Parallel Monte Carlo Study on Caffeine-DNA Interaction in Aqueous Solution.

M.D. Kalugin[1] and A.V. Teplukhin[2]

[1]*Institute of System Programming, Russian Academy of Sciences, Moscow, Russia*
shaman@isp.ras.ru

[2]*Institute of Mathematical Problems in Biology, Russian Academy of Sciences, Pushchino, Russia*
tepl@impb.psn.ru

Abstract

Monte Carlo simulation of the caffeine-DNA interaction in aqueous solution at room temperature was carried out using parallel calculations on supercomputer. Very large simulation boxes were used containing superhelical B-DNA fragment surrounded by caffeine and water molecules. The most probable binding sites of caffeine molecules on the DNA surface as well as structural features of the respective caffeine-DNA complexes were revealed for several solutions' concentrations.

1. Introduction

Caffeine-DNA interaction in aqueous solution is studied intensively both theoretically and experimentally. The primary aim of all of these investigations is to elucidate mechanisms of caffeine activity as "drug interceptor" or/and "DNA protector" [1-4] at molecular level. Using NMR or IR/UV-spectroscopy techniques provides a great deal of such data. It should be remembered that most part of structural information results from solving of mathematically ill-posed problem of spectra interpretation, e.g. each workable structure of molecular aggregates should be taken into account. The best way to do this or at least to stimulate researcher's intuition – is computer simulation. More realistic models need to be constructed to enhance usability of simulation results in this case. So, the aim of this paper is to reveal regularities in spatial arrangement of caffeine molecules in the immediate vicinity of the DNA double helix as well as in bulk aqueous solution.

DNA usually reacts with a drug in two ways. Drug molecule can act as intercalator inserting between two adjacent base pairs of DNA or can merely bind to atoms on the DNA surface depending on its chemical structure. The structure of DNA suffers great distortions in the former case; otherwise, small ligand molecules only can slightly disturb geometry of the DNA double helix (e.g., by changing its grooves width). According to the IR-spectroscopy data [5] caffeine is capable to interact with the DNA phosphates or with the external atomic groups of the nucleic acid bases. This gives us a great opportunity to drastically diminish computational time by abandoning the internal degrees of freedom of the DNA specimen. So, DNA was treated as a solid body in the course of calculations.

Unlike the majority of enzymes DNA double helix doesn't provide well-defined compact pockets to ligand molecules. Therefore drug-DNA affinity not only depends on availability of specific combination of hydrophilic and/or hydrophobic atomic groups at the binding site but also can be modulated by the local geometry of the DNA grooves. Besides, it is known that caffeine associates to form dimers and higher aggregates in aqueous solution [6]. Thus, large simulation box is needed to freely accommodate these constituents. Also steps should be taken to receive statistically reliable computer simulation data. Of course, we can repeat the computation several times using arbitrary initial data. However it is expedient to arrange several identical DNA fragments in tandem to calculate any required data by averaging available binding sites over. As a result the simulation box is growing in size.

Computer simulation has become a highly important tool in theoretical investigations of structure-function relationships governing the fundamental biochemical processes in living cells. Thus, measurable properties are usually obtained in course of averaging over the set of samples generated by either deterministic (molecular dynamics equations) or stochastic (Monte Carlo) methods. In general, these averages converge slowly with the increasing of the length of simulations. So, using of the sequential routines for sampling becomes a very time consuming procedure well before

simulated system size achieves desirable value. Moreover, computer memory required for storing all the necessary data arrays may exceed available resources. Parallel calculations on multiprocessor hardware have a great potential for solving these kinds of problems.

A new approach (in the context of Monte Carlo method) to computer simulation of mesoscopic aggregates (10^9– plus atoms) has been recently worked out by our group of specialists in high performance compilers, parallel programming, molecular simulation and biophysics in effective collaboration. The approach is based on the original parallel algorithm [7] that allows execution of the Metropolis sampling routine [8] for several particles (e.g., atoms, water molecules, rigid molecular fragments, etc. depending on number of available processors) simultaneously. To do this we should distribute simulation box between available processors (domain decomposition strategy) so, that each will monitor the particles in the smallest region of simulated system giving us an opportunity to do several Monte Carlo trials (random displacement and, if required, rotation) simultaneously. Of course, neighboring processors should exchange some data to treat a particle motion from one region to another correctly, but this leads to negligible growing of calculation time. A high scalable software package was developed using both Fortran77 + MPI 1.2 [9] and ParJava [10] programming tools to study caffeine in aqueous solution.

In the following discussion, the results of parallel Monte Carlo study on caffeine-DNA interactions in aqueous solution at room temperature are presented for three different concentrations (dilute, concentrated and oversaturated).

2. Methods of Calculations

2.1. Simulation of water via conventional (consecutive) Monte Carlo algorithm.

Standard Metropolis sampling procedure [8,11] is commonly employed for the calculation of various thermodynamical or/and structural characteristics of water solution by Monte Carlo method. The most typical simulation box (cubic unit cell containing several hundreds of water molecules) with periodical boundary conditions is used and the closest image technique [11] is implemented to calculate the energy of intermolecular interaction using atom-atom potential functions with spherical cutoff (Coulombic and van der Waals terms are considered). The box size, the temperature and the number of molecules do not change during the particular simulation (NVT-ensemble). Initial position of water molecules may be chosen randomly within the unit cell.

To make a long story short, standard (consecutive) Metropolis NVT-sampling at given temperature usually follows the next pattern: the particular molecule is randomly displaced according with the periodical boundary condition, rotated, and then the energy increment is calculated. If a move is downhill in the energy a new position is accepted; otherwise a new position should be accepted with a probability corresponding to Boltzmann factor of the energy difference. If the uphill move is rejected, the molecule is retained at its former position and the former configuration is recounted as a new state in Markov's chain. We should repeat the procedure with the next molecule from here on. Maximal displacement per one trial is normally 0.2 Angström, and maximal rotation angle is 0.125 radians. For more details see Ref. 11.

2.2. Simulation of water using parallel Monte Carlo algorithm.

Computer simulation of biopolymers in aqueous solution requires a simulation box of greater volume to house many millions of atoms. The standard (consecutive) Metropolis sampling procedure does not suit this case. Parallel calculations using multiprocessor hardware have evidently much potential for resolving this problem. The most important features of parallel Monte Carlo algorithm [7] implemented in the present work are briefly outlined below.

Primarily we should share the simulation box between available processors (domain decomposition strategy), each will monitor molecules in the smallest region (P-box) of the simulated system. Hence, multiprocessor supercomputer can perform several Monte Carlo trials simultaneously giving drastic speed up to the simulation. To ensure the algorithm correctness all the simultaneously moved molecules should be statistically independent. Besides, the data needed to calculate the energy of interactions between the molecules of the particular P-box with those contained in the adjacent P-boxes must be stored in the processor unit memory and updated timely.

To meet the requirements we proceed as follows: each P-box is subdivided into 8 identical L-boxes (Figure 1). The L-box edge must be greater than $R_c+2\delta$, where R_c is a spherical cutoff radius commonly used to shorten intermolecular interaction and δ is the maximal value of the molecular displacement per one Monte Carlo move. At every instant all processors are only working with molecules contained in the L-boxes located at the same corners of the respective P-boxes (Figure 2). All the processors must pick their L-boxes

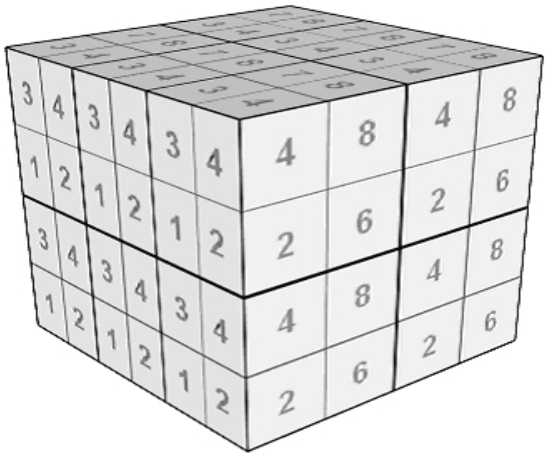

Figure 1: Fragment of simulation box formed by 12 P-boxes is presented. L-box subdivision (finest lines) as well as numbering scheme (digits) are shown.

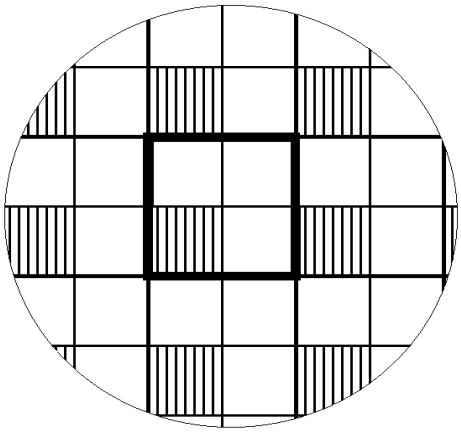

Figure 2: Fragment of XY-section of simulation box is presented showing one P-box (fat square) and 8 its neighbors placed at the same layer. L-boxes simultaneously picked by processors to do sampling procedure are shaded (only first variant of four possible for this layer is presented).

in the same order from 1 to 8 (see numbering in Figure 1) and concurrently. As a result, the distance between any of simultaneously moved molecules is always greater than R_c, and they cannot influence each other.

Data on molecules of 64 L-boxes (8 of its own and 56 from the neighboring P-boxes) should be stored in the memory of each processor unit and updated timely to handle molecular movement between adjacent P-boxes correctly and to calculate 'interprocessor' contribution to the interaction energy. Thus, owing to potential functions cutoff the memory required per processor unit is only determined by constant R_c and does not depend on the volume of the simulation box itself. The same is true for maximal length of interprocessor messages as well as for the amount of calculations per processor.

It should be pointed out that the number of messages sent/received by every processor to exchange the data with its neighbors is always a constant. Starting from the models requiring 27 (3x3x3 set) processors or more, each processor should generally exchange messages with its 26 neighbors. Since interprocessor communication is relatively slow operation, it is more advantageous to implement the three-stage updating procedure [12] where only three pairs of exchanges are required (Figure 3).

It should be also noted that only three exchange operations are needed in the special case of 8 processors (2x2x2 set) because of periodical boundary conditions.

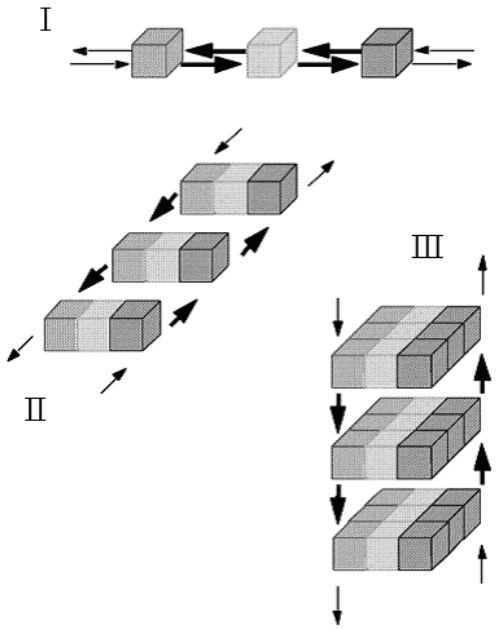

Figure 3: Interprocessor communication scheme. In the first step, each processor exchanges the data with neighbors in the +X and -X directions. In step 2, each processor exchanges the 3 sets of data it now possesses (it's own, and those from the +/- X-neighbors) with the neighbors in the +Y and -Y directions. Each processor now possesses 9 sets of data and can carry out step 3, exchanging the data sets with the processors in the +Z and -Z directions. Totally 3 pairs of exchanges are needed. Adopted from [12].

The parallel Monte Carlo algorithm can be briefly outlined as follows, on this basis:

1. Each processor unit stores only the minimal allowable (no less than four-fold value of $R_c+2\delta$ per one dimension) domain of the simulation box in its memory. This domain is subdivided into 64 identical L-boxes (3D arrangement looks like 4x4x4). Domains possessed by adjacent processors are partially overlapped. Metropolis sampling procedure is applied only to molecules located within the domain's core (a union of 8 central L-boxes named above as P-box).
2. L-box selection. All processors are working on molecules within a particular L-box in the domain's core. The order in which L-boxes are chosen is the same for each processor.
3. Standard Metropolis sampling procedure is applied to molecules in the selected L-box.
4. Three-stage updating procedure is executed. Some molecules can not only change its coordinates, but move to another L-box as well.
5. While desirable number of iterations is not reached go to stage 2.

To speed the computations up it is useful to repeat the stage 3 several times before proceed with stage 4. Neighbor list application is very advantageous in this case.

It should be noted that the usage of blocking *SendRecv* routines of MPI library to exchange data at stage 4 plays a very important role in general synchronization of the computations over the processors set. Owing to data dissimilarity (or for any other reason) minor 'discordance' in the processors work still remains and positively influences on the overall performance by lowering the probability of peak overloading of interprocessor communication hardware.

The initial configuration (free from very close interatomic contacts) is prepared using the program designed to generate atomic coordinates for the molecules randomly and uniformly distributed at given density over the simulation box. This procedure is carried out in parallel by the set of processors identical to that, used in the further simulation. Each processor generates atomic coordinates for the molecules within the respective P-box only and then adjacent processors exchange data. There is a data file for each processor where atom coordinates of the respective domain are stored.

It should be noted that the approach presented here has no formal restrictions on the simulation box size (Figure 4, here **np** ranges from 2^3 to 17^3. The smallest simulation box contains 5514 atoms and the largest one – 3386673 atoms).

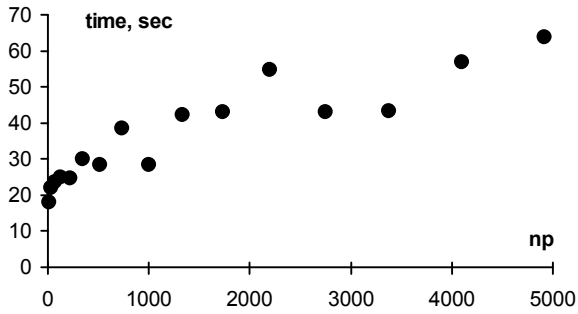

Figure 4: The time taken to complete 1000 all-molecular cycles of Metropolis procedure on set of "water boxes" arranged in the order of increasing size. Here **np** is at once the number of identical (by size) P-boxes required to build the particular water box and the number of used processor units. Each P-box contains, on average, of 230 water molecules.

2.3. Using parallel Monte Carlo algorithm for more complicated objects.

Proceeding with the software development designed for study of biopolymers in aqueous solution, we run into a variety of problems. Only simple models are taken into consideration in this work. A biopolymer molecule centered in the simulation box is treated like a set of small 'motionless' atomic groups, electrically neutral, if possible. In case of DNA there are naturally organized arrangements of its atoms in the form of nucleic acid bases, deoxyribose, and phosphate fragments confined within the sphere of radius R_b (2-3 Å). Thus each fragment should be added to the particular P-box depending on coordinates of geometrical center of the respective atomic group.

These fragments are kept motionless through Metropolis sampling procedure and water-DNA fragment interaction energy is calculated if the distance between their geometrical centers is not more than $R_{bc}=R_b+R_c$. The R_c value is equal to 9.0 Å and water R_b is set to be zero in this work. In the general case we have $R_{bc}(i,j)=R_b(i)+R_b(j)+R_c$, where *i* and *j* are fragment's types (caffeine and water molecules are also included). The L-box edge size must be greater than maximal of these R_{bc}s by 2δ. We use absolute values of the atomic partial charges (commonly expressed in the units of positron's charge) as weighting factors when calculating the geometrical center coordinates.

The initial positions of water molecules are obtained from the previously prepared water box of appropriate size by discarding water molecules located too closely from the solute atoms in this case.

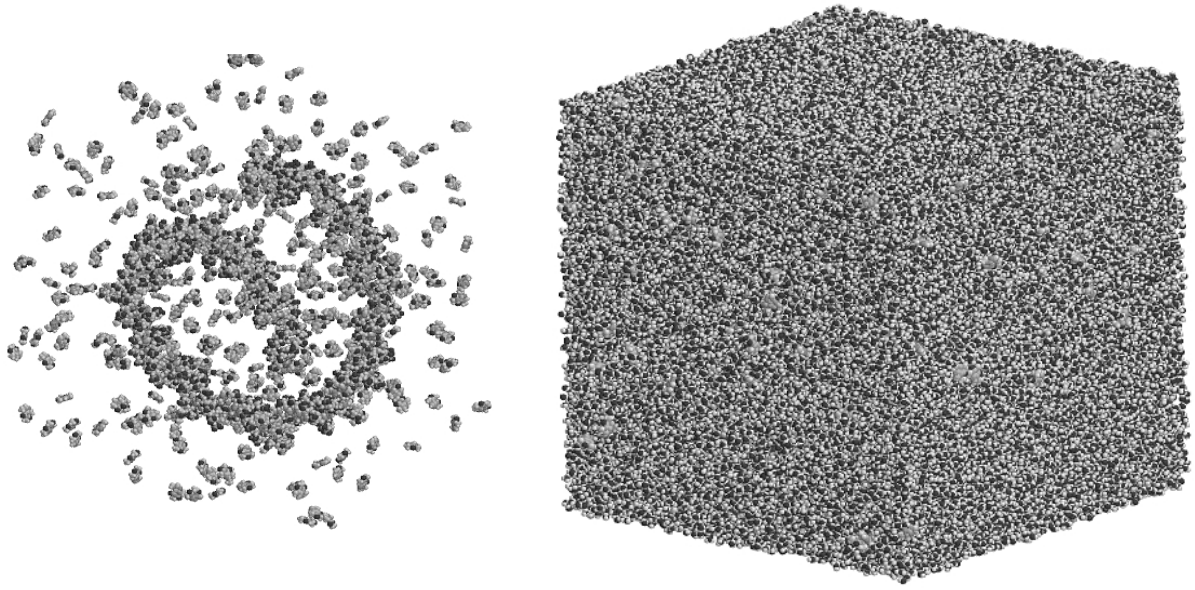

Figure 5: The simulation box arrangement (DNA in 0.1M caffeine aqueous solution at 300K). One of instantaneous molecular configuration at the end of simulation (van der Waals representation): on left image only caffeine molecules and DNA are presented and on right image all atoms (including water ones) are shown.

2.4. Molecular geometries, potential functions, and simulation box makeup.

We used three-centered solid body-like water model from [13]. All caffeine molecules were also treated like solid bodies with molecular geometry the same as described in [14]. Motionless DNA duplex $(GAGAAAAAGA)_{14}$:$(TCTTTTTCTC)_{14}$ consisting of 140 nucleotide pairs (14 double helical turns) and forming a loop of superhelix (134 Å in diameter and 76 Å pitch measured between the appropriate phosphates) was centered in the cubic simulation box with the edge size of 180 Å.

Coordinates of DNA atoms were calculated by applying 'curvilinear' replication procedure to the double-helical decamer GAGAAAAAGA: TCTTTTTCTC (original coordinates as well as the respective transformation parameters were kindly presented to AVT by Prof. V.B. Zhurkin). It should be noted that the minor groove size [15] of this superhelical 140-mer DNA duplex is periodically (for 14 times) varies from its minimal value at the end of A-tract to its maximum near the beginning of A-tract. In addition two 5'-terminal $-CH_2-PO_3$ groups were replaced by $-CH_3$ ones as well as two 3'-terminal oxygen atoms by hydroxyl groups. Thus the DNA specimen consisted of 8914 atoms.

Potential functions for water-water, water-caffeine and water-DNA interactions are those described and used in [13,15,16]. Energy of caffeine-caffeine and caffeine-DNA interactions was calculated using parameterization data from [17]. The partial atomic charges were calculated by combination of simple MO-LCAO and Huckel's methods using the parameters taken from [18].

Three computer simulation experiments were completed to study the caffeine-DNA interactions in dilute, concentrated, and oversaturated aqueous solutions (caffeine concentrations were approximately equal to 0.05M, 0.1M, and 0.2M). To do this, three initial configurations of the simulation box were prepared containing the same DNA specimen and, respectively: 175 and 192869, or 350 and 187224, or 700 caffeine and 181730 water molecules. So, each studied system totally contains 591721, 578986, and 570904 atoms respectively. The general view of the simulation box is presented in Figure 5.

To achieve equilibrium, the initial segment of Markov's chain, comprised of 5 million trials for each water molecule, was discarded from consideration. The averaging was carried out over the chains providing up to 25 million trials for each water molecule. Three sets of 216 (P-box 3D arrangement looks like 6x6x6) processor units were used to perform calculations on MVS100K supercomputer (http://www.jscc.ru).

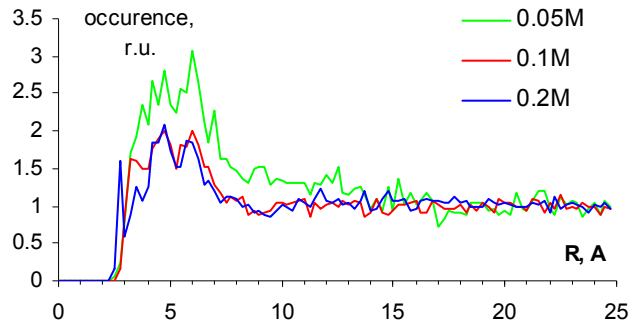

Figure 6: Occurrence (in relative units) of the caffeine molecules *vs.* the distance between closest caffeine-DNA atoms pair. Data for three values of caffeine concentration in aqueous solution are shown.

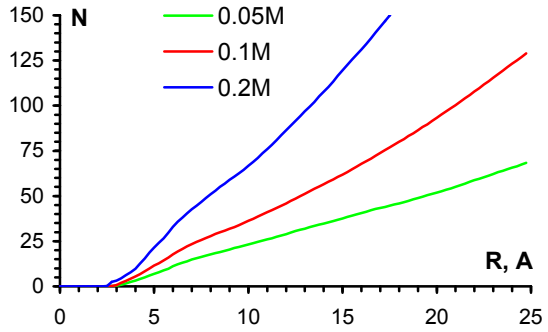

Figure 7: The number of caffeine molecules located near DNA, so that the distance between closest caffeine-DNA atoms pair is not more than R.

3. Structural features of caffeine-DNA interactions.

It is conveniently to deal with radial distribution function while studying local ordering in a simple liquid, and yet this is not a good way for the investigation of spatial correlations in the caffeine-DNA subsystem. To characterize the position of the particular caffeine molecule against the DNA molecule it is more properly to use the distance between the closest pair of its atoms (the first atom belongs to caffeine while the other one to the DNA, hydrogen atoms are not taken into consideration) while calculating the distribution function. This function has a cylindrical symmetry at the distances smaller than the DNA superhelix radius. Furthermore, a cylindrical core of 9 Å radius should be excluded to normalize the distribution function properly. In other words, we plot the caffeine distribution function originating from the surface of the curvilinear tube passing coaxially through the DNA double helix. We use 9 Å instead of standard B-DNA radius value to take the DNA grooves into account because they do actually reduce the diameter of double helix.

The occurrence (relative units; unity level corresponds to the uniform distribution) of caffeine molecules near the DNA surface was calculated as a function of the distance between closest caffeine-DNA atoms pair and is presented in Figure 6. First of all a highly populated (relative to the uniform distribution) shell about 4 Å thick around DNA was revealed regardless the solution concentration. Next were several clear-cut maxima at the distances of 3.5, 4.7, and 6 Å, respectively. A sharp peak at 2.75 Å may be associated with H-bond formation between caffeine and the DNA amino- and OH-groups. As it follows from our calculations this takes place mainly in highly concentrated (oversaturated) solutions.

Total amount of caffeine molecules forming the shell of the given thickness around the 140-mer DNA can be evaluated using data displayed in Figure 7. Thus, there are, on average, only 0.1, or 0.2, or 2.4 caffeine molecules forming at a time H-bonds with the DNA H-donor groups in 0.05M, 0.1M, or 0.2M solutions, respectively. In order to reveal the geometry of the most probable caffeine-DNA aggregates we should analyze a representative set of instantaneous configurations of the simulation box.

As to a particular case of individual instantaneous configuration we should extract coordinates only for the caffeine molecules bounded with DNA and the coordinates of DNA atoms itself using the geometrical criterion of proximity based on data shown in Figure 6 (distance between the closest caffeine-DNA atoms pair does not exceed 3.65 Å). The representative set was prepared by extracting the data from instantaneous configuration obtained after every 100000th all-molecular cycle of Metropolis procedure for each studied system. We collected three sets of 120 instantaneous configurations of the caffeine "shell" formed around the DNA 140-mer (for three solution concentrations, respectively). Stereo image of one of these sets of the 'DNA caffeine shell' (0.1M solution) is presented in Figure 8.

The main structural types of caffeine-DNA associates forming in aqueous environment (Figure 9) and the most preferential binding sites were revealed as well as some correlation with nucleotide sequence was found by close examination out of the representative sets mentioned above.

Thus, configurations like *m1* or *m2* constitute the most important structural class and occur in the narrowed portions of the DNA minor groove (5'-AAAGA sites). Here *m1* represents complexes stabilized by an H-bond formation between guanine

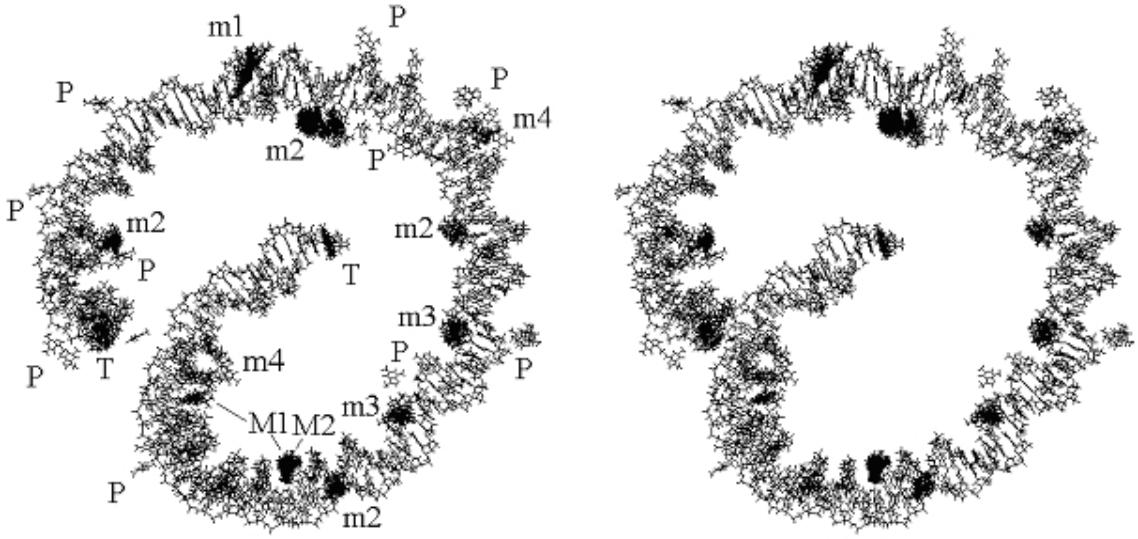

Figure 8: Configurational ensemble of the caffeine-DNA associates in 0.1M aqueous solution.

NH_2-group and any hydrophilic atom of caffeine molecule while *m2* stands for similar complexes without an H-bond. It should be noted that *m1*-type structure prevails in highly concentrated solution while *m2*-type dominates at mean and low concentrations.

Configurations like *m3*, *m4* and *P* (Figures 8 and 9) may be pooled to form the second structural class characterized by the external (without inserting into the minor groove) mode of binding to the DNA sugar-phosphate backbone. Common to these structures are the closest van der Waals contacts between the caffeine methyl groups and aliphatic gaps in the DNA sugar-phosphate chains. A choice between *m3* and *m4* type is a matter of convention because the plane of caffeine's aromatic ring is often situated in the intermediate position. The probability of bidentate (*m3* and *m4*) complex formation slightly reduces with the increasing of caffeine concentration and, on the contrary, the ratio of *P*-associates growing rapidly. The widest portions of the DNA minor groove (5'-GAAA sites) are the most preferential for the structures of this class.

The third structural class includes aggregates of *M1* and *M2* types (Figures 8 and 9) formed mainly in the narrowest portions of the DNA major groove (5'-GAAA sites). Main structural feature of *M1*-like complex is H-bond between any hydrophilic atom of caffeine molecule and the DNA NH_2-groups belonged to adenine or cytosine nucleic acid base. The *M2*-type binding occurs in a variety of fashions where both bridge-like constructions (*via* hydrophobic contacts between caffeine CH3-groups and methyl group of thymine and/or deoxyribose C2'(H_2)-group) and some kinds of π–complex (thymine C5–CH_3 chemical bond points to the center of associated caffeine and is orthogonal to the plane of its aromatic ring) are presented.

Aggregates of *T*-type in the vicinity of terminal parts of the DNA double helix constitute a special class of 'restricted use' because there are no reliable experimental data on structure of the DNA duplex ends in aqueous solution. In this case either any or both molecular complexes are stacked and T-shaped structures are formed between the caffeine and terminal nucleic acid base pair (see 'T-ensemble', Figure 9). The probability of these complexes formation is growing rapidly as caffeine's concentration increases in the solution. An H-bond formation between caffeine and H-donor 3'-terminal OH-groups occurs rarely.

4. Conclusions

A very large simulation box can be used due to the distributed parallel calculations. This allows setting up the methodology of computer-aided drug design to the next level when binding affinity is directly evaluated by calculating the ratio of the host-ligand pairs formed in the simulation box containing a great quantity of ligands and host molecules (or binding sites). On this way time-consuming procedure of free energy increment calculation is not needed. Besides, our method makes it possible to take into consideration a further important contribution to free energy of the system originating from the ligand-ligand interactions (as well as host-host, and multiparticle terms) when treating a specimen with typical concentration of solutes.

In conclusion we should note that our results are in good agreement with experimental data. In particular, it was observed [5] that caffeine interacts with sugar-phosphate backbone as well as NH_2-groups exposed to the DNA grooves.

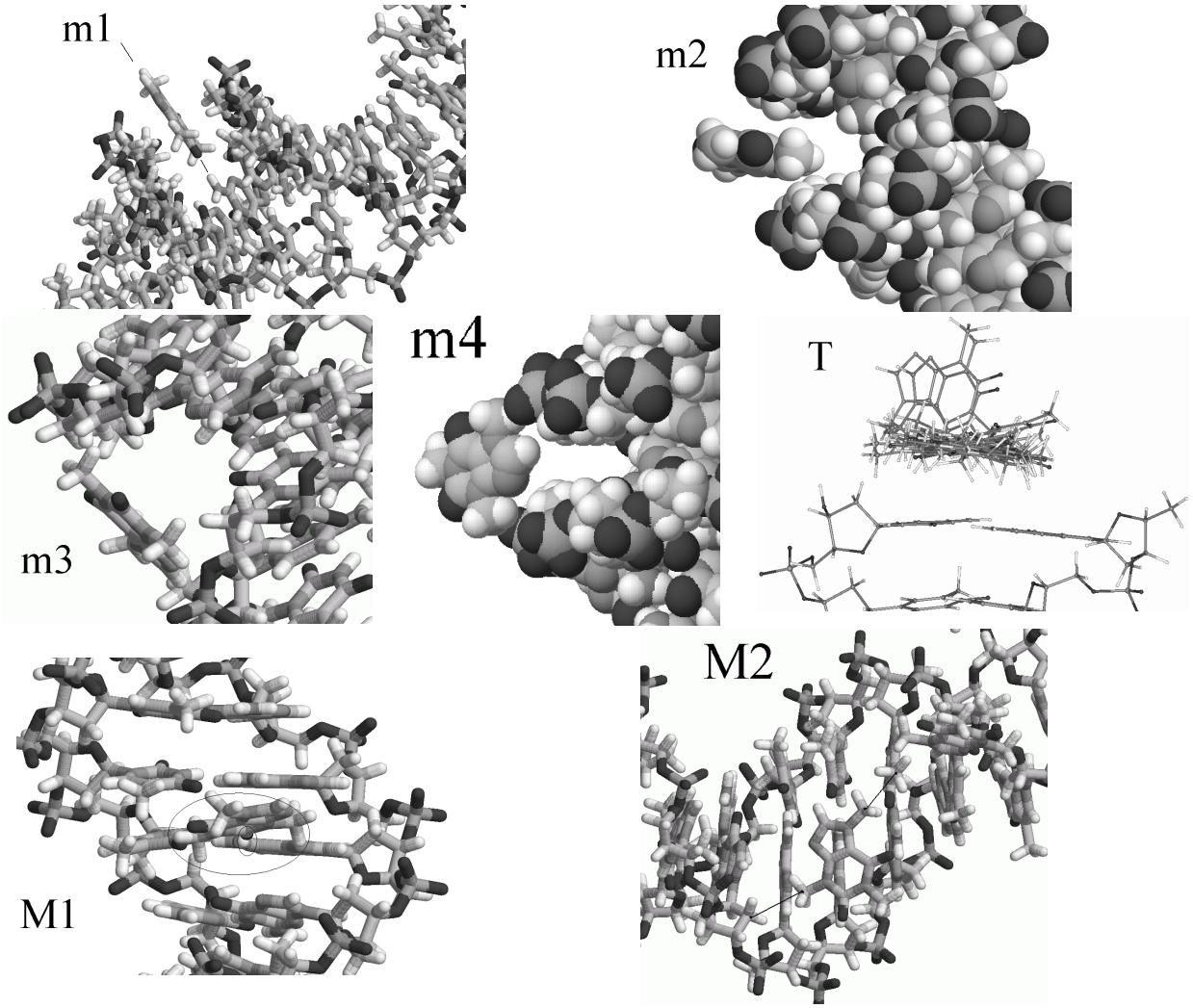

Figure 9: The main structural types of the caffeine-DNA associates formed in aqueous environment.

5. References

[1] Larsen R.W., Jasuja R., Hetzler R., Muraoka P.T., Andrada V.G., Jameson D.M., *Biophys. J.*, 70:443-452, 1996

[2] Traganos F., Kapuscinski J., Darzynkiewicz Z., *Cancer Res.*, 51:3682-3689, 1991

[3] Davies D.B., Veselkov D.A., Djimant L.N., Veselkov A.N., *Eur. Biophys. J.*, 30:354-366, 2001

[4] Lyles M.B., Cameron I.L., *Cell Biol. Int.*, 26:145-154, 2002

[5]Shestopalova A.V., *The PhD Thesis*. Inst. Phys AS GSSR, Tbilisi, 1987.

[6] Falk M., Gil M., Iza N., *Can. J. Chem.*, 68:1293-1299, 1990

[7] Teplukhin A.V., *Mat. Model.*, 16(11):15-24, 2004

[8] Metropolis N.A., Rosenbluth A.W., Rosenbluth M.N., Teller A.H., Teller E., *J. Chem. Phys.,* 21:1087-1092, 1953

[9] Snir M., Otto S., Huss-Lederman S., Walker D., Dongarra J., *MPI: The Complete Reference.* MIT Press, Boston, 1996.

[10] Ivannikov V., Gaissaryan S., Avetisyan A., Padaryan V., In proc. of 10th EuroPVM/MPI conference. Venice, Sept. 2003, *LNCS,* v. 2840, pp. 491-494.

[11] Allen M.P., Tildesley D.J., *Computer simulation of liquids*. Oxford University Press, N.Y., 1987.

[12] Heffelfinger G.S., *Comput. Phys. Commun.*, 128:219-237, 2000

[13] Poltev V.I., Grokhlina T.I., Malenkov G.G., *J. Biomol. Struct. Dyn.*, 2:413-429, 1984

[14] Sutor D.J., *Acta Cryst.*, 11:453-458, 1958

[15] Teplukhin A.V., Zhurkin V.B., Jernigan R., Poltev V.I., *Mol. Biol.*, 30(1,pt2):75-84, 1996

[16] Teplukhin A.V., Malenkov G.G., Poltev V.I., *J. Biomol. Struct. Dyn.*, 16:289-300, 1998

[17] Poltev V.I., Shulyupina N.V., *J. Biomol. Struct. Dyn.*, 3:739-765, 1986

[18] Berthod H., Pullman A., *J. Chim. Phys*., 62:942-946 1965