Prediction of Interaction Energies of Substituted Hydrogen-Bonded Watson-Crick Cytosine:Guanine(8X) Base Pairs

ARTICLE *in* THE JOURNAL OF PHYSICAL CHEMISTRY B · APRIL 2009

Impact Factor: 3.3 · DOI: 10.1021/jp8071926 · Source: PubMed

2 AUTHORS, INCLUDING:

Paul L A Popelier
The University of Manchester
190 PUBLICATIONS **7,124** CITATIONS

SEE PROFILE

THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

Prediction of Interaction Energies of Substituted Hydrogen-Bonded Watson#Crick Cytosine:Guanine Base Pairs

Chunxia Xue, and Paul L. A. Popelier

J. Phys. Chem. B, **2009**, 113 (10), 3245-3250• DOI: 10.1021/jp8071926 • Publication Date (Web): 18 February 2009

Prediction of Interaction Energies of Substituted Hydrogen-Bonded Watson−Crick Cytosine:Guanine^{8X} Base Pairs

Chunxia Xue and Paul L. A. Popelier*

Manchester Interdisciplinary Biocentre (MIB), 131 Princess Street, Manchester M1 7DN, Great Britain, and School of Chemistry, University of Manchester, Oxford Road, Manchester M13 9PL, Great Britain

Received: August 12, 2008; Revised Manuscript Received: January 13, 2009

We investigated the variation in the interaction energy between the Watson−Crick hydrogen-bonded DNA base pairs guanine and cytosine (G^{8X}:C), where guanine is substituted in the C8 position by 37 different functional groups. Base pairs were optimized at the B3LYP/6-311+G(2d,p) level. A base pair complex containing a more strongly electron-withdrawing group remarkably forms a more stable base pair with C. Multivariate linear regression provided a quantitative relationship between the interaction energies and descriptors generated by the quantum chemical topology (QCT) approach. The descriptors were sampled from the monomers only, not the supermolecular base pair complexes. A model with $r^2 = 0.96$ and a root-mean-square (rms) value of 0.6 kJ/mol was obtained for a training set of 28 base pair complexes. The model was tested by an external test set of 9 complexes, yielding $r^2 = 0.99$ and an rms value of 0.2 kJ/mol. The results indicated that the bonds $C_6{=}O_6$ and $N_2{-}H_2$ at the hydrogen-bonded frontier of the guanine derivatives play an important role in transmitting the substituent effects. A linear correlation between substitution energies and Hammett constants (σ_m) was also obtained for all 37 substituents, yielding $r^2 = 0.82$ and an rms value of 1.2 kJ/mol. The model based on QCT descriptors can therefore be used for the prediction of the interaction energy of the base pair G^{8X}:C, strictly based on data for the G^{8X} monomers only.

1. Introduction

Watson−Crick- (WC-) type base pairs are formed between complementary hydrogen-bonding acceptor and donor sites of nucleic acid bases cytosine (C), guanine (G), adenine (A), and thymine (T) [uracil (U) in RNA]. This process is fundamental for molecular recognition in duplex formation of DNA,[1] which plays a key role in the working of the genetic code.[2] Thus, it is not surprising that this system of five simple building blocks continues to be a paradigm. Modification of nucleic acid bases has attracted extensive interest because most of them have biological activities.[3,4] For example, halogenated pyrimidines have been synthesized as potential antitumor, antibacterial, and antiviral agents.[5] Substituted nucleic acid bases have also been used in computer-aided molecular design to improve the stability of base pairs.[6]

Hobza and co-workers reviewed the computational work on the interaction energies of natural nucleic acid base pairs.[7,8] Some time ago,[9] it was proposed that the hydrogen bonds in DNA base pairs are basically electrostatic in nature.[10] However, recent work[11−13] by Bickelhaupt et al. revealed that, in WC base pair complexes, donor−acceptor σ-orbital interactions (i.e., charge transfer) between N or O lone pairs on one base and N−H σ^* orbitals on the other base compare in strength to the electrostatic interaction term.

For many years, substituent effects have been commonplace in covalently bonded systems, where they find numerous applications in medical chemistry, biochemistry, organic chemistry, and material chemistry.[14] Studies of substituent effects have contributed much to our understanding of chemical mechanisms and established chemical models and notions.[15] However, so far, little is known on substituent effects in the context of noncovalent interactions. This is crucial to understanding biological processes that rely on these effects in terms of structure and function. Nevertheless, it should be mentioned that, unlike molecular systems, supramolecular systems usually show complicated substituent effects.[16] Presumably, this is caused by the complexity and weakness of interactions simultaneously involved in a supramolecular system. It appears that experimental methods are challenged[17,16] in separating the effect exerted by each single interaction. Computational studies, even on simplified subsystems,[18] might help in the quest for a better understanding of mechanistic supramolecular chemistry.

A small number of studies on substituents affecting hydrogen bonding of modified DNA base pairs have been carried out, by Kawahara et al.,[6] Guerra et al.,[19,20] and Meng et al.[21−23] In these works, the numbers of substituents studied were limited. Moreover, no quantitative structure−interaction energy relationship was developed for the substituted base pairs, based on descriptors calculated from the monomer's structure alone. Very recently, in a companion to the current article,[24] we looked at the effects of 42 substituents on the interaction energy of the $C^{5X,6X}$:G base pair. This set of electron-withdrawing and -donating groups, introduced at the cytosine C6 and C5 positions, led to quantitative models based on quantum chemical topology (QCT).[25,26]

In the present work, we investigate the effects on WC hydrogen bonding when substituents are introduced at the guanine C8 position (denoted G^{8X}). In particular, we analyze how the hydrogen-bond length and strength are affected by replacing hydrogen atom H8 in G by as many as 36 diverse electron-withdrawing and -donating groups. The purpose is to determine the substituent effects caused by these groups, which can be instructive in modifying G interacting with C. Furthermore, we wish to develop a quantitative model relating the interaction energy to the substituted structure of the G monomer.

* Corresponding author.

10.1021/jp8071926 CCC: $40.75
Published on Web 02/18/2009

TABLE 1: Hydrogen-Bond Lengths (Å) and Energies (kJ/mol) of G^{8X}:C Base Pairs

no.	substituent	$O_6\cdots N_4$	$N_1\cdots N_3$	$N_2\cdots O_2$	ΔE^{HB}	BSSE	$\Delta\Delta E^a$	σ_m	predicted $\Delta\Delta E^b$	predicted $\Delta\Delta E^c$
1	−NO	2.84	2.93	2.89	−111.97	2.32	−7.89	0.62	−5.92	−7.72
2	$-NO_2$	2.84	2.94	2.90	−111.25	2.31	−7.17	0.71	−7.07	−6.72
3[d]	−COCl	2.84	2.94	2.91	−109.63	2.32	−5.55	0.51	−4.53	−5.49
4	−CN	2.83	2.94	2.91	−109.40	2.28	−5.32	0.56	−5.16	−5.16
5	−COH	2.83	2.94	2.92	−107.90	2.27	−3.81	0.35	−2.50	−4.06
6	$-CBr_3$	2.83	2.95	2.92	−107.48	2.36	−3.39	0.28	−1.61	−3.34
7[d]	$-CCl_3$	2.82	2.95	2.92	−107.40	2.34	−3.32	0.40	−3.13	−3.17
8	−COOH	2.82	2.95	2.92	−106.82	2.30	−2.74	0.37	−2.75	−2.97
9	$-CH_2F$	2.81	2.95	2.94	−106.64	2.27	−2.55	0.12	0.42	0.06
10	$-CHBr_2$	2.82	2.95	2.93	−106.61	2.31	−2.53	0.31	−1.99	−2.78
11[d]	$-CHCl_2$	2.82	2.95	2.93	−106.49	2.30	−2.40	0.31	−1.99	−2.46
12	$-CHF_2$	2.82	2.95	2.93	−106.31	2.29	−2.23	0.29	−1.73	−2.23
13	−F	2.81	2.95	2.94	−106.26	2.30	−2.17	0.34	−2.37	−1.83
14	−Cl	2.81	2.95	2.93	−106.18	2.30	−2.09	0.37	−2.75	−1.77
15[d]	−Br	2.81	2.95	2.93	−106.11	2.30	−2.02	0.39	−3.00	−1.93
16	$-COCH_3$	2.82	2.95	2.93	−106.08	2.28	−1.99	0.38	−2.88	−2.46
17	$-COOCH_3$	2.82	2.95	2.93	−105.78	2.28	−1.69	0.37	−2.75	−2.04
18	−CCH	2.81	2.95	2.94	−105.63	2.26	−1.55	0.21	−0.72	−1.90
19[d]	$-CF_3$	2.82	2.95	2.92	−105.50	2.31	−1.41	0.43	−3.51	−0.90
20	$-COOC_2H_5$	2.82	2.95	2.93	−105.47	2.29	−1.39	0.37	−2.75	−1.77
21	$-COOC_3H_7$	2.82	2.95	2.93	−105.39	2.28	−1.31	0.37	−2.75	−1.66
22	$-CONH_2$	2.82	2.95	2.93	−104.82	2.27	−0.74	0.28	−1.61	−1.48
23[d]	−SH	2.80	2.95	2.94	−104.46	2.27	−0.37	0.25	−1.23	−0.77
24	−H	2.80	2.95	2.95	−104.08	2.25	0.00	0.00	1.95	0.17
25	$-CHCH_2$	2.80	2.95	2.95	−103.99	2.26	0.09	0.06	1.19	−0.36
26	−OH	2.79	2.95	2.95	−103.86	2.28	0.22	0.12	0.42	−0.15
27[d]	$-CH_3$	2.80	2.95	2.95	−103.32	2.25	0.76	−0.07	2.84	0.95
28	$-OCH_3$	2.79	2.95	2.96	−103.22	2.28	0.86	0.12	0.42	0.65
29	$-CH_2Br$	2.81	2.95	2.93	−103.14	2.27	0.94	0.12	0.42	1.00
30	$-t\text{-}C_4H_9$	2.79	2.95	2.95	−102.99	2.28	1.09	−0.10	3.22	1.52
31[d]	$-OC_2H_5$	2.79	2.95	2.96	−102.98	2.28	1.10	0.10	0.68	0.95
32	$-CH_2Cl$	2.81	2.95	2.94	−102.77	2.27	1.32	0.11	0.55	1.40
33	$-C_2H_5$	2.80	2.95	2.95	−100.13	2.25	3.95	−0.07	2.84	3.85
34	$-n\text{-}C_3H_7$	2.80	2.95	2.95	−99.97	2.24	4.11	−0.06	2.71	3.94
35[d]	$-n\text{-}C_4H_9$	2.80	2.96	2.96	−99.96	2.21	4.12	−0.08	2.96	3.99
36	$-i\text{-}C_3H_7$	2.80	2.95	2.95	−99.92	2.24	4.16	−0.04	2.46	4.07
37	$-i\text{-}C_4H_9$	2.80	2.96	2.96	−99.82	2.23	4.27	−0.08	2.96	4.17

[a] Substitution energies calculated with Gaussian 03. [b] Substitution energies predicted with the Hammett constant σ_m. [c] Substitution energies predicted with the MLR model using QCT descriptors. [d] Test set.

Note that we did not involve the properties of the bond critical points (BCPs) in the supermolecular complexes, in particular, those of the hydrogen bonds. Such an analysis would detract from the main aim of the article, which is to demonstrate that *monomeric* BCP properties can be linked to supermolecular energies. The only information we need from the supermolecular calculations is the energy, not the BCP properties.

2. Computational Details

2.1. Substituents. Table 1 lists the hydrogen-bond lengths and energies of unsubstituted guanine and 36 substituted guanines, and Figure 1 provides the general structure and atom-labeling scheme. The geometry of each isolated monomer and supermolecular complex (G^{8X}:C, WC type) was optimized at the B3LYP level[27,28] with the 6-311+G(2d,p) basis set, using the Gaussian 03 program.[29] The corresponding wave function of each monomer was also generated. Previous work[30,31] confirmed that this level is reliable, as it correctly predicts intermolecular vibrational frequencies. The interaction energies of the complexes were corrected for the basis set superposition error (BSSE) by the counterpoise method.[32] Hereafter, we refer to the molecular interaction energy with the BSSE correction as ΔE^{HB}. The quantity ΔE^{HB} is defined as the difference between the total energy of a given base pair and the energies of the corresponding monomers (eq 1). The more negative ΔE^{HB}, the more the stable base pair complex. The quantity $\Delta\Delta E$, which we call the substituent energy, can be seen as the effect the substituent has on ΔE^{HB} (eq 2). A more negative $\Delta\Delta E$ value means a more stable complex than the unsubstituted G:C base pair.

$$\Delta E^{HB} = E(G^{8X}{:}C) - [E(G^{8X}) + E(C)] \quad (1)$$

$$\Delta\Delta E = \Delta E^{HB}(G^{8X}{:}C) - \Delta E^{HB}(G{:}C) \quad (2)$$

Geometrical optimization was carried out for both monomers and complexes without constraints. Nonplanarity of the exocyclic amino moiety in the isolated bases has been reported before.[7] Although the differences in the energies derived from the planar and nonplanar structures are small (0.6−1.6 kJ/mol depending on the level of theory[33]), this energy difference is not negligible considering the span of substituent energies and therefore had to be taken into account.

2.2. QCT Descriptors. Loosely speaking, bond critical points (BCPs) appear[34] between two nuclei that are bonded. BCPs are points in real three-dimensional space where the gradient of the electron density vanishes ($\nabla\rho = 0$) and where the Hessian of ρ has two negative eigenvalues ($\lambda_1 < \lambda_2 < 0$) and one positive eigenvalue ($\lambda_3 > 0$). Two paths of steepest ascent through ρ originating at a given BCP terminate at the two nuclei that the BCP connects. A collection of such paths forms a molecular

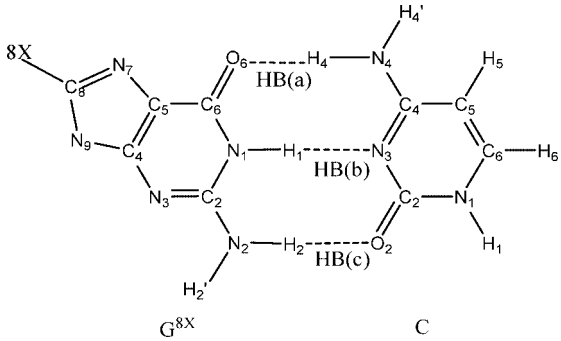

Figure 1. Structure and atomic labeling scheme of substituted WC G^{8X}:C base pairs.

graph, which acts as a "topological shorthand" of the molecule being analyzed. A molecular graph can be supplemented[35] by physical properties evaluated at all BCPs. The electron density itself is such a property, which has been related to bond order through an observed exponential relationship.[36] Another property is the Laplacian of the electron density, which is denoted by $\nabla^2\rho$ and is defined as the sum of the Hessian eigenvalues. The Laplacian measures the local (pointlike) degree of concentration or depletion of ρ in space. A third property is the ellipticity ε, defined as $(\lambda_1/\lambda_2) - 1$. It is always positive at the BCP because $\lambda_1 < \lambda_2 < 0$. The ellipticity measures ρ's deviation from circular symmetry in the plane spanned by the eigenvectors associated with λ_1 and λ_2. Typically, the fourth and fifth properties are two forms of kinetic energy density, denoted by $K(\mathbf{r})$ and $G(\mathbf{r})$, respectively. They are defined as

$$K(\mathbf{r}) = -\frac{1}{4}N\int \mathrm{d}\tau' \,(\psi^*\nabla^2\psi + \psi\nabla^2\psi^*) \tag{3}$$

$$G(\mathbf{r}) = \frac{1}{2}N\int \mathrm{d}\tau'(\nabla\psi^*\cdot\nabla\psi) \tag{4}$$

where $\int \mathrm{d}\tau'$ denotes an integration over the spin coordinates of all N electrons except one and Ψ is the wave function. Interpreting $K(\mathbf{r})$ in chemical terms is not straightforward, although useful formulas describing its link to the Laplacian and the more "classical" kinetic energy $G(\mathbf{r})$ can be found elsewhere.[37] Finally, the equilibrium bond length, R_e, can be regarded as a sixth "BCP property". It is different from the other properties in that it is not a property (or density function) evaluated at the BCP. Still, it can be associated with a BCP because a BCP is topologically connected to two nuclei, which, of course, have an internuclear distance R_e. These BCP properties can be conveniently collected into a six-component vector $(\rho,\nabla^2\rho,\varepsilon,K,G,R_e)$ that describes each bond.

After generation of the ab initio wave function of each *monomeric* nucleic acid base, the wave function file was read by a local version of the program MORPHY.[38,39] This program extracts the required bond descriptors and exports them into a format that is convenient for subsequent statistical analysis. Note that no topological information from the supermolecular wave functions is generated.

BCP properties have featured under the quantum topological molecular similarity (QTMS) method.[40−42] Over the years, QTMS has provided successful quantitative relationships between a variety of activities and QCT descriptors. Among the series of QTMS studies, one finds ecological applications,[43−47] medicinal applications,[48−52] and quantitative structure−property relationships.[44,53−61] It is clear that QCT descriptors manage to

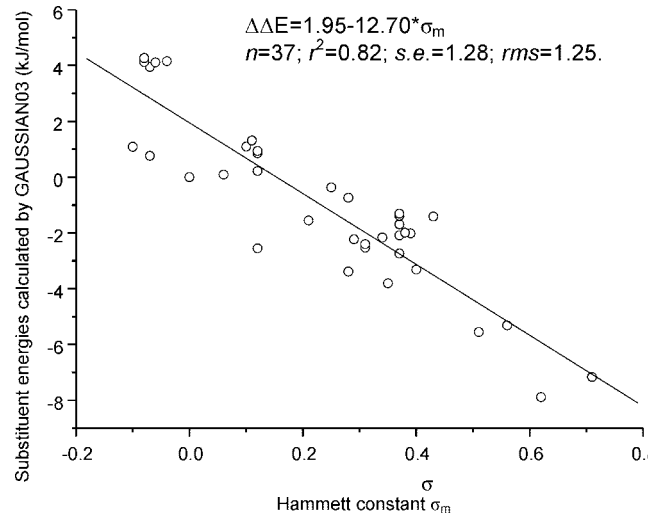

Figure 2. Relationship between the substituent energies (kJ/mol) and the Hammett constant σ_m.

capture electronic effects in a discerning manner. Hence, they can replace Hammett constants,[56,62] both the original and their later extensions.

2.3. Feature Selection and Regression Analysis. The data set was split into a training set and a true test set in order to facilitate external validation. Predictions were made for the test set after regression models had been built from the data in the selected training set. Correlation analysis of the descriptors was performed first, once the descriptors had been generated. According to this analysis, pairs of descriptors with an absolute correlation coefficient $|r|$ above the threshold value of 0.85 contain at least one descriptor that needs to be eliminated. Subsequently, we used descriptor-screening methods to select the most relevant descriptor. This is an important step in the construction of a predictive model. Then, stepwise multiple linear regression (MLR) accomplished the feature selection method choosing the subset of molecular descriptors.[63,64] In the current stepwise model-building technique (for regression designs with a single dependent variable), one identifies an initial model and repeatedly alters the model from the previous step by adding (forward step) or removing (backward step) a predictor variable. The search terminates when the stepping procedure does not further improve the model. For the initial linear regression step, we used the best single predictor, which is the most significant variable. Next, we added descriptors one at a time, always adding the one that most improved the fit, until the fit did not significantly improve. We constructed the regression equation once all of the significant variables had been determined. The number of variables retained in the model is based on the levels of significance assumed for inclusion of variables within the model and their exclusion from it.

After descriptor selection, MLR determines the linear model, given in the equation

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n \tag{5}$$

where Y is the dependent variable, X_i $(i = 1-n)$ represents the descriptors, with coefficients b_i $(i = 1-n)$, and b_0 is the intercept. The model's performance was described by means of the fitting power of the parameters (r^2), the standard error of the estimates (s) and root-mean-square (rms) error. The F value of the Fisher test and t-test values were also determined at a significance level of 0.05. Any developed model, based on a designed training

TABLE 2: Correlation Equations for the Linear Model Based on QCT Descriptors[a]

descriptor	unstandardized coefficients	unstandardized error	standardized coefficients	*t*-test	sig
constant	−126.28	37.24		−3.39	0.00
$\varepsilon_{C_6=O_6}$	−705.74	169.82	−0.26	−4.16	0.00
$G_{N_2-H_2}$	3865.81	308.91	0.77	12.51	0.00

[a] Training set: $n = 28$, $r^2 = 0.96$, $F = 328.43$, standard error = 0.61 (kJ/mol), rms error = 0.58 (kJ/mol), $N = 28$. Test set: $n = 9$, $r^2 = 0.99$, rms error = 0.24 (kJ/mol).

set, was externally validated by evaluating the prediction errors and the rms error and r^2. Good predictive properties are an additional indication that chance correlation has been avoided. The SPSS program[65] was used for feature selection and MLR analysis.

3. Results and Discussion

3.1. Substitution Effects Calculated ab Initio. Table 1 lists the 37 substituents and their corresponding counterpoise-corrected hydrogen-bond energies (ΔE^{HB}), the substitution effects ($\Delta\Delta E$), and the basis set superposition errors (BSSEs). The interaction energy of the natural G:C base pair (no. 24) is 104.08 kJ/mol. From no. 1 (NO) to no. 23 (SH), the values of $\Delta\Delta E$ are all negative, which means that these modified base pairs are more stable than unmodified G:C. From no. 25 ($CHCH_2$) to no. 37 (i-C_4H_9), the $\Delta\Delta E$ values are all positive, indicating that these base pairs are less stable than G:C. The most stable and unstable complexes were $G^{8\text{-NO}}$:C and $G^{8\text{-}i\text{-}C_4H_9}$:C, with interaction energies of −111.97 and −99.82 kJ/mol, respectively. Within this range of 12.15 kJ/mol, there is a homogeneous distribution of data. This is important for the reliability of the linear regression. From Table 1, a remarkable tendency was observed, namely, that G^{8X} derivatives containing strong electron-withdrawing groups (such as F, CN, or NO_2) form a more stable base pair with C, which is in agreement with earlier work.[6,22]

Table 1 also lists the hydrogen-bonds length of G^{8X}:C. We calculated the lengths of the three hydrogen bonds $O_6\cdots N_4$ [HB(a)], $N_1\cdots N_3$ [HB(b)], and $N_2\cdots O_2$ [HB(c)] in G:C to be 2.80, 2.95, and 2.95 Å, respectively. This corresponds to a short−long−long pattern. The corresponding lengths measured by experimental X-ray crystallography are 2.91, 2.95, and 2.86 Å (long−long−short), respectively.[66] A dedicated study[13] attributed this discrepancy, also noted at the BP86/TZ2P level, to the molecular environment (water, sugar hydroxyl groups, counterions) of the base pairs in the experimentally studied crystals. From Table 1, it can also be seen that, in G^{8X} derivatives containing an electron-withdrawing group, most HB(a) bonds are elongated, and most HB(c) bonds are contracted. For G^{8X} derivatives containing an electron-donating group, most HB(a) and HB(c) bond lengths differ very little from those in G:C. Finally, no substituent has much influence on the HB(b) bond lengths.

3.2. Substitution Energies Calculated with the Hammett Constant σ_m. The Hammett substituent constants, which quantify the electron-withdrawing or -donating capabilities of a given substituent, have enjoyed much success as structural descriptors in quantitative structure−activity/property relationships.[67] First, we seek a relationship between the substituent energies and the Hammett constant σ_m. A fairly linear correlation was found, as shown in eq 6 and in Figure 2

$$\Delta\Delta E = 1.95\ (\pm 0.32) - 12.70\ (\pm 1.00)\sigma_m \qquad (6)$$

where $N = 37$, $r^2 = 0.82$, standard error = 1.28 (kJ/mol), and rms error = 1.25 (kJ/mol).

The negative slope means that the substituent energy decreases with increasing σ_m, corresponding to more stable base pairs. Substituent energies predicted by eq 6 are listed in Table 1.

Although eq 6 has a fairly good correlation coefficient, the standard error (1.28 kJ/mol) and rms error (1.25 kJ/mol) are quite large compared to the magnitude of the substitution energies found here. The Hammett methodology, which is empirical in nature, can be replaced by simple ab initio descriptors. Therefore, in the next step, we make use of QCT descriptors.

3.3. Results of MLR Model Based on QCT Descriptors. The BCP properties, calculated for a given substituted G, act as independent variables, and the corresponding substituent energy (calculated by Gaussian) is the dependent variable. After the correlation analysis of the descriptors, the stepwise regression routine provided a linear model. This model predicts substitution energies for G^{8X}:C from calculated BCP properties. The data set was divided randomly into two subsets: the test set (nos. 3, 7, 11, 15, 19, 23, 27, 31, and 35) and the training set (nos. 1, 2, 4−6, 8−10, 12−14, 16−18, 20−22, 24−26, 28−30, 32−34, 36, and 37) (9 and 28 points, respectively). The training set was used to build the MLR model, and the test set was used to evaluate its prediction ability. The best linear model contains two QCT descriptors. The regression coefficients of the descriptors are listed in Table 2. The linear correlation coefficient value of the two descriptors is 0.78, which means that the descriptors are independent in this MLR analysis because this value is less than the threshold of 0.85. The two descriptors selected were the ellipticity of the $C_6{=}O_6$ BCP ($\varepsilon_{C_6=O_6}$) and the kinetic energy density $G(\mathbf{r})$ of the $N_2{-}H_2$ BCP ($G_{N_2-H_2}$). In Table 2, the values for "sig" are all zero, which indicates that the selected descriptors are significant predictors of substituent energies. External prediction results were obtained for the test set, listed in Table 1, confirming the predictive capability of the model. This model gives an rms error of 0.58 kJ/mol for the training set, 0.24 kJ/mol for the test set, and 0.52 kJ/mol for the whole set. The corresponding correlation coefficients (r^2) are 0.96, 0.99, and 0.97, respectively. Through the comparison of the r^2 and rms values for the total data set, we can see that the performance of the model based on QCT descriptors is much better than that of the model based on the Hammett constant σ_m. Figure 3 shows the substitution energies obtained by the QCT model versus those calculated by Gaussian 03, for both the training set and the test set.

The two descriptors in the MLR model correspond to atoms involved in hydrogen bonding. Atom O_6 is involved in HB(a), and atom H_2 is involved in HB(c). The descriptor $\varepsilon_{C_6=O_6}$ receives a negative coefficient in the regression; this indicates that the substituent energy $\Delta\Delta E$ decreases, and hence the substituted base pair becomes more stable, with increasing ellipticity of the $C_6{=}O_6$ bond. On the other hand, $G_{N_2-H_2}$ receives a positive coefficient in the regression, indicating that the base pair become less stable upon an increase in $G_{N_2-H_2}$. The higher the *t*-test value (Table 2), the more influence the descriptor has on the

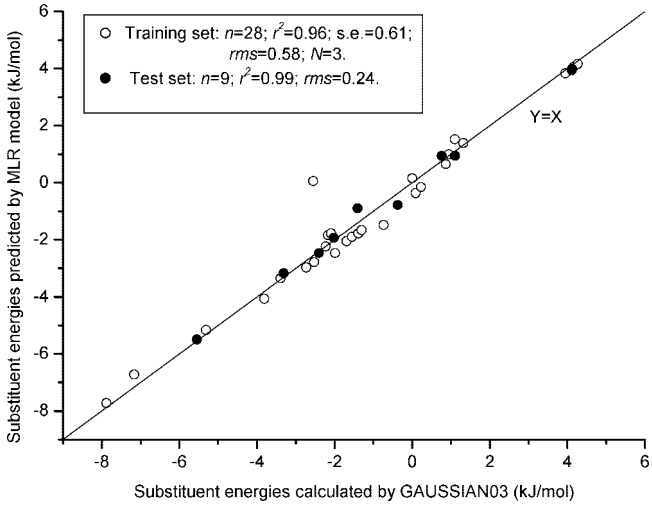

Figure 3. Substituent energies ($\Delta\Delta E$) predicted by the MLR model (based on QCT descriptors) versus ab initio energies.

substituent energy. Hence, $G_{N_2-H_2}$ has more influence. Equation 7 summarizes how substituent energies can be predicted from

$$\Delta\Delta E = -705.74\varepsilon_{C_6=O_6} + 3865.81G_{N_2-H_2} - 126.28 \quad (7)$$

descriptors obtained from a newly substituted G monomer where the descriptors are inserted in atomic units and the energy emerges in kJ/mol. The importance of kinetic energy is remarkable, and the physical origin of its importance is unfortunately not clear at this stage. It should be emphasized that it is property of the monomer, not of the complex, for example, in the context of hydrogen-bond analysis of intermolecular critical points.

4. Conclusions

Quantum chemical calculations were performed on 37 Watson−Crick base pairs formed between 8-position-substituted G and unmodified C. The presence of an electron-withdrawing group on the 8-position of G forms a more stable base pair with C. Linear regression between the substituent energies and the Hammett constant σ_m yielded $r^2 = 0.82$ and a root-mean-square (rms) error of 1.25 kJ/mol. However, MLR using only two QCT descriptors provided $r^2 = 0.96$ and an rms error of 0.58 kJ/mol for the training set of about three-quarters of the compounds. For the external test set of the (randomly chosen) remaining compounds, we obtained $r^2 = 0.99$ and an rms error 0.24 kJ/mol. Compared to the full energy range of all compounds, this corresponds to an error of about 2%. Hence, this model can be used to the estimate stabilities of Watson−Crick G^{8X}:C base pairs. Because only monomeric data are required, this model provides a savings in computational time. Moreover, unlike for Hammett constants, no empirical data are necessary, which enables predictions for novel substituents.

Acknowledgment. The authors thank the European Union for allocating an "Individual Fellowship" (Marie Curie, FP6-2004-Mobility, Proposal no. 021966-DUPLEX).

References and Notes

(1) Watson, J. D.; Crick, F. H. *Nature* **1953**, *171*, 737.
(2) Brenner, S.; Jacob, F.; Meselson, M. *Nature* **1961**, *190*, 576.
(3) Sulkowska, A.; Rownicka, J.; Bojko, B.; Sulkowski, W. *J. Mol. Struct.* **2003**, *133*, 651.
(4) Lamsabhi, M.; Alcami, M.; Mo, O.; Bouab, W.; Esseffar, M.; Abboud, J. L. M.; Yanez, M. *J. Phys. Chem. A* **2000**, *104*, 5122.
(5) Morris, S. M. *Mutat. Res.* **1993**, *297*, 39.
(6) Kawahara, S.; Uchimaru, T. *Eur. J. Org. Chem.* **2003**, 2577.
(7) Hobza, P.; Sponer, J. *Chem. Rev.* **1999**, *99*, 3247.
(8) Hobza, P.; Zahradnik, R.; Mueller-Dethlefs, K. *Collect. Czech. Chem. Commun.* **2006**, *71*, 443.
(9) Morokuma, K. *Acc. Chem. Res.* **1977**, *10*, 294.
(10) Sponer, J.; Jurecka, P.; Hobza, P. *J. Am. Chem. Soc.* **2004**, *126*, 10142.
(11) Guerra, C. F.; Bickelhaupt, F. M. *Angew. Chem., Int. Ed.* **1999**, *38*, 2942.
(12) Guerra, C. F.; Bickelhaupt, F. M.; Snijders, J. G.; Baerends, E. J. *Chem. Eur. J.* **1999**, *5*, 3581.
(13) Guerra, F. C.; Bickelhaupt, F. M.; Snijders, J. G.; Baerends, E. J. *J. Am. Chem. Soc.* **2000**, *122*, 4117.
(14) Hansch, C.; Leo, A.; Taft, R. W. *Chem. Rev.* **1991**, *91*, 165.
(15) Anslyn, E. V.; Dougherty, D. A. *Modern Physical Organic Chemistry*; University Science Books: Sausalito, CA, 2006.
(16) Ying, F.; Lei, L.; Qing-Xiang, G. *Chem. Res. Chin. Univ.* **2002**, *18*, 348.
(17) Kool, E. T. *Annu. Rev. Biomol. Struct.* **2001**, *30*, 1.
(18) Sinnokrot, M. O.; Sherrill, C. D. *J. Am. Chem. Soc.* **2004**, *126*, 7690.
(19) Guerra, C. F.; van der Wijst, T.; Bickelhaupt, F. M. *Struct. Chem.* **2005**, *16*, 211.
(20) Guerra, C. F.; van der Wijst, T.; Bickelhaupt, F. M. *Chem. Eur. J.* **2006**, *12*, 3032.
(21) Meng, F.; Liu, C.; Xu, W. *Chem. Phys. Lett.* **2003**, *373*, 72.
(22) Meng, F.; Wang, H.; Xu, W.; Liu, C. *Int. J. Quantum Chem.* **2005**, *104*, 79.
(23) Meng, F.; Wang, H.; Xu, W.; Liu, C.; Wang, H.; Xu, W.; Liu, C. *Chem. Phys.* **2005**, *308*, 117.
(24) Xue, C. X.; Popelier, P. L. A. *J. Phys. Chem. B* **2008**, *112*, 5257.
(25) Popelier, P. L. A. *Atoms in Molecules. An Introduction*; Pearson Education: London, 2000.
(26) Bader, R. F. W. *Atoms in Molecules. A Quantum Theory*; Oxford University Press: Oxford, U.K., 1990.
(27) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev.* **1988**, *B37*, 785.
(28) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
(29) 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 03*, revision C.02, Gaussian, Inc.: Wallingford, CT, 2004.
(30) Joubert, L.; Popelier, P. L. A. *Phys. Chem. Chem. Phys.* **2002**, *4*, 4353.
(31) Nir, E.; Kleinermanns, K.; de Vries, M. S. *Nature* **2000**, *408*, 949.
(32) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553.
(33) Sponer, J.; Florian, J.; Hobza, P.; Leszczynski, J. *J. Biomol. Struct. Dyn.* **1996**, *13*, 827.
(34) Popelier, P. L. A. *Chem. Phys. Lett.* **1994**, *228*, 160.
(35) Popelier, P. L. A. Molecular similarity and complimentarity based on the theory of atoms in molecules. In *Molecular Similarity in Drug Design*; Dean, P. M., Ed.; Chapman and Hall: London, 1995; p 215.
(36) Bader, R. F. W.; Slee, T. S.; Cremer, D.; Kraka, E. *J. Am. Chem. Soc.* **1983**, *105*, 5061.
(37) Bader, R. F. W.; Preston, H. J. T. *Int. J. Quantum Chem.* **1969**, *3*, 327.
(38) Popelier, P. L. A. *MORPHY98: A program for the topological analysis of the electron distribution*; written by P. L. A. Popelier with contributions from R. G. A. Bone, University of Manchester, Manchester, U.K., 1998.
(39) Popelier, P. L. A. *Comput. Phys. Commun.* **1996**, *93*, 212.
(40) Popelier, P. L. A. *J. Phys. Chem. A* **1999**, *103*, 2883.
(41) O'Brien, S. E.; Popelier, P. L. A. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 764.
(42) Popelier, P. L. A. Quantum topological molecular similarity—Past, present and future. In *EuroQSAR2002: Designing Drugs and Crop*

Protectants: Processes, Problems and Solutions; Ford, M., Livingstone, D. J., Dearden, J., van de Waterbeemd, H., Eds.; Blackwell: Oxford, U.K., 2003; p 130.

(43) Chaudry, U. A.; Popelier, P. L. A. *J. Phys. Chem. A* **2003**, *107*, 4578.

(44) Popelier, P. L. A.; Chaudry, U. A.; Smith, P. J. *J. Chem. Soc., Perkin Trans. 2* **2002**, 1231.

(45) O'Brien, S. E.; Popelier, P. L. A. Quantum Molecular Similarity: Use of Atoms in Molecules Derived Quantities as QSAR Variables. In *ECCOMAS Proceedings*, Barcelona, Spain, 2000.

(46) O'Brien, S. E. Quantum Molecular Similarity, an Atoms in Molecules Approach. Ph.D. Thesis, Department of Chemistry, University of Manchester, Manchester, U.K., 2000.

(47) Roy, K.; Popelier, P. L. A. *QSAR Comb. Sci.* **2008**, *27*, 1006.

(48) O'Brien, S. E.; Popelier, P. L. A. *J. Chem. Soc., Perkin Trans. 2* **2002**, 478.

(49) Popelier, P. L. A.; Chaudry, U. A.; Smith, P. J. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 709.

(50) Smith, P. J.; Popelier, P. L. A. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 135.

(51) Roy, K.; Popelier, P. L. A. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 2604.

(52) Mohajeri, A.; Hemmateenejad, B.; Mehdipour, A.; Miri, R. *J. Mol. Graph. Modell.* **2008**, *2008*, 1057.

(53) Alsberg, B. K.; Marchand-Geneste, N.; King, R. D. *Chemom. Intell. Lab. Syst.* **2000**, *54*, 75.

(54) Alsberg, B. K.; Marchand-Geneste, N.; King, R. D. *Anal. Chim. Acta* **2001**, *446*, 3.

(55) Chaudry, U. A.; Popelier, P. L. A. *J. Org. Chem.* **2004**, *69*, 233.

(56) Smith, P. J.; Popelier, P. L. A. *Org. Biomol. Chem.* **2005**, *3*, 3399.

(57) Singh, N.; Loader, R.; O'Malley, P. J.; Popelier, P. L. A. *J. Phys. Chem. A* **2006**, *110*, 6498.

(58) Hemmateenejad, B.; Mohajeri, A. *J. Comput. Chem.* **2008**, *29*, 266.

(59) Esteki, M.; Hemmateenejad, B.; Khayamian, T.; Mohajeri, A. *Chem. Biol. Drug Des.* **2007**, *70*, 413.

(60) Buttingsrud, B.; Alsberg, B.; Astrand, P.-O. *J. Comput. Chem.* **2007**, *28*, 2130–2139.

(61) Buttingsrud, B.; Alsberg, B. K.; Astrand, P.-O. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2226–2233.

(62) Smith, P. J. Quantum Chemical Topological Properties as Electronic Descriptors in Quantitative Structure−Activity/Property Relationships. Ph.D. Thesis, Department of Chemistry, University of Manchester, Manchester, U.K., 2003.

(63) Darlington, R. B. *Regression and Linear Models*; McGraw-Hill: New York, 1990.

(64) Massart, D. L.; Vandeginste, B. M. G.; Buydens, L. M. C.; De Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics. Part A.*; Elsevier: Amsterdam, The Netherlands, 1997.

(65) *SPSS*, version 10.0.7; SPSS Inc.: Chicago, IL, 2000.

(66) Saenger, W. *Principles of Nucleic Acid Structure*; Springer-Verlag: New York, 1984.

(67) Hammett, L. P. *Physical Organic Chemistry*; McGraw-Hill: New York, 1940.

JP8071926