Chemoinformatics Analysis of Results Generated with Quantum Mechanical-Combinatorial Approach

Maciej Haranczyk, John Holliday, Maciej Gutowski

The University Of Sheffield.

HERIOT WATT UNIVERSITY

Introduction

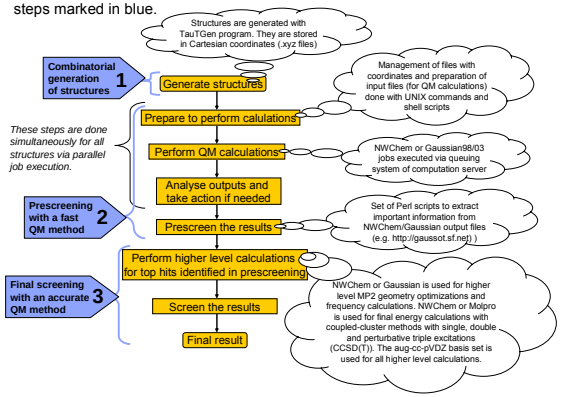
Abstract: Our recent studies on charged nucleic acid bases suggest that the most stable ionic tautomers might be structurally different from the most stable neutral species. In most of the cases these molecular systems cannot be predicted with common chemical knowledge. For example, we were able to identify the most stable tautomers of anionic nucleic acid bases among not-studied so far enamine-imine tautomers.

Therefore, we have described a hybrid quantum mechanical-combinatorial (HQM-C) procedure for identification of 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 identified most stable tautomers where analyzed at two levels. First their geometries and properties where calculated with quantum chemistry methods. Second, the whole set of tautomers was analyzed using chemoinformatics methods. 2D substructure features of a set of nucleic acid bases tautomers were coded into Boolean arrays using the BCI Fingerprint toolkit. They were later compared using various similarity coefficients, clustered using a hierarchical aggregate group-average algorithm and studied using substructure analysis approach. We were able to identify structural features that seem to be correlated with high stability of particular ionic tautomers.

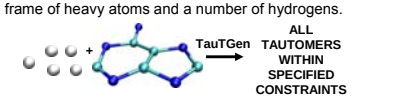
Hybrid QM-C Approach

Hybrid quantum mechanical-combinatorial (HQM-C) approach allows for screening of combinatorially generated library of structures, which properties are calculated with QM methods. The algorithm (flowchart below) consist of 3 major steps marked in blue.



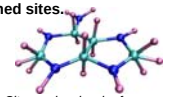
Generation of Tautomers

TauTGen – the TauTomer Generator program builds desired tautomers from a frame of heavy atoms and a number of hydrogens.



TauTGen generates all possible combinations how to distribute all available hydrogens among all defined sites.

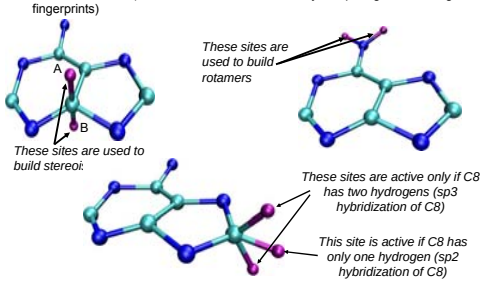
Number of combinations = $\binom{N_{sites}}{N_{hydrogens}}$



Sites and molecular framework

For each generated combination TauTGen checks constraints in the following order:

- constraints on the maximum and minimum number of hydrogens at each atom of molecular framework
  site constraints (see below)
- stereoconfiguration check to reduce the number of stereoisomers to a set of
- diastereoisomers (Enantiomer detection is done by comparing stereoconfiguration fingerprints)



Future of HQM-C

Tautomers

Complication level

- **-Proton transfer**
- **-Developed tools: TauTGen, GOT** — 2005-2006
- **-Characterization at the DFT level**
- **-Library size: ~500 molecules**

Congeners

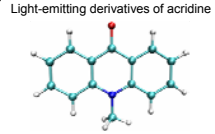
- **-Substitution with halogen atom**
- **-Software tools: ConGener** — 2007
- **-Characterization at semiempicical level**
- **-Library size: ~100000 molecules**

Any derivatives (application in molecular design)

- **-Substitution with any group**
- **-Tools: in preparation** — 2008 (?)
- **-Characterization at semiempirical or DFT level**
- **-Library size > 2000000**

Various Persistent Organic Pollutants exist in the environment as families of halogen substituted congeners (molecules based on the same carbon skeleton but different by a substitution pattern). To facilitate computational studies on these species we developed ConGENER software package. The main program combinatorially generates all congeners of a parent molecule. The quantum-mechanical calculations provide descriptors used for QSAR/QSPR calculations to perform environmental toxicity studies.

Light-emitting derivatives of acridine

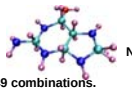


Unsubstituted 9-acridone emits blue light

Derivative of 9-acridone emitting green light.

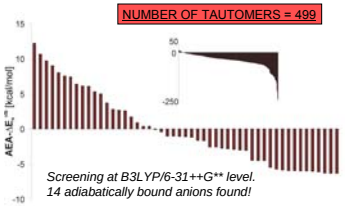
The preliminary experiments conducted in our group reveal that a substitution of the nucleus of 9-acridinone with electron-donating substituents (like methoxy- or methyl- groups) leads to a batochromic shift of the maximum band of the relative emission spectrum. However, only a limited number of derivatives was tested in this context. **In order to develop new systems that can be potentially useful in biomedical analyses we plan to explore the chemical space using HQM-C approach. Excited states will be characterized at the QM level.**

Guanine⁻



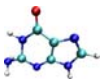
5 hydrogen atoms and 23 sites give 33649 combinations.

NUMBER OF TAUTOMERS = 499

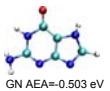


*Screening at B3LYP/6-31++G** level.*
14 adiabatically bound anions found!

Most stable neutrals

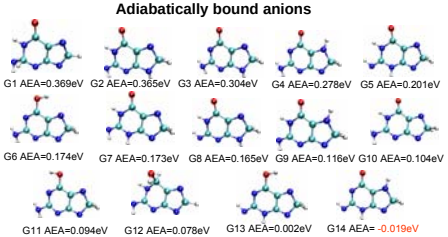


G AEA=-0.459 eV

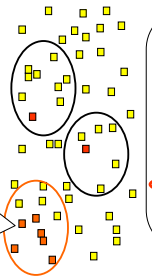


GN AEA=-0.503 eV

Adiabatically bound anions



Hierarchical group-average agglomerative clustering and substructure analysis (based on weighted fingerprints) was performed for a reduced set of 165 tautomers.



Most of adiabatically bound anions are grouped in one cluster

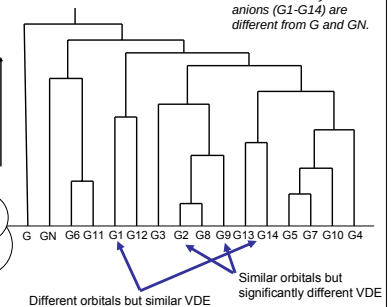
G(N2,N1,N3,2C8), G(N2,N1,N7,2C8), G(N9,N2,N1,2C8), G(2N2,N1,2C8), **G(N2,N3,N7,2C8)**, **G(N9,N2,N3,2C8)**, **G(2N2,N3,2C8)**, G(N9,N2,N7,2C8), **G(2N2,N7,2C8)**, **G(N9,2N2,2C8)**.

Common feature is the presence of two hydrogens on C8

Clustering using Tanimoto similarity

#Cl	#Cla	Clusters (n_act/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

In the later step of clustering one cluster with all top five (and 7 in total) most stable tautomers is formed.



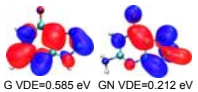
Dendrogram confirms that adiabatically bound anions (G1-G14) are different from G and GN.

The substructural features unique to adiabatically bound anions of guanine are:

- Absent hydrogen at C2, when there is a hydrogen at C4 or C6
- Absent hydrogen at C4, when there is a hydrogen at C2 or C6 or N7
- Absent hydrogen at C5, when there is a hydrogen at N3 or N1 or C8 or N9
- Absent hydrogen at C6, when there is a hydrogen at C2 or C8 or N7
- Absent hydrogen at C8, when there is a hydrogen at C5 or C6

Analysis of SOMOs

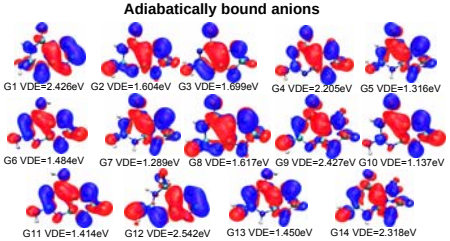
Anions of the most stable neutrals



G VDE=0.585 eV GN VDE=0.212 eV

G and GN anions seems to have significant antibonding character comparing to anions G1-G14.

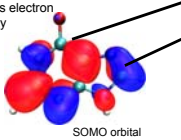
Adiabatically bound anions



Orbital Holograms

Orbital holograms are created by Bader analysis of the excess electron density

N1	C2	N3	C4	…			
0.1	0.15	0.23	0.01	0.3	0.2	0.01	0.02



SOMO orbital

Similarity between orbital holograms A and B is calculated using Euclidean distance

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

Summary

The hybrid quantum mechanical-combinatorial approach was used to identify low energy tautomers of anionic guanine. Many previously unknown tautomers turned out to be more stable than any of the previously known tautomers.

We employed chemoinformatics techniques to characterize these new species. We identified structural features that seem to be correlated with high stability of particular anionic tautomers. We could demonstrate that the majority of the most stable anions are very similar as they are grouped together during clustering. At the same time, the analysis of SOMO electron density demonstrated that these new species are significantly different from the anions of the most stable neutrals.

The future extensions of HQM-C approach are presented.

Acknowledgements

This work was supported by the: (i) Polish State Committee for Scientific Research (KBN) Grant N204 127 31/2963, (ii) European Union COST P9 Program (P9-02569 and P9-02841 STSM grants). M.H. holds the Foundation for Polish Science (FNP) award for young scientists. The calculations were performed at the Academic Computer Center in Gdańsk (TASK) and at the Molecular Science Computing Facility (MSCF) in the William R. Wiley Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the U.S. Department of Energy's Office of Biological and Environmental Research and located at the Pacific Northwest National Laboratory, which is operated by Battelle for the US Department of Energy.