Enhanced 3D-Databases: A Fully Electrostatic Database of AM1-Optimized Structures

Bernd Beck,[†] Anselm Horn,[†] John E. Carpenter,[§] and Timothy Clark*[,†]

Computer-Chemie-Centrum, Institut für Organische Chemie, Friedrich-Alexander-Universität Erlangen-Nürnberg, Nägelsbachstrasse 25, D-91052 Erlangen, Germany, and Cray Research Center, 665 Lone Oak Drive, Eagan, Minnesota, 56121

Received July 24, 1998

In a feasibility study, a single conformation 3D version of the Maybridge database (53 471 compounds) has been produced using geometries optimized with AM1 semiempirical MO-theory. The database entries include full electrostatic information within the NAO-PC model and can be used to generate spectroscopic and physical properties using established QSPR models. The data were generated from the original database using custom cleanup software to remove database inconsistencies and, for instance, to isolate the "interesting" ion of ion pairs, 2D to 3D conversion using CORINA and subsequent geometry optimization using VAMP. The complete geometry optimization run was carried out in less than 15 h elapsed time on a Silicon Graphics Origin 2000 with 126 processors. The total failure rate for the structure cleanup, 2D to 3D conversion, and geometry optimization steps was around 1%.

Recent years have seen a resurgence of interest in 3D QSAR[1−3] and QSPR[4] techniques for activity and property prediction, although the problem of multiple conformations remains to be solved. Nevertheless, it has become evident that conventional 2D-database technology is no longer adequate for routine screening of large amounts of data using 3D techniques. Furthermore, conventional atom-monopole based electrostatics are increasingly being recognized to be less than adequate for many QSAR, QSPR, and docking applications,[5] so that there is clearly a need for a modern, compact but accurate electrostatic description of molecules that can be used to store a detailed description of the molecular electrostatics in a database. Quantum mechanical techniques are one solution to this problem, although they are traditionally thought of as being too CPU-intensive for application to complete databases. In this study, we report on the feasibility of using semiempirical MO-theory to provide optimized 3D structures and detailed accurate electrostatics for a complete moderate sized database.

PREPROCESSING THE DATABASE

For this test we used the entire Maybridge database[6] of 53 471 molecules, which was provided by the Maybridge Chemical Company. In order to eliminate data inconsistencies, such as several different types of entry for ion pairs, the entire database was preprocessed with TARZAN,[7] a program system written especially for this project. The individual consistency checks and corrections performed by TARZAN are listed in Table 1. This preprocessing yielded a total set of 53 256 molecules; only 215 compounds were not treated correctly. One hundred seventy-eight of these could not be protonated automatically (the structures were stored as implicit ion pairs), four files contained more than three molecules, in eight files an unknown (to TARZAN) counterion was found, the wrong charges were assigned to five compounds, and finally in 17 2D-entries unresolvable inconsistencies were found.

2D TO 3D CONVERSION

The preprocessed 2D SD-files were passed to CORINA[8,9] for conversion to 3D coordinates, which were used as the starting geometries for the semiempirical geometry optimizations. The 53 256 surviving compounds from the TARZAN preprocessor required 6 h 43 min CPU time on a 90 MHz Silicon Graphics R8000 Power Challenge (one processor) with only 41 2D to 3D conversion failures. The CORINA runs used the −*drs* option in order to eliminate counterions (but see tosylates in Table 1). Since this run, we have extended TARZAN to make it capable of performing this part of the preprocessing. The processed structures were tested additionally for unpaired electrons, which would indicate a processing error.

SEMIEMPIRICAL OPTIMIZATIONS

All geometry optimizations were performed with the standard AM1 Hamiltonian.[10] A prerelease version of VAMP 7.0[11] was used in a single processor version on a 126-processor 195 MHz R10000 Silicon Graphics Origin 2000 in Eagan, MN. The VAMP standard Eigenvector Following (EF)[12] optimizer was chosen throughout in a modified form that uses a force field initial guess for the Hessian matrix in order to increase optimization speed significantly.[13] Furthermore, we used Cartesian, rather than internal, coordinates. Our version of the EF optimizer is generally a little slower using Cartesian coordinates than with a well chosen set of internal coordinates, but these optimizer options were chosen for reliability, rather than speed. Large geometry changes during the optimization process can lead to undefined dihedral angles when using internal coordinates. Although this problem is generally fixed at runtime by

[†] Friedrich-Alexander-Universität Erlangen-Nürnberg.
[§] Cray Research Centre.

10.1021/ci9801318 CCC: $15.00
Published on Web 10/13/1998

Table 1. Consistency Checks and Corrections Performed by TARZAN

no. of molecules found in the sd-files	checks and corrections
$N = 1$	check for covalently bonded counterion(s), e.g. RNa or PR4X; if present remove the bond
	(NEW: remove the counter ion(s)) an set the correct charge
	check elements against parametrization of the chosen Hamiltonian
	determine charge of the molecule
	create VAMP input file
$N = 2,3$	determine main molecule (largest molecule in the file)
	Check for covalently bonded counterion(s), e.g. RNa or PR4X; if present remove the bond and set the correct charge
	check the main molecule for tosylate or any other defined counterion, if so replace main molecule and start check again
	check elements in the main molecule against parametrization of the chosen Hamiltonian
	determine charge of the main molecule
	charge consistency check between the main molecule and the found counterion(s)
	new: remove counterion(s)
	create VAMP input file
	optional: try to protonate or deprotonate main molecule if the counterion is protonated/deprotonated
$N > 3$	write out an error message to the .log file
files produced by TARZAN:	
	sd-file with the "cleaned up" 2D compound
	log-file containing control, warning or error messages
	dat-file containing the VAMP keyword line

VAMP, we chose the more conservative coordinate system in order to eliminate as many potential problems as possible. Using Cartesian coordinates also eliminated the need to convert the CORINA output coordinates into Z-matrices. Standard CORINA output SD-files were used throughout as input. Apart from possible problems with the coordinate system, a second potential source of failure for a semi-empirical MO-optimization is nonconvergence of the self-consistent field (SCF) calculation at some stage in the optimization. In order to avoid this problem as far as possible, VAMP was set up to use the IIS converger of Badziag and Solms[14] throughout. Although IIS is often slower to converge than Pulay's scheme[15] or DIIS,[16] we have found it to be extremely reliable. In fact, SCF convergence was not a problem during the conversion of the database. Nevertheless a maximum of 2000 iterations within one SCF cycle was chosen. Additionally a time limit of 300 min was imposed for optimizations in order to avoid extremely long runs for molecules with very flat potential surfaces.

Calculations were carried out one molecule per processor in two parallel jobs using 63 processors each and in batches of 5000 compounds under the Irix 6.4 operating system. The runtime option *dplace −migration 0 −propagate* was used to turn off the dynamic scheduling normally used by the system, so that jobs ran locally on their own processors. As the memory in VAMP 7.0 is completely dynamically allocated, excessive swapping or memory bottlenecks were not experienced. Visual observation of the system performance suggested that all processors were running at or above 99% utilization during the run.

The database was processed within a 16 h exclusive block access to the machine on October 18th 1997. After initial installation and tuning of the run conditions, the first two batches of 5000 molecules were started at 07.00 a.m. Central European Time. These 10 000 compounds were allowed to optimize, resulting in some "tail" time in which molecules that ran for significantly longer than the average were processed. After all but a few of these jobs had finished, the next batch was started. The 53 215 compounds that survived the TARZAN and CORINA stages were subjected to this procedure, and the last individual molecule finished at 21:15 p.m. Central European Time.

The VAMP runs produced a total of 15 program-related failures or 0.03% of the total number of molecules. An additional 37 optimizations failed because of an error in a development version of a system library. These errors could not be reproduced on our local machine. Several runs exceeded the preset CPU-time limit either because of the size of the molecules involved or because there were a large number of soft internal rotations, which generally cause problems with optimizations in Cartesian coordinates. Seventeen arsenic containing compounds could not be calculated by VAMP, because As is not parametrized in AM1. Two hundred fifty-one molecules "dissociated" on optimization. This means that one or more bonds in the compounds lengthened or were broken (bond order below 0.5). Most of these structures contain one or more $-SO_3$, $-SO_2NR$, or $-N_3$ groups. The 52 932 successful optimizations required an average CPU (user + system) time of 96.8 s per molecule, which combined with the total elapsed time of 14 h 15 min for 126 processors gives a machine usage to produce useful results of 79%. Note, however, that the procedure of running the database in a series of batches of compounds decreases the machine usage to well below the maximum possible.

The procedure used to proceed from 2D to 3D optimized structures is summarized in Scheme 1. The resulting VAMP sd-files require 451.2 MByte disk space and include the optimized 3D structure, information about the calculation (number of optimization cycles, Hamiltonian), point group, energies, detailed electrostatics, polarizability tensor, ionization potential, electron affinity, and the bond orders. Thus, the compactness of the electrostatic representation used makes this type of detailed database feasible.

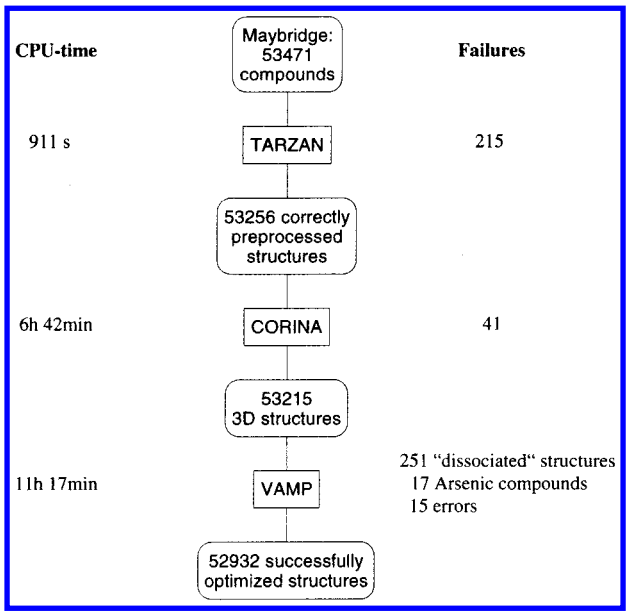
PROPERTIES CALCULATIONS

The optimized structures were used as input for PROPHET,[17] in order to calculate the molecular and atomic properties listed in Table 2. PROPHET requires as input the atom types, the coordinates, the bond order matrix, and the stored electrostatic information of the compound and therefore relies on the results of prior quantum mechanical calculations. In this work, we have used AM1, but DFT or *ab initio* methods could also be used. The detailed informa-

Table 2. Molecular and Atomic Properties Calculated by the Current Version of PROPHET

atomic and molecular properties	(1) VESPA atomic charges[18,19]
	(2) molecular multipoles up to the octupole
	(3) atomic multipoles up to the quadrupole
	(4) molecular SES-surface,[20−22] volume, and globularity[23]
	(5) ESP-histogram describing the distribution of the electrostatic potential over the surface points
	(6) Politzer parameters:[24]
	−number of surface points with positive/negative potential
	−ESPmin, ESPmax
	−mean positive/negative potential
	−positive/negative variance
	−total variance
	−balance parameter
	−product of total variance and balance
	(7) pharmaceutically relevant local dipoles:[25] hydrogen bond donor/acceptor and aromatic vectors (normal to the plane); three different options to calculate the acceptor vectors:
	(a) use all NAO-PCs[26,27]
	(b) use only the NAO-PCs describing the lone pairs
	(c) "rabbit ears"
QSPR models included	(1) log *P* prediction[28]
	(2) ^{13}C estimation[29]

Scheme 1



tion stored in the sd files is adequate for all QSPR-models developed for PROPHET thus far.

The compact stored electrostatic information of the molecule is used to set up the NAO-PCs (natural atomic orbital-point charges)[26,27] in order to describe the electron distribution around the heavy atoms. These NAO-PCs are then used to calculate the electrostatic potential of the molecule and the atomic and molecular multipole moments. The architecture of this program allows the inclusion of new properties or updates very easily so that in a following run only those properties that are new or modified will be calculated.

For the 52 932 compounds that were successfully optimized by VAMP, the properties listed in Table 2 were calculated within 9 h (on 53 processors of a 64-processor 195 MHz R10000 Silicon Graphics Origin 2000 in Eagan, Minnesota), whereby the most time-consuming step is the calculation of the VESPA charges.[18,19] Considering the fact that during this first test 1000 compounds were submitted to one processor, we expect that by using optimized control procedures the properties could have been calculated in less than 8 h. The properties calculations for all 52 932 structures occurred without error.

The calculated properties were appended to the VAMP sd files, which then require 805.3 MByte storage space.

SUMMARY AND OUTLOOK

This initial feasibility study has demonstrated that semiempirical MO-techniques can be applied to databases of the order of hundreds of thousands of compounds and that a very detailed electrostatic description of the molecules can be stored well within the limits of currently available disk space. The software involved in the complete raw database to optimized structure conversion process has been shown to be reliable enough for such applications. The data stored is sufficient for both the application and the development of both QSPR and QSAR models. We are now developing improved QSPR and QSAR models to take advantage of the information contained in such databases.

The major remaining problem is that of treating multiple conformations for flexible molecules. At present the most promising approach seems to be to generate a limited library of the likely most stable conformations for each flexible molecule using programs such as CORINA or COBRA[30] that can generate a list of "reasonable" conformations. A more difficult problem is using these conformational libraries for QSAR models without providing too many degrees of freedom in the model. At present, most of the existing experimental data does not seem to justify multiconformational QSPR models.

In this work we have optimized the molecular structures with AM1 in order to investigate the feasibility of such a step. This may not be necessary, and databases could be constructed either on the basis of the CORINA structures themselves or with structures optimized by molecular mechanics. Our present QSPR models, which were trained using AM1-optimized structures, perform significantly worse if the alternative geometries are used. This would, however, not necessarily be true if the models were trained for other geometries. However, we emphasize here that the quantum mechanical optimization step does not represent a significant

hurdle in terms of CPU-time on modern machines and that future semiempirical methods will give significantly improved accuracy. Thus, the consequent use of a quantum mechanical method both for the structure and the electrostatic properties of the molecules seems justified.

ACKNOWLEDGMENT

This project was supported by Oxford Molecular Ltd. We thank Johnny Gasteiger and Jens Sadowski for support for CORINA and the Cray Research Center in Eagan, MN for providing us CPU-time on their 126 processor Origin 2000.

REFERENCES AND NOTES

(1) Böhm, H.-J.; Klebe, G.; Kubinyi, H. In *Wirkstoffdesign: Der Weg zum Arzneimittel*; Spektrum Akademischer Verlag: Heidelberg, 1996; Chapter IV, pp 361−466.
(2) Greco, G.; Novellino, E.; Martin, Y. C. 3D-QSAR Methods. In *Designing Bioactive Molecules: Three-Dimensional Techniques and Applications;* Martin, Y. C.; Willet, P., Eds.; American Chemical Society: Washington, DC, 1998; Chapter 10, pp 219−254.
(3) *Classical and Three-Dimensional QSAR in Agrochemistry;* Hansch, C., Toshio F., Eds.; American Chemical Society: Washington, DC, 1995.
(4) Hansch, C.; Leo, A. In *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.
(5) Dixon, P. M.; Blaney, J. M. Docking: Predicting the Structure and Binding Affinity of Ligand-Receptor Complexes. In *Designing Bioactive Molecules: Three-Dimensional Techniques and Applications;* Martin, Y. C., Willet, P., Eds.: American Chemical Society: Washington, DC, 1998; Chapter 8, pp 175−198.
(6) Maybridge Chemicals Company Ltd.: Trevillet, Tintagel, Cornwall PL34 OHW, England.
(7) Beck, B.; Hennemann, M. TARZAN, Erlangen, 1998; unpublished results.
(8) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic Generation of 3D-Atomic Coordinates for Organic Molecules. *Tetrahedron Comp. Method*. **1990**, *3*, 537−547.
(9) Sadowski, J.; Gasteiger, J. Corina v. 1.8; Oxford Molecular: Medawar Centre, Oxford Science Park, Oxford, OX4 4GA, England.
(10) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model *J. Am. Chem. Soc.* **1985,** *107,* 3902−3909.
(11) Clark, T.; Alex, A.; Beck, B.; Chandrasekhar, J.; Gedeck, P.; Horn, A.; Hutter, M.; Rauhut, G.; Sauer, W.; Steinke, T. Vamp 7.0; Oxford Molecular Ltd.: to be released in August 1998.
(12) Baker, J. An Algorithm for the Location of Transition States *J. Comput. Chem.* **1985**, 7, 385−395.
(13) Horn, A. New Semiempirical MO-Methods: A Combined QM/MM Ansatz for Geometry-Optimization. Diploma, Erlangen, 1997.
(14) Badziag, P.; Solms, F. An improved SCF iteration scheme. *Computers Chem.* **1988**, *12*, 233−236.
(15) Pulay, P. Improved SCF Convergence Acceleration. *J. Comput. Chem.* **1982**, 3, 556−560.
(16) Csaszar, P.; Pulay, P. Geometry Optimization by Direct Inversion in the Iterative Subspace. *J. Mol. Struct. (Theochem)* **1984**, *114*, 31−34.
(17) Beck, B.; Clark T. PROPHET, Erlangen, 1998; unpublished results.
(18) Beck, B.; Clark, T.; Glen, R. C. A Detailed Study of VESPA Electrostatic Potential-Derived Atomic Charges. *J. Mol. Model. (electronic edition)* **1995**, *1*, 176−187.
(19) Beck, B.; Clark. T.; Glen, R. C. VESPA: A New, Fast Approach to Electrostatic Potential-Derived Atomic Charges from Semiempirical Methods. *J. Comput. Chem.* **1997**, *18*, 744−756.
(20) Heiden. W.; Goetze, T.; Brickmann, J. Fast Generation of Molecular Surfaces from 3D Data Fields with an Enhanced "Marching Cube" Algorithm. *J. Comput. Chem.* **1993**, *14*, 246−250.
(21) Marsili, M. In *Physical Property Prediction in Organic Chemistry*; Jochum, C., Hicks, M. G., Sunkel, J., Eds., Springer: Berlin-Heidelberg, 1988; p 249.
(22) Pascual-Ahuir, J. L.; Silla, E.; Tuñon, I. GEPOL: An Improved Description of Molecular Surfaces. III. A New Algorithm for the Computation of a Solvent-Excluding Surface. *J. Comput.Chem.* **1994**, *15*, 1127−1138.
(23) Meyer, A. Y. The Size of Molecules. *Chem. Soc. Rev.* **1986**, *15*, 449−475.
(24) Murray, J. S.; Lane, P.; Brinck, T.; Paulsen, K.; Grice, M. F.; Politzer, P. Relationships of Critical Constants and Boiling Points to Computed Molecular Surface Properties. *J. Phys. Chem.* **1993**, *97*, 9369−9373.
(25) Clark, T.; Henneman, M.; Beck, B. HOBO: An Alignment-Free 3D-QSAR Technique Based on Hydrogen-Bond Dipoles. Paper in preparation.
(26) Rauhut, G.; Clark, T. Multicenter Point Charge Model for High-Quality Molecular Electrostatic Potentials from AM1 Calculation. *J. Comput. Chem.* **1993**, *14*, 503−509.
(27) Beck, B.; Rauhut, G.; Clark, T. The Natural Atomic Orbital Point Charge Model for PM3: Multipole Moments and Molecular Electrostatic Potentials. *J. Comput. Chem.* **1994**, *15*, 1064−1073.
(28) Breindl, A.; Beck, B.; Clark, R.; Glen, R. C. Prediction of the n-Octanol/Water Partition Coefficient, logP, Using a Combination of Semiempirical MO-Calculations and a Neural Network. *J. Mol. Model.* (*electronic edition*) **1997**, *3*, 142−155.
(29) Clark, T.; Breindl, A.; Rauhut, G. A Combined Semiempirical MO/Neural Net Technique for Estimating ^{13}C Chemical Shifts. *J. Mol. Model.* (*electronic edition*) **1995**, *1*, 22−35.
(30) Cobra v. 3.21; Oxford Molecular: Medawar Centre, Oxford Science Park, Oxford, OX4 4GA, England.

CI9801318