Coarse-Grained Modelling of Protein-DNA Complexes

Kierran Falloon [1], Oliver Henrich [1], Aram Davtyan [2], Garegin Papoian [2]

[1] Department of Physics, University of Strathclyde, Glasgow, UK,
[2] DeepOrigin, South San Francisco, CA, USA

Introduction

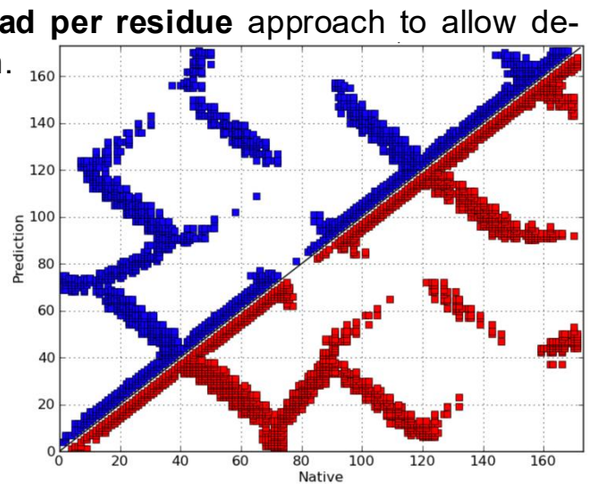
In recent times, the use of computational prediction and modelling in the study of biological macromolecules has been gaining traction to affordably fill gaps in experimental data. Presented here is a coupling of two leading coarse-grained models of protein and DNA respectively, implemented within the LAMMPS MD framework [1].

AWSEM-MD

AWSEM [2] adopts a **three bead per residue** approach to allow de-novo protein structure prediction.

It combines the use of:

- **Physical potentials** such as a harmonic backbone and hydrogen bonding.
- **Direct** and **water-mediated** interactions.
- **Structural biases** such as fragment memory, which are sequence-dependent.

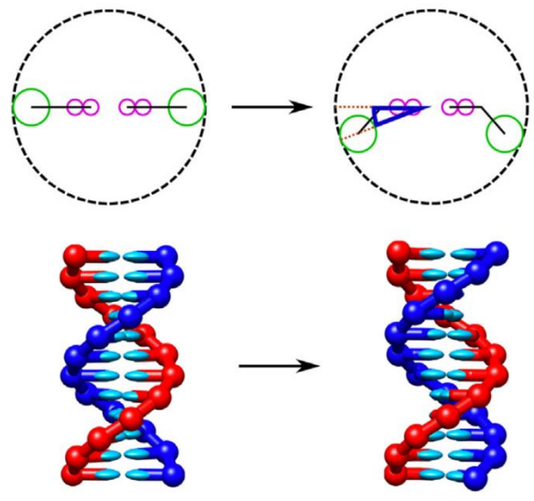


Predicted contact map from AWSEM *showing close agreement to native.*

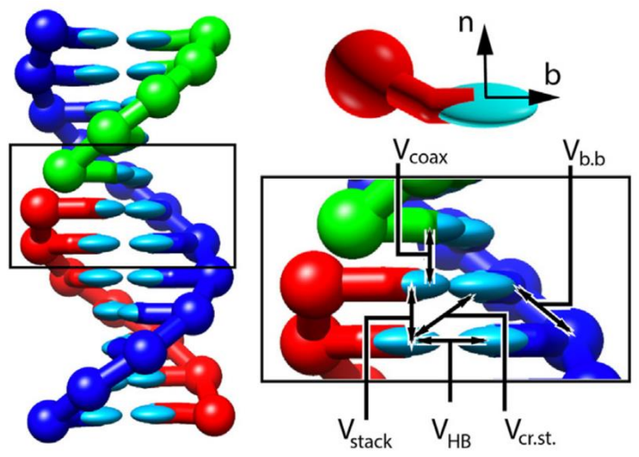
oxDNA Model

The oxDNA(2) model [3] consists of **rigid nucleotides** with **three interaction sites.** The model encompasses:

- A **FENE backbone**.
- **Hydrogen bonding** potentials.



Altering backbone bonding site from oxDNA1 reproduces the major and minor grooves in oxDNA2.



Overview of bonded and pair interactions in oxDNA2.

- **Stacking** potentials to model the tendency to form coplanar and cross-planar stacks.
- **Excluded volume** to to disallow atom overlapping.
- Offset backbone positioning to recreate **major-minor grooving**.
- **Debye-Hückel** electrostatics**.**

The Protein-DNA Potential

For a pair $\alpha\beta$, the excluded-volume interaction takes form of a **WCA potential**:

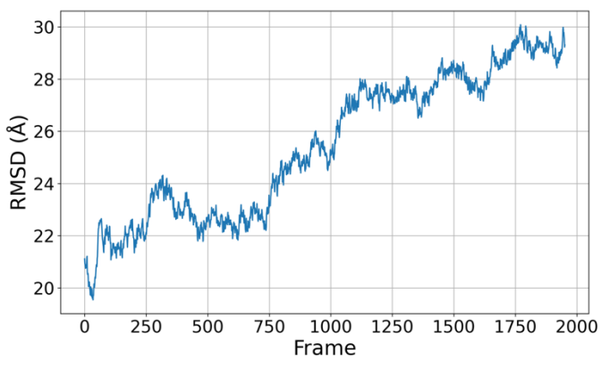
$$U_{\alpha\beta}^{\mathrm{WCA}} = \begin{cases} 4\epsilon_{\alpha\beta}\left[\left(\frac{\sigma_{\alpha\beta}}{r_{\alpha\beta}}\right)^{9} - \left(\frac{\sigma_{\alpha\beta}}{r_{\alpha\beta}}\right)^{6}\right] - U_{\alpha\beta}^{\mathrm{LJ}}(r_{\alpha\beta}^{m};\ \sigma_{\alpha\beta},\ \epsilon_{\alpha\beta}), & r_{\alpha\beta} \le r_{\alpha\beta}^{c_{LJ}}\ \ \epsilon_{\alpha\beta} = 0.03\ \mathrm{kcal/mol} \\ 0, & r_{\alpha\beta} > r_{\alpha\beta}^{c_{LJ}}\ \ \sigma_{\alpha\beta} = 5.7\ \text{Å} \end{cases}$$

Protein-DNA electrostatics are modeled through a **Debye-Hückel potential** to emulate screening:

$$U_{\alpha\beta}^{\mathrm{DH}} = \begin{cases} C \cdot q_{\alpha\beta}\frac{\exp(-\kappa r_{\alpha\beta})}{r_{\alpha\beta}}, & r_{\alpha\beta} \le r_{\alpha\beta}^{c_{DH}} \\ 0, & r_{\alpha\beta} > r_{\alpha\beta}^{c_{DH}} \end{cases}$$

$$C = \frac{e^2}{4\pi\epsilon_0\epsilon_r}, \quad \kappa = \lambda_{\mathrm{D}}(T,\ I)^{-1} = \sqrt{\frac{\epsilon_0\epsilon_r k_B T}{2N_A e^2 I}}^{-1}$$

Initially, unit charges were used. Protein charge lies on the C_β-backbone atom and is dependent on amino acid. DNA charge lies on the backbone interaction site and is negative.



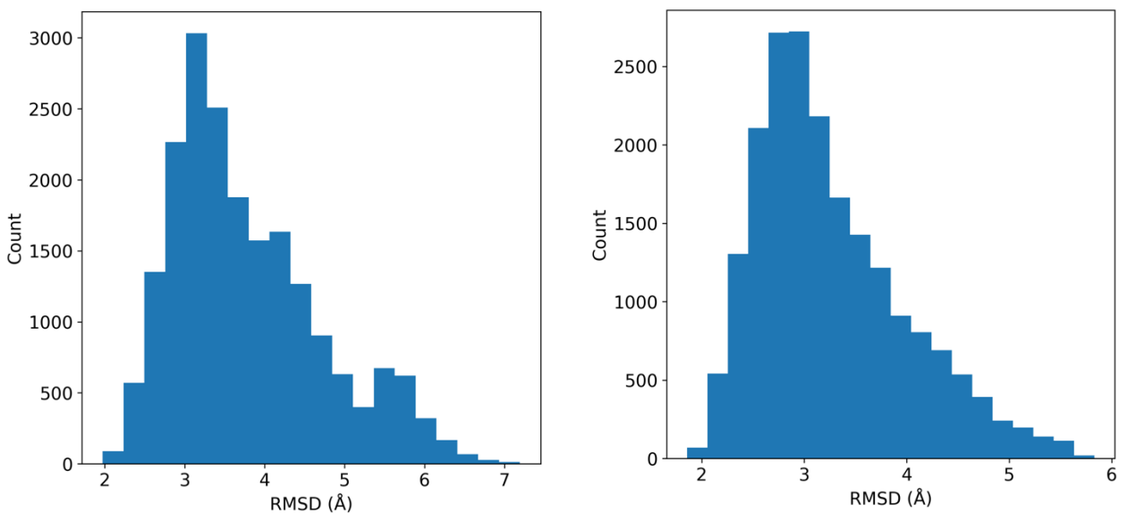
RMSD time series of model output at 300K. 1 frame = 1000 timesteps.

Optimisation

In order to not over-bias toward PDB structure, the function to minimise has a global minima at a nonzero point:

$$S = (\mathrm{RMSD} - T_{\mathrm{RMSD}})^2$$

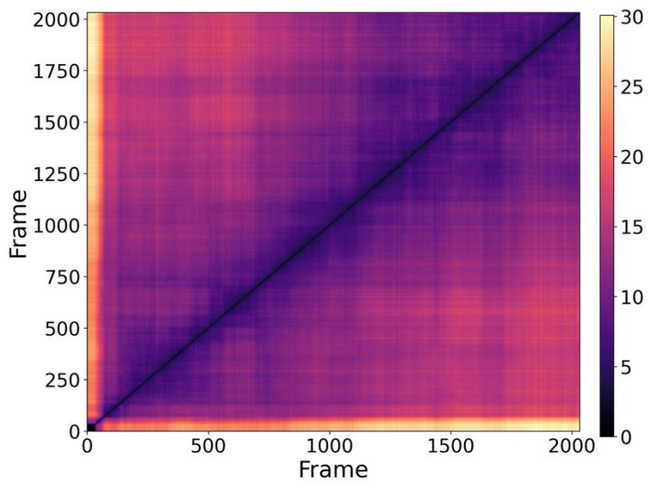
Basinhopping [4] with L-BFGS-B was used to iteratively converge on a parameter space with minima corresponding to experimentally viable states.



Histograms of sampled RMSD of region of interest before (left) and after (right) parameter optimisation with a target of 3Å.

Sampling Methodology

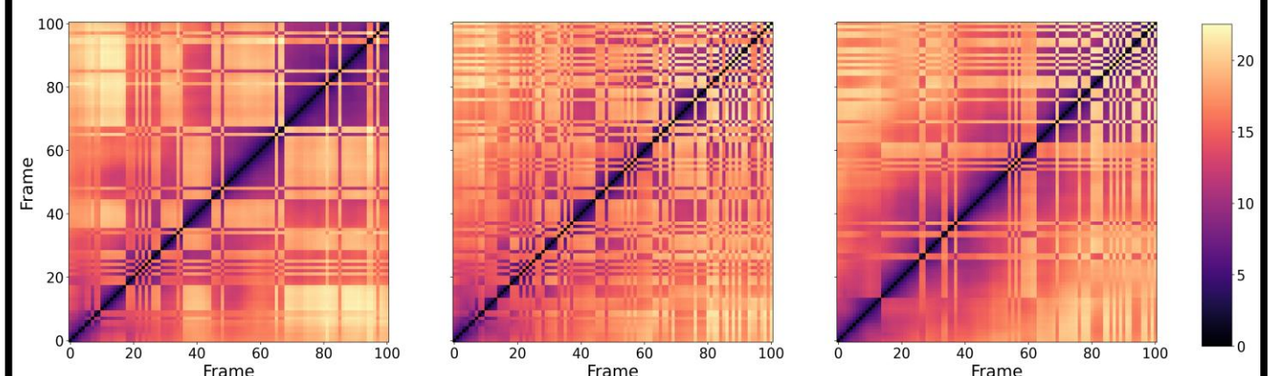
Due to having an abundance of experimental data and limited sequence specificity in structure formation, nucleosome structures were chosen for initial training of the model. However, the inherent stability of these structures lead to poor sampling efficiency. To improve this, Replica Exchange Molecular Dynamics (REMD) was employed.



All-to-all RMSD, showcasing stability for ~2000 frames after initial minimization.

Through estimating the integrated autocorrelation time of the RMSD, an efficient workflow was developed to deduce the optimal equilibration time by maximising the remaining number of statistically independent samples while minimising the number of frames to discard.

$$\tau_{\mathrm{int}} = 1 + 2\sum_{t=1}^{N}\left(1 - \frac{t}{N}\right)\hat{c}(t)$$



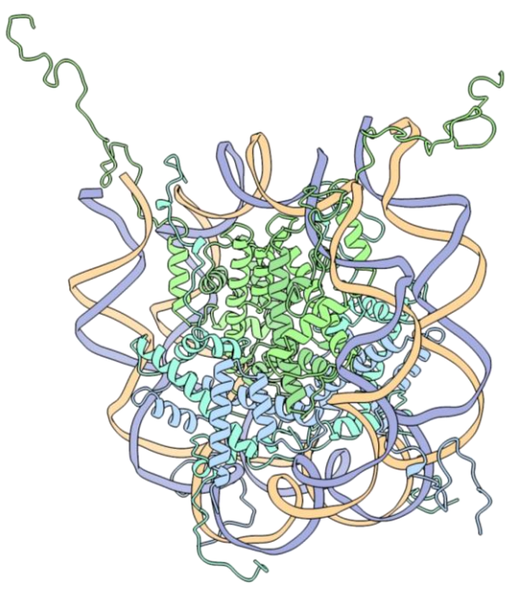
All-to-all RMSDs for temperatures 293.15K, 300K and 306.85K, highlighting the diverse configuration landscape created through REMD.

Choosing 5 independent seeds and temperatures in the range $290 \le T \le 310$K allows effective sampling of configuration space for a given structure. The success measure used is the average RMSD of the dynamical states (of a subset of atoms) when compared to the reference PDB structure:

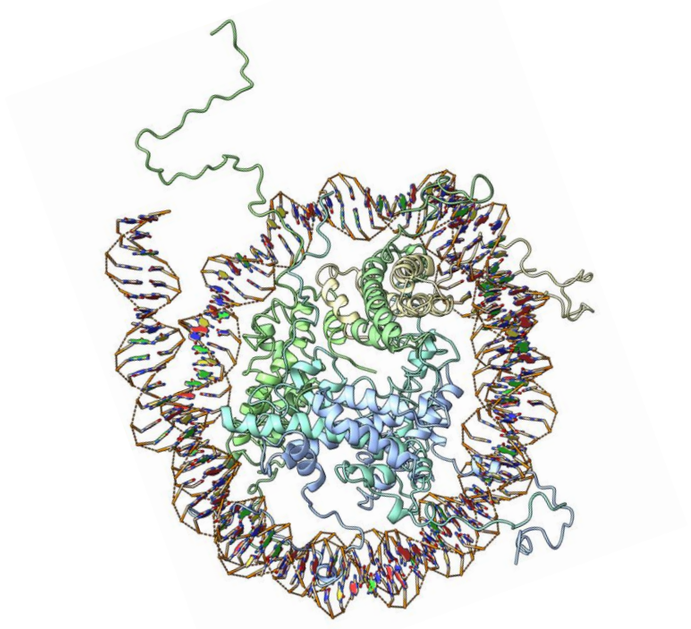
$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|\mathbf{r_i} - \mathbf{r_c}|^2}$$

Results

When employing steered optimisation, Basinhopping has performed demonstrably better in finding sensible parameter changes where local optimisation methods alone struggled. After five optimisation iterations on four unique nucleosome structures, a cost function value corresponding to an average RMSD of approximately 5.7Å was achieved.



Experimental PDB structure (1KX5)



Back-mapped structure reached through dynamical simulation under the implemented protein-DNA potential.

Conclusions

- Efficient sampling of highly stable systems has been realized through REMD. Analysis of correlation metrics guided the selection of optimal strategies during both equilibration and sampling.
- The ability to steer parameter change to allow for retention of biophysical mechanisms not captured directly in experimental comparison data has been demonstrated through a mixture of global and local optimisation methods.
- With ongoing refinement of the existing parameters, the model holds considerable promise for achieving high structural accuracy.

References

[1] A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. In't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, et al., Computer Physics Communications 271, 108171 (2022);
[2] A. Davtyan, N. P. Schafer, W. Zheng, C. Clementi, P. G. Wolynes, and G. A. Papoian, The Journal of Physical Chemistry B 116, 8494 (2012);
[3] T. Ouldridge, A. Louis, J. Doye, *J. Chem. Phys.* **134**, 085101 (2011); P. Sulc, F. Romano, T. Ouldridge, L. Rovigatti, J. Doye, A. Louis, *J. Chem. Phys.* **137**, 135101 (2012);
[4] D. J. Wales and J. P. Doye, "Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms," The Journal of Physical Chemistry A 101, 5111–5116 (1997).