson B-form double helix.

Quantifying such sequence-dependent modulations via computational modelling and experimental verification is my (and several other) group's immediate goal.

“DNA cyclization is potentially the most powerful approach for systematic quantitation of sequence-dependent DNA bending and flexibility.”

First sentence of abstract in Zhang & Crothers, Biophysical J. **84** (2003)

DNA minicircle cyclization rates, i.e. probability of loop formation, are exquisitely dependent upon sequence, particularly intrinsic curvatures, at scales of 50 – 250 bp. Such in vitro cyclization experiments measuring J-factors à la Shore-Baldwin, Crothers group, Kahn, Widom, Vologodskii, etc, are currently a *very* active experimental area.

Maddocks take home messages Part 1:

- The inverse problem to obtain detailed sequence-dependent parameter sets directly from such experiments is hopeless, too many parameters to fit, 10 independent dimer steps, 136 independent tetramer steps, and, depending on the level of detail in your model, 9 or 21 or more shape and stiffness parameters to find for each of the 10 or 136 possible steps.

- But the experiments provide an excellent reality check on a given set of model parameters. And given a set of model parameters and the appropriate computational approach the direct problem of predicting J-factors is not so hard.

Maddocks group current approach:

- Take short scale Molecular Dynamics atomistic description of 10-20 bps as the ‘truth’. (Could perhaps instead take crystal structure or NMR data as the ‘truth’.)

- Extract parametrizations, or constitutive relations, for coarser-grained sequence-dependent models, rigid base, or rigid base-pair, or continuum elastic descriptions, and pass up to longer scales

- Use continuum mechanics techniques to make numerically efficient computations with error control, at the longer scales.

- Seek quantitative comparison between model predictions and experimental data to test the model validity. Ultimately a check on the MD potentials.

Talk is structured backwards:

- First review older work of Manning, JHM, Kahn (1996, J. Chem Phys) that shows how to use continuum mechanics to compute minicircle shapes given constitutive relations for sequence-dependent rigid base models

- Then discuss ongoing, and unfinished, work on how to extract constitutive relations from MD.

Coarse-graining and Statistical Mechanics

Basic modelling question for DNA in solution is to understand how the sequence affects the configuration-space equilibrium distribution of various things according to the measure

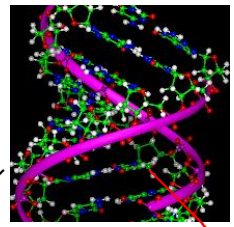
$$\frac{1}{Z}\exp[-\beta U(q)]J(q)$$

for a choice of configuration variable q, potential U, Jacobian J, partition function Z, and temperature scale β.

Often need to study a conditional distribution subject to possibly nonlinear side conditions, such as nonlocal closure (or cyclization) constraints on the configuration q.

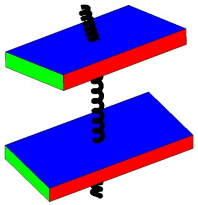
The choice of the configuration variable q is the choice of level of coarse (or fine) graining in your model–e.g. atomistic, rigid-XX, or continuum.

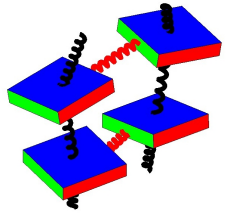
Next Message: It is more efficient to compute at the bottom level!



Chain of Rigid Base-Pairs

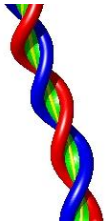
Chain of Rigid Bases





Elastic Rod

Elastic Birod



Want to give a sketch of a justification of the claim that computations with the continuum models at the bottom are better:

A first step in understanding equilibrium distributions is to understand

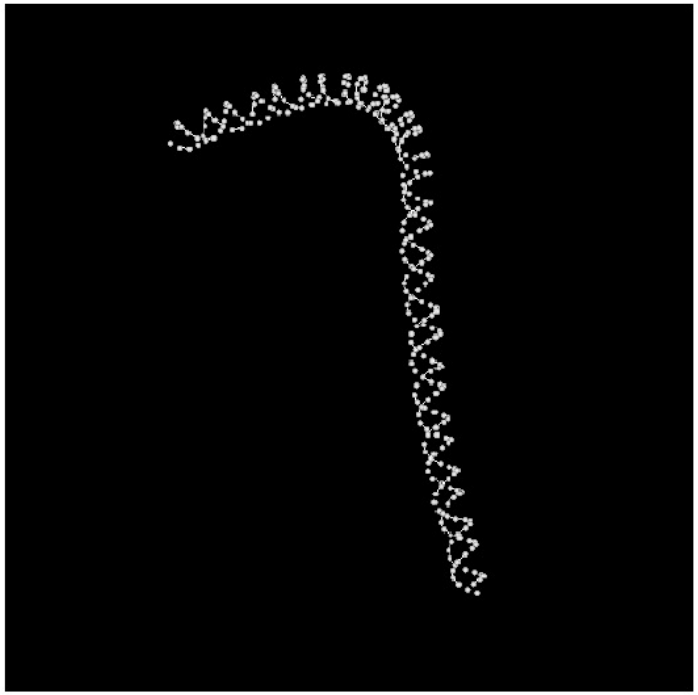
- the location and number of their peaks (constrained local minima of U)
- the height of energy barriers (constrained saddles of U)

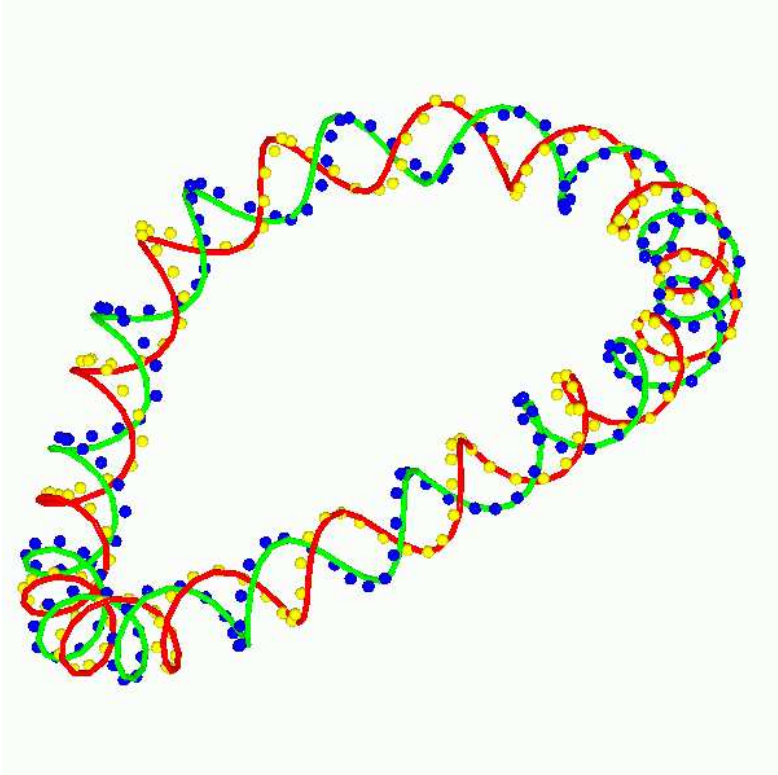
Thus the $T = 0$ computation of understanding critical points of $U(q)$ subject to nonlinear, nonlocal closure conditions, can give good finite T information.

And, as pointed out by Zhang & Crothers in the particular context of DNA mini-circles, can get entropic corrections according to a harmonic approximation in each constrained well of the potential energy, at least in discrete models.

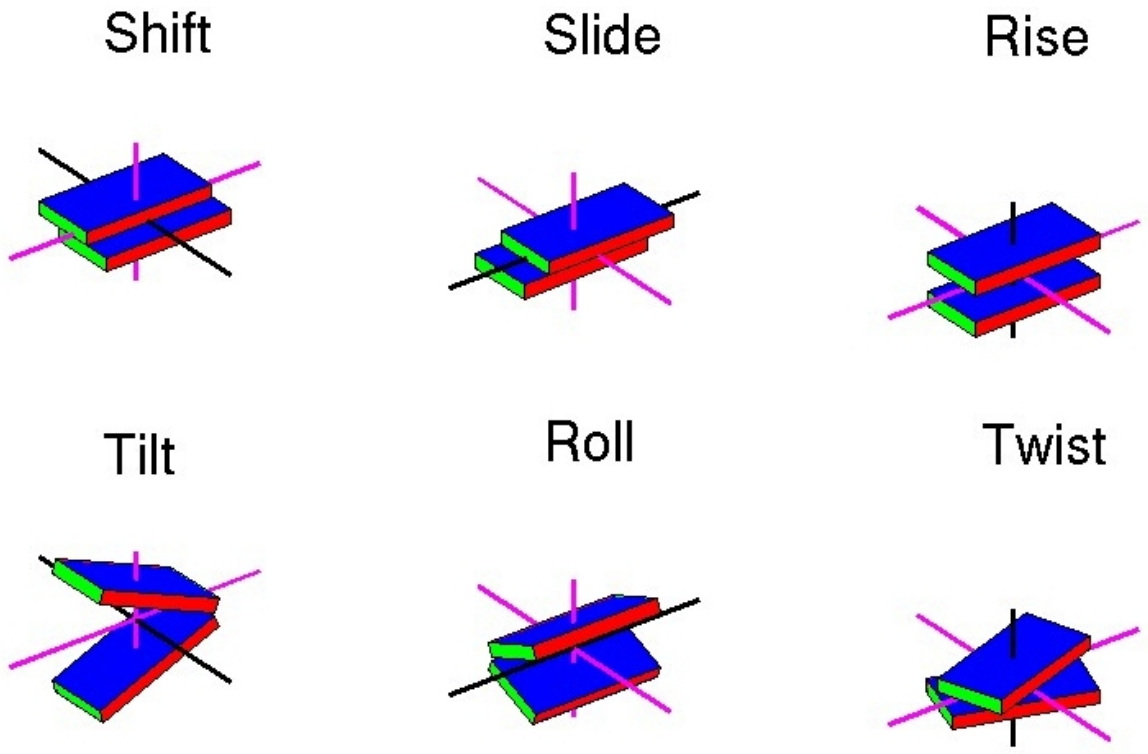
(Continuum version of that theory in the specific context of DNA using theory of path-integrals is ongoing work of L. Cotta-Ramusino.)

Then the basic question in predicting J-factors is the following. Given an oligomer described by a model at say the rigid base-pair level, compute the minimum energy shape when it is closed (cyclized) at some given link.





The rigid base-pair description of DNA involves the Inter-Base Pair Deformations



Three translational strains v_i^k, $i = 1, 2, 3$ at each junction $k = 1, \ldots, N$, and three rotational strains u_i^k, $i = 1, 2, 3$ at each junction $k = 1, \ldots, N$.

A ‘standard’ energy is that the translation variables are constrained to take given values $\hat{v}_i^k$ while there is an energy of bending of the form

$$\sum_{k=1}^{N} W^k(\mathrm{u}^k - \hat{\mathrm{u}}^k),$$

where

$$W^k(\mathrm{u}^k - \hat{\mathrm{u}}^k) = \frac{1}{2} \sum_{\mathrm{i,j}=1}^{3} \mathrm{K}_{\mathrm{ij}}^{\mathrm{k}}(\mathrm{u}_{\mathrm{i}}^{\mathrm{k}} - \hat{\mathrm{u}}_{\mathrm{i}}^{\mathrm{k}})(\mathrm{u}_{\mathrm{j}}^{\mathrm{k}} - \hat{\mathrm{u}}_{\mathrm{j}}^{\mathrm{k}}),$$

(Nine unknown parameters per base pair step in this model.)

The sequence-dependence of unstressed shape is supposed to be captured in different values of $\hat{v}_i^k$ and $\hat{u}_i^k$ for different junctions k dependent upon the nearby (dimer or tetramer) sequence.

Several published ‘wedge angle sets’, ie tables of sequence-dependent values of these intrinsic shape parameters.

The sequence dependence of stiffness is supposed to be captured in the values of K_{ij}^k in the same way. Not so much known about the appropriate values.

Leaving aside the question of how accurate this rigid-base pair model actually is (pretty clear now that it isn't so accurate), pose the question of how to compute the equilibrium shapes of configurations that have been cyclized at various links.

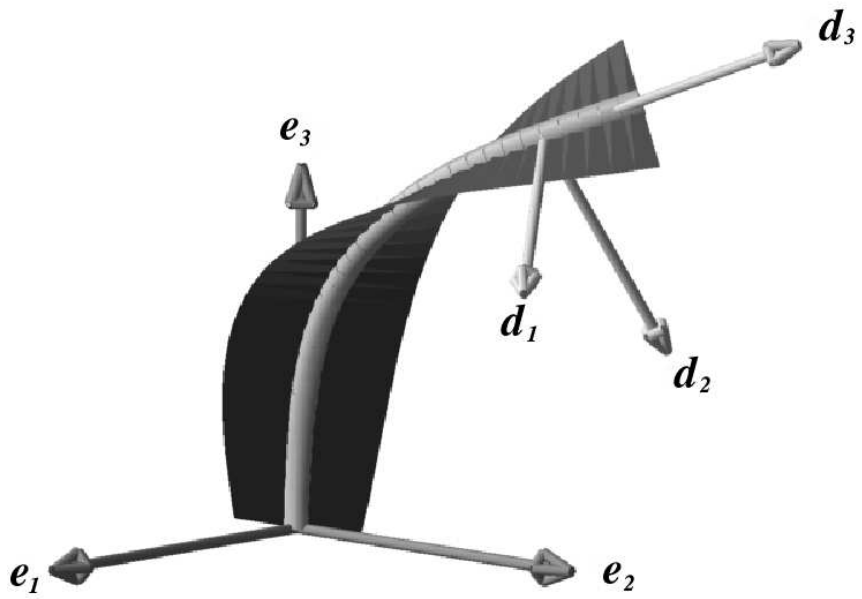
Taking the continuum limit of this system is straightforward. Just recognize the sum as being a finite difference approximation to an integral and interpolate the coefficients.

In this case the continuum limit is an inextensible, unshearable elastic rod.

An inextensible, unshearable elastic rod has the kinematics

$$\mathbf{d_i'} = \mathbf{u} \times \mathbf{d_i}, \quad \mathbf{i} = \mathbf{1}, \mathbf{2}, \mathbf{3}.$$

The components $u_i(s) := \mathbf{u}(s) \cdot \mathbf{d}_i(s)$ with respect to the frame $\{\mathbf{d}_i\}$ are the bending ($i = 1, 2$) and twisting ($i = 3$) strains. We gather the three components u_i in a triple u.



...and an assumed quadratic energy density

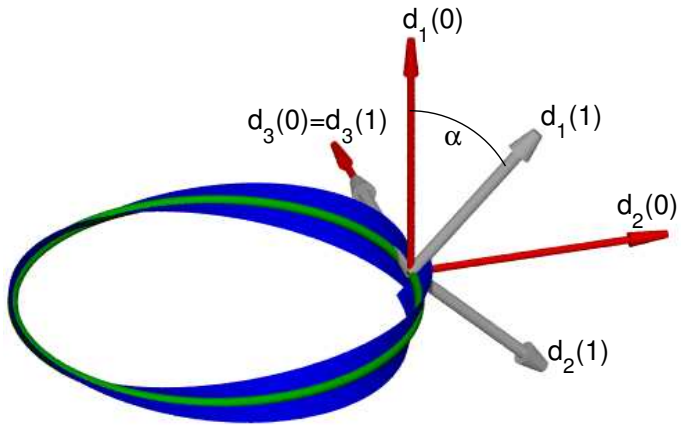
$$W(u - \hat{u}, s) = \frac{1}{2} \sum_{i,j=1}^{3} K_{ij}(s)(u_i - \hat{u}_i)(u_j - \hat{u}_j),$$

where $\hat{u}(s) = (\hat{u_1}(s), \hat{u_2}(s), \hat{u_3}(s))$ are the components of strain in the minimum energy unstressed configuration and the $K_{ij}(s)$ form the stiffness matrix.

Then the equilibrium configurations of a minicircle are the critical, or stationary, points of the calculus of variations problem

$$\int_0^1 W(\mathrm{u} - \hat{\mathrm{u}}, \mathrm{s}) - \mathbf{n} \cdot \mathbf{d}_3 \, \mathrm{ds},$$

subject to appropriate (nonlinear and parameter dependent) two-point boundary conditions.



The equilibrium conditions are the first-order necessary conditions, or Euler-Lagrange equations, associated with this calculus of variations problem, i.e. a two-point boundary value problem for a second-order system of ODE.

Because there is a scalar parameter, namely the angle α, expect locally one dimensional families of equilibria

As W is assumed convex, there is an associated Hamiltonian form of the two-point boundary value problem, with Hamiltonian

$$W^*(\mathrm{m}, \mathrm{s}) + \mathrm{m} \cdot \hat{\mathrm{u}}(\mathrm{s}) + \mathrm{n}_3,$$

where W^* is the conjugate function to W, for example pure quadratic with coefficient matrix K^{-1} if W is quadratic.

But most importantly the shape parameters $\hat{\mathrm{u}}(s)$ always enter the Hamiltonian linearly. The transform of a function of a shifted argument is a shift of the transform!

At this stage the continuum rod model is neither more nor less coarse-grained than the original rigid base-pair model–it all depends on the length scale on which the coefficient functions in the Hamiltonian vary.

Can quantitatively change the level of coarse graining within the continuum model by averaging or filtering the coefficients of the Hamiltonian (but you must average the Hamiltonian not the Lagrangian...)

In particular can always write the Hamiltonian as

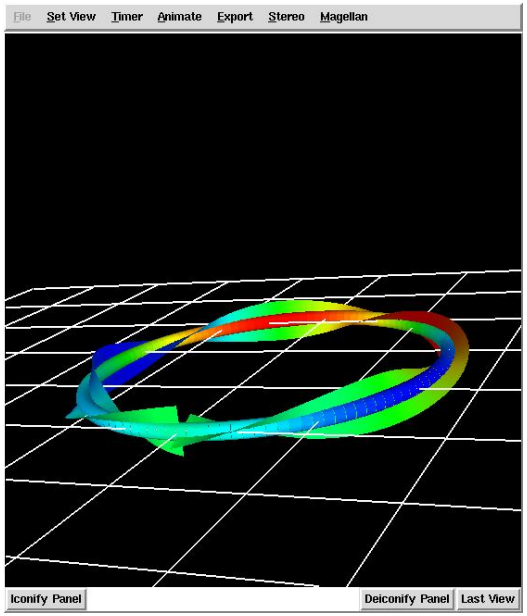
$$H_0 + \epsilon H_1$$

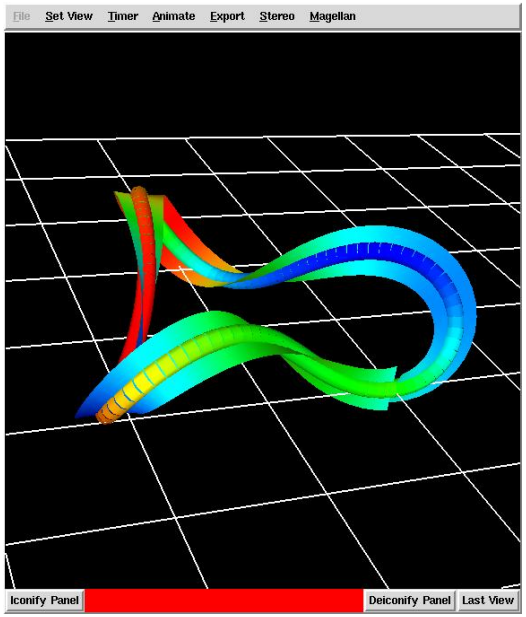
where H_0 is a Hamiltonian corresponding to a uniform and isotropic rod, and ϵ is a homotopy or continuation parameter that deforms to the actual Hamiltonian of interest in the model.

There are physically natural choices for the integrable H_0 that can be constructed from a multi-scale homogenization exploiting the high-intrinsic twist $\hat{u}_3$ as a large parameter.

Key idea: After passing to the continuum limit, first compute with H_0 being the appropriate ‘vanilla’ elastica rod model of Euler, Bernoulli and Kirchhoff (AKA as the WLC or TWLC)–i.e. uniform, with isotropy (register symmetry). K_{ij} diagonal and constant, û vanishing.

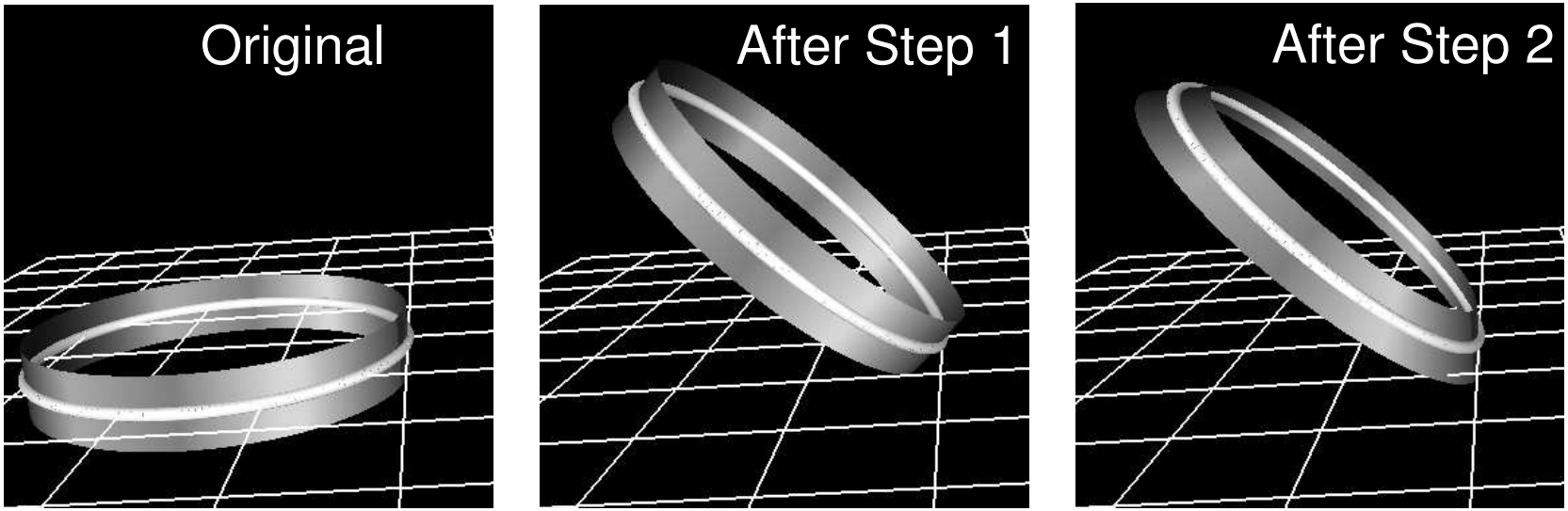
Good news: In the integrable case there are explicitly known solutions (twisted circles) and the entire bifurcation diagram is connected. Can therefore easily march anywhere.





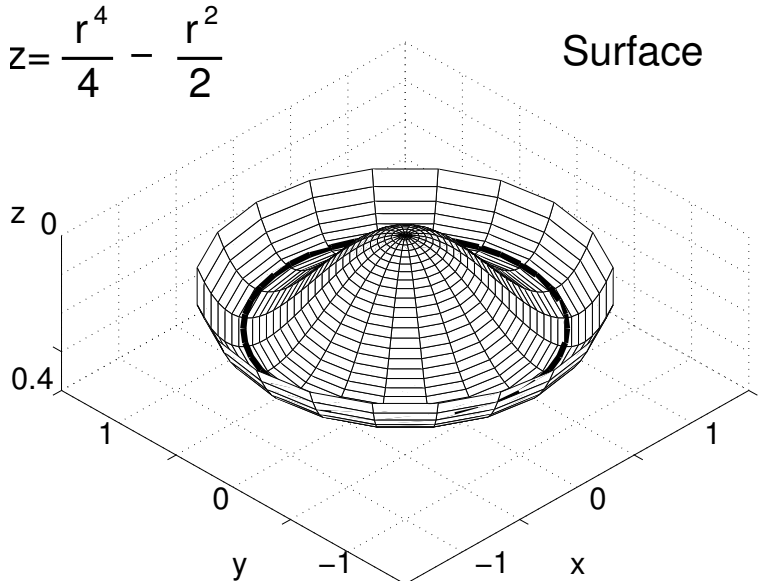
Bad news: The symmetries that imply complete integrability also imply that all solutions of the boundary value problem arise as parts of non-isolated families– circles of twisted circles, torii of non-planar equilibria. Makes the numerical computations a little delicate.

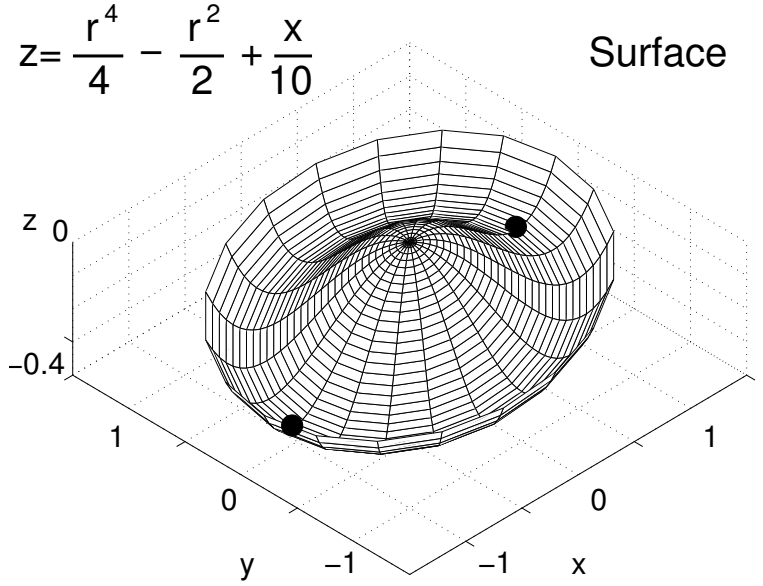
For example the non-isolation associated with register is illustrated by..

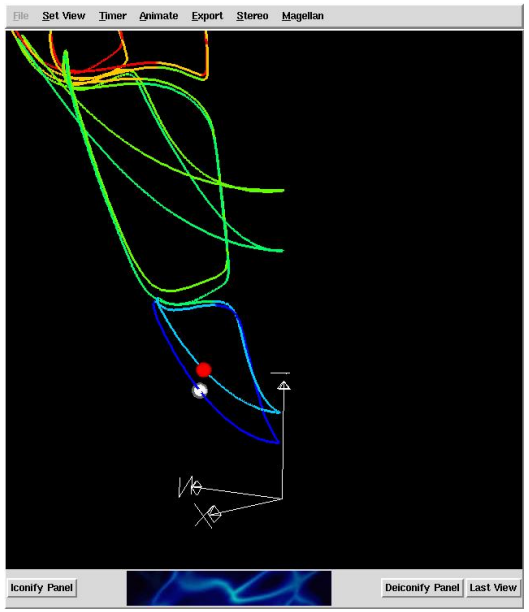


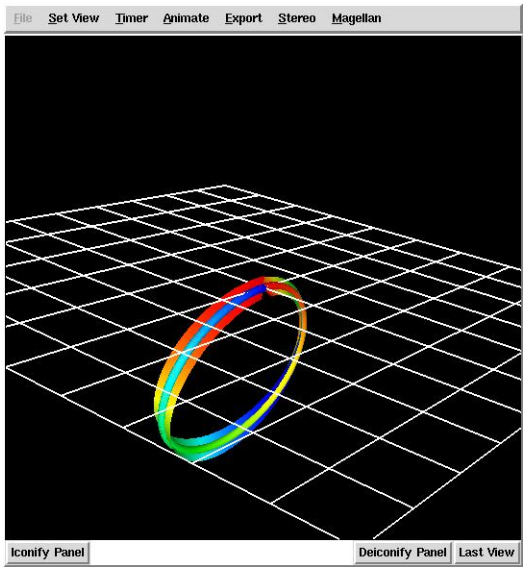
Nevertheless can use (Melnikov or Poincare function) analysis to understand how to break symmetries by the addition of small sequence dependent intrinsic curvatures, and small sequence dependent variations in stiffnesses.

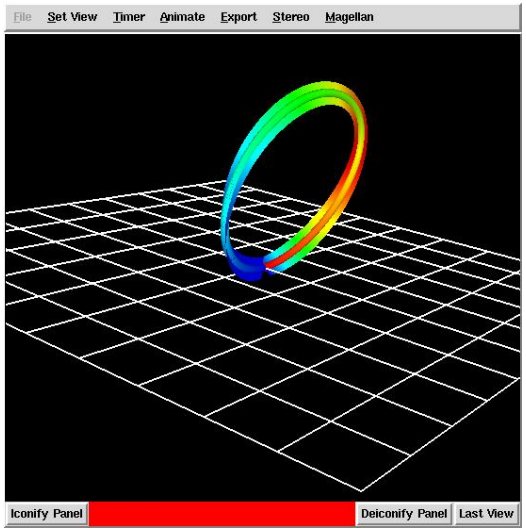
Evaluate the perturbation H_1 of the Hamiltonian along the symmetries of the families of equilibria, and the equilibria that persist are precisely the stationary points of the perturbation along the symmetry orbit.



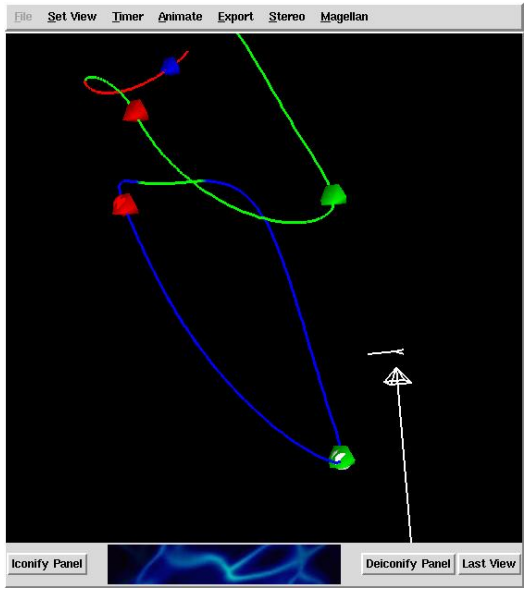


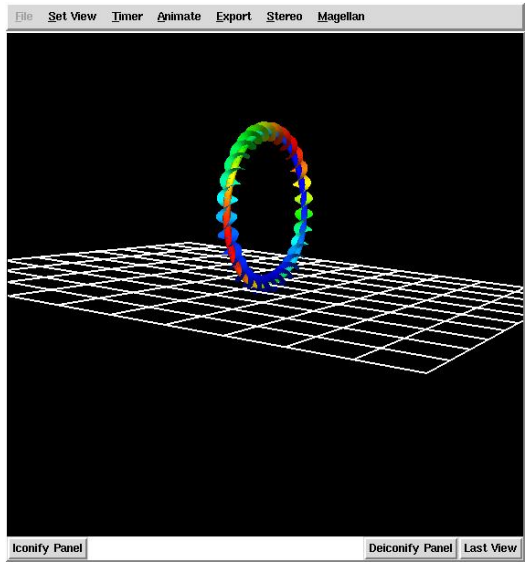

File Set View Timer Animate Export Stereo Magellan
Iconify Panel
Deiconify Panel
Last View

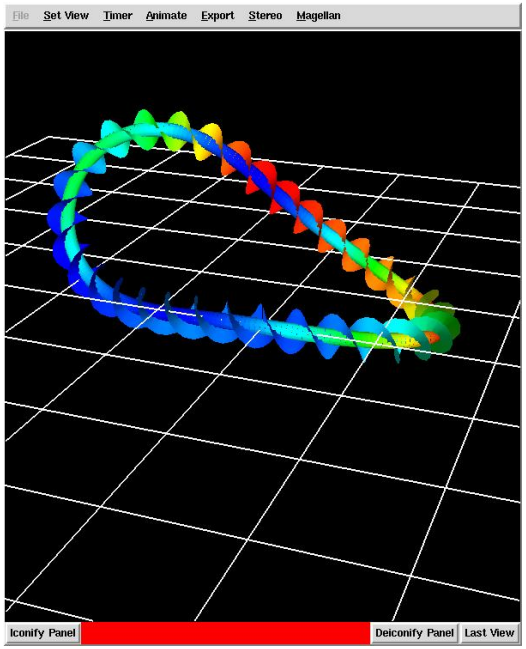

File Set View Timer Animate Export Stereo Magellan
Iconify Panel
Deiconify Panel
Last View


File Set View Timer Animate Export Stereo Magellan
Iconify Panel
Deiconify Panel
Last View

Finally numerical homotopy to whatever are the finite size sequence dependent effects you believe are in $U(q)$.







Given $U(q)$ this is all a doddle, and can be entirely automated, for example 6K bifurcation diagrams were calculated for various sequences.

Moreover can use the symmetry breaking analysis to design (given an energy model) anomolous sequences that should display critical behaviours and multi-peaked distributions of various things.

Perhaps most importantly there are no significant errors anywhere between the discrete rigid base-pair model and the continuous model.

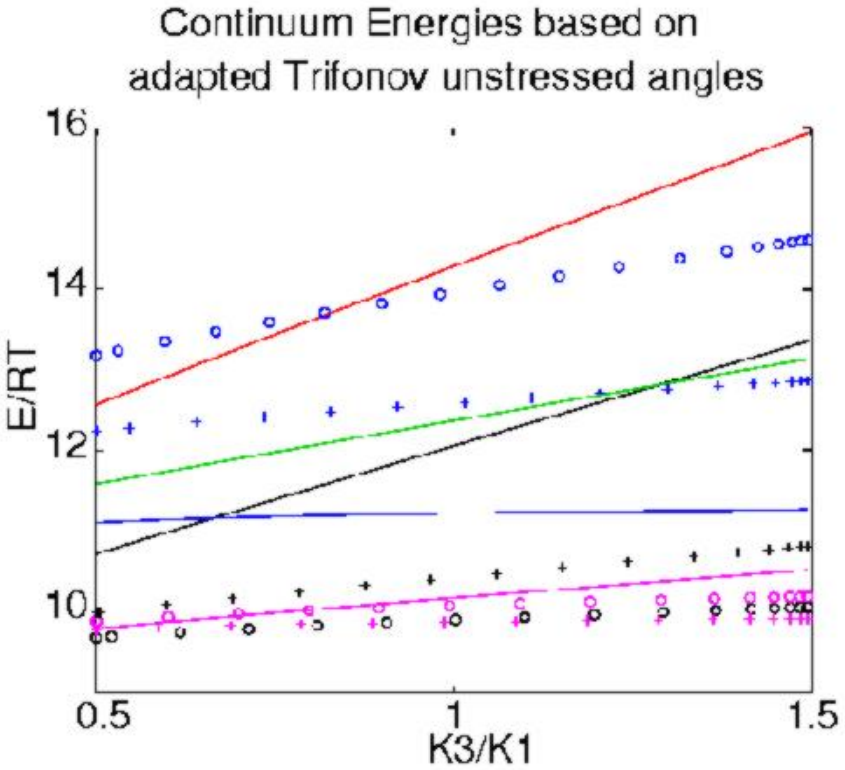
Continuum and Discrete Comparison

Convert continuum equilibrium back to discrete shape and use it as initial guess for discrete equilibrium computation—compare continuum and discrete energies and shapes.

Name	E_{disc}	E_{cont}	% Diff.	RMS config. diff. (Ang.)
12A09	0.10746	0.10740	0.05	0.59
09T09	0.11701	0.11714	0.11	0.64
13A09	0.09469	0.09457	0.13	0.58
17A11	0.09682	0.09654	0.29	0.63
11A17	0.09860	0.09817	0.43	0.58
15A09	0.08305	0.08280	0.30	0.58
11T15	0.07732	0.07730	0.03	0.61
09A17	0.07976	0.07947	0.36	0.58
08A17	0.07414	0.07390	0.32	0.59
08T15	0.07501	0.07487	0.22	0.61
09T15	0.07292	0.07284	0.11	0.60

Errors between the models and experiment are another story!

Can predict ordering of good versus bad cyclizers in agreement with experiment, but highly dependent upon what rigid-base pair model you use.



DNA	Symbol	$-\ln J = \Delta G/RT$
12A09	o	20.9
09T09	—	20.9
13A09	+	20.1
17A11	—	19.1
11A17	—	17.8
15A09	—	17.2
11T15	—	16.2
09A17	+	15.7
08A17	o	15.0
08T15	o	14.8
09T15	+	14.7

Conclusions of Part 1:

- Continuum computations are just as accurate as Rigid-XX models
- Continuum formulations allow analysis and understanding that can used to design interesting experiments
- Continuum models are not intrinsically very coarse-grained. The length scale of the variation in the coefficients sets the level of coarse graining. And within a continuum model can vary the level of coarse-graining.
- To get more quantitative comparison with experiment need better constitutive relations to be passed up from the rigid-XX model!

The Passage from Atomistic to Rigid-XX

Problem at hand is therefore how to determine a sufficiently good sequence-dependent Rigid-XX potential function $U(q)$?

Note: Because DNA is so long and thin, while finite geometry effects such as ring closure conditions must be treated as nonlinear, it is nevertheless plausible that to a good approximation the potential $U(q)$ may be treated as quadratic (at least in some circumstances e.g. no kinks..).

Idea:

- generate atomistic-level time-series via a Molecular Dynamics simulation. Expensive, but in principle only need to do it a few (?) times, and the DNA oligomer need not be so long.

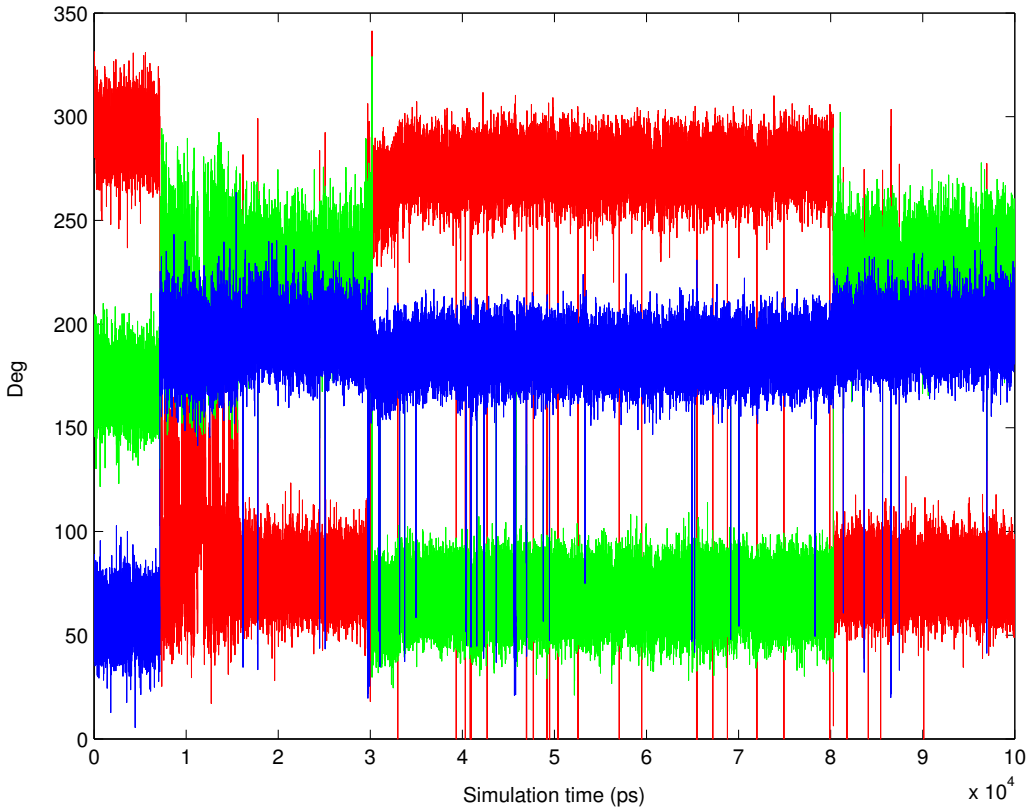
- reduce to a time-series of Rigid-XX variables q characterizing the configuration of the oligomer by fitting the coarse grain variables at constant t snapshots to the atomistic coordinates. (In particular the solvent is gone.)

- compute “accurate” coefficients for an assumed Rigid-XX quadratic potential $U(q) = (q - \hat{q}) \cdot K(q - \hat{q})$ from appropriate averages along the coarse-grain time-series

The general idea of making a comprehensive study of DNA properties as a function of its sequence, once and for all(?) lead to the Ascona B-DNA Consortium or ABC. Initially ten groups divided up the work, and shared all the simulation data for a set of entirely compatible 15 ns simulations of 39 15-mers that contain each of all possible 136 independent sequence tetramers at least twice (and away from the ends).

And of course there were many surprises. For example....

Extensions of some of the ABC MD runs showed previously unobserved cis-trans flips in various of the back-bone angles with residency times of tens of nanoseconds. (Here alpha, beta, gamma at step 11 of a poly-AT 15mer.)

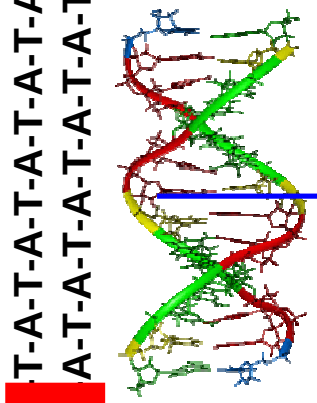


alpha-gamma flips appear all the time in crystal structures but not clear if they are well described in the Amber MD potentials.

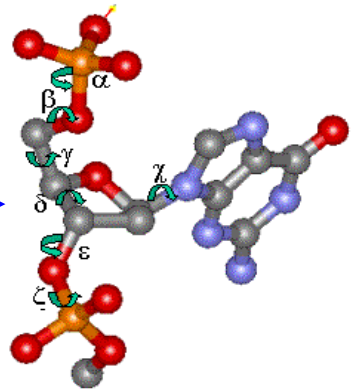
Observation of the base-pair parameters reveals that they are apparently slaved to the backbone angles. After a sharp transition or flip in the backbone angles the inter-base-pair parameters slowly evolve apparently to a new potential well.

But the backbone flips are relatively rare so can try to make a quadratic fit in each well.

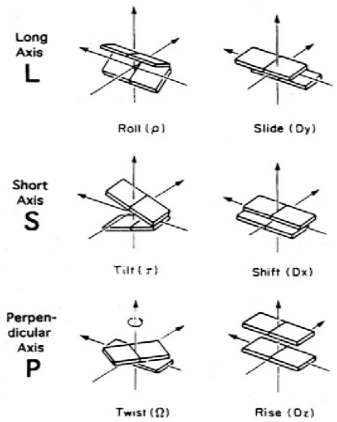
Analysis of MD-data for DNA-segment



part of backbone

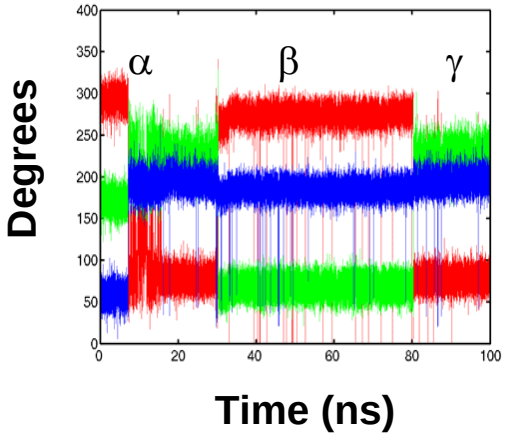


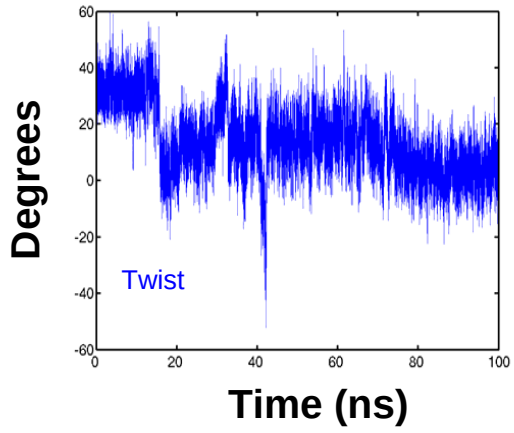
base-pair coordinates



Gonzalez and J.H.Maddocks:
Extracting parameters for base-pair level models of DNA from molecular dynamics simulations
Theor.Chem.Acc. 106:76-82, 2001

Step 9 (AT), 2nd strand





Basic idea is that for a quadratic potential $U(q) = (q - \hat{q}) \cdot K(q - \hat{q})$

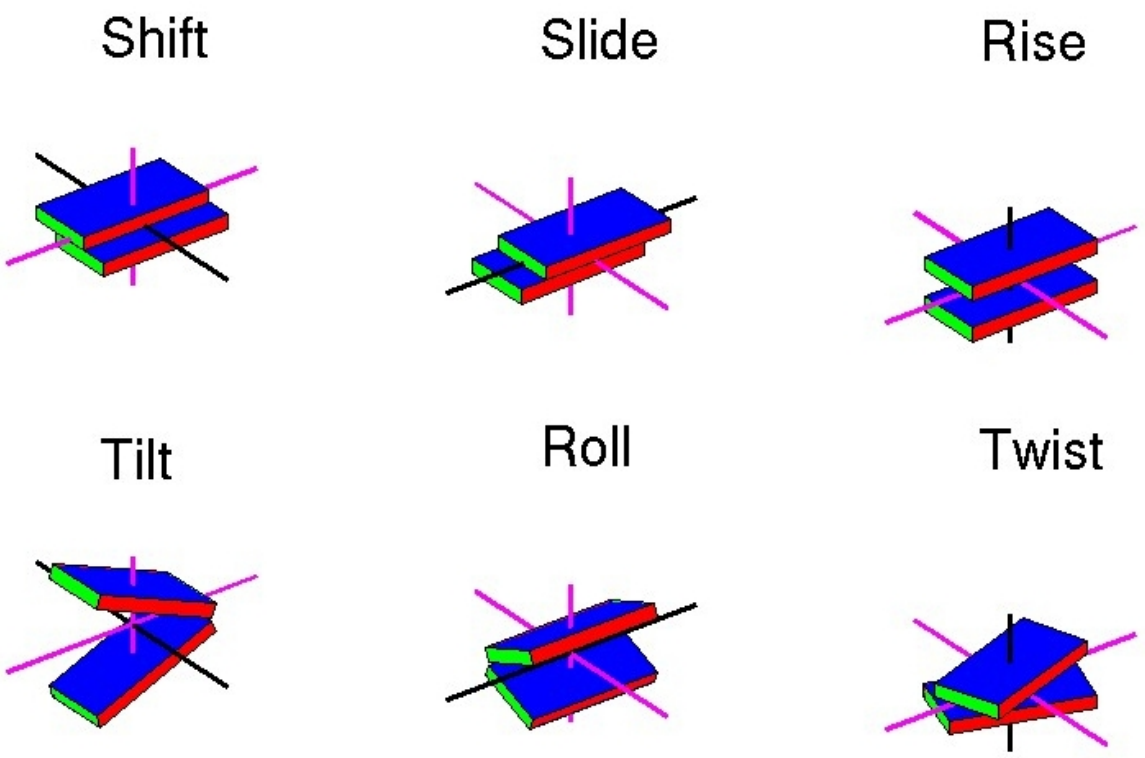
$$\left\langle \frac{q}{J(q)} \right\rangle = \hat{q} \quad \text{and} \quad \left\langle \frac{q_i\, q_j}{J(q)} \right\rangle = K^{-1}_{ij}$$

where $\langle \cdot \rangle$ denotes expectation according to the measure

$$\frac{1}{Z} \exp[-\beta U(q)] J(q).$$

Then one replaces the configuration space average over a (hopefully long enough) time series that (hopefully) stays in, but explores all of, the quadratic potential well.

Inter-Base Pair Deformations

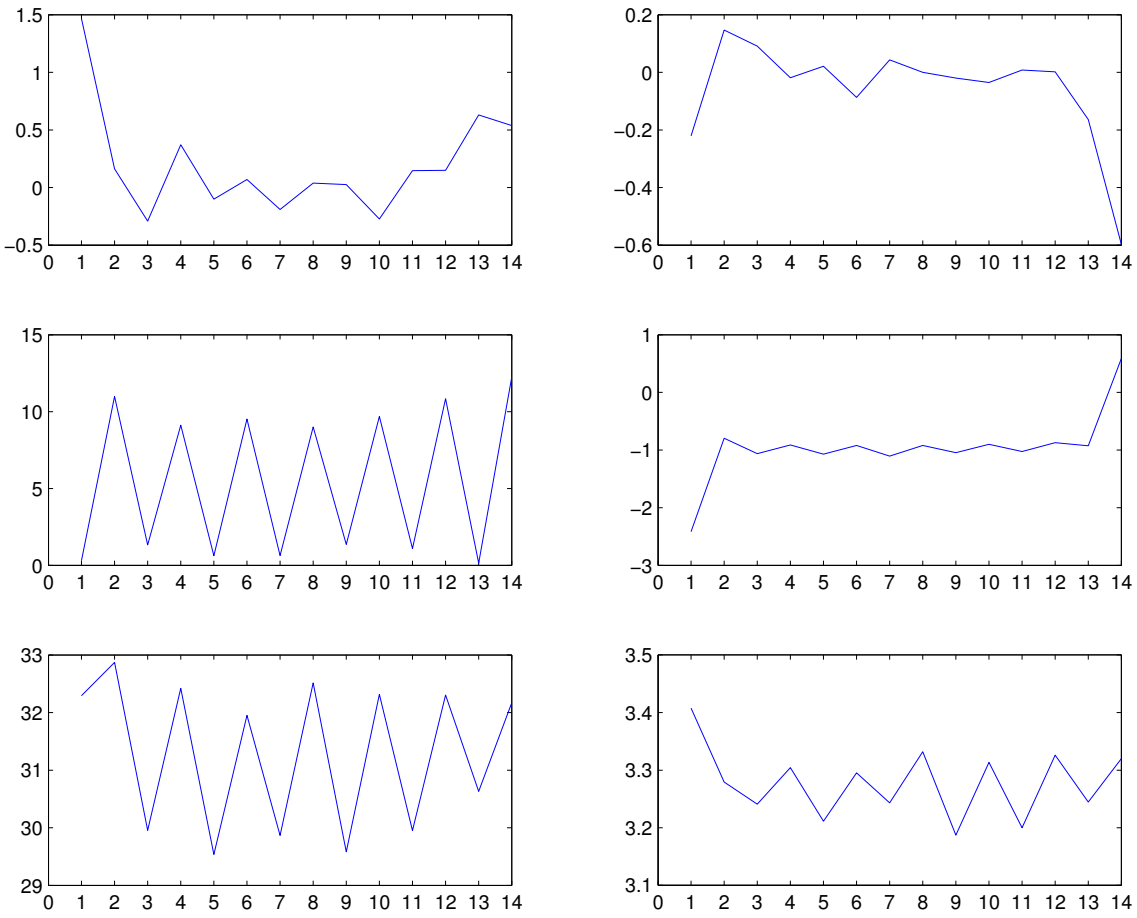


For example with q chosen to correspond to a rigid base-pair model with variables for an oligomer with N junctions we are lead to $\hat{q}$ being a $6N$ vector of averages and a $6N \times 6N$ covariance matrix with inverse K the stiffness matrix.

It is important to note that nothing need be assumed *a priori* about bandedness or sparsity of the stiffness matrix K–every rigid body could in principle be coupled to every other one, and what the couplings actually are is one of the main points of interest.

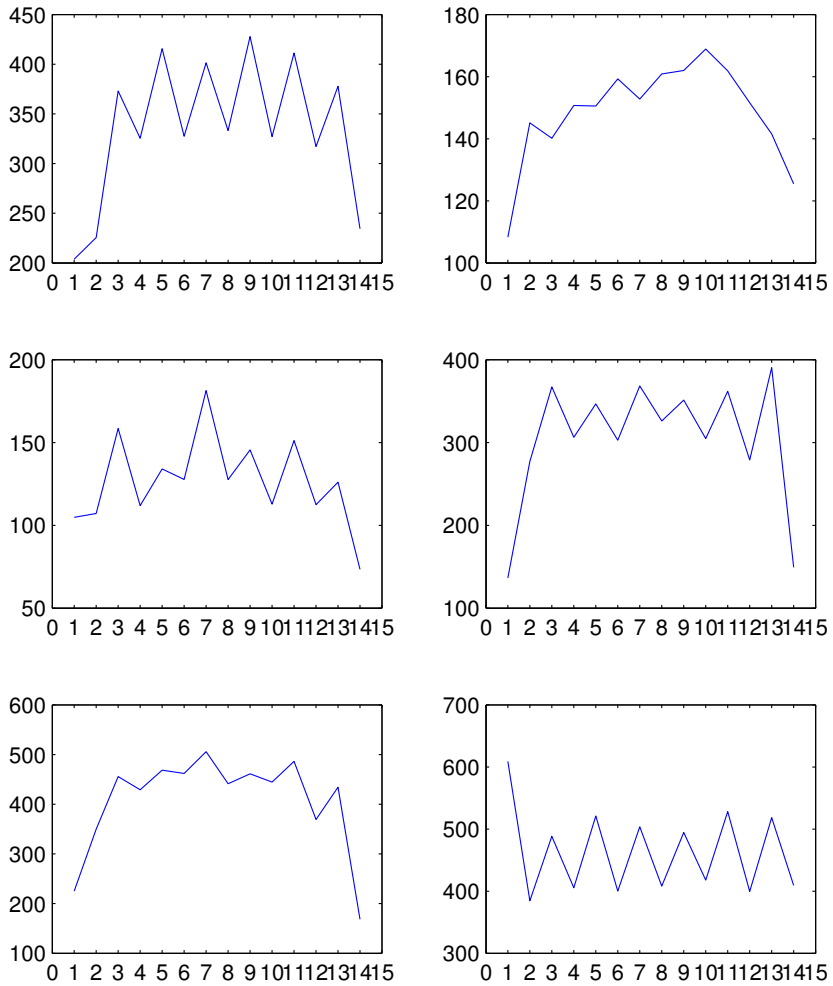
Of course we can do no better than the accuracy of the MD simulation and its potentials (in this case Amber), although the extraction techniques would not change with changed potentials.

For the first seven nanoseconds, i.e. within one well of the backbone, of the poly AT sequence with 14 junctions, the average strains $\hat{q}$ look plausible...



(tilt, roll, twist on left, shift, slide, rise on right)

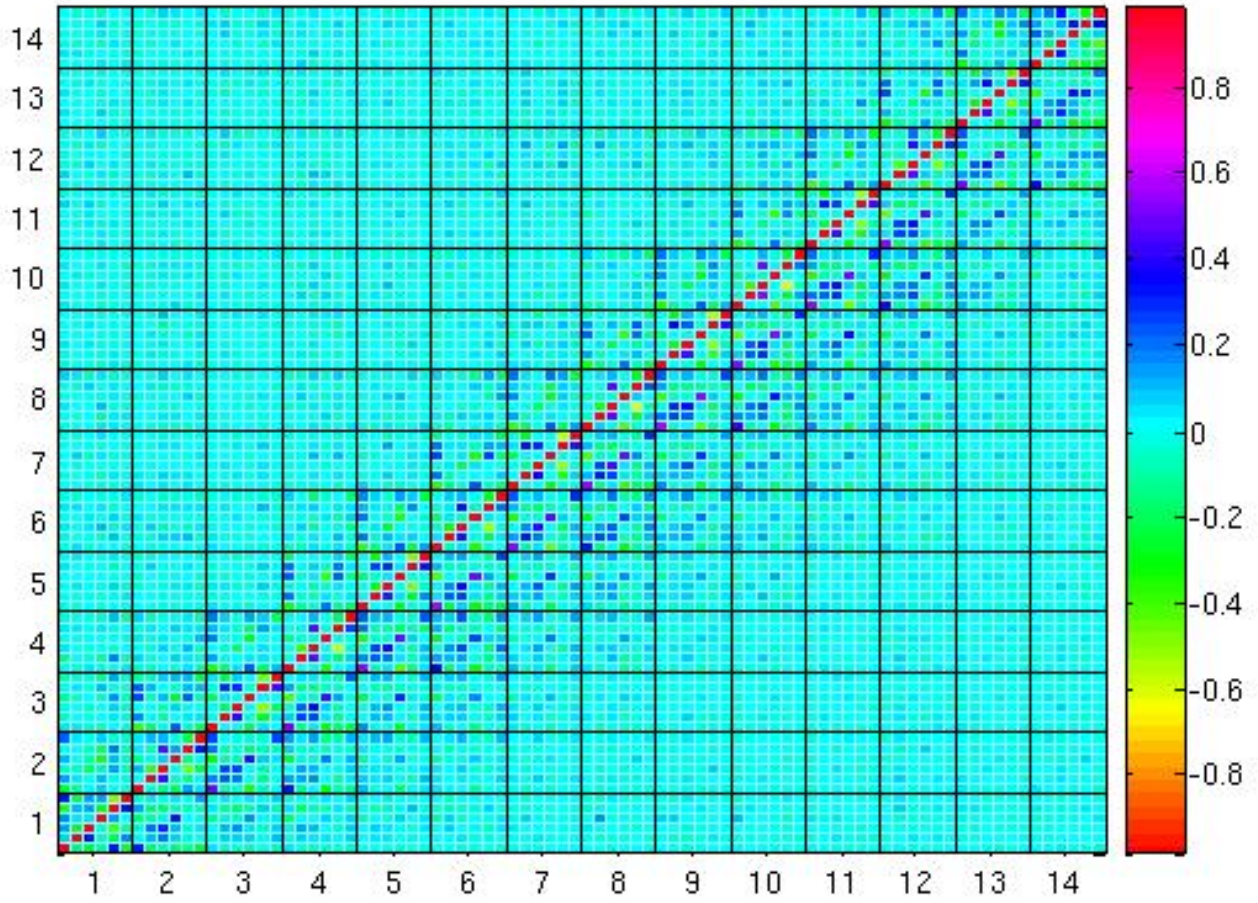
As do the diagonal entries of the covariance matrix



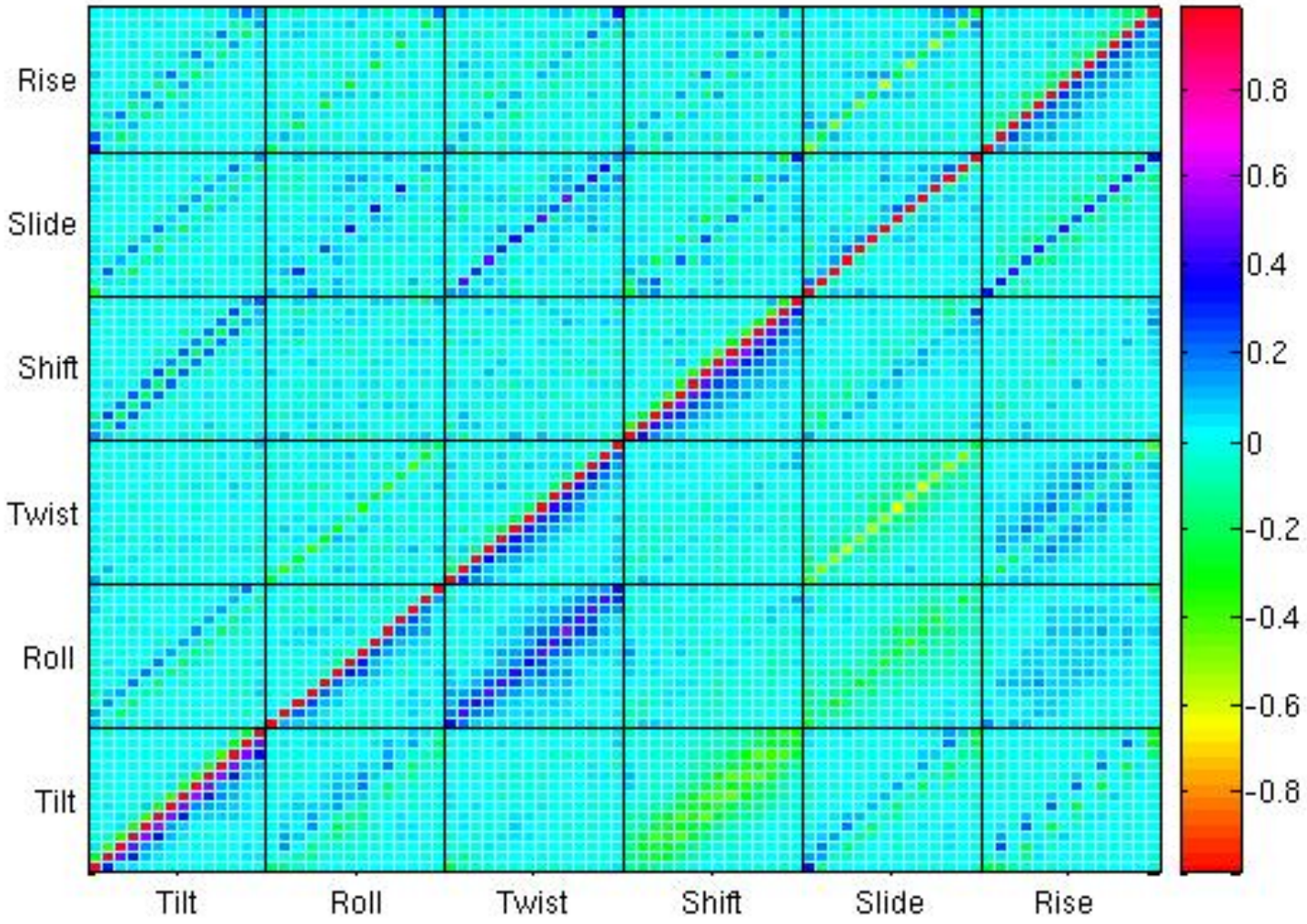
(tilt, roll, twist on left, shift, slide, rise on right)

But the sparsity pattern of the matrices exhibited surprises....

Top left the correlation matrix (i.e. covariance with diagonals scaled (non-dimensionalized) to 1, variables grouped by six, junction by junction for 14 junctions), bottom right its inverse, ie the stiffness matrix

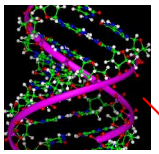


The eye sees different patterns when each of the six junction parameters is grouped. Top left covariance (diagonals scaled to 1, parameter by parameter), bottom right stiffness



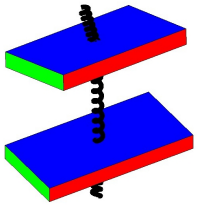
Conclusion is that in a rigid base-pair model the discrete stiffness matrix is at best tri-diagonal, and even this is not very convincing. In particular the stiffness matrix is less-banded than the covariances. Not so plausible, and not a finite difference stencil of a ‘standard’ elastic rod model.

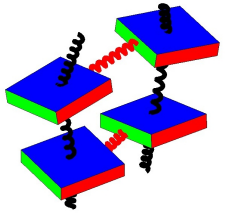
Message: Sequence-Dependent Coarse-graining at scales of tens of base-pairs is NOT self-consistent down the left column!



Chain of Rigid Base-Pairs

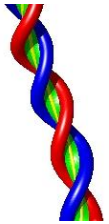
Chain of Rigid Bases





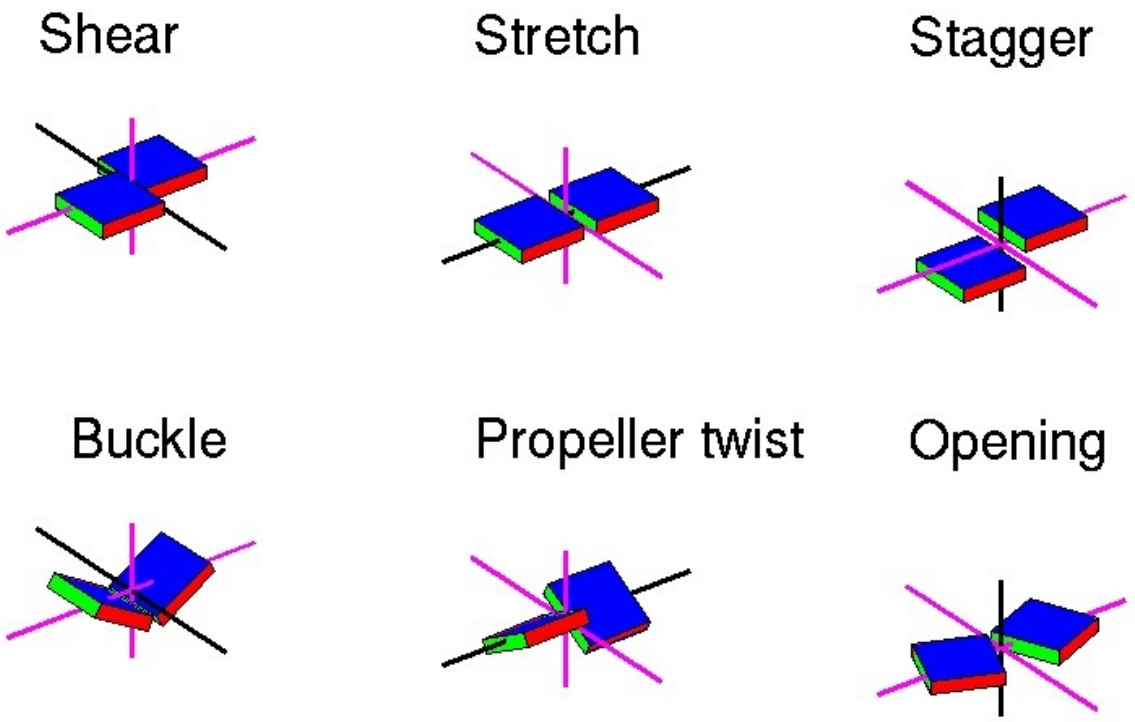
Elastic Rod

Elastic Birod



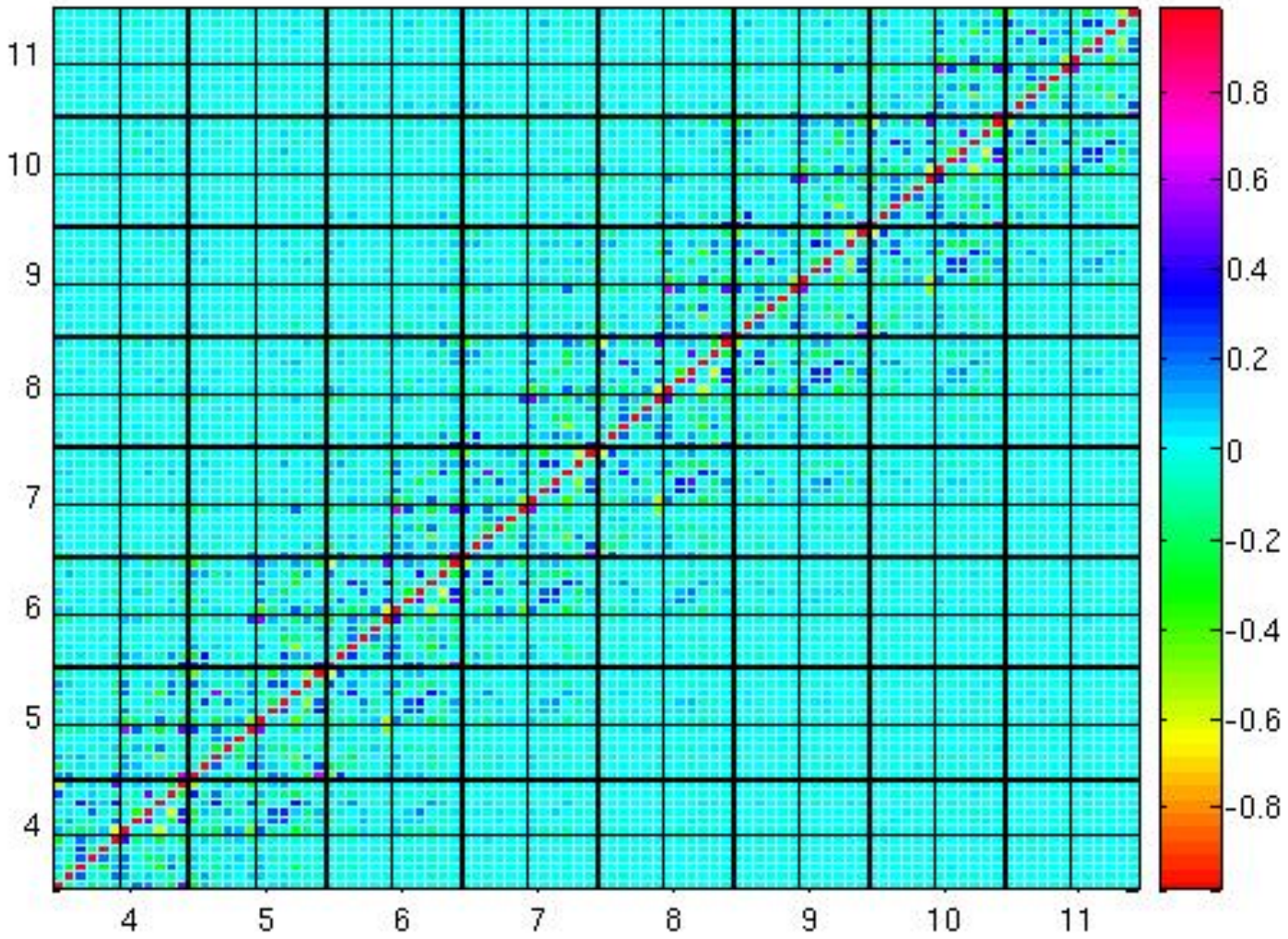
Alternatively if the configuration q is chosen to correspond to a Rigid Base model with Inter-Base pair variables plus...

Intra-Base Pair Deformations

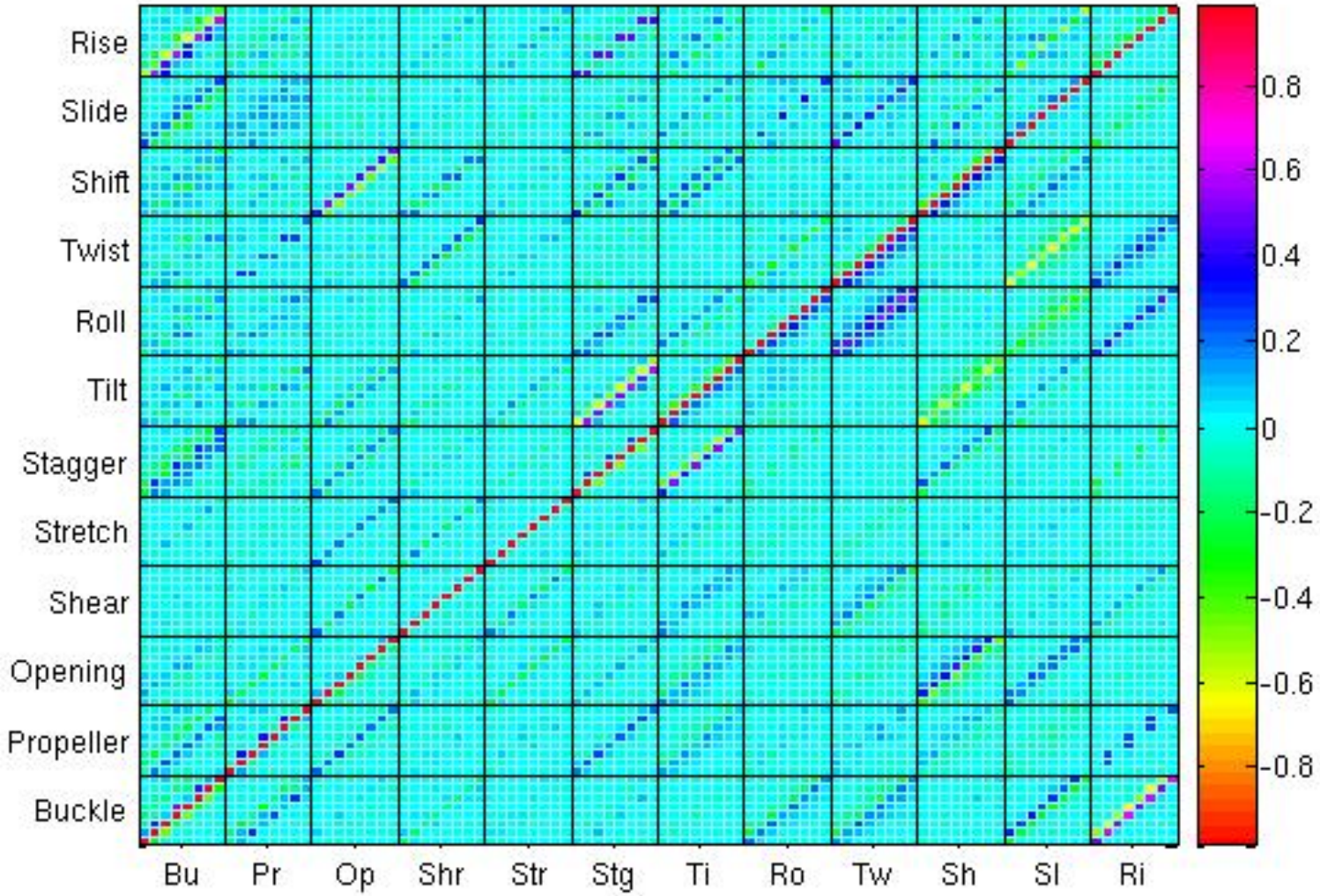


then for an oligomer with N junctions we are lead to a $12N \times 12N$ covariance matrix and its inverse, here submatrices for an alternating AT sequence with 14 junctions....

Top left covariance (diagonals scaled to 1, ordered base by base, submatrix for bases 4 through 11), bottom right stiffness



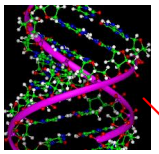
and when re-ordered parameter by parameter get....



The sparsity pattern is very close to a nearest neighbour model in which each base interacts with and only with its five nearest neighbours.

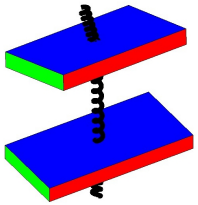
The stencil is also a natural finite difference approximation to an *elastic birod*, which is a new continuum mechanics theory tailored to model double stranded DNA in the continuum limit (so that efficient numerics can be applied).

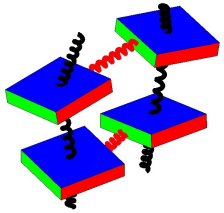
Message: Sequence-Dependent Coarse-graining at scales of tens of base-pairs is self-consistent down the right column!



Chain of Rigid Base-Pairs

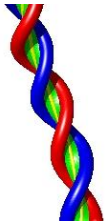
Chain of Rigid Bases





Elastic Rod

Elastic Birod



Conclusions Part 2:

- Coarse graining atomistic to rigid base-pair to elastic rod does not seem to be self-consistent at scales of tens of base pairs.

- Coarse graining atomistic to rigid base to elastic birod does seem to be self-consistent at scales of tens of base pairs, but we need to understand the new birod model, compute minicircle bifurcation diagrams, and to check comparison with experiment.

- Now seem set to extract sequence-dependent parameters for a nearest neighbour rigid-base/birod theory that coarse grains a particular MD potential.

- Naturally associated with a sequence-dependence of the model parameters on tetramers.

- Lots to do. For example need to understand back-bone flips to be able to identify quadratic wells.

- Further coarse-graining from birod to rod is in principle perfectly possible at longer length and time scales, so an elastic rod model can still be perfectly acceptable in addressing questions at longer scales.

- A quadratic energy rigid-base/birod theory offers promise of predicting the initiation of melting or unstacking via focussing effects associated with long length scale deformations.

- Range of validity of quadratic energy delicate. Seems to work quite well for 158 bp minicircles.

Work spans more than a decade, and ranges from mature results to current observations and efforts.

Many inter-connecting contributions from many people....

Articles available from http://lcvmwww.epfl.ch or google John Maddocks

Based on multiple collaborations:

- O. Gonzalez, G. Stoll, group of C. Schuette: Extracting coarse-grain constitutive relations from Molecular Dynamics data
- R. Lavery group, L. Heffler, F. Lankas: producing and visualizing Molecular Dynamics simulations
- J. Kahn, M. Samala, A. Amzallag: making DNA minicircles and measuring cyclization rates
- J. Dubochet, A. Stasiak, J. Bednar: taking stereo Cryo-EM minicircle images
- A. Amzallag, C. Vaillant, M. Jacob & M. Unser group: Minicircle reconstruction from Cryo images
- R. Manning, R. Paffenroth, K. Hoffman, P. Furrer, Y. Li: Computation and visualization of DNA minicircle bifurcation diagrams
- S. Kehrbaum, S. Rey: High-twist homogenization
- M. Moakher: Rigid-base and Elastic Birod models

And in passing looking at MD shows that rigid bases are a much better approximation than rigid base-pairs..

(miniTRAJ.mpg)