Quantum Mechanical Energy-Based Screening of Combinatorially Generated Library of Tautomers. TauTGen: A Tautomer Generator Program

Maciej Harańczyk[†,‡] and Maciej Gutowski*[,†,‡,§]

Department of Chemistry, University of Gdańsk, 80-952 Gdańsk, Poland, Chemical Sciences Division, Fundamental Sciences Directorate, Pacific Northwest National Laboratory, Richland, Washington 99352, and Chemistry-School of Engineering and Physical Sciencs, Heriot-Watt University, Edinburgh EH14 4AS, U.K.

Received June 27, 2006

We describe a procedure of finding low-energy tautomers of a molecule. The procedure consists of (i) combinatorial generation of a library of tautomers, (ii) screening based on the results of geometry optimization of initial structures performed at the density functional level of theory, and (iii) final refinement of geometry for the top hits at the second-order Möller−Plesset level of theory followed by single-point energy calculations at the coupled cluster level of theory with single, double, and perturbative triple excitations. The library of initial structures of various tautomers is generated with TauTGen, a tautomer generator program. The procedure proved to be successful for these molecular systems for which common chemical knowledge had not been sufficient to predict the most stable structures.

INTRODUCTION

Computer programs are nowadays successfully used by chemists in the areas of drug and materials design,[1,2] petroleum chemistry,[3] and modeling of pollutants contaminating natural ecosystems.[4] Powerful computers and efficient computational methods are critical in virtual screening of libraries of compounds. Virtual-screening methods have been primarily applied for drug design, but nowadays the spectrum of possible applications is becoming much broader. For example, they have been used in the design of molecular receptors with binding sites that complement metal ion guests.[5] Extended and diverse libraries of compounds are typically developed using combinatorial chemistry methods.

For molecules in the library one can assign descriptors—numbers that describe structure and chemical properties. Common examples of descriptors are molecular weight and number of H-bond donors and acceptors, whereas examples of properties are ionization potential, deprotonation energy, or solvent accessible surface. Typical QSAR software packages allow calculating thousands of descriptors. Having a library of compounds with determined descriptors and properties, the virtual screening needs to be performed in order to rank the compounds in the library against the requested pattern.

Because of the large size of combinatorially generated libraries, the calculation of descriptors and properties has to be fast. This implies that typically only very approximate computational methods are employed. Among electronic structure methods, semiempirical and tight-binding approaches provide the very first choice. Qualitatively different situation is encountered when studying tautomeric equilibria. In these cases only one but very high quality descriptor is required—free energy resulting from an accurate energy at $T = 0$ K and thermal corrections. Tautomers are products of intramolecular proton-transfer reactions which are one of the most basic chemical processes in biology. The basic building blocks of living cells—amino acids and nucleic acid bases (NABs)—might undergo tautomeric reactions. The tautomeric form of each NAB that predominates in DNA or RNA is called the canonical form. The intra- as well as intermolecular tautomerizations involving NABs have long been suggested as critical steps in mutations of the DNA genetic material.[6−8] The energy differences between structurally different tautomers might be as small as a fraction of a kcal/mol. Thus a meaningful study of the relative stability of tautomers typically requires employing the most accurate ab initio methods. For example, the relative energies of tautomers of cytosine in the gas phase predicted at the second-order Möller−Plesset (MP2) level of theory differs qualitatively from the relative energies predicted at the coupled cluster of theory with single, double, and perturbative triple excitations (CCSD(T)), suggesting that only the latter method might be accurate enough for determining the relative stability of close-lying tautomers.[9]

In the studies of tautomeric equilibrium, the first selection of potentially important tautomers is typically done based on common organic chemistry knowledge. There are also software tools available for generation of tautomers.[10−12] Typically, they first identify proton donor and acceptor sites. Next, a library of compounds is generated with various tautomers resulting from proton transfers between electronegative atoms, such as N or O.

Recently we have been studying anions of nucleic acid bases which are expected to be important in radiation induced mutagenesis. Contrary to earlier experimental and computational predictions, we demonstrated that the most stable valence anions of pyrimidine bases, such as 1-methylcytosine,[13] uracil,[14] and thymine,[15] do not result from proton transfer between electronegative atoms, N or O. Instead they

* Corresponding author e-mail: m.gutowski@hw.ac.uk.
† University of Gdańsk.
‡ Pacific Northwest National Laboratory.
§ Heriot-Watt University.

10.1021/ci6002703 CCC: $37.00
Published on Web 12/29/2006

result from enamine-imine transformations, i.e., a proton is transferred between a NH site to a carbon site.

Some of these valence anions proved to be adiabatically bound with respect to the most stable tautomers of neutral NABs. It was an important finding because so far it was believed that the only adiabatically bound anions of NABs have a dipole-bound character.[16] The importance of valence anions results from the fact that dipole-bound anions are strongly perturbed by other atoms or molecules, and their relevance in condensed phase environments is questionable. Our discovery of adiabatically bound valence anions of pyrimidine NABs was facilitated by a series of studies on proton-transfer reactions in anionic complexes of NABs with various proton donors.[17,18]

Our initial studies were focused on pyrimidine NABs because the number of potentially relevant tautomers was managable—a few tens of structures.[13−15] A natural next step would be to study analogous anionic tautomers for the purine NABs: guanine and adenine. Unfortunately, the number of tautomeric structures for which we would like to perform prescreening using the DFT level of theory becomes as large as 500−700. Interestingly, this is not so much the computer time but rather the human time required to prepare, run, and analyze the calculations, which becomes prohibitive. Clearly this problem should be handled using a hybrid approach involving both combinatorial and accurate quantum chemical methods. In addition, the most promising tautomers might result from enamine-imine transformations of canonical tautomers, i.e., a proton is transferred between N and C atoms. These possibilities are not taken into account in the available software for generation of tautomers, in which only typical proton donor and acceptor sites, i.e., N(O)H and O(N) centers, respectively, are engaged in proton-transfer steps, and the resulting tautomers are enumerated accordingly.[10−12] To overcome these limitations, we have developed a new program for generation of tautomers, TauTGen, which is described in the next section. This program builds all possible tautomers from a molecular framework (the core) and a specified number of hydrogen atoms. The hydrogens are attached to the sites specified by a user and a library of tautomers is combinatorally generated within a user-defined list of constraints. The prescreening is performed based on the results of DFT geometry optimizations. We call this approach "energy-based virtual screening" because, in contrast to the "structure-based virtual screening", the most stable tautomers are the target of this screening. The geometries of the top hits identified in the B3LYP energy-based screening were further optimized at the MP2 level of theory, and final energies were calculated at the CCSD(T) level. A good measure of success of this approach is our finding about valence anions of guanine. This base, which was believed to have the smallest electron affinity among nucleobases,[19] supports at least 13 anionic tautomers, which are *adiabatically bound* with respect to the neutral canonical tautomer.[20] The most stable anion of guanine is adiabatically bound by as much as 8.5 kcal/mol. Using the same approach, we found at least one anion of adenine that is adiabatically bound with respect to the neutral canonical tautomer, the adiabatic electron affinity is 0.9 kcal/mol. The approach applied to cytosine demonstrated that it does not support an adiabatically bound valence anion. Finally, we have performed a preliminary screening of tautomers of cationic uracil to explore the posibility of formation of unusual tautomers. Although we have not found those, we demonstrated that the relative energy differences between the most stable tautomers are much smaller for the cationic than for the neutral species.

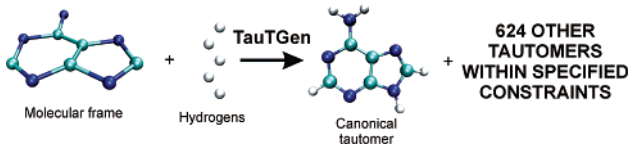

Figure 1. TauTGen uses a fixed frame of heavy atoms and a given number of hydrogen atoms to create tautomers of the resulting molecular system.

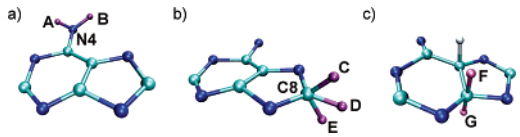

Figure 2. Information needed to define sites for hydrogen attachment. The sites are marked with letters A−G.

METHODS

TauTGen − A Program To Create a Library of Tautomers. The TauTGen program was written in the C programming language with the purpose of generating a library of tautomers for a given molecule. TauTGen constructs tautomers from a molecular frame built of heavy atoms and a given number of hydrogens (Figure 1). The user has to provide geometry of the molecular frame and to specify the minimum and maximum number of hydrogen atoms connected to each heavy atom. Sites for placement of hydrogen atoms are also defined by the user. To define a site, the user has to provide the following information:

- Name - a string of characters used to build up a filename for each tautomer
- A point where the hydrogen atom is to be placed. The point is defined relative to the fixed molecular frame
- Information which heavy atom is the holder of this site (connectivity information)
- The required total number of hydrogen atoms assigned to the heavy atom which would make the specific site available for occupation (a site constraint)
- Stereoconfiguration information, which tells the program if occupying a particular site will lead to the *R* or *S* configuration of the connected heavy atom.

Special care is taken to precisely name the sites. These names are used to create the names of tautomers that are later used as the filenames. For example, sites A and B (Figure 2a) are named "N4cis" and "N4trans" to distinguish possible rotamers resulting from rotation of the N4H imino group. The connectivity information is used to count the number of hydrogen atoms at each heavy atom, N_s. The number of available sites for hydrogen might be 2 even when $N_s = 1$, see for example the C6 site of adenine, Figure 3a and Table 1. Each site has a defined constraint, which tells for which values of N_s the site becomes available for occupation. This is what we mean by the site constraint. This option is used to build proper hybridizations of heavy atoms. For example, the C and E sites (Figure 2b) are occupied only when $N_s = 2$ for C8. Then C8 attains the sp^3

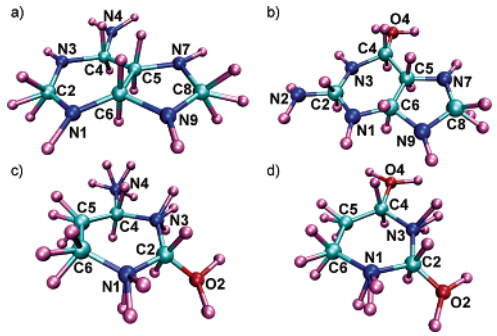

Figure 3. Molecular frameworks of adenine (a), guanine (b), cytosine (c) and uracil (d) with all sites for hydrogen attachment. The total number of sites differs from the number of sites in Tables 1−4 because some sites overlap.[40]

Table 1. Set of Constraints Used When Searching for the Most Stable Tautomers of Anionic Adenine

	min. and max. no. of hydrogen atoms at heavy atom		no. of available sites for each no. of hydrogens at heavy atom (N_s=1 and 2)		
atom	min.	max.	$N_s = 1$	$N_s = 2$	asymmetric atom
N1	0	1	1		
C2	0	2	1	2	
N3	0	1	1		
C4	0	1	2		yes
N4	1	2	2	2	
C5	0	1	2		yes
C6	0	1	2		yes
N7	0	1	1		
C8	0	2	1	2	
N9	0	1	1		

hybridization. On the other hand, the D site is occupied only when $N_s = 1$ for C8—the sp^2 hybridization is then assigned to C8.

If a user wants to generate stereoisomers, then two sites have to be used for each asymmetric atom in order to describe the *R* and *S* configurations. In the case of planar or nearly planar NABs the sites F and G that are "below" and "above" the molecular plane might be distinct (Figure 2c). Each of these sites bears additional information describing the configuration, e.g., 1 or 2 for the "above" or "below" configuration, respectively.

As soon as the framework, available sites, the total number of hydrogen atoms $N_{hydrogens}$, and all constraints are defined, TauTGen generates all possible distributions of $N_{hydrogens}$ hydrogens among N_{sites} sites. For each distribution TauTGen checks whether all applied constraints are respected. The constraints are checked in the following order:

- Constraints on the maximum and minimum number of hydrogens connected to each heavy atom
- Site constraints; check if the sites are used consistently with the actual values of N_s
- Stereoconfiguration; check whether other enantiomer has already been generated (this check is not done by default).

Each new distribution needs to pass all these checks to become an entry in the library of tautomers.

The stereoconfiguration check is done by a separate routine that detects enantiomers of a given distribution. If an enantiomer of the previously generated stereoisomer has been built, the distribution is rejected so the final set of stereoisomers consists of diasteroisomers only. The following steps are parts of the stereoconfiguration check:

- A stereoconfiguration fingerprint is assigned to each new distribution. The fingerprint contains information if hydrogens occupying stereosensitive sites are above or below the molecular plane. In other words, we keep track whether the involved heavy atoms are *R* or *S*.
- An inverse stereoconfiguration fingerprint is created for the distribution. It is then compared against the stereoconfiguration fingerprints of all previously generated stereoisomers of the same tautomer.
- If there is no match between the fingerprints, the current distribution is a diastereoisomer of the previously generated stereoisomers, and it is accepted to the library. If there is a match, then the current distribution is an enantiomer and hence it is rejected.

Finally, TauTGen generates filenames and saves atomic coordinates of each member of the library to a separate file. The filename is a string of these site names that were used to build up the molecule. If proper site names are defined, the filename can uniquely name the molecular structure and discriminate various rotamers of the same tautomer. To facilitate files management, we sometimes divide structures among groups and subgroups based on the values of N_s for preselected heavy atoms. If stereoisomers were generated, the tautomer name is supplemented with a stereoconfiguration label: e.g. "Z_nml" where n, m, and l are names of these sites that are on the same side of the molecular plane.

The source code of TauTGen is available free of charge and can be downloaded from the Sourceforge Internet archive.[21] The manual of TauTGen is available online and includes examples of input files.

Screening. We used simple UNIX shell scripts to automate the screening procedure. The initial geometries of molecular structures are expressed in Cartesian coordinates and stored in typical xyz files. They are used to build input files to the Gaussian03[22] program using a csh shell script. Initial screening is performed at the DFT level of theory with a B3LYP exchange-correlation functional[23] and 6-31+G** basis set for adenine and uracil and 6-31++G** basis set for guanine and cytosine. A tendency of B3LYP to overestimate the excess electron binding energy helps to avoid false negatives when screening for adiabatically bound anions. The 6-31+G** and 6-31++G** basis sets[22] have an advantage that the time required to perform geometry optimization for a NAB is acceptable. This choice of the method and the basis sets was also supported by our earlier experience with calculation of adiabatic electron affinities (AEAs) for some pyrimidine NABs.[13−15]

It is known that "buckling" of the ring of a NAB might increase the electronic stability of the anion, because the excess electron typically occupies a π^* orbital.[13−15] For this reason, all initial structures of anions were built from buckled molecular frames. In the case of about 15% of generated structures, the initial try of the self-consistent field procedure (SCF) failed to converge. In these cases we applied one, or a combination of up to four approaches: (a) start the calculation from orbitals generated with a smaller basis set (3-21G or 6-31G*), (b) start the calculation from orbitals generated in water solution simulated with the IEF-PCM method and the cavity built up using the United Atom (UA0) model,[24] (c) try to converge the SCF procedure using a quadratically converging algorithm, and (d) start the calculation from a slightly distorted geometry (the distortion was

Table 2. Set of Constraints Used When Searching for the Most Stable Tautomers of Anionic Guanine

	min. and max. no. of hydrogen atoms at heavy atom		no. of available sites for each no. of hydrogens at heavy atom (N_s=1 and 2)		asymmetric atom
	min.	max.	$N_s = 1$	$N_s = 2$	
N1	0	1	1		
C2	0	1	2		yes
N2	1	2	2	2	
N3	0	1	1		
C4	0	1	2		yes
O4	0	2	2	2	
C5	0	1	2		yes
C6	0	1	2		yes
N7	0	1	1		
C8	1	2	1	2	
N9	0	1	1		

introduced by performing 2 optimization steps but for the neutral molecule). In consequence, we recorded only a few cases when the SCF procedure failed to converge for the initial structure of the anion. All screening calculations of adenine, cytosine, uracil, and some of guanine were performed using Gaussian03[22] on dual Intel Itanium2 nodes. The remaining tautomers of guanine were calculated using NWChem[25] on an SGI Altix computer. For ionic systems, the B3LYP geometry optimizations were followed by single point calculations for neutral systems at the optimal ionic geometries.

We developed Gaussian Output Tools (GOT) scripts[26] to analyze output files from Gaussian03. The GOT scripts are written in the Practical Extraction and Report (Perl) language and can extract final energies, geometries, and forces from the Gaussian03 output files. Analogous scripts were developed for NWChem output files. Other shell scripts were used to identify and restart the calculations of tautomers for which the SCF or geometry optimization failed to converge. The final B3LYP energies for the cationic, neutral, and anionic species were copied to a Microsoft Office Excel spreadsheet, which was used to calculate the relative energies as well as AEAs and electron vertical detachment energies (VDEs) (anions) and adiabatic and vertical ionization potentials (cations). The spreadsheet was also used to sort the molecular structures according to their relative energies. In some cases we found that two or more initial structures converged to the same energy. We have analyzed these cases, in addition to the most stable tautomers, using the Molden software package.[27] In all these cases, the same energy resulted from the same converged structure.

The Final Energy Calculations for Anions. The anionic tautomers with positive values of AEA determined at the B3LYP level were further optimized at the MP2 level of theory with augmented correlation-consistent polarized basis sets of double-ζ quality (AVDZ).[28] The final single-point calculations were performed at the coupled cluster level of theory with single, double, and noniterative triple excitations (CCSD(T)/AVDZ)[29] at the optimal MP2 geometries. The open-shell CCSD(T) calculations were carried out at the R/UCCSD(T) level. In this approach, a restricted open shell Hartree−Fock calculation was initially performed to generate the set of molecular orbitals, and the spin constraint was relaxed in the coupled cluster calculation.[30] The 1s orbitals of carbon, nitrogen, and oxygen atoms were excluded from the MP2 and coupled-cluster treatments.

The relative energies of the anion with respect to the most stable tautomer of the neutral were first corrected for the energies of zero-point vibrations to derive the values of AEA. Next, thermal corrections as well as the entropy terms, calculated at the MP2 level for $T = 298$ K and $p = 1$ atm in the harmonic oscillator-rigid rotor approximation, were included to derive the relative stability in terms of Gibbs free energy.

The MP2 geometry optimizations and frequency calculations were performed with Gaussian03[22] and the CCSD(T) calculations with the MOLPRO[31] package. The codes were run on clusters of dual Intel Itanium2 nodes with and without Quadrics interconnect.

RESULTS

Calculations for Anionic Purine Nucleic Acid Bases. We have tested our algorithms and the TauTGen program on anionic tautomers of adenine and guanine. A set of constraints used for adenine is presented in Table 1. We defined 20 sites available for hydrogen attachment (Table 1 and Figure 3a). Fourteen sites were active for heavy atoms with $N_s = 1$. Within these 14 sites, 2 sites were available to build rotamers of the N4 imino group, and 2 sites were available for each of the C4, C5, and C6 atoms to build stereoisomers with different positions of hydrogens in relation to the molecular plane. An additional 6 sites were available to build tautomers with two hydrogens at the C2, N4, and C8 atoms. Similar constraints were defined for guanine and are presented in Table 2 and Figure 3b. In the case of guanine we created 23 sites available for hydrogen attachment. Seventeen sites were available for heavy atoms with $N_s = 1$. Within these 17 sites, 4 sites were available to build rotamers of the N2 imino and O4 hydroxy groups, and 2 sites were available for each of the C2, C4, C5, and C6 atoms to build stereoisomers with different positions of hydrogens in relation to the molecular plane. Additional 6 sites were available to build tautomers with two hydrogen atoms at N2, O4, and C8.

The final molecular frames and sites are displayed in Figure 3a,b. The preparation time of a TauTGen input file was estimated to be about 15 min for each of the bases. The majority of this time is consumed by manually drawing the sites using the Molden software package[27] and naming them. This process could be automated if a larger number of molecules had to be studied.

Within these constraints TauTGen generated 625 unique structures for adenine and 499 structures for guanine. In the course of generation of tautomers of adenine, TauTGen generated initially 15 504 distributions of five hydrogen atoms among 20 sites, from which only 9192 tautomers passed a check for the minimum and maximum number of hydrogens at each heavy atom. Only 1148 of them passed the site constraint check. This number was later reduced to 625 in the course of the stereoconfiguration check. In the case of guanine the TauTGen program generated initially 33 649 distributions, which were later reduced to 9768, 907, and finally to 499 in the series of constraint checks.

The 499 structures of guanine were optimized at the B3LYP/6-31++G** level, and the 625 structures of adenine were optimized at the B3LYP/6-31+G** level. An average calculation time for one structure was about 4 and 3 h for

guanine and adenine, respectively, on a dual Intel Itanium2 node. With this speed of calculations, one needs less than 2000 node hours to screen guanine or adenine at the DFT level. A parallel execution of jobs might further shorten the "wall time" required for screening. Indeed, it took us about 2 weeks to screen 625 tautomers of adenine with an unprivileged access to a 128 dual Itanium2 nodes cluster in the TASK academic computer center in Gdańsk.[32] The wall time includes the time when jobs waited in the queuing system. A time scale of 2 weeks is comparable with the time required to perform one MP2/AVDZ calculation of numerical frequencies for anionic guanine on the same dual Itanium2 node.

With the goal being the determination of adiabatically bound anions of purine bases, we compared the final B3LYP energies for anions with the B3LYP energy of the neutral canonical tautomer at its optimal geometry. The histograms presenting the resulting AEA values for all structures are presented in Figure 4. It might be seen that the values of AEA smoothly decrease for about 90% of the structures. A sudden decrease of AEA for the remaining 10% of the structures is related to the fact that some of these structures decompose in the course of geometry optimization. The three most stable structures for each ionic nucleobase are presented in Figure 5.

In case of guanine and the B3LYP/6-31++G** level of theory, we found 14 anionic tautomers which were *more stable* than the canonical neutral. All of them were further studied at the MP2 and CCSD(T) levels with the AVDZ basis set. These calculations revealed that 13 tautomers *support adiabatically bound anions*. The most stable anion is characterized by an AEA of 8.5 kcal/mol. Moreover, three tautomers have a hydrogen atom at N9, where a sugar unit is connected when guanine is incorporated into DNA or RNA.[33]

From the 625 anionic tautomers of adenine only one has lower energy than the neutral canonical tautomer at the B3LYP/6-31+G** level of theory. This tautomer is presented in Figure 5a together with other two most stable tautomers. The following MP2 geometry optimization and the single point CCSD(T)/AVDZ energy calculation revealed that the tautomer is adiabatically bound with respect to canonical neutral by 0.9 kcal/mol. The chemical aspects of this study will be described elsewhere.[33]

Calculation for Anionic Cytosine. We have also applied our algorithms and the TauTGen program to identify the most stable anionic tautomers of cytosine. This time, because of the smaller molecular size, we decided to explore a larger tautomeric space than in the case of purine bases. We have included tautomers that have two hydrogen atoms on the nitrogen atoms in the ring (N1, N3), see Figure 3c. We have also included tautomers with a protononed amino group. A set of constraints used for cytosine is presented in Table 3. We defined 27 sites available for the attachment of hydrogens (Table 3 and Figure 3c). Twelve sites were active for heavy atoms with $N_s = 1$. Within these 12 sites, 2 sites were available to build rotamers of the O2 hydroxy group, and 2 sites were available to build rotamers of the N4 imino group. Two sites were available for each of the C2 and C4 atoms to build stereoisomers with different positions of hydrogens in relation to the molecular plane. Additional 12 sites were available to build tautomers with two hydrogens at the N1,

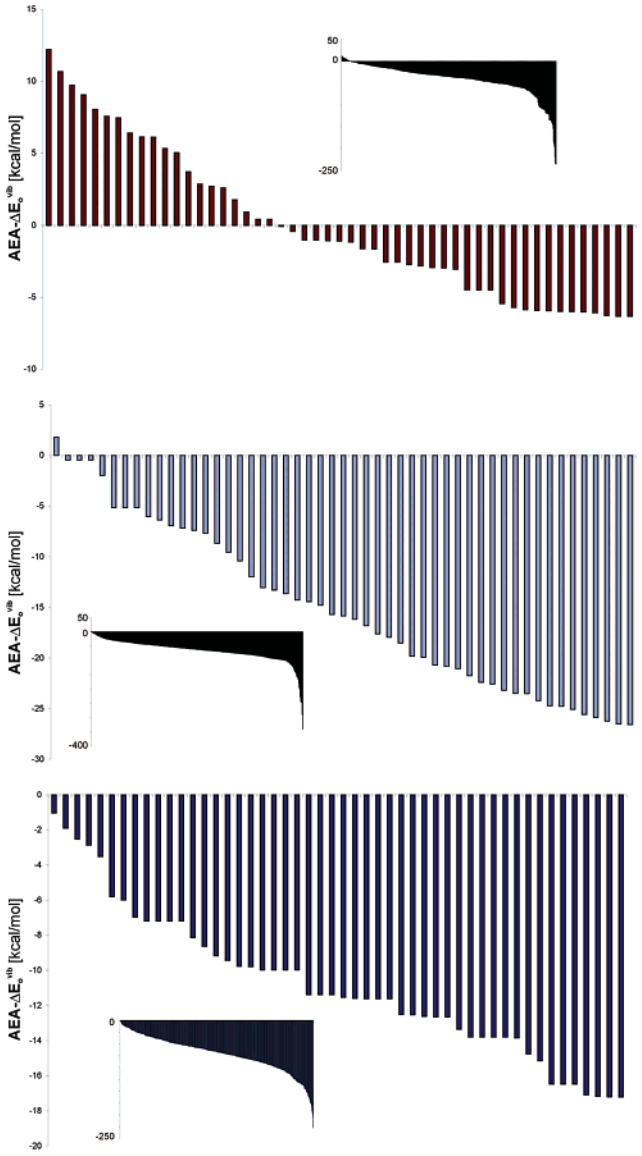

Figure 4. Adiabatic electron affinity (AEA) for tautomers of guanine (top), adenine (middle), and cytosine (bottom) calculated at the B3LYP/6-31++G**, B3LYP/6-31+G**, and B3LYP/6-31++G** levels of theory, respectively. The values of AEA for the 50 most stable tautomers are presented on larger plots. The smaller plots present the AEA of all tautomers. The tautomers are ordered according to the decreasing value of AEA.

O2, N3, N4, C5, and C6 atoms. Three sites were used to build a protonated amino group with 3 hydrogens attached to the N4 atom. The final molecular frame and sites are displayed in Figure 3c.

Within these constraints TauTGen generated 753 unique tautomers. TauTGen generated initially 80 730 distributions of five hydrogen atoms among 27 sites, from which 73 309 tautomers passed a check for the minimum and maximum number of hydrogens at each heavy atom. Only 1255 of them passed the site constraint check. The number was later reduce to the final 753 tautomers in the stereoconfiguration check.

The 753 structures were optimized at the B3LYP/6-31++G** level. An average calculation time per one structure was less than 3 h on a dual Intel Itanium2 node. The total time of calculations at the DFT level did not exceed 2000 node hours. Similarly to the case of purine bases, we tried to identify adiabatically bound anions of cytosine by

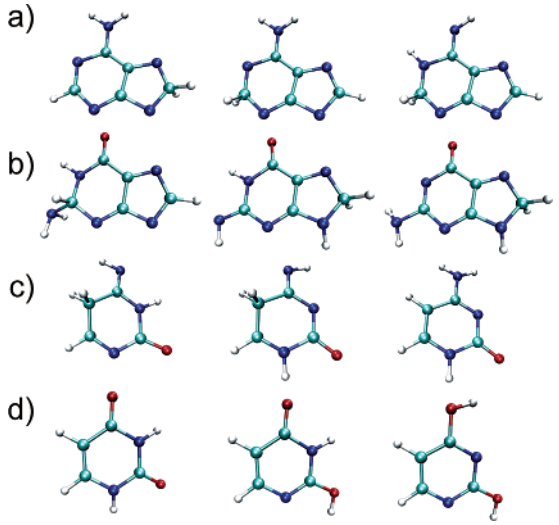

Figure 5. Top three most stable tautomers of anions of adenine (a), guanine (b), cytosine (c), and cations of uracil (d).

Table 3. Set of Constraints Used When Searching for the Most Stable Tautomers of Anionic Cytosine

	min. and max. no. of hydrogen atoms at heavy atom		no. of available sites for each no. of hydrogens at heavy atom (N_s=1, 2 and 3)			asymmetric
atom	min.	max.	$N_s = 1$	$N_s = 2$	$N_s = 3$	atom
N1	0	2	1	2		
C2	0	1	2			yes
O2	0	2	2	2		
N3	0	2	1	2		
C4	0	1	2			yes
N4	0	3	2	2	3	
C5	0	2	1	2		
C6	0	2	1	2		

comparing the final B3LYP energies for anions with the B3LYP energy of the neutral canonical tautomer at its optimal geometry. A histogram presenting the resulting AEA values for all structures is presented in the bottom of Figure 4. The values of AEA smoothly decrease for about 80% of the structures. A sudden decrease of the AEA for the remaining 20% structures is related to the fact that some of these structures decompose in the course of geometry optimization.

In case of cytosine and the B3LYP/6-31++G** level of theory, we have not found any anionic tautomer which would be adiabatically bound with respect to the canonical neutral tautomer. The three most stable structures are presented in Figure 5c. The eight most stable tautomers were further studied at the MP2 and CCSD(T) levels with the AVDZ basis set. These calculations revealed that the most stable valence anions of cytosine result from enamine-imine transformations of the canonical tautomer, but none of them is adiabatically bound. Still, the value of AEA is as small as −0.3 kcal/mol and the accompanying value of VDE is 1.1 eV. Moreover, the most stable anionic tautomer that is also biologically relevant, i.e., has a hydrogen atom at N1, where a sugar unit is connected, is characterized by a AEA of −2.1 kcal/mol.[33] Again, this tautomer results from an enamine-imine transformation of the canonical tautomer.

Calculations for Cationic Uracil. Our preliminary results for ionized uracil are presented below. Because we have no information about any unconventional tautomer being significantly stable, we decided to explore a broad tautomeric space. We have included tautomers which have two hydrogen atoms at the nitrogen and carbon atoms in the ring. We have also included tautomers with protononed hydroxy groups. A full set of constraints used for uracil is presented in Table 4. We defined 24 sites available for hydrogen attachment (Table 4 and Figure 3d). Twelve sites were active for heavy atoms with $N_s = 1$. Within these 10 sites, 4 sites were available to build rotamers of the O2 and O4 hydroxy groups. Another 4 sites, two on each of C2 and C4 atoms, were placed to built stereoisomers. Additional 12 sites were available to build tautomers with two hydrogens at the N1, O2, N3, O4, C5, and C6 atoms. The final molecular frame and sites are displayed in Figure 3d.

Table 4. Set of Constraints Used When Searching for the Most Stable Tautomers of Ionized Uracil

	min. and max. no. of hydrogen atoms at heavy atom		no. of available sites for each no. of hydrogens at heavy atom (N_s=1 and 2)		asymmetric
atom	min.	max.	$N_s = 1$	$N_s = 2$	atom
N1	0	2	1	2	
C2	0	1	2		yes
O2	0	2	2	2	
N3	0	2	1	2	
C4	0	1	2		yes
O4	0	2	2	2	
C5	0	2	1	2	
C6	0	2	1	2	

TauTGen generated initially 10 626 distributions of four hydrogen atoms among 24 sites, from which 9919 tautomers passed a check for the minimum and maximum number of hydrogens at each heavy atom. Only 624 of them passed the site constraint check. The latter number was later reduced to 392 with enantiomer detection and removal routine.

The 392 structures were optimized at the B3LYP/6-31+G** level. An average calculation time was more than 2 h per structure on a dual Intel Itanium2 node. The total time of calcluations at the DFT level did not exceed 1000 node hours. The canonical tautomer proved to be the most stable for U^+. The three most stable structures are presented in Figure 5d. In Figure 6 we plot relative energies (ΔE) of tautomers determined with respect to the most stable (canonical) tautomer. A histogram presenting the values of ΔE is presented in Figure 6. It might be seen that the values of ΔE smoothly increase for about 90% of the structures. The last 10% of the structures correspond to these tautomers that decompose to very unstable species.

The six most stable cationic tautomers, together with the corresponding neutrals, were further studied at the B3LYP/6-31++G** level of theory. The major findings so far are (a) the most stable tautomer of cationic uracil is canonical and (b) the relative energy differences between the most stable tautomers are much smaller for the cationic than for the neutral species (see the bottom of Figure 6). These findings will be verified at the MP2 and CCSD(T) levels of theory in a separate study.[33]

CONCLUSIONS

We have proposed a procedure of finding the most stable tautomers of a molecule by performing energy-based screening of combinatorially generated tautomers. The procedure

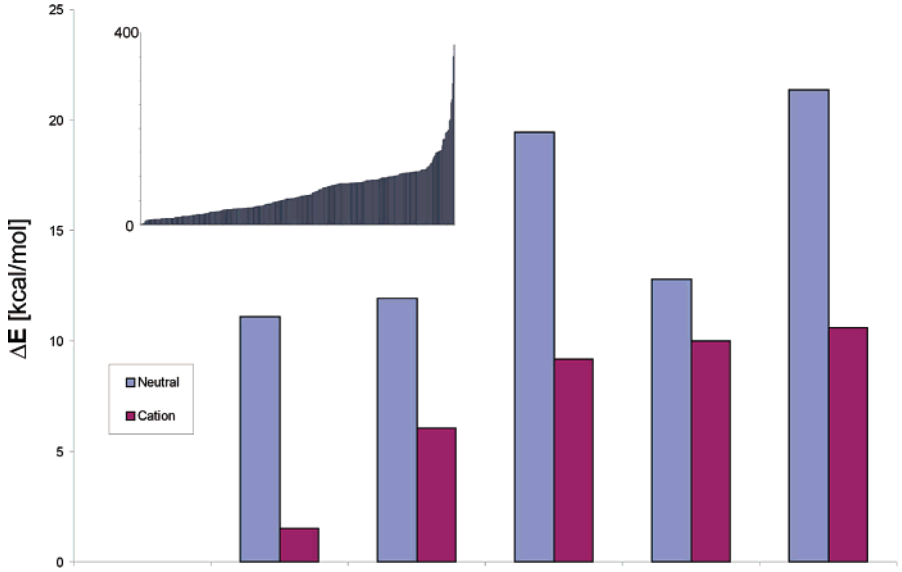

Figure 6. The relative energies of tautomers of cationic uracil calculated at the B3LYP/6-31+G** level of theory (top). The values are calculated with respect to the most stable tautomer (canonical), and they are sorted in ascending order. The values for the most stable tautomers of the neutral and cationic species are presented on a larger plot (bottom).

consists of 3 steps: (i) combinatorial generation of structures with the TauTGen program, (ii) prescreening based on DFT energies of optimized structures, and (iii) high level electronic structure calculations with thermal corrections to enthalpy and entropy for the most stable structures obtained in step (ii).

TauTGen program has been developed to generate all tautomers including those resulting from enamine-imine transformations. It builds molecules from a given molecular frame of heavy atoms and a specified number of hydrogen atoms. Hydrogen atoms are attached at the sites of possible hydrogen placement, which are specified by the user. These sites are combinatorially occupied to generate all possible tautomers. Then each distribution of hydrogens is tested against a set of constraints on (i) the maximum and minimum number of hydrogen atoms at each of the heavy atoms and (ii) the proper occupation of each site. These "occupation" tests are followed by a stereoconfiguration test. "Chemical intuition" is used only to define the sites of possible hydrogen atom attachment and to set constraints on the maximum and minimum number of hydrogens attached to each heavy atom. This approach gives more flexibility than other tautomer generation programs, which typically use chemical intuition to identify sites involved in intramolecular proton transfer.

The proposed procedure of finding low-energy tautomers proved to be very efficient and successful when applied to charged nucleic acid bases. In the case of adenine, guanine, and cytosine we found many anionic tautomers that are more stable than the anions of canonical tautomers. Moreover, in the case of guanine, which was believed not to support an adiabatically bound valence anion, we found 13 anions that are bound. In the case of adenine we found one adiabatically bound valence anion. In the case of cytosine, we demonstrated that it does not support an adiabatically bound anion, but the most stable tautomers of its valence anion result from enamine-imine transformations of the canonical tautomer.

The proposed procedure of finding low-energy structures has also been applied to ionized uracil. The six most stable tautomers were found to be the same as in the case of neutral species. However, the relative energy differences between the most stable tautomers are much smaller for the cationic than for the neutral species.

A time scale for performing energy-based screening of a combinatorially generated library of tautomers was noted. It takes about 15 min to manually prepare an input file for the TauTGen program for a system of the size of adenine or guanine. It takes about 2 s to generate a library of tautomers on Intel Pentium4 1.6 Ghz. The Gaussian03 B3LYP/6-31+G** energy-based screening for the computationally most demanding library, 625 structures of adenine, took less than 2000 node hours on the dual Intel Itanium2 nodes. Since the jobs can be executed in parallel, the wall time of the screening may be shortened to a few days. It makes the proposed procedure very attractive to solve problems of the relative stability of tautomers, when common chemistry knowledge fails to predict the most stable structures.

It is well-established that the environment may have tremendous impact on the relative stability of tautomers. Polar solvents may stabilize tautomers that are otherwise unstable in the gas phase. An evident example of such stabilization is formation of zwitterionic forms of amino acids in water solutions. This phenomenon is essential for enzyme catalysis that is based on preorganized electrostatics.[34] Therefore, the next version of our approach will have a new capability to perform searches for the most stable tautomers in water solution or any other solvent. This would require using the solvent reaction field model of Tomasi et al.[24] or the Langevin dipole method of Florian and Warshel[35] when performing quantum chemical calculations. The methods for calculations of the free energy of solvation by explicit solvent treatment and thermodynamic perturbation/integration could also be used to simulate solvent effects.[36−39] These calculations are typically more time-consuming than gas-phase calculations but will provide results better tailored to the users' interests.

Other extensions of TauTGen will include applications of structure similarity tools. Additional tests will be implemented to exclude topological symmetry duplicates, which

currently can sneak into the library with input files created in an unfortuitous manner. They will also automate the process of analysis of optimized molecular geometries at the stage of prescreening. For instance, a user will be able to follow which tautomer decomposed or converged to another tautomer during the geometry optimization procedure.

Our goal is to extend the combinatorial-computational approach to deal with problems more complex than searching for the most stable tautomer. Typically a molecular designer has a specific property in mind, such as emission/absorption wavelength, proton affinity, etc., and searches for stable molecular structures that would display the desirable value of the targeted molecular property. Our main goal is to develop algorithms and software tools that would facilitate combinatorial searches of this type based on the results of quantum chemical calculations. We would also like to improve the management of thousands of files generated on the course of ab initio calculations by using a database system. In future projects we would also like to avoid the time-consuming "manual" analysis of the output structures for the top hits identified in the screening. Instead, we will develop adequate tools to automate this task.

ACKNOWLEDGMENT

Stimulating discussions with Janusz Rak and Tomasz Puzyn are gratefully acknowledged. This work was supported by the (i) Polish State Committee for Scientific Research (KBN) Grants DS/8221-4-0140-6 (M.G.) and N204 127 31/2963 (M.H.) and (ii) U.S. DOE Office of Biological and Environmental Research, Low Dose Radiation Research Program (M.G.). M.H. is thankful for financial support from the European Union Social Funds ZPORR/2.22/II/2.6/ARP/U/2/O5. M.H. is a holder of the award from the Foundation for Polish Science (FNP). Computing resources were available through (i) the Academic Computer Center in Gdańsk (TASK), (ii) a Computational Grand Challenge Application grant from the Molecular Sciences Computing Facility in the Environmental Molecular Sciences Laboratory (EMSL), and (iii) the National Energy Research Scientific Computing Center (NERSC). The EMSL is funded by DOE's Office of Biological and Environmental Research. Pacific Northwest National Laboratory is operated by Battelle for the U.S. DOE under Contract DE-AC06-76RLO 1830.

REFERENCES AND NOTES

(1) Xu, J.; Hagler, A. Chemoinformatics and Drug Discovery. *Molecules* **2002**, *7*, 566−600.

(2) Jorgensen, W. L. The Many Roles of Computation in Drug Discovery. *Science* **2004**, *303*, 1813−1818.

(3) Perry J. K. *Abstr. Pap. Am. Chem. Soc.* **1997**, *213(1*), 176.

(4) Puzyn, T.; Falandysz, J. Computational estimation of logarithm of n-octanol/air partition coefficient and subcooled vapor pressures of 75 chloronaphthalene congeners. *Atmos. Environ.* **2005**, *39*, 1439−1446.

(5) Hay, B. P.; Firman, T. K. HostDesigner: A Program for the de Novo Structure-Based Design of Molecular Receptors with Binding Sites that Complement Metal Ion Guests. *Inorg. Chem.* **2002**, *41*, 5502−5512.

(6) Lowdin, P. O. Proton Tunneling in DNA and its Biological Implications. *Rev. Mod. Phys.* **1963**, *35*, 724−732.

(7) Estrin, D. A.; Paglieri, L.; Corongiu, G. A. Density Functional Study of Tautomerism of Uracil and Cytosine. *J. Phys. Chem.* **1994**, *98*, 5653−5660.

(8) Morpugo, S.; Bossa, M.; Morpugo, G. O. Ab initio study of intramolecular proton transfer reactions in cytosine. *Chem. Phys. Lett.* **1997**, *280*, 233−238.

(9) Fogarasi, G. Relative Stabilities of Three Low-Energy Tautomers of Cytosine: A Coupled-Cluster Electron Correlation Study. *J. Phys. Chem. A* **2002**, *106*, 1381−1390, and references cited therein.

(10) Sayle, R.; Delany, J. Canonicalization and Enumeration of Tautomers, in Innovative Computational Applications. Institute for International Research, Sir Francis Drake Hotel, San Francisco, October 25−27, 1999.

(11) TAUTOMER, developed and distributed by Molecular Networks GmbH, Erlangen, Germany. http://www.mol-net.com.

(12) Pospisil, P.; Ballmer, P.; Scapozza, L.; Folkers, G. Tautomerism in computer-aided drug design. *J. Recept. Signal Transduction Res.* **2003**, *23*, 361−371.

(13) Harańczyk, M.; Rak, J.; Gutowski, M. Stabilization of very rare tautomers of 1-methylcytosine by an excess electron. *J. Phys. Chem. A* **2005**, *109*, 11495−11503.

(14) Bachorz, R. A.; Rak, J;. Gutowski, M. Stabilization of very rare tautomers of uracyl by an excess electron. *Phys. Chem. Chem. Phys.* **2005**, *7*, 2116−2125.

(15) Mazurkiewicz, K.; Bachorz, R. A.; Gutowski, M.; Rak, J. On the unusual stability of valence anions of thymine based on very rare tautomers. A computational study. *J. Phys. Chem. B* **2006**, *110*, 24696−24707.

(16) Hendricks, J. H.; Lyapustina, S. A.; de Clercq, H. L.; Snodgrass, T. J.; Bowen, K. H. Dipole bound, nucleic acid base anions studied via negative ion photoelectron spectroscopy. *J. Chem. Phys.* **1996**, *104*, 7788−7791.

(17) Gutowski, M.; Dąbkowska, I.; Rak, J.; Xu, S.; Nilles, J. M.; Radisic, D.; Bowen, K. H., Jr. Barrier-free intermolecular proton transfer in the uracil-glycine complex induced by excess electron attachment. *Eur. Phys. J. D* **2002**, *20*, 431−439.

(18) Harańczyk, M.; Bachorz, R. A.; Rak, J.; Gutowski, M.; Radisic, D.; Stokes, S. T.; Nilles, J. M.; Bowen, K. H. Excess Electron Attachment Induces Barrier-Free Proton Transfer in Binary Complexes of Uracil with H_2Se and H_2S but Not with H_2O. *J. Phys. Chem. B* **2003**, *107*, 7889−7895.

(19) Li, X.; Cai, Z.; Sevilla, M. D. DFT Calculations of the Electron Affinities of Nucleic Acid Bases: Dealing with Negative Electron Affinities. *J. Phys. Chem. A* **2002**, *106*, 1596−1603.

(20) Harańczyk, M.; Gutowski, M. Finding Adiabatically Bound Anions of Guanine through Combinatorial-Computational Approach. *Angew. Chem. Int. Ed*. **2005**, *44*, 6585−6588.

(21) TauTGen: Tautomer Generator Program. Available at http://tautgen-.sf.net.

(22) *Gaussian 03, Revision C.02*; Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A, Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. Gaussian, Inc.: Wallingford, CT, 2004.

(23) (a) Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* ***1988***, *38*, 3098−3100. (b) Becke, A. D. Density-functional thermochemistry. III. The role of exact Exchange. *J. Chem. Phys.* **1993**, *98*, 5648−5652. (c**)** Lee, C.; Yang, W.; Paar, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37*, 785−789.

(24) Tomasi, J.; Persico, M. Molecular Interactions in Solution: An Overview of Methods Based on Continuous Distributions of the Solvent. *Chem. Rev.* **1994**, *94*, 2027−2094.

(25) (a) Straatsma, T. P.; Aprà, E.; Windus, T. L.; Bylaska, E. J.; de Jong, W.; Hirata, S.; Valiev, M.; Hackler, M.; Pollack, L.; Harrison, R.; Dupuis, M.; Smith, D. M. A; Nieplocha, J.; Tipparaju V.; Krishnan, M.; Auer, A. A.; Brown, E.; Cisneros, G.; Fann, G.; Früchtl, H.; Garza, J.; Hirao, K.; Kendall, R.; Nichols, J.; Tsemekhman, K.; Wolinski, K.; Anchell, J.; Bernholdt, D.; Borowski, P.; Clark, T.; Clerc, D.; Dachsel, H.; Deegan, M.; Dyall, K.; Elwood, D.; Glendening, E.; Gutowski, M.; Hess, A.; Jaffe, J.; Johnson, B.; Ju, J.; Kobayashi, R.; Kutteh, R.; Lin, Z.; Littlefield, R.; Long, X.; Meng, B.; Nakajima, T.; Niu, S.; Rosing, M.; Sandrone, G.; Stave, M.; Taylor, H.; Thomas, G.; van Lenthe, J.; Wong, A.; Zhang, Z.; *NWChem, A Computational Chemistry Package for Parallel Computers, Version 4.6*; Pacific

Northwest National Laboratory: Richland, WA 99352-0999, U.S.A., 2004. (b) Kendall, R. A.; Aprà, E.; Bernholdt, D. E.; Bylaska, E. J.; Dupuis, M.; Fann, G. I.; Harrison, R. J.; Ju, J.; Nichols, J. A.; Nieplocha, J.; Straatsma, T. P.; Windus, T. L.; Wong, A. T. High Performance Computational Chemistry: an Overview of NWChem a Distributed Parallel Application. *Comput. Phys. Commun.* **2000**, *128*, 260−283.

(26) GOT: Gaussian Output Tools. Available at http://gaussot. sf.net.

(27) Schaftenaar, G.; Noordik, J. H. Molden: a pre- and post-processing program for molecular and electronic structures. *J. Comput.-Aided Mol. Des.* **2000**, **14**, 123−34.

(28) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *J. Chem. Phys.* **1992**, *96*, 6796−6806.

(29) Taylor, P. R. In *Lecture Notes in Quantum Chemistry II*; Roos, B. O., Ed.; Springer-Verlag: Berlin, 1994.

(30) (a) Knowles, P. J.; Hampel, C.; Werner, H.-J. Coupled cluster theory for high spin, open shell reference wave functions. *J. Chem. Phys.* ***1994***, *99*, 5219−5227. (b) Deegan, J. J. O.; Knowles, P. J. Perturbative corrections to account for triple excitations in closed and open shell coupled cluster theories. *Chem. Phys. Lett.* **1994**, *227*, 321−326.

(31) Amos, R. D.; Bernhardsson, A.; Berning, A.; Celani, P.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Hampel, C.; Hetzer, G.; Knowles, P. J.; Korona, T.; Lindh, R.; Lloyd, A. W.; McNicholas, S. J.; Manby, F. R.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pitzer, R.; Rauhut, G.; Schütz, M.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T.; Werner, H.-J. MOLPRO, a package of ab initio programs designed by H.-J. Werner and P. J. Knowles, version 2002.1.

(32) Academic Computer Center in Gdañsk (TASK). http://www.task.gda.pl.

(33) Harańczyk, M.; Gutowski, M. Manuscript in preparation.

(34) Warshel, A. *Computer Modeling of Chemical Reactions in Enzymes and Solutions*; John Wiley and Sons: New York, 1997.

(35) Florian, J.; Warshel, A. Langevin Dipoles Model for ab Initio Calculations of Chemical Processes in Solution: Parametrization and Application to Hydration Free Energies of Neutral and Ionic Solutes and Conformational Analysis in Aqueous Solution. *J. Phys. Chem. B* **1997**, *101*, 5583−5595.

(36) Boresch, S.; Karplus, M. The Meaning of Component Analysis: Decomposition of the Free Energy in Terms of Specific Interactions. *J. Mol. Biol.* **1995**, *254*, 801−807.

(37) Hansson, T.; Oostenbrink, C.; van Gunsteren, W. F. Molecular dynamics simulations. *Curr. Opin. Struct. Biol.* **2002**, *12*, *190*−196.

(38) Chandrasekhar, J.; Jorgensen, W. L. Energy Profile for a Nonconcerted S_N2 Reaction in Solution. *J. Am. Chem. Soc.* **1985**, *107*, 2974−2975.

(39) van Gunsteren, W. F.; Berendsen, H. J. C. Computer Simulation of Molecular Dynamics: Methodology, Applications, and Perspectives in Chemistry. *Angew. Chem. Int. Ed.* **1990**, *29*, 992−1023.

(40) E.g. there are 4 sites connected to N4 of adenine, but 2 sites that describe the amino group (2 hydrogen atoms at N4) overlap with another 2 sites describing rotamers of the N4 imino group.

CI6002703