The Abdus Salam
International Centre for Theoretical Physics

United Nations
Educational, Scientific and
Cultural Organization

IAEA
International Atomic Energy Agency

2038-32

Conference: From DNA-Inspired Physics to Physics-Inspired Biology

1 - 5 June 2009

Multi-Scale Modelling of DNA: Parametrizations of Coarse-Grain Sequence-Dependent Models of DNA Mechanics

John H. MADDOCKS
Ecole Polytechnique Federale de Lausanne
Institut de Mathematiques B, 1015
Lausanne
SWITZERLAND

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

ICTP Trieste May 2005 !!!!!

Good general principle:

The answer should depend on the question...

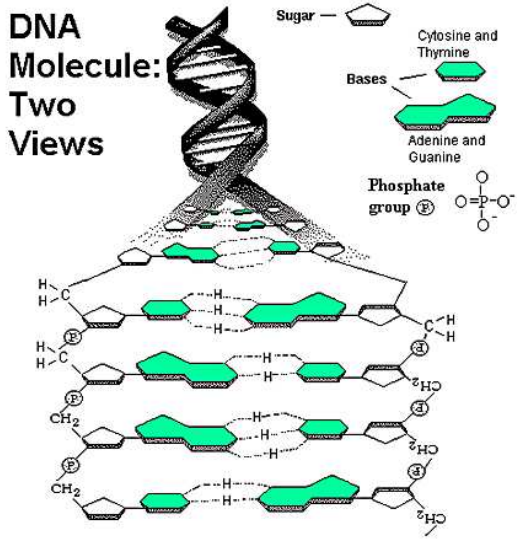
Thus to identify the appropriate level of coarse-graining in a model need to understand the scale of the experimental data that is to be modelled.

Today will discuss:

- fully-atomistic scales of a few bases of DNA
- Rigid-XX models at a few tens of bases in length
- and mention in passing continuum models appropriate at many tens or a few hundreds of bases in length on upward.

MOST IMPORTANT AND FUNDAMENTAL POINT

Because of the different stackings between purines and pyrimidines, and because of the different numbers of hydrogen bonds, the base pair sequence modulates both the intrinsic shape, and local stiffness properties of the double helix in a biologically significant way.



Today:

- Discuss ongoing, and unfinished, work on how to extract sequence-dependent constitutive relations from MD, at a few tens of bases,

Coarse-graining and Statistical Mechanics

Main point in coarse graining DNA is to eliminate an explicit treatment of the solvent. Then the dynamics becomes a stochastic Langevin system of the general form

$$\dot{q} = M^{-1}(q)p = H_p \qquad \dot{p} = -H_q - D(q)H_p + \Sigma(q)\dot{W}(t)$$

where the configuration variable q is the choice of level of coarse (or fine) graining in your model–e.g. atomistic, rigid-XX, or continuum, p is the associated conjugate momentum, $H(q,p) = \frac{1}{2}p\,M^{-1}(q)p + U(q)$ is the Hamiltonian, $U(q)$ is the 'potential' energy, $M(q)$ is the mass matrix, and the effects of the solvent are reduced to the (viscous) damping matrix $D(q)$ and stochastic forcing matrix $\Sigma(q)$ applied to the derivative of the white noise $W(t)$.

The explicit dependence on q is necessary when angle (or other non Cartesian) coordinates are present as is the case for Rigid-XX descriptions.

For atomistic or Rigid-XX configuration variables this is a (big) system of stochastic ordinary differential equations, and in the continuum case it is a (small) system of stochastic partial differential equations.

With an appropriate fluctuation-dissipation relation between $D(q)$ and $\Sigma(q)$ this Langevin system always has the stationary configuration space distribution

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

for a choice of configuration variable q, potential $U(q)$, Jacobian, or metric correction factor, $J(q)$, partition function Z, and temperature scale β.

In particular the Boltzmann distribution is *NOT* natural for DNA mechanics (although the variations in, and thus effects of, the additional term $J(q)$ may be quite small at some scales).

Will discuss:

- how to obtain an approximation of a sequence-dependent $U(q)$ appropriate for Rigid-XX levels of coarse graining from Molecular Dynamics simulations

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 in the double helix ...).

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 using a coarse-graining rule such as the Tsukuba convention. (In particular the solvent is gone.)

- compute "accurate" coefficients for an assumed Rigid-XX quadratic potential $U(q) = (q - \hat{q}) \cdot K(q - \hat{q})$ by equating appropriate averages along the coarse-grain time-series with analytic expressions involving the unknown coefficients.

Most simplistic idea is that for a quadratic potential $U(q) = (q - \hat{q}) \cdot K(q - \hat{q})/2$

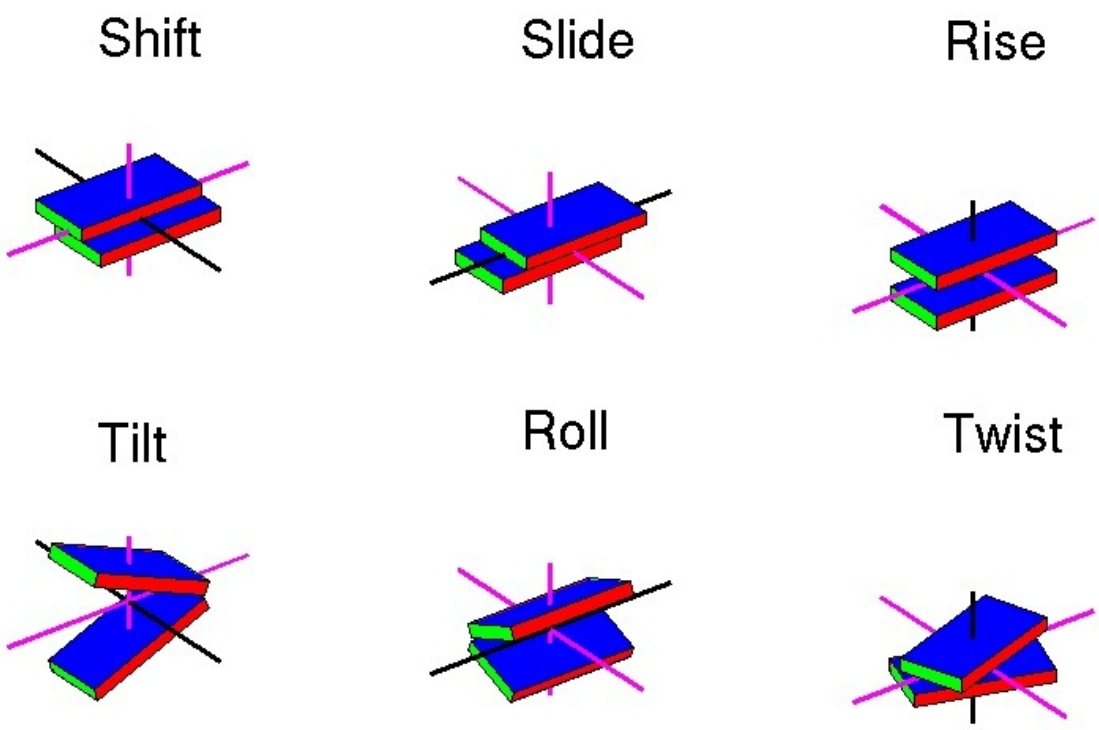
$$\left\langle \frac{q}{J(q)} \right\rangle = \hat{q} \quad \text{and} \quad \left\langle \frac{q_i\, q_j}{J(q)} \right\rangle = K^{-1}_{ij}$$

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

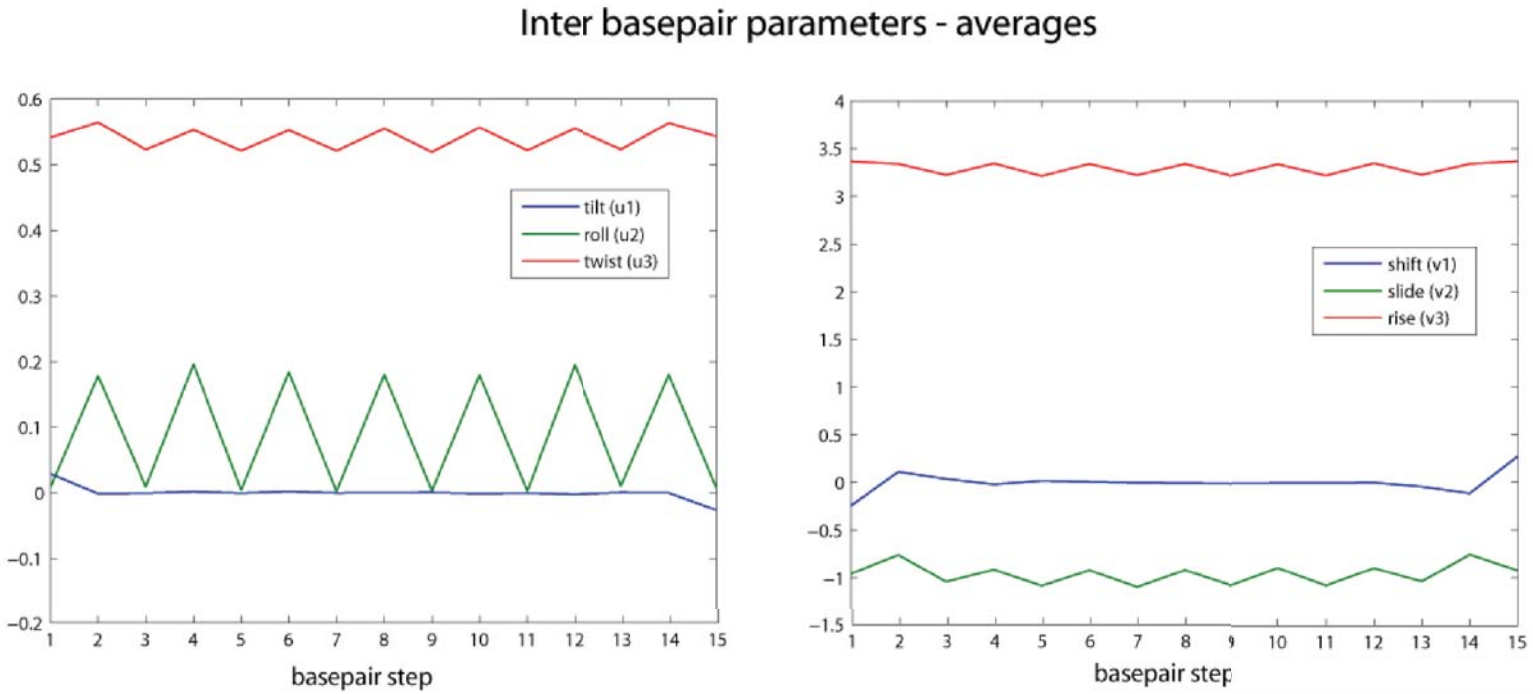


For example with q chosen to correspond to a rigid base-pair model with variables for an oligomer with N junctions we are lead to $\hat{q}$ being a $6N$ vector of averages and a $6N \times 6N$ covariance matrix with inverse K the stiffness matrix.

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.

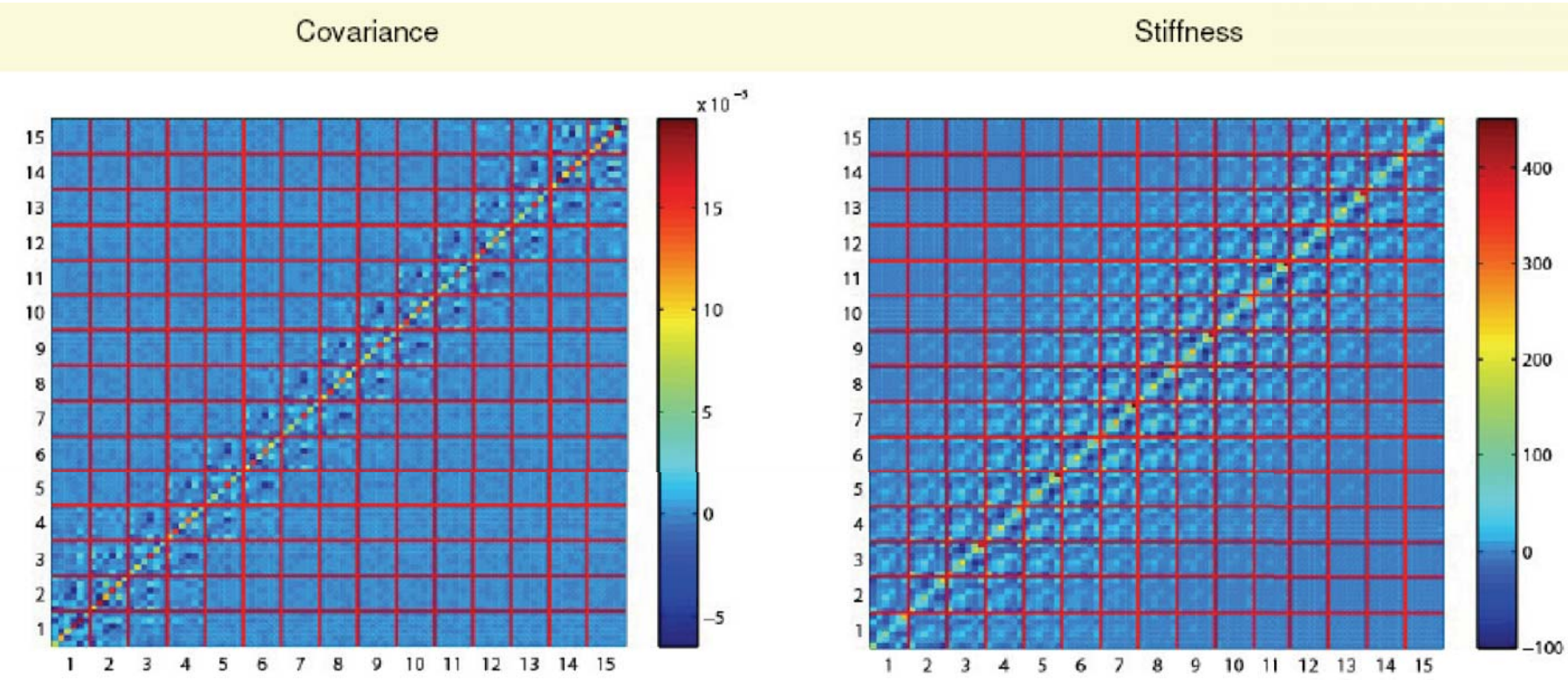
Of course in this approach we can do no better than the accuracy of the fine grained MD simulation and its potentials (in this case Amber), although the extraction techniques would not change with changed potentials.

For a 180 nanosecond simulation of a palindromic poly AT sequence (capped at ends with CG pairs and filtered to remove all alpha-gamma flips, and any configuration with a broken hydrogen bond anywhere in the oligomer, ie within one B-form well) the average strains $\hat{q}$ pass all necessary symmetry conditions



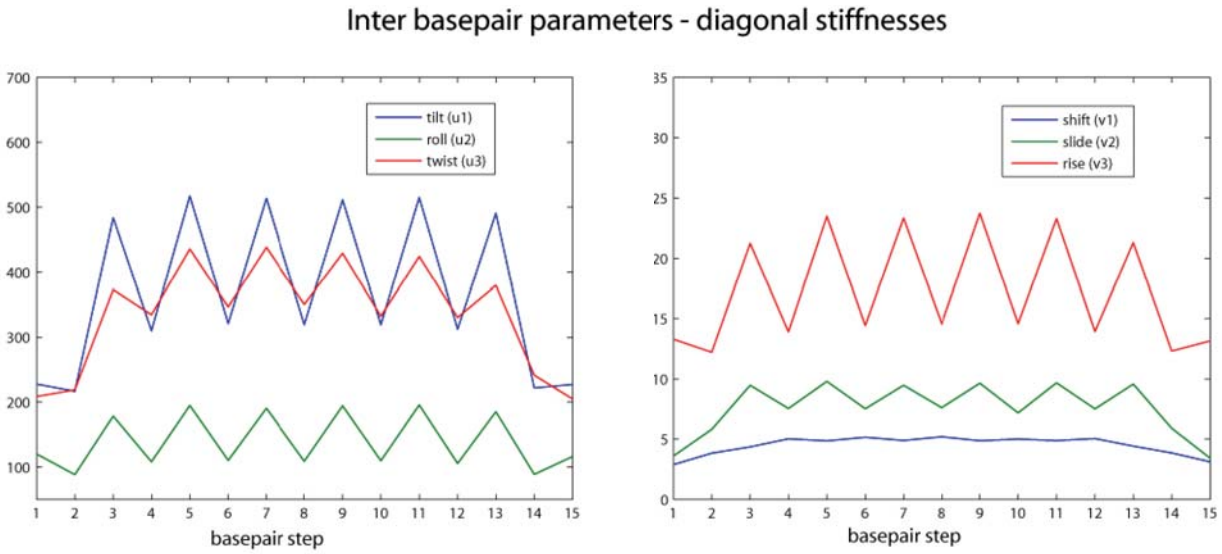
(tilt, roll, twist on left, shift, slide, rise on right)

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

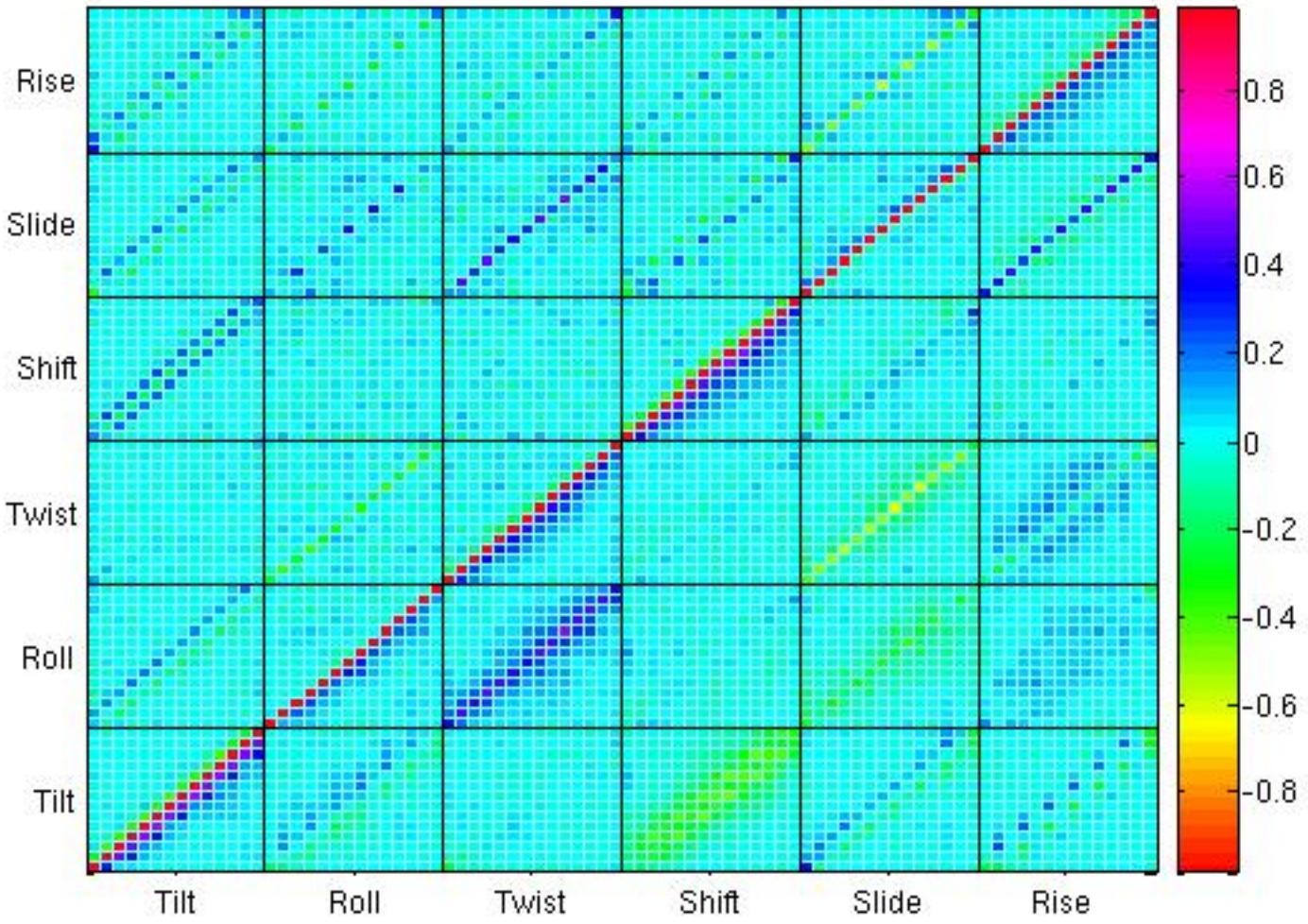


On left the rigid base pair covariance matrix, which is highly banded, and on the right its inverse, the stiffness matrix K, which is much less banded.

To focus on the sparsity pattern we can nondimensionalize and scale with the diagonal entries, and also reorder entries to group by each parameter along the sequence

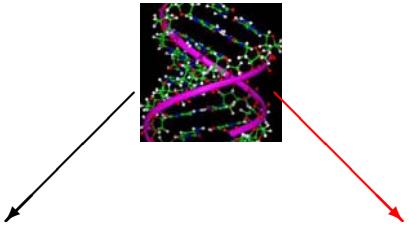


Top left covariance (diagonals scaled to 1, parameter by parameter), bottom right stiffness each of the six junction parameters grouped.



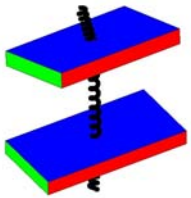
Conclusion is that in a rigid base-pair model the discrete stiffness matrix is at best tri-diagonal, and even this is not very convincing. In particular the stiffness matrix is less-banded than the covariances. Not so plausible, and not a finite difference stencil of a ‘standard’ elastic rod model.

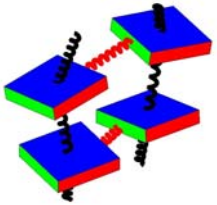
Message: Sequence-Dependent Coarse-graining at scales of tens of base-pairs is NOT self-consistent down the left column!



Chain of Rigid Base-Pairs

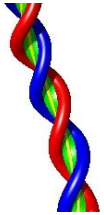
Chain of Rigid Bases





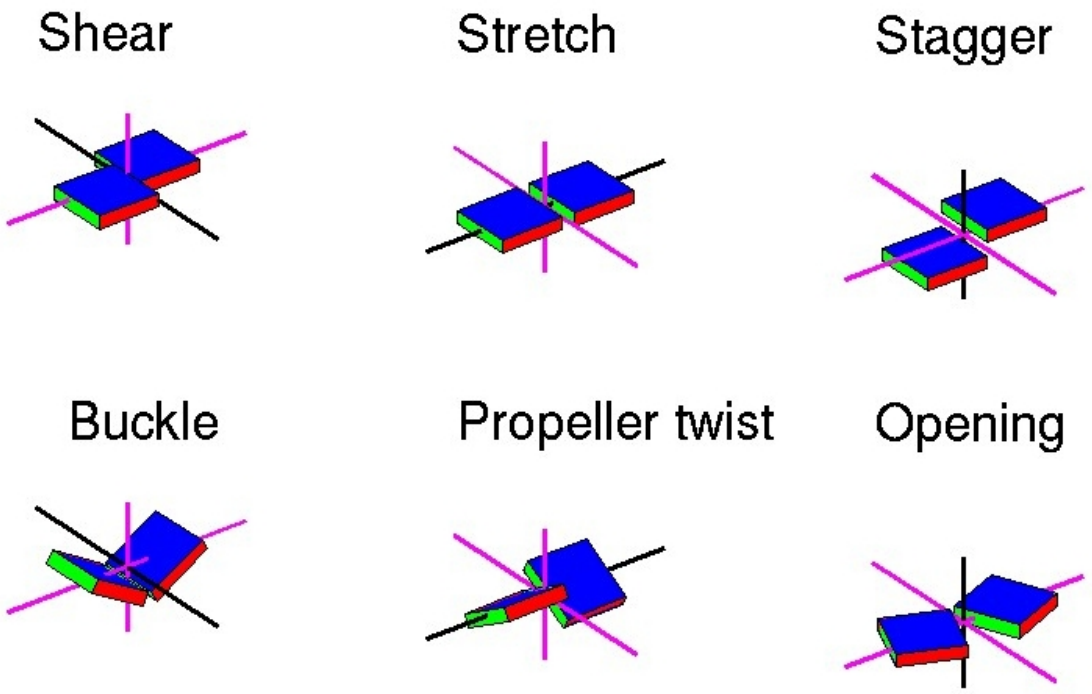
Elastic Rod

Elastic Birod

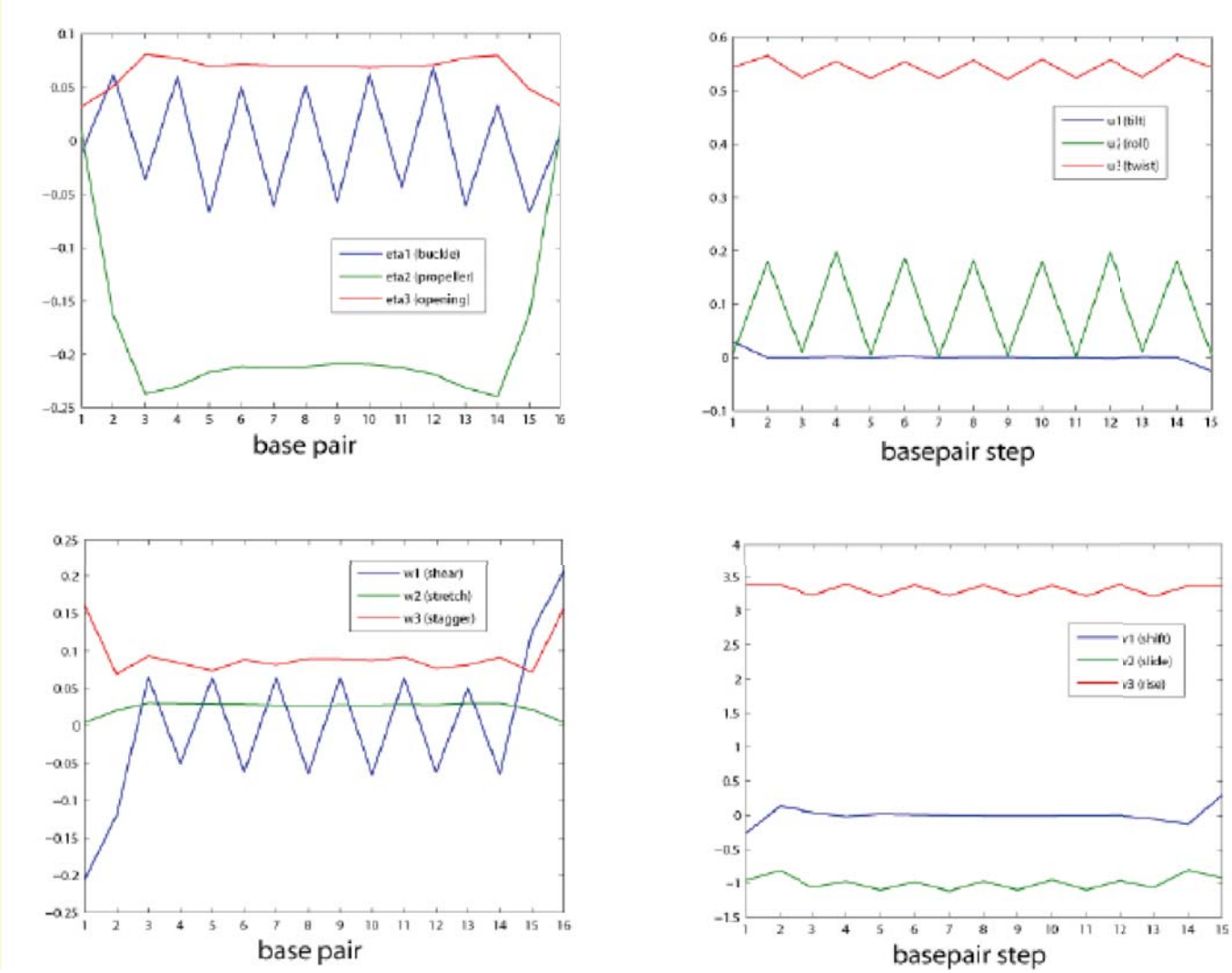


Alternatively if the configuration q is chosen to correspond to a Rigid Base model with Inter-Base pair variables plus...

Intra-Base Pair Deformations

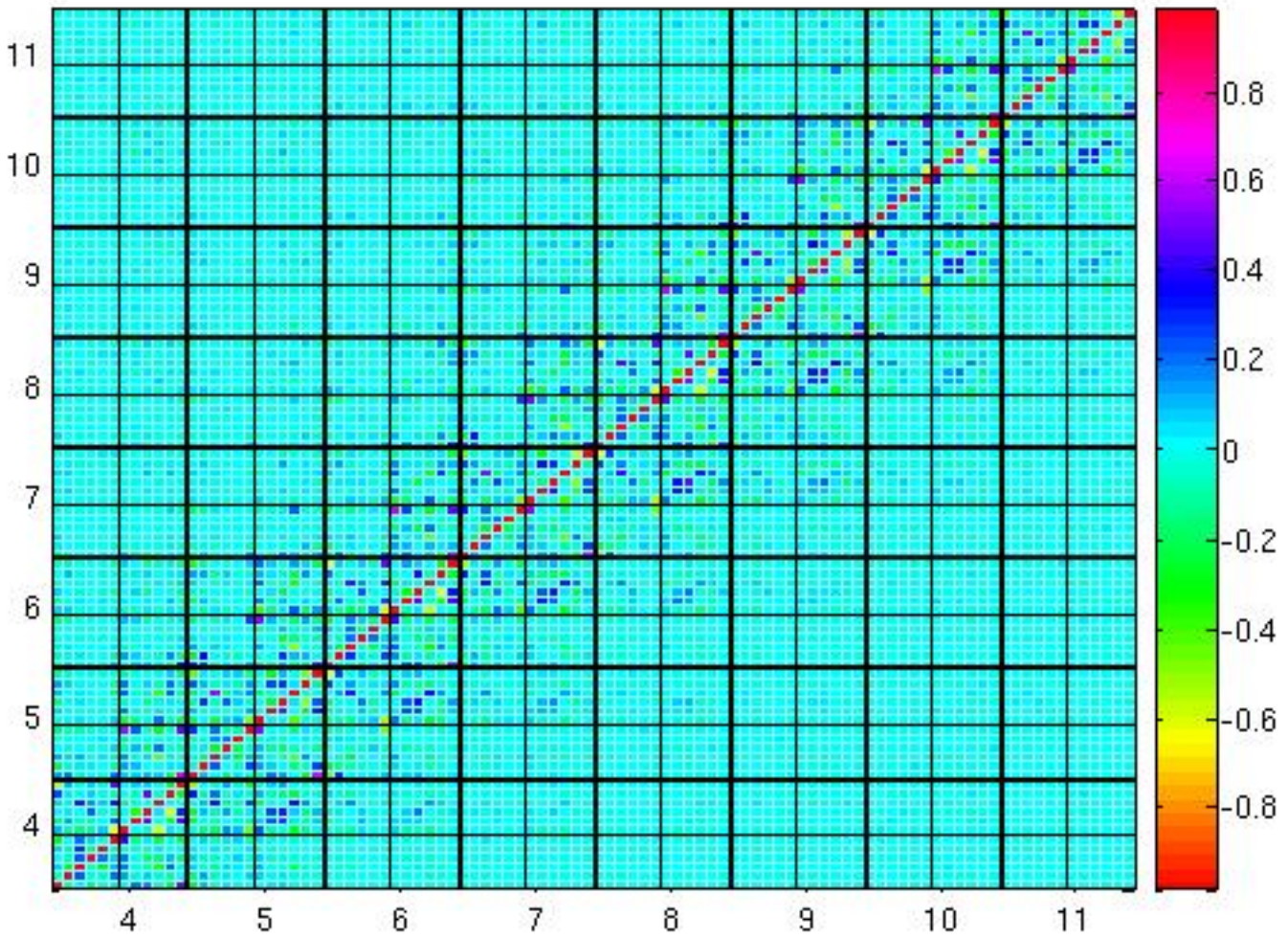


then for an oligomer with N junctions we are lead to a $12N \times 12N$ covariance matrix and its inverse, here submatrices for an alternating AT sequence with 14 junctions....

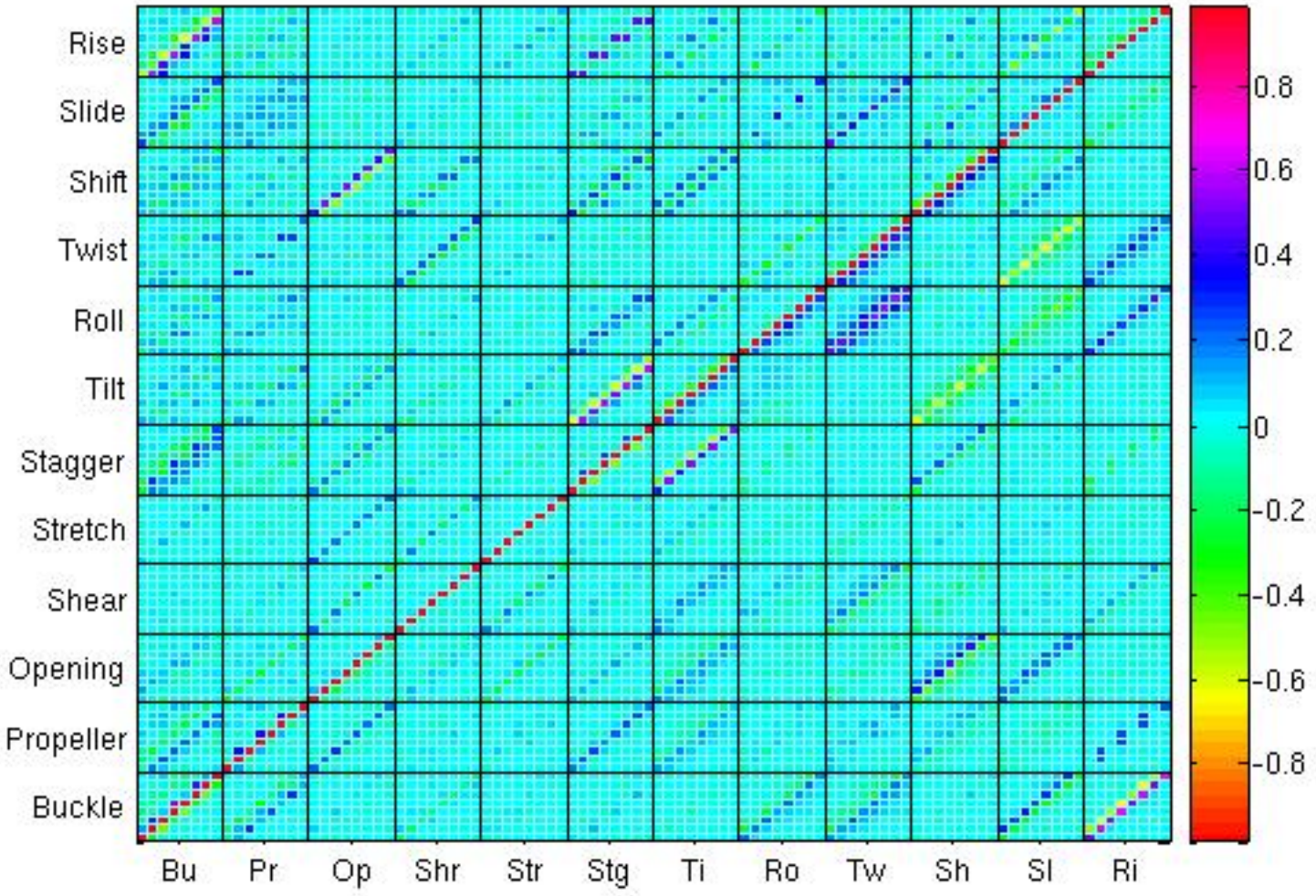


On the left averages of the six intra base-pair variables for each of the 16 base-pairs in a $G(AT)_7C$ oligomer, and on the right the six inter-base-pair parameters for each of the 15 junctions. Rotations above, translations below.

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



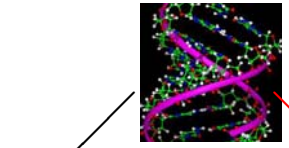
and when re-ordered parameter by parameter get....



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

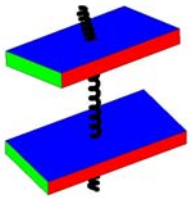
The stencil is also a natural finite difference approximation to an *elastic birod*, which is a new continuum mechanics theory tailored to model double stranded DNA in the continuum limit (so that efficient numerics can be applied).

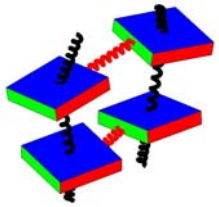
Message: Sequence-Dependent Coarse-graining at scales of tens of base-pairs is self-consistent down the right column!



Chain of Rigid Base-Pairs

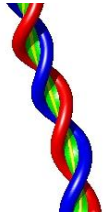
Chain of Rigid Bases





Elastic Rod

Elastic Birod



Conclusions thus far:

- Coarse graining MD atomistic simulations of a poly (AT) oligomer to a rigid base-pair model/elastic rod model is not self-consistent at scales of tens of base pairs.

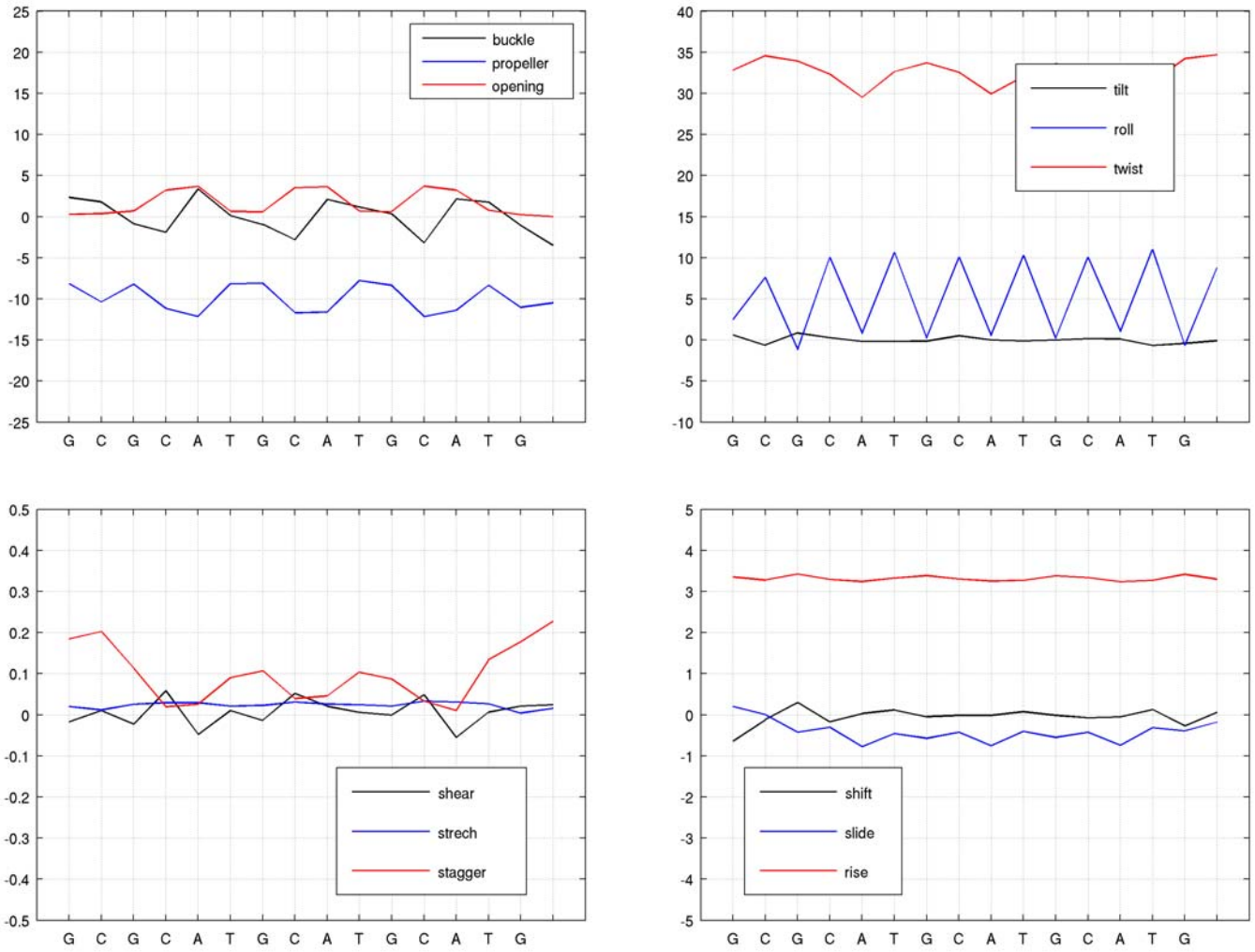
- Coarse graining MD atomistic simulations of a poly (AT) oligomer to a rigid base/elastic birod model does seem to be self-consistent at scales of tens of base pairs.

Perhaps the issue is related to:

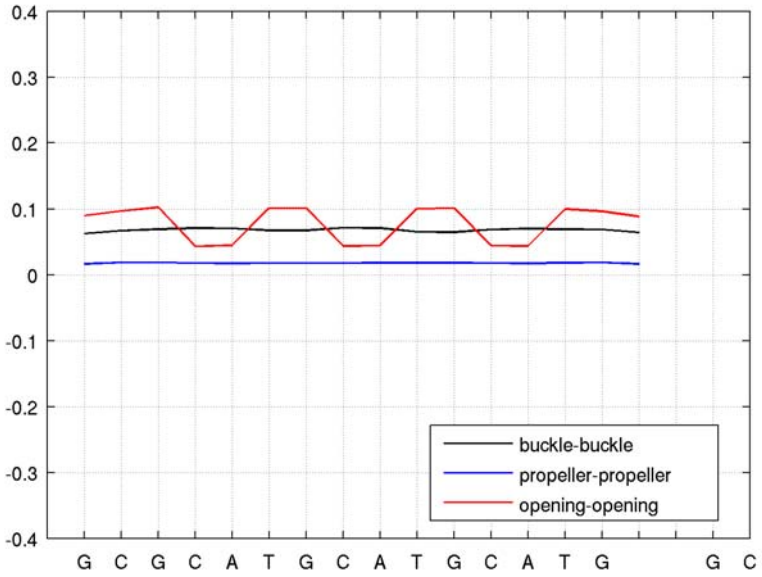
- The particular poly(AT) sequence

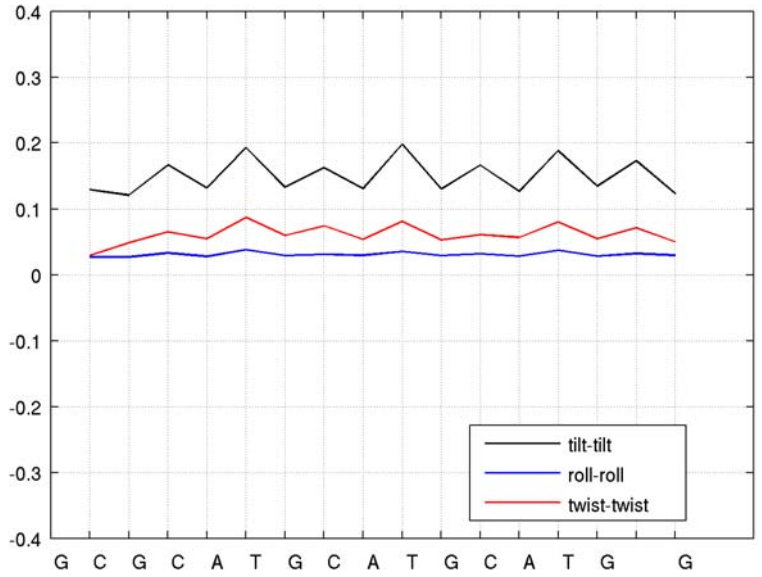
- The particular MD protocol

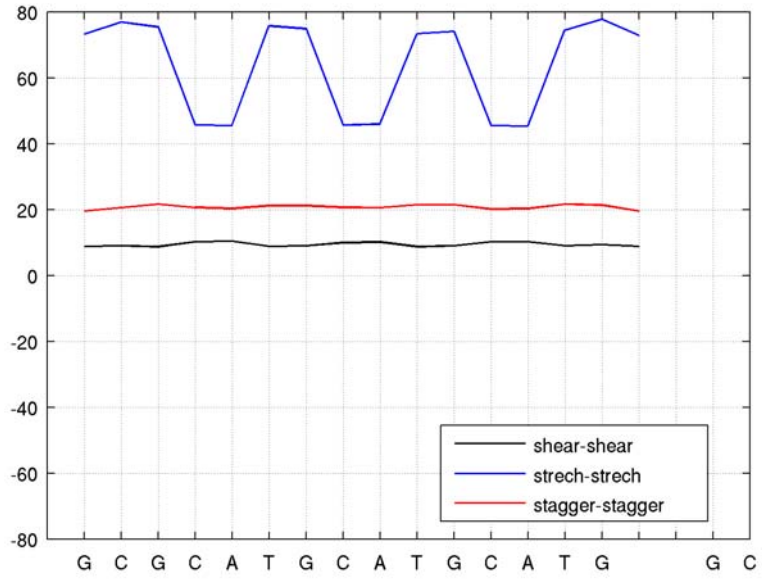
Next refute these possibilities by considering a poly (ATGC) sequence from the ABC II data set.

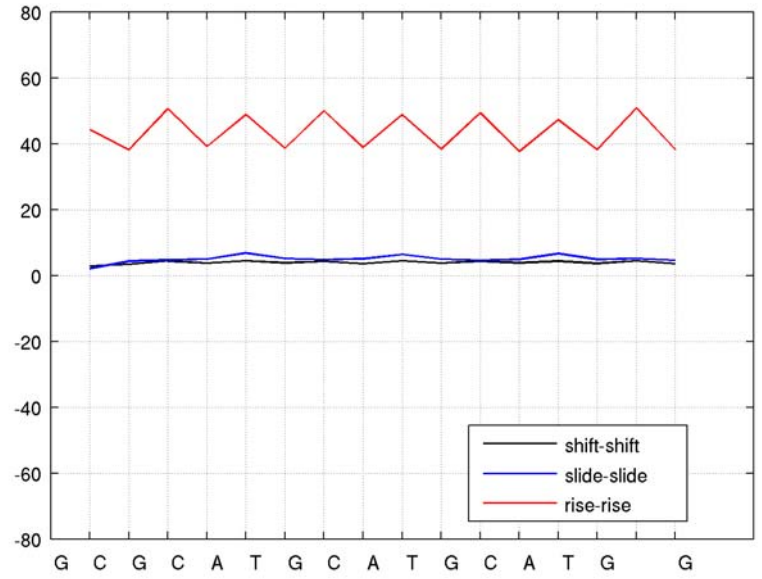


On the left averages of the six intra base-pair variables for each of the 18 base-pairs in a $GCGC(ATGC)_3GC$ oligomer, and on the right the six inter-base-pair parameters for each of the 17 junctions. Rotations above, translations below.



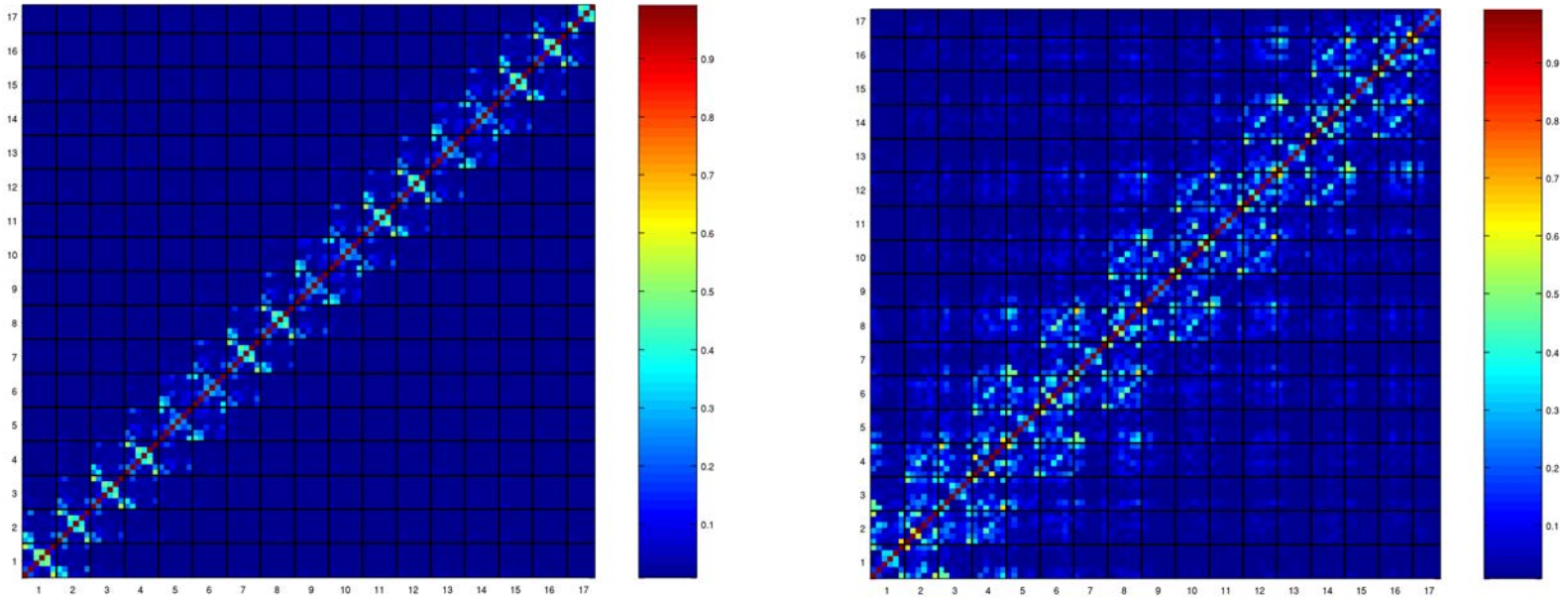






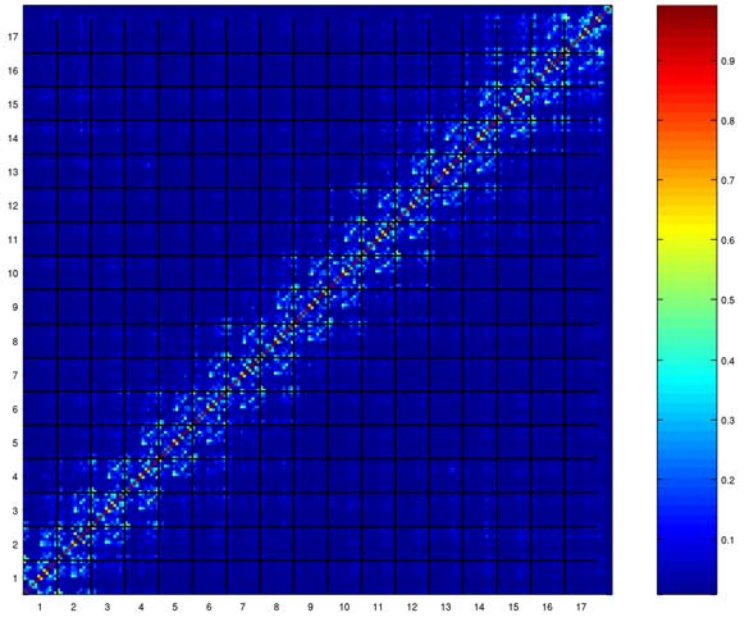
Diagonal stiffnesses for all 12 inter and intra base pair parameters

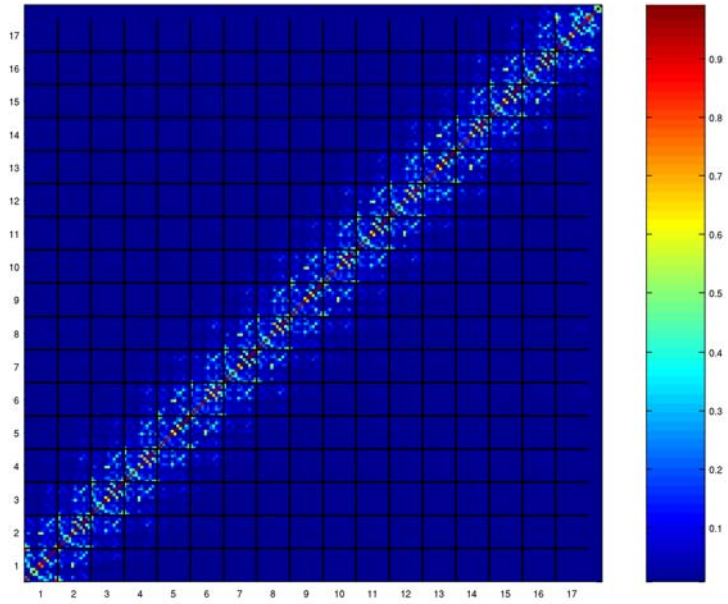
Covariance left and stiffness matrices (diagonals scaled to 1, ordered base by base) for rigid base pair model



Again we have the implausible observation that the stiffness matrix for a rigid base pair model is less banded than the correlation matrix.

Covariance left and stiffness matrices (diagonals scaled to 1, ordered base by base) for rigid base model





And again for a rigid base model the bandwidth of correlations and stiffnesses are rather similar.

The next possibility is that both MD protocols are wildly wrong. We refute that by making a comparison of diagonal blocks of the covariance matrix with analogous covariances drawn from crystal structures.

More specifically we compare the averages and covariances of the rigid base pair parameters for a particular step, which we arbitrarily chose to be the (Gp)CpA(pT) step between the 8th and 9th base pair in our MD simulation, with the analogous quantities from two crystal structure ensembles of CpA step parameters extracted by Olson et al from a) naked DNA crystals and b) DNA-protein complex crystals. Olson *et al* PNAS 1998,

`http://rutchem.rutgers.edu/~olson/pdna.html`

First averages or shape parameters (degrees and angstroms)

	Twist	Tilt	Roll	Shift	Slide	Rise
MD values	33.4	0.54	10.57	-0.03	-0.42	3.29
Naked B-DNA	37.7	0.2	1.7	-0.01	1.47	3.26
Protein - DNA complexes	37.3	0.5	4.7	0.09	0.53	3.33

With exception of the slide parameter all are really rather ‘close’. But more importantly for our purposes the three values of each parameter are of the same order of magnitude.

Second we compare the two Olson *et al* 6×6 covariances with the corresponding diagonal sub-block of the MD simulation. Rather than reporting all entries comparison is made first between the six eigenvalues

MD values	0.09	0.34	0.58	25.11	29.37	86.33
Naked B-DNA	0.02	0.09	0.22	6.00	9.22	118.12
Protein - DNA complexes	0.05	0.27	0.54	12.43	20.21	49.68

And then between the normalized eigenvectors by computing $X_1^T X_2$ and $X_1^T X_3$

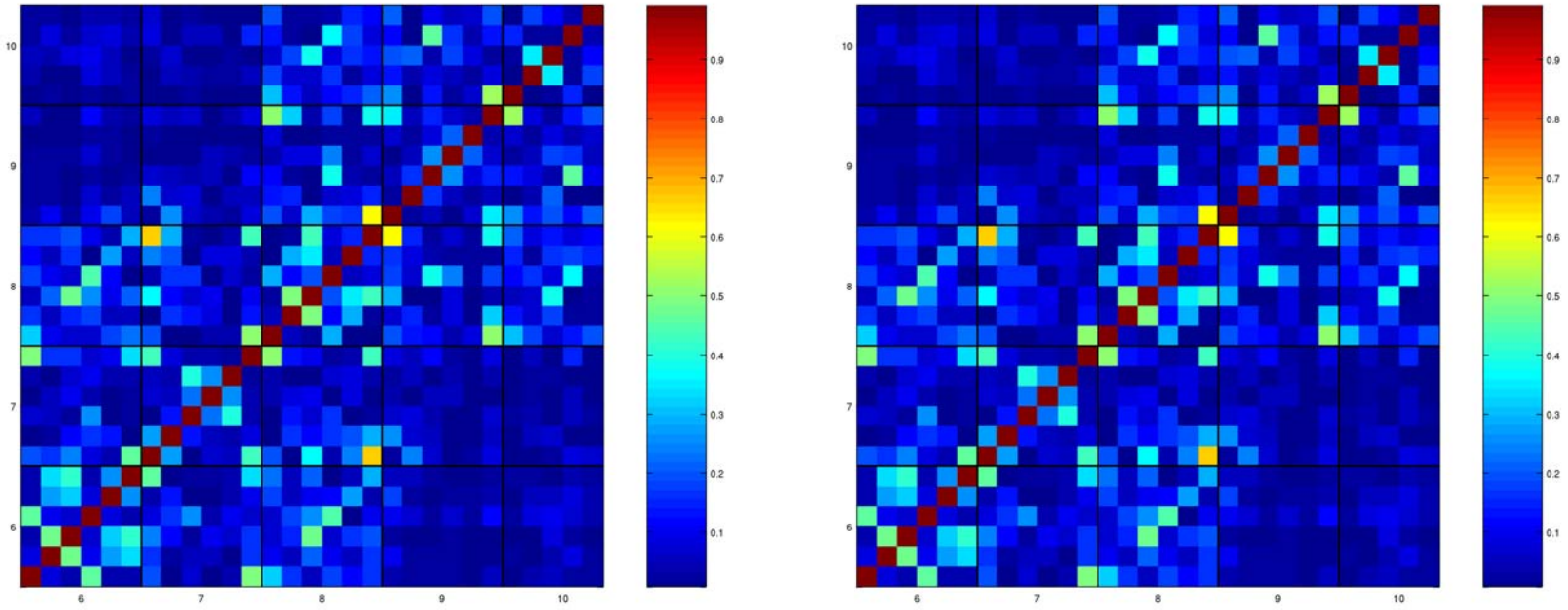
$$\begin{pmatrix} 0.98 & 0.18 & -0.03 & 0.01 & -0.02 & 0.00 \\ -0.01 & -0.12 & -0.99 & -0.06 & -0.02 & -0.03 \\ -0.18 & 0.97 & -0.12 & -0.00 & 0.04 & 0.05 \\ -0.03 & 0.04 & -0.00 & 0.39 & -0.92 & -0.08 \\ 0.00 & -0.03 & -0.06 & 0.92 & 0.38 & 0.09 \\ 0.00 & -0.05 & -0.02 & -0.06 & -0.11 & 0.99 \end{pmatrix} \quad \begin{pmatrix} 0.99 & 0.07 & 0.13 & 0.01 & -0.02 & 0.00 \\ -0.07 & -0.58 & 0.81 & 0.03 & -0.01 & -0.00 \\ -0.13 & 0.81 & 0.57 & -0.02 & -0.01 & 0.06 \\ 0.01 & -0.04 & 0.01 & -0.99 & -0.05 & 0.11 \\ 0.02 & -0.01 & 0.00 & -0.02 & 0.97 & 0.25 \\ 0.00 & 0.04 & 0.03 & -0.12 & 0.25 & -0.96 \end{pmatrix}$$

If eigenvectors are close, then up to permutation and sign, these two matrices should be close to the identity, and they both are.

The point is not to make a detailed entry by entry comparison. Rather these numbers indicate that the on diagonal 6×6 block covariances are well within an order of magnitude between the MD simulations and the two crystal ensembles, which gives an experimental reality check on the MD simulations.

To my knowledge no-one has looked at off-diagonal covariances in crystal structure covariances, but in MD the off diagonal are significant.

Covariance left and stiffness matrices (diagonals scaled to 1, ordered base by base) for rigid base pair model junctions 6 through 10



Conclusion super diagonal blocks have significant entries.

And ignoring the off-diagonal covariances has a significant effect on estimates of stiffness:

$$\frac{\|S - S_{Diag}\|}{\|S_{Diag}\|} = 4.84,$$

where S is the 6x6 block of the full stiffness matrix of the rigid base model, and S_{Diag} is the inverse of the 6x6 diagonal block of the covariance matrix.

In other words need to invert the global matrix to get good estimates of model parameters.

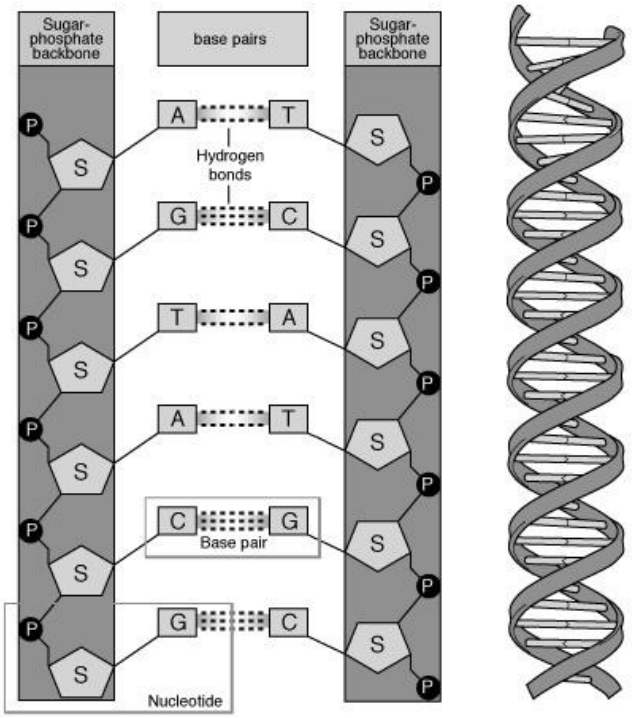
Conclusions:

- Coarse graining MD atomistic simulations of an oligomer to a localized rigid base-pair model/elastic rod model is not self-consistent at scales of 10–20 base pairs.

- Coarse graining MD atomistic simulations of an oligomer to a rigid base/elastic birod model does seem to be self-consistent at scales of 10–20 base pairs.

- 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.



Unlike rigid base-pair model, a rigid base model is pre-stressed, and sequence effects can propagate.

Remarks

- 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. High intrinsic twist plays a crucial role.

- 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. But cannot hope to capture either backbone flips or kinking.

- Once in a continuum setting semi-classical path integral methods allow very efficient evaluation of things like looping probabilities.

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, J. Curuksu, D, Perkevicuite: producing and visualizing Molecular Dynamics simulations
- S. Kehrbaum, S. Rey: High-twist homogenization
- M. Moakher: Rigid-base and Elastic Birod models
- L. Cotta-Ramusino: Semi classical path-integrals and entropic corrections in Continuum Models