

The cgDNA sequence-dependent coarse-grain model of dsDNA: Bridging the scales from Molecular Dynamics to Bioinformatics

John Maddocks + many co-authors mentioned as we go along

Laboratory for Computation and Visualization in Mathematics and Mechanics



Funding:



SWISS NATIONAL SCIENCE FOUNDATION



Einstein Stiftung Berlin
Einstein Foundation Berlin

IPAM, Learning Physical Models

October 30, 2019

Talk Summary I

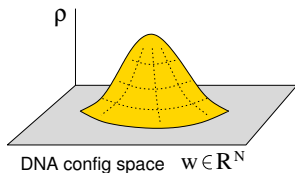
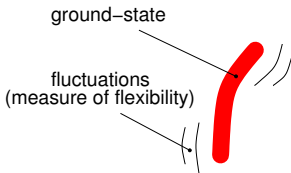
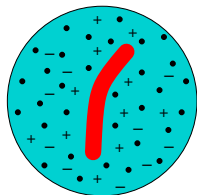
- No machine learning
- Instead an important(?) example which I suspect could benefit from machine learning techniques
- Pattern/feature recognition in ensembles (millions or billions of members) of multivariate (banded) Gaussian PDFs each in dimension N (N of the order a few hundred to a few thousand, band width 42)
- The ensembles of Gaussians are generated from a coarse-grain equilibrium statistical mechanics model of DNA called *cgDNAloc* that is used to scan over genomic length scales to try to identify short sub-sequences with exceptional mechanical properties. Enhance bioinformatics with mechanics.
- Might consensus protein binding site sequences share common mechanical properties?

Talk Summary II

- The *cgDNA* family of models have a machine learning flavour in that they have up to 20K parameters to be fit.
- But have to predict 4^n sequences with n of order a few 10s. (Not just two types of elephant.)
- In the unlikely event of there being time at the end, make some remarks about how to estimate dinucleotide dependent parameter sets for the coarse-grain banded Gaussian from a small library of long duration, atomistic (i.e. fine grain) MD simulations of short DNA fragments.

Statistical mechanical modelling of DNA

We want a predictive coarse-grain model for the sequence-dependent equilibrium distribution, including ground-state structure and flexibility, of a dsDNA fragment of any given sequence S and for a parameter set \mathcal{P} modelling given solvent conditions .

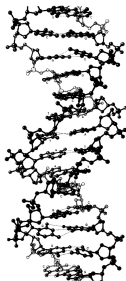
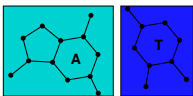
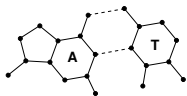
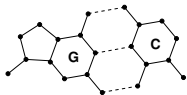
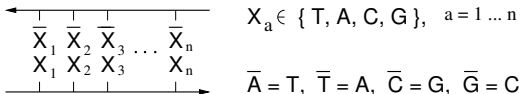


$$\rho(w; S, \mathcal{P}) = \frac{1}{Z} e^{-\beta U(w; S, \mathcal{P})}$$

w configuration coordinates
 ρ probability density function
 U free energy
 Z, β constants

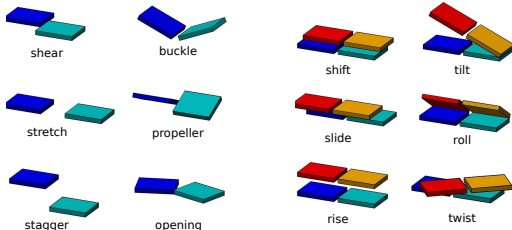
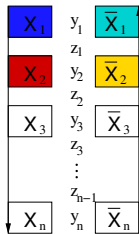
Rigid-base configuration variables

cgDNA is a coarse-grain model in which each base is explicitly described as a distinct rigid body; backbones are only considered implicitly. *cgDNA+* adds explicit treatment of phosphate groups.



Rigid Base Configuration Coordinates

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



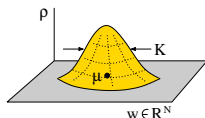
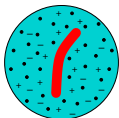
$$(x_a, \eta_a) =: y_a \in \mathbb{R}^6 \text{ intra bp}$$

$$(v_a, u_a) =: z_a \in \mathbb{R}^6 \text{ inter bp}$$

The oligomer coordinate vector is $w = (y_1, z_1, \dots, z_{n-1}, y_n) \in \mathbb{R}^N$.
 Concretely, use (small modifications of) Curves+ coordinates,
 Lavery et al (NAR, 2009) implementation of Tsukuba embedding
 of frames in atoms of each base, rotations via Cayley vectors.

The *cgDNA* model

The internal energy is approximated as quadratic so that the equilibrium distribution is a (high-dimensional) Gaussian.



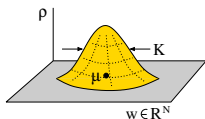
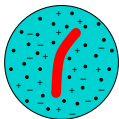
$$\rho(w) = \frac{1}{Z} e^{-\beta U(w)}, \quad U(w) = \frac{1}{2} (w - \mu) \cdot K (w - \mu)$$

$\mu = \mu(S, \mathcal{P}) \in \mathbb{R}^N$ ground-state configuration ($N = 12n - 6$)

$K = K(S, \mathcal{P}) \in \mathbb{R}^{N \times N}$ banded stiffness matrix $K = K^T > 0$.

The *cgDNA* model

The internal energy is approximated as quadratic so that the equilibrium distribution is a (high-dimensional) Gaussian.



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

$\mu = \mu(S, \mathcal{P}) \in \mathbb{R}^N$ ground-state configuration ($N = 12n - 6$)

$K = K(S, \mathcal{P}) \in \mathbb{R}^{N \times N}$ banded stiffness matrix $K = K^T > 0$.

cgDNA provides a variety of parameter sets \mathcal{P} that allow explicit construction of $\mu(S, \mathcal{P})$ and $K(S, \mathcal{P})$ for oligomers of arbitrary length sequence S . Currently each parameter set \mathcal{P} estimated from a library of fine grain, large scale Molecular Dynamics simulations.

The *cgDNA* model

Article and software download:

<http://lcvwww.epfl.ch/cgDNA/>

- O. Gonzalez, D. Petkeviciute, jhm, J. Chem. Phys. (2013) (basic model in gory detail)
- D. Petkeviciute, M. Pasi, Gonzalez, jhm, Nucleic Acids Research (2014) (lite version plus *cgDNA1.0* Matlab scripts)
- O. Gonzalez, M. Pasi, D. Petkeviciute, J. Glowacki, jhm, Multiscale Model. Simul. (2017) (math stuff about parameter estimation)

Web interface <http://cgDNAweb.epfl.ch>

- L. de Bruin, jhm, Nucleic Acids Research (2018, web server issue), primarily for interactive visualisation of ground states
- *cgDNA2.0* revision of Matlab scripts strictly compatible with cgDNAweb along with a Python implementation (Patelli, Zwahlen, Sharma).

Quick reminder of an elementary linear algebra computation

If A and B are symmetric matrices with $A + B$ invertible, and a and b are vectors, then the sum of two shifted quadratic forms can be written as a single shifted quadratic form plus a constant:

$$(x - a) \cdot A(x - a) + (x - b) \cdot B(x - b) = (x - c) \cdot C(x - c) + \text{const.}$$

where

$$C = A + B \quad (x_i x_j \text{ coefficients}) \quad c = C^{-1}(Aa + Bb) \quad (x_i \text{ coefficients})$$

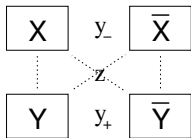
so that the value of the overall shift c involves the inversion of the matrix sum $(A + B)$ applied to Aa and Bb .

In our context, first average “forces” Aa and Bb , and then compute an effective ground state displacement c .

The *cgDNA2.0* dinucleotide step model

Based on localised interaction energies in each base-pair junction
each with localised sequence-dependence

Ten independent dinucleotide cross-junction interaction energies



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

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

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

$$\sigma_d^{XY} := K_d^{XY} \mu_d^{XY}$$

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

And allow end blocks to have different dimer dependence $K_d^{5'XY}$
and $\sigma_d^{5'XY}$ (end-fraying).

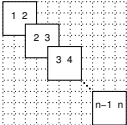

$XY3'$ blocks implied by $5'XY$ blocks via Crick-Watson reading
symmetry.

The *cgDNA* (2.0) model

Each dinucleotide-dependent parameter set \mathcal{P} is estimated by fitting first and second moment statistics to a library of (large scale) Molecular Dynamics simulations:

$$\mathcal{P} = \{K^{XY}, K^{5'XY}, \sigma^{XY}, \sigma^{5'XY}\}, \quad X, Y \in \{A, C, G, T, M, N\}.$$

Stiffness and shape reconstruction:

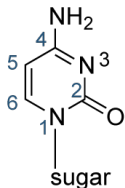
$$K(S, \mathcal{P}):$$

$$\sigma(S, \mathcal{P}):$$


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

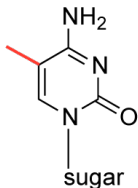
Formula comes from summing over junction energies. The matrix inversion means that the oligomer ground state μ has a nonlocal dependence on sequence. Linear algebra expression of **frustration**.

Two things to know: I. Epigenetic base modifications

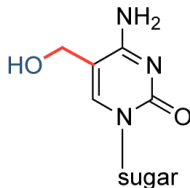
The C base in a CG base pair are often methylated when they occur in CpG dinucleotide steps



cytosine
(C)



5-methylcytosine
(mC)

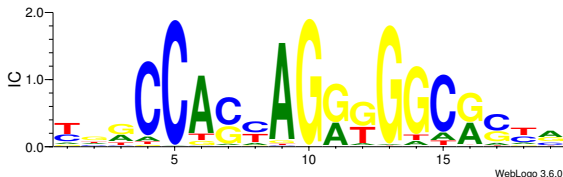


5-hydroxymethylcytosine
(hmC)

CpG steps generally under-represented in the genome except in CpG islands documented in bioinformatics databases. They frequently occur in promotor regions for genes, and whether or not they are methylated is known to be very important for gene expression.

Two things to know: II. Sequence logos, a standard tool in bio-informatics

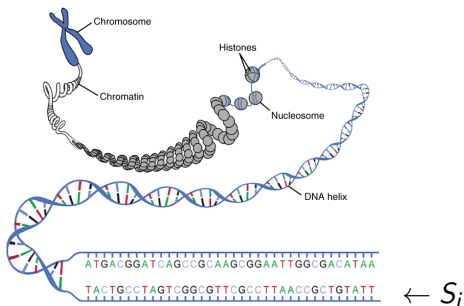
- Example: over 930 known CTCF transcription factor binding sites from JASPAR ¹ data base



¹Mathelier, et al. Nucleic Acids Research, 2014.

Identifying mechanically exceptional DNA sub-sequences inside a chromosome

Typically a chromosome sequence S is of length $10^5 - 10^8$ bp. Way beyond the length scale where it is either feasible or sensible to compute a *cgDNA* pdf. Instead aim to compute *cgDNA/loc* marginal PDFs for (short) sub-sequences $S_i \subset S$. (PhD thesis T. Zwahlen)



Dealing with shape nonlocality

Computing (Gaussian) marginal PDFs of Gaussians is a standard thing, and a marginal of a banded Gaussian is even itself also banded, which means that computing *cgDNA* marginals is very fast.

However, one of the strengths of the *cgDNA* model is the nonlocal sequence-dependence of the ground state shape $\mu(S)$, and this has to be accounted for by taking a marginal of a sub-sequence within a sub-sequence with additional known flanking base pairs:

$$\rho_{loc}(w | S_i \subset S'_i \subset S) = \frac{1}{Z} e^{-(w - \mu_{loc}) \cdot K_{loc}(w - \mu_{loc})}$$

Application of *cgDNAloc*: locating mechanically exceptional DNA sequences

Objective: given a long DNA sequence S , e.g. a chromosome, pick a fixed length (say 10 - 200 bp) and locate sites/sub-sequences S_i of the given length with exceptional mechanical properties in the sense that the PDF $\rho_{loc}(w|S_i)$ is far from a sequence-averaged PDF $\bar{\rho}_{loc}(w)$ for a fragment of the same length.

Application of *cgDNAloc*: locating mechanically exceptional DNA sequences

Objective: given a long DNA sequence S , e.g. a chromosome, pick a fixed length (say 10 - 200 bp) and locate sites/sub-sequences S_i of the given length with exceptional mechanical properties in the sense that the PDF $\rho_{loc}(w|S_i)$ is far from a sequence-averaged PDF $\bar{\rho}_{loc}(w)$ for a fragment of the same length.

- How to compare two probability distributions?

Application of *cgDNAloc*: locating mechanically exceptional DNA sequences

Objective: given a long DNA sequence S , e.g. a chromosome, pick a fixed length (say 10 - 200 bp) and locate sites/sub-sequences S_i of the given length with exceptional mechanical properties in the sense that the PDF $\rho_{loc}(w|S_i)$ is far from a sequence-averaged PDF $\bar{\rho}_{loc}(w)$ for a fragment of the same length.

- How to compare two probability distributions?
- How to compute the sequence-averaged PDF $\bar{\rho}_{loc}(w)$?

Comparing two PDFs

One (but only one) standard way to compare two probability distributions ρ_1, ρ_2 on \mathbb{R}^n is to compute the **Relative entropy** or **Kullback-Leibler divergence** between them, which is defined as

$$KL(\rho_1, \rho_2) = \int_{\mathbb{R}^n} \rho_1(x) \ln \left(\frac{\rho_1(x)}{\rho_2(x)} \right) dx.$$

If both ρ_i are (multivariate) Gaussians with means μ_i and inverse covariances or stiffnesses K_i , $i = 1, 2$ there is an explicit formula for the integration (which is important for us for computational time of evaluation):

$$KL(\rho_1, \rho_2) = \frac{1}{2} \left[\text{tr} (K_2 K_1^{-1}) - \ln \frac{\det K_2}{\det K_1} - n \right] + \frac{1}{2} (\mu_1 - \mu_2) \cdot K_2 (\mu_1 - \mu_2)$$

Computing average PDFs

Given an ensemble $\{\rho_i\}$ of PDFs with means and covariances μ_i , K_i , $i = 1, \dots, m$, compute a Gaussian average ρ_{av} from the ensemble mean and covariance μ_{av} , K_{av} by averaging shapes and covariances:

$$\mu_{av} = \frac{1}{m} \sum_{i=1}^m \mu_i, \quad K_{av}^{-1} = \frac{1}{m} \left(\sum_{i=1}^m K_i^{-1} + \mu_i \otimes \mu_i \right) - \mu_{av} \otimes \mu_{av}.$$

Computing average PDFs

Given an ensemble $\{\rho_i\}$ of PDFs with means and covariances μ_i , K_i , $i = 1, \dots, m$, compute a Gaussian average ρ_{av} from the ensemble mean and covariance μ_{av} , K_{av} by averaging shapes and covariances:

$$\mu_{av} = \frac{1}{m} \sum_{i=1}^m \mu_i, \quad K_{av}^{-1} = \frac{1}{m} \left(\sum_{i=1}^m K_i^{-1} + \mu_i \otimes \mu_i \right) - \mu_{av} \otimes \mu_{av}.$$

For an ensemble of banded *cgDNAloc* PDFs $\{\rho_{loc}(S_i)\}$ (say along a chromosome), define a **banded cgDNAloc average** ρ_{bav} (over the ensemble S_i) by best fitting to a Gaussian with a *banded* stiffness matrix. Know how to do this using a maximum entropy fit.

Application of *cgDNAloc*: locating mechanically exceptional DNA sequences, made precise

Objective: given a long DNA sequence S , find fixed length (10 - 200 bp) sub-sequences $S_i \subset S$ with exceptional mechanical properties.

Application of *cgDNAloc*: locating mechanically exceptional DNA sequences, made precise

Objective: given a long DNA sequence S , find fixed length (10 - 200 bp) sub-sequences $S_i \subset S$ with exceptional mechanical properties.

- In the sense for each $S_i \subset S$, compute $KL(\rho_{loc}^i, \rho_{bav})$, where

Application of *cgDNA/loc*: locating mechanically exceptional DNA sequences, made precise

Objective: given a long DNA sequence S , find fixed length (10 - 200 bp) sub-sequences $S_i \subset S$ with exceptional mechanical properties.

- In the sense for each $S_i \subset S$, compute $KL(\rho_{loc}^i, \rho_{bav})$, where
 - ρ_{loc}^i is the cgDNA marginal pdf of $S_i \subset S$,

Application of *cgDNA/loc*: locating mechanically exceptional DNA sequences, made precise

Objective: given a long DNA sequence S , find fixed length (10 - 200 bp) sub-sequences $S_i \subset S$ with exceptional mechanical properties.

- In the sense for each $S_i \subset S$, compute $KL(\rho_{loc}^i, \rho_{bav})$, where
 - ρ_{loc}^i is the cgDNA marginal pdf of $S_i \subset S$,
 - ρ_{bav} is the reference (banded) **sequence averaged** cgDNA pdf

Application of *cgDNA/loc*: locating mechanically exceptional DNA sequences, made precise

Objective: given a long DNA sequence S , find fixed length (10 - 200 bp) sub-sequences $S_i \subset S$ with exceptional mechanical properties.

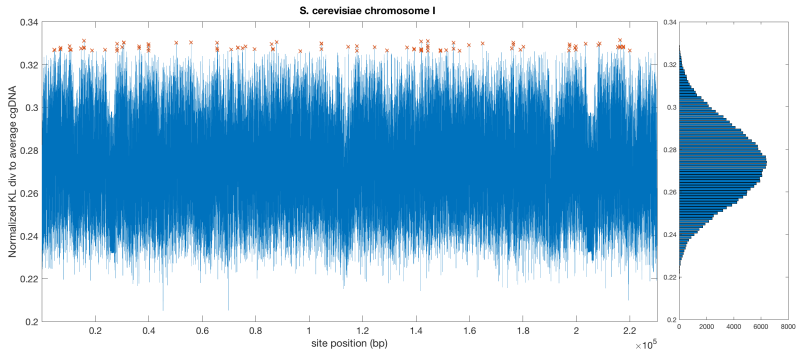
- In the sense for each $S_i \subset S$, compute $KL(\rho_{loc}^i, \rho_{bav})$, where
 - ρ_{loc}^i is the cgDNA marginal pdf of $S_i \subset S$,
 - ρ_{bav} is the reference (banded) **sequence averaged** cgDNA pdf
- Select S_i with extreme high values of KL divergence.

Application of *cgDNA/loc*: locating mechanically exceptional DNA sequences, made precise

Objective: given a long DNA sequence S , find fixed length (10 - 200 bp) sub-sequences $S_i \subset S$ with exceptional mechanical properties.

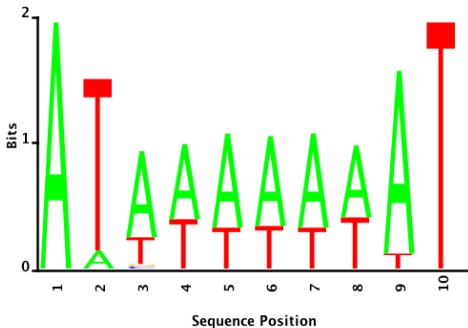
- In the sense for each $S_i \subset S$, compute $KL(\rho_{loc}^i, \rho_{bav})$, where
 - ρ_{loc}^i is the cgDNA marginal pdf of $S_i \subset S$,
 - ρ_{bav} is the reference (banded) **sequence averaged** cgDNA pdf
- Select S_i with extreme high values of KL divergence.
- Plot sequence logos of the outliers to see if there is a pattern.

S. cerevisiae genome, 10bp sites



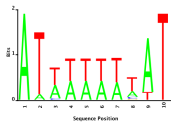
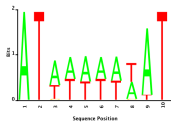
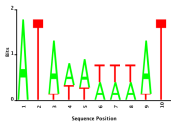
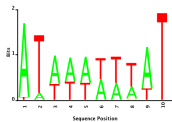
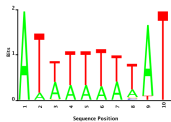
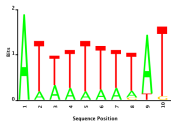
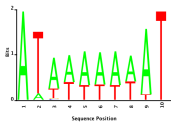
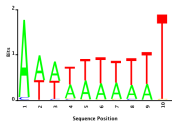
S. cerevisiae chr II, 10bp sites outliers

Sequence logos of high outlier sites (3σ or $KL \gtrsim 0.33$):

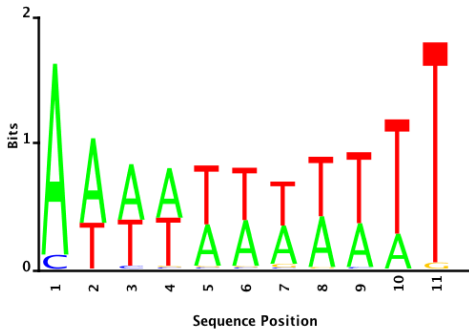


(Sequence logos show base composition frequencies at each location with height weighted by information content. Sometimes straight frequencies clearer.)

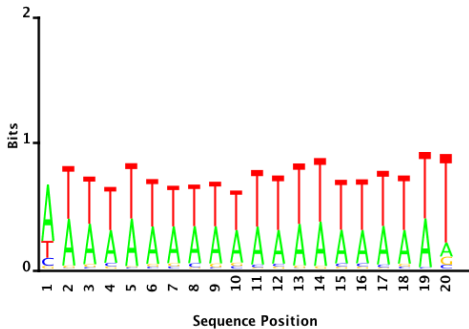
S. cerevisiae chr I-VIII, 10bp sites outliers



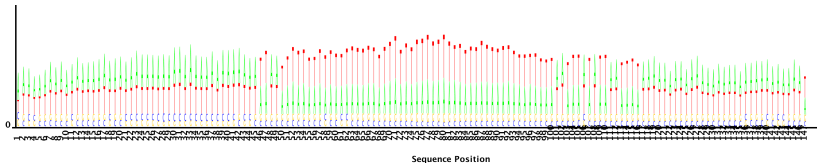
S. cerevisiae chr I, 11bp sites



S. cerevisiae chr I, 20bp sites

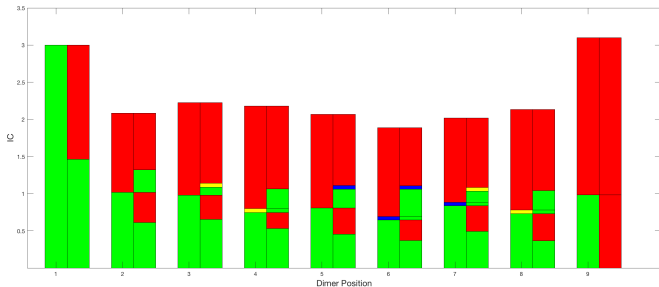


S. cerevisiae chr III, 147bp sites



S. cerevisiae chr I, 10bp sites outliers

A step further: look at dinucleotide sequence logos (because *cgDNA* parameters are dimer-dependent)

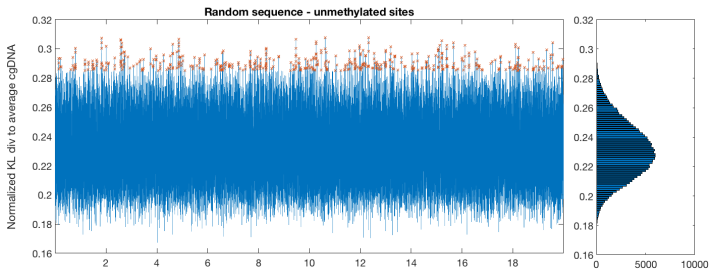


Green = A, Blue = C, Yellow = G, Red = T

Reveals strong preference for AA/TT steps.

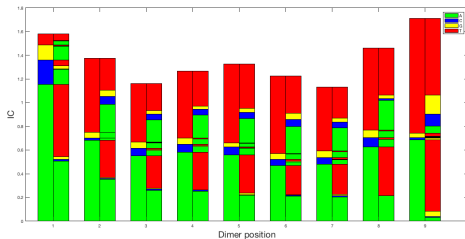
Are biological chromosomes very different from random 'chromosomes' ?

Analogous plots for 200K bp random sequence.



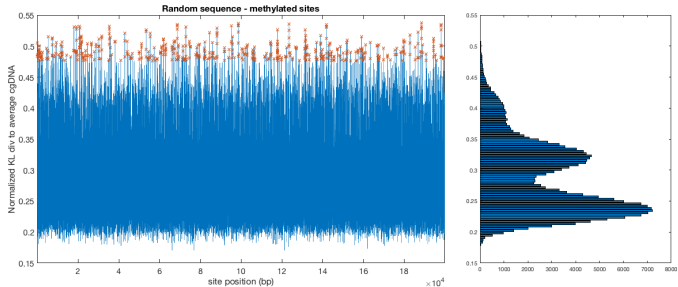
Outliers noticeably lower $KL \gtrsim 0.29$

Random sequence outlier dinucleotide sequence logo

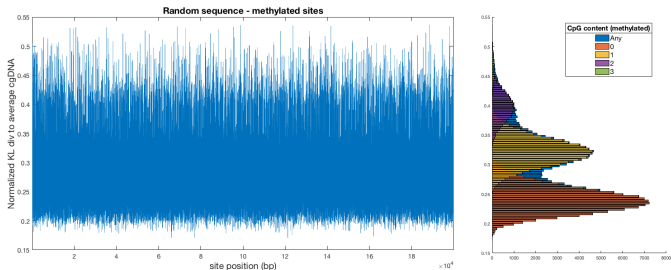


What about epigenetics?

- Keep average PDF over **unmethylated** 200K bp random sequence.
- Then **all** CpG steps are methylated
- outliers in sliding window at 99.7 percentile, $KL \gtrsim 0.47$

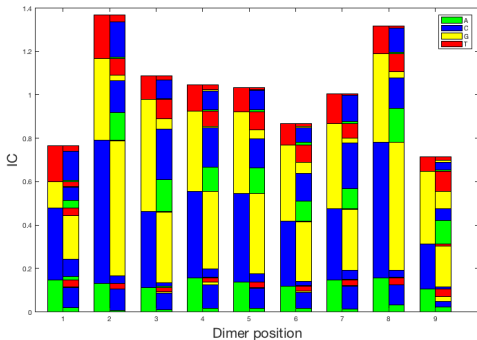


Methylated CpG 10bp sites in random 'chromosome'



Each methylated CpG step increases 'distance' to average.

Random chr, 10bp methylated sites outliers-sequence logo



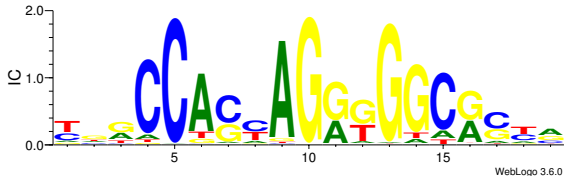
Outlying sequence switches from being AA/AT rich to being methylated CpG rich

Application II: Probing consensus protein binding sites (a first stab)

- Instead of searching sites far from average, can look for sites that are **close** (mechanically) to a particular sequence motif

Application II: Probing consensus protein binding sites (a first stab)

- Instead of searching sites far from average, can look for sites that are **close** (mechanically) to a particular sequence motif
- Example: CTCF transcription factor binding sites from JASPAR² data base



Application II: Probing consensus protein binding sites (a first stab)

- 930 fragments (≈ 200 bp), each containing one CTCF binding site

acttgctgtaagcatggggggggggttgatggtgggttttatgtaCTACAGGTGGTcaagtattatggtaccgtacaataattcatggtggctggcagtaatgtacgaaataca
tagcggttgtgatgggtgagcaataactgggtgtaaccaaatctgctcccatgaaagaacagagaatagtta

atgactgctgtaagggtgggtaggttgtggtatcctagtgggtgagggggtggttggagttgcagttgatggtgatagttgaaggttgattgctgactgctgtaagcatgggggggggggt
ttgatggtgggtgggttttatgtaCTACAGGTGGTcaagtattatggtaccgtacaataattcatggtggc

ggatgacccccctcagataggggtccctgaccacatcctccgtgaaatcaataatcccgCACAAAGAGTGCTactctcctcgctccgggccataaacacttgggggtagctaaagtga
actgtatccgacatctggttctactctcagggccataaagcctaaatagccccacgcttccctaaataagacatcacgatg

atggcttccatgaggtcctaagctccccaaagctgggtcaacctcctcgctacttggcccttctccagttgtcctgggcactgatggtCACCAGGTGGCgatgtgccaggaggac
gctctgtgccatctatgccaanaatcaacactacactcttccctcctcaggggtgcaggaatccggtccagaccag

aatccagggcaccgagaaagtccagcagccctgagcctgctgacttgcaatgacagggactggcCGGGAGATGGCagccctgtcccaggtcgggtgggtcctcagccgcag
ttcgtgcacggccatcgggaggcagtcaggaacactcactgatgtagcaactcagcctaaaagcagttttattataatacat

Application II: Probing consensus protein binding sites (a first stab)

- 930 fragments (≈ 200 bp), each containing one CTCF binding site

```
acttgctgtaagcatgggggggggggttgatggtgggtttttatgtaCTACAGGTGGTcaagtattatggtaccgtacaataattcatggtggctggcagtaatgtacgaaataca  
tagcgggtggtgaggggagcaataactgggggtaacccaaatctgctcccatgaaagaacagagaatagttta
```

```
tatgactgtaaggggggtaggtttgtggtatcctagtggggagggggtggaagtgcagttgatggtgatagttgaaggttgattgctgactgctgtaagcatggggggggggg  
ttgatggtgggtttttatgtaCTACAGGTGGTcaagtattatggtaccgtacaataattcatggtggc
```

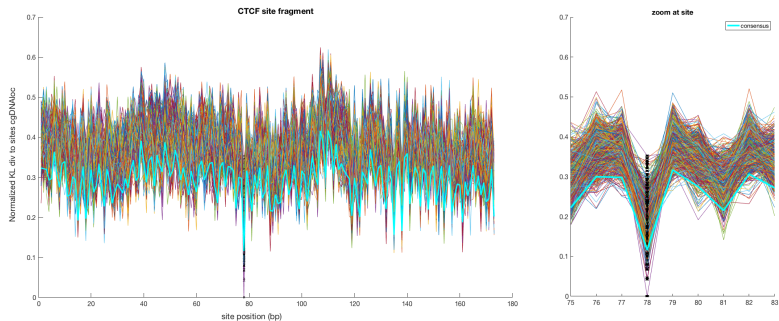
```
ggatgacccccctcagataggggtcccttgaccacatcctccgtaaatcaataatcccgCACAAAGAGTGCTactcctcctcctccgggccataaacacttgggggtagctaaagta  
acgtatccgacatctggttctactctcaggggccataaagcctaaatagccccacacgttccccctaaataagacatcacgatg
```

```
atggcttccatgaggtcctaagctccccaaagctgggtcaacctcctcgtactctgccccttctccaggttgcctgggcactgatggtCACCAGGTGGCgatgtgccaggaggac  
gctctgtgccatctatgcccataaacactacactctcctcctcctcctcaggggtgcaggaatccggtccagaccag
```

```
aatccagggcaccagagaaagtccagcagccctgagcctgctcacttgcaagtacagggactggcCGGGAGATGGCagccctgtcccaggtcgggtgggtccctcagccgcag  
ttcctgcacggccatcgggaggcagtcaggaacactcactgatgtagcaactcagcctaaaagcagttttattataatacat
```

- Use pdf averaging over **binding site sequences** to construct a 'consensus' *cgDNA*/loc distribution for CTCF binding site.

Application II: Probing consensus protein binding sites (a first stab)



Plot of KL divergences obtained by scanning one of the 930 fragments with each of the *cgDNA/loc* PDFs for all of the 930 CTCF binding sites, plus the consensus averaged *cgDNA/loc*.

Not entirely convincing

Perhaps potentials in the MD simulations not sufficiently accurate, or presence of multi-valent ions important, or ...

Can check whether *cgDNA* predictions of shape and stiffness are close enough to MD predictions. Can do better by making the coarse grain model less coarse. *cgDNA+* a rigid-base plus rigid-phosphates model (Patelli thesis, 2019)

Current conjecture is that asking that K-L divergence is small is too strong a condition. Should instead try to identify the features in common between binding site PDFs and non-binding site PDFs. Machine learning?

Picking level of coarse-graining and estimating parameters

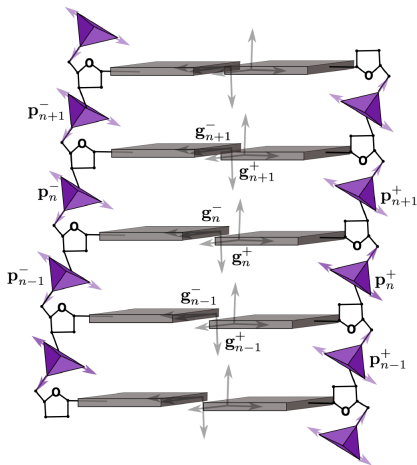
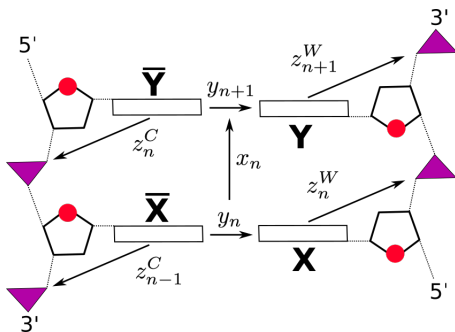


Figure 8.2 – Schematic representation of two the interacting strands representation of double stranded DNA. The sugar ring is shown but is not explicitly modelled.

Coordinates for the base-phosphate group interaction



- For an interior dimer step:

$$\left(z_{n-1}^C, y_n, z_n^W, x_n, z_n^C, y_{n+1}, z_{n+1}^W \right) \in \mathbb{R}^{42}$$

- For the 5'-end dimer step: $\left(y_1, z_1^W, x_1, z_1^C, y_2, z_2^W \right) \in \mathbb{R}^{36}$

Pattern of the inverse covariance (stiffness matrix) (1/3)

Rigid basepair degrees of freedom

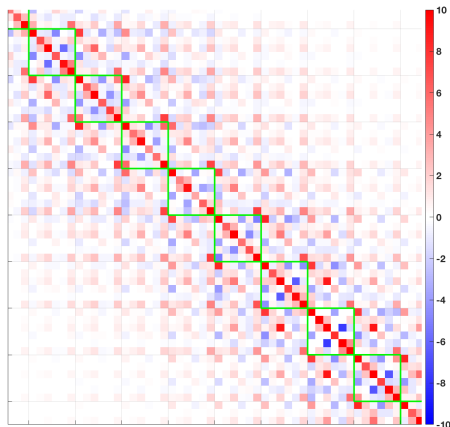


Figure: Inverse covariance observed from MD simulation.

Pattern of the inverse covariance (stiffness matrix) (2/3)

Rigid base degrees of freedom

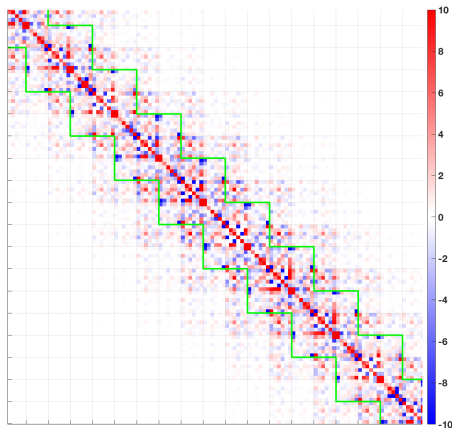


Figure: Inverse covariance observed from MD simulation.

Pattern of the inverse covariance (stiffness matrix) (3/3)

Rigid base and rigid phosphate degrees of freedom

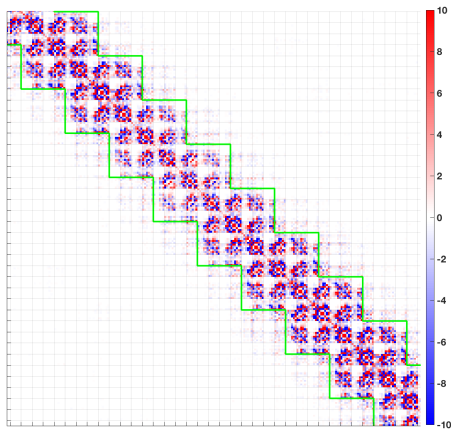


Figure: Inverse covariance observed from MD simulation.

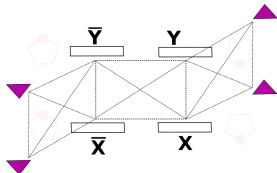
The cgDNA+ dinucleotide model (1/2)

Ten independent interior dinucleotide cross-junction interaction energies

$$U_I(w_I) = \frac{1}{2}(w_I - \mu^{XY}) \cdot K^{XY}(w_I - \mu^{XY}),$$
$$\mu^{XY} \in \mathbb{R}^{42}, K^{XY} = [K^{XY}]^T > 0 \in \mathbb{R}^{42 \times 42}$$

Sixteen independent 5'-end dinucleotide cross-junction interaction energies

$$U_E(w_E) = \frac{1}{2}(w_E - \mu^{5'XY}) \cdot K^{5'XY}(w_E - \mu^{5'XY}),$$
$$\mu^{5'XY} \in \mathbb{R}^{36}, K^{5'XY} = [K^{5'XY}]^T > 0 \in \mathbb{R}^{36 \times 36}$$



The cgDNA+ dinucleotide model (2/2)

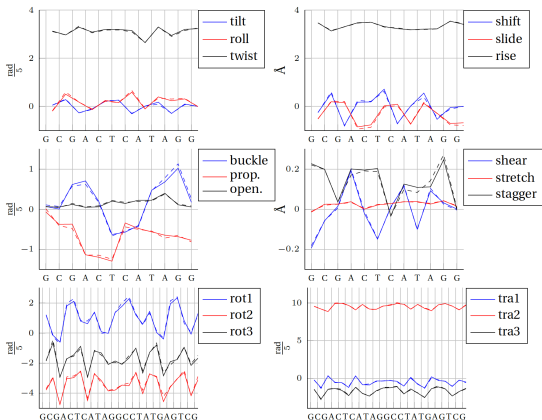
By summing the dinucleotide contributions over all junctions along a DNA fragment of sequence S , we obtain the quadratic free energy

$$U(w; S, \mathcal{P}) = \frac{1}{2}(w - \mu) \cdot K(w - \mu) + \text{const},$$
$$\mu = \mu(S, \mathcal{P}) \in \mathbb{R}^N, \quad K = K^T = K(S, \mathcal{P}) > 0 \in \mathbb{R}^{N \times N}$$

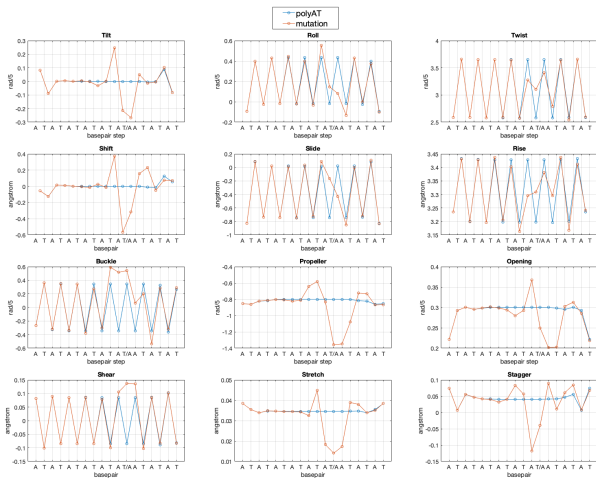
where \mathcal{P} is the cgDNA+ model parameter set and $N = 24n - 18$. Thus for any sequence the cgDNA+ model reconstructs the following Gaussian PDF:

$$\rho(w; S, \mathcal{P}) = \frac{1}{Z} \exp \{-\beta U(w; S, \mathcal{P})\}$$

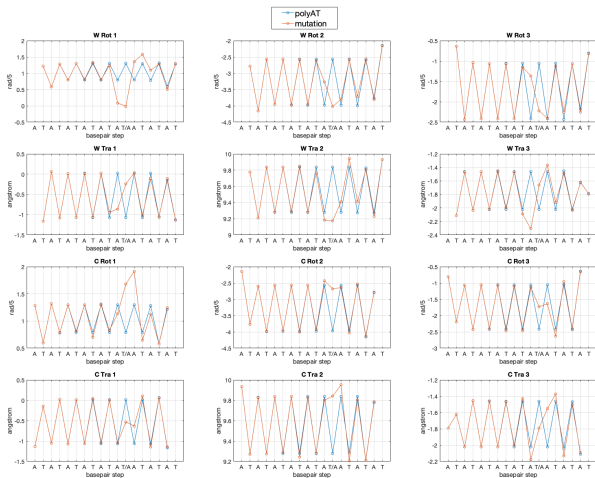
Now the fit to MD predictions of ground states (and stiffness) is spectacularly good e.g. ($3\mu\text{s}$ palindromic data)



And nonlocality is strong (intra/inter coordinates)



And nonlocality is strong (phosphate coordinates)

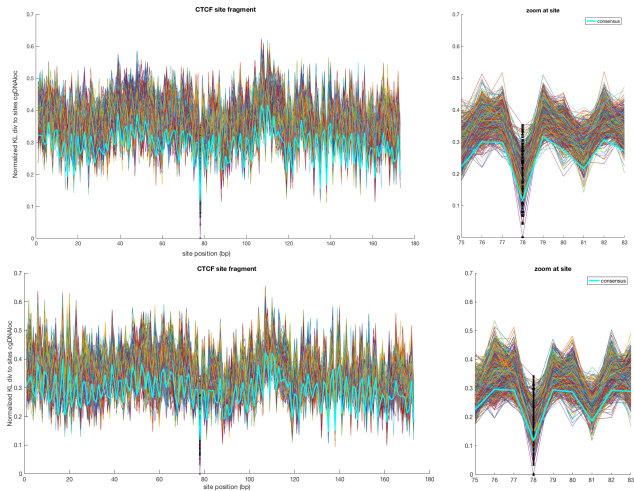


Can repeat all of above analyses

And signals are all equally or more strong, and conclusions are unaltered, except . . .

CTCF binding site alignment signal remarkably unchanged.

Top cgDNA, bottom cgDNA+



How to get a cgDNA+ parameter set

We need data ...

Our parameter set is trained on MD simulation using AMBER and the parambsc1 force field and the ABC protocol. Our training set (so far) is composed by 16 palindromic 24 basepair long oligomers each simulated for $3\mu s$.

How to get a cgDNA+ parameter set

We need data ...

Our parameter set is trained on MD simulation using AMBER and the parmbsc1 force field and the ABC protocol. Our training set (so far) is composed by 16 palindromic 24 basepair long oligomers each simulated for $3\mu s$.

We need good numerics ...

The parameter set is the solution of a nonlinear optimization problem involving Kullback-Leibler divergence in approximately 15K dimensions (cgDNA 'only' 1.5K dimensions)

How to get a cgDNA+ parameter set

We need data ...

Our parameter set is trained on MD simulation using AMBER and the parambsc1 force field and the ABC protocol. Our training set (so far) is composed by 16 palindromic 24 basepair long oligomers each simulated for $3\mu s$.

We need good numerics ...

The parameter set is the solution of a nonlinear optimization problem involving Kullback-Leibler divergence in approximately 15K dimensions (cgDNA 'only' 1.5K dimensions)

We need Fisher Information matrix ...

We use a gradient flow pre-conditioned by an inverse Hessian of the K-L divergence, ie the Fisher information evaluated at the training set data. That is a quasi-Newton method pre-conditioned with the (generalised) inverse of the Fisher information, that is evaluated only once. Converges in about an hour on a desktop machine.

Error from MD in shape: cgDNA vs cgDNA+

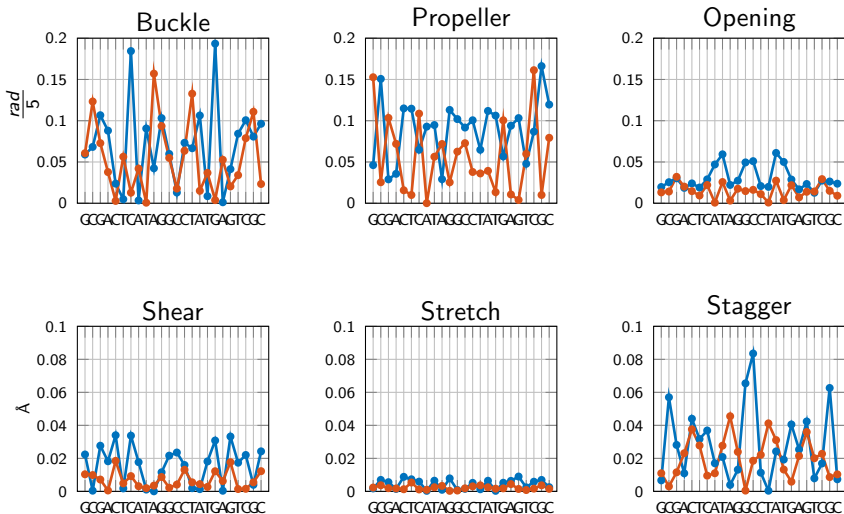


Figure: ■ : $|MD - cgDNA|$ ■ : $|MD - cgDNA+|$

Error from MD in shape: cgDNA vs cgDNA+

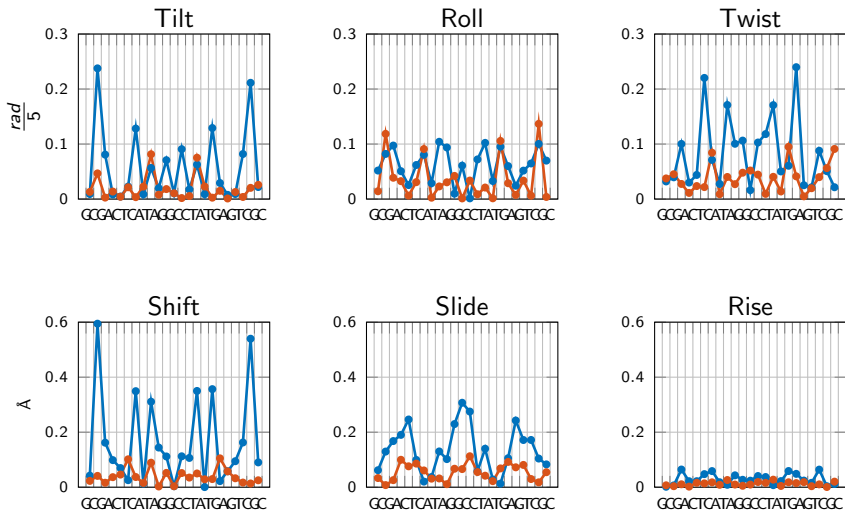


Figure: ■ : $|MD - cgDNA|$ ■ : $|MD - cgDNA+|$

Conclusions specific to coarse grain models of DNA

- Now have a hierarchy of coarse-grain models: rigid base pair, rigid base, rigid base plus rigid phosphate
- The more fine grain, the more banded are the observed precision matrices.
- The more fine grain, the (much) larger the associated parameter set
- Provided a Fisher Information pre-conditioner is used can still fit large *cgDNA+* parameter sets.
- Taking marginals goes down the chain, and sequence dependence becomes less localised—marginals of banded Gaussians are not banded.
- Currently considering Bayesian and Akaike Information Criteria for model selection, how fine grain is justified.

General and Classic Mathematical question

How to best estimate the observed ensemble mean and covariance (\hat{w}, C) by a mean and structured covariance (\hat{w}_s, C_s) where the stiffness matrix $K_s := C_s^{-1}$ is restricted to have a prescribed sparsity pattern?

Introduce the notation that $[[\cdot]]$ is a symmetric index set of a square matrix, and $[[\cdot]]'$ is the complementary index set.

And then write $[[K_s]]' = 0$ to indicate that the non vanishing entries of K_s are constrained to be contained in the index set $[[\cdot]]$.

Motivated by DNA application, particularly interested in cases where the index set $[[\cdot]]$ corresponds to overlapping diagonal sub-blocks, possibly of differing sizes.

Reminder: entropy and relative entropy of PDFs

The (negative of) the entropy of a PDF ρ is

$$\int \rho(w) \ln \rho(w) dw.$$

The relative entropy or Kullback-Leibler divergence (pre-distance) between two PDFs ρ_* and ρ_o is

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

For Gaussians there is the explicit formula (where $:$ is the usual matrix inner-product)

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

Approach I: Impose sparsity pattern via a maximum relative entropy fit

Take $\rho_o(w)$ to be the previously described maximum entropy/maximum likelihood Gaussian found by observing all means and all covariances from the ensemble $\{w_i\}$. Associated stiffness matrix K will in general be dense.

Then can impose a stiffness sparsity pattern by minimising the Kullback-Leibler divergence $D(\rho_*, \rho_o)$ over all Gaussian PDFs ρ_* with a banded stiffness matrices K_s (trivially $\hat{w}_s = \hat{w}$).

Approach I: Numerics (Gonzalez, Petkeviciute, jhm, J. Chem.Phys., 2013)

The K-L objective functional can be written as an explicit function of the entries in K_S . For our DNA examples approximately $4K$ unknowns per oligomer. Numerically very robust, both to optimisation algorithm and choice of initialisation, and we always observed convergence to a unique optimiser for each oligomer.

However did not consider any general mathematical existence or uniqueness result, in part because it is not evident that the maximum relative entropy fit between two Gaussians is the 'best' characterisation. Could use other measures of distance between two Gaussians, or another approach entirely.

Approach II: Enforcing a Prescribed Sparsity Pattern using Maximum Entropy (O. Gonzalez et al, Multiscale Model. Simul. (2017))

Find the PDF $\rho_m(w)$ of maximal entropy subject to the constraints of a) all of its first moments being prescribed, but now b) only some of its covariances $[[C]]$ being prescribed, where $[[\cdot]]$ is a given index set.

There are related literatures in both statistics (Dempster 1960s, and later Luenberger, Cover, ...) and matrix completion (Johnson et al 1980s, ...).

Approach II: Known results

When C is positive definite, and $[[\cdot]]$ includes all diagonal entries, then there is a unique entropy maximiser, the maximiser is Gaussian, and its stiffness matrix K_m satisfies the first order necessary (and also sufficient) conditions

$$[[K_m]]' = 0 \quad \text{and} \quad [[K_m^{-1}]] = [[C]].$$

If instead it is *assumed* that the desired distribution is a banded Gaussian, then you arrive at the same necessary conditions on its parameters starting from either Kullback-Leibler divergence, with the Gaussian with sample mean and covariance in the *first* argument, and the banded Gaussian to be fit as the second argument, or from Maximum Likelihood for the banded Gaussian directly from the ensemble of observations.

An explicit algorithm for computing the maximum absolute entropy fit in the special case that $[[\cdot]]$ is overlapping blocks

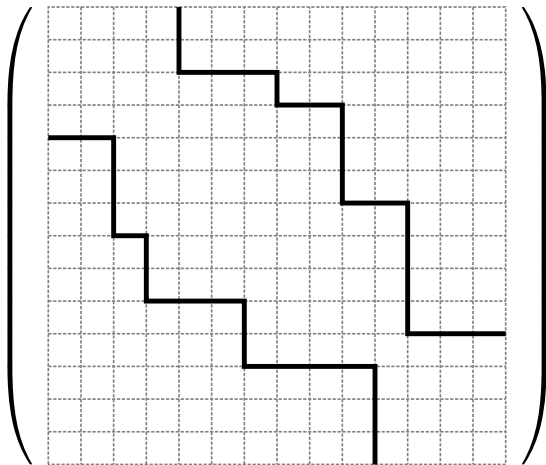
For general imposed sparsity patterns $[[\cdot]]$ not aware of efficient algorithms for high dimensional problems to be able to solve the first order conditions to find K_m .

$$[[K_m]]' = 0 \quad \text{and} \quad [[K_m^{-1}]] = [[C]].$$

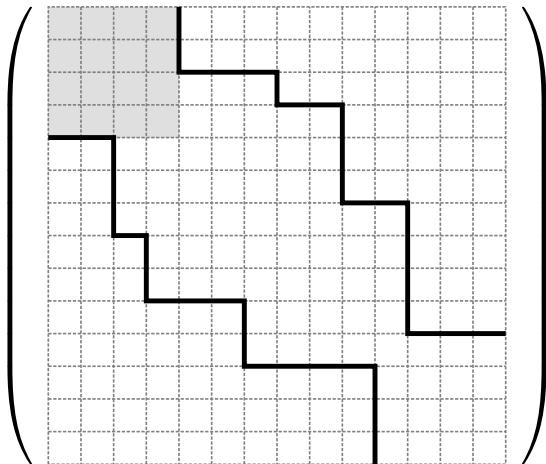
One easy case is when $[[\cdot]]$ is block diagonal, then everything decouples and K_m is block diagonal and the blocks of K_m are just the inverses of the sub-blocks of $[[C]]$

As part of his Phd work Glowacki found an analogous explicit algorithm to compute K_m for the particular class of *overlapping block* sparsity. Alas for us, the proof is same as appears in the book Graphical Models by S. Lauritzen, but result couched in very different language. Also a linear algebra result by Johnson et al.

An overlapping squares index set $[[\cdot]]$

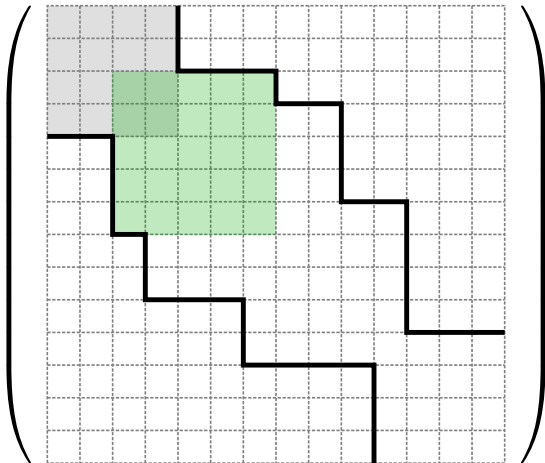


First pass: invert each sub-block of C and write to corresponding block in K



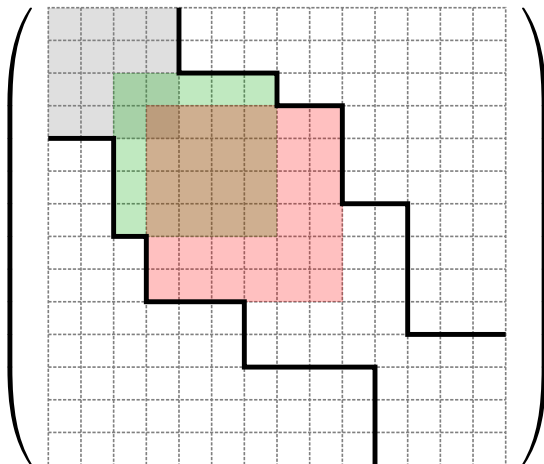
First pass: invert each sub-block of C and write to corresponding block in K

Add entries in overlap regions



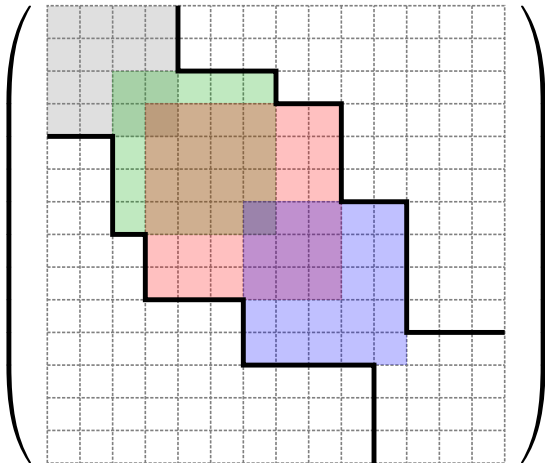
First pass: invert each sub-block of C and write to corresponding block in K

Add entries in overlap regions



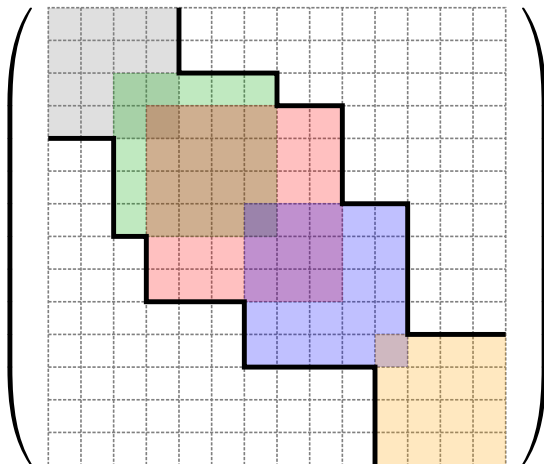
First pass: invert each sub-block of C and write to corresponding block in K

Add entries in overlap regions



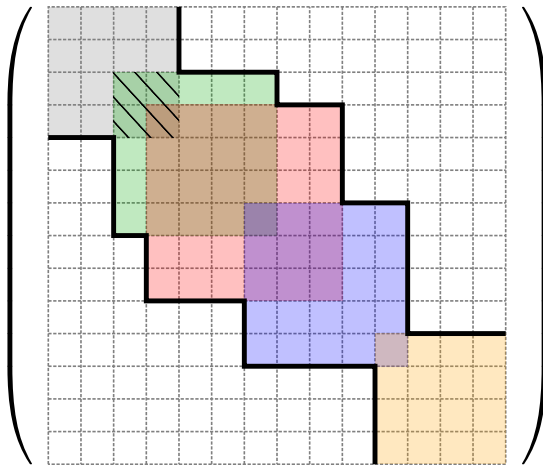
First pass: invert each sub-block of C and write to corresponding block in K

Add entries in overlap regions



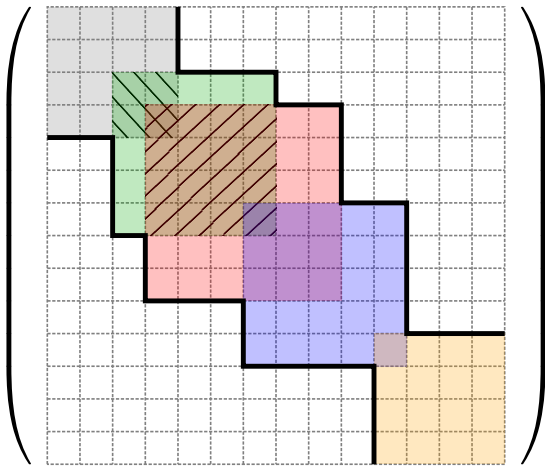
Second pass: correct for overlaps

Invert each overlap of adjacent sub-blocks of C and *subtract* from corresponding block in K



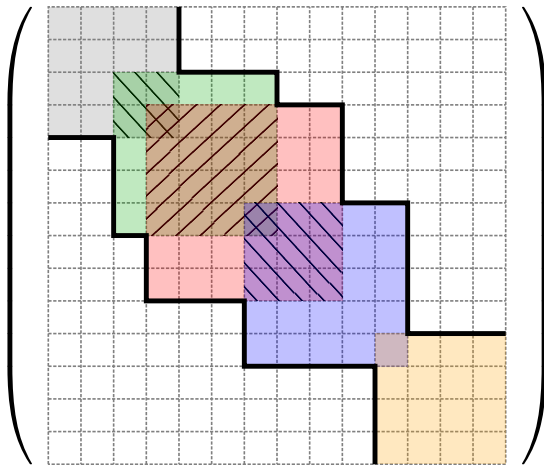
Second pass: correct for overlaps

Invert each overlap of *adjacent* sub-blocks of C and *subtract* from corresponding block in K



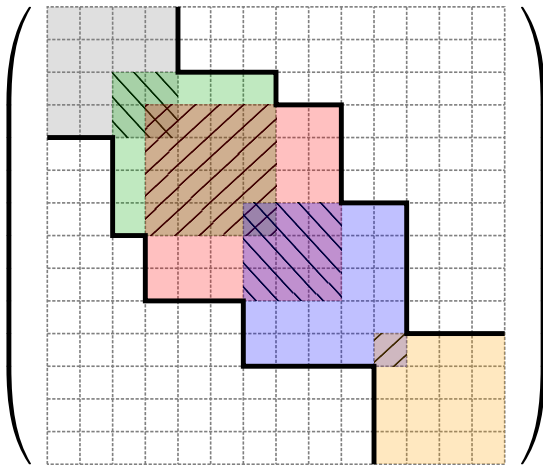
Second pass: correct for overlaps

Invert each overlap of adjacent sub-blocks of C and *subtract* from corresponding block in K



Second pass: correct for overlaps

Invert each overlap of adjacent sub-blocks of C and *subtract* from corresponding block in K

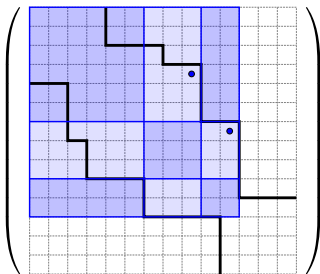


Proof

Is easy once you formulate the result. Schur complement formulas for inverse of 2×2 block partitioned symmetric matrix, combined with induction.

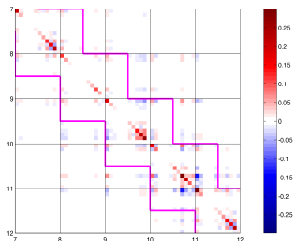
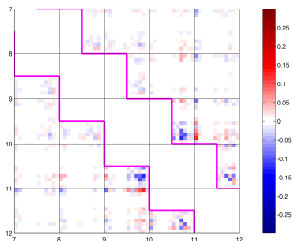
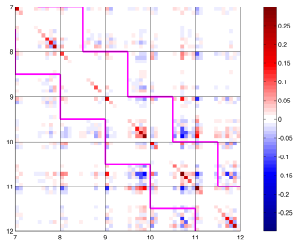
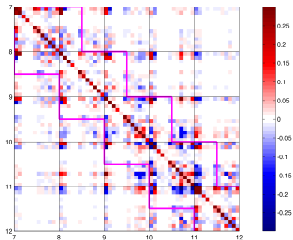
Seems like result should be more widely known.

$$[[K_m^{-1}]]' \neq [[C]]'$$



The condition on a symmetric matrix C such that its inverse satisfies $[[C^{-1}]]' = 0$ is that the off-diagonal blocks are minimal possible rank (generalises a result of G. Strang for the tridiagonal case). Can accordingly recursively modify entries of C outside the stencil $[[\cdot]]$, but not as simple as algorithm to find K .

Covariances: MD vs. Maximum Entropy vs. Maximum Relative Entropy



Stiffnesses: MD vs. Maximum Entropy vs. Maximum Relative Entropy

