DNA and Big Data: 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

CIS, Get to know your neighbours

July 5, 2021
The Natural Philosophy of Genomes: Statistical Mechanics meets Bioinformatics

- Major scientific accomplishments of the 21st century (so far) include publication of whole genome DNA sequences:
  - Individual genomes (personalised medicine, ongoing)
  - Epigenetics, modified base alphabets, and tissue specific epigenomes (ongoing)
- Realm of bioinformatics (lists of billions of letters), and annotated bioinformatics: which (sub-)sequences are responsible for what and when. An amazing amount of information is known.
- But HOW do changes in sequence effect changes in function?
- JHM (and many others) of the view that multi-scale mathematical modelling of the statistical mechanics of dsDNA needed to gain better understanding of how things actually work in molecular biology.
The Natural Philosophy of Genomes:
Statistical Mechanics meets Bioinformatics

- Major scientific accomplishments of the 21st century (so far) include publication of whole genome DNA sequences:
  - Individual genomes (personalised medicine, ongoing)
  - Epigenetics, modified base alphabets, and tissue specific epigenomes (ongoing)

- Realm of bioinformatics (lists of billions of letters), and annotated bioinformatics: which (sub-) sequences are responsible for what and when. An amazing amount of information is known.

- But HOW do changes in sequence effect changes in function?

- JHM (and many others) of the view that multi-scale mathematical modelling of the statistical mechanics of dsDNA needed to gain better understanding of how things actually work in molecular biology.
The Natural Philosophy of Genomes: Statistical Mechanics meets Bioinformatics

- Major scientific accomplishments of the 21st century (so far) include publication of whole genome DNA sequences:
  - Individual genomes (personalised medicine, ongoing)

- But HOW do changes in sequence effect changes in function?
- JHM (and many others) of the view that multi-scale mathematical modelling of the statistical mechanics of dsDNA needed to gain better understanding of how things actually work in molecular biology.
The Natural Philosophy of Genomes: Statistical Mechanics meets Bioinformatics

- Major scientific accomplishments of the 21st century (so far) include publication of whole genome DNA sequences:
  - Individual genomes (personalised medicine, ongoing)
  - Epigenetics, modified base alphabets, and tissue specific epigenomes (ongoing)

- But how do changes in sequence effect changes in function?
- JHM (and many others) of the view that multi-scale mathematical modelling of the statistical mechanics of dsDNA needed to gain better understanding of how things actually work in molecular biology.
The Natural Philosophy of Genomes: Statistical Mechanics meets Bioinformatics

- Major scientific accomplishments of the 21st century (so far) include publication of whole genome DNA sequences:
  - Individual genomes (personalised medicine, ongoing)
  - Epigenetics, modified base alphabets, and tissue specific epigenomes (ongoing)
- Realm of bioinformatics (lists of billions of letters), and annotated bioinformatics: which (sub-) sequences are responsible for what and when. An amazing amount of information is known.
- But HOW do changes in sequence effect changes in function?
- JHM (and many others) of the view that multi-scale mathematical modelling of the statistical mechanics of dsDNA needed to gain better understanding of how things actually work in molecular biology.
The Natural Philosophy of Genomes: Statistical Mechanics meets Bioinformatics

- Major scientific accomplishments of the 21st century (so far) include publication of whole genome DNA sequences:
  - Individual genomes (personalised medicine, ongoing)
  - Epigenetics, modified base alphabets, and tissue specific epigenomes (ongoing)
- Realm of bioinformatics (lists of billions of letters), and annotated bioinformatics: which (sub-) sequences are responsible for what and when. An amazing amount of information is known.
- But **HOW** do changes in sequence effect changes in function?
The Natural Philosophy of Genomes: Statistical Mechanics meets Bioinformatics

- Major scientific accomplishments of the 21st century (so far) include publication of whole genome DNA sequences:
  - Individual genomes (personalised medicine, ongoing)
  - Epigenetics, modified base alphabets, and tissue specific epigenomes (ongoing)

- Realm of bioinformatics (lists of billions of letters), and annotated bioinformatics: which (sub-) sequences are responsible for what and when. An amazing amount of information is known.

- But HOW do changes in sequence effect changes in function?

- JHM (and many others) of the view that multi-scale mathematical modelling of the statistical mechanics of dsDNA needed to gain better understanding of how things actually work in molecular biology.
Talk Summary: Big Data I

- Pattern/feature recognition in ensembles (millions or billions of members) of multivariate (banded in stiffness/precision matrix) Gaussian PDFs each in dimension $N$ ($N$ of the order a few hundred to a few thousand, band width 42)

- The ensembles of Gaussians generated from a tool called cgDNAloc that can be used to scan a short window over genomic length scales to identify short sub-sequences with exceptional (statistical) mechanical properties. Enhance bioinformatics with mechanics.

- cgDNAloc is a marginal of a coarse-grain equilibrium statistical mechanics model of dsDNA called cgDNA+.
Talk Summary: Big Data I

- Pattern/feature recognition in ensembles (millions or billions of members) of multivariate (banded in stiffness/precision matrix) Gaussian PDFs each in dimension $N$ ($N$ of the order a few hundred to a few thousand, band width 42)
- The ensembles of Gaussians generated from a tool called $cgDNAloc$ that can be used to scan a short window over genomic length scales to identify short sub-sequences with exceptional (statistical) mechanical properties. Enhance bioinformatics with mechanics.
Talk Summary: Big Data I

- Pattern/feature recognition in ensembles (millions or billions of members) of multivariate (banded in stiffness/precision matrix) Gaussian PDFs each in dimension $N$ ($N$ of the order a few hundred to a few thousand, band width 42)

- The ensembles of Gaussians generated from a tool called $cgDNAloc$ that can be used to scan a short window over genomic length scales to identify short sub-sequences with exceptional (statistical) mechanical properties. Enhance bioinformatics with mechanics.

- $cgDNAloc$ is a marginal of a coarse-grain equilibrium statistical mechanics model of dsDNA called $cgDNA+$
Talk Summary: Big Data II

• The sequence-dependent \( \text{cgDNA}^+ \) family of coarse-grain models have a machine learning flavour in that they have 20K+ parameters to be fit.
Talk Summary: Big Data II

- The sequence-dependent $\textit{cgDNA}^+$ family of coarse-grain models have a machine learning flavour in that they have $20\text{K}^+$ parameters to be fit.
- But have to predict $4^n$ sequences with $n$ of order a few 10s. ($n = 10$ is $\approx 500\text{K}$ independent sequences, in four letter alphabet.)
• The sequence-dependent cgDNA+ family of coarse-grain models have a machine learning flavour in that they have 20K+ parameters to be fit.

• But have to predict $4^n$ sequences with $n$ of order a few 10s. ($n = 10$ is $\approx 500K$ independent sequences, in four letter alphabet.)

• cgDNA+ has dinucleotide-step sequence-dependent parameter sets that predict an equilibrium distribution in the form of a banded Gaussian for any sequence.
The sequence-dependent \( cgDNA^+ \) family of coarse-grain models have a machine learning flavour in that they have 20K+ parameters to be fit.

But have to predict \( 4^n \) sequences with \( n \) of order a few 10s. \((n = 10 \text{ is } \approx 500K \text{ independent sequences, in four letter alphabet.})\)

\( cgDNA^+ \) has dinucleotide-step sequence-dependent parameter sets that predict an equilibrium distribution in the form of a banded Gaussian for any sequence.

The parameter sets are estimated from a small (currently 16 member) training library of long duration (10 \( \mu \text{s} \) simulations or \( 10^7 \) snapshots saved from \( 10^{10} \) time steps), atomistic (i.e. fine grain) MD simulations (in dimension \( \approx 50K \)) of short (currently 24bp palindromes) DNA fragments. One library (corresponding to one MD simulation protocol) is \( 1.6 \times 10^8 \) snapshots in dimension 558/50K (corresponding to dry/wet).
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
$\rho$ probability density function
$U$ free energy
$Z, \beta$ constants
The *cgDNA* family of models

Article and software download:
http://lcvmwww.epfl.ch/cgDNA/

- Original *cgDNA* model and applications jhm with O. Gonzalez, D. Petkevicuite, M. Pasi, J. Glowacki, A. Grandchamp, J. Mitchell
The *cgDNA* family of models

Article and software download:
http://lcvmwww.epfl.ch/cgDNA/

- Original *cgDNA* model and applications jhm with O. Gonzalez, D. Petkevicuïte, M. Pasi, J. Glowacki, A. Grandchamp, J. Mitchell

- Plus A. Patelli, and R. Sharma for *cgDNA+*
The *cgDNA* family of models

Article and software download:
http://lcvmwww.epfl.ch/cgDNA/

- Original *cgDNA* model and applications jhm with O. Gonzalez, D. Petkevicuīte, M. Pasi, J. Glowacki, A. Grandchamp, J. Mitchell
- Plus A. Patelli, and R. Sharma for *cgDNA+*
- And T. Zwahlen for *cgDNAloc* and applications
The \textit{cgDNA} family of models

Article and software download:
http://lcvmwww.epfl.ch/cgDNA/

- Original \textit{cgDNA} model and applications jhm with O. Gonzalez, D. Petkevicuite, M. Pasi, J. Glowacki, A. Grandchamp, J. Mitchell
- Plus A. Patelli, and R. Sharma for \textit{cgDNA}+
- And T. Zwahlen for \textit{cgDNAloc} and applications

Web interface http://cgDNAweb.epfl.ch

The *cgDNA* family of models

Article and software download:
http://lcvmwww.epfl.ch/cgDNA/

- Original *cgDNA* model and applications jhm with O. Gonzalez, D. Petkevicu
te, M. Pasi, J. Glowacki, A. Grandchamp, J. Mitchell
- Plus A. Patelli, and R. Sharma for *cgDNA+*
- And T. Zwahlen for *cgDNAloc* and applications

Web interface http://cgDNAnet.epfl.ch

- *cgDNA+* Matlab and Python scripts strictly compatible with cgDNAnet (Patelli, Zwahlen, Sharma).
Coarse grain configuration variables

cgDNA+ is a coarse-grain model in which each base and each phosphate group is explicitly described as a distinct rigid body.

\[ X \in \{ T, A, C, G \}, \quad a = 1 \ldots n \]

\[ \text{A = T, } \bar{T} = A, \quad \text{C = G, } \bar{G} = C \]
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 \( 12n - 6 \) DOF.

\[
(x_a, \eta_a) =: y_a \in \mathbb{R}^6 \text{ intra bp} \quad \quad (v_a, u_a) =: z_a \in \mathbb{R}^6 \text{ inter bp}
\]

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.
Plus coordinates for the base-phosphate group interaction

- For an interior dimer step:
  \[
  (z_{n-1}^C, y_n, z_n^W, x_n, z_n^C, y_{n+1}, z_{n+1}^W) \in \mathbb{R}^{42}
  \]

- For the 5'-end dimer step:
  \[
  (y_1, z_1^W, x_1, z_1^C, y_2, z_2^W) \in \mathbb{R}^{36}
  \]
The *cgDNA+* model

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

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

\[ \mu = \mu(S, P) \in \mathbb{R}^N \quad \text{ground-state configuration} \quad (N = 24n - 18) \]

\[ K = K(S, P) \in \mathbb{R}^{N \times N} \quad \text{banded stiffness matrix} \quad K = K^T > 0. \]

where the parameter set \( P \) allows explicit construction of \( \mu(S, P) \) and \( K(S, P) \) for oligomers of arbitrary sequences \( S \) in the given alphabet, and of an arbitrary length \( n \) (i.e. the number of base-pair levels).
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 coarse-graining context, first average “forces” $Aa$ and $Bb$, and then compute an effective ground state displacement $c$. 
The cgDNA+ dinucleotide step model I

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

Ten independent dinucleotide cross-junction interaction energies

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

\[
\mu_d^{XY} \in \mathbb{R}^{42}, \quad K_d^{XY} \in \mathbb{R}^{42 \times 42}
\]

\[
\mathbb{R}^{42} \ni w_d = (p_w^-, y_-, p_c^-, z, p_w^+, y_+, p_c^+)
\]

End blocks have different dimension and dimer-dependence \(K_d^{5'XY}\) and \(\sigma_d^{5'XY}\) (missing phosphate and end-fraying).

\(\bar{Y} \bar{X} 3'\) blocks implied by \(5'XY\) blocks via Crick-Watson reading symmetry.
The \textit{cgDNA$+$} dinucleotide step model II

Each dinucleotide-dependent parameter set $\mathcal{P}$ is of the form

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

Estimate $\mathcal{P}$ from a variational principle involving Kullback-Leibler divergence from the large-scale MD time-series estimates $\rho_j$ of oligomer-based, first and centred second moment statistics for each sequence $S_j$ in the training library

$$\mathcal{P} = \text{argmin} \sum_j KL(\rho_{cgDNA+}(S_j, \mathcal{P}), \rho_j)$$

This minimisation is in $20K+$ dimensions. There is a lot of detail, but a gradient flow pre-conditioned by an inverse Hessian of the K-L divergence, i.e. the Fisher information evaluated at the training set data, typically converges in about an hour on a desktop machine. Thus a quasi-Newton method pre-conditioned with the (generalised) inverse of the Fisher information, that is evaluated only once, is very efficient.
The \( cgDNA^+ \) dinucleotide step model III

Stiffness and shape reconstruction:

\[
K(S, P), \quad \sigma(S, P): \quad \mu(S, P) = K(S, P)^{-1}\sigma(S, P)
\]

Formula comes from summing over junction energies. The matrix inversion means that the oligomer ground state \( \mu \) has a nonlocal dependence on sequence. Linear algebra expression of frustration.
Now the fit to MD predictions of ground states (and stiffness) is spectacularly good e.g. 10μs simulation not in training set data.
And nonlocality is strong (intra/inter coordinates)
Is $\text{cgDNA}^+$ close to reality?

cgDNA$^+$ can be no more accurate than the underlying MD simulations, and MD simulations are themselves just a model. Quantitative direct checks against experiment are challenging. But reason to be optimistic, e.g. a third Big Data set is the PDB structures of Protein/dsDNA co-crystals. Comparing statistics of tetramer dependent ground state shapes gives close to identical clustering (with a Mahalonobis distance with sequence averaged stiffness).

In particular, in both dendograms derived from the two independent data sets, tetramer ground state shapes $\mu$ are clustered by their central dimer step, except for YR steps, where $Y$ denotes a Pyrimidine ($C$ or $T$, a one ring base) and $R$ denotes a Purine ($A$ or $G$, a two ring base).

Figure produced in collaboration with Sharma, W. Olson and L. Czapla.
Is \textit{cgDNA}+ close to reality?
Scanning chromosomes

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)

*Picture source:https://cnx.org/contents/9TxHOD3O@4/The-Nucleus-and-DNA-Replication*
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:

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

In fact the above tetramer dendograms already involved $cgDNAloc$ marginal groundstates in an average flanking sequence context.
Computing average PDFs

Given an ensemble \( \{\rho_i\} \) of PDFs with means and covariances \( \mu_i, K_i, i = 1, \ldots, m \), compute an average Gaussian pdf \( \rho_{av} \) with mean and (centred) covariance \( \mu_{av}, K_{av} \) obtained by averaging shapes and covariances:

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

or

\[
K_{av}^{-1} = K_H^{-1} + \Sigma_\mu
\]

where the two terms on the RHS are the inverse of the harmonic average stiffness, and the Shape Covariance matrix.

For ensembles of \textit{cgDNA}+ Gaussians the eigenvectors of \( K_H^{-1} \) and \( \Sigma_\mu \) always seem to be rather close. This need not be the case for ensembles of general Gaussians as the two quantities are independent.
The case of an average of banded Gaussians

As already remarked, the cgDNAloc marginal of a banded cgDNA+ Gaussian is always automatically banded. However the Gaussian average of banded Gaussians need not be banded.

For an ensemble of banded cgDNAloc PDFs \( \{ \rho_{\text{loc}}(S_i) \} \) (say along a chromosome), define a **banded cgDNAloc average** \( \rho_{\text{bav}} \) (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.
Two things to know: 1. Epigenetic base modifications

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

\[
\text{cytosine (C)} \quad \text{5-methylcytosine (mC)} \quad \text{5-hydroxymethylcytosine (hmC)}
\]

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 19bp CTCF transcription factor binding sites from JASPAR \(^1\) data base

Sequence logos show base composition frequencies at each location with height weighted by information content, or base 2 relative entropy, against base pair index. Sometimes straight frequencies clearer.

Exhaustive $k$-mers + 5bp random flanking, $k = 1, \ldots, 9$, but show $k = 4$ case.

Use metric or linear kernel PCA. Instead of diagonalising covariance matrix $\Sigma_{\mu}$, solve generalised eigenproblem

$$\Sigma_{\mu}x = \lambda K^{-1}_H x,$$

Corresponds to changing the local metric to $K_H$, then project (with standard inner product) onto generalized eigenvectors as principal components.

Use hierarchical, single-linkage algorithm to identify clusters.
Properties of Metric PCA

Metric PCA on shape covariance is invariant under linear change of scale.

k-mer metric PCA

- exhibits a spectrum gap after $k$th generalized eigenvalue
- yield $2^k$ clusters in projection onto $k$ principal components
- clusters match R/Y alphabet encoding

nothing true for standard k-mer PCA applied to shape covariance $\Sigma_\mu$. 
4-mer shapes
4-mer shapes
Cluster sequence logos
Application II of cgDNAloc: locating mechanically exceptional DNA sequences

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 $\text{KL}(\rho_i^{loc}, \rho_{bav})$, where
  • $\rho_i^{loc}$ is the cgDNA marginal pdf of $S_i \subset S$,
  • $\rho_{bav}$ 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.
Application II of \textit{cgDNAloc}: locating mechanically exceptional DNA sequences

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 \( KL(\rho^i_{loc}, \rho_{bav}) \), where
Application II of cgDNAloc: locating mechanically exceptional DNA sequences

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 $KL(\rho^i_{loc}, \rho_{bav})$, where
  - $\rho^i_{loc}$ is the cgDNA marginal pdf of $S_i \subset S$,
Application II of \textit{cgDNAloc}: locating mechanically exceptional DNA sequences

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 $KL(\rho^{i}_{loc}, \rho_{bav})$, where
  - $\rho^{i}_{loc}$ is the cgDNA marginal pdf of $S_i \subset S$,
  - $\rho_{bav}$ is the reference (banded) sequence averaged cgDNA pdf
Application II of cgDNAloc: locating mechanically exceptional DNA sequences

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 $KL(\rho^i_{loc}, \rho_{bav})$, where
  - $\rho^i_{loc}$ is the cgDNA marginal pdf of $S_i \subset S$,
  - $\rho_{bav}$ is the reference (banded) sequence averaged cgDNA pdf

- Select $S_i$ with extreme high values of KL divergence.
Application II of \textit{cgDNAloc}: locating mechanically exceptional DNA sequences

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 $KL(\rho_{\text{loc}}^i, \rho_{\text{bav}})$, where
  - $\rho_{\text{loc}}^i$ is the cgDNA marginal pdf of $S_i \subset S$,
  - $\rho_{\text{bav}}$ 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 window, 1bp steps
S. cerevisiae chr II, 10bp window outliers

Sequence logos of high outlier sites ($3\sigma$ or $KL \gtrapprox 0.33$):
S. cerevisiae chr I-VIII, 10bp window 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 \textit{cgDNA}+ parameters are dimer-dependent)

\textbf{Green} = A, \textbf{Blue} = C, \textbf{Yellow} = G, \textbf{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 $KL \approx 0.29$
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.
Outlying sequence switches from being AA/AT rich to being methylated CpG rich
Application III: 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

\[\text{WebLogo 3.6.0} \quad \begin{array}{cccc}
    & G & A & C \\
  0.0 & C & A & T \\
  1.0 & T & A & G \\
  2.0 & A & G & C \\
\end{array}\]

Application III: 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 JASPAR \(^2\) data base

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

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

```plaintext
acttgtcatgcaggggggggggtttgatggtgggtgtttttagtaCTACAGGTGTCaagtatttatgttacccgtcataatattcatgtggctggcagtaatgtaagaattaca
tgcggtttgtgatgggtagtcacaataactggtggtgtaaccqaatatctgctcccatagaagtaacagagaaatagttta
tatgtactgttaaggggtgggtaggttgtgtatccctagttgggtgaagggtgggttagctgggtccttgccagtggatgatgtgtataggtgtaaggtgtgatcgtactgtcgtgtaaggcatgggggggggtl
tgggtggtgggtgtttttagtaCTACAGGTGTCaagtatttatgttacccgtcataatattcatgtggc
ggatgaccoccccctccataaggggtcccttgagaccacctatcctcgtgaatatcaatctccgCACAAGATGCatctctctctcgcctccggccataaacacttgggggtagctaaagtga
ergtatgctcagcactggtttcctacttcagggccataaagcctaataaagccacaagtttcccttaataaagacatcaagatg
atggcttcctcataggcctcaagcctcctcccaagctggtctccctctccgtcctctctccgtctggttgctggcactgtggtCACCAGGTGCGatgtgctcccagggagagac
gctgctgtgcctctctcataagggccaaaactcacaactctcattctcctgctcaggggtcggggaatactgtcctcagaccag
aatacaggccacccqagaaagtccagcagcagccgtgagctctgctcaatgacagagactgacCGGGAGATGCGaccctgtccccaggtgggttgctcctcagccacag
ctcttgctacgcctgctgccaggggctgctcaggaacactcactgtgtgtgtagcaactcagccctaaaaagcagttttataatacat
```
Application III: Probing consensus protein binding sites (a first stab)

- **930 fragments (≈ 200 bp), each containing one CTCF binding site**
  
  - `actgctgtgaagctggtggggttttgattggttgggttgattgtaCTACAGGTGTcaagttttatagttgtaacgtaaattttctggtggtcgcagtaatgtacgaatataca`
  - `tagcgggtgtgattggtgagtcgtaactggttggttaacctaaatgcctccccatgaagacacagaagatgtaa`
  - `gtgagtggtggtgttattgtacCTACAGGTGTcaagttttatagttgtaacgtaaattttctggtggtcgcag`
  - `ggatgacccccctcagataagggtctccttgacaccacctctcctgcggaaatctataatcctccgCAACAAGAGTGCTactctctctccgggcctaaacactctggggtagctaaatgattaactgtatcgcactcgggccataaagcctaatagcctcccaacagttcccttailaaataagacacatcaacagtg`
  - `atatggtcttcgttagggctcatctgaagctcctccagcgccttcctgcgtctggtggtcgcagtaactggttggtgagtcgtaactggttggttaacctaaatgcctccccatgaagacacagaagatgtaa`
  - `ggtgagtggtggtgttattgtacCTACAGGTGTcaagttttatagttgtaacgtaaattttctggtggtcgcag`
  - `aatgccgcttcgcaacgcgagaatctcagcagcgctggtggtcgcagtaactggttggtgagtcgtaactggttggttaacctaaatgcctccccatgaagacacagaagatgtaa`

- **Use pdf averaging over binding site sequences to construct a ‘consensus’ cgDNAloc distribution for CTCF binding site.**
Application III: Probing consensus protein binding sites (a first stab)

Plot of KL divergences obtained by scanning one of the 930 fragments with each of the cgDNAloc PDFs for all of the 930 CTCF binding sites, plus the consensus averaged cgDNAloc.
Conclusion: encouraging but not entirely convincing

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

Current conjecture is that asking that K-L divergence to be small at a binding site is too strong a condition. Should instead try to identify the features in common between binding site PDFs compared to non-binding site PDFs. Currently in first stages of applying machine learning techniques to attempt this classification problem (with J. Quer, supported in Berlin by the Einstein Foundation).
Conclusions and Observations I

- The \textit{cgDNA\textdagger+} coarse grain model accurately predicts sequence-dependent equilibrium distributions of dsDNA when compared to data taken from MD simulation. Estimation of each \textit{cgDNA\textdagger+} parameter set involves a large scale computation of a small training library of short oligomers. But once a parameter set has been estimated the \textit{cgDNA\textdagger+} model itself easily allows predictions for millions of long sequences.

- MD simulations with monovalent counter ions are probably rather close to reality. The case of multivalent counter ions is less clear, and is (alas) biologically important.

- For comparing with bioinformatics data the local marginalized model \textit{cgDNAloc} is the appropriate tool, including versions where averaging is carried out over the flanking sequence to the window of the marginal.
Conclusions and Observations II

• *cgDNAloc* predicts $k$ dominant modes in a generalised PCA analysis of ground state shape within a window of length $k$ base pairs. This corresponds to a strong shape clustering in the purine-pyrimidine alphabet.

• *cgDNAloc* predicts that statistical mechanical outlying sequences in the four letter standard sequence alphabet are A,T rich.

• But epigenetically modified Cs completely overwhelm this affect, and the outliers correspond to the number of modified CpG steps

• Some hope for better understanding indirect read out and how binding site sequence affects affinities to transcription factor binding proteins.

• Still lots to do ...