Structure and Singly Occupied Molecular Orbital Analysis of Anionic Tautomers of Guanine

Maciej Haranczyk,[a,b] John Holliday,[b] Peter Willett,[b] Maciej Gutowski[a,c]

[a]*Department of Chemistry, University of Gdańsk, 80-952 Gdańsk, Poland*
[b]*Department of Information Studies, University of Sheffield, Sheffield S1 4DP,UK*
[c]*Chemistry-School of Engineering and Physical Sciencs, Heriot-Watt University, Edinburgh EH14 4AS, UK*

Abstract: Recently we reported the discovery of adiabatically bound anions of guanine which might be involved in the processes of DNA damage by low-energy electrons and in charge transfer through DNA. These anions correspond to some tautomers that have been ignored thus far. They were identified using a hybrid quantum mechanical-combinatorial approach in which an energy-based screening was performed on the library of 499 tautomers with their relative energies calculated with quantum chemistry methods.

In the current study we analyze the adiabatically bound anions of guanine in two aspects: 1) the geometries and excess electron distributions are analyzed and compared with anions of the most stable neutrals to identify the sources of stability; 2) the chemical space of guanine tautomers is explored to verify if these new tautomers are contained in a particular subspace of the tautomeric space. The first task involves the development of novel approaches – the quantum chemical data like electron density, orbital and information on its bonding/antibonding character are coded into holograms and analyzed using chemoinformatics techniques. The second task is completed using substructure analysis and clustering techniques performed on molecules represented by 2D fingerprints.

The major conclusion is that the high stability of adiabatically bound anions originates from the bonding character of the π orbital occupied by the excess electron. This compensates for the antibonding character that usually causes significant

buckling of the ring. Also the excess electron is more homogenously distributed over both rings than in the case of anions of the most stable neutral species. In terms of 2D substructure, the most stable anionic tautomers generally have additional hydrogen atoms at C8 and/or C2 and they don't have hydrogen atoms attached to C4, C5 and C6. They also form an "island of stability" in the tautomeric space of guanine.

E-mails: maharan@chem.univ.gda.pl; j.d.holliday@sheffield.ac.uk;

 p.willett@sheffield.ac.uk; m.gutowski@hw.ac.uk

Graphical abstract

Structure and Singly Occupied Molecular Orbital Analysis of Anionic Tautomers of

Guanine

Maciej Haranczyk,[a,b] John Holliday,[b] Peter Willett,[b] Maciej Gutowski[a,c]

[a]*Department of Chemistry, University of Gdańsk, 80-952 Gdańsk, Poland*
[b]*Department of Information Studies, University of Sheffield, Sheffield S1 4DP,UK*
[c]*Chemistry-School of Engineering and Physical Sciencs, Heriot-Watt University, Edinburgh EH14 4AS, UK*

Quantum chemistry and chemoinformatics methods are used to characterize adiabatically bound anions of guanine and to determine the structural features responsible for their stability.
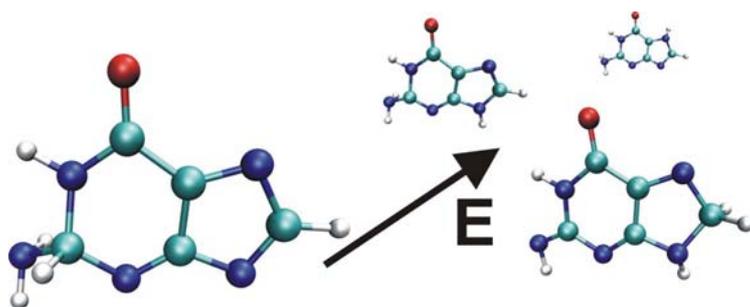
1. **Introduction**

Anions of nucleic acid bases (NAB) can be formed by trapping low-energy electrons produced in living cells by high-energy radiation. Recent experiments suggested that single and double strand breaks develop in DNA exposed to low-energy electrons.[1] Furthermore, charged nucleobases play a critical role in electron and hole transfer in DNA.[2-4] Therefore anionic states of nucleic acid bases have been intensively studied both experimentally and theoretically. [5,6,7,8,9]

The research on the gas-phase NABs conducted in recent years suggested that the most stable and adiabatically bound anions have a dipole-bound rather than valence character.[10] The dipole-bound anions are, however, strongly perturbed by other atoms or molecules and their relevance in condensed phase environments is questionable. All studies have concentrated only on canonical tautomers (ones present in the DNA molecule) and on a few other, previously reported amino-oxo and imino-hydroxy tautomers in which proton is transferred between electronegative atoms, N or O. In contrary we demonstrated that the most stable valence anions of nucleic acid bases, such as 1-methylcytosine[11], uracil,[12] thymine,[13] guanine[14] and adenine[15] correspond to tautomers that result from transformations in which a proton is transferred from an NH site to a carbon site. Moreover, some of these valence anions proved to be adiabatically bound with respect to the most stable tautomers of neutral NABs and to be significantly more stable than dipole-bound anions.

Our initial searches for the most stable anionic tautomers focused on pyrimidine NABs because the number of potentially relevant tautomers was manageable – a few tens of structures.[11,12,13] The number of analogous anionic tautomers for the purine NABs (guanine and adenine), for which we wanted to perform pre-screening using the density functional level of theory (DFT), was as large

as 500-700. This is problematic not so much because of the computer time but rather the human time required to prepare, run, and analyze the calculations, which becomes prohibitive. To overcome these limitations, we developed a hybrid approach involving both combinatorial and accurate quantum chemical methods.[16] The procedure involves: (i) combinatorial generation of a library of tautomers; (ii) pre-screening based on the results of geometry optimization of initial structures performed at the DFT level of theory; and (iii) final refinement of geometry for the top hits at the second order Møller-Plesset level of theory (MP2) followed by single-point energy calculations at the coupled cluster level of theory with single, double, and perturbative triple excitations (CCSD(T)).[17] The library of initial structures of various tautomers is generated with TauTGen,[18] a tautomer generator program developed by us. This program builds all possible tautomers from a molecular framework and a specified number of hydrogen atoms. In contrast to other existing tautomer enumerator programs, it can easily generate as many as 1000 tautomers for molecules of the size of purine bases.

A good measure of the success of our hybrid quantum mechanical-computational approach is our investigation of the valence anions of guanine. This base was believed to have the smallest electron affinity among nucleobases.[19] In the only one available experimental study, Burrow et al. reported a vertical electron affinity of -0.46 eV for guanine and assigned it to a hydroxyl (enol) tautomer.[20] At the same time, our initial study at the CCSD(T) level of theory[21] predicted a negative value of adiabatic electron affinity (AEA, calculated with respect to the neutral canonical tautomer) of the canonical tautomer of -0.459 eV (G, Fig. 1) and value of -0.503 eV for the most stable neutral tautomer (GN, Fig. 1). However, after screening 499 tautomers of guanine, we identified 14 adiabatically bound anions at the

B3LYP/6-31++G** level. These structures, G1-G14, are presented in Fig. 1 and named in decreasing order of stability. Moreover, we demonstrated that the first 13 of these anionic tautomers are adiabatically bound with respect to the neutral canonical tautomer at the CCSD(T) level of theory. The most important properties - the calculated values of AEA and electron vertical detachment energy (VDE) for all 16 important tautomers - are summarized in Table 1. The new species (G1-G14) represent tautomers that are highly unstable as neutral tautomers (Table 1). The most stable anion of guanine (G1, Fig. 1) is adiabatically bound by as much as 0.369 eV (ca. 8.5 kcal/mol), and the most loosely bound adiabatically bound anion (G13, Fig. 1) is characterized by an AEA of only 0.002 eV. The G14 tautomer, with AEA of 0.017 eV kcal/mol at the DFT level employed for screening, has the negative value of - 0.019 eV at the CCSD(T) level of theory. A detailed discussion of the biological relevance of the new tautomers, suggested possible formation pathways as well as the experimental confirmation of our findings (a photoelectron spectrum) have been presented elsewhere. [22]   Simultaneously, the same combinatorial-computational approach was employed in searches of the most stable anionic tautomers of adenine. In this case, we found one anion that is adiabatically bound with AEA of 0.04 eV (0.9 kcal/mol). The finding has been also confirmed in photoelectron spectroscopy experiment performed by Bowen's group.[15]

The application of the hybrid quantum mechanical-computational approach proved to be successful in the identification of the most stable tautomers. At the same time it brought new challenges for computational chemists. For example, how to analyze tens of structures characterized at the high level of theory in the last step of our approach? The natural step is to use chemoinformatics techniques, which have been developed to deal with large amount of chemical data. This, however, brings

another challenge: how to process quantum chemical (QC) data using existing chemoinformatics tools?

In this contribution we present the steps we have taken to meet these challenges. The approaches we have developed combine data from QC calculations (e.g., orbitals, electron density and geometries) with chemoinformatics analysis methods (e.g., similarity calculations and clustering). They are applied to the study the 16 important tautomers of guanine. By comparing the similarity of the excess electron distribution represented by vectors (called *holograms*) derived using Bader's electron density analysis, we demonstrate that the anions of the most stable neutral tautomers are significantly different from the most stable anionic tautomers. The geometries of the latter are nearly planar due to the bonding character of π orbital occupied by the excess electron (Fig. 2). The latter is also more homogenously distributed over the fragments of the most stable anionic tautomers.

In addition, we applied chemoinformatics methods to study the tautomeric space of anionic guanine, rather than limiting ourselves to particular chemical structures. The sets of adiabatically bound and adiabatically unbound tautomers were compared to identify the set of structural features determining the stability. To accomplish this, we coded 2D substructure features present in each of 165 tautomers into Boolean vectors (called *fingerprints*). Then weighted and modal fingerprints were generated to represent groups of adiabatically bound and unbound anions. Substructural analysis was carried out based on occurrence of particular substructure features represented in the groups' fingerprints. The analysis suggested that the characteristic features of the set of adiabatically bound anions are the absence of hydrogen atoms at C4, C5 and C6 carbons.

Additionally, the 165 anionic tautomers were clustered according to their relative similarity. We observed that most of the adiabatically bound anions cluster together, suggesting the existence of an "island of stability" in the chemical space of guanine tautomers. For example, we were able to find a cluster containing 24 elements including 7 out of the top 10 most stable tautomers. These 7 tautomers correspond to the 5 most stable tautomers (equal to 3% of the total tautomers) and to two further adiabatically bound anions from the second part of top 10 list. When compared with canonical guanine, the characteristic substructural features of tautomers in this cluster are additional hydrogen atoms at C8 and/or C2 atoms.

2. Methods

2.1 Geometries of the most stable tautomers

The details of the hybrid quantum mechanical-combinatorial procedure for finding the most stable tautomers of anionic guanine have been described elsewhere. When generating tautomers with the TauTGen program, we imposed the following constraints: 0 or 1 hydrogen atom on N1, C2, N3, C4, C5, C6, N7, N9; 1 or 2 hydrogen atoms on N2 or C8; 0, 1 or 2 hydrogens on O4. With these constraints we generated 499 initial structures including the E, Z stereoisomers (depending on the configuration of asymmetric C2, C4, C5 and C6 carbons) and rotamers of imino and hydroxyl groups. As mentioned before, we found that 2.75% (14) of all screened structures are adiabatically stable at the B3LYP/6-31++G** level, with the range of stability being 0.017-0.529 eV. The notation Gn (n=1...14) is used to name these anionic species according to the descending AEA values reported in the previous study and presented in Table 1. These anionic structures, as well as anions of the canonical tautomer (G, Fig. 1) and of the most stable neutral tautomer (GN, Fig. 1)

were further optimized and harmonic frequencies calculated at the second order Møller-Plesset (MP2) level. The aug-cc-pVDZ basis set (AVDZ)[23] was used. The 1s orbitals of carbon, nitrogen and oxygen atoms were excluded from electron correlation treatments (MP2 and CCSD(T)).

The Cartesian coordinates of these guanine tautomers obtained at the MP2 level are provided as Supporting Information. The calculations were performed using Gaussian03[24] and NWChem[25] software packages on clusters of dual Intel Itanium nodes. The molecular structures and orbitals were plotted using the VMD program.[26]

The important structural feature of all 16 anionic tautomers of guanine is the buckling of the molecule. The geometrical parameters related to buckling are the dihedral angles presented in Table 2. One can compute the similarity between the buckling modes (S_{BM}) of different tautomers using the Euclidean distance:

$$S_{BM} = \left[\sum_{i=1}^{N} (\gamma_{iA} - \gamma_{iB})^2 \right]^{\frac{1}{2}},$$

where γ_{iA} and γ_{iB} are the i-th dihedral angle related to the buckling of tautomer A and B, respectively.

Having defined a similarity measure between buckling modes, all pairwise similarities can be calculated, and clustering then performed to group the most similarly buckled tautomers. The clustering method applied here is a sequential agglomerative hierarchical non-overlapping (SAHN) method.[27] This was implemented using the stored-matrix algorithm, so named because the starting point is a matrix of all pairwise similarities between tautomers in the set to be clustered. Each cluster initially corresponds to an individual structure (singleton). As clustering proceeds, each cluster may contain one or more molecules. Eventually, there evolves one cluster that contains all tautomers. At each iteration, a pair of clusters is merged

(agglomerated) and the number of clusters decreases by 1. The various SAHN methods that are available differ in the way in which the similarity between clusters is defined. Here we calculate it as the arithmetic average of similarities between all pairs of tautomers (which we refer to subsequently as the hierarchical group-average agglomerative clustering, HGAA). The progress of clustering the 16 anionic tautomers is presented as a dendrogram in Fig. 3.

2.2 Analysis of singly occupied molecular orbital

2.2.1 Distribution of singly occupied molecular orbital density

The π orbitals occupied by an excess electron in the 16 important anionic tautomers were presented in Fig. 2. The major differences in the distribution of the excess electron can be identified by comparing the plots. However, to get the quantitative information we developed a novel approach, in which the electron density contribution coming from the singly occupied molecular orbital (SOMO) is assigned to heavy atoms using Bader analysis.[28,29] Bader analysis defines a unique way of dividing molecules into atoms. The definition of an atom is based purely on the electronic charge density. The atoms are divided by so-called zero flux surfaces, which are 2-D surfaces on which the charge density is a minimum perpendicular to the surface. Having defined the surface limiting atom, the charge density is integrated over the volume occupied by particular atoms.

An *orbital hologram* is the vector the components of which hold information about population of the excess electron on each heavy atom. The orbital holograms obtained with this approach are presented in Table 3. We calculate similarity between two orbitals by calculating the Euclidian distance D_{AB} between orbital holograms A and B:

$$D_{AB} = \left[\sum_{i=1}^{N} (x_{iA} - x_{iB})^2 \right]^{\frac{1}{2}},$$

where x_{iA} and x_{iB} are i-th component of orbital holograms A and B, respectively.

Having defined a similarity measure between orbitals, all pairwise similarities can be calculated and then HGAA clustering (as described in the previous section) is performed to group the most similar orbitals.

The progress of clustering the 16 SOMO orbitals of the most important anionic tautomers is presented as a dendrogram in Fig. 4. The information that is available from orbital clustering contributes to our understanding of the binding modes of the excess electron. The similar shape of the electron density distribution corresponding to the SOMO orbital suggests the similar nature of the corresponding electronic state of the molecule.

During the development of the method for comparing molecular orbitals, we tested several alternative approaches. At first, we employed grid-based similarity calculations. The similarity between two orbitals was defined by the Euclidian distance of two grids, i.e., the sum of the absolute values of the differences in values of wavefunctions at corresponding points on both grids. The obvious disadvantage in this approach is the need to find the maximum overlap between the two wavefunctions before the similarity calculation can be performed. We have also tested the orbital hologram approach with a different similarity measure – the Manhattan distance, DM, defined as:

$$DM_{AB} = \sum_{i=1}^{N} \left| x_{iA} - x_{iB} \right|,$$

where x_{iA} and x_{iB} are the i-th component of orbital holograms A and B, respectively. All the orbital comparison methods discussed in this section gave qualitatively the same results.

In addition to the analysis of the excess electron distribution using clustering of orbital holograms we have checked how the excess electron is distributed among the main regions of the guanine molecule. For this purpose we divided the molecule into three regions: I – 6-member ring excluding C5 and C6 atoms, II – C5 and C6 atoms common for both 5- and 6-member rings, and III – 5-member ring excluding C5 and C6. For each of the regions I-III, the corresponding components of the orbital hologram were summed, giving information on the excess electron distribution in those regions. The standard deviation of values obtained for regions I-III indicates whether or not the excess electron is homogenously distributed among the regions. The results of this analysis are presented in Table 4.

2.2.2 Bonding and antibonding character of π orbital

As can be seen in Fig. 2, the π orbitals occupied by the excess electron in the 16 important anionic guanine tautomers have partly a bonding and partly an antibonding character. The major differences in the distribution of bonding and antibonding areas of SOMO orbital can be identified by visual inspection. To verify if the bonding and antibonding character of the π orbital correlates with a particular tautomer's stability, we developed an approach that can quantitatively measure the bonding or antibonding character. This is done by summing the contributions over the chemical bonds present in the molecular framework built from the heavy atoms.

In this approach, the SOMO orbital (Fig. 5a) and the corresponding electron density are studied separately for the parts above (+) and below (-) the molecular

plane (Fig. 5b). For each part above or below the plane, for each atom, the charge is integrated over the space occupied by particular atoms (where the atomic volumes are defined by the Bader analysis described in the previous section). The charge integration is done separately for the electron density coming from the positive and negative parts of the SOMO orbital. Then the two contributions are subtracted to give partial charge δ^{X*} (where X is atom, and * is + or − depending on, respectively, part above or below the molecular plane) (Fig. 5b). The partial charge δ^{X*} is then divided by the number of bonds connecting atom X to the framework atoms.

Having the partial charges, δ^{X*}, for atoms, we can calculate the bonding or antibonding character on each bond. In the case of two bonded atoms, A and B (Fig. 5b), the sign of δ^{A+} and δ^{B+}, and δ^{A-} and δ^{B-} is checked to ensure that they are the same. If so, then there is a bonding interaction between A and B that can be quantified by:

$$AB= |\delta^{A+}+\delta^{B+}| + |\delta^{A-}+\delta^{B-}|.$$

If δ^{A+} and δ^{B+}, and δ^{A-} and δ^{B-} have different signs, then there is an antibonding interaction that is measured by:

$$AB= -1*(|\delta^{A+}|+|\delta^{B+}| + |\delta^{A-}|+|\delta^{B-}|).$$

Having defined a method to measure the bonding/antibonding character of the SOMO orbital for each bond, we can define a vector, the components of which hold this information for all bonds present in the molecule. We will refer to these vectors as *bonding character holograms* (Table 5). Similarly to the orbital holograms defined in the previous section, the similarity between bonding character holograms is calculated using the Euclidean or Manhattan distance. Both similarity measures give qualitatively the same results in this case. The bonding character holograms are

clustered using the HGAA method described previously and the corresponding dendrogram is presented in Fig. 6.

The total bonding and antibonding character of the SOMO orbital can be calculated as a sum of, respectively, bonding and antibonding contributions over all the components of a bonding character hologram. These summed values indicate what part of the excess charge contributes to the bonding and antibonding effect, respectively, and is presented in Table 6.

In the presented approach to the calculation of the bonding and antibonding contribution for each bond, the quantity is calculated even if there is a large disproportion in the excess electron density between the atoms forming the bond. Therefore we have created another implementation of the approach. Here, the bonding and antibonding contribution for each bond are calculated only if the excess electron density on the atom with the smaller integrated density is more than a threshold value, t, for the integrated density on the other atom. The bonding character hologram calculated for a threshold value of 10% is presented in Table 7. The total bonding and antibonding characters of SOMO orbital density obtained with the same threshold, as well as a threshold of 5%, are presented in Table 6. Moreover, the clustering was repeated for the holograms obtained with a threshold value of 10%, and the resulting dendrogram is presented in Fig. 7.

2.3 2D Substructure Analysis

We investigated the tautomer structure-stability relationship (SSR) by various chemoinformatics techniques including substructural analysis and clustering. The SSR analysis was carried out on a reduced set of 165 tautomers (because of software limitations, multiple stereoisomers and rotamers were removed from a set of 499

tautomers as they would become redundant in the fingerprint representation presented in this section, which does not take into account the spatial orientation of the bonded atoms). These tautomers were regenerated with the TauTGen program. We can precisely name the structures using a notation G(a,b…e), where a, b…e are the atoms to which the hydrogen atoms are attached: an example of this notation is presented in Fig. 1. It might be noted that some structures from the original set, like G6 and G11, are represented by only one structure, G(2N2,O4,2C8), in the reduced set.

All tautomers were represented by fingerprints - Boolean arrays indicating the presence or absence of 2D structural fragments specified in a dictionary of fragments. The latter was generated from the 2D connection tables of all tautomers in the library using a 2D descriptor generator program. The updated version (0.04) of TauTGen program was developed to export connection tables, however all bonds are currently assumed to be single. The generation of substructure dictionary as well as fingerprints themselves were performed using the BCI fingerprint package available from Digital Chemistry.[30] This software was not specifically written to work with structures differing only in the position of hydrogen atoms so it does not consider hydrogen positions in substructural fragments. Therefore, to process the library of tautomers we had to use a technical "trick" and substitute all hydrogen atoms with fluorine atoms (which did not exist in the initial dataset).

The substructure fragments used in BCI fingerprints fall into the following main fragment families:

- Augmented Atom: a central atom and its immediate neighbours, with the connecting bonds
- Atom Sequence: a linear sequence of connected atoms and bonds traced through the molecule

15

- Atom Pair: two atoms (including details of atom type, number of non-hydrogen connections), and the topological distance between them by the shortest path[31]

- Ring Composition: a sequence of atoms and bonds around a single ring from the Extended Set of Smallest Rings (ESSR)[32]

- Ring Fusion: a sequence of ring-connectivities around a single ring from the ESSR

The substructure dictionary derived from our set of 165 tautomers contained 1143 fragments (meaning that 1143-bit fingerprints were generated). In the latter step, to further extend the fingerprints, we introduced three types of nitrogen atoms (type 1 (amino/imino): N2; type 2 (in 6-member ring): N1 and N3; type 3 (in 5-member ring): N7 and N9). This allowed us to generate 1492 bit fingerprints, which are called extended-fingerprints throughout this paper.

Having the tautomers represented by fingerprints, one can analyse them by comparing particular bits in the fingerprints, comparing the weighted value of a bit occurring in a subset of structures or comparing bits in modal fingerprints generated for a subset of structures: a weighted fingerprint is generated by summation of the corresponding bits in a set of structures and dividing by the number of structures in the set; and a modal fingerprint[33] can be derived from a weighted fingerprint by setting a bit to 1 if the average value of a bit is higher than a threshold, or 0 otherwise.

A similarity coefficient can be defined between structures coded by the fingerprints. There are many different similarity coefficients, with the Tanimoto coefficient, S_T, being probably the most popular. S_T is defined:

$$S_T = \frac{a}{(a+b+c)}$$

where *a* is the number of bits common to two fingerprints, and *b* and *c* are the number of bits set only in the first or the second fingerprint, respectively. Besides the Tanimoto, we tested 12 other similarity coefficients as defined in Ref. 34, namely: Russell/Rao, Simple Matching, Baroni-Urbani/Buser, Ochiai/cosine, Kulczynski(2), Forbes, Fossum, Simpson, Pearson, Yule, Stiles and Dennis.

Having defined a similarity measure, molecules can be clustered by their relative similarity with various methods (see Refs. 27 and 35 for reviews). The main clustering method used to cluster the tautomers was the HGAA method described in the previous section; the other clustering methods that we tried are k-means and Jarvis-Patrick, which are described in Ref. 27 and which are implemented in the BCI fingerprint software package.

3. Results

3.1 Geometry and buckling modes

In the case of the 16 important tautomers of anionic guanine, the excess electron occupies the π molecular orbital and might cause buckling of the molecular framework. The latter helps to compensate the antibonding effect of the SOMO orbital.[36] The dihedral angles determining non-planarity are presented in Table 2. The buckling is strongest in the case of the anion of the canonical tautomer, which is mainly buckled in the 6-member ring region. The dihedral angle C6N1C2N3 of 28 degrees clearly indicates strong buckling. In the case of the GN tautomer, in turn, buckling is significant in the 5-member ring region confirmed by the C6C5N7C8 angle of 13 degrees. As will be shown in the following section, the buckling modes of G and GN correlate strongly with the excess electron localization on the 6- and the 5-membered ring, respectively.

Interersting, all the G1-14 tautomers are much less buckled than G and GN, with deviations of dihedral angles from either 0º or 180º rarely exceeding 5º. The G1 and G12 seem to be exceptions: however, the large deviations of dihedral angles here originate from the change in hybridization of the C2 and C4 atom, respectively, from sp^2 to sp^3. Therefore these cases cannot be compared with buckling coming from the antibonding effect of the occupied π orbital. Reasonably small non-planarity of all adiabatically bound anions implies that the antibonding character of SOMO orbital is much less evident than in the case of G and GN, or, alternatively, that the bonding character of the SOMO orbital is dominant. These hypotheses will be reviewed in the following sections as they would also justify the high stability of G1-G14 species.

The dendrogram in Fig. 3 presents the clustering of tautomers based on the dihedral angles defining buckling. It can be seen clearly that G, GN, G1 and G12 are most diversely buckled as they are clustered in the last four steps of the clustering process. During this process, at the stage of 7 clusters (marked with a dotted line in Fig. 3) we can see that besides the four diverse singleton tautomers, there are three more clusters containing similarly buckled tautomers: these clusters are: (G2,G3,G8), (G4,G9,G14) and (G5,G6,G7,G10,G11,G13).

The tautomers included in these clusters have similar 2D substructural features suggesting that the latter have a major influence on the buckling mode. The tautomers in the (G2,G3,G8) cluster have two hydrogen atoms at C8 and a hydrogen at N9. The G4, G9 and G14 in the next cluster have two hydrogen atoms at C8 atom and a hydrogen at N7. Finally the tautomers in the (G5,G6,G7,G10,G11,G13) cluster have two hydrogen atoms at C8 and have no hydrogen atoms on either N7 or N9. Interestingly, each of these three clusters contains tautomers with similar values of VDE: ca. 1.6, 2.3 and 1.3 eV, respectively. The big difference between the mean VDE

values for (G4,G9,G14) and the two remaining clusters correlates with the high instability of neutral G4, G9 and G14 reported in Table 1.

3.2 SOMO orbital density analysis for 16 important tautomers

Fig. 2 shows that the excess electron does not seem to be distributed homogeneously in the various tautomers. The plots prepared for G and GN also suggest that these two tautomers are significantly different from the adiabatically bound anions of guanine. To verify these observations we performed clustering of the excess electron density represented by the orbital holograms described in section 2.2, and the resulting dendrogram is shown in Fig. 4.

The clustering confirms the very high similarity between the orbitals of rotamers as they are clustered in the first three steps (G2 and G8, G5 and G7, G6 and G11). Looking down from the top of the dendrogram, it is clear that the orbitals of G and GN are dissimilar from the adiabatically bound anions as their branches join other branches during the last four clustering steps. Looking at the middle of the dendrogram (as denoted by the dotted line in Fig. 4), we can identify seven clusters containing tautomers with a similar distribution of the excess electron: G, GN and G14 are singletons, and the remaining clusters are (G6,G11,G13), (G4,G5,G7,G10), (G1,G12), (G2,G3,G8,G9). The tautomers of the (G6,G11,G13) and (G1,G12) clusters have similar values of VDE. In the case of the remaining (G4,G5,G7,G10) and (G2,G3,G8,G9) clusters, the values of VDE vary from 1.1 to 2.4 eV, although significant similarity in the excess electron density is reported. The discrepancies in VDE have their origin in the difference of stability of the neutral counterparts (Table 1). Another interesting pattern can be observed in the case of the G1 and G14 pair: here, the difference in stability of both anions and neutrals is almost 0.4 eV while the

19

values of VDE differ by only 0.1 eV. Therefore we expect that these anions bind the excess electron in a "different way" (stronger by 0.3 eV), which is reflected in the dissimilarity of the SOMO density distribution.

Besides looking at the excess electron distribution among atoms, we calculated the excess electron distribution among three fragments of the guanine molecule (Table 4). The G and GN tautomers distinguish themselves from amongst the 16 tautomers as they have the most heterogeneous distribution of the excess electron, with the majority localized in fragment I and III, respectively. The G1 and G12 pair also exhibits a significant disproportion in distribution of the excess electron mainly localized in the N7C8N9 region. The G6 and G11 is another pair with an heterogeneous distribution of the excess electron, which is mainly localized in the region I. Both pairs are elements of the two clusters described above. In the case of the remaining tautomers, the excess electron is homogeneously distributed among the three regions of the molecule. These homogenous distributions of the excess electrons are related to small molecular buckling, as already reported in the previous section. They also correlate with the bonding and antibonding character of the SOMO orbital, as discussed in the following section.

3.3 Bonding and Antibonding character of SOMO orbital

The bonding character holograms of the SOMO orbitals of the 16 important guanine tautomers, and their summed total bonding and antibonding characters are presented in Tables 5 and 6, respectively.

In general, the bonding character holograms of SOMO orbitals seem to be a good numerical representation of the plots of SOMO orbitals presented in Fig. 2. For example, in the case of the G tautomer, the bonding character can be observed

between C2N2, N7C8, and the antibonding character in the N1C2N3 fragment is clearly reflected in the G bonding character hologram. Moreover, the latter suggests the bonding character of SOMO in the C4C5C6 area, which is not clearly visible with the selected contour spacing used to plot the SOMO in Fig. 2. All tautomers, with the exception of GN and G12, have a bonding character in the region of C4C5C6. Some bonds, like C5N7 and C8N9, always have an antibonding character in all 16 tautomers.

The SOMO orbital character in different tautomers is revealed during clustering of the bonding character holograms, as shown in Fig. 6. The G, GN, G1 and G12 are different from the remaining tautomers as they are clustered in the last three steps of the agglomeration process. At the seven-clusters level, other important clusters are (G6,G11,G13), (G4,G5,G7,G10), (G2,G3,G8) and (G9,G14). The formation of clusters similar to the first three was also observed when clustering the orbital holograms, verifying the high degree of correlation between the two approaches.

As mentioned in the Methods section, we introduced a threshold parameter that could affect the estimation of the bonding and antibonding characters. We repeated the clustering for the bonding character holograms generated with t=10%, and the dendrogram is presented in Fig. 7. The results we obtained suggested that five tautomers (G, GN, G1, G12 and G14) are diverse and different from the others. The remaining tautomers fitted into three clusters with the already recognizable pattern of elemental distribution: (G6,G9, G11,G13), (G4,G5,G7,G10), (G2,G3,G8). However, G9 is clustered together with G6, G11 and G13, whereas it was clustered with G2, G3 and G8 when clustering using orbital holograms. This fact suggests that although G9

has a similar electron distribution to the first group, the bonding and antibonding pattern of the SOMO orbital fits to the second group.

The total bonding and antibonding character of the SOMO orbital calculated for all 16 tautomers is presented in Table 6. The bonding character of SOMO in adiabatically bound anions, G1-G14, is in general larger than in G and GN. This finding is strengthened by the fact that the observation holds even if the value of the threshold is changed to 5% or 10%. The total antibonding character of SOMO for guanine tautomers does not seem to follow any pattern. The G1 and G12 tautomers are distinct in the sense that they have the largest bonding character and, at the same time, the lowest antibonding character. The G1 and G12 tautomers are also distinct from the others because the bonding character of SOMO in the area of C5C6N9 and N7C8 can be observed together with no significant antibonding character in the remaining parts of the molecule (Fig. 2).

The analysis of the total bonding character provides evidence of correlation between the stability of the adiabatically bound anions and the bonding character of the SOMO orbital. It also supports the conclusion that a high planarity of adiabatically bound anions (no significant buckling) originates from the large bonding character of the π orbital occupied by the excess electron. Surprisingly, the antibonding character of SOMO orbital does not seem to correlate with either the stability or the planarity of the 16 guanine tautomers.

3.4 Substructural analysis of the library of tautomers

The structure-stability relationships analysis was started by checking whether any tautomer in the library of 165 tautomers had a unique structural feature (i.e., not shared with any other tautomer). Seven tautomers of this kind were identified (each

22

with one unique bit set) but none of them was an adiabatically bound anion, or was within 20% of the most stable species.

Because we were not able to find any particular structural feature present in all adiabatically bound anions, we looked at the set of tautomers as belonging to two groups: the adiabatically bound tautomers (10 tautomers, which we will call *active group*) and the remaining set of 155 tautomers (*inactive group*). For each group we calculated modal and weighted fingerprints derived from the extended fingerprints. Depending on the threshold used when calculating the modal fingerprints, we are able to identify 184 (threshold 50%), 117 (60%), 100 (70%), 71 (80%) and 66 (90%) substructural features unique to the active group. The most natural approach would be to use a 50% threshold, which means that a particular 2D substructure feature is considered to be existing (or not existing) when it is found in more than 50% of the tautomers in the library. However, even with the tightest threshold of 90%, 66 substructural features seem to be too many to be analyzed one by one.

When comparing the weighted fingerprints of the active and inactive groups, we looked for substructural features (bits) for which the value differed by 0.59 (as it was the first threshold to identify any feature), i.e., a particular feature is considered unique to a group when it is present 59% more often in one group than the other. For example, if one group has no molecules with the feature, the other group has to have this feature in at least 59% of molecules. We were able to identify 1 non-redundant feature absent in the group of adiabatically bound anions. It is:

1. Absent hydrogen atom separated by 3 bonds from carbon connected with 4 atoms (none of the adiabatically bound anions has this feature)

When we lowered the threshold to 0.58, we could identify six additional non-redundant features absent in the active group (only 1 in 10 adiabatically bound anions has this substructure feature). They are:

2. Absent carbon atom connected to four other atoms (2 carbons, 1 hydrogen and 1 nitrogen)

3. Absent sequence H-C-C, where both carbons are part of a ring

4. Absent sequence H-C-C-N, where C-C-N is part of the ring and N is in a 5-member ring

5. Absent sequence H-C-C-N-C, where C-C-N-C is part of the ring and N is in a 5-member ring

6. Absent sequence H-C-C-N-C-N, where C-C-N-C-N are part of the ring and both Ns are part of a 5-member ring

7. Absent sequence H-C-N-C-C-H, where C-N-C-C are part of a ring and N is in a 5-member ring.

The above substructure features coded in the fingerprints can be translated to more "chemist-friendly" form. For example, the substructural features unique to adiabatically bound anions of guanine are:

1. Absent hydrogen at C2, when there is a hydrogen at C4 or C6 (pt. 1)

2. Absent hydrogen at C4, when there is a hydrogen at C2 or C6 or N7 (pt. 1)

3. Absent hydrogen at C5, when there is a hydrogen at N3 or N1 or C8 or N9 (pt. 1)

4. Absent hydrogen at C6, when there is a hydrogen at C2 or C8 or N7 (pt. 1)

5. Absent hydrogen at C8, when there is a hydrogen at C5 or C6 (pt. 1)

6. Absent hydrogen at C5 (pt. 2)

7. Absent hydrogen at C4, C5 or C6 (pts. 3 and 4 and 5)

8. Absent hydrogen at C5 and C6 (pt. 6)

9. Absent hydrogen at C5 or C6, when there is a hydrogen at C8 (pt. 7)

Caution has to be kept when deriving general organic chemistry rules for predicting the stability of tautomers based on the appearance of substructural features. For example, if the features above were applied as strict rules, only 9 out of 10 adiabatically bound anions would be predicted - tautomer G(N2,C8,C4,N3,O4) would not fulfil rule 7 (pts. 3-5) - and 55 out of 155 adiabatically unbound anions would be considered bound. Thus, rather than the rules being absolute criteria, they should be regarded as features, that are much more likely to appear in the set of adiabatically bound anions than in the set of remaining tautomers.

When the threshold for comparing weighted fingerprints was lowered further to 0.5, we could identify ten additional substructures unique to each group, but such a long list of features makes them difficult to analyse by eye. It might, however, be possible to apply machine learning methods and use these features to teach the computer to recognize the most stable tautomers: this possibility will be explored in future studies.

3.5 Cluster analysis

In the other approach to structure-stability relationships, analysis we clustered the set of 165 tautomers with the aim of identifying a cluster (or clusters) with a high concentration of adiabatically bound anions. Such cluster(s) would correspond to an island of stability in the chemical space of guanine tautomers. The tautomers were represented with extended fingerprints and the similarities calculated using the Tanimoto coefficient. The snapshots of HGAA clustering at 165, 80, 60, 50, 30 and 15 clusters are presented in Table 8. At each level of clustering we note the number of

clusters containing adiabatically bound anions (active clusters) and the composition of these clusters (Table 8). At the beginning of the clustering, there are 10 active clusters as all tautomers are singletons. As clustering proceeds, larger active clusters are formed, with higher concentrations of adiabatically bound anions. For example, at the level of 60 clusters, two active clusters already contain more than one adiabatically bound tautomer. As clustering proceeds further, a dominant active cluster is formed. For example, at the level of 30 clusters, this dominant cluster contains 6 adiabatically bound tautomers and only 4 less stable ones; and at the final level considered (15 clusters), this dominant cluster has 24 elements and contains 7 out of 10 adiabatically bound anions, including the 5 most stable tautomers (3% of the most stable anions). This cluster contains the following tautomers (adiabatically bound anions marked in bold):

G(N2,N1,C2,N3,C8), G(N2,N1,C2,N7,C8), G(N9,N2,N1,C2,C8), G(2N2,N1,C2,C8), G(N2,N1,C2,2C8), **G(N2,N1,N3,2C8)**, G(N2,N1,N7,2C8), G(N9,N2,N1,2C8), G(2N2,N1,2C8), G(N2,C2,N3,N7,C8), G(N9,N2,C2,N3,C8), **G(2N2,C2,N3,C8)**, G(N2,C2,N3,2C8), G(N9,N2,C2,N7,C8), G(2N2,C2,N7,C8), G(N2,C2,N7,2C8), G(N9,2N2,C2,C8), G(N9,N2,C2,2C8), **G(N2,N3,N7,2C8)**, **G(N9,N2,N3,2C8)**, **G(2N2,N3,2C8)**, G(N9,N2,N7,2C8), **G(2N2,N7,2C8)**, **G(N9,2N2,2C8)**.

The common features of the tautomers in this cluster are:

1. At least two hydrogen atoms distributed among C2 and C8 (0 or 1 atoms at C2, 1 or 2 atoms at C8)

2. No hydrogen atoms at C4, C5 or C6

3. No hydroxyl group

The remaining 3 out of 10 adiabatically bound anions, namely G(2N2,O4,2C8), G(N2,N3,C4,O4,C8) and G(N2,N1,O4,2C8), were found in three other clusters containing 3, 8 and 18 elements. These species are structurally distinct from the tautomers in the dominant active cluster.

The majority of adiabatically bound anions represented by extended fingerprints can be clustered together as they lie close to each other in the chemical space. In other words, there is a set of structural features that make them similar to each other and at the same time dissimilar to the remaining tautomers in the chemical space. These 2D structural features may determine susceptibility to the excess electron and they will be summarized in the following sections.

Besides the HGAA method, we tried two other clustering methods, k-means and Jarvis-Patrick employing Tanimoto similarity measure. We also tested 13 different similarity coefficients combined with HGAA. These combinations led to clusters with lower concentration of adiabatically bound anions.

4. Summary and Discussion

The results on anionic tautomers of guanine presented in the previous sections were collected at two levels of resolution. At the first level, the analysis of quantum chemical data allowed us to identify features that either group together or distinguish between the 16 important tautomers. At the second level, the chemical space of tautomers was studied to show that the most stable tautomers lay close to each other in a particular area of this space.

The analysis of the quantum chemical data (namely, the excess electron density, the character of the SOMO orbital, and the geometrical parameters related

with buckling) as well as 2D substructure information for the 16 important anionic tautomers shows that:

- The G1-G14 tautomers are distinguished from the anions of the most stable neutrals, G and GN, as they have a more homogeneous distribution of the excess electron among the fragments of the molecule. The geometries of the G1-G14 are nearly planar due to the greater bonding character of the π orbital occupied by the excess electron. The common structural feature of G1-G14 that distinguishes them from G and GN is an additional hydrogen atom at C2, C4 or C8 when compared with the canonical structure. The most stable anions are also characterized by values of VDE in the range of 1.1 - 2.5 eV whereas the VDEs of G and GN are only 0.6 and 0.2 eV, respectively.

- The G1 and G12 tautomers are significally different from the remaining 12 most stable tautomers. When compared to the canonical tautomer, they have an additional hydrogen atom at C2 or C4, respectively. They are very unstable as neutral species. Therefore both G1 and G12 have large VDE values of ca. 2.4 and 2.5 eV, respectively. The buckling patterns of these molecules also distinguishes them from the others. The tautomers are also different from the others in terms of excess electron distribution. The total bonding character of the SOMO orbital is reported to be highest for these two tautomers.

- The biologically relevant tautomers, G2, G3 and G8, with hydrogen atom at N9 atom and two hydrogen atoms at C8, seem to be very similar to each other, in terms of 2D structure, the buckling mode, the excess electron distribution, the bonding character of SOMO and values of VDE of ca. 1.6 eV. They are however different from the following groups.

- The G5, G7 and G10 form another group of similar tautomers with two hydrogens at C8 but no hydrogen at either N7 or N9. They are similar to each other in terms of the buckling mode, the excess electron distribution, the bonding character of SOMO and values of VDE of ca. 1.2 eV.

- The G6, G11 and G13 form another group of similar tautomers with two hydrogens at C8, no hydrogen at either N7 or N9, but with a hydroxyl O4H group. They are similar to each other in terms of the buckling mode, the excess electron distribution, the bonding character of SOMO and values of VDE of ca. 1.4 eV.

- The G4, G9 and G14 tautomers are similar to each other in terms of 2D substructure (hydrogen at N7, two hydrogens at C8), the buckling mode and large values of VDE of ca. 2.4 eV. The latter correlated with significant instability of neutral counterparts. In terms of the SOMO density and SOMO bonding/antibonding character, however, they are more similar to other groups than they are to each other.

The possibility to identify subgroups in a small set of 16 tautomers demonstrates a high correlation between the properties that we have considered. Obviously it is expected that the 2D structure of the molecule correlates with the excess electron density. The excess electron density is defined by the SOMO orbital, which defines the buckling mode. However, what is new and unexpected is that the protonation states of particular sites seem to have a larger effect on the excess electron density and related properties than do the others. These sites are C2, C4 and C8 carbons as well as N7 and N9 nitrogens.

The analysis of the tautomeric space of guanine suggests that the most stable tautomers have unique structural features. Some of these features could be identified

using the substructural analysis approach. In the presented case the features are the absence of hydrogen atoms at C4, C5 and C6. Substructure analysis is, however, a statistical approach and the results can only be viewed as a suggestion of "more likeness" rather than a definite basis for categorisation. For example, the G12 tautomer has a hydrogen atom at C4, and thus does not follow the rule derived above.

The clustering technique does not provide any structural information directly. It does, however, suggest that some unique features of the most stable anions are present as the formation of clusters with a high concentration of these species is observed. This suggests the existence of an "island of stability" in the tautomeric space, which may be used in the future to develop faster methods for the identification of the most stable tautomers. For example, one could reduce the number of calculations required to screen the tautomeric space to find the most stable species. Such a reduction could be achieved in the following steps:

- Generate *t* tautomers and the corresponding fingerprints.

- Perform hierarchical agglomerative clustering, stopping at one cluster.

- Choose a level of *P* clusters.

- Select *p* molecules, one molecule from each of *P* clusters.

- Run quantum chemical calculations for the *p* molecules to obtain their relative energy.

- Perform energy-based screening of the *p* molecules to get the *m* most stable molecules representing *M* clusters.

- Analyze the dendrogram representing the clustering. Identify *S* clusters at the level of *F* clusters (*F<P*) that contain *M* clusters.

- Run quantum chemical calculations for all molecules (*s*) contained in the *S* clusters.

- Perform energy-based screening of the *s* molecules to get the most stable tautomers.

The efficiency of such a procedure can be estimated using the data collected from the clustering of guanine tautomers presented in Table 8. For example, $P=80$ clusters are selected and the energy based screening is performed for 80 representative molecules. In the worst case, we would find only two adiabatically bound anions (as only two clusters have 100% concentration of adiabatically bound anions). We select only one ($m=1$), the most stable molecule in the set of 80, and trace it in the dendrogram up to the level of 15 clusters ($F=15$). In this case only one cluster ($S=1$) is selected. The cluster has 24 elements and we need to characterize 23 of them at the QC level (one is already characterized). This procedure would require us to perform QC calculations for 103 tautomers instead of 165, giving 37.6% of CPU time saving while retrieving all the five most stable tautomers (and seven adiabatically bound anions total)! If the safer option of $m=3$ is selected, we would end up with 126 calculations (80 at first stage and remaining elements of $S=3$ clusters) – 23.6% of CPU time saving and 8 adiabatically bound anions retrieved. Our ongoing work on anions of adenine and cytosine suggests that such optimized search procedures would successfully identify the most stable tautomers of these molecules: the general application of such procedure would, however, need further investigation. Such a procedure might be valuable for an initial rough exploration of tautomeric spaces of large molecules, or molecules for which little is known about the chemistry (either due to the nature of the molecule or the environment in which it is placed).

5. Conclusions

The hybrid quantum mechanical-combinatorial approach was employed to identify adiabatically bound anions of guanine. The set of 499 tautomers (including rotamers and stereoisomers) was pre-screened at the DFT level and 14 adiabatically bound anions were identified. 13 of them were verified at the CCSD(T) level of theory.

In this study, 16 important anionic tautomers of guanine, the 14 mentioned in the paragraph above and 2 anions of the most stable neutral tautomers, were analyzed in terms of the molecular geometry, single-occupied molecular orbital and its bonding/antibonding character, and the related excess electron density. To perform this analysis we developed SOMO orbital holograms and bonding character holograms, vectors containing information about the excess electron distribution related to these properties and derived using Bader's population analysis. By comparing the similarity of the excess electron density represented by orbital holograms, we demonstrated that the anions of the most stable neutral tautomers are significantly different from the most stable anionic tautomers. We observed a more homogeneous distribution of the excess electron among fragments of the molecule in the adiabatically bound anions. The bonding character of the π orbital occupied by the excess electron is greater in the latter, as indicated by the analysis of the total bonding character calculated from the bonding character holograms. As a result of this, the geometries of the most stable anions are nearly planar. The 14 most stable anionic tautomers were compared using the criteria of the buckling mode, the excess electron distribution and the bonding/antibonding character of SOMO. Five groups of similar tautomers could be identified. The correlation is observed between the protonation state of C2, C4, C8, N7 and N9 atoms and the assignment of a tautomer to a particular

group under the considered criteria. For example, the most stable tautomer forms its own group as it is different from others in terms of the considered criteria. All biologically relevant tautomers with hydrogen at N9 are found to be in one group.

Chemoinformatics methods, including substructural analysis and clustering, were used to identify the set of structural features that might determine the stability. The 2D substructure features of a set of 165 tautomers (excluding rotamers and stereoisomers) were coded into Boolean arrays (fingerprints), and then weighted fingerprints generated to represent groups of adiabatically bound and unbound anions. Substructural analysis based on the occurrence of particular substructure features represented in these fingerprints suggested that, in general, there are no hydrogens present at C4, C5 or C6 in the set of adiabatically bound anions.

Additionally, the hierarchical agglomerative clustering of the library of tautomers suggested that most of the adiabatically bound anions are very similar in terms of 2D structure, as represented by the fingerprints. For example, we identified a cluster of 24 tautomers including seven adiabatically bound anions (and all the five most stable tautomers). When compared with the canonical tautomer, the distinct substructural features of these tautomers are additional hydrogen atoms at C8 and/or C2 atoms. Formation of clusters with high concentration of the most stable tautomers, proves the existence of an "island of stability" in the tautomeric space. This information may be used in the future to develop more efficient methods for the identification of the most stable tautomers.

Supporting Information Available: Geometries (Cartesian coordinates) of G, GN and G1-G14 tautomers obtained at the MP2/AVDZ level of theory.

Figure 1. 16 important tautomers of guanine identified in previous studies. Two alternative notations are provided.

G
G(2N2,C8,N3,N9)

GN
G(2N2,C8,N3,N7)

G1
G(2N2,C8,C2,N3)

G2
G(N2,2C8,N3,N9)

G3
G(2N2,2C8,N9)

G4
G(2N2,2C8,N7)

G5
G(N2,2C8,N1,N3)

G6
G(2N2,2C8,O4)

G7
G(N2,2C8,N1,N3)

G8
G(N2,2C8,N3,N9)

G9
G(N2,2C8,N3,N7)

G10
G(2N2,2C8,N3)

G11
G(2N2,2C8,O4)

G12
G(N2,C8,C4,N3,O4)

G13
G(N2,2C8,N1,O4)

G14
G(N2,2C8,N1,N7)
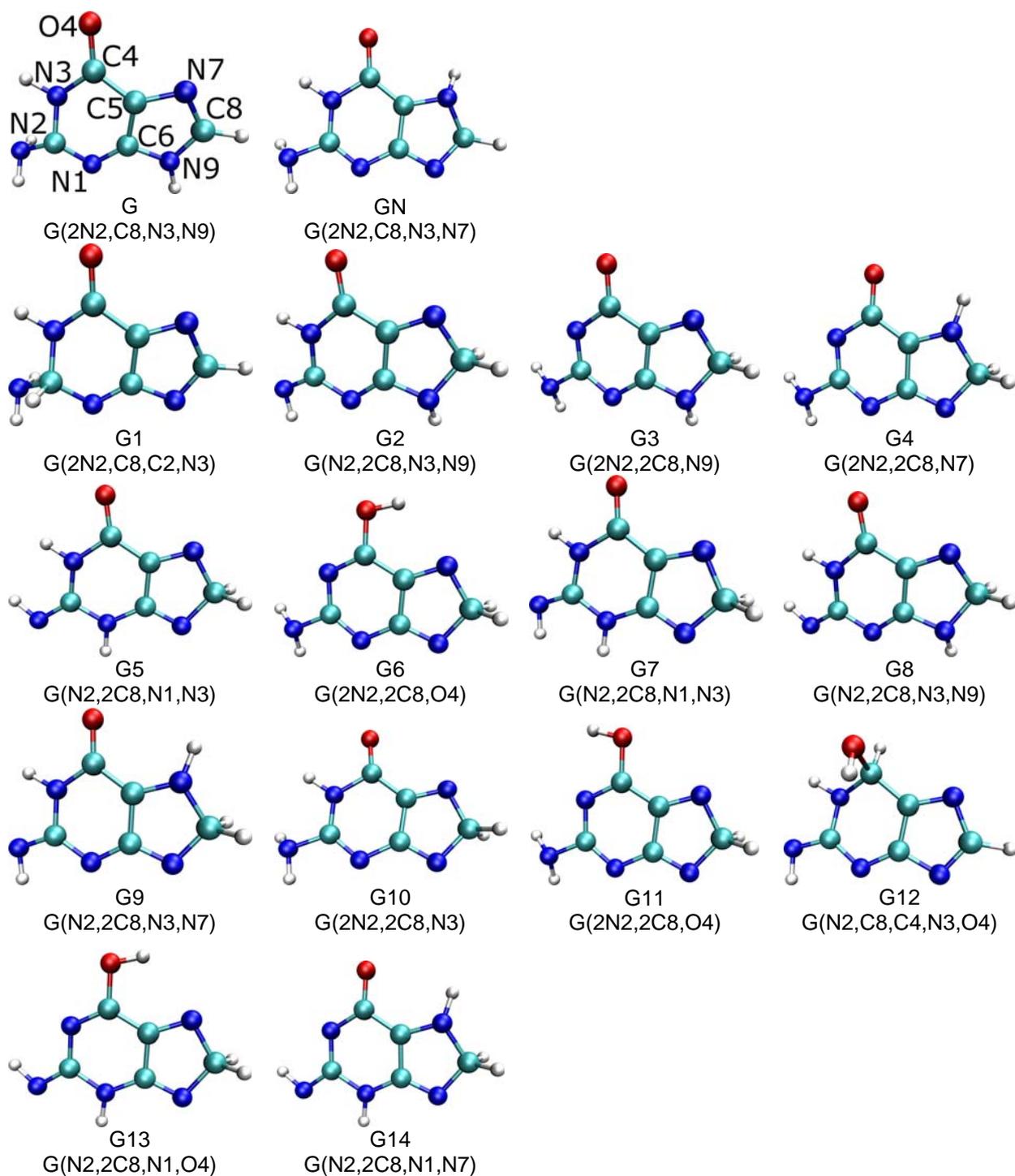
Figure 2. Singly occupied molecular orbitals of 16 tautomers of guanine plotted with a spacing of 0.03 bohr$^{-3/2}$.
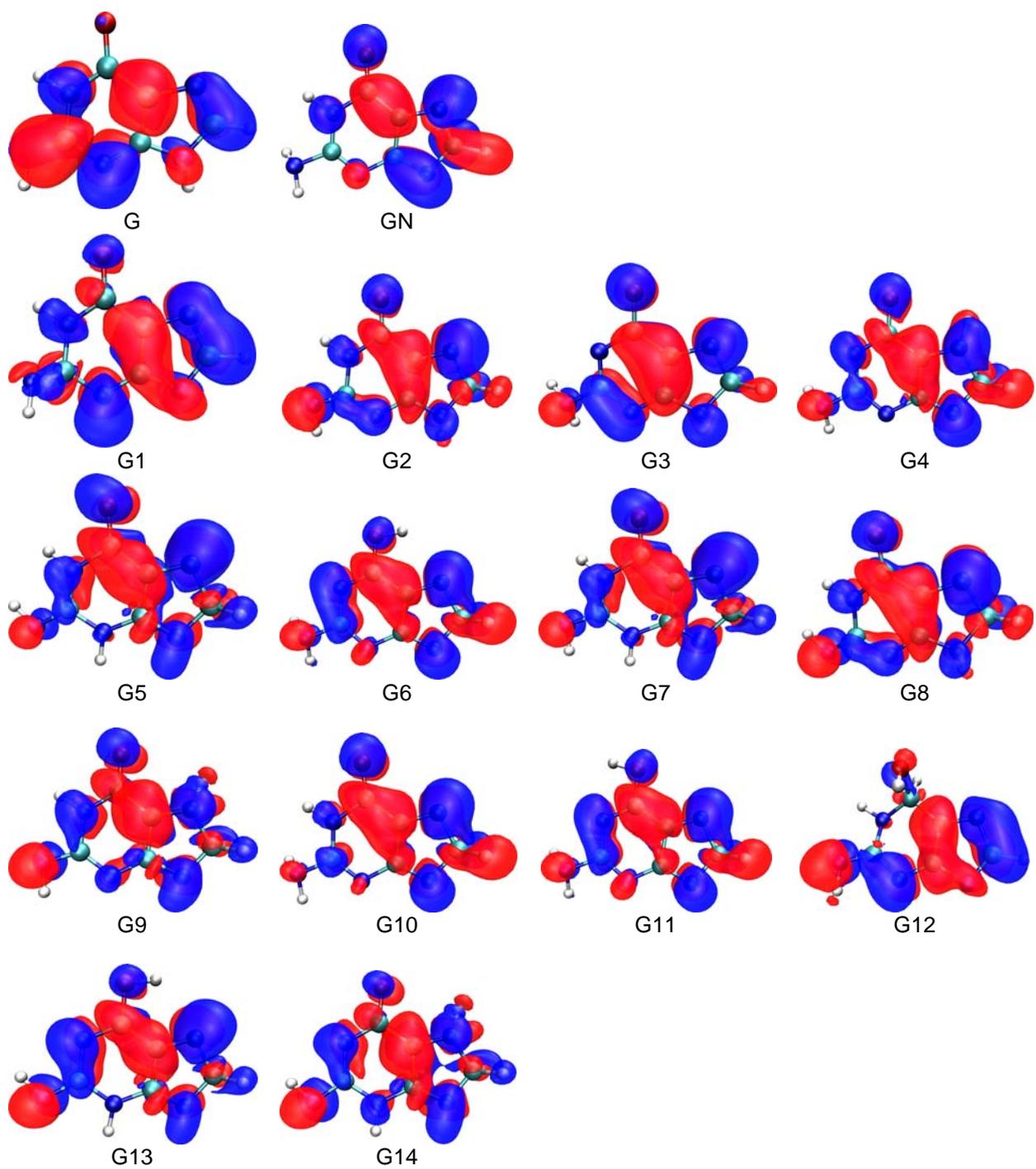
Figure 3. Dendrogram presents HGAA clustering of 16 important anionic tautomers of guanine in terms of buckling mode of the molecule. The dotted horizontal line represents the seven-cluster partition.
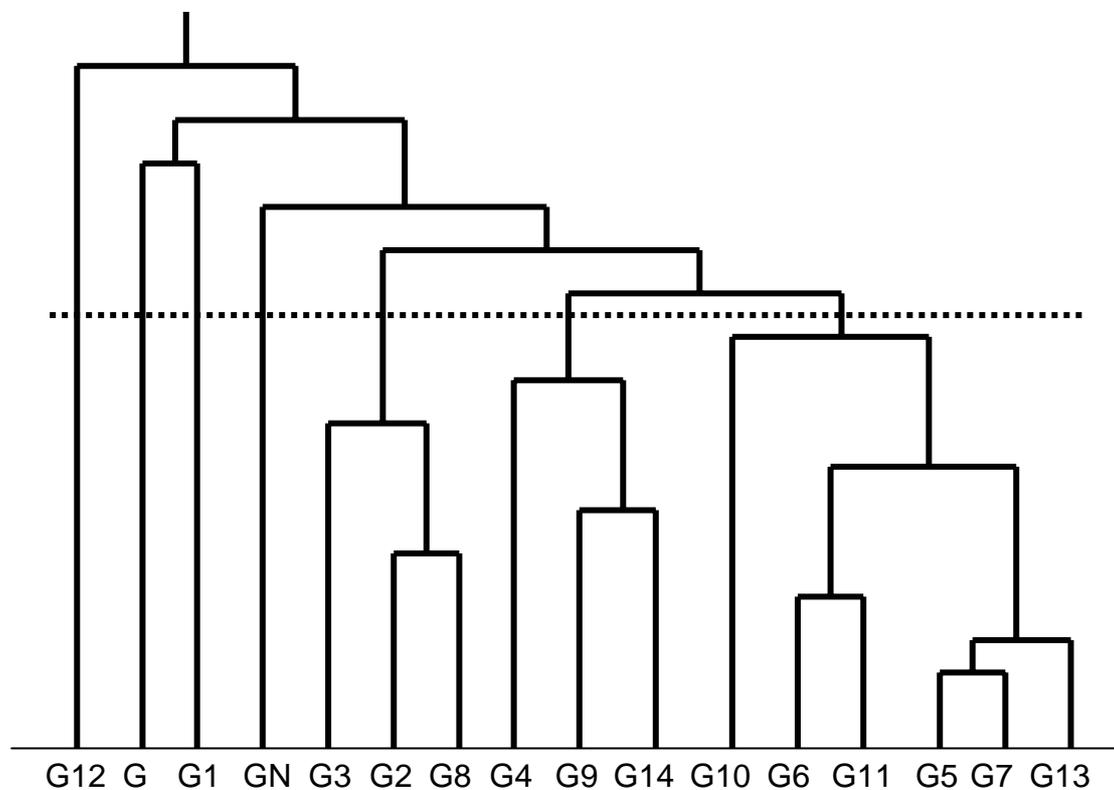
Figure 4. Dendrogram presents HGAA clustering of SOMO orbital holograms of 16

important anionic tautomers of guanine. The dotted horizontal line represents the

seven-cluster partition.
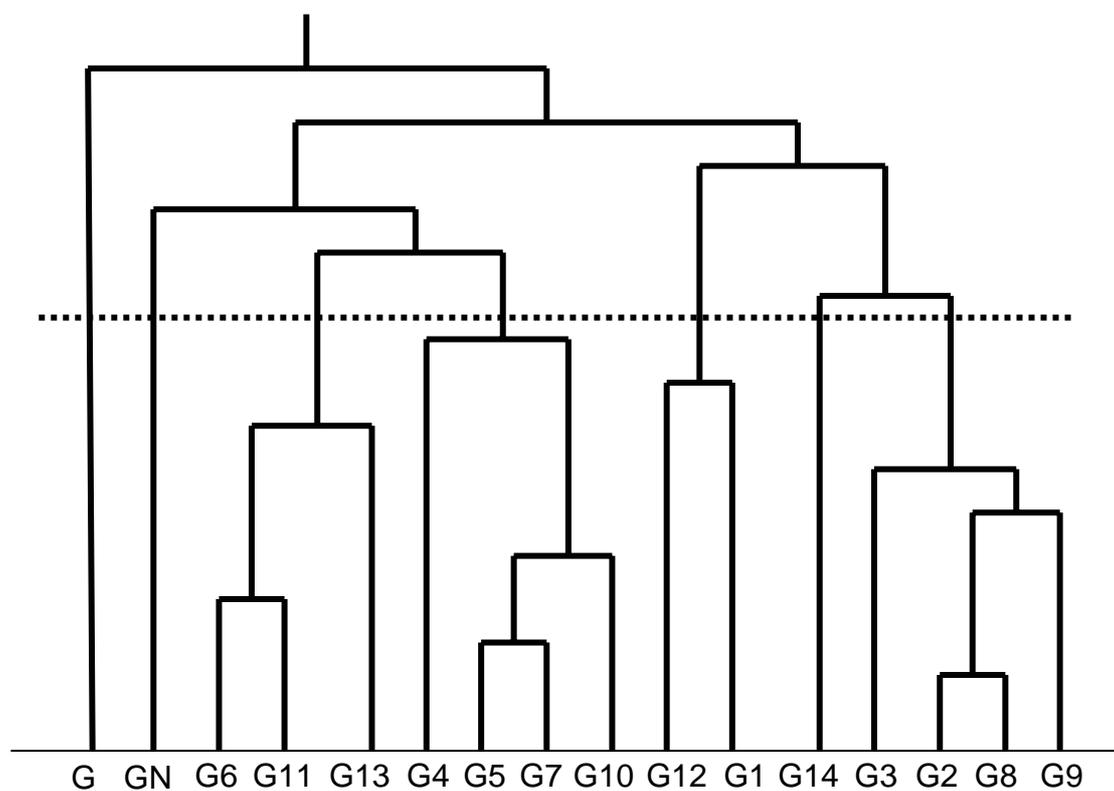
Figure 5. Division of SOMO orbital for the purpose of calculating bonding and antibonding effect on a chemical bond between A and B.
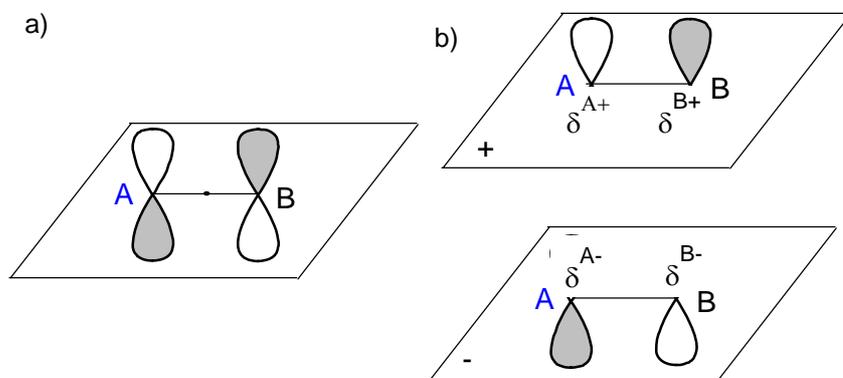
Figure 6. Dendrogram presents HGAA clustering of bonding character holograms of 16 important anionic tautomers of guanine. The dotted horizontal line represents the seven-cluster partition.
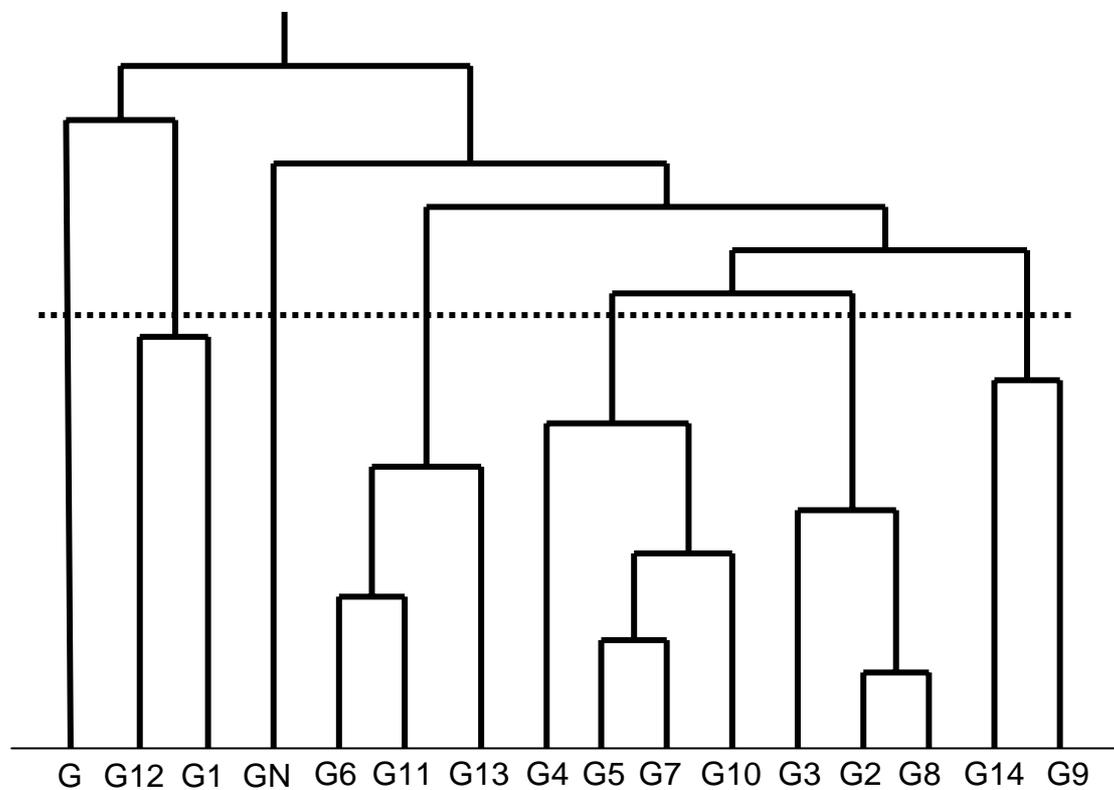
Figure 7. Dendrogram represents HGAA clustering of bonding character holograms of 16 impotrant anionic tautomers of guanine obtained with threshold of 10% . The dotted horizontal line represents the seven-cluster partition.
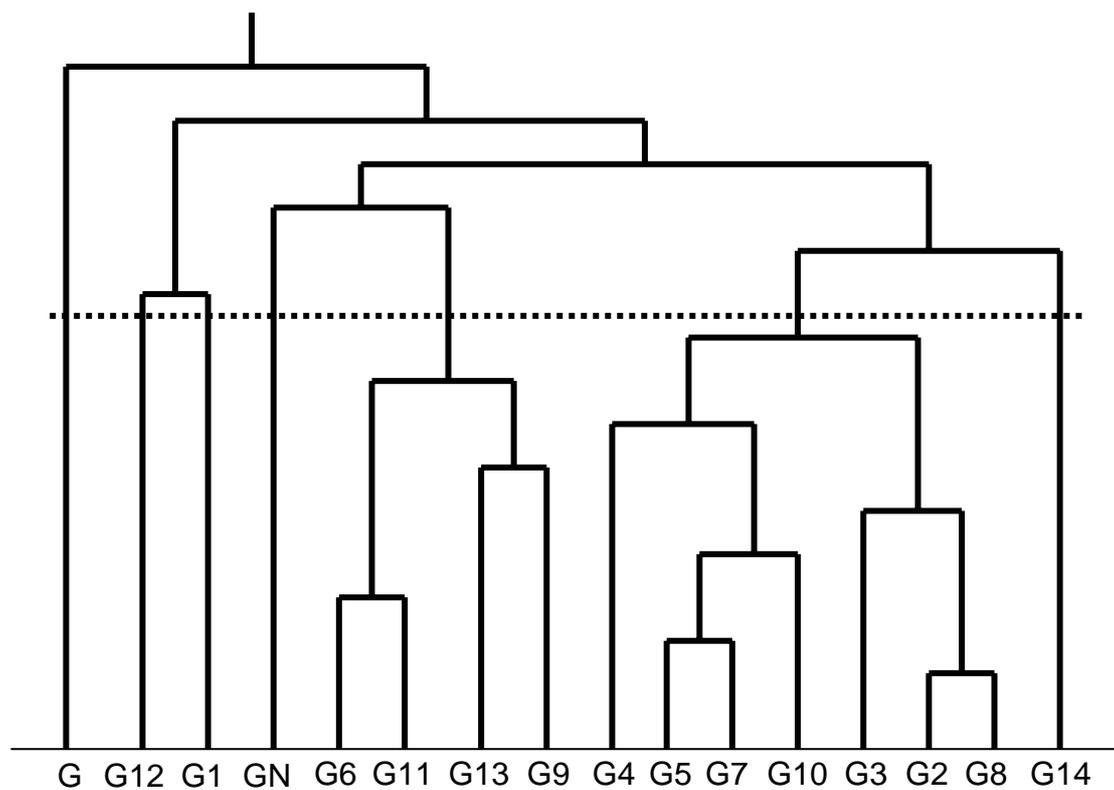
Table 1. The relative energies and relative energies corrected for zero-point vibration energy of 16 neutral tautomers of guanine together with adiabatic electron affinities (AEA, calculated with respect to the canonical tautomer) and electron vertical detachment energies (VDE) for the anionic tautomers of guanine. All energies in eV calculated at the CCSD(T)/AVDZ level with zero-point vibration energy calculated at the MP2/AVDZ level.

|	Neutral		Anions	
	ΔE	$\Delta(E+E_{zpv})$	AEA	VDE
G	0.000	0.000	-0.459	0.585
GN	-0.033	-0.025	-0.503	0.212
G1	1.819	1.771	0.369	2.426
G2	1.015	0.984	0.365	1.604
G3	1.085	1.044	0.304	1.699
G4	1.725	1.676	0.278	2.205
G5	0.965	0.920	0.201	1.316
G6	1.171	1.112	0.174	1.484
G7	0.965	0.920	0.173	1.289
G8	1.242	1.201	0.165	1.617
G9	2.118	2.069	0.116	2.427
G10	0.919	0.865	0.104	1.137
G11	1.192	1.125	0.094	1.414
G12	2.241	2.181	0.078	2.542
G13	1.323	1.277	0.002	1.450
G14	2.161	2.115	-0.019	2.318

Table 2. Dihedral angles related to the buckling of the guanine molecule. The geometries of the tautomers were optimized at the MP2/AVDZ level of theory.

Tautomer	Dihedral angles								
	C4C5C6N1	C5C6N1C2	C6N1C2N3	C6N1C2N2	C6C5C4N3	C6C5C4O4	N1C6C5N7	N7C5C6N9	C6C5N7C8
G	-7.82	-7.48	27.89	163.44	1.44	184.47	180.51	0.97	-0.84
GN	1.02	-0.78	0.96	183.60	-1.20	178.49	175.53	-5.06	13.05
G1	-5.31	-8.94	27.68	148.78	-0.97	182.75	180.43	-0.51	0.04
G2	3.25	-1.34	-2.05	178.52	-1.44	178.98	183.00	4.42	5.23
G3	2.49	-1.12	-0.36	181.88	-2.26	178.88	182.86	4.50	4.73
G4	-2.58	0.64	1.64	183.12	2.29	182.71	177.85	-2.45	5.02
G5	-0.04	0.11	-0.11	179.90	-0.02	179.97	179.97	-0.01	0.01
G6	-0.20	-0.49	1.08	183.19	0.52	180.84	180.39	-0.01	-0.06
G7	-0.02	0.07	-0.07	179.95	-0.02	180.00	179.99	0.00	0.00
G8	3.90	-1.49	-2.69	177.94	-1.67	178.70	183.61	4.75	5.06
G9	-1.72	0.62	0.98	180.46	0.97	181.17	177.20	-2.85	6.08
G10	1.53	-2.03	1.94	184.04	-0.70	178.87	180.78	-0.08	-0.29
G11	0.14	-0.59	0.82	183.06	0.21	180.33	180.38	-0.01	0.05
G12	2.32	5.52	9.37	185.65	-22.04	101.88	179.48	0.24	0.14
G13	0.00	0.01	-0.01	180.01	0.00	180.00	180.00	0.00	0.00
G14	-2.54	1.27	0.33	179.91	2.11	182.34	176.20	-3.80	7.90

Table 3. SOMO orbital holograms obtained for 16 anionic tautomers of guanine.

Tautomer	Excess charge on atoms of molecular framework										
	N1	C2	N3	C4	C5	C6	N7	C8	N9	N2	O4
G	0.298	0.172	0.121	0.012	0.133	0.009	0.084	0.034	0.041	0.088	0.007
GN	0.026	0.000	0.062	0.086	0.152	0.008	0.153	0.198	0.222	0.002	0.091
G1	0.132	0.016	0.036	0.024	0.174	0.067	0.335	0.062	0.123	0.007	0.025
G2	0.116	0.000	0.045	0.053	0.271	0.058	0.213	0.030	0.045	0.100	0.070
G3	0.143	0.001	0.038	0.039	0.243	0.093	0.190	0.032	0.079	0.047	0.096
G4	0.021	0.001	0.087	0.034	0.307	0.030	0.272	0.067	0.097	0.036	0.049
G5	0.009	0.000	0.069	0.110	0.255	0.028	0.207	0.058	0.115	0.027	0.122
G6	0.045	0.001	0.217	0.114	0.165	0.016	0.147	0.061	0.155	0.033	0.046
G7	0.011	0.000	0.066	0.099	0.261	0.033	0.211	0.059	0.120	0.028	0.112
G8	0.109	0.000	0.047	0.060	0.271	0.051	0.215	0.030	0.040	0.098	0.079
G9	0.109	0.000	0.075	0.042	0.264	0.008	0.217	0.031	0.016	0.192	0.047
G10	0.016	0.000	0.074	0.127	0.222	0.025	0.193	0.063	0.129	0.013	0.139
G11	0.053	0.000	0.231	0.134	0.139	0.013	0.130	0.061	0.158	0.028	0.051
G12	0.190	0.000	0.005	0.009	0.119	0.090	0.291	0.063	0.165	0.054	0.013
G13	0.008	0.000	0.200	0.102	0.216	0.017	0.187	0.049	0.098	0.083	0.040
G14	0.031	0.000	0.160	0.013	0.269	0.028	0.219	0.043	0.059	0.167	0.011

Table 4. Excess electron distribution over fragments I-III of guanine molecule.

Tautomer	Excess charge on fragments			
	I*	II**	III***	Std. div.
G	0.698	0.142	0.159	0.258
GN	0.267	0.160	0.573	0.175
G1	0.239	0.241	0.520	0.132
G2	0.383	0.329	0.288	0.039
G3	0.364	0.336	0.300	0.026
G4	0.227	0.337	0.436	0.086
G5	0.337	0.283	0.380	0.040
G6	0.456	0.181	0.363	0.114
G7	0.317	0.294	0.390	0.041
G8	0.393	0.322	0.284	0.045
G9	0.465	0.272	0.264	0.093
G10	0.369	0.247	0.384	0.062
G11	0.498	0.153	0.349	0.141
G12	0.272	0.209	0.519	0.134
G13	0.433	0.233	0.334	0.082
G14	0.382	0.297	0.320	0.036

*6-member ring (atoms: N1,C2,N2,N3,C4 and O4)

** fragment common for both 6- and 5- member ring (atoms: C5 and C6)

*** 5-member ring (atoms: N7,C8 and N9)

Table 5. Bonding character holograms for 16 tautomers of anionic guanine. The positive number means bonding character, otherwise antibonding character for particular bond between atoms of molecular frame.

Tautomer	Bond between atoms of molecular frame											
	N1C2	C2N3	N3C4	C4C5	C5C6	C6N1	C5N7	N7C8	C8N9	N9C6	C2N2	C4O4
G	-0.181	-0.095	-0.044	0.048	0.047	-0.128	-0.064	0.038	-0.028	0.014	0.053	-0.007
GN	-0.007	-0.016	-0.045	0.079	-0.052	-0.007	-0.104	-0.152	-0.183	0.086	0.000	-0.083
G1	-0.053	-0.018	-0.021	0.066	0.080	-0.070	-0.186	0.159	-0.060	0.051	0.006	-0.028
G2	0.049	0.013	-0.031	0.108	0.110	-0.069	-0.150	-0.073	-0.028	-0.034	-0.096	-0.073
G3	0.051	-0.001	0.013	0.094	0.112	-0.082	-0.139	-0.073	-0.038	-0.054	-0.025	-0.096
G4	0.005	0.038	-0.049	0.114	0.112	-0.014	-0.199	-0.129	-0.079	-0.056	-0.026	-0.049
G5	0.000	-0.017	-0.053	0.122	0.094	-0.009	-0.149	-0.092	-0.085	-0.066	-0.025	-0.119
G6	-0.015	0.083	-0.121	0.093	0.060	0.020	-0.104	-0.079	-0.106	-0.082	-0.015	-0.054
G7	0.000	-0.017	-0.050	0.120	0.098	-0.011	-0.152	-0.094	-0.087	-0.070	-0.026	-0.111
G8	0.048	0.013	-0.033	0.110	0.107	-0.065	-0.150	-0.073	-0.026	-0.030	-0.093	-0.080
G9	-0.054	-0.032	-0.046	0.102	0.091	-0.057	-0.161	-0.087	-0.022	-0.010	-0.191	-0.047
G10	-0.002	0.013	-0.055	0.116	0.082	0.011	-0.137	-0.095	-0.094	-0.072	-0.003	-0.130
G11	-0.021	0.086	-0.130	0.091	0.051	0.025	-0.090	-0.073	-0.108	-0.083	-0.012	-0.061
G12	0.064	0.000	-0.002	0.000	0.069	-0.094	-0.163	0.155	-0.081	0.079	-0.045	0.002
G13	0.000	0.080	-0.114	0.106	0.078	-0.006	-0.127	-0.078	-0.072	-0.054	-0.079	-0.048
G14	-0.013	-0.079	-0.084	0.094	0.099	-0.022	-0.165	-0.096	-0.048	-0.037	-0.167	-0.012

Table 6. The total bonding (TOT) and total antibonding (TOT*) character of SOMO

orbital derived from bonding character hologram with threshold of 0, 5 and 10%.

Tautomer	Threshold 0%		Threshold 5%		Threshold 10%	
	TOT	TOT*	TOT	TOT*	TOT	TOT*
G	0.200	-0.548	0.200	-0.420	0.105	-0.420
GN	0.165	-0.649	0.079	-0.574	0.079	-0.574
G1	0.362	-0.436	0.362	-0.436	0.362	-0.436
G2	0.280	-0.553	0.217	-0.457	0.217	-0.457
G3	0.270	-0.508	0.206	-0.482	0.206	-0.482
G4	0.268	-0.603	0.226	-0.576	0.226	-0.576
G5	0.216	-0.615	0.216	-0.565	0.216	-0.565
G6	0.257	-0.577	0.173	-0.547	0.173	-0.465
G7	0.218	-0.618	0.218	-0.564	0.218	-0.564
G8	0.279	-0.550	0.218	-0.457	0.218	-0.457
G9	0.193	-0.708	0.102	-0.374	0.102	-0.374
G10	0.222	-0.588	0.209	-0.583	0.209	-0.583
G11	0.253	-0.578	0.167	-0.546	0.167	-0.463
G12	0.370	-0.384	0.307	-0.340	0.307	-0.340
G13	0.264	-0.578	0.184	-0.499	0.106	-0.493
G14	0.193	-0.722	0.193	-0.463	0.099	-0.380

Table 7. Bonding character holograms for 16 tautomers of anionic guanine derived using threshold of 10%. The positive number means bonding character, otherwise antibonding character for particular bond between atoms of molecular frame.

Tautomer	Bond between atoms of molecular frame											
	N1C2	C2N3	N3C4	C4C5	C5C6	C6N1	C5N7	N7C8	C8N9	N9C6	C2N2	C4O4
G	-0.181	-0.095	-0.044	0.000	0.000	0.000	-0.064	0.038	-0.028	0.014	0.053	-0.007
GN	0.000	0.000	-0.045	0.079	0.000	-0.007	-0.104	-0.152	-0.183	0.000	0.000	-0.083
G1	-0.053	-0.018	-0.021	0.066	0.080	-0.070	-0.186	0.159	-0.060	0.051	0.006	-0.028
G2	0.000	0.000	-0.031	0.108	0.110	-0.069	-0.150	-0.073	-0.028	-0.034	0.000	-0.073
G3	0.000	-0.001	0.000	0.094	0.112	-0.082	-0.139	-0.073	-0.038	-0.054	0.000	-0.096
G4	0.000	0.000	-0.049	0.114	0.112	-0.014	-0.199	-0.129	-0.079	-0.056	0.000	-0.049
G5	0.000	0.000	-0.053	0.122	0.094	0.000	-0.149	-0.092	-0.085	-0.066	0.000	-0.119
G6	0.000	0.000	-0.121	0.093	0.060	0.020	-0.104	-0.079	-0.106	0.000	0.000	-0.054
G7	0.000	0.000	-0.050	0.120	0.098	0.000	-0.152	-0.094	-0.087	-0.070	0.000	-0.111
G8	0.000	0.000	-0.033	0.110	0.107	-0.065	-0.150	-0.073	-0.026	-0.030	0.000	-0.080
G9	0.000	0.000	-0.046	0.102	0.000	0.000	-0.161	-0.087	-0.022	-0.010	0.000	-0.047
G10	0.000	0.000	-0.055	0.116	0.082	0.011	-0.137	-0.095	-0.094	-0.072	0.000	-0.130
G11	0.000	0.000	-0.130	0.091	0.051	0.025	-0.090	-0.073	-0.108	0.000	0.000	-0.061
G12	0.000	0.000	-0.002	0.000	0.069	-0.094	-0.163	0.155	-0.081	0.079	0.000	0.002
G13	0.000	0.000	-0.114	0.106	0.000	0.000	-0.127	-0.078	-0.072	-0.054	0.000	-0.048
G14	0.000	0.000	0.000	0.000	0.099	-0.022	-0.165	-0.096	-0.048	-0.037	0.000	-0.012

Table 8. Progress of HGAA clustering of 165 tautomers of guanine represented with extended fingerprints. The each level of clustering (N_{Cl}) the number of active clusters is noted (N_{ACl}) and composition of these clusters (number of adiabatically bound anions (n_a) and the total number of element in this cluster (n_{tot}). Active clusters with high concentration of adiabatically bound anions are marked with bold font.

N_{Cl}	N_{ACl}	Active cluster composition (n_a/n_{tot})
165	10	1/1 1/1 1/1 1/1 1/1 1/1 1/1 1/1 1/1 1/1
80	8	1/3 1/3 **2/3** 1/3 1/3 1/1 1/2 **2/2**
60	7	1/3 1/3 **3/6** 1/4 1/1 1/2 **2/2**
50	6	1/3 1/2 **3/6** 1/2 1/1 **3/4**
30	5	1/3 1/10 1/8 1/4 **6/10**
15	4	1/3 **7/24** 1/18 1/8

References

1. Boudaïffa, B.; Cloutier, P.; Hunting, D.; Huels, M. A.; Sanche, L. Science 2000, 287, 1658-1660.

2 (a) Cai, Z. ; Sevilla, M.D. J Phys Chem. B 2000, 104, 6942-6949; (b) Messer, A.; Carpenter, K.; Forzley, K.; Buchanan, J.; Yang, S.; Razskazovskii, Y.; Cai, Z.; Sevilla, M.D. J Phys Chem B 2000, 104, 1128-1136; (c) Cai, Z.; Gu, Z.; Sevilla, M.D. J Phys Chem B 2000, 104, 10406-10411.

3. Berlin, Y.A.; Burin, A.L.; Ratner, M.A. J Am Chem Soc 2001, 123, 260-268 and references cited therein.

4. Bixo, M.; Jortner, J. J Phys Chem A 2001, 105, 10322-10328 and references cited therein.

5. Seidel, C.A.M.; Schulz, A.; Sauer, M.H.M. J Phys Chem 1996, 100, 5541-5553.

6. Aflatooni, K.; Gallup, G.A.; Burrow, P.D. J Phys Chem A 1998,102,6205-6207.

7. Periquet, V.; Moreau, A.; Carles, S.; Schermann, J.P.; Desfrancüois, C. J Electron Spectrosc Relat Phenom 2000, 106, 141-151.

8. Li, X.; Cai, Z.; Sevilla, M.D. J Phys Chem A 2002, 106, 1596-1603.

9. Wesolowski, S.S.; Leininger, M.L.; Pentchew, P.N.; Schaefer III, H.F. J Am Chem Soc 2001, 123, 4023-4028.

10. Hendricks, J.H.; Lyapustina, S.A.; de Clercq, H.L.; Snodgrass, T.J.; Bowen, K.H. J Chem Phys 1996, 104, 7788-7791.

11. Haranczyk, M.; Rak, J.; Gutowski, M. J Phys Chem A 2005, 109, 11495-11503.

12. Bachorz, R.A.; Rak, J.; Gutowski, M. Phys Chem Chem Phys 2005, 7, 2116-2125.

13. Mazurkiewicz, K.; Bachorz, R.A.; Rak, J.; Gutowski, M. J Phys Chem B 2006, 110, 24696-24707.

14. Haranczyk, M.; Gutowski, M. Angewandte Chemie Int Ed 2005, 44, 6585-6588.

15 Haranczyk, M.; Gutowski, M.; Li, X.; Bowen K.H. Proc Natl Acad Sci (PNAS) 2007, 104, 4804-4807.

16. Haranczyk, M.; Gutowski M. J Chem Inf and Model 2007, 47, 686-694.

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

18. Freely available at http://tautgen.sourceforge.net

19. Li, X.; Cai, Z.; Sevilla, M.D. J Phys Chem A 2002, 106, 1596-1603.

20. Aflatooni, K.; Gallup, G.A.; Burrow, P.D. J Phys. Chem A 1998, 102, 6205-6207.

21. Haranczyk, M.; Gutowski, M. J Am Chem Soc 2005, 127, 699-706.

22. Haranczyk, M.; Gutowski, M.; Li, X.; Bowen K.H. submitted to Physical Review Letters.

23. Kendall, R.A.; Dunning Jr., T.H.; Harrison, R.J. J Chem Phys 1992, 96, 7696-6806.

24. Gaussian 03, Revision C.02, Frisch, M.J.; Trucks, G.W.; Schlegel, H.B.; Scuseria, G.E.; Robb, M.A.; Cheeseman, J.R.; Montgomery, Jr., J.A.; 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.

25. (a) Straatsma, T.P.; Apra, 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, 2004, Pacific Northwest National Laboratory, Richland, Washington 99352-0999, USA; (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. Computer Phys Comm 2000, 128, 260-283.

26. Humphrey, W.; Dalke, A.; Schulten, K. J Mol Graph 1996, 14, 33-38.

27. Downs, G.M.; Barnard, J.M. Clustering methods and their uses in computational chemistry, Reviews in Computational Chemistry, vol. 18, chapter 1, edited by K.B. Lipkowitz and D.B. Boyd, Wiley-VCH, 2002

28. Henkelman, G.; Arnaldsson, A.; Jónsson, H. Comput Mater Sci 2006, 36, 254-360.

29. Sanville, E.; Kenny, S. D.; Smith, R.; Henkelman G. J Comp Chem 2007, 28, 899-908.

30. http://www.digitalchemistry.co.uk

31. Carhart, R.E.; Smith, D.H.; Venkataraghavan, R. J Chem Inf Comput Sci 1985, 25, 64-73.

32. Downs, G.M.; Gillet, V.J.; Holliday, J.D.; Lynch, M.F. J Chem Inf Comput Sci 1989, 29, 207-214.

33. Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. J Chem Inf Comput Sci 2004, 44, 1177-1185.

34. Holliday, J.D.; Hu, C-Y.; Willett, P. Combi Chem High Throughput Screening 2002, 5, 155-166.

35. Willett, P. Similarity and clustering in chemical information systems; Research Studies Press, 1987.

36. Gutowski, M.; Dabkowska, I.; Rak, J.; Xu, S.; Nilles, J.M.; Radisic, D.; Bowen, K.H. Eur Phys J D 2002, 20, 431-439.