

# Three-Dimensional Model of a Selective Theophylline-Binding RNA Molecule

C.-S. Tung\*†, T. I. Oprea†, G. Hummer†‡ and A. E. García†

† Theoretical Biology and Biophysics (T-10), MS K710 and ‡Center for Nonlinear Studies (CNLS), Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545 USA

A three-dimensional (3D) model for an RNA molecule that selectively binds theophylline but not caffeine is proposed. This RNA, which was found using SELEX (Jenison *et al.*, 1994), is 10 000 times more specific for theophylline ( $K_D=320$  nM) than for caffeine ( $K_D=3.5$  mM), although the two ligands are identical except for a methyl group substituted at N7 (present only in caffeine). The binding affinity for ten xanthine-based ligands was used to derive a comparative molecular field analysis model ( $R^2=0.93$  for three components, with cross-validated  $R^2$  of 0.73), using the SYBYL and GOLPE programs. A pharmacophoric map was generated to locate steric and electrostatic interactions between theophylline and the RNA binding site. This information was used to identify putative functional groups of the binding pocket and to generate distance constraints. On the basis of a model for the secondary structure (Jenison *et al.*, 1994), the 3D structure of this RNA was then generated using the following method: each helical region of the RNA molecule was treated as a rigid body; single-stranded loops with specific end-to-end distances were generated. The structures of RNA-xanthine complexes were studied using a modified Monte Carlo algorithm. The detailed structure of an RNA–ligand complex model, as well as possible explanations for the theophylline selectivity are discussed.

**Keywords:** RNA folding; binding specificity; molecular modeling; QSAR; theophylline; SELEX

## Introduction

Large variability and flexibility of nucleic acid molecules make them ideal to design specific sequences for performing various tasks. The SELEX procedure (Tuerk and Gold, 1990) allows the selection of nucleic acid molecules that bind molecular targets with high affinity and specificity. For example, a selected RNA molecule that folds into a pseudoknot motif can serve as a ligand that inhibits the function of the human immunodeficiency virus type I reverse transcriptase (Tuerk *et al.*, 1992). Another RNA molecule was selected as a ligand to basic fibroblast growth factor to inhibit receptor binding (Jellinek *et al.*, 1993). Other RNA molecules were found that have high-binding affinity to small molecular ligands such as ATP (Sassanfar and Szostak, 1993) and a variety of organic dyes (Ellington and Szostak, 1990). Single-stranded DNA molecules that bind and inhibit human thrombin were also found using the SELEX procedure (Bock *et al.*, 1992). SELEX was proposed as an alternative route for drug discovery, since this procedure is capable of selecting nucleic acids that recognize and bind to specific molecular targets (Jenison *et al.*, 1994).

Recently, RNA molecules with high affinity for theophylline, but not caffeine, were obtained using the SELEX procedure. Theophylline, a naturally occurring methyl xanthine, is currently used as bronchodilator in the treatment of airway obstructive diseases. Chemical differentiation from other related xanthines, e.g. caffeine

and theobromine, when monitoring blood levels of theophylline, is important for avoiding toxicity problems. All theophylline-binding RNA molecules share a similar secondary structure (Jenison *et al.*, 1994) that includes a conserved CCU bulge and a six bases symmetric internal loop flanking a conserved three base-pair stem. On the basis of equilibrium filtration analysis with [ $^{14}\text{C}$ ]theophylline, the binding affinities of these RNA molecules are similar to those observed for monoclonal antibodies raised against the particular antigen (Poncellet *et al.*, 1990). Detailed physicochemical and nuclear magnetic resonance (NMR) studies were reported for one of these RNA molecules, labelled mTCT8-4 (Jenison *et al.*, 1994). Thermal denaturation studies showed that mTCT8-4 forms a stable folded structure with a  $T_m$  of 72°C. The binding affinity of ten xanthine derivatives to TCT8-4 RNA was also established in competition studies using radiolabeled theophylline (see Table 1 in Jenison *et al.*, 1994).

To understand the molecular basis of theophylline–RNA specificity, a series of ten ligands binding to TCT8-4 and their physicochemical properties using molecular modeling procedures was examined. On the basis of pharmacophoric pattern matching, a three-dimensional (3D) quantitative structure–activity relationship (QSAR) of the ten xanthine derivatives using the comparative molecular field analysis (CoMFA) (Cramer III *et al.*, 1988a) method was developed.

On the basis of a set of structural constraints generated using the pharmacophoric map and stereochemistry, a 3D model in atomic detail was constructed for the RNA molecule (TCT8-4). A complete set of reduced coordinates

\* Author to whom correspondence should be addressed.

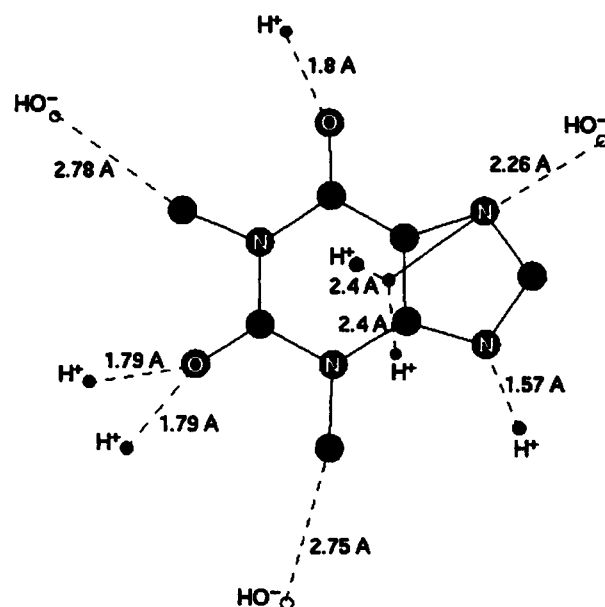
developed in the authors' laboratory was used for the modeling of nucleic acid structures (Soumpasis and Tung, 1988; Tung, 1993; Tung *et al.*, 1994). This method has the advantage that every parameter in the set has intuitive physical meaning, therefore it is ideal for modeling nucleic acid molecules with limited tertiary structure information.

## Methods

### Structure-activity relationship

The ten xanthine derivatives have been comparatively analyzed, in order to extract maximum information about the nature of the ligands and their molecular recognition features. These features were used to gain insight in the binding mechanism of the ligands to the target RNA molecule, and to determine the characteristic elements required for binding (i.e. the pharmacophore). The pharmacophoric pattern of theophylline was determined using the XED (extended electron definition) procedure (Vinter, 1994), which explicitly treats pi-electrons and lone pairs of polar groups in the molecular mechanics framework of the COSMIC90 force field (Morley *et al.*, 1991). After XED points are added to a given molecule having AM1-derived atom-centered charges, the local minima for the interaction with three probes ( $H^+$ ,  $HO^-$  and  $H_2O$ ) are located, similar to the GRID (Goodford, 1985) method. The geometric location of the  $H^+$  and  $HO^-$  minima was used to define a pharmacophoric map for theophylline, as presented in Fig. 1.

Structural models of the xanthine derivatives were calculated using the AM1 Hamiltonian, as implemented in Mopac 6.0 (Stewart, 1990), with full geometry optimization (keywords: AM1 EF). For these optimized molecules, the free energy of solvation was estimated using the Amsol 4.0 program (Cramer and Truhlar, 1992) (keywords AM1 SM2



**Figure 1.** The pharmacophoric pattern of theophylline obtained using the XED procedure. Hydrogen atoms were omitted for clarity. Interaction energies between probe atoms and theophylline are given in the Results section.

1SCF). Since purine, the scaffold of xanthine derivatives, can exist in two tautomeric structures (proton can migrate from N7 to N9—see Table 1), N7–N9 tautomers were theoretically investigated for these compounds (except 1,3 dimethyl-uric acid, 7-methylxanthine, theobromine and caffeine). Mopac AM1 and Amsol AM1-SM2 heats of formation (HoF) were used to compare the stability of these compounds, in conjunction with calculated and measured Log *P*, the octanol–water partition coefficient (Leo, 1993). Log *P* was obtained from values from Daylight CIS.

The relationship between chemical structure and binding affinity of xanthine derivatives to mTCT8-4 was studied using two 3D QSAR programs, CoMFA and GOLPE

**Table 1.** Chemical structures, calculated and experimental properties of xanthine derivatives used in the competitive assay with TCT 8-4 RNA. For each property m is measured, c is calculated and p is predicted; n.a.—not available. See explanations in text

Compound	R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>	R <sub>4</sub>	R <sub>5</sub>	Log <i>P</i> (m)	p <i>K</i> <sub>a</sub> (m)	Δ <i>G</i> <sub>sol</sub> (c)	p <i>K</i> <sub>a</sub> (m)	p <i>K</i> <sub>a</sub> (p)
Theophylline	CH <sub>3</sub>	CH <sub>3</sub>	H	H	—	−0.02	8.81	−19.367	0.495	−0.377
CP-theophylline	propionate	CH <sub>3</sub>	H	H	—	−0.12	n.a.	−24.165	0.032	−1.024
Xanthine	H	H	H	H	—	−0.73	7.53	−22.384	−0.930	−0.775
1-Methylxanthine	CH <sub>3</sub>	H	H	H	—	−0.27	7.70	−20.982	−0.954	−0.954
3-Methylxanthine	H	CH <sub>3</sub>	H	H	—	−0.72	8.10	−20.705	−0.301	0.404
7-Methylxanthine	H	H	CH <sub>3</sub>	H	—	−0.89	8.33	−18.441	−2.778	−3.814
Theobromine	H	CH <sub>3</sub>	CH <sub>3</sub>	H	—	−0.78	10.00	−16.554	−2.778	−2.898
1,3-Dimethyluric acid	CH <sub>3</sub>	CH <sub>3</sub>	H	O	H	−0.52	n.a.	−19.337	−3.041	−3.452
Hypoxanthine <sup>a</sup>	H	—	—	H	H	−1.11	1.91	−25.334	−1.690	−1.196
Caffeine	CH <sub>3</sub>	CH <sub>3</sub>	CH <sub>3</sub>	H	—	−0.07	0.60	−15.680	−3.544	−2.544

<sup>a</sup>Hypoxanthine has no oxygen substituted at C2.

(Generating Optimal Linear PLS Estimations). A brief description of these methods is included. The CoMFA (Cramer III *et al.*, 1988a) method postulates that observed variations in experimental activity among different compounds may be explained with calculated or measured physicochemical properties, and are dependent on the compounds intermolecular interactions with the binding site. In CoMFA, these non-covalent interactions are represented as the steric (Lennard-Jones) and electrostatic (Coulombic) energies of a ligand with a grid of regularly spaced probe atoms. Both steric and electrostatic energy are tabulated for each molecule at every grid point. The resulting matrix is statistically analyzed using partial least squares (PLS) (Wold *et al.*, 1993). CoMFA fields are thus related to the binding affinity. This procedure emphasizes receptor characteristics that were probed by the initial data set.

The following settings were used for CoMFA: 1.5 Å spaced grid (18 × 15 × 10 Å); probe atom: Csp<sup>3</sup>; dielectric constant of 1.0; a +/− 20 kcal/mol cut-off value with no electrostatic interactions at steric overlap points (i.e. electrostatic field values are not considered at grid points overlapping with ligand atoms). The regression analyses were performed using the Sybyl (SYBYL™) implementation (Cramer III *et al.*, 1988b) of the PLS algorithm (Wold *et al.*, 1993), with cross-validation (leave-one-out model), and three principal components. A 0.4 kcal/mol SDEV (standard deviation) column filter was applied. The AM1-optimized geometries with AM1-derived partial charges were used to generate the CoMFA models. Non-hydrogen atoms were superimposed on the purine scaffold of theophylline.

The predictive ability of this model was evaluated by calculating  $q^2$ :

$$q^2 = 1 - \frac{PRESS}{SD}, \quad (1)$$

$$PRESS = \sum (Y_p - Y_a)^2, \quad (2)$$

$$SD = \sum (Y_a - Y_m)^2, \quad (3)$$

where *PRESS* is the sum of squared deviations between predicted ( $Y_p$ ) and actual ( $Y_a$ ) activities, and *SD* is the sum of the squared deviations between actual activities and the mean ( $Y_m$ ) biological activity value for all molecules. The best predictive model yields a  $q^2$  of 1, but values can go to zero (or below) for models that predict values equal to, or worse than  $Y_m$ .

GOLPE (Baroni *et al.*, 1993) was used in conjunction with CoMFA to further refine the PLS model using variable selection procedure according to fractional factorial design (Cruciani *et al.*, 1993). The internal predictivity of the xanthine model system was determined by three-random groups cross-validation. In this procedure, the data set is randomly divided into three equal groups, one set being excluded from the model and predicted. The process was repeated 25 times, using recalculation of weights. Results were also compared with the leave-one-out cross-validation.

The distribution of both steric and electrostatic fields was studied in GOLPE, and a 0.4 SDEV filter was applied to each field. The steric/electrostatic weighting ratio, automatically suggested in CoMFA, was applied: 1.0 for the steric, and 0.55 for the electrostatic field. GOLPE was used to correlate macroscopic properties of the ligands with the binding affinity,  $pK_D$ , using the PLS algorithm.

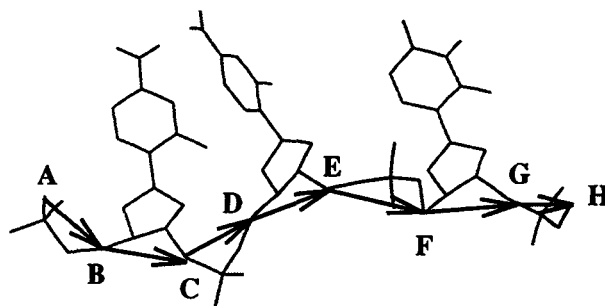
## RNA folding

From the predicted secondary structure, the whole RNA molecule can be divided into two motifs (helices and loops). To simplify the problem, all helices in the molecule were modeled as rigid bodies assuming an ideal A-RNA conformation. Six parameters (three rotational and three translational) are required to describe the position and orientation of each helix in space with respect to one another. By positioning and orienting in space all helices, the lengths of all the interconnecting loops are determined. To model the structures of these interconnecting loops, an algorithm that is capable of generating single-stranded loops with fixed end-to-end distances, was developed.

If  $S_n$  indicates the vector for C5'-O3' of the  $n$ th sugar and  $P_n$  indicates the vector for O3'-C5' between the  $n$ th and the  $(n+1)$ th sugars, the span of the three-base loop can be represented as seven connecting vectors, shown in Fig. 2. This group of seven vectors can be further decomposed into two long vectors (*AD* and *EH*) and one short vector (*DE*). The end-to-end distance ( $d$ ) of three connecting bonds is related to the torsional angle ( $\alpha$ ) of the three bonds according to the following equation (Tung, 1993):

$$d^2 = l_1^2 \sin^2 \theta_1 + l_2^2 \sin^2 \theta_2 + (l_1 \cos \theta_1 + l_3 \cos \theta_2 - l_2)^2 - 2l_1 l_3 \sin \theta_1 \sin \theta_2 \cos \alpha. \quad (4)$$

where  $l_1$ ,  $l_2$  and  $l_3$  are the three bond lengths and  $\theta_1$  and  $\theta_2$  are the two bond angles. For a specific end-to-end distance, a torsional angle ( $\alpha$ ) can be found that satisfies the condition as specified in equation (4). For example, the structure of the ADEH segment can be completely defined using three lengths ( $l_{AD}$ ,  $l_{DE}$ ,  $l_{EH}$ ), two angles ( $\theta_{ADE}$ ,  $\theta_{DEH}$ ) and one torsional angle ( $\phi$ ) corresponding to the rotation of the ADEH segment with respect to the end-to-end pseudobond (A-H). In a similar way, the two short segments (ABCD and EFGH) can each be described by a set of six parameters. A



**Figure 2.** Schematic representation of a three-base single-stranded loop. Seven vectors were used to represent the structure of the loop such that it can easily fit to a specified end-to-end distance. In this reduced representation, 28 parameters are required to completely describe the structure of the three-base loop.



Soumpasis *et al.*, 1990), B to Z (Soumpasis, 1984; García and Soumpasis, 1989), and double helix to hairpin (García *et al.*, 1990) transitions.

Under low salt conditions ( $\approx 10$  mM NaCl), when the Debye length is much larger than the characteristic length of oligomers, the Debye screening potential behaves approximately like  $1/\epsilon R$ , where  $\epsilon$  is the dielectric coefficient of the solution at a given temperature, and  $R$  is the distance between a pair of charged groups. This approximation has been used in the modeling of unusual DNA hairpins (Gupta, 1993a), centromeric (Catasti, 1994) and telomeric (Gupta, 1993b) DNA and the relative stability between parallel and antiparallel (García *et al.*, 1994) DNA structures. This simple approximation gives the correct weight of Coulombic interactions relative to other interactions used in the modeling of biomolecules. This is appropriate for polyelectrolytes under low salt solution concentration, and when the secondary structure (i.e. hydrogen bonding pattern) of the molecule is known or assumed. Consistent with this model, the electrostatic interactions were screened with a dielectric constant of 80 corresponding to that of bulk water. The inclusion of a screened Coulomb potential effectively treats the ions as a cloud. Therefore, ions such as  $Mg^{2+}$  were not explicitly included in the calculation.

Selected structures for the drug-RNA complexes were subject to energy minimization using AMBER (Pearlman *et al.*, 1991). No cut off was applied for the non-bonded interactions. The geometric parameters and charges of the theophylline were determined as described in the previous section. Further refinement of the intramolecular potentials of theophylline (e.g. fitting of vibrational spectra) were considered unnecessary for this study, where the intermolecular interactions of RNA and theophylline are dominant.

## Results and Discussion

### Structure-activity relationships

Prior to modeling the pharmacophoric pattern and the structure-affinity relationship, the local minima for each structure and the tautomers of susceptible compounds was examined. Carboxy-propyl (CP)-theophylline has two conformational minima (using AM1): at  $-91.68$  and  $-91.57$  kcal/mol, respectively. The carboxy-propyl (CP) linker is perpendicular to the purine scaffold, above and below the plane. The calculated free energy of solvation,  $\Delta G_{\text{sol}}$  (using AM1-SM2), is equal for both conformers, since they have similar energies in water:  $-115.85$  and  $-115.72$  kcal/mol, respectively. Thus, both conformers of the CP linker can be important in the binding process and were considered in the modeling of mTCT8-4 RNA.

The N9 tautomers of theophylline, CP-theophylline and 3-methylxanthine are not energetically relevant, in either the gas phase (AM1) or water (AM1-SM2) (see Table 2). The N9 tautomers of xanthine and 1-methylxanthine seem to gain importance in aqueous solution as opposed to in the gas phase:  $\Delta \text{HoF}$  of N7 compounds in 3 kcal/mol higher in the gas phase, but only 2 kcal/mol higher in water. Since the

N9 tautomers  $\Delta G_{\text{sol}}$  is better than that of N7 ones, it is possible that tautomeric structures become relevant for these two compounds, but with a smaller molar fraction than N7 tautomers. Calculated Log  $P$  values are identical for both N7 and N9 tautomers, and do not significantly differ from the measured values (see Table 2).

The N9 tautomer of hypoxanthine has a lower HoF than the N7 tautomer, both in the gas phase ( $\Delta \text{HoF} \approx 1$  kcal/mol) and in solvent ( $\Delta \text{HoF} \approx 2.45$  kcal/mol). The latter suggests

**Table 2. Force-field parameters for theophylline and caffeine.**  
Forcefield parameters are defined by the energy functions:  $E_b = K_b (l - l_0)^2$ ,  $E_{\theta} = K_{\theta} (\theta - \theta_0)^2$ ,  $E_{\phi} = V/N [1 - \cos (n\phi - \phi_0)]$  and  $E_{\text{HB}} = A/r^{12} - B/r^{10}$  for bond lengths, bond angles and dihedral angles and hydrogen bonds respectively. Parameters are defined as in Weiner *et al.*, 1986.

#### (a) Bond parameters.

Bond	$K_b$	$l_0$
H-N4	440	1.01
N4-C4	440	1.38
N4-C7	520	1.40
C4-H4	340	1.08
C4-N5	520	1.33
N5-C6	350	1.40
C6-N6	350	1.40
C6-C7	520	1.35
N6-C8	340	1.47
N6-C5	440	1.43
C8-H8	340	1.09
C5-O5	570	1.22
C5-O7	440	1.48
N6-C4	440	1.38
N6-C7	520	1.40

#### Table 2(b). Angle parameters

Angle	$K_{\theta}$	$\theta_0$
H-N4-C4	70	127.2
H-N4-C7	70	126.9
N4-C4-H4	35	121.7
N4-C4-N5	70	113.4
N4-C7-C6	70	105.4
N4-C7-C5	70	131.8
C4-N5-C6	70	102.6
C4-N4-C7	70	105.9
H4-C4-N5	35	124.9
N5-C6-N6	70	127.5
N5-C6-C7	70	112.7
C6-N6-C8	70	117.7
C6-N6-C5	70	123.9
C6-C7-C5	70	122.8
N6-C8-H8	70	109.5
N6-C5-O5	80	121.7
N6-C5-N6	70	114.8
N6-C6-C7	70	119.9
C8-N6-C5	70	118.3
H8-C8-H8	35	109.5
C5-N6-C5	70	123.9
N6-C5-C7	70	114.7
O5-C5-C7	80	123.6
C4-N6-C7	70	105.2
C4-N6-C8	70	127.6
7-N6-C8	70	127.2
N6-C7-C6	70	105.7
N6-C4-H4	35	121.2
N6-C4-N5	70	114.0
N6-C7-C5	70	131.6

**Table 2(c) Dihedral angle parameters**

Dihedral	N	V <sub>0</sub>	φ <sub>0</sub>	n
X-N4-C7-X	4	6.60	180	2
X-N4-C4-X	4	6.70	180	2
X-C4-N5-X	2	20.00	180	2
X-C6-C7-X	4	16.30	180	2
X-C5-C7-X	4	4.40	180	2
X-N5-C6-X	2	5.10	180	2
X-C6-N6-X	4	6.60	180	2
X-N6-C5-X	4	6.60	180	2
X-N6-C8-X	6	0	0	3
X-X-N4-H		1.0	180	2
X-X-C4-H4		0.0	180	2
X-X-N6-C8		0.0	180	2
X-X-C5-O5		10.5	180	2

**Table 2(d). Hydrogen bond parameters**

Atom pair	A	B
H N5	7557.0	2385.0
H2 N5	7557.0	2385.0
H3 N5	4019.0	1409.0
HO N5	7557.0	2385.0
HS N5	14184.0	3082.0
HW N5	7557.0	2385.0

**Table 2(e) Lennard-Jones parameters**

Atom	R	ε
N	1.85	0.09
C	1.85	0.12
O	1.60	0.20

**Table 2(f) Partial Charges (in esu)**

Atom	Theophylline	Caffeine
N1	-0.5319	-0.2899
CN1	-0.1794	-0.2398
HCN1	0.1231	0.1285
C2	0.9673	0.7519
O2	-0.4432	-0.5580
N3	-0.7823	-0.4728
CN3	0.0936	-0.1226
HCN3	0.0807	0.1029
C4	0.6902	0.7898
O4	-0.6128	-0.5480
C5	-0.1845	-0.4364
C6	0.6337	0.4883
HCN7	—	0.1710
H7/CN7	0.4137	-0.4261
N7	-0.5025	0.0855
C8	0.2974	0.2317
H8	0.1435	0.1355
N9	-0.6139	-0.5961

that N9 hypoxanthine has a higher molar fraction in aqueous solution. Both tautomers were included in the CoMFA study. However, the 3D QSAR models were not able to distinguish between the two structures. For hypoxanthine, the N3-N7 and N3-N9 tautomers were also examined (Table 2). These tautomers are not energetically relevant, in either the gas phase ( $\Delta H_{\text{of}} \approx 7$  and 14 kcal/mol) or aqueous solution ( $\Delta H_{\text{f}} \approx 5.15$  and 8.5 kcal/mol), when compared with that of the N9 tautomer HoF. The calculated Log *P* values ( $-0.59$  for both the N3-N7 and N3-N9 tautomers) are different from the measured one ( $-1.11$ ). The difference from the experimental value is significantly higher for

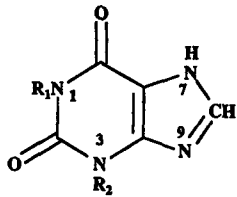
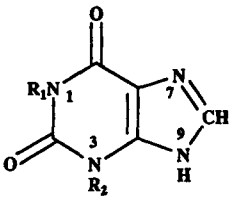
the N3-N7 and N3-N9 tautomers ( $\Delta \text{Log } P \approx 0.5$ ) than the N7 and N9 ones ( $\Delta \text{Log } P \approx -0.1$ ). On the basis of the gas phase and solvent energy, as well as calculated Log *P* estimates, it is concluded that these tautomers are unlikely to occur in the system of interest, and have not used them in the QSAR studies.

The pharmacophoric map of theophylline (Fig. 1) was determined using the XED procedure. Three hydroxide minima were observed, at less than 3 Å, in coplanarity with the bicyclic compound. One minimum, considered energetically relevant ( $-39.5$  kcal/mol), is situated between N7 and C8, somewhat closer to the former. This minimum is not located directly above the N7 hydrogen due to the slightly acidic character of the C8 hydrogen. Two other  $\text{HO}^-$  minima, located above the N1 and N3 methyl groups at more than 2.7 Å, were not included in the constraint map due to the weak nature of their interaction with hydroxide (6.5–7.5 kcal/mol). These minima, a consequence of the slightly positive N-substituted methyl hydrogens, do not represent significant molecular recognition features. Among the six proton minima observed, the strongest interaction is at N9 ( $-28.5$  kcal/mol), which is the major site for tautomeric shifts. Both carbonyl oxygens exhibit strong proton affinity: two minima at O2 are located above the oxygens' lone pairs ( $-22.5$  kcal), whereas one is located above O6 ( $-19.5$  kcal). The other O6 lone pair does not have a minimum due to the proximity of the N7 hydrogen. Two other minima ( $-23.9$  kcal/mol) are located above and below the bicyclic plane, between C5 and C6, and are caused by the pi-electron cloud. These minima were included in the constraint map, after merging the positions of the O2 proton minima into a single location. The distances indicated in the pharmacophoric map were used as distance constraints for the RNA-folding procedure.

The binding affinity of xanthine derivatives to mTCT8-4 was correlated with several macroscopic physicochemical properties: measured Log *P* and  $\text{p}K_{\text{a}}$  and calculated  $\Delta G_{\text{sol}}$ , respectively. There was no correlation between  $\text{p}K_{\text{D}}$  and Log *P* ( $R^2=0.08$ , standard error of estimate 1.294,  $n=10$ ), nor between  $\text{p}K_{\text{D}}$  and  $\text{p}K_{\text{a}}$  ( $R^2=0.00$ , standard error of estimate 1.34,  $n=8$ ). This first result excludes hydrophobicity as the driving force in the process of binding of xanthine derivatives to mTCT8-4. This important result is supported by the polar nature of the ligands (except for methyl groups, they contain polar moieties). This prompted the authors to focus on the electrostatic interactions suggested by the pharmacophoric pattern. The lack of correlation between  $\text{p}K_{\text{a}}$  and  $\text{p}K_{\text{D}}$  implies that the binding process is not dependent on the protonation state of the ligands, without implying that the binding process is pH independent. A weak correlation was observed between  $\Delta G_{\text{sol}}$  and  $\text{p}K_{\text{D}}$  ( $R^2=0.345$ , standard error of estimate 1.092,  $n=10$ ). Far from being significant, this result shows that poor solubility is, in some cases, related to reduced affinity for mTCT8-4. Indeed, three compounds mentioned to have poor solubility (Jenison *et al.*, 1994), 7-methylxanthine, theobromine and 1,3-dimethyl uric acid, have a small  $\Delta G_{\text{sol}}$  and a high  $\text{p}K_{\text{D}}$ . This estimate does not explain, however, why caffeine and theophylline, compounds with similar  $\Delta G_{\text{sol}}$ , have higher solubility, but totally different binding affinities.

The absence of correlation between the aforementioned

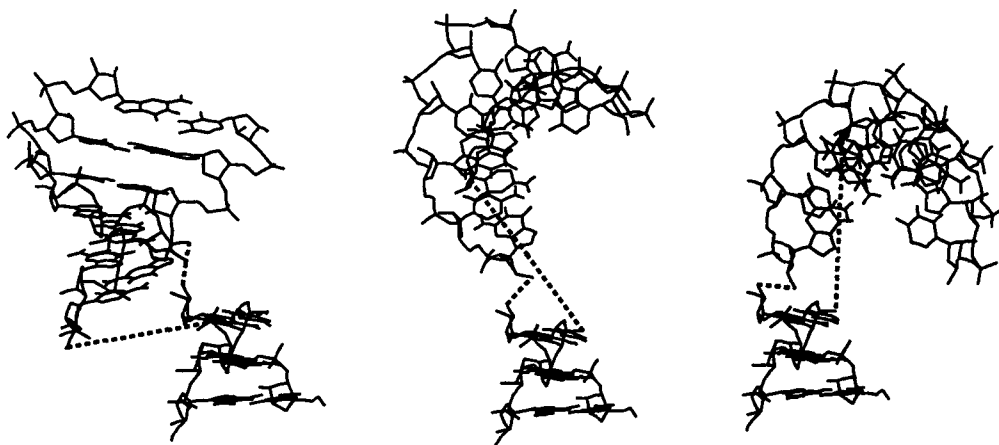
**Table 3.** Calculated properties for N7 (left) and N9 (right) tautomers of xanthine derivatives

							
Compound	R <sub>1</sub>	R <sub>2</sub>	Log P(m)	Clog P	H <sub>f</sub> (g)	H <sub>f</sub> (s)	ΔG <sub>so</sub> (c)
Theophylline-N7	CH <sub>3</sub>	CH <sub>3</sub>	-0.02	-0.06	4.781	-14.856	-19.367
Theophylline-N9	CH <sub>3</sub>	CH <sub>3</sub>	—	-0.06	8.245	-12.057	-20.303
CP-theophylline-N7	propionate	CH <sub>3</sub>	-0.12	0.12	-91.683	-115.848	-24.165
CP-theophylline-N9	propionate	CH <sub>3</sub>	—	0.12	-87.908	-112.860	-24.952
Xanthine-N7	H	H	-0.73	-0.70	-8.316	-30.700	-22.384
Xanthine-N9	H	H	—	-0.70	-4.113	-28b.671	-24.557
1-Methylxanthine-N7	CH <sub>3</sub>	H	-0.27	-0.34	-1.448	-22.430	-20.982
1-Methylxanthine-N9	CH <sub>3</sub>	H	—	-0.34	2.094	-20.176	-23.079
3-Methylxanthine-N7	H	CH <sub>3</sub>	-0.72	-0.69	-2.203	-22.908	-20.705
3-Methylxanthine-N9	H	CH <sub>3</sub>	—	-0.69	1.053	-20.298	-21.351
Hypoxanthine-N7	H	—	-1.11	-1.20	44.453	20.456	-23.797
Hypoxanthine-N9	H	—	—	-1.20	43.326	17.992	-25.334
Hypoxanthine-N3	—	H	—	-0.59	50.389	23.146	-27.243
Hypoxanthine-N3N9	—	H	—	-0.59	57.309	26.456	-30.853

Note: Four possible tautomers of hypoxanthine have been investigated. CLogP is the estimated LogP calculated using PCModules (from Daylight CIS). HoF(g) and HoF(s) are the AM1 (gas phase) and AM1-SM2 (solvent) heats of formation.

macroscopic descriptors and  $pK_D$  warranted the use of the CoMFA method. The model system of ten xanthine derivatives was analysed with the CoMFA-GOLPE procedure. Using principal component analysis (Wold *et al.*, 1987), three principal components were found to explain 63.14 per cent of the variance. The initial CoMFA run, using 2112 variables, yields the following three component model:  $R^2=0.956$ , standard error of estimate 0.281,  $q^2=0.341$  and standard error of prediction 1.09 (using the leave-one-out procedure). This model, while suggesting common trends in the series, lacks robustness. Therefore, this model was submitted to a variable selection procedure using the fractional factorial design strategy (implemented in GOLPE). The model retained 48 of 2112 variables, and

the three-component model explains 87.1 per cent of the variance. This model yields improved robustness ( $R^2=0.932$ , standard error of estimate 0.352,  $q^2=0.731$  and standard error of prediction 0.699 with leave-one-out cross-validation). The increase in internal consistency is also reflected in the internal predictivity, as determined by three-random groups cross-validation:  $q^2=0.598$ , standard error of prediction 0.855. The predicted activities of this latter model are listed in Table 3. While these do not reflect the best predictive model (leave-one-out values are slightly better), they are likely to reflect better the predictive ability of the model, in the absence of an external (test) set (Oprea and García, 1996)—the process being mimicked by exclusion of one third of the compounds. Because the training set



**Figure 4.** Different arrangements of the two helices in mTCT8-4. The three base-pair helix is fixed in space as shown in the lower part of the figure. The six base-pair helix is allowed to sample different positions and orientations such that the two connecting distances (between the two helices shown as two dotted lines) are constrained to 3.5 Å (the length of a phosphate group) and 16.6 Å (the averaged length of a three-base single-stranded loop), respectively.

includes only ten compounds, this model was used only to examine trends in the RNA–ligand binding process.

Graphical analysis of CoMFA fields revealed the following trends: a favorable electrostatic interaction, located around N7–C8–N9, which is likely to be occupied by negative charges in the binding site. This accounts for the poor affinity of 1,3-dimethyl uric acid, and for the higher affinity of hypoxanthine, xanthine, 1-methylxanthine and 3-methylxanthine. This CoMFA result is supported by the authors' RNA model (see later) and by the  $\text{HO}^-$  minimum above N7–C8, suggested in the theophylline pharmacophoric map (Fig. 1). In their analysis (Jenison *et al.*, 1994), Jenison and co-workers suggest 'a binding pocket involving a steric boundary at the C8 region'. On the basis of the results here, it is suggested that the RNA binding site contains a negatively charged polar group facing the C8 region of the ligand. It is further predicted that uric acid derivatives will have a  $K_D$  in the higher micromolar range.

Areas of reduced steric bulk tolerance in CoMFA are located around N7, accounting for the poor binding affinity of 7-methyl substituted compounds (theobromine, 7-methylxanthine and caffeine). Areas of beneficial steric interaction are located around N3 (explaining the over-predicted affinity of 3-methylxanthine) and considerably less at N1 (theophylline is underpredicted by this model). Since no chemical variation occurs at C6 and C2 (except for hypoxanthine), CoMFA fields do not reveal the importance of electrostatic interactions at these two positions. This does not constitute a problem to the authors' understanding of the mechanism of binding, since these atoms were used when superimposing the ten compounds (and are part of the pharmacophore).

In summary, the lack of correlation with hydrophobicity and solvation parameters suggests that electrostatic interactions play a major role in the binding process. CoMFA results corroborate the pharmacophoric map, indicating the importance of electrostatic interaction points in defining the distance constraints.

## RNA folding

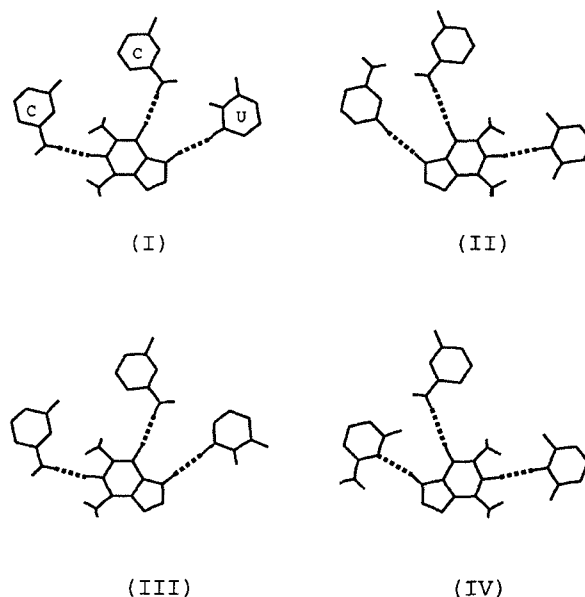
The binding between theophylline and mTCT8-4 is specific, hence the intercalating mode of binding in the duplex region of the RNA molecule is ruled out. The possibility that the UCU hairpin loop is involved in the binding site was ruled out because of the lack of sequence homology and length in this loop. The two possibilities for the theophylline binding site in mTCT8-4 are: (1) the totally conserved CCU bulge and (2) the highly conserved six-bases symmetric internal loop. On the basis of the proton NMR spectra (Jenison *et al.*, 1994), the stoichiometry of binding is one theophylline per RNA molecule. The authors favor the notion of a unique binding site for theophylline in the RNA molecule. The CCU bulge is completely homologous among the set of RNA molecules that bind theophylline. In addition, the appearance of two imino proton resonances at 11 and 15 ppm upon binding of theophylline was interpreted as a G forming a non-standard base pair and a C protonated at the imino group (Jenison *et al.*, 1994). Both interpretations are

consistent with the collapse of the six-bases symmetric bulge into a stem region when theophylline is bound to the RNA. Therefore, it is concluded that the CCU bulge is part of the binding site.

Other parts of the RNA molecule interacting directly with the drug, or in close proximity of the binding site include a three base-pair helix (r(CCA)-r(UGG)) and a six base-pair helix (r(GCAUCG)-r(UGAUGC)). The CCU loop and the two flanking helices constitute the part of the RNA molecule that was modeled in detail. Using the reduced set of coordinates described in the Methods section, 40 parameters are required for the description of the drug–RNA complex (28 parameters for the CCU loop, six parameters for the six base pairs helix and six parameters for theophylline).

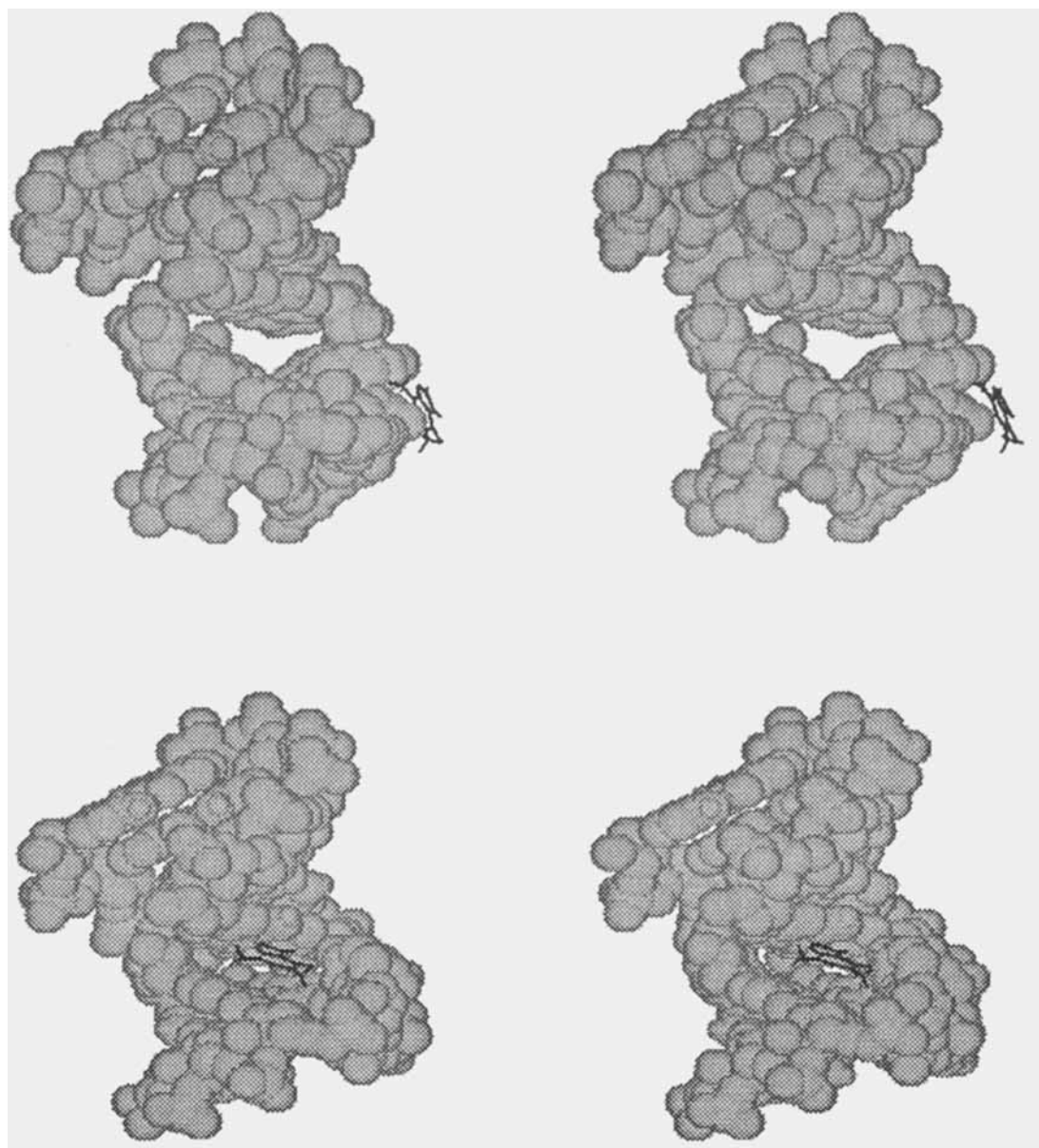
The authors have studied the end-to-end distances for all three-base spans in the crystal structure of tRNA<sup>phe</sup> [TRNA04 in Protein Data Bank (Bernstein *et al.*, 1977)]. The authors found that the average end-to-end distance ranged from 10 Å to 20 Å. This average end-to-end distance (16.6 Å) was used as an initial constraint for the three-base loop. The optimum distance for phosphate groups (3.5 Å for O3'–C5' of two neighbouring sugars) was used as another constraint. This allowed the authors to position and orient the two helices with respect to each other. Many different arrangements for the two helices satisfy both the imposed constraints as shown in Fig. 4. This indicates that the RNA molecule is highly flexible at the CCU bulge region in the unbound state. This observation is consistent with the notion that binding of theophylline stabilizes a conformation that is otherwise flexible. Once the initial arrangement of the two helices is selected, the distance constraints for the three-base loop are removed. Structural constraints deduced from the pharmacophoric map shown in Fig. 1 were used as a guideline for modeling the drug–RNA complex.

Fig. 5 shows four different possibilities for the three H-



**Figure 5.** Four H-bond patterns between theophylline and the CCU loop. On the basis of the pharmacophoric map, the three strong H-bond donors/acceptors on theophylline are O2, O6 and N9. These three were chosen to form H-bonds with the CCU loop.



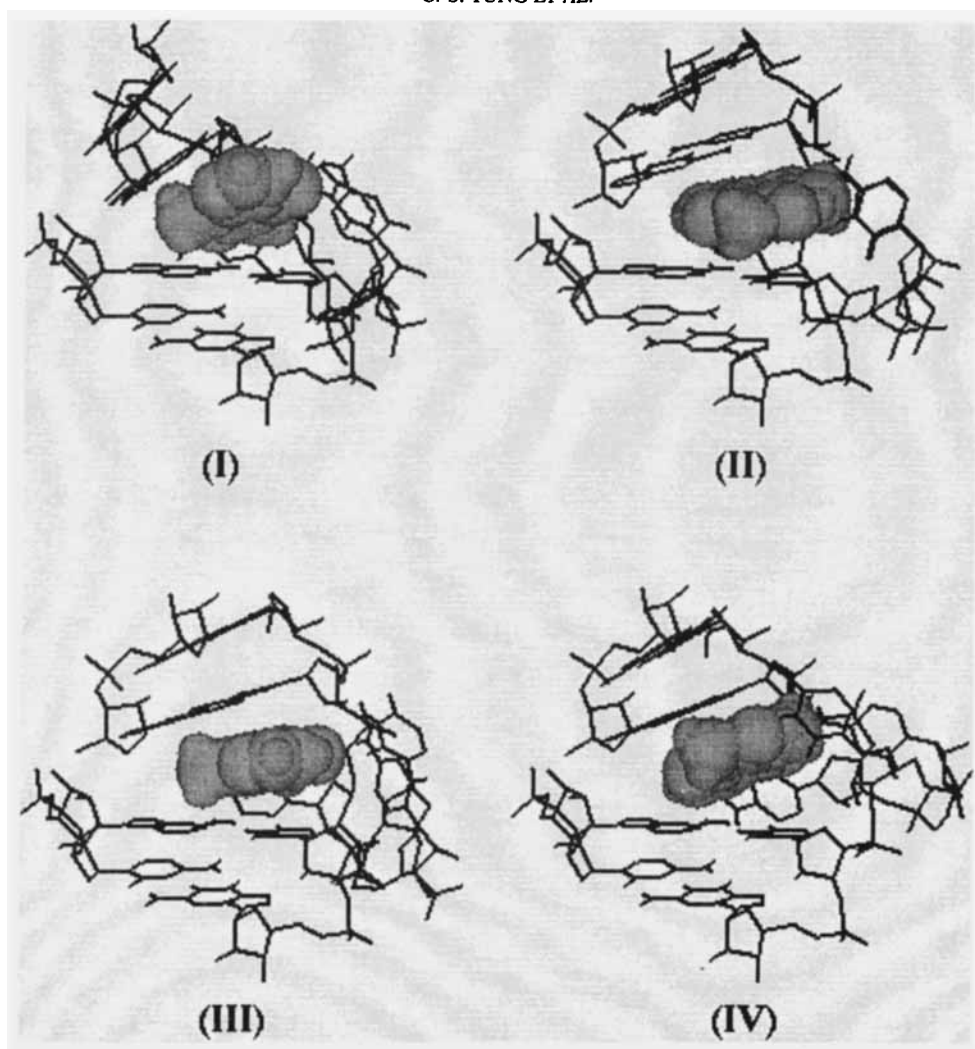


**Figure 6.** Structure of the CCU loop, the two flanking helices and the theophylline (stereo). The top panel shows a selected random structure of the group while the lower panel shows the structure after 5000 cycles of Monte Carlo simulation. The Monte Carlo simulation not only positioned the theophylline inside the binding cavity but also reduced the cavity size to improve the interaction between the drug and the RNA molecule.

bonds between theophylline and the CCU loop. In addition to these H-bonds, atoms H3 of U29 and HN2B of G11 were constrained to satisfy the two proton minima observed above and below the theophylline ring (see Fig. 1). The structure of the four different binding modes was modeled with the modified Monte Carlo algorithm. The result of a typical run is shown in Fig. 6. The initial conformation, selected from a pool of random structures is shown on the top. The equilibrated structure (final conformation) is shown on the bottom. The equilibrated structures of different runs were ranked according to their total energies (the conformational energy plus the constraint energy). The lowest energy structure of a particular set of runs was selected for that specific binding mode. The structures of the four different binding modes are plotted in Fig. 7.

#### Atomic model refinement

The structures corresponding to the four different binding modes were subject to further refinement using AMBER (Pearlman *et al.*, 1991). On the basis of 5000 steps of unconstrained minimization, structure II showed the lowest energy and was chosen for further analysis. This choice was also motivated by the agreement with previously identified constraints. In particular, structure II has the 1-methyl group pointing outwards such that the CP linker can be used to connect CP-theophylline to the selection column in the SELEX procedure. In this conformation, the CP linker is pointing above the bicyclic ring. In addition, the N7 position is buried in the binding pocket and involved in a hydrogen bond. These results are consistent with the negative



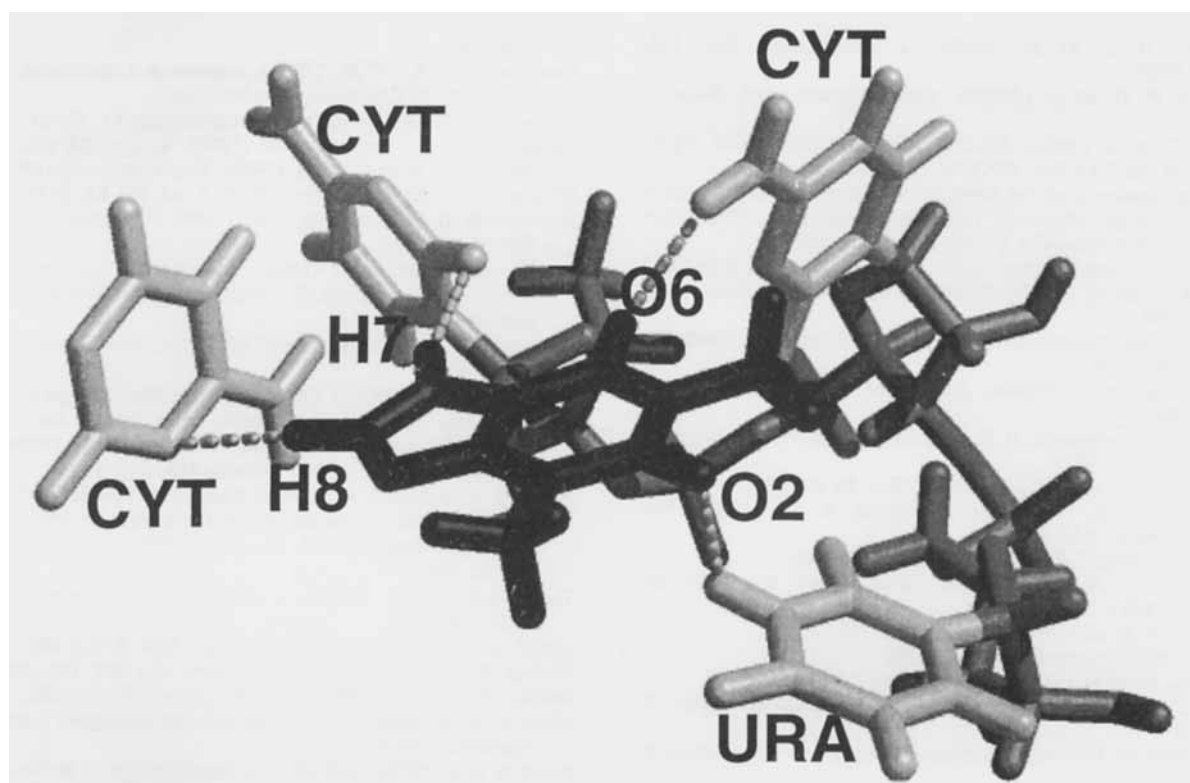
**Figure 7.** Four different modes of binding between the theophylline and the RNA molecule. Theophylline is plotted as space-filling model, while RNA is plotted as a stick model. For clarity, only residues closed to the binding pocket are shown. Each of the four panels corresponds to the lowest energy Monte Carlo equilibrated structure of a pool of 100 selected random structures.

selection against caffeine, theobromine and 7-methylxanthine, which have the N7 hydrogen replaced by a methyl group. This methyl group causes unfavourable steric and electrostatic interactions if the binding shows similar overall structural properties.

In the mTCT8-4 sequence, the C15pG16 end of the RNA is closed by a UCU loop (Jenison *et al.*, 1994). During unconstrained minimization, the base pair at the end tends to fold into the major groove. However, this hardly affects the structural properties of the binding pocket. Nevertheless, to avoid these problems the terminal base pair was constrained. Another problem encountered was that two of the three hydrogen bonds between theophylline and RNA bases are partly lost or substituted by hydrogen bonds with RNA backbone atoms. This is attributed to the significantly reduced electrostatic interactions when using a physically motivated dielectric constant of 80. With a dielectric constant of 1, charge interactions contribute considerably to hydrogen-bond energies. Therefore, distance constraints were applied for the hydrogen bonds between theophylline and the RNA bases.

Fig. 8 shows a capped-sticks representation of the

binding pocket after 50 000 steps of minimization (rms-deviation of the forces 0.006 kcal/mol-Å). Theophylline is tightly packed into the binding pocket. It is stacked with an adjacent guanine of the long helix. On the N3-C4-N9 side, theophylline is packed against an adenine of the short helix. Fig. 8 shows the hydrogen-bond interactions of O2, O6 and H7 with the CCU bulge. Also shown is the connecting ribose-phosphate backbone. The hydrogen H7 is interacting with a cytosine. If caffeine, 7-methylxanthine or 3,7-dimethylxanthine are to replace theophylline, no space will be available for the methyl group at the 7 position. This is consistent with the drop in binding affinity observed for these compounds (see Table 3). As shown in Fig. 8, the positively charged hydrogen H8 is interacting with a nitrogen of a base-paired cytosine, leaving no room for an oxygen at the 8 position, as in 1,3-dimethyl uric acid. This is consistent with the electrostatic CoMFA fields (previously described) and may explain the lack of binding affinity for this ligand. Thus, the energy-minimized structure of the drug-RNA complex can explain the weak binding of derivatives with large substitutes at the 7 and 8 positions under the assumption of similar binding modes.



**Figure 8.** Interactions of theophylline (black) with the CCU bulge and a base-paired cytosine in the energy-minimized structure. Bases are shown in grey, the ribose-phosphate backbone connecting the CCU bulge is shown in dark grey.

The two methyl groups are facing the opening and are partly exposed to solvent. Addition of the CP linker (used in the SELEX procedure) to the C1 carbon is possible without large structural changes. The minimized structure is therefore in agreement with the important pharmacophoric and SELEX-related constraints.

## Conclusion

Results from different molecular modeling techniques converged in obtaining the 3D structural model of a RNA molecule in the absence of extensive folding information. Small-molecule methods such as pharmacophoric mapping and structure-activity relationships of known ligands were used to provide structural constraints for RNA folding. The putative binding site for theophylline was identified by analyzing the available RNA sequences using secondary structure prediction methods and known experimental data. These preliminary results were instrumental in providing the structural details. A method to construct 3D model structures of folded RNA was developed using both stereochemistry and experimental observation. This folding method is general, conceptually simple and easy to use.

With this method, structural models of the RNA molecule mTCT8-4 were constructed. Further atomic refinement of these structural models was achieved by tailoring force field parameters for theophylline. The final model of the mTCT8-4 RNA molecule is in agreement with that of experimental observation, and may provide further insight into the molecular mechanisms of RNA-theophylline specificity. This model can be tested by NMR techniques: specific isotope-labeled C26, C27 and U28; and theophylline can be used to verify the hydrogen bond pattern described in Fig. 8. Coordinates for the model proposed in this paper are available from the authors via WWW site <http://www.isbr.lanl.gov/theory/coordinates>.

## Acknowledgements

The authors thank Dr Martijn Huynen for discussions on RNA secondary structure prediction. Dr David Weininger is acknowledged for permission to use the Daylight software and access to the MedChem94 database. Professor Sergio Clementi (Perugia, Italy) and Tripos Inc. (St Louis, MO) provided software. This work was supported by the US Department of Energy.

## References

- Baroni, M. *et al.* (1993). *Quant. Struct.-Act. Relat.* **12**, 9–20.
- Bernstein, F. C. *et al.* (1977). *J. Mol. Biol.* **112**, 535–542.
- Bock, L. C. *et al.* (1992). *Nature* **355** 564–566.
- Besler, B. H., Merz, K. M. and Kollman, P. A. (1990). *J. Comp. Chem.* **11**, 431–439.
- Catasti, P. *et al.* (1994). *Biochemistry* **33**, 3819–3830.
- Cramer, C. and Truhlar, D. (1992). *J. Comput. Aided Mol. Des.* **6**, 629–666.

- Cramer III, R. D. et al. (1988a). *J. Am. Chem. Soc.* **110**, 5959–5967.
- Cramer III, R. D. et al. (1988b). *Quant. Struct.- Act. Relat.* **7**, 18–25.
- Cruciani, G. et al. (1993). *3D-QSAR in Drug Design*, ed. by H. Kubinyi, pp. 551–564. ESCOM, Leiden.
- Daylight programs and the MedChem94 database are available from Daylight Chemical Information Systems, 18500 Von Karman Ave suite 450, Irvine, CA 92715 (1994).
- Ellington, A. D. and Szostak, J. W. (1990). *Nature* **346**, 818–822.
- Frisch, M. J. et al. (1992). *Gaussian 92*. Gaussian, Inc., Pittsburgh, PA.
- García, A. E. and Soumpasis, D. M. (1989). *Proc. Natl. Acad. Sci.* **86**, 3160–3164.
- García, A. E. et al. (1990). *J. Biomol. Struct. and Dyn.* **8**, 173–186.
- García, A. E., Soumpasis, D. M. and Jovin, T. M. (1994). *Biophys. J.* **66**, 1742–1755.
- Goodford, P. J. (1985). *J. Am. Chem. Soc.* **28**, 849–856.
- Gupta, G., García, A. E. and Hiriyanna, K. T. (1993a). *Biochemistry* **32**, 948–960.
- Gupta et al. (1993b). *Biochemistry* **32**, 7098–7103.
- Jellinek, D. et al. (1993). *Proc. Natl. Acad. Sci. USA* **90**, 11227–11231.
- Jenison, R. D. et al. (1994). *Science* **263**, 1425–1429.
- Leo, A. J. (1993). *Chemical Reviews* **93**, 1281–1306.
- Metropolis, N. et al. (1953). *J. Chem. Phys.* **21**, 1087–1092.
- Morley, S. D. et al. (1991). *J. Comput. Aided Mol. Des.* **5**, 475–504.
- Muthukumar, M. (1994a). *Macromol. Theory and Simulations* **3**, 61–71.
- Muthukumar, M. (1994b). *J. Chem. Phys.* **100**, 7796–7803.
- Opera, T. I. and García, A. E. (1996). *J. Comput. Aided Mol. Des.* In press.
- Pearlman, D. A. et al. (1991). *Amber 4.0* Technical report, University of California, San Francisco.
- Poncelet, S. M. et al. (1990). *J. Immunoassay* **11**, 77–88.
- Sassanfar, M. and Szostak, J. W. (1993). *Nature* **364**, 550–553.
- Schildkraut, C. and Lifson, S. (1965). *Biopolymers* **3**, 195–208.
- Soumpasis, D. M. (1984). *Proc. Natl. Acad. Sci.* **81**, 5116–5120.
- Soumpasis, D. M. and Tung, C.-S. (1988). *J. Biomol. Struct. and Dyn.* **6**, 397–420.
- Soumpasis, D. M. et al. (1990). *Theoretical Biochemistry and Molecular Biophysics*, R. Lavery and D. L. Beveridge (Eds.), Adenine Press, NY.
- Srivastava, D. and Muthukumar, M. (1994). *Macromolecules* **27**, 1461–1465.
- Stewart, J. J. P. (1990). *J. Comput. Aided Mol. Design* **4**, 1–105.
- SYBYL™ and CoMFA are available from Tripos, Inc., 1699 S Hanley Rd, Suite 303, St Louis, MO 63144, USA (1994).
- Turek, C. and Gold, L. (1990). *Science* **249**, 505–510.
- Turek, C. et al. (1992). *Proc. Natl. Acad. Sci. USA* **89**, 6988–6992.
- Tung, C.-S. (1993). *Computation of Biomolecular Structures*. D. M. Soumpasis and T. M. Jovin (Eds.) 87–97, Springer-Verlag, NY.
- Tung, C.-S. et al. (1994). *J. Biomol. Struct. and Dyn.* **11**, 1327–1344.
- Vinter, J. (1994). *J. Comput. Aided Mol. Des.* **8**, 653–668.
- Weiner, S. J. et al. (1984). *J. Am. Chem. Soc.* **106**, 765–784.
- Weiner, S. J. et al. (1986). *J. Comput. Chem.* **7**, 230–252.
- Wold, S. et al. (1987). *Chemometrics and Intelligent Laboratory Systems* **2**, 37–52.
- Wold, S. et al. (1993). *3D-QSAR in Drug Design*. H. Kubinyi (Ed.), 523–550, ESCOM, Leiden.
- Zuker, M. (1989). *Science*, **244**, 48–52.