DISENTANGLING THE ROLE OF TETRANUCLEOTIDES IN THE SEQUENCE-DEPENDENCE OF DNA CONFORMATION: A MOLECULAR DYNAMICS APPROACH

MARCOS J. ARAÚZO-BRAVO

Department of Biosciences and Bioinformatics, Kyushu Institute of Technology, Iizuka, Fukuoka, 820-8502, Japan, E-mail: marara@bse.kyutech.ac.jp

SATOSHI FUJII

Department of Chemistry and Biochemistry, Kyushu University, Fukuoka, Japan, E-mail: fujii@takenaka.cstm.kyushu-u.ac.jp

HIDETOSHI KONO

Neutron Research Center and Center for Promotion of Computational Science and Engineering, Japan Atomic Energy Research Institute, 8-1, Umemidai, Kizu-cho, Soraku-gun, Kyoto, 619-0215 PRESTO, Japan Science and Technology Agency, 4-1-8, Honcho, Kawaguchi City, Saitama, 332-0012, Japan, E-mail: kono@apr.jaeri.go.jp

AKINORI SARAI

Department of Biosciences and Bioinformatics, Kyushu Institute of Technology, Iizuka, Fukuoka, 820-8502, Japan, E-mail: sarai@bse.kyutech.ac.jp

Sequence-dependence of DNA conformation plays an essential role in the protein-DNA recognition process during the regulation of gene expression. Proteins recognize specific DNA sequences not only directly through contact between bases and amino acids, but also indirectly through sequence-dependent conformation of DNA. To test to what extent the DNA sequence defines the DNA structure we analyzed the conformational space of all unique tetranucleotides. The large quantity of data needed for this study was obtained by carrying out molecular dynamics simulations of dodecamer B-DNA structures. Separate simulations were performed for each of the possible 136 unique tetranucleotides at the dodecamer centers and the simulated trajectories were transformed into the DNA conformational space. This allowed us to explain the multimodal conformational state of some dinucleotides as aggregations of tetranucleotide conformational states that have such a dinucleotide inside their center. We proposed simple models to express in a linear way how the different bases that embrace a central dinucleotide perturb its conformational state, emphasizing how the conformational role of each base depends on its relative position (left, central, right) in the final tetranucleotide, and how the same peripherical base plays a different role depending on which is the central dinucleotide. These models allow us to establish an index to quantify the degree of context-dependence, observing an increasing context-dependence from the average base-pair step conformations AA/TT, CG, AC/GT (context-independent), AG/CT, AT, GC, GG/CC (weakly context-dependent), and GA/TC, CA/TG, TA (context-dependent).

1. Introduction

The idea that sequence defines DNA structure has gained acceptance, and thus the root of sequence dependent conformational variations has become an important problem. Results from crystallographic screens to address this problem indicate that variations from mean structural features may provide proteins with the information required for indirect readout, and for specifying altered structures.[16] Coarse preditions of the DNA structure from nucleic sequence using knowledge-based techniques[2] are possible, but such an approach requires data of enough quantity and quality. To test to what extent the DNA structure is determined by its sequence we made a systematic analysis of an interaction range of 3 base-pair steps long —tetranucleotide— level. We analyzed the conformational space of the all the 136 unique tetranucleotides. Since in the current structure databases there are not enough data to perform a reliable statistical analysis over all the possible tetranucleotides, we generated the large quantity of data necessary for this study by Molecular Dynamics (MD) simulations. We tried to envisage the perturbations induced in every central dinucleotide conformational state by all the possible bases that embrace a central dinucleotide and to analyze the reasons for the multimodal conformational states underlined by several authors through the study of crystal structures and computational techniques.[13]

2. Methods

We have generated dodecamer B-DNA sequences 5'-CGCGW_lXYZ_rCGCG-3', where $\{W_l, X, Y, Z_r\} \in \mathcal{N} = \{A,C,G,T\}$. Each sequence has one of the 136 unique tetranucleotides at its center, and the terminals are always the CGCG tetranucleotide that gives higher stability to the ensemble. Initial DNA structures were built based on the Arnott B-DNA model[3] with the nucgen module in the AMBER packages 6 and 7.[14,8] Using the Leap module of the package, the initial DNA structures were solvated with the TIP3P water molecules[9] so that the DNA molecule could be covered with at least a 9 Å water-layer in each direction in a truncated octahedral unit cell. For the neutralization of the system, 22 K^+ ions were added at favorable positions and then 17 K^+ and 17 Cl^- ions were added so that the salt concentration of the system would be 0.15 M. First a 1000-steps minimization for water molecules and ions with fixed DNA structure was taken, followed by a further 2500-steps minimization for the entire system to remove the large strains in the system. The cutoff used for the van der Waals interactions was 9.0 Å. The particle mesh Ewald method (PME)[7] was used for calculating the full electrostatic energy of a unit cell. After the minimization, the entire system was linearly heated up from zero to 300 K with a weak harmonic restraint to the initial coordinates on DNA (10 kcal/mol) during 20 ps of MD simulation under NVT condition. Further, a 100 ps of molecular simulation was carried out, keeping the weak DNA restraint for the equilibration of the system under NPT condition at 300 K. MD simulation for each of the 136 unique sequences was then carried out to sample the DNA conformations for 2 ns with NPT condition. The temperature was controlled to be 300 K by Berendsen's algorithm[4] with a coupling time of 1 fs, which was set to be the same as the time step of the MD simulation to produce a canonical ensemble of DNA conformations.[11] The SHAKE algorithm[15] was used on bonds involving hydrogen.

The force field parameters used for the MD was from Wang *et al.* (parm99).[17]

A sampling period of 2 ns is not always enough time to reach the stationary state. For the case of the AATT and ACGA, 10 ns simulations were performed instead of 2 ns. Thus, we confirmed that 2 ns were enough to stabilize the AAT structure, but for the ACGA at least 5 ns were necessary. More MD are being carried out to optimize the sampling period for each one of the 136 different tetranucleotide structures. In all cases, to obtain the final ensemble, we used the last 1 ns trajectories, where the system was sampled at every 1 ps (1000 conformations).

To perform the conformational analysis, the DNA molecule was approximated as an elastic object, with 6 degrees of freedom θ_i within a fixed geometry of bases. The local conformation of the DNA was identified at each location of a base-pair (from complementary strands) in terms of known deformations such as base-pair step translations Shift, Slide, Rise, and base-pair step rotations Tilt, Rolls and Twist.[12,6] In the current analysis we use the conformational parameters of the central dinucleotide calculated with the program 3DNA.[10] Since symmetric properties exist, from all the possible 256 tetranucleotides a subset of 136 are unique. Similarly, from all the possible 16 dinucleotides only 10 are unique. Since the conformational coordinates are calculated using one of the DNA strands,[10] the Shift and Tilt coordinates of the other DNA strand are inverted for the symmetric steps. Then, special care should be taken in the case of Shift and Tilt conformational coordinates when dealing with symmetries.

In order to reproduce the dinucleotide conformational states from the tetranucleotide ones, the dinucleotide XY MD data are calculated as the union of all the tetranucleotides W_lXYZ_r that have the dinucleotide XY at their center, $\{W_l, X, Y, Z_r\} \in \mathcal{N} = \{\text{A,C,G,T}\}$

$$XY = \bigcup_{l}^{\mathcal{N}} \bigcup_{r}^{\mathcal{N}} W_l XY Z_r \tag{1}$$

3. Results

3.1. *Statistical Analysis of the Aggregation of Tetranucleotide Conformational States*

In order to study how the tetranucleotide conformational states aggregate to produce the dinucleotide ones, for each set of the 1000 states in which each one of the 136 unique tetranucleotides evolves in its MD simulated trajectory, we calculated the gravity center of each 6 base-pair conformational coordinates. Then we aggregated the tetranucleotide data that have the same central dinucleotide using Eq. (1). For the 6 conformational coordinates of the 10 dinucleotide aggregates we calculated the gravity center μ, the standard deviation σ (of the gravity center of the tetranucleotide set that forms the aggregate), the tetranucleotide Tet_{max} that induces the maximum perturbation Δ_{max}, where the perturbation is $\Delta = |\mu - \mu_{Tet}|$ (μ_{Tet} is the gravity center of the tetranucleotide *Tet*). All these values are summarized in Table 1.

At first glance, from an observation of the average values μ of the conformational state of each dinucleotide in Table 1, it is clear that each DNA sequence induces a different structural conformational state, e.g. the Shift ranges from -0.45 Å for GA to 0.18 Å for AC, or the Twist ranges from 25.87° for CA to 36.64 ° for GC.

In the longer tetranucleotide range, we observe how the bases that embrace the central dinucleotide, to form a tetranucleotide, perturb the conformational state of their central dinucleotide in a non-uniform way. This phenomenon is quantified through the standard deviation σ, e.g. the CG Twist has a high dispersion of 4.8°, where the most disturbing tetranucleotide is GCGG, whereas for the AG Twist the dispersion is only 1.7°.

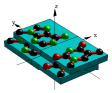
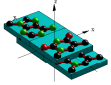
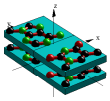
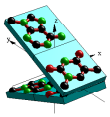
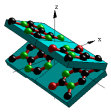
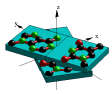
3.2. *Multimodal Conformation State of the Central Base-Pairs*

The breaking down of the dinucleotide conformational space within the tetranucleotide space allows us to explain the multimodal behavior of several dinucleotide steps already pointed out in the literature.[13] To disentangle the dinucleotide conformational space we used scatterplots and analyzed the conformational distribution pattern of all the tetranucleotides that aggregate at the same central dinucleotide.

The bidimensional scatterplots of the coordinates pairs with more salient features were chosen from all 15 possible pairs of combinations of the 6 conformational coordinates θ_i, shown in Figure 1. The left side panels of the figure present examples with unimodal conformational distributions, whereas the examples in the right side show multimodal distributions. The histograms and the equipotential ellipses were also calculated in the scatterplots. The ellipses are projections of the six-dimensional equi-potential surfaces on the respective base-pair plane obtained from the 2×2 covariance matrices; these contours correspond to energies of 4.5 k_BT ("$3\Delta\theta$ ellipses").[10] We emphasize the role of the different tetranucleotides that have the same central dinucleotide, coloring their dot distribution with the same color. The color code grades in the scale from blue to red for ordered couples of peripherical bases (AXYA, AXYC, AXYG, AXYT, CXYA, CXYC, CXYG, CXYT, GXYA, GXYC, GXYG, GXYT, TXYA, TXYC, TXYG, TXYT). We use the same color scheme for the corresponding "$3\Delta\theta$ ellipses". We observe in the right side panels of Figure 1 how the ellipses that lie in a dissimilar way to the global distribution surround peripherical dots with a uniform color. Thus, the peripherical conformational states belong to the same tetranucleotides. Then, the trajectory of each DNA structure evolves generally around the same conformational energy local minimum, and the same structure does not oscillate between different local minima. The aggregation of the trajectories around different gravity centers produced by structures with the same dinucleotide center but with different neighbors is the cause for emerging multimodal distributions in the MD dinucleotides conformational states. The bimodal (GA, GG, CG) and three-modal (TA) distributions are due to the superposition of tetranucleotide modes with different gravity centers. This means that the modes of some dinucleotides are split by their tetranucleotide modes.

The bistable behavior of the steps involving G|C nucleotides (CG, GC and GG/CC) has been already reported based both on computational models[13] and on MD simulations.[5] Packer *et al.*[13] proposed the electrostatic interactions as the reason for this behavior. Our

Table 1. Average μ, standard deviation σ, maximum perturbation Δ_{max}, and tetranucleotide of maximum perturbation Tet_{max} for the 6 conformational coordinates of each central unique dinucleotide. The symbol / separates symmetric dinucleotides.

		Shift, Å	Slide, Å	Rise, Å	Tilt °	Roll °	Twist °
							
CG	μ	-0.012	-0.336	3.268	-0.044	9.029	26.989
	σ	0.435	0.290	0.235	1.494	1.512	4.838
	Δ_{max}	0.731	0.693	0.454	3.478	2.188	8.248
	Tet_{max}	GCGC	ACGT	ACGT	GCGC	ACGA	GCGG
CA	μ	-0.287	-0.443	3.260	0.578	9.606	25.865
	σ	0.336	0.286	0.208	1.094	1.417	4.690
	Δ_{max}	0.692	0.578	0.482	2.629	2.761	10.661
	Tet_{max}	TCAC	ACAT	ACAT	TCAC	TCAC	TCAA
TA	μ	-0.119	-0.268	3.249	-0.209	8.455	29.026
	σ	0.441	0.347	0.154	1.153	1.834	2.713
	Δ_{max}	0.834	0.705	0.392	2.535	3.278	4.969
	Tet_{max}	ATAA	ATAT	ATAT	TTAC	TTAC	ATAT
AG	μ	-0.193	-0.907	3.451	-2.294	3.030	32.214
	σ	0.189	0.274	0.085	0.809	1.472	1.741
	Δ_{max}	0.373	0.558	0.180	1.647	3.224	4.033
	Tet_{max}	CAGG	TAGC	TAGC	CAGG	CAGG	CAGG
GG	μ	-0.179	-0.961	3.547	0.160	4.995	33.024
	σ	0.388	0.538	0.112	1.388	1.175	2.840
	Δ_{max}	0.728	0.903	0.219	3.310	1.838	6.660
	Tet_{max}	GGGC	CGGA	CGGA	GGGC	AGGA	CGGT
AA	μ	-0.228	-0.495	3.345	-2.394	1.196	34.465
	σ	0.254	0.231	0.071	0.593	1.474	2.484
	Δ_{max}	0.506	0.384	0.132	1.207	4.344	6.356
	Tet_{max}	GAAG	TAAG	TAAG	TAAT	TAAG	TAAG
GA	μ	-0.447	-0.253	3.392	-1.095	2.401	36.280
	σ	0.373	0.489	0.081	1.785	1.549	2.685
	Δ_{max}	0.723	0.901	0.238	3.057	3.451	6.442
	Tet_{max}	CGAT	CGAG	AGAT	CGAT	GGAT	AGAA
AT	μ	0.126	-0.970	3.268	0.404	0.045	31.148
	σ	0.218	0.171	0.069	0.895	1.733	2.434
	Δ_{max}	0.351	0.272	0.139	2.082	4.072	4.838
	Tet_{max}	TATA	TATG	TATG	TATA	AATG	TATG
AC	μ	0.176	-0.858	3.360	-0.835	1.072	32.022
	σ	0.266	0.190	0.077	1.061	1.368	1.923
	Δ_{max}	0.764	0.396	0.158	2.725	2.559	3.952
	Tet_{max}	AACC	CACG	AACT	AACC	AACG	AACC
GC	μ	-0.061	-0.610	3.422	0.165	0.311	36.635
	σ	0.312	0.353	0.098	0.989	1.608	2.456
	Δ_{max}	0.541	0.658	0.211	1.563	3.701	4.429
	Tet_{max}	TGCA	AGCC	GGCC	AGCG	GGCC	AGCG

MD simulations results agree in general with the results of Packer *et al.*[13] as is shown in Araúzo *et al.*[1] The explanation of how the dinucleotide multimodal conformational states

arised from the perturbations induced by their neighbors is complementary to the molecular mechanism of the sequence-dependence based on electrostatic interactions during the stacking process, proposed by Packer *et al.*,[13] for the dinucleotide steps such as GG/CC with an intrinsic bimodal feature due to electrostatic interaction. Our results suggest that the final conformational energy local minimum of the central dinucleotide could be induced by the interactions with its neighbors.

3.3. *Quantification of the Influence of the Neighbor Bases over the Central Base-Pairs*

To measure the degree to which every set of 3 dinucleotide steps interacts to form the conformational state of each tetranucleotide, we propose simple linear models. These models inverse the dinucleotide aggregation Eq. (1) under the hypothesis that each tetranucleotide conformational state can be explained as a function of 3 dinucleotides

$$W_l XY Z_r = f_{XY}\left(W_l X, XY, Y Z_r\right) \tag{2}$$

As an initial approach, we model such a function as a linear one and use the minimal square method to estimate the linear combination coefficients. We are interested to measure the degree to which each of all the possible dinucleotides that can embrace a central dinucleotide interacts to perturb the conformational state of the central one. This allows reinterpreting of the dinucleotide aggregation Eq. (1) as a function of the dinucleotides that perturb a central one instead of the original function of aggregation of tetranucleotides. This is done substituting in Eq. (1) the tetranucleotide expression given by Eq. (2)

$$XY = \bigcup_{l}^{\mathcal{N}} \bigcup_{r}^{\mathcal{N}} f_{XY}\left(W_l X, XY, Y Z_r\right) \tag{3}$$

where to shorten the notation, the 6-dimensional conformational states of the peripherical dinucleotides $W_l X$ and $Y Z_r$ will be denoted from now on as W_l and Z_r, respectively. With this notation we try to emphasize how the left and right neighbors perturb the conformational state of the central dinucleotide. Approximating the functions f_{XY} with linear models, finally we obtain

$$XY \approx \sum_{l}^{\mathcal{N}} w_l \cdot W_l + xy \cdot XY + \sum_{r}^{\mathcal{N}} z_r \cdot Z_r \tag{4}$$

where each uppercase symbol, W_l, XY, Z_r, represents the 6-dimensional conformational vector of the corresponding left, central and right dinucleotides, whereas the lowercase symbols, w_l, xy, z_r, stand for the regression coefficients estimated with the minimal square method. With the symbol $\approx$ we want to emphasize that this method is only an approximation, since we are interested in obtaining a rough idea of the contribution of each dinucleotide in the perturbation of the central one, and not to do prediction of DNA conformational states. For such a task, non-linear techniques such as neural networks can be more

accurate. We perform 10 linear regressions, one for each unique dinucleotide XY. Each model has 9 parameters, 4 (a_l, c_l, g_l, t_l) accounting for the perturbations that the 4 different bases in the left side can induce in the central dinucleotide, 1 (xy) accounting for the way in which the central dinucleotide counteracts the perturbation, and other 4 (a_r, c_r, g_r, t_r), accounting for the perturbations induced from the right side. Thus, we estimate 90 parameters in total. In order to obtain these parameters, we group all the tetranucleotides with the same central dinucleotide in the same model. Thus, groups of 16 or 10 members arise depending on the symmetries. In each model we use simultaneously the 6 conformational coordinates. To estimate the model parameters, the dependent term is the average conformational state μ_{Tet} of the tetranucleotide (data shown in Araúzo *et al.*,[1]) the independent terms are the average conformational states μ shown in the first row of each dinucleotide in Table 1. For example, a model without symmetric components, such as AA, has 16 members, thus providing 96 data to estimate its 9 parameters. A model with symmetric components, such as AT, provides 60 data. With this procedure we obtain finally the following 10 linear models

$$
\begin{aligned}
AA &= -0.03A_l - 0.03C_l - 0.11G_l - 0.05T_l + 0.99AA_c + 0.07A_r + 0.07C_r + 0.03G_r + 0.10T_r \\
AC &= -0.12A_l - 0.12C_l - 0.11G_l - 0.07T_l + 1.06AC_c + 0.07A_r + 0.05C_r + 0.05G_r + 0.05T_r \\
AG &= -0.07A_l - 0.10C_l - 0.10G_l - 0.10T_l + 1.01AG_c + 0.03A_r + 0.08C_r + 0.10G_r + 0.07T_r \\
AT &= +0.16A_l + 0.18C_l + 0.08G_l + 0.11T_l + 0.92AT_c + 0.02A_r - 0.05C_r - 0.10G_r - 0.08T_r \\
CA &= +0.32A_l + 0.14C_l + 0.10G_l + 0.08T_l + 0.99CA_c - 0.26A_r - 0.14C_r - 0.15G_r - 0.07T_r \\
CG &= +0.08A_l - 0.12C_l - 0.08G_l - 0.10T_l + 1.00CG_c - 0.00A_r + 0.04C_r + 0.06G_r - 0.04T_r \\
GA &= -0.16A_l - 0.07C_l - 0.11G_l - 0.04T_l + 1.28GA_c - 0.30A_r - 0.20C_r - 0.19G_r - 0.19T_r \\
GC &= -0.20A_l - 0.10C_l - 0.15G_l - 0.14T_l + 1.08GC_c + 0.09A_r + 0.03C_r + 0.08G_r + 0.14T_r \\
GG &= -0.20A_l - 0.19C_l - 0.22G_l - 0.18T_l + 1.26GG_c - 0.12A_r - 0.03C_r - 0.15G_r - 0.01T_r \\
TA &= +0.26A_l + 0.10C_l + 0.17G_l + 0.19T_l + 1.02TA_c - 0.26A_r - 0.21C_r - 0.23G_r - 0.12T_r
\end{aligned}
\tag{5}
$$

These equations summarize the disentangling of the perturbation of each of the 10 unique dinucleotides by all their possible neighbors in our MD simulation data. They show how the conformational role of each base depends on its relative position (left, central, right) in the final tetranucleotide, e.g. an A to the left side of AC (a_l=-0.12) causes a global decrease of the native conformational coordinates of AC, whereas an A to the right side of AC (a_r=+0.07) increases the coordinates. Also, Eqs. 5 show how the same peripherical base plays a different role depending on which is the central dinucleotide, e.g. a C to the left side of CA (c_l=+0.14) increases the coordinates, whereas a C to the left side of GG (c_l=-0.19) decreases the coordinates. The mean absolute errors (MAE) of the models range from 0.58 for AA to 1.08 for CG.

The 10 linear models, Eqs. 5, allow us to establish a simple index δ that quantifies the degree of context-dependence of each central dinucleotide. This is done subtracting from each central linear regression parameter of each model xy the absolute value of the sum of the peripherical parameters w_l, z_r, and normalizing dividing by the central parameter

$$\delta_{xy} = \left(xy - \sum_{l}^{\mathcal{N}} |w_l| - \sum_{r}^{\mathcal{N}} |z_r| \right) / xy \tag{6}$$

The higher the δ_{xy} is, the more independent is the central dinucleotide conformational state of its neighbors. Thus, this δ_{xy} allows us to classify on a quantitative basis the dinucleotides in the following way, according to the increasing context-dependence: AA/TT, CG, AC/GT (context-independent), AG/CT, AT, GC, GG/CC (weakly context-dependent), and GA/TC, CA/TG, TA (context-dependent). Currently, we are in the process of validating Eqs. (5) with crystal structure data. When more crystal structures become available in structural databases, Eqs. (5) can also be derived from real data (at the actual growth speed of such databases this can happen quite soon). In theory, it is also possible to perform the above analysis for each independent conformational state, by modeling each conformational state with a different model. In this way 60 models will arise. Here such a problem is not tackled since we are interested in the analysis of the global conformational state, but such an approach can be interesting in order to build conformational prediction models.

4. Conclusions

This work described an analysis of the deformability along 6 general base-pair step conformational coordinates of all 136 distinct DNA tetranucleotide duplex sequences based on MD simulations. It complements previous statistical efforts for experimental dinucleotide duplexes by Olson *et al.*[12] The MD results show that the multimodality in the conformational state of several dinucleotide steps observed in crystal data can be explained as the aggregration of the conformational states of the tetranucleotides that had at their center the same dinucleotide. Even for the cases in which the bistability of GG/CC seemed to be an intrinsic dinucleotide property derived from the bimodal distribution of the electrostatic interaction, [13] the different neighbors pushed the conformational state to one of the two local minima. These results suggest that sequence defines structure, but does in a complex way, since the same neighbor perturbs the conformational state of each central dinucleotide in a different manner. The conformational multimodality plays an important role in the DNA recognition since the different conformational modes induced by the neighbors of a central base-pair step can work as a signal for the binding of protein or other ligand. Currently, we are carrying out an analysis to classify the different types of perturbations that emanate in 3 dinucleotide interactions assembling each of the 136 unique tetranucleotides.

Acknowledgments

M.J. Araúzo-Bravo would like to acknowledge the Japanese Society for the Promotion of Science (JSPS) for supporting him for this research. This work is supported in part by Grants-in-Aid for Scientific Research 16014219 and 16041235 (A. Sarai) and 16014226 (H. Kono) from Ministry of Education, Culture, Sports, Science and Technology in Japan. We thank Prof. N. Go for encouraging this work and providing useful comments. Part of the MD calculations were carried out using ITBL computer facilities at JAERI.

References

1. M. J. Araúzo-Bravo, S. Fujii, H. Kono, S. Ahmad, and A. Sarai. Sequence-dependent conformational energy of DNA derived from molecular dynamics simulations: Toward understanding the indirect readout mechanism in protein-DNA recognition. *Journal of the American Chemical Society*, 2005. In press.
2. M. J. Araúzo-Bravo and A. Sarai. Knowledge-based prediction of DNA atomic structure from nucleic sequence. *Genome Informatics*, 16(2), December 2005. In press.
3. A. Arnott and D. W. Hukins. Refinement of the structure of B-DNA and implications for the analysis of X-ray diffraction data from fibers of biopolymers. *Journal of Molecular Biology*, 81(2):93–105, December 1973.
4. H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and A. DiNola. Molecular dynamics with coupling to an external bath. *Journal of Chemical Physics*, 81:3684–3690, 1984.
5. D. L. Beveridge, G. Barreiro, K. S. Byun, D. A. Case, S. B. Dixit T. E. Cheatham III and, E. Giudice, F. Lankaš, R. Lavery, J. H. Maddocks, R. Osman, E. Seibert, H. Sklenar, G. Stoll, K. M. Thayer, P. Varnai, and M. A. Young. Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I. Research design and results on d(C_pG) steps. *Biophysical Journal*, 87:3799–3813, December 2004.
6. R. E. Dickerson, M. Bansal, C.R. Calladine, S. Diekmann S., W. N. Hunter, O. Kennard, E. Kitzing, R. Lavery, H. C. M. Nelson, W.K. Olson, and W. Saenger. Definitions and nomenclature of nucleic acid structure parameters. *Nucleic Acids Research*, 17(5):1797–1803, 1989.
7. U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen. A smooth particle mesh Ewald method. *Journal of Chemical Physics*, 103:8577–8593, 1995.
8. T. E. Cheatham III and M. A. Young. Molecular dynamics simulation of nucleic acids: Successes, limitations and promise. *Biopolymers*, 56:232–256, 2001.
9. W. L. Jorgensen. Transferable intermolecular potential functions for water, alcohols and ethers. Application to liquid water. *Journal of the American Chemical Society*, 103:335–340, 1981.
10. X. J. Lu and W. K. Olson. 3DNA: A software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Research*, 31(17):5108–5121, 2003.
11. T. Morishita. Fluctuation formulas in molecular-dynamics simulations with the weak coupling heat bath. *Journal of Chemical Physics*, 113(8):2976–2982, 2000.
12. W. K. Olson, M. Bansal, S. K. Burley, R. E. Dickerson, M. Gerstein, E. C. Harvey, U. Heinemann, X. J. Lu, S. Neidle, Z. Shakked, H. Sklenar, M. Suzuki, C. S. Tung, E. Westhof, C. Wolberger, and H. M. Berman. A standard reference frame for the description of nucleic acid base pair geometry. *Journal of Molecular Biology*, 313(1):229–237, 2001.
13. M. J. Packer, M. P. Dauncey, and C. A. Hunter. Sequence-dependent DNA structure: Dinucleotide conformational maps. *Journal of Molecular Biology*, 295:71–83, 2000.
14. D. A. Pearlman, D. A. Case, J. W. Caldwell, W. R. Ross, T. E. Cheatham III, S. DeBolt, D. Ferguson, G. Seibel, and P. Kollman. AMBER, a computer program for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to elucidate the structures and energies of molecules. *Computer Physics Communications*, 91:1–41, 1995.
15. J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alcanes. *Journal of Computational Physics*, 23:372–336, 1977.
16. A. Sarai and H. Kono. Protein-DNA recognition patterns and predictions. *Annual Review of Biophysics and Biomolecular Structure*, 34:379–398, June 2005.
17. J. M. Wang, P. Cieplak, and P. A. Kollman. How well does a restrained electrostaic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry*, 21:1049–1074, 2000.

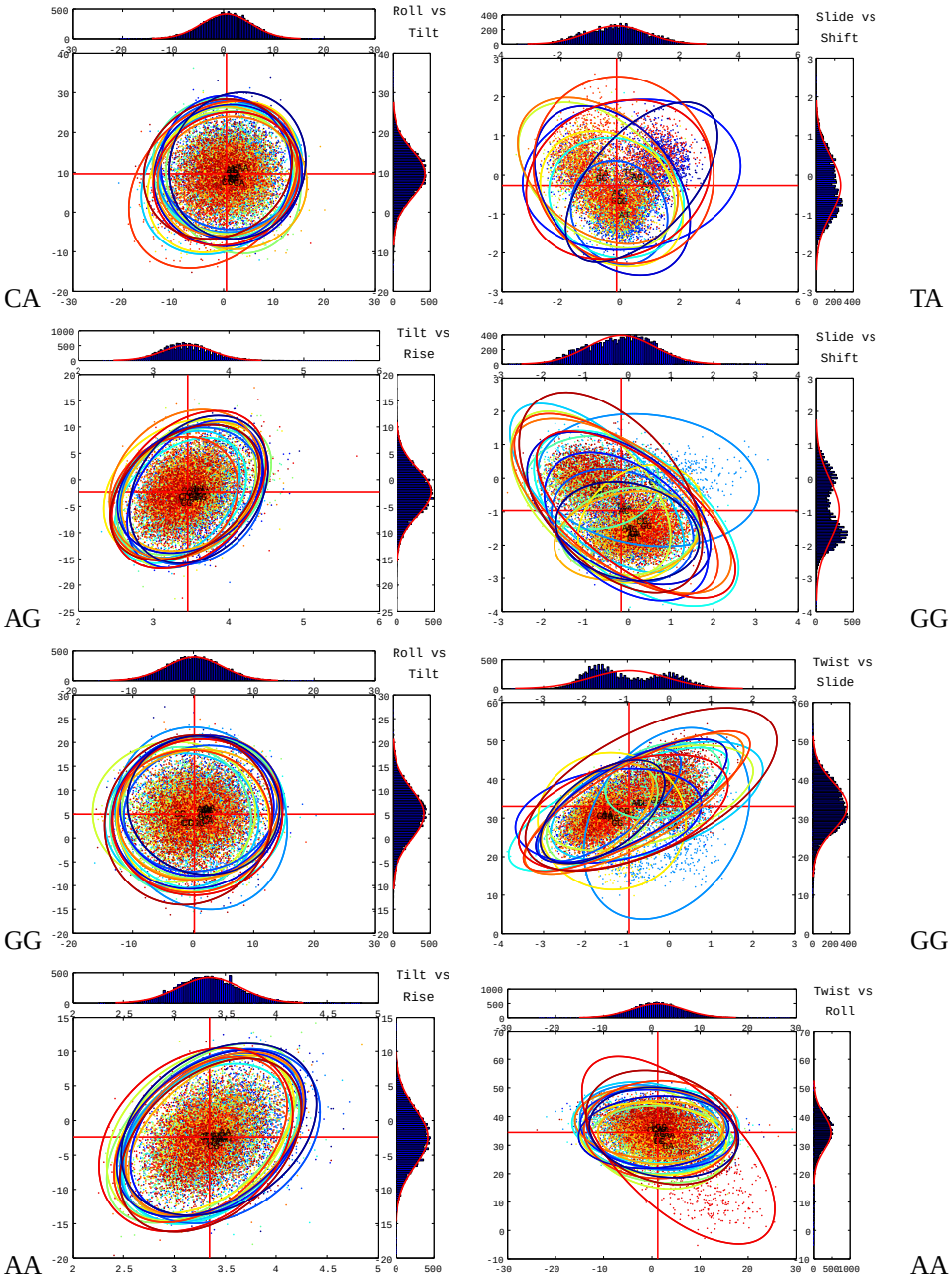


Figure 1. Scatterplots of some unimodal (left) and multimodal (right) dinucleotide conformational distributions as aggregations of tetranucleotides.