Rapid Prediction of Chemical Metabolism by Human UDP-glucuronosyltransferase Isoforms Using Quantum Chemical Descriptors Derived with the Electronegativity Equalization Method

Michael J. Sorich,*,† Ross A. McKinnon,† John O. Miners,‡ David A. Winkler,§ and Paul A. Smith‡

School of Pharmacy and Medical Sciences, University of South Australia, Frome Road, Adelaide, SA, 5000, Australia, Department of Clinical Pharmacology, Flinders University and Flinders Medical Centre, Bedford Park SA, Australia, and Division of Molecular Science, CSIRO, Clayton, Victoria, Australia

Received June 8, 2004

This study aimed to evaluate in silico models based on quantum chemical (QC) descriptors derived using the electronegativity equalization method (EEM) and to assess the use of QC properties to predict chemical metabolism by human UDP-glucuronosyltransferase (UGT) isoforms. Various EEM-derived QC molecular descriptors were calculated for known UGT substrates and nonsubstrates. Classification models were developed using support vector machine and partial least squares discriminant analysis. In general, the most predictive models were generated with the support vector machine. Combining QC and 2D descriptors (from previous work) using a consensus approach resulted in a statistically significant improvement in predictivity (to 84%) over both the QC and 2D models and the other methods of combining the descriptors. EEM-derived QC descriptors were shown to be both highly predictive and computationally efficient. It is likely that EEM-derived QC properties will be generally useful for predicting ADMET and physicochemical properties during drug discovery.

Introduction

Quantum chemical (QC) descriptors have proven useful for the prediction of many molecular biological and physicochemical properties of interest to the pharmaceutical industry. These include molecular toxicity, absorption, metabolism, receptor binding, octanol/water partition coefficients, stability, pK_a, and chromatographic retention times.[1−13]

Numerous types of descriptors have been developed to capture QC properties. Prominent among these are atomic charges, polarizability, molecular orbital energies, superdelocalizabilities, dipole moments, frontier orbital densities, and molecular quantum similarity measures.[5−7,14] While global QC descriptors characterizing the molecule as a whole generally have wider applicability, atomic or local properties (such as atomic charges or frontier orbital electron densities) are often useful for studying structural analogues and understanding regioselectivity.[5,8,15,16] Recently, an electronegativity equalization method (EEM) has been described for the fast connectivity- and geometry-dependent calculation of density functional theory (DFT) molecular and atomic properties.[17,18] These properties include molecular equalized electronegativity, molecular hardness, molecular softness, atomic charges, atomic Fukui functions, and atomic softness.[17,18] EEM generally allows calculation of these QC properties in a fraction of the time required for ab initio and semiempirical approaches. For example, atomic charges may be calculated in the order of a million molecules of reasonable size in 1 h using a personal computer.[17,18] This potentially allows for QC properties to be used as part of high-throughput in silico screens in drug discovery programs.

Global QC indices describe properties of the whole molecule. Molecular equalized electronegativity refers to the change in energy of the molecule as the number of electrons in the systems is perturbed and thus can be approximated by the average of the molecular ionization energy and electron affinity.[19,20] Simplistically, molecular electronegativity can be considered as a measure of the tendency of a molecule to attract electrons. The hardness global reactivity index is defined by the change in molecular electronegativity with perturbation in electron number and provides an indicator of overall stability of the system.[21,22] The final global reactivity descriptor, molecular softness, is inversely related to the molecular hardness.[19,20,22]

Local QC indices describe properties of atoms in a molecule. The atomic charge provides a local reactivity index that is defined as the difference between the nuclear charge and the total (inner and valence) electron density attributed to the atom. Whereas a neutral, isolated atom has a net charge of zero, the formation of bonds in a molecule leads to a redistribution of the valence electron density (governed by the electronegativity of the atoms). This usually results in areas of charge imbalance (between the positive charge associated with the nucleus and the immediately surrounding electron charge) within the molecule.[17] The Fukui function is a space distribution that describes the change in electron density as the total number of electrons in the molecular system is altered.[15,21,22] This indicates how incoming or outgoing electrons are redistributed in various regions of the molecule. As such, the atom-condensed Fukui function (the Fukui function integrated over the region attributed to an atom) is useful as an indicator of relative atomic susceptibility to electrophilic or nucleophilic attack.[15,21,22] Softness is

* To whom correspondence should be addressed. Phone: +61 8 8302 2034. Fax: +61 8 8302 2389. E-mail: michael.sorich@unisa.edu.au.
† University of South Australia.
‡ Flinders University and Flinders Medical Centre.
§ CSIRO.

10.1021/jm0495529 CCC: $27.50
Published on Web 09/15/2004

the Fukui function scaled by the molecular hardness.[19] Consequently reactivity values may be compared between atoms in different molecules using the atom-condensed softness.

Conjugation with glucuronic acid, derived from the cofactor UDP-glucuronic acid, is an essential phase II clearance mechanism for drugs from all therapeutic classes.[23] Moreover, glucuronidation serves as an elimination mechanism for many endogenous compounds (e.g., bilirubin, bile and fatty acids, steroid hormones), dietary chemicals, and environmental pollutants and facilitates the excretion of the products of phase I metabolism. Glucuronidation reactions are catalyzed by the enzyme UDP-glucuronosyltransferase (UGT). Consistent with its broad substrate profile, UGT exists as a superfamily of enzymes. Sixteen functional human UGT isoforms have been identified to date, and all but one of these (UGT2A1, an enzyme expressed in nasal epithelium) have been classified in just two subfamilies (UGT1A and UGT2B) based on amino acid sequence identity.[24] The individual isoforms exhibit distinct, but overlapping, substrate selectivities and differ in terms of regulation of expression and drug−drug interactions.[24]

Although structure−function studies involving UGT were first reported over 50 years ago,[25] it is only recently that the physicochemical and structural features influencing the metabolism of diverse chemicals by individual isoforms have been explored.[25−30] Notably, three pattern recognition methodologies were compared for their ability to classify chemicals as substrates or nonsubstrates for 12 human UGT isoforms, using simple 2D chemical descriptors.[29] In general, the support vector machine methodology was found to generate the most predictive models.[29] While UGT 1A3, 1A4, 1A6, 1A7, 1A9, 1A10, 2B4, and 2B17 were predicted well (>75% test set predicted correctly), scope remained for improvement particularly with the remaining isoforms. Furthermore, it was unclear whether the suboptimal prediction of glucuronidation for these isoforms was due to the complex chemical data sets, insufficient data, or noisy data. Additionally, physicochemical interpretation of the models generated with the 2D descriptors was not possible because of the nature of chemical properties/descriptors and pattern recognition methods used.

Subsequently, a multiple pharmacophore methodology was designed, implemented, and applied to determine structural features associated with substrates and nonsubstrates of human UGT isoforms.[30] The models generated using this approach were more interpretable but capable of modeling only UGT 1A6, 1A7, 1A9, and 2B4 well. Pharmacophore models generally only account for the three-dimensional distribution of simple chemical features such as hydrogen bond donors and acceptors and hydrophobic regions.[31] Thus, it was anticipated that a more accurate representation of the electrostatic distribution of the molecule would likely improve predictivity for the suboptimally modeled isoforms.[24]

UGT catalyzes the conjugation of lipophilic chemicals containing a suitable acceptor functional group (typically −OH, −COOH, $-NR_x$) with UDP-glucuronic acid according to a second-order nucleophilic substitution mechanism.[32] This suggests that QC descriptors capturing information on molecular and atomic nucleophilicity and pK_a may correlate well with metabolism by UGT. Previous studies in rats using congeneric series of chemicals demonstrated that QC properties have a significant influence on the extent of chemical glucuronidation in vivo,[2,33−35] further supporting this hypothesis.

This study aimed to assess the ability of EEM-derived QC descriptors to generate predictive structure−activity relationships. The second aim was to explore the use of electronic properties to improve the prediction of metabolic reactions catalyzed by 12 human UGT isoforms.

Materials and Methods

Data Sets. As described in previous work,[29,30] 12 isoform-specific data sets of substrates and nonsubstrates for each UGT isoform, ranging in size from 50 to 250 chemicals, were collated from the literature. Data were generated from assays utilizing a single recombinant UGT isoform.

Assessment of Models. To assess the ability of the models generated to predict the metabolism of new chemicals (generalization performance), 30% of each data set was randomly chosen as the test set. The remaining 70% of chemicals were used to generate the models. The same training and test sets were used as described in previous work,[29,30] allowing for performance comparison of the two methods described here with models generated from generic 2D chemical descriptors and pharmacophores. The test set was not used in any way to influence the training or selection of the models. Indeed, the test set was predicted and compared against the known experimental result only after the models were completely defined. Unless defined otherwise, the generalization performance of the models was expressed as the percent of test set chemicals that was correctly predicted (both substrates and nonsubstrates).

QC and 2D Descriptors. Chemical structures for the 523 chemicals in the 12 data sets were constructed using ChemDraw (CambridgeSoft, MA). Three-dimensional structures were optimized by minimizing the universal force field empirical energy, as implemented in Cerius2 (Accelrys, CA). Equalized molecular electronegativity, molecular hardness, molecular softness, atomic charge, atomic softness, and atomic Fukui function values were calculated for each chemical using in-house software implementing the EEM algorithm outlined in Bultinck et al.[17,36]

The EEM procedure is based on the electronegativity equalization principle. This asserts that when molecules are formed, atoms with initially different electronegativities combine in such a way that the electronegativities of the molecular atoms become equal, thereby yielding the molecular equalized electronegativity.[22,36] Electron transfer takes place from atoms with lower electronegativity to those with higher electronegativity, the latter reducing their electronegativity value and the former increasing it.[22] Applying EEM to an n-atom molecule, the atomic charges and the molecular electronegativity are determined by solving a set of $n + 1$ linear equations.[17] By equilibration of the individual atomic electronegativities to the molecular electronegativity, n of these equations are obtained. The remaining equation comes from constraining the sum of the molecular charge to the total molecular charge.[17] The Fukui function and local softness are calculated in a similar manner.[17] The parametrization of this method is currently limited to chemicals containing only H, C, O, N, and F atoms.[17] Thus, chemicals with atoms other than these were excluded from the data sets used in this study. Molecular descriptors capturing the distribution of the atomic QC properties (Table 1) were subsequently calculated for each chemical using in-house software written in Python.

The 2D descriptors used in this study were identical to those described in Sorich et al.[29] Briefly, the 67 2D descriptors calculated comprised counts of simple chemical fragments (atom types,[37] functional groups, and rings), eigenvalue descriptors,[38] and connectivity indices (vertex degree and valence vertex degree).[39] These were originally chosen on the basis of simplicity, ease of calculation, and diverse representation of chemical properties. The 2D models described here differ from

Table 1. Quantum Chemical Molecular Descriptors

Equalized molecular electronegativity, molecular hardness, and molecular softness
Most positive and most negative charge on any atom and specific atoms (H, C, O, N, F)
Maximum charge separation between any atom and between atoms of the same element (H, C, O, N, F)
Sum of squares of charge of all atoms and atoms of the same element (H, C, O, N, F)
Mean of positive atomic charges, negative atomic charges and absolute atomic charges
Relative positive charge (max negative charge/sum of negative charges) and relative negative charge (min negative charge/sum of negative charges)
Most positive and most negative atom-condensed Fukui function on any atom and specific atoms (H, C, O, N, F)
Most positive and most negative atom-condensed softness on any atom and specific atoms (H, C, O, N, F)

those previously reported[29] in the following ways: (a) variable selection was not used; (b) only chemicals that were parametrized for EEM were included.

Descriptor Preprocessing. Prior to model generation with the pattern recognition methods, data were processed to improve prediction performance. Descriptors containing minimal variance were excluded. Specifically, any descriptor where 90% or more of the values were the same across the data set was considered redundant. All remaining descriptors were scaled to set the mean value to zero and the variance to 1, allowing all descriptors to have equal weighting in the training process.

Support Vector Machine (SVM). Classification models were generated using the ν-SVM methodology.[40] Briefly, this algorithm operates by fitting a hyperplane such that the largest margin is formed between two classes of chemicals while minimizing the classification errors. Nonlinearity in a data set is accounted for with kernel functions, which map the input vectors to some higher dimensioned space such that a hyperplane can be found with reduced classification errors.[41] The SVM models were generated using the LIBSVM implementation[42] of the ν-SVM algorithm.[40] For all the data sets, the radial basis function kernel was used with the default value of the γ parameter (=1/(number of descriptors)) and the ν parameter set to 0.1.

Cluster Analysis−Genetic Algorithm−Partial Least Squares Discriminant Analysis (Cluster−GA−PLSDA). A second set of classification models were based on partial least squares discriminant analysis. A combination of cluster analysis and genetic algorithm optimization was used to choose a small, diverse, and relevant subset of descriptors for the model. The details of this algorithm are given in Sorich et al.[30] This method has the advantage that the models are interpretable, providing the descriptors are interpretable.

Combining QC and 2D Models. With the aim of improving predictive ability, three different methods ("Maximum", "Combined", and "Consensus") were used to combine information contained in the QC and 2D descriptors. The "Maximum" method involved selecting either the 2D or the QC model for each isoform based on which gave the maximum test set performance. The "Combined" method involved regenerating the models using the combined 2D and QC descriptors. Finally, the "Consensus" method entailed determination of consensus between the predictions of the QC and 2D models. When the two models were in consensus, the chemical was predicted; otherwise, the chemical was labeled as "uncertain". A paired *t*-test was used to compare the predictive ability of the best method to all others.

Results

The predictive ability of the QC descriptors is displayed in Table 2. By use of the SVM methodology, 78% of chemicals in the test sets were predicted correctly, on average, compared to 73% for models generated with cluster−GA−PLSDA. Statistical analysis of the data in Table 2 using a paired *t*-test, indicated that there was no statistically significant difference between the mean predictive ability ($p = 0.08$) of the two methods used to discriminate substrates and nonsubstrates.

Table 2. Performance Comparison (Percentage of Compounds Predicted Correctly) for QC Descriptors Using Two Pattern Recognition Methods

UGT isoform	SVM,[a] %	cluster−GA−PLSDA,[b] %
1A1	85	68
1A3	89	71
1A4	83	85
1A6	67	71
1A7	79	74
1A8	77	61
1A9	80	64
1A10	80	70
2B4	83	83
2B7	64	71
2B15	67	72
2B17	80	80
average	78	73
median	80	71

[a] Support vector machine. [b] Cluster analysis−genetic algorithm−partial least squares discriminant analysis.

The test set results of the QC models generated by the SVM method are presented in further detail in Table 3. On comparison of the test set prediction accuracies to that expected by chance, the overall, substrate, and nonsubstrate prediction accuracies were all found to be highly statistically significant ($p \ll 0.001$, one-sided paired *t*-test). On average, the substrates in the test sets were predicted marginally better than the nonsubstrates; however, this was not statistically significant when measured with a paired *t*-test ($p = 0.31$). There was no significant (linear) correlation between the number of chemicals in the data set (for each isoform) and the percent of the test set predicted correctly ($r = -0.14$, $p = 0.65$).

The use of three different methods to combine information in the 2D and EEM descriptors is explored in Table 4. As shown in the "Combined" column, regenerating the models using the combination of both 2D and QC descriptors does not improve performance. While the "Maximum" method improved predictivity, the "Consensus" method performed best. In fact, the "Consensus" approach resulted in a statistically significant improvement in predictivity over all other models (i.e., 2D or QC descriptors alone and the "Combined" and "Maximum" methods of combining QC and 2D descriptors). The breakdown of the "Consensus" model test set predictions is elaborated in Table 5.

Table 6 displays the details of the models (>70% of test set predicted correctly) generated with the cluster−GA−PLSDA method using the EEM descriptors. These models contained between one and five EEM descriptors. The right-hand column details the relative contribution of each descriptor in the model. The absolute number indicates the relative size of the contribution,

Table 3. Performance of Models Generated for Each UGT Isoform by Support Vector Machine Using QC Descriptors

UGT isoform	no. of chemicals in data set	% substrates	% of test set predicted correctly		
			all chemicals	substrates	non-substrates
1A1	174	39	85	81	88
1A3	156	76	89	94	67
1A4	156	55	83	78	94
1A6	161	41	67	72	64
1A7	65	40	79	57	92
1A8	104	78	77	95	40
1A9	176	65	80	86	67
1A10	147	50	80	86	74
2B4	131	31	83	75	87
2B7	196	65	64	73	36
2B15	125	42	67	60	71
2B17	53	45	80	70	100
average	137	52	78	77	73
median	152	48	80	77	73

Table 4. Percent of Test Set Predicted Correctly

	2D, %	QC, %	Combined, %	Maximum, %	Consensus, %
1A1	64	85	77	85	88
1A3	84	89	87	89	90
1A4	88	83	85	88	90
1A6	78	67	75	78	83
1A7	74	79	74	79	81
1A8	68	77	81	77	77
1A9	82	80	80	82	87
1A10	80	80	80	80	86
2B4	88	83	86	88	88
2B7	71	64	62	71	73
2B15	64	67	67	67	70
2B17	73	80	80	80	90
average	76	78	78	80	84
median	76	80	80	80	87
p value[a]	0.002	0.001	0.001	0.002	

[a] *p* value was determined by comparing the particular model to the "Consensus" model using a paired *t*-test.

and the sign denotes how the descriptor influences the likelihood of the chemical being a substrate. For example, in the UGT1A3 model, as the maximum separation of charge between nitrogen atoms in the molecule increases, the molecule becomes more likely to be a nonsubstrate. Conversely, as the minimum charge of a hydrogen atom in the molecule increases, the compound is more likely to be a substrate. Furthermore, the maximum nitrogen charge separation has a larger influence than the minimum hydrogen charge.

Discussion

Predictive Capability of QC Descriptors. For 9 of the 12 isoforms modeled, the combination of QC descriptors and SVM was able to produce models with good predictivity (Tables 2 and 3). The average predictive ability across the 12 isoforms was 78%, compared to 76% using SVM with the 2D descriptors[29] and 72% using multiple pharmacophores.[30] This indicates that the descriptors based on the QC properties capture information that is crucial to chemical glucuronidation catalyzed by the various human UGT isoforms.

Cluster−GA−PLSDA Models. QSAR models are generally generated for two reasons: prediction and interpretation. In this paper we primarily describe a system for building predictive models suitable for virtual screening. The cluster−GA−PLSDA methodology was also used to generate predictive models (Table 2). However, distinct from the SVM models, such models are linear and contain only a small number of descriptors. Therefore, the combination of this method with interpretable descriptors should generate models that provide insight into the glucuronidation reaction. The glucuronidation reaction is thought to proceed according to a second-order nucleophilic substitution mechanism.[32] It was originally anticipated that the descriptors containing information on the softness and Fukui function properties would be the most important predictors of glucuronidation because these properties are related to atomic nucleophilicity. However, the dominance of the

Table 5. Test Set Details of Consensus 2D QC Model

UGT isoform	% of test set predicted correctly			no. uncertain	% uncertain
	all chemicals	substrates	non-substrates		
1A1	88	83	91	19	36
1A3	90	97	57	4	9
1A4	90	85	100	6	13
1A6	83	69	95	16	31
1A7	81	60	91	3	16
1A8	77	100	33	5	16
1A9	87	91	79	9	16
1A10	86	89	83	8	18
2B4	88	82	90	2	5
2B7	73	86	42	14	25
2B15	70	70	71	9	25
2B17	90	83	100	5	33
average	84	83	78		
median	87	84	86		

Table 6. Relative Contribution of Descriptors to the Cluster−GA−PLSDA Models

	% of test set predicted correctly			
UGT isoform	all chemicals	substrates	nonsubstrates	relative contribution of descriptor to model
1A3	71	75	56	max nitrogen charge separation (−64) min hydrogen charge (+36)
1A4	85	91	75	max hydrogen charge (−38) min atomic charge (−26) max oxygen charge separation (−22) max nitrogen charge (−8) min oxygen charge (+5)
1A6	71	33	91	molecular softness (−55) max nitrogen charge separation (−45)
1A7	74	71	75	max oxygen charge (−55) molecular softness (−44)
1A10	70	86	57	min hydrogen charge (+100)
2B4	83	92	80	max carbon charge (−39) mean absolute charge (+34) max atomic charge (−16) min nitrogen charge (+11)
2B7	71	76	57	max atomic softness (−44) max nitrogen charge separation (+32) carbon charge sums of squares (−24)
2B15	72	80	67	max oxygen softness (+39) min hydrogen charge (+34) max hydrogen softness (+28)
2B17	80	90	60	atomic charge sums of squares (−57) max oxygen charge (−43)

descriptors based on atomic charge suggests that molecular recognition of the substrate (in addition to chemical reactivity of the substrate) is an important determinant of chemical glucuronidation (Table 6). Thus, the interplay of molecular recognition and chemical reactivity effects obfuscates the unambiguous interpretation of the physicochemical roles of the descriptors. It is likely that the application of local QC descriptors to predict the site of glucuronidation will result in a clearer understanding of the distinct chemical properties influencing substrate binding and chemical reactivity.[24] Some descriptors were found in more than one model, that is, for more than one isoform. In all but one case, the descriptors had a consistent effect (increased likelihood of substrate vs nonsubstrate) in the different models (Table 6). This may indicate the existence of certain chemical features that consistently affect the ability of chemicals to be glucuronidated by multiple UGT isoforms.

Combination of 2D and QC Descriptors. Since the SVM models generated with QC descriptors or 2D descriptors resulted in good predictivity, the combination of both sets of descriptors was attempted. Regenerating the models using the combination of both descriptors did not improve predictivity. This outcome is probably a result of incorporating too many descriptors for the limited data available. The optimal method for combining 2D and QC descriptors in this situation was found to be a consensus approach. According to this approach, if the 2D and QC models do not agree with each other, then the chemical is classified as "uncertain". By use of this approach, many of chemicals previously predicted incorrectly by the 2D and/or QC models were classified as "uncertain", thereby significantly improving the percent of test set chemicals predicted correctly (see Table 4). As shown in Table 5, both substrates and nonsubstrates are predicted significantly better using the consensus method over the 2D and QC models in isolation. The disadvantage of this method is that the glucuronidation of up to 30% of chemicals cannot be predicted (i.e., labeled "uncertain").

One important source of misclassification in the models presented here is the number and distribution of substrates and nonsubstrates in the training and test sets. For an average sized data set (approximately 100 chemicals), there would only be about 30 molecules in the test set. Similar arguments apply for the training set with perhaps only 35 or so substrates to cover the diversity of "substrate space". Clearly, larger data sets will produce better models. The current work suggests that the paucity and noise of the data are probably the most important factors affecting the model accuracy rather than the desciptor efficiency.

It appears that the test sets of UGT2B7 and UGT2B15 are consistently predicted with lower accuracy than the other UGT isoforms (Table 4). UGT2B7 is known for its ability to metabolize chemicals of highly divergent structure, and thus, the inferior test set prediction may be a function of a more complex structure−activity relationship. The reason underlying the inferior performance of UGT2B15 in silico models is unclear.

Comparison with Previous Modeling of UGT Using QC Descriptors. QC properties have been used previously to predict aspects chemical metabolism by UGT. In 1992, the relative conjugation of a congeneric series of 14 substituted benzoic acids with glucuronic acid and taurine in the rat was reported.[2] Of the 39 primarily semiempirical molecular orbital derived QC descriptors calculated for use in the study, the two most important were found to be partial atomic charge meta to the carboxylic acid and electrophilic superdelocalizability at the same position. Two later reports applied the same methodology for classification of the metabolic fate of other chemical classes. Urinary excretion of sulfate and glucuronide conjugates of 16 substituted phenols was investigated in the rat and classified using a number of semiempirical QC descriptors.[33] In addition to classification, linear regression techniques were used

to derive models capable of quantifying the amount of sulfate and glucuronide conjugates excreted in urine. In a similar manner, the urinary excretion of glycine and glucuronide conjugates of 24 substituted benzoic acids was investigated in the rat.[34] The HOMO (highest occupied molecular orbital) energy, log *P* and electrophilic superdelocalizability on the aromatic ring were highlighted as important properties influencing chemical conjugation. The inclusion of a further 22 benzoic acids helped highlight the importance of partial atom charges.[35] The current study extends the earlier work by using the QC properties to predict glucuronidation by individual UGT isoforms. This study is also differentiated by the large and diverse data set of chemicals used to train the models and the advanced pattern recognition methodologies used. Furthermore, the QC properties used here were calculated by a methodology that is more computationally efficient by orders of magnitude compared to the previously employed ab initio or semiempirical methodologies. Thus, the work reported here provides the basis for an efficient high-throughput screening method to predict chemical glucuronidation.

Importance of EEM in Drug Discovery in Silico Screening. To the best of the authors' knowledge, this is the first study reporting the application of EEM-derived DFT chemical properties in a structure−activity relationship. While QC properties have shown great utility in many QSAR studies, their use in the pharmaceutical industry has been limited by the associated computational overhead. With the growing need to generate in silico models capable of predicting very large numbers of structurally diverse chemicals, QC properties have been largely ignored. Because of the recent development of EEM QC properties, the application of quantum chemistry to in silico screening in drug discovery programs is now realistic.[18] Furthermore, this study demonstrates that the EEM calculated QC properties contain useful information for the prediction of biochemical properties such as chemical metabolism.

The main limitation of the EEM method is that only H, C, N, O, and F are currently parametrized. This resulted in the exclusion of chemicals containing other elements from the analyses described here. It is common for new methods requiring parametrization to initially restrict the basis set to cover only the most common cases. Generally, if the method is useful, the range of atom types parametrized increases quickly, and it is expected that this will be the case with the EEM methodology.

QC Properties for the Prediction of Site of Glucuronidation. QC properties have been used previously to predict "site of reactivity" for cytochrome P450 substrates with reasonable success.[3] The prediction of glucuronidation regioselectivity, and hence the structure of the metabolite(s), would be of significant value because of the altered toxicological and pharmacological effects of metabolites and the challenge of characterizing metabolite structure in a high-throughput setting. Furthermore, it is likely that in the absence of regioselectivity data for a large number of chemicals it will prove to be very difficult to combine understanding of structural/steric (e.g., from pharmacophore analyses) and electronic (this study) determinants of chemical binding and metabolic turnover by UGT.

This study aimed to discriminate between substrates and nonsubstrates of each human UGT isoform. Since the molecular descriptors derived from the QC properties were effective in differentiating substrates from nonsubstrates, it is very likely that the QC properties calculated with EEM (charge, softness, and/or Fukui function) are important determinants of glucuronidation likelihood at nucleophilic sites of the molecule. Indeed, the atom-based nature of many QC properties calculated using EEM makes this method well suited for prediction of regioselectivity.

Conclusions

In silico models were built to predict chemical glucuronidation based on three global (equalized electronegativity, molecular hardness, and molecular softness) and three local (atomic charge, Fukui function, and atomic softness) QC properties calculated with EEM. This method allows calculation of these properties at a fraction of the computational expense of other approaches such as ab initio and semiempirical methods. Consequently, QC properties may be used for in silico screening in drug discovery programs. This study is the first reported use of EEM-derived QC descriptors in a structure−activity relationship. The results presented here indicate that EEM QC descriptors can be used to generate highly predictive and computationally efficient in silico models.

The use of descriptors derived from QC properties was explored for the prediction of substrates and nonsubstrates of UGT isoforms. Models built with SVM demonstrated a small improvement in predictivity over previous attempts using 2D chemical descriptors. Models generated using the cluster−GA−PLSDA methodology resulted in partially interpretable models of slightly inferior predictive ability. By combination of the 2D and QC models with a consensus approach, significant improvement in overall, substrate, and nonsubstrate predictivity was possible. This model is capable of predicting substrates of UGT isoforms with approximately 84% success.

Acknowledgment. This work was funded by a grant from the National Health and Medical Research Council of Australia.

References

(1) Tuppurainen, K.; Lotjonen, S.; Laatikainen, R.; Vartiainen, T.; Maran, U.; et al. About the mutagenicity of chlorine-substituted furanones and halopropenals. A QSAR study using molecular orbital indices. *Mutat. Res.* **1991**, *247*, 97−102.

(2) Ghauri, F. Y.; Blackledge, C. A.; Glen, R. C.; Sweatman, B. C.; Lindon, J. C.; et al. Quantitative structure−metabolism relationships for substituted benzoic acids in the rat. Computational chemistry, NMR spectroscopy and pattern recognition studies. *Biochem. Pharmacol.* **1992**, *44*, 1935−1946.

(3) Cnubben, N. H.; Peelen, S.; Borst, J. W.; Vervoort, J.; Veeger, C.; et al. Molecular orbital-based quantitative structure−activity relationship for the cytochrome P450-catalyzed 4-hydroxylation of halogenated anilines. *Chem. Res. Toxicol.* **1994**, *7*, 590−598.

(4) Gaudio, A. C.; Korolkovas, A.; Takahata, Y. Quantitative structure−activity relationships for 1,4-dihydropyridine calcium channel antagonists (nifedipine analogues): A quantum/classical approach. *J. Pharm. Sci.* **1994**, *83*, 1110−1115.

(5) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-chemical descriptors in QSAR/QSPR Studies. *Chem. Rev.* **1996**, *96*, 1027−1043.

(6) Waller, C. L.; Evans, M. V.; McKinney, J. D. Modeling the cytochrome P450-mediated metabolism of chlorinated volatile organic compounds. *Drug Metab. Dispos.* **1996**, *24*, 203−210.

(7) Norinder, U.; Svensson, P. Descriptors for Amino Acids Using Molsurf Parametrization. *J. Comput. Chem.* **1998**, *19*, 51−59.
(8) de Groot, M. J.; Ackland, M. J.; Horne, V. A.; Alex, A. A.; Jones, B. C. A novel approach to predicting P450 mediated drug metabolism. CYP2D6 catalyzed N-dealkylation reactions and qualitative metabolite predictions using a combined protein and pharmacophore model for CYP2D6. *J. Med. Chem.* **1999**, *42*, 4062−4070.
(9) Basak, S. C.; Grunwald, G. D.; Gute, B. D.; Balasubramanian, K.; Opitz, D. Use of statistical and neural net approaches in predicting toxicity of chemicals. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 885−890.
(10) Agatonovic-Kustrin, S.; Beresford, R.; Yusof, A. P. M. Theoretically-derived molecular descriptors important in human intestinal absorption. *J. Pharm. Biomed. Anal.* **2001**, *25*, 227−237.
(11) Girones, X.; Gallegos, A.; Carbo-Dorca, R. Antimalarial activity of synthetic 1,2,4-trioxanes and cyclic peroxy ketals, a quantum similarity study. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 1053−1063.
(12) Gross, K. C.; Seybold, P. G.; Peralta-Inga, Z.; Murray, J. S.; Politzer, P. Comparison of quantum chemical parameters and Hammett constants in correlating p*K*(a) values of substituted anilines. *J. Org. Chem.* **2001**, *66*, 6919−6925.
(13) Girones, X.; Carbo-Dorca, R. Using molecular quantum similarity measures under stochastic transformation to describe physical properties of molecular systems. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 317−325.
(14) Gallegos, A.; Robert, D.; Girones, X.; Carbo-Dorca, R. Structure−toxicity relationships of polycyclic aromatic hydrocarbons using molecular quantum similarity. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 67−80.
(15) Langenaeker, W.; Demel, K.; Geerlings, P. Quantum-chemical study of the Fukui function as a reactivity index: Part 2. Electrophilic substitution on monosubstituted benzenes. *J. Mol. Struct.* **1991**, *234*, 329−342.
(16) Clare, B. W. QSAR of benzene derivatives: comparison of classical descriptors, quantum theoretic parameters and flip regression, exemplified by phenylalkylamine hallucinogens. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 611−633.
(17) Bultinck, P.; Langenaeker, W.; Lahorte, P.; De Proft, F.; Geerlings, P.; et al. The electronegativity equalization method I: Parameterisation and validation for atomic charge calculations. *J. Phys. Chem.* **2002**, *106*, 7887−7894.
(18) Bultinck, P.; Carbo-Dorca, R. Algebraic relationships between conceptual DFT quantities and the electronegativity equalization hardness matrix. *Chem. Phys. Lett.* **2002**, *364*, 357−362.
(19) Chandra, A. K.; Nguyen, M. T. Use of local softness for the interpretation of reaction mechanisms. *Int. J. Mol. Sci.* **2002**, *3*, 310−323.
(20) Chandrakumar, K. R. S.; Pal, S. The concept of density functional theory based descriptors and its relation with the reactivity of molecular systems: A semi-quantitative study. *Int. J. Mol. Sci.* **2002**, *3*, 324−337.
(21) Li, Y.; Evans, J. N. S. The Fukui function: a key concept linking frontier molecular orbital theory and the hard−soft-acid−base principle. *J. Am. Chem. Soc.* **1995**, *117*, 7756−7759.
(22) Geerlings, P.; De Proft, F. Chemical reactivity as described by quantum chemical methods. *Int. J. Mol. Sci.* **2002**, *3*, 276−309.
(23) Miners, J. O.; Mackenzie, P. I. Drug glucuronidation in humans. *Pharmacol. Ther.* **1991**, *51*, 347−369.
(24) Miners, J. O.; Smith, P. A.; Sorich, M. J.; McKinnon, R. A.; Mackenzie, P. I. Predicting human drug glucuronidation parameters: Application of in vitro and in silico modeling approaches. *Annu. Rev. Pharmacol. Toxicol.* **2004**, *44*, 1−25.
(25) Smith, P. A.; Sorich, M. J.; Low, L. S. C.; McKinnon, R. A.; Miners, J. O. Modelling metabolism by UDP-glucuronosyltransferases. *J. Mol. Graphics Modell.* **2004**, *22*, 507−517.
(26) Sorich, M. J.; Smith, P. A.; McKinnon, R. A.; Miners, J. O. Pharmacophore and quantitative structure activity relationship modelling of UDP-glucuronosyltransferase 1A1 (UGT1A1) substrates. *Pharmacogenetics* **2002**, *12*, 635−645.
(27) Smith, P. A.; Sorich, M. J.; McKinnon, R. A.; Miners, J. O. In silico insights: chemical and structural characteristics associated with UDP-glucuronosyltransferase (UGT) substrate selectivity. *Clin. Exp. Pharmacol. Physiol.* **2003**, *30*, 836−840.
(28) Smith, P. A.; Sorich, M. J.; McKinnon, R. A.; Miners, J. O. Pharmacophore and quantitative−structure activity relationship modeling: Complementary approaches for the rationalization and prediction of UDP-glucuronosyltransferase 1A4 substrate selectivity. *J. Med. Chem.* **2003**, *46*, 1617−1626.
(29) Sorich, M. J.; Miners, J. O.; McKinnon, R. A.; Winkler, D.; Burden, F. R.; et al. Comparison of linear and non-linear classification algorithms for the prediction of drug and chemical metabolism by human UDP-glucuronosyltransferase isoforms. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2019−2024.
(30) Sorich, M. J.; Miners, J. O.; McKinnon, R. A.; Smith, P. A. Multiple pharmacophores for the investigation of human UDP-glucuronosyltransferase isoform substrate selectivity. *Mol. Pharmacol.* **2004**, *65*, 301−308.
(31) Greene, J.; Kahn, S.; Savoj, H.; Sprague, P.; Teig, S. Chemical function queries for 3D database search. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1297−1308.
(32) Radominska-Pandya, A.; Czernik, P. J.; Little, J. M.; Battaglia, E.; Mackenzie, P. I. Structural and functional studies of UDP-glucuronosyltransferases. *Drug Metab. Rev.* **1999**, *31*, 817−899.
(33) Homes, E.; Sweatman, B. C.; Bollard, M. E.; Backledge, C. A.; Beddell, C. R. Prediction of urinary sulphate and glucuronide conjugate excretion for substituted phenols in the rat using quantitative structure−metabolism relationships. *Xenobiotica* **1995**, *25*, 1269−1281.
(34) Cupid, B. C.; Beddell, C. R.; Lindon, J. C.; Wilson, I. D.; Nicholson, J. K. Quantitative structure−metabolism relationships for substituted benzoic acids in the rabbit: prediction of urinary excretion of glycine and glucuronide conjugates. *Xenobiotica* **1996**, *26*, 157−176.
(35) Cupid, B. C.; Holmes, E.; Wilson, I. D.; Lindon, J. C.; Nicholson, J. K. Quantitative structure−metabolism relationships (QSMR) using computational chemistry: pattern recognition analysis and statistical prediction of phase II conjugation reactions of substituted benzoic acids in the rat. *Xenobiotica* **1999**, *29*, 27−42.
(36) Bultinck, P.; Langenaeker, W.; Carbo-Dorca, R.; Tollenaere, J. P. Fast calculation of quantum chemical molecular descriptors from the electronegativity equalization method. *J. Chem. Inf. Comput. Sci.* **2002**, *43*, 422−428.
(37) Winkler, D.; Burden, F. R.; Watkins, A. J. R. Atomistic topological indices applied to benzodiazepines using various regression methods. *Quant. Struct.−Act. Relat.* **1998**, *17*, 14−19.
(38) Burden, F. R. A Chemically Intuitive Molecular Index Based on the Eigenvalues of a Modified Adjacency Matrix. *Quant. Struct.−Act. Relat.* **1997**, *16*, 309−314.
(39) Kier, L. B.; Murray, W. J.; Hall, L. H. Molecular connectivity. 4. Relationships to biological activities. *J. Med. Chem.* **1975**, *18*, 1272−1274.
(40) Scholkopf, B.; Smola, A. J.; Williamson, R. C.; Bartlett, P. L. New support vector algorithms. *Neural Comput.* **2000**, *12*, 1207−1245.
(41) Burges, C. J. C. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Min. Knowl. Discovery* **1998**, *2*, 121−167.
(42) Chang, C.-C.; Lin, C.-J. LIBSVM: A library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm (software, 2003).

JM0495529