Sequence effects on B-DNA

JOBIM 2015 - Clermont-Ferrand

Marco Pasi

Bioinformatics: Structures and Interactions
BMSSSI, IBCP, Lyon

Applied Maths Section
EPFL, Lausanne
Introduction

- DNA: a passive vector of information?

- Protein-DNA interactions: indirect readout & DNA allostery

- The “second genetic code”

- DNA counterion atmosphere

Lavery, Q. Rev. Biophys. 2004
Kim et al., Science 2013
Sobetzko et al., Mol. BioSyst. 2013
Manning, Q. Rev. Biophys. 1978
The Ascona B-DNA Consortium

The most prominent labs in the field of DNA simulations put their efforts together in 2002 to carry out molecular dynamics simulations on a broad enough spectrum of B-DNA sequences to understand the extent and nature of sequence effects on structure and structural fluctuations.

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)
- J. Sponer (Brno)
The ABC strategy

<table>
<thead>
<tr>
<th>39 x 18-mers</th>
<th>GC-CD-ABCD-ABCD-ABCD-GC</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCAGTTATGAGTGATGAGC</td>
<td>GCTACTATTATTATTAACTGC</td>
</tr>
<tr>
<td>GCCAAGAAAGGAAGGAGC</td>
<td>GCATCGTGCTGCTCGCTGC</td>
</tr>
<tr>
<td>GCCTAGCTAGCTAGCTGC</td>
<td>GCTGTAGGTAGGTAGGTGC</td>
</tr>
<tr>
<td>GCCGACGCGAGCGCAGGC</td>
<td>GCCTAGCTAGCTAGCTGC</td>
</tr>
<tr>
<td>GCTGAGTGATGAGTGAGC</td>
<td>GTCTAGCTACGTACGTAGC</td>
</tr>
<tr>
<td>GCTCAGTCACTAGTCACTGC</td>
<td>GCACGAACGACGACGACGC</td>
</tr>
<tr>
<td>GCCGTCGTCGTCGTCGTCGC</td>
<td>GCACGAACGACGACGACGC</td>
</tr>
<tr>
<td>GCCGCCGGCGGCGGCCGCG</td>
<td>GCACGAACGACGACGACGC</td>
</tr>
<tr>
<td>GCAGCAAGCAAGCAAGGC</td>
<td>GCCTGATGTGATGTGATGTGC</td>
</tr>
</tbody>
</table>

136 distinct tetranucleotide sequences

(new miniABC: 13 x 18-mers optimal (de Bruijn))

Occurrences in the NDB
The ABC protocol

Amber parmbsc0 + SPC/E
150 mM KCl (Smith-Dang)
11,500 water molecules
37,000-47,000 atoms
PME, PBC
1 μs simulation per oligomer

39 oligomers (18bp)
136 tetranucleotides
1.0 μs/oligo
35 μs = 35 M obs.
9.0 TB data

Lavery et al., NAR 2010
Rigid-base and Birod configuration variables (Lankas et al., PCCP, 2009)

We consider a coarse-grain model in which each base is modeled as a separate rigid body; backbones are only considered implicitly.

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

$A = T, \ T = A, \ C = G, \ G = C$

$n$

$a = 1 \ldots n$

Helical parameters

- Shear
- Stretch
- Stagger
- Shift
- Slide
- Rise
- Buckle
- Propeller
- Opening
- Tilt
- Roll
- Twist

MD analysis: Curves+, Canal, Canion

Lavery et al., NAR 2009

Blanchet et al., NAR 2011
Sequence effects on averages

Slide

Roll

Rise

Twist

bp step

RR  RY  YR  YY
B-DNA polymorphism

Non-Gaussian and multi-peaked helical parameter distributions. In fact, not only the nature of the dinucleotide steps, but also the nature of their fragments, colored according to its sequence on the Watson strand. The extreme values of variance for each dinucleotide are indicated by the colored horizontal lines: R..R (red), R..Y (green), Y..R (blue), Y..Y (orange). The values for the different families of tetranucleotide closest to the center of each oligonucleotide and, for clarity, three inter-BP, as well as all intra-BP, parameters do not significantly deviate from Gaussian behavior in any sequence context.

Evidence of conformational transitions. It should be stressed that each distribution plotted in Figure 3 shows that while all RR steps, and twist for all RR and YR steps. The other three inter-BP, as well as all intra-BP, parameters do not significantly deviate from Gaussian behavior in any sequence context.

Evidence of conformational transitions. The conclusions of our analysis are summarized in Figure 12. The results are: (i) non-Gaussian, or multi-peaked, behaviors can be observed in the other panels of Figure 2; (ii) the 3′/5′ diastereomeric preferences (g–) or trans (t), which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which represent minima of the corresponding potential energy surface, are known as BI and characteristic of canonical B-DNA, or (g+, gauche) and (O3′/H9256′−P) diastereomeric preferences, which rep
Sequence-dependent backbone dynamics

Flanking sequence:

Bl := ε, ζ = t, g-
BII := ε, ζ = g-, t

R = Purine (A, G)
Y = Pyrimidine (C, T)
Interactions in the backbone

Flanking sequence

- RR
- RY
- YR
- YY

percentage HB C8-H...O3'

bp step

percentage BII

R . . R
R . . Y
Y . . R
Y . . Y

Flanking sequence

3'

H8

O3'

C8

G

5'

G

G
Curvilinear Helicoidal Coordinates

Figure 1. Left: Schematic view of the curvilinear helicoidal coordinates (CHC). An ion (red dot) is described by a distance $D$ along the curved helical axis (black line), a radial distance $R$ from the axis and an angle $A$ from a reference vector which tracks the helical twist of the nucleic acid. At the base pair levels, this vector corresponds approximately to the long axis of the base pairs and points toward the 5′-3′ strand. Consequently, $A \approx 90^\circ$ places an ion in the minor groove and $A \approx 270^\circ$ places an ion in the major groove. Right: Isodensity surfaces (red) for the phosphorus atoms of the AGCT oligomer analyzed in the CHC system, then mapped into Cartesian space using the average helical axis of the oligomer (black line). The nucleotides are colored to indicate the base sequence (G: blue, C: green, A: red, T: orange). All isodensity plots were obtained using Chimera (39, 40).

Overall motions of the double helix will necessarily lead to some 'blurring' of the distributions associated with ions that are strongly interacting with DNA. Alternatively, one can count ion contacts with specific atoms of DNA using a chosen distance cutoff. This gives good local information, but is not well adapted to ions positioned, or fluctuating, between clusters of atoms, and it is clearly not adapted to analyzing overall distributions.

We now propose a new approach to analyze the distribution of ions, or molecules, around DNA based on the use of a natural coordinate system for DNA, namely curvilinear helicoidal coordinates (CHC). Since our algorithm, Curves+, provides a method for obtaining the helical axis of any DNA conformation, it is relatively easy to extend this analysis to determine the positions of ions or molecules around DNA in terms of their location with respect to the helical axis (using the base pair positions to define steps along the molecule and the local position of the helical grooves). Making this choice implies that overall helical deformations including twisting, bending and stretching can be eliminated from the analysis of ion densities. It is then possible to plot ion distributions radially from the curvilinear helical axis (i.e. as a cylindrical distribution function), along the length of the helical axis, or around the (curvilinear) helical axis, either for the entire molecule or for any chosen zone. We can use this approach to calculate average ion populations or local concentrations close to any chosen part of the molecule and also analyze the convergence of the...
Figure 3. Average phosphorus distributions calculated using CHC for the 1 μs AGCT trajectory and plotted in various planes: DA (top left), DR (top right), RA (bottom left). The bottom right plot shows the RA plane for the sugar C1' atoms from the same trajectory. The blue to red color scale represents increasing molarity.

DR and RA plots show the average radius of the phosphorus atoms from the helical axis as a white line and as a white circle respectively. DA and RA plots show the minor groove limits, defined by the average C1' positions, as a white line and as white radial vectors respectively. RA plots also have a vertical radial vector indicating the center of the major groove.
Figure 4. $K^+$ distributions along the helix. Two representations of the $K^+$ atmosphere around the AGAG (a), CGCG (b), GGGG (c) and AAAAA (d) oligomers. (a–d), left: Cartesian $K^+$ isomolarity surfaces at 15 M (red) and 5 M (green mesh) reconstructed from CHC histograms with respect to the average structure, shown in stick representation and colored according to sequence (A = red, C = green, G = blue, T = orange); the 5′ and 3′ ends of the Watson strand are marked. (a–d), right: $K^+$ population along the oligomer (5′-3′ direction upward) within the major (left column) and minor (right column) grooves, using the color scheme defined at the top: population increases from blue to red in steps of 0.05. Each base pair step is split into two half steps; see Figure 1 for a precise definition of the grooves.

Pasi et al., NAR (2015)
Convergence of ion populations

Figure 4

Time average populations within the DNA grooves for the unique base pair steps (T8pA9, A9pG10, G10pC11) belonging to the central tetranucleotide of the AGCT oligomer for increasing durations (ns) of the molecular dynamics trajectory.
K⁺ populations in the grooves

**Major Groove**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>G.G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A.A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G.A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A.G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G.C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A.C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C.G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C.A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A.T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T.G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T.A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C.T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C.C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C.G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T.T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Minor Groove**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>G.G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A.A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G.A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A.G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G.C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A.C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C.G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C.A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A.T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T.G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T.A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C.T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C.C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C.G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T.T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Population**

- **Major Groove**
  - Population range: 0.1 to 0.9
- **Minor Groove**
  - Population range: 0.0 to 0.6
CG dynamics

Time -45 ps
Cation entry in the minor groove
Time -30 ps
High twist → low twist
Time 0
Low slide → high slide
C8H8⋯O3' formation
Minor groove widening

CpG in high twist
GpR in low slide
No cation in the groove
Groove is narrower
C8H8⋯O3' not formed

CpG in low twist
GpR in high slide
Cation inside groove
Groove is wider
C8H8⋯O3' formed

We found that the ions in the inner region of the minor groove for the CG step have the strongest effect. In the case of K+, they increase the probability of having BII/BII by a factor of 3 (from 0.27 to 0.80), meaning that when there is an ion present in that region of the CG step there is an 80% probability of BII/BII and LT. In the case of Na+, the presence of an ion in the inner minor groove increases the probability of BII/BII by a factor of 2. Ions in the outer minor or major groove of the CG step disfavor significantly the BII/BII state, while the ions at the adjacent steps, in the inner or outer regions, have only weak effects.

CONCLUSION
We have carried out studies of the unique polymorphism of CG base pair step at an unprecedented level of detail. We have found that the HT/LT conformational transition is the result of a complex choreography of changes. Evidence from very extensive MD simulations indicates that the entrance of cations into the CG minor groove initiates the twist transition. This transition in turn involves BI↔BII changes in the phosphodiester backbone of the 3′-side of the CG step and transitions in the slide of neighboring GR steps. The HT↔LT equilibrium is strongly dependent on the sequence context, and this is found to be linked to the appearance of C8H8⋯O3' interactions that stabilize the LT state for certain tetranucleotide sequences. Consequently, CG steps have unique conformational properties that can be finely tuned by the sequence environment. This work confirms and explains the link between polymorphism and cation dynamics, previously described in only a few unrelated experimental (9, 10) and theoretical studies (11, 18). Ions play a significant role in a concerted and synchronized conformational choreography, controlling not only the thermodynamics, but also the kinetics, of the transitions.

The unique conformational properties of the CG step may be part of the explanation of its significant underrepresentation within the human genome (Supplementary Figure S15). This observation has traditionally been explained by the tendency of CG steps to be methylated, which favors cytosine to thymine mutations as the result of deamination (23). However, such an explanation is not complete, since cytosine underrepresentation also occurs in genomes where this nucleobase is not methylated (S. Cerevisiae, C. Elegans and D. Melanogaster, see Supplementary Figure S16). It seems then that the unique conformational properties of CG steps may have an impact on their genomic frequency, by John Maddocks on October 6, 2014

Dans et al., NAR (2014)
Transcription Factor Binding Sites (TFBS)

Experimental TFBS determination

- SELEX-seq
- HT-SELEX
- AC/TCAGGGTAC
- CTTACGCAGCG
- TATTCGTCAGTG
- AGCGTCAGACCT
- ATCGTCATCGTC
- TCATCGTCAGTG
- PBM
- CSI arrays
- ChIP-seq

Sequence consensus

- Bases
- Backbone
- Ions

Physical consensus

PAX5
Acknowledgements

Richard Lavery, IBCP

Thank you

John Maddocks, EPFL

Tom Cheatham  Pablo Dans  Alexis Michon
Analyzing water distributions

Radial distribution H2O

Molarity vs. R (Å)
K+ dynamics in the grooves

GGGGG major groove

AAAAA minor groove
K+ dynamics in the grooves (2)
Figure 3: Stereoview sum $(2F_o - F_c)$ electron density of structure X contoured at 1.0σ surrounding the hydrated magnesium ion, the partial spermine molecule, and the minor groove sodium and water molecules that interact with DNA bases.
Physical consensus

In vitro protein-DNA interactions

<table>
<thead>
<tr>
<th>DNA binding assays</th>
<th>SELEX-seq HT-SELEX</th>
<th>PBM CSI arrays</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ACGTCAGGGGTAC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CTTACGTCAGCCG</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TATTCGTCAGTG</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AGCGTCAGGACTT</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ATTCGTATCGTC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TCATCGTCAGT</td>
<td></td>
</tr>
</tbody>
</table>

Arbitrary sequence

CTTAGCCAGCCAGTCTCCATCAGTCCACAGGA

Potential in vivo binding sites

Actual in vivo binding sites

- Bias in the in vitro dataset
- Other cellular processes (e.g. compaction)
- Lack of physical information

Slattery et al., 2014
**miniABC: optimal packing of 4-mers**

1) AACGTGCTATGGAA
2) AATAAGTACCAGGA
3) AGAAACAGCTCTGC
4) AGGCGCAAGACTGA
5) ATTGGGGACACTAC
6) GAACGTCAAAGGTTG
7) GCCGAATGTAAATT
8) GGAGGGCCGGGTGG
9) GTTAGATTAAAATT
10) TACGGGATCGAGA
11) TGATATACGATGCA
12) TGGCATGAAGCGAC
13) TTGTGACGGCTAGG