A rigid-base model for DNA structure prediction

O. Gonzalez

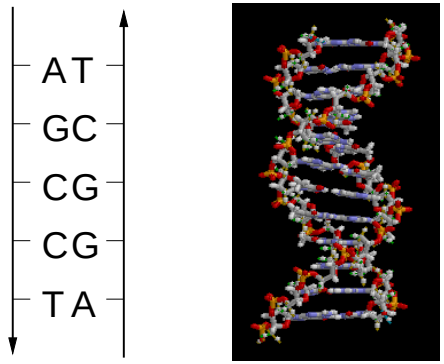
Introduction

Objective. To develop a model to predict the structure and flexibility of standard, B-form DNA from its sequence.

Introduction

Objective. To develop a model to predict the structure and flexibility of standard, B-form DNA from its sequence.

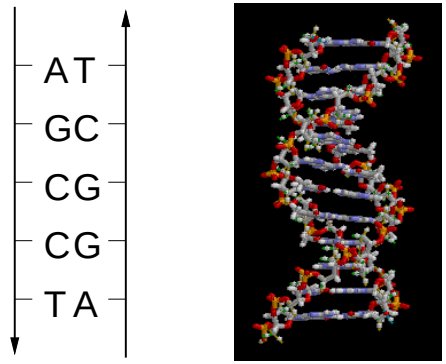
Idealized structure proposed by Watson and Crick, 1953.



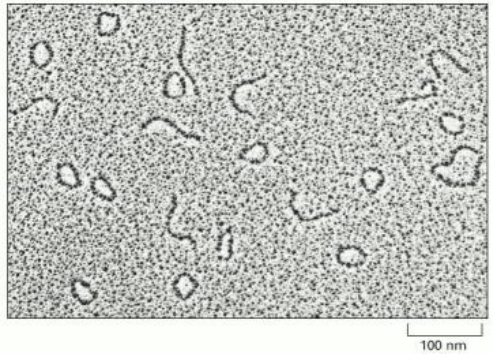
Introduction

Objective. To develop a model to predict the structure and flexibility of standard, B-form DNA from its sequence.

Idealized structure proposed by Watson and Crick, 1953.



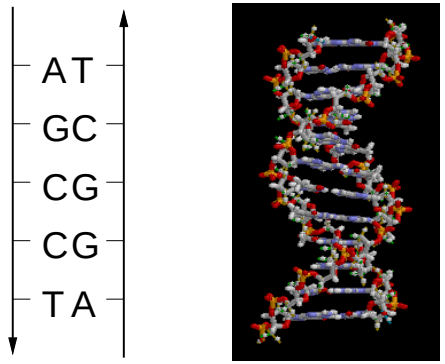
Actual structure depends strongly on sequence, circa 1980.



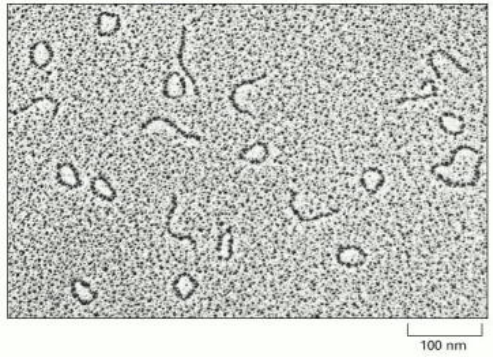
Introduction

Objective. To develop a model to predict the structure and flexibility of standard, B-form DNA from its sequence.

Idealized structure proposed by Watson and Crick, 1953.



Actual structure depends strongly on sequence, circa 1980.



Background. 30 years of history; lack of data hindered progress; recent construction of large MD dataset is making it accessible.

Ascona B-DNA Consortium

Contributing labs.

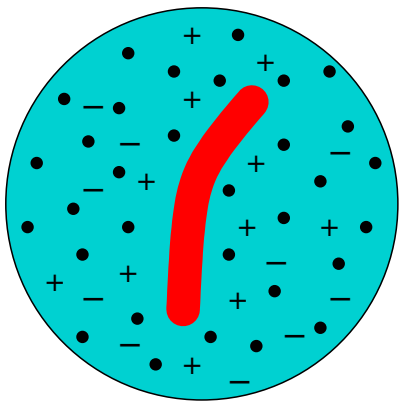
D. Beveridge (Wesleyan)
T. Bishop (Tulane)
D. Case (Rutgers)
T. Cheatham (Utah)
B. Jayaram (Delhi)
F. Lankas (Prague)
C. Laughton (Nottingham)
R. Lavery (Lyon)
J. Maddocks (Lausanne)
M. Orozco (Barcelona)
R. Osman (Mt. Sinai)
A. Perez (Barcelona)
H. Sklenar (Berlin)
J. Sponer (Brno)
M. Young (Berkeley)
...

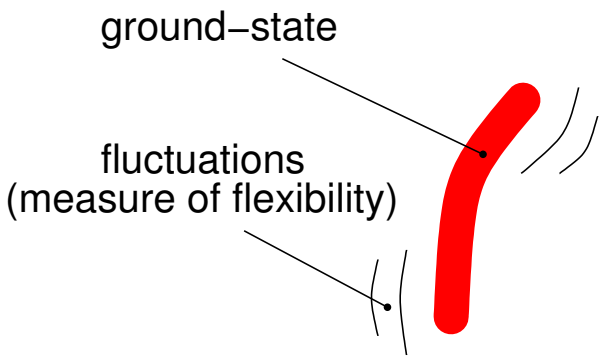
MD dataset. (consortium+local)

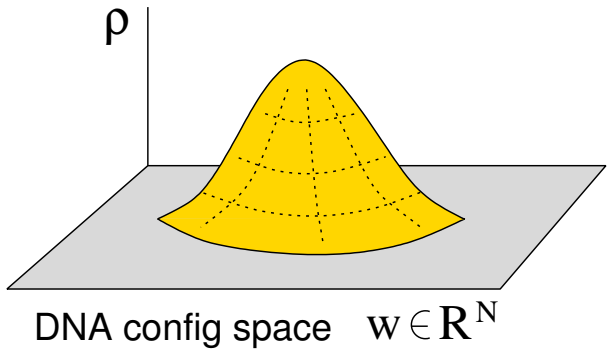
- over 50 different DNA oligomers (12-18 bp each)
- all 136 unique tetramer sub-sequences represented
- all 10 unique dimer end-sequences represented
- standard simulation protocol w/AMBER program
- all simulations w/explicit water and counterions
- 50-200 ns simulation time for each oligomer

Basic problem

Under fixed solvent conditions, we seek a model to predict the ground-state structure and flexibility of any given DNA oligomer.





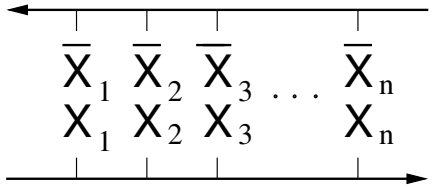


$$\rho(w) = \frac{1}{Z} e^{-\beta U(w)}$$

w	configuration coordinates
$\rho(w)$	probability density function
$U(w)$	free energy
Z, β	constants

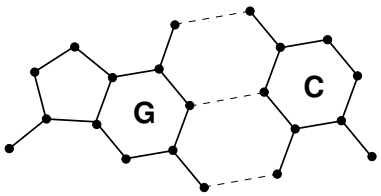
Rigid-base representation

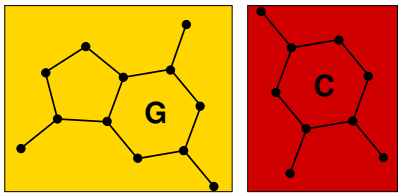
We consider a model in which each base is modeled as a separate rigid body; side-chains are not considered explicitly.

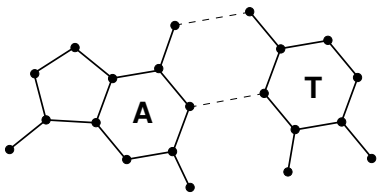


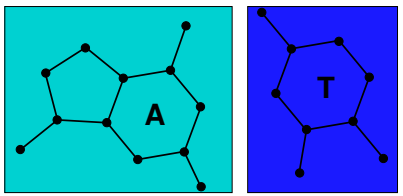
$X_a \in \{ T, A, C, G \}, \quad a = 1 \ldots n$

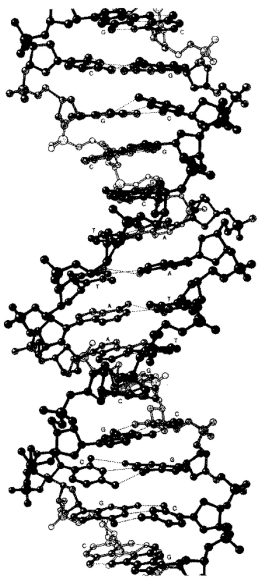
$\overline{A} = T, \ \overline{T} = A, \ \overline{C} = G, \ \overline{G} = C$

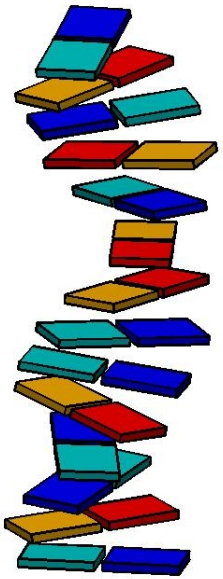






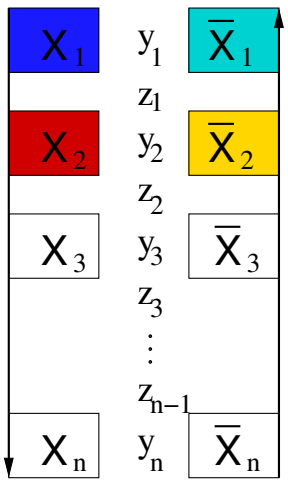


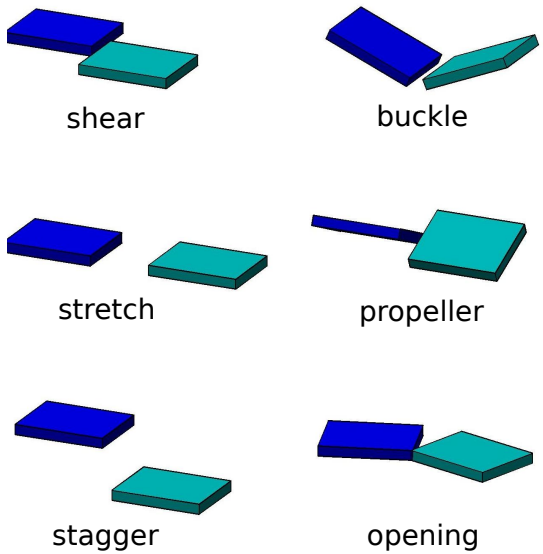


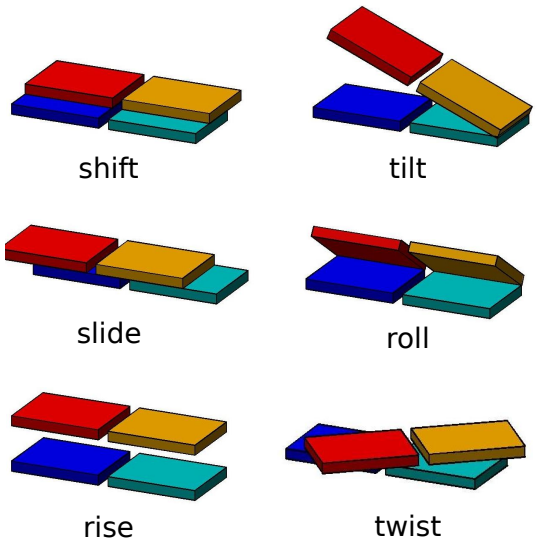


Configuration coordinates

An oligomer with n basepairs has $6n$ intra-basepair and $6(n-1)$ inter-basepair degrees of freedom; a total of $N = 12n - 6$.







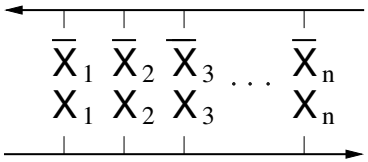
$y_a \in \mathbb{R}^6$ intra

$z_a \in \mathbb{R}^6$ inter

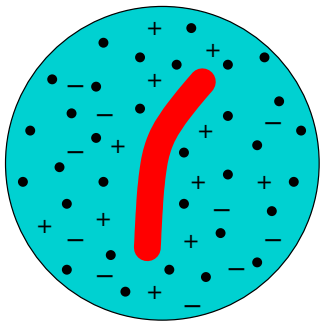
The oligomer coord vector is $w = (y_1, z_1, \ldots, z_{n-1}, y_n) \in \mathbb{R}^N$.

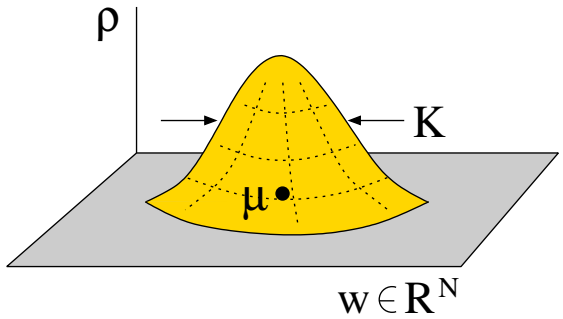
Free energy

Motivated by observed data, we consider a model in which the free energy is quadratic.



$S = X_1\ X_2\ X_3\ \cdots\ X_n$





$$\rho(w) = \frac{1}{Z} e^{-\beta U(w)}, \qquad U(w) = \frac{1}{2}(w-\mu)\cdot K(w-\mu)$$

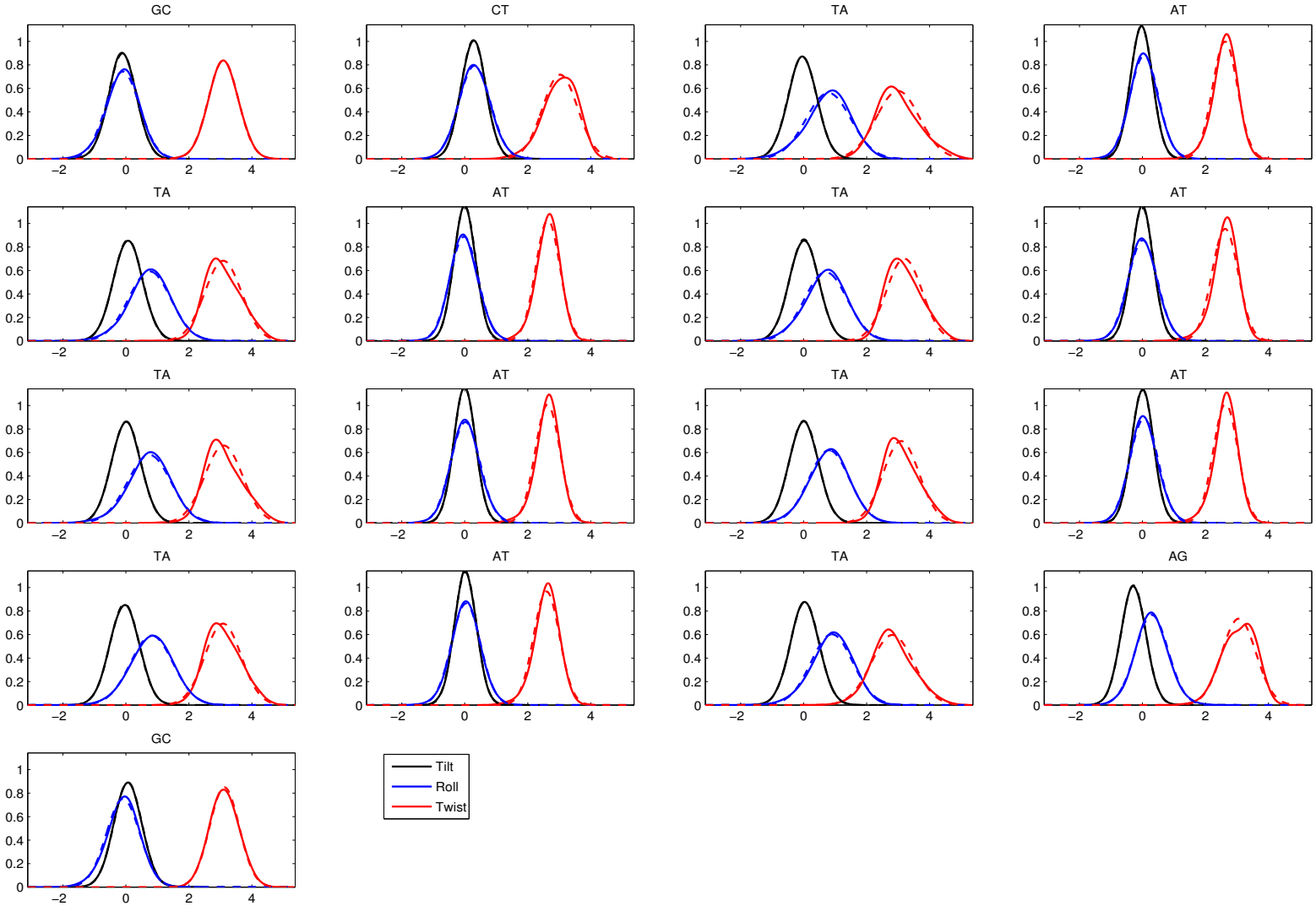
$\mu = \mu(S) \in \mathbb{R}^N$ ground-state configuration

$K = K(S) \in \mathbb{R}^{N\times N}$ ground-state stiffness

We seek explicit approximations to the functions $\mu(S)$ and $K(S)$.

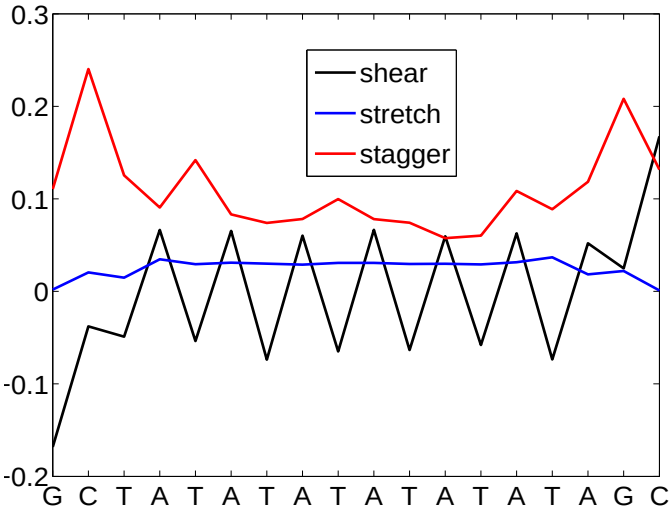
Sample data: coordinate marginals

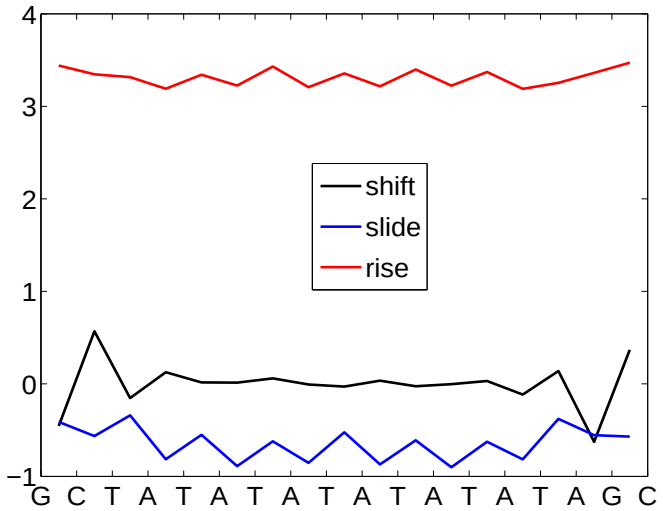
S=GCTATATATATATATAGC

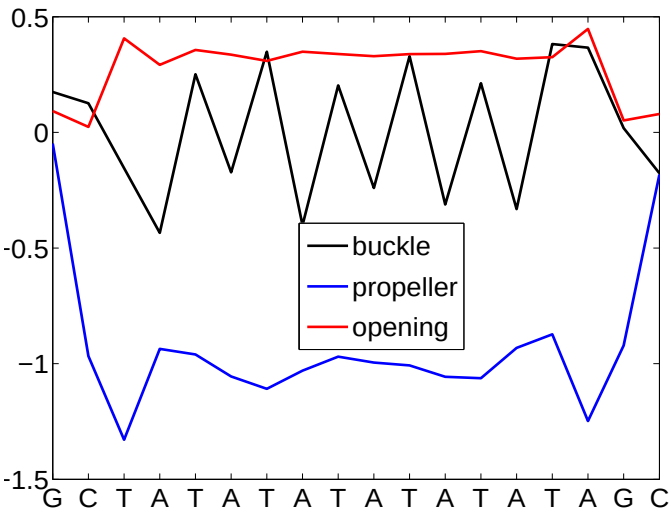


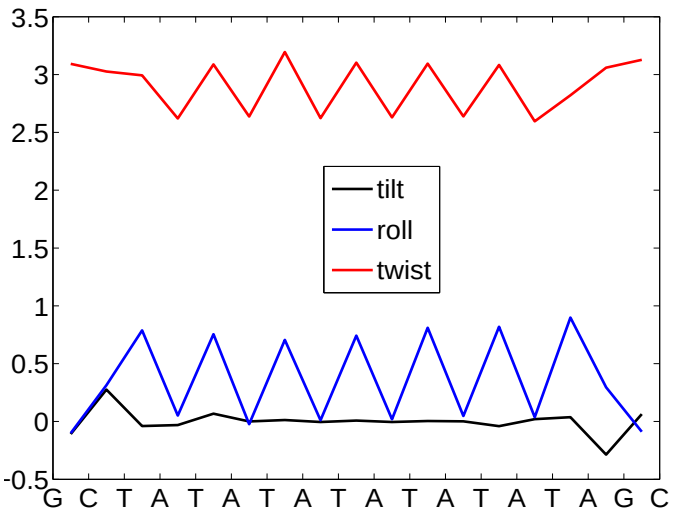
Sample data: ground-state configuration

S=GCTAT**A**TATATATATAGC



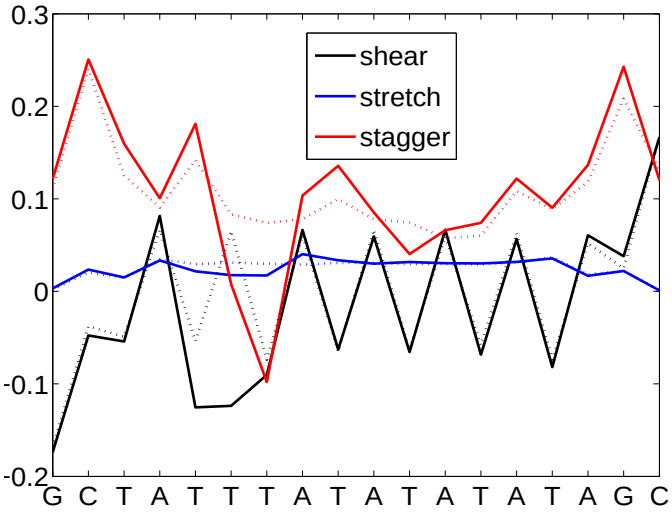


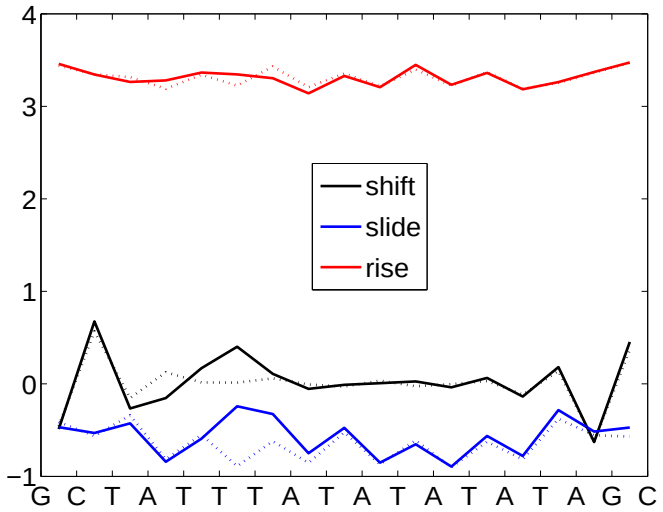


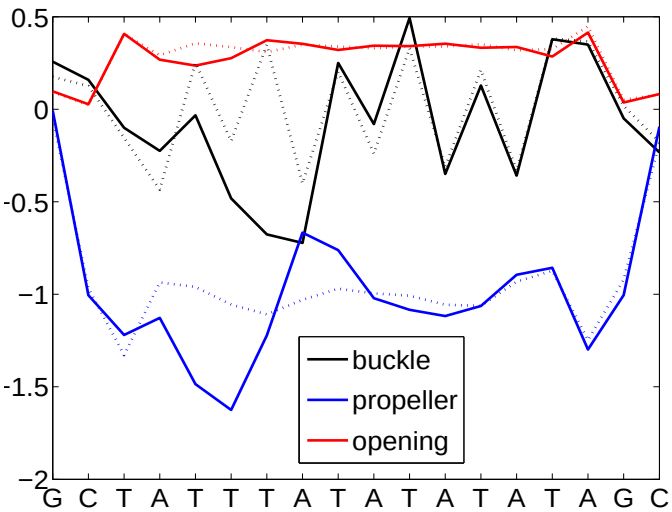


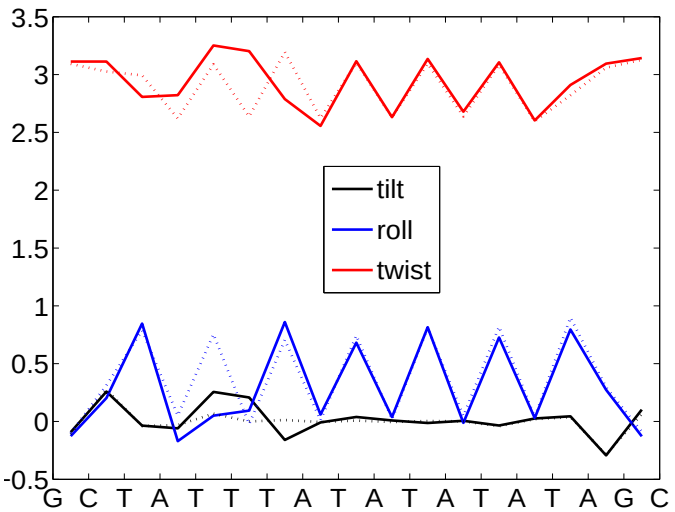
Sample data: ground-state configuration

S=GCTAT**T**TATATATATAGC



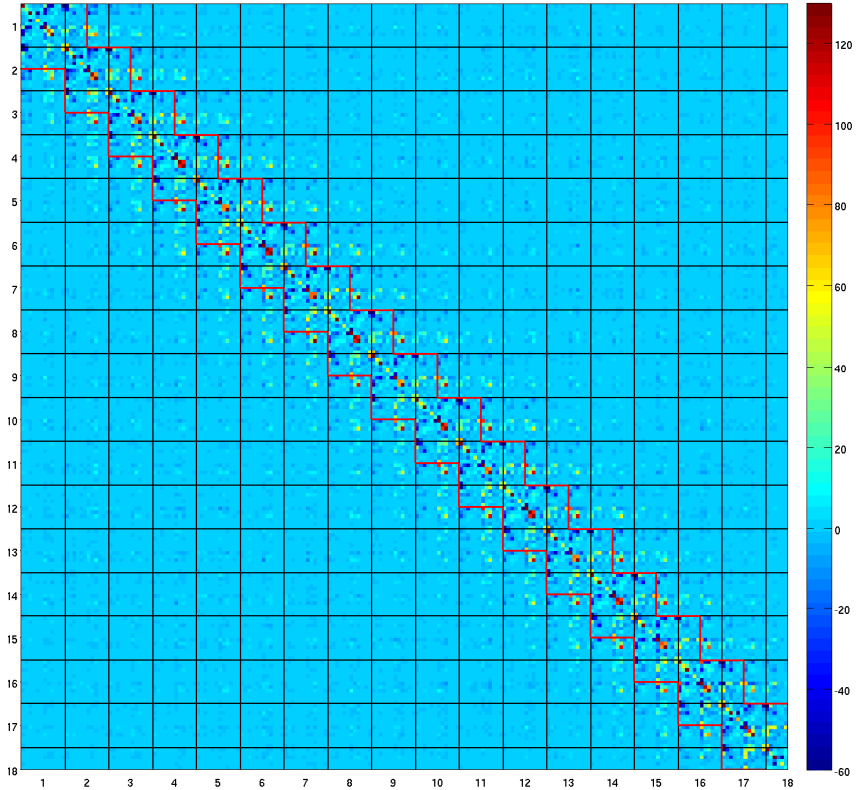


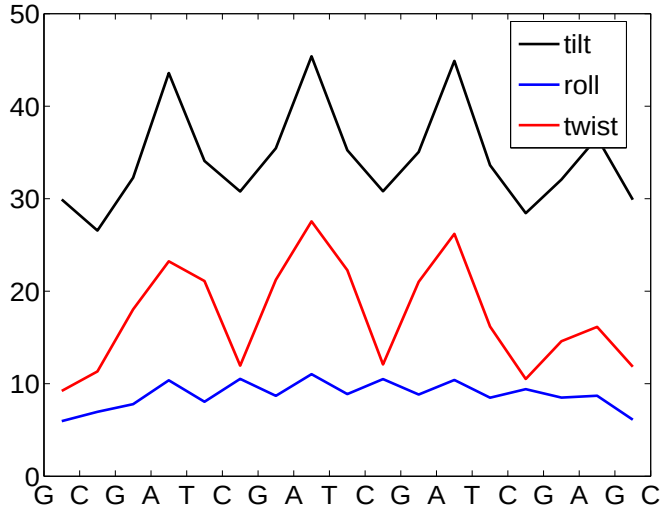




Sample data: ground-state stiffness

S=GCGATCGATCGATCGAGC

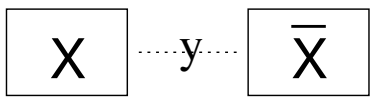




A monomer/dimer based model

We consider a model based on two types of interaction energies.

monomer interaction energy



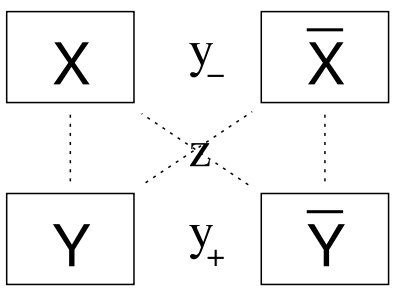
$w_m = y \in \mathbb{R}^6$

$$U_m = \tfrac{1}{2}(w_m - \mu_m^X) \cdot K_m^X (w_m - \mu_m^X)$$

$$\mu_m^X \in \mathbb{R}^6, \quad K_m^X \geq 0 \in \mathbb{R}^{6\times 6}$$

$$X \in \{T, A, C, G\}$$

dimer interaction energy



$w_d = (y_-, z, y_+) \in \mathbb{R}^{18}$

$$U_d = \tfrac{1}{2}(w_d - \mu_d^{XY}) \cdot K_d^{XY} (w_d - \mu_d^{XY})$$

$$\mu_d^{XY} \in \mathbb{R}^{18}, \quad K_d^{XY} \geq 0 \in \mathbb{R}^{18\times 18}$$

$$X, Y \in \{T, A, C, G\}$$

A monomer/dimer based model

By summing the monomer/dimer contributions along an oligomer, we obtain the energy

$$U(w) = \frac{1}{2}(w-\mu)\cdot K(w-\mu) + C$$

$$\mu = \mu(S,\mathcal{P}), \qquad K = K(S,\mathcal{P}), \qquad C = C(S,\mathcal{P}),$$

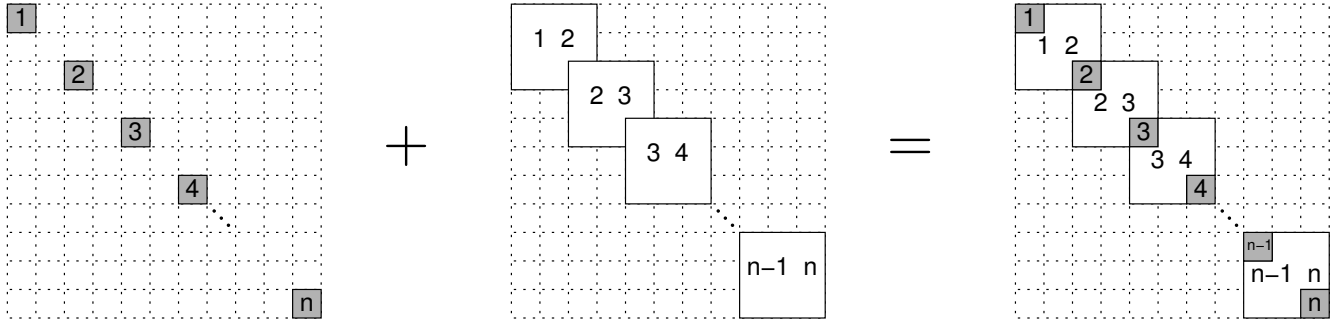
where S is the oligomer sequence and $\mathcal{P}$ is the model parameter set

$$\mathcal{P} = \begin{cases} K_m^X, & \sigma_m^X := K_m^X \mu_m^X, & X \in \{T,A,C,G\} \\ K_d^{XY}, & \sigma_d^{XY} := K_d^{XY} \mu_d^{XY}, & X,Y \in \{T,A,C,G\}. \end{cases}$$

A monomer/dimer based model

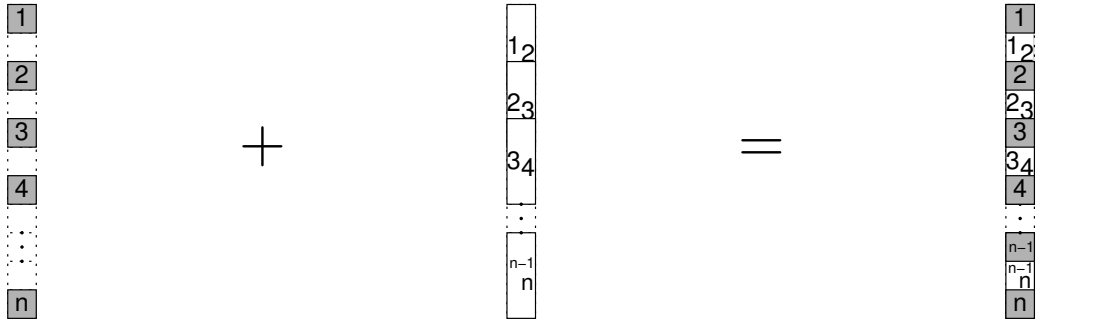
The ground-state configuration $\mu(S, \mathcal{P})$ and stiffness $K(S, \mathcal{P})$ are determined by $S = X_1 X_2 \cdots X_n$ and $\mathcal{P} = \{K_m^X, \sigma_m^X, K_d^{XY}, \sigma_d^{XY}\}$.

K:



6x6 monomers | 18x18 dimers | NxN oligomer

σ:



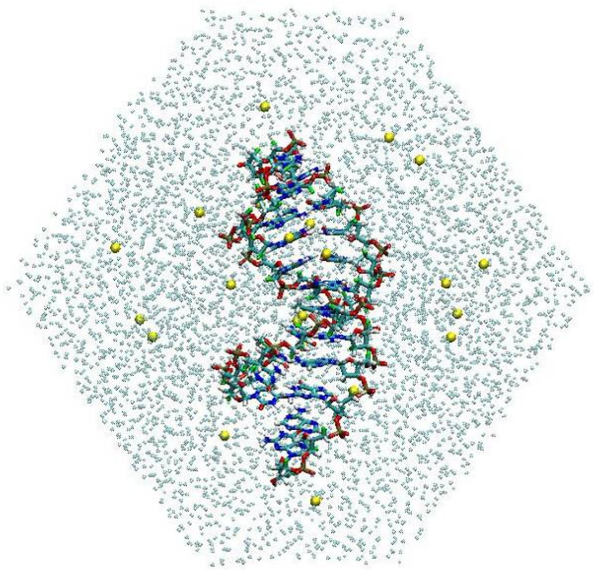
6x1 monomers | 18x1 dimers | Nx1 oligomer

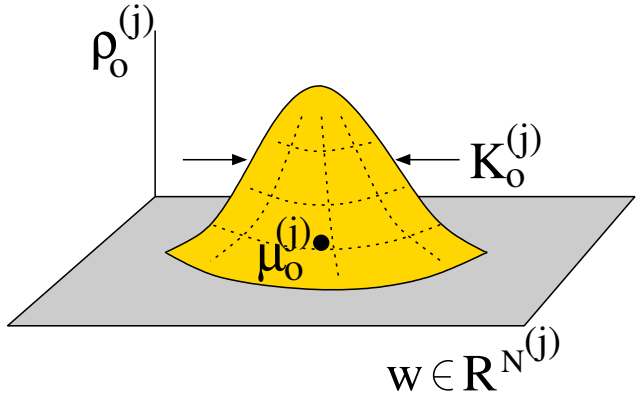
μ:

$$\mu(S, \mathcal{P}) = K(S, \mathcal{P})^{-1} \sigma(S, \mathcal{P})$$

Data for parameter estimation

To estimate the parameter set $\mathcal{P}$, we used a database of MD-observed probability densities $\rho_o^{(j)}(w)$ for sequences $S^{(j)}$, $j = 1, \ldots, J$.





$$\rho_o^{(j)}(w) = \frac{1}{Z^{(j)}} e^{-\beta U_o^{(j)}(w)}$$

$$U_o^{(j)}(w) = \frac{1}{2}(w - \mu_o^{(j)}) \cdot K_o^{(j)}(w - \mu_o^{(j)})$$

Functional for parameter estimation

A best-fit parameter set $\mathcal{P}$ can be obtained by minimizing the objective functional

$$\mathcal{F}(\mathcal{P}) = \sum_{j=1}^{J} D(\rho(S^{(j)}, \mathcal{P}), \rho_o^{(j)}),$$

where D is the Kullback-Leibler divergence (pre-distance)

$$D(\rho_*, \rho_o) = \int \rho_*(w) \ln \left[\frac{\rho_*(w)}{\rho_o(w)} \right] \, dw.$$

For Gaussians,

$$D(\rho_*, \rho_o) = \frac{1}{2} \left[K_*^{-1} : K_o + (\mu_* - \mu_o) \cdot K_o(\mu_* - \mu_o) - \ln \left(\frac{\det K_o}{\det K_*} \right) - I : I \right].$$

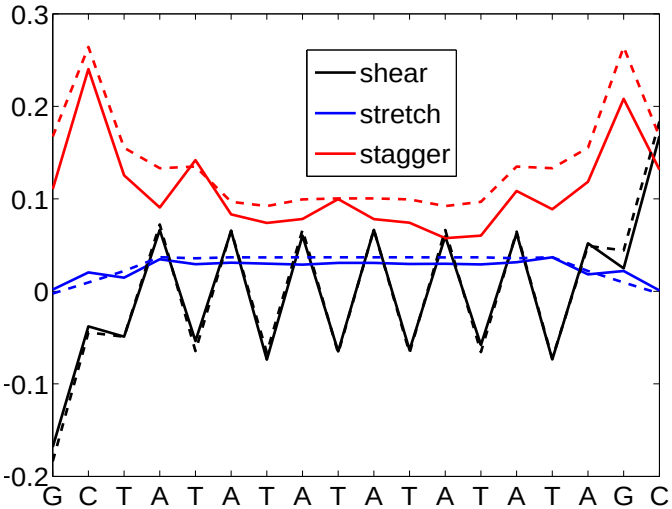
Results

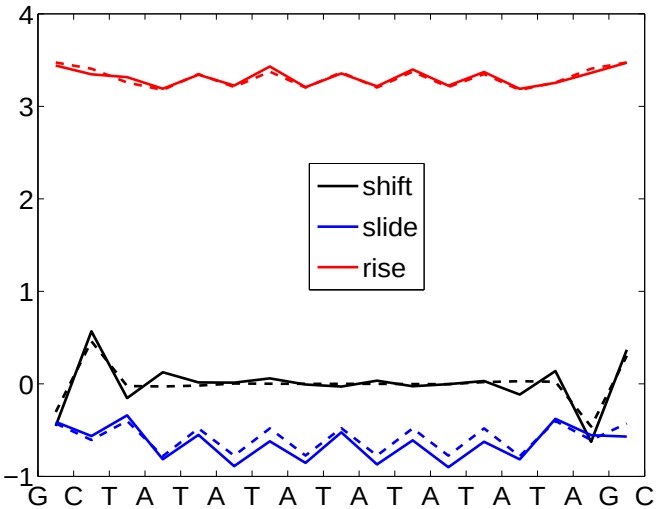
A best-fit parameter set $\mathcal{P}$ was obtained from the MD dataset via numerical minimization of the Kullback-Leibler functional.

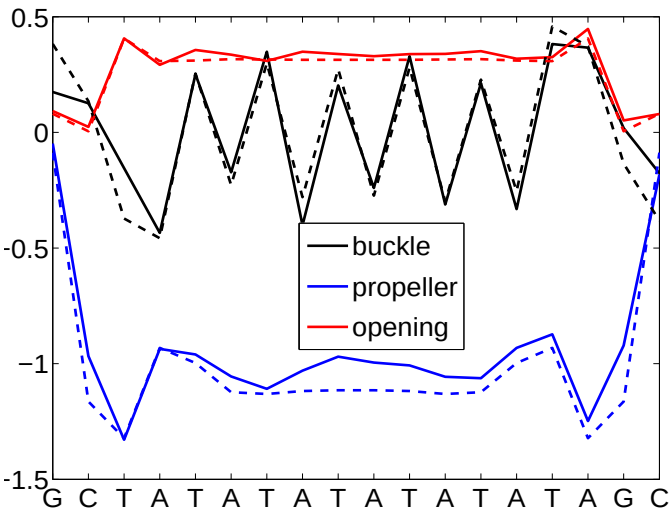
The parameter set $\mathcal{P}$ allows us to predict the ground-state configuration $\mu(S, \mathcal{P})$ and stiffness $K(S, \mathcal{P})$ for any sequence S.

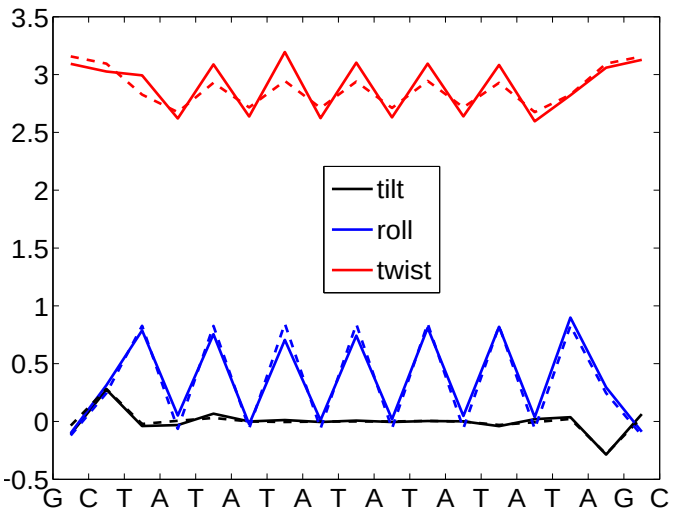
MD vs Model: ground-state configuration

S=GCTAT**A**TATATATATAGC



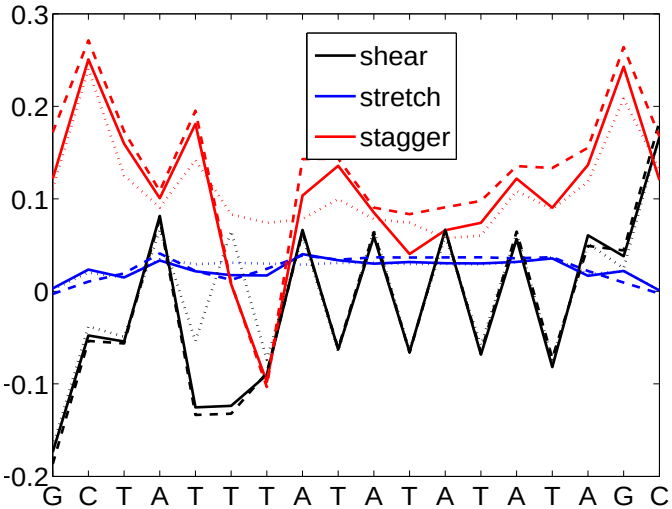


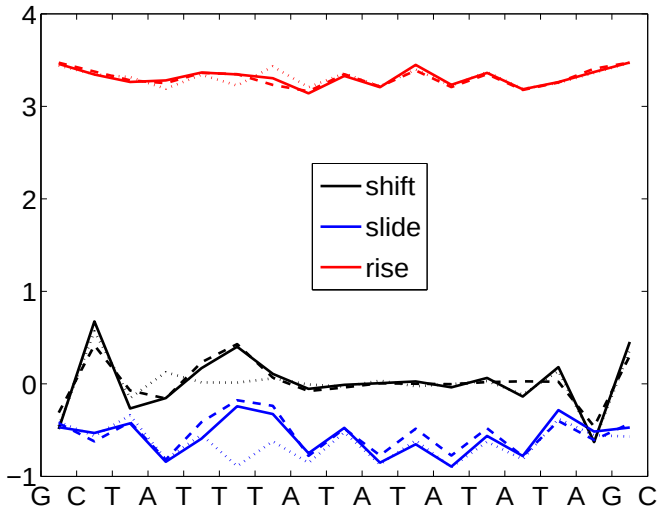


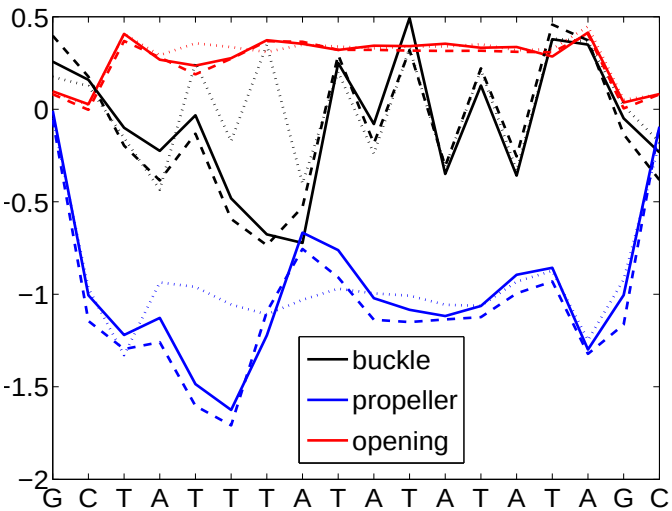


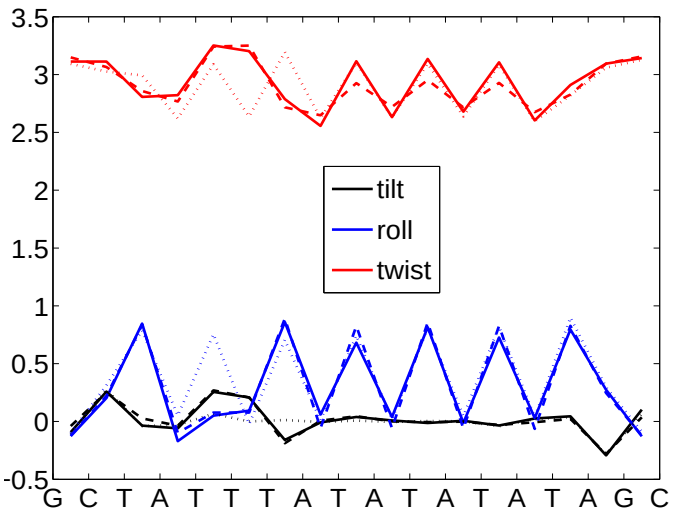
MD vs Model: ground-state configuration

S=GCTAT**T**TATATATATAGC



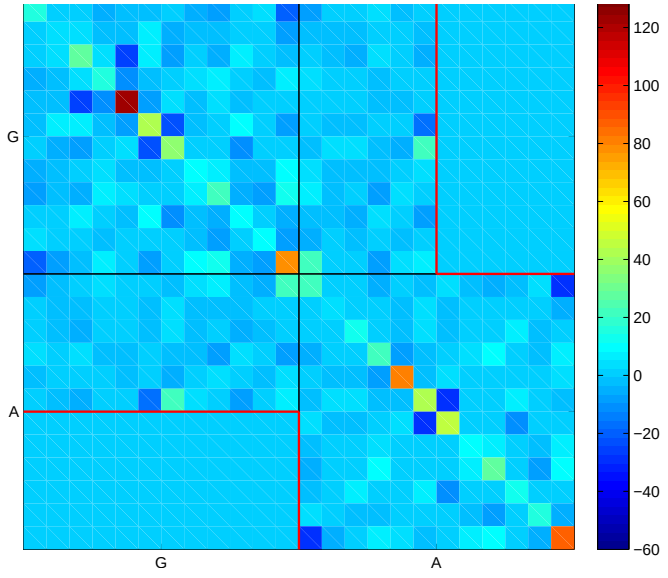




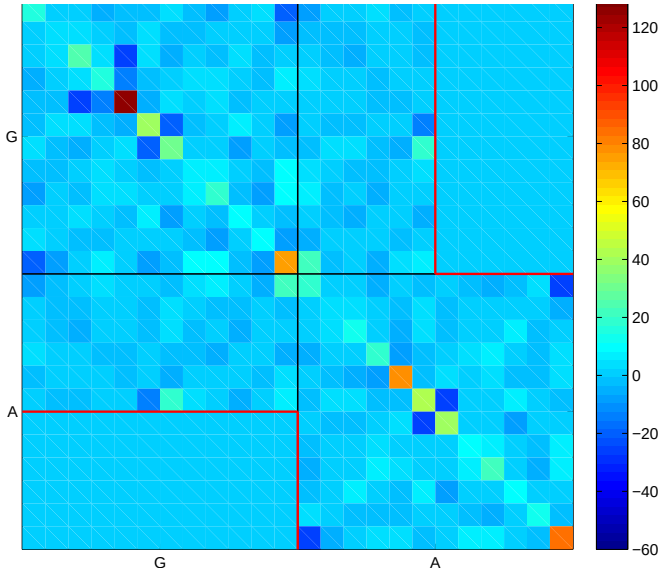


MD vs Model: ground-state stiffness

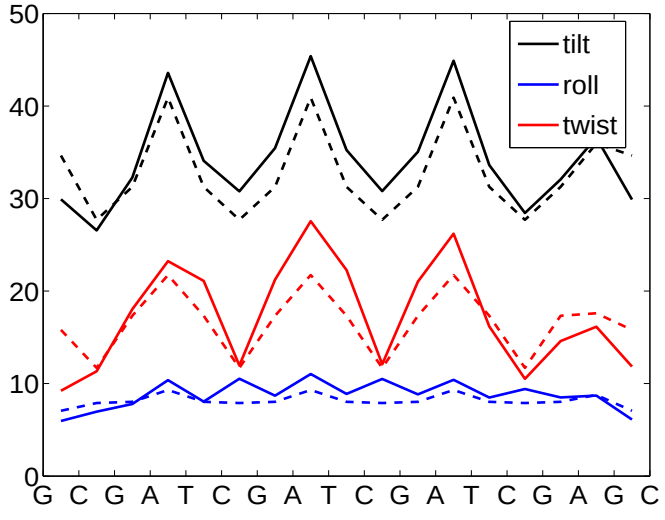
S=GCGATCGATCGATCGAGC



MD



Model



Summary

• A model to predict the ground-state configuration and flexibility of B-form DNA from its sequence has been developed.

• The model can resolve sequence-effects both within and between oligomers.

• The model was parametrized using MD and its predictive capabilities have been tested against MD.

• The model provides a way to quantify the intrinsic pre-stress or *frustration* in DNA.

• The model suggests non-local dependence of ground-state on sequence is due to pre-stress.

Thank You