PCCP

Parameterization of the electronegativity equalization method based on the charge model 1

G. Menegon,*[a] K. Shimizu,[b] J. P. S. Farah,[b] L. G. Dias*[a] and H. Chaimovich[a]

[a] *Department of Biochemistry, Institute of Chemistry, University of São Paulo, São Paulo, SP, Brazil. E-mail: garantes@iq.usp.br and lgdias@iq.usp.br*
[b] *Department of Chemistry, Institute of Chemistry, University of São Paulo, São Paulo, SP, Brazil*

Received 17th July 2002, Accepted 25th October 2002
First published as an Advance Article on the web 6th November 2002

Fast calculation of charge distributions in molecules is feasible in the electronegativity equalization method, EEM. Atomic electronegativities and hardnesses, fundamental parameters in EEM, were obtained here by using CM1 atomic charges at semiempirical PM3 level as targets. A new optimization approach composed of Genetic and Simplex algorithms is also described. The correlation between EEM and CM1 charges improved considerably (correlation coefficient improved from 0.931 to 0.977, standard deviation from 0.079 to 0.032 and Fisher's F from 31 627 to 102 977, for 4093 data points) in comparison to previous EEM parameters (L. G. Dias *et al.*, *Chem. Phys.*, 2002, **282**, 237, ref. 23). Atomic parameters obtained here are discussed and compared to other EEM schemes and to parameters derived from empirical approaches.

I. Introduction

The electronegativity equalization method, EEM, is a semiempirical density-functional theory[1] where atomic electronegativity and hardness are parameterized for the calculation of charge distribution in molecules.[2] The EEM can also be used for calculation of bond charges (located at sites chosen by appropriate criteria)[3] in a refined level of electronic distribution.

Reference targets for the parameterization of EEMs can be obtained in several ways.[2–17] The preferred target forms are: charge models;[2–11] interaction energies of small clusters, when the construction of many-body force fields is investigated;[12–15] or experimental data, such as valence state ionization potentials, electronic affinities or NMR coupling constants.[16,17]

Among the charge models, those derived from electrostatic potential and the CM1 and CM2 models are attractive since they are less sensitive to basis set,[9,10,18,19] reproduce intermolecular electrostatic interactions (essential ingredient for simulations of condensed phase systems) and can be used to estimate the electrostatic contribution to the solvation free energy.[20] The CM1[18] and CM2[19] are defined by a parameterization procedure that takes as input charges from a population analysis of a wave function and maps[18,19] them to reproduce charge-dependent observables obtained from experiment (or from converged quantum mechanical calculations on small molecules).

We have recently described a method, GBEEM, to estimate solvent-induced charge redistribution.[23] The GBEEM combined the EEM with the generalized Born model. The comparison between GBEEM and CM1 indicated GBEEM can generate atomic charges comparable to CM1 but suggested that a new set of EEM atomic parameters would be more efficient.

The EEM has much lower computational costs than the CM1, CM2 and potential-derived charge models. Hence, it has promising applications for large biomolecular systems[5,21,22] and for simulation of condensed phase systems.[12–15,22]

Here we parameterize atomic electronegativities and hardnesses of molecules containing C, H, N and O in different organic functions. The new set improved the relationship between the EEM and CM1 charge models.

II. Electronegativity equalization method

In the EEM, the energy of a molecule, E, is expressed as a function of its N-atomic charges, q_i, atomic electronegativities, χ_i^*, and hardnesses, η_i^*:[2,4]

$$E = \sum_{i=1}^{N}\left(E_i^{\circ} + \chi_i^* q_i + \frac{\eta_i^*}{2} q_i^2\right) + \frac{1}{2}\sum_{i=1}^{N}\sum_{\substack{j=1\\(j\neq i)}}^{N}\eta_{ij}(r_{ij})q_i q_j \quad (1)$$

The electrostatic interaction terms (modified Coulomb interaction terms), $\eta_{ij}(r_{ij})$, take into account the structure of molecules. They are expressed in the Klopman–Ohno–Mataga–Nishimoto approximation:[4,24]

$$\eta_{ij}(r_{ij}) = \frac{1}{\sqrt{r_{ij}^2 + \left(\frac{2}{\eta_i^* + \eta_j^*}\right)^2}} \quad (2)$$

Using the principle of electronegativity equalization:

$$\chi_1\left(\equiv\frac{\partial E}{\partial q_1}\right) = \chi_2\left(\equiv\frac{\partial E}{\partial q_2}\right) = \chi_3\left(\equiv\frac{\partial E}{\partial q_3}\right) = \ldots = \chi_N\left(\equiv\frac{\partial E}{\partial q_N}\right) = \chi \quad (3)$$

and the constraint in the total molecular charge, Q:

$$\sum_{i=1}^{N} q_i = Q \quad (4)$$

a set of N simultaneous linear equations, which solution yields N atomic charges, can be obtained.

This journal is © The Owner Societies 2002

III. Parameterization procedure

Training set

The training set consisted of 250 neutral molecules containing only C, H, O and N in different organic functions: alkane, alkene, alkyne, ether, aldehyde, ketone, ester, nitrile, amine, amide, carboxylic acid and alcohol. No aromatic or conjugated systems were included.

Geometries and target charges

Optimized ground-state geometries and CM1 charges were obtained in the vacuum at the semiempirical PM3 level by the AMSOL program.[25]

Nonlinear optimization

The Genetic[26,27] and Simplex[28] algorithms are used to fit parameters and to explore multi-dimensional parameter surfaces without previous knowledge of the explored functional form and its derivatives. The Genetic algorithm, GA, does not converge as fast as the Simplex algorithm to local stationary points. On the other hand, the Simplex is not as efficient as GA to find global minima or maxima. Here we combined GA[26,27] and Simplex[28] algorithm for optimization.

The evolution calculated by GA typically involved a population of 30 individuals for at least 300 generations. The fitness was evaluated as the root-mean-square deviations between the EEM charges and the target CM1 charges calculated at PM3 level over 4903 atoms of the training set. The deviations were weighted by the inverse of the occurrence of the atom type to assure each element type (C, H, O and N) had the same influence over the total fitness.

An individual is defined as a set of atomic electronegativities and hardnesses, which were allowed to vary respectively by 30% and 50% from the initial parameters.[23] Children were generated by uniform parental crossover (0.5 probability). Mutations both of the normal and creep types were activated (probabilities of 0.08 for both types). The best individuals were chosen by tournament selection. Elitism (replication of the best individual) and niching (sharing) were activated during all generations. Variations of 0.01 units in real numbers were allowed. After the GA run, the final set of 9 (number of parameters + 1) best individuals was placed in the Simplex algorithm taken from Numerical Recipes in Fortran 77.[28] The Simplex was run for 500 iterations using the same limits fixed in the GA. The Simplex iterations stopped when changes in the fitness were smaller than a given threshold, *e.g.* 10^{-5}. Finally, the best-fit individual was placed back into the GA and survived for 50 generations, at least.

IV. Results and discussion

Fig. 1 relates CM1 data with charges calculated by EEM-PD[23] (see Table 1 for acronym definitions) and with the parameters set optimized here, EEM-CM1. Three tendency lines are shown. The dotted line assumes identity between target and calculated charges. The other two lines are regressions between CM1 with EEM-CM1 (solid) and EEM-PD (dashed) charges. It is evident that our parameterization procedure succeeded in reaching an essential identity between CM1 and EEM-CM1 sets (compare solid and dashed line with dotted line). This coincidence is noteworthy since the two charge models (CM1 and EEM) are conceptually different.

The relationship between EEM-CM1 residuals and CM1 charges is presented with different symbols to help in the identification of functional groups in Fig. 2.

Residuals higher than 0.15 (absolute value) represent the following organic functions: esters, aldehydes, nitriles, and alkynes.

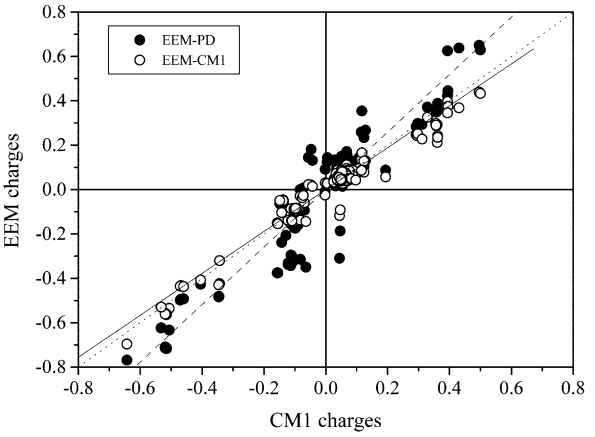

Fig. 1 EEM charges *versus* CM1 charges. The dashed and solid lines are the linear fits to the EEM-PD ($y = ax + b$, $a = 1.30$, $b = -6.1 \times 10^{-7}$, $R = 0.931$, SD = 0.079, $F = 31\,627$, N = 4903) and EEM-CM1 ($y = ax + b$, $a = 0.94$, $b = -5.0 \times 10^{-7}$, $R = 0.977$, SD = 0.032, $F = 102\,977$, $N = 4903$) charges, respectively. The dotted line is the identity line.

The higher residuals in esters correspond to O atoms. In aldehydes, the carbonylic carbons have the higher residuals. In nitriles and alkynes compounds, C_α and C(sp) presented the higher residuals, respectively.

The present EEM-CM1 parameters (Tables 1 and 2) are within atomic electronegativities, χ^*, and hardnesses, η^*, determined previously.[23]

The χ^* show the expected order.[1] In molecules, the $\chi_H{}^*$ was smaller than $\chi_C{}^*$ in all EEM parameterizations (Table 1). The reverse is observed for isolated atoms.[29] This is an effect of covalent bond formation involving hydrogen atoms. It is also necessary to assign different atomic electronegativities when negatively and positively charged hydrogens are present in the molecule[30] (in Tables 1 and 2, only positively charged hydrogen parameters are shown).

The following sequence for the atomic electronegativities was generally seen in the EEMs: $\chi_O{}^* > \chi_N{}^* > \chi_C{}^* > \chi_H{}^*$ (Table 1). The exception was the EEM-M,[2] where $\chi_N{}^* > \chi_O{}^*$.

The η^* sequences are different for each EEM. In EEM-M, $\eta_H{}^* > \eta_N{}^* > \eta_O{}^* > \eta_C{}^*$. In QEq-PD[4] and EEM-PD, $\eta_O{}^* > \eta_H{}^* > \eta_N{}^* > \eta_C{}^*$. In QEq-M,[4] $\eta_O{}^* > \eta_N{}^* > \eta_H{}^* > \eta_C{}^*$. And finally, in EEM-CM1, $\eta_O{}^* > \eta_H{}^* > \eta_C{}^* > \eta_N{}^*$.

Komorowski used refractive indices to obtain local hardnesses.[31] His resulting sequence was: $\eta_H{}^* \cong \eta_N{}^* > \eta_O{}^* > \eta_C{}^*$. *. The atomic hardnesses also depend on the coordination number, for example, $\eta_{C(sp3)}{}^* > \eta_{C(sp2)}{}^* > \eta_{C(sp)}{}^*$.

Table 1 Atomic electronegativities used in different EEM parameterizations

	Electronegativity/kcal mol^{-1} e^{-1}				
Atom	EEM-M[a]	QEq-PD[b]	QEq-M[b]	EEM-PD[c]	EEM-CM1[d]
H	101.67	101.98	89.33	101.98	89.58
C	131.00	116.99	119.38	116.99	107.36
N	244.43	178.42	174.05	178.42	156.45
O	196.02	190.92	205.55	190.92	231.93

[a] EEM-M was parametrized using Mulliken charges at RHF/STO-3G level.[2] [b] QEq-PD and QEq-M are Bakowies and Thiel's EEM that were parametrized using potential-derived and Mulliken charge models at RHF/6-31G* level, respectively.[4] [c] The initial set.[23] The EEM-PD has the same atomic electronegativities as the QEq-PD. [d] This work.

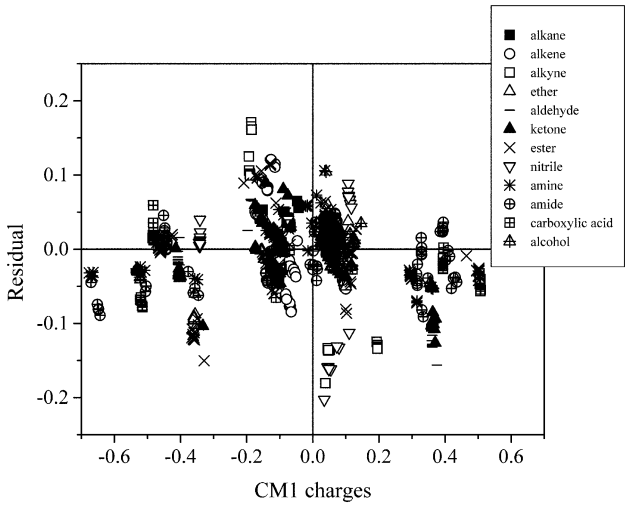

Fig. 2 Residual *versus* CM1 charges. The organic functions are presented with different symbols (see insert).

The EEM-M has atomic hardnesses in a very similar order to Komorowski method. All the others EEMs have the O atoms as the hardest.

Komorowski's method is one of the many different partitioning schemes for the calculation of atomic properties in molecules.[32–34] Other partitioning schemes can be compared to EEMs. For example, Miller's atomic hybrid polarizability method[34] exchanges the order of N and O hardnesses in comparison to Komorowski values. Hence one has to proceed with caution when comparing atomic properties derived from different schemes.

The atomic charges of two relevant biological molecules, a tetra-peptide, Gly-Ala-Ser-Ala, and a saccharide, tri-glucose were calculated to validate the EEM-CM1 in a unbiased test. It should be noted that the training set did not contain these two molecules, or even molecules with more than one organic function: amide, alcohol and carboxylic acid (peptide); alcohol and acetal (saccharide). The two molecular geometries were optimized in vacuum at the PM3 level before charge calculations. Schematic representations of the tetra-peptide and tri-glucose optimized geometries are presented in Fig. 3.

The regression between EEM-CM1 and PM3-CM1 charges is shown in Fig. 4. The two charge models are well correlated. This result demonstrates that the parameters obtained here are useful and transferable to molecules outside the training set. In addition, calculations with molecules containing multiple organic functions and functions not included in the traning set, *e.g.* acetals, are feasible with the EEM-CM1 set.

Table 2 Atomic hardnesses used in different EEM parameterizations

	Hardness/kcal mol^{-1} e^{-2}				
Atom	EEM-M[a]	QEq-PD[b]	QEq-M[b]	EEM-PD[c]	EEM-CM1[d]
H	317.62	319.17	319.17	338.32	424.40
C	208.71	232.09	258.89	246.02	328.93
N	304.09	299.08	325.87	317.02	309.59
O	255.58	344.35	371.15	365.01	477.13

[a] EEM-M was parametrized using Mulliken charges at RHF/STO-3G level.[2] [b] QEq-PD and QEq-M are Bakowies and Thiel's EEM. They were parametrized using potential-derived and Mulliken charge models at RHF/6-31G* level, respectively.[4] [c] The initial set.[23] The EEM-PD atomic hardenesses are 6% greater than the QEq-PD [d] This work.

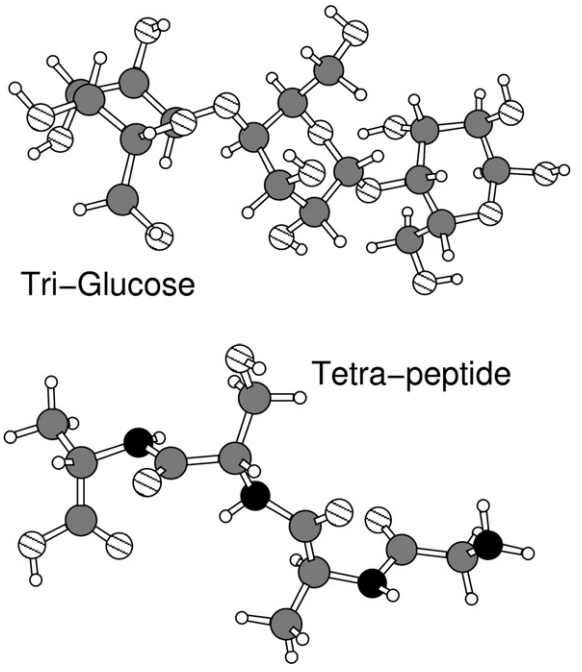

Fig. 3 Structure of the tri-glucose and tetra-peptide (Gly-Ala-Ser-Ala). Small, black, gray and hatched circles are hydrogen, nitrogen, carbon and oxygen atoms, respectively.

The EEM-CM1 calculations for the tetra-peptide were faster than the single-point PM3-CM1 calculations by a factor of a thousand times in a PC Linux Athlon 800 MHz.

V. Conclusions

The EEM-CM1 has an improved general performance for the fast calculation of atomic charges. At present, the extension of the parameter set to aromatic compounds and the inclusion of other organic functions, *e.g.*, compounds containing F, Cl, Br and I, are under development. Additional constraints and an extended set containing specific parameters for the organic functions can further improve the correspondence between EEM and CM1.

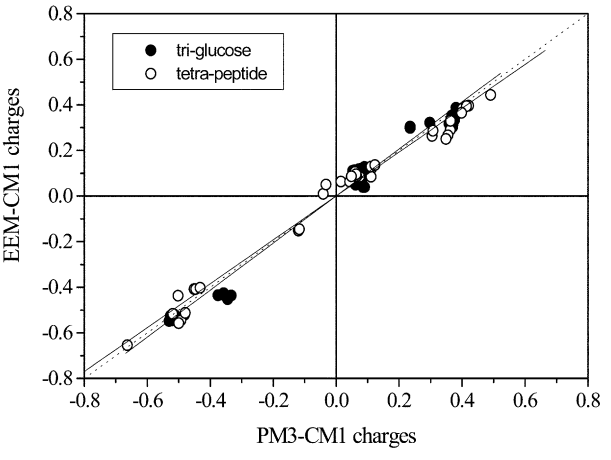

Fig. 4 EEM-CM1 charges for the tri-glucose ($y = ax + b$, $a = 1.03$, $b = 3.3 \times 10^{-17}$, $R = 0.992$, SD = 0.038, $N = 66$) and the tetra-peptide ($y = ax + b$, $a = 0.96$, $b = -2.8 \times 10^{-6}$, $R = 0.991$, SD = 0.042, $N = 41$) compared to the CM1 charges. The solid lines are linear fits and the dotted line is the identity line.

Acknowledgements

This work was supported by the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, 98/10066-7). L. G. Dias is a postdoctoral fellow of FAPESP (99/07688-9). G. Menegon (99/04072-7) and K. Shimizu (01/05852-8) are FAPESP graduate fellows. J.P.S. Farah is grateful to LCCA-CCE-USP (Quantsol Project) for computational facilities.

References

1 R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules*, Oxford University Press, New York, 1989.
2 W. J. Mortier, S. K. Ghosh and S. Shankar, *J. Am. Chem. Soc.*, 1986, **108**, 4315.
3 Z.-Z. Yang and C.-S. Wang, *J. Phys. Chem. A*, 1997, **101**, 6315.
4 D. Bakowies and W. Thiel, *J. Comput. Chem.*, 1996, **17**, 87.
5 K.-H. Cho, Y. K. Kang, K. T. No and H. A. Scheraga, *J. Phys. Chem. B*, 2001, **105**, 3624.
6 R. S. Mulliken, *J. Chem. Phys.*, 1955, **23**, 1833.
7 P.-O. Lowdin, *J. Chem. Phys.*, 1950, **18**, 365.
8 A. E. Reed, R. B. Weinstock and F. Weinhold, *J. Chem. Phys.*, 1985, **83**, 735.
9 L. E. Chirlian and M. M. Francl, *J. Comput. Chem.*, 1987, **8**, 894.
10 C. I. Bayly, P. Cieplak, W. D. Cornell and P. A. Kollman, *J. Phys. Chem.*, 1993, **97**, 10 269.
11 R. F. W. Bader, *Adv. Quantum Chem.*, 1981, **14**, 63.
12 M. C. C. Ribeiro and L. C. J. Almeida, *J. Chem. Phys.*, 2000, **113**, 4722.
13 M. C. C. Ribeiro and L. C. J. Almeida, *J. Chem. Phys.*, 1999, **110**, 11 445.
14 Y.-P. Liu, K. Kim, B. J. Berne, R. A. Friesner and S. W. Rick, *J. Chem. Phys.*, 1998, **108**, 4739.
15 T. Komatsuzaki and I. Ohmine, *Mol. Simul.*, 1996, **16**, 321.
16 W. J. Mortier, K. V. Genechten and J. Gasteiger, *J. Am. Chem. Soc.*, 1985, **107**, 829.
17 F. De Proft, W. Langenaeker and P. Geerlings, *J. Phys. Chem.*, 1993, **97**, 1826.
18 J. W. Storer, D. J. Giesen, C. J. Cramer and D. G. Truhlar, *J. Comput.-Aided Mol. Des.*, 1995, **9**, 87.
19 J. Li, T. Zhu, C. J. Cramer and D. G. Truhlar, *J. Phys. Chem. A*, 1998, **102**, 1820.
20 C. C. Chambers, G. D. Hawkins, C. J. Cramer and D. G. Truhlar, *J. Phys. Chem.*, 1996, **100**, 16 385.
21 Y. Cong and Z.-Z. Yang, *Chem. Phys. Lett.*, 2000, **316**, 324.
22 J. L. Banks, G. A. Kaminski, R. H. Zhou, D. T. Mainz, B. J. Berne and R. A. Friesner, *J. Chem. Phys.*, 1999, **110**, 741.
23 L. G. Dias, K. Shimizu, J. P. S. Farah and H. Chaimovich, *Chem. Phys.*, 2002, **282**, 237.
24 R. F. Nalewajski, J. Korchowiec and Z. Zhou, *Int. J. Quantum Chem. Quantum Chem. Symp.*, 1988, **22**, 349.
25 G. D. Hawkins, D. J. Giesen, G. C. Lynch, C. C. Chambers, I. Rossi, J. W. Storer, J. Li, T. Zhu, P. Winget, D. Rinaldi, D. A. Liotard, C. J. Cramer, D. G. Truhlar, *AMSOL-version6.6*, 1999.
26 D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, Reading, 1989.
27 D. L. Carroll's World Wide Web site: http://cuaerospace.com/carroll/ga.html, accessed in May/2002.
28 W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, *Numerical Recipes in Fortran 77: The Art of Scientific Computing*, 2nd edn., Cambridge University Press, New York, 1992.
29 R. G. Pearson, *Inorg. Chem.*, 1988, **27**, 734.
30 H. Toufar, K. Nulens, G. O. A. Janssens, W. J. Mortier, R. A. Schoonheydt, F. De Proft and P. Geerlings, *J. Phys. Chem.*, 1996, **100**, 15 383.
31 L. Komorowski, *Chem. Phys.*, 1987, **114**, 55.
32 B. Martin, P. Gedeck and T. Clark, *Int. J. Quantum Chem.*, 2000, **77**, 473.
33 K. T. No, K. H. Cho, M. S. Jhon and H. A. Scheraga, *J. Am. Chem. Soc.*, 1993, **115**, 2005.
34 K. J. Miller, *J. Am. Chem. Soc.*, 1990, **112**, 8533.