COMPUTER SOFTWARE REVIEWS

CODESSA Version 2.13 for Windows

Ovidiu Ivanciuc†

Department of Organic Chemistry, Faculty of Chemistry, Polytechnic University of Bucharest, Splaiul Independentei 313, 77206 Bucharest, Romania

Received July 24, 1996

INTRODUCTION

CODESSA (Comprehensive Descriptors for Structural and Statistical Analysis) version 2.13 for Windows is a new software product destinated for the computation of a wide range of molecular descriptors and for their use in QSPR (Quantitative Structure−Property Relationships) and QSAR (Quantitative Structure−Activity Relationships) studies. CODESSA was developed under the direction of Professor A. R. Katritzky (University of Florida) and Professor M. Karelson (University of Tartu) and is distributed by Semichem.

Semichem was founded in 1992 with the goal to sustain and promote the quantum chemistry models developed by Professor Michael J. S. Dewar and incorporated in AMPAC. CODESSA uses the quantum chemical results provided by the AMPAC output files and computes more than 450 molecular descriptors. The molecular descriptors can be used to develop QSPR/QSAR models by the use of various statistical techniques: MLR (Multi-Linear Regression), PCA (Principal Component Analysis), and PLS (Partial Least-Squares).

For those interested in establishing quantitative structure−property models, CODESSA is a flexible software tool which generates a large set of theoretical descriptors, allows an easy manipulation of the patterns, and uses sound statistical methods to obtain QSAR models.

SYSTEM REQUIREMENTS, INSTALLATION, AND DOCUMENTATION

The system requirements of CODESSA are the following: an IBM PC 386, 486, Pentium or compatible, 4 MB of RAM, 4 MB of free hard disk space, a VGA or SuperVGA video display, a mouse, and Microsoft Windows 3.1 or higher version. The evaluation of CODESSA presented in this review is based on installation and testing on an IBM PC 486 DX2 computer at 66 MHz with 16 MB of RAM.

The installation of CODESSA is easy to perform and follows the usual procedure of Windows-based programs. The printed documentation of CODESSA consists of three manuals: a Reference Manual which is an overview of the computation of the theoretical descriptors and of the statistical methods used in CODESSA; a User Manual which presents the menus in CODESSA, with a short description and pictures of the windows; a Training Manual, which gives details on the use of CODESSA. The Training Manual would be more useful if it is used to learn CODESSA with some examples provided on the distribution disk.

† E-mail : o_ivanciuc@chim.upb.ro.

THEORETICAL DESCRIPTORS OF THE MOLECULAR STRUCTURE

CODESSA uses two kinds of molecular structure files: files containing the constitutional (topological) description of the molecules and files containing the results of a quantum chemical computation of the molecules. The constitutional description can be supplied in several standard formats: the MDL molfile format, the HyperChem HIN format, and the SYBYL MOL format. The quantum chemical results can be taken from AMPAC or MOPAC output files. For the evaluation of CODESSA I have used HyperChem HIN files and AMPAC 5/MOPAC 6 output files.

The theoretical descriptors of the molecular structure computed in CODESSA form six groups: constitutional, topological, geometrical, electrostatic, quantum-chemical, and thermodynamic.

The constitutional descriptors reflect the number and types of atoms, bonds, and rings in the molecule. They are simple descriptors, but with a long tradition in additive QSPR models.

The topological descriptors set contains a selection of the most used indices: the Wiener index W, the Randić, Kier, and Hall connectivity indices, the Kier shape and flexibility indices, and the information content indices.

Using the three-dimensional molecular structure, CODESSA computes a number of geometrical descriptors: moments of inertia, shadow indices, molecular volume and surface area, and gravitation indices.

The empirical partial charges, computed on the basis of atomic electronegativities, generate the set of electrostatic descriptors: minimum and maximum partial charge, topological electronic index, and charged partial surface area indices.

The most important and large set of theoretical descriptors computed in CODESSA is the set of quantum-chemical descriptors: charge distribution indices, valency indices, and energy-related descriptors.

A set of thermodynamic descriptors is computed on the basis of a second AMPAC/MOPAC file obtained with a different set of keywords.

STATISTICAL MODELS

The preliminary statistical analysis of the descriptors or molecular properties offers important information for the selection of the potential useful descriptors. It consists in the computation of a wide range of one- and two-dimensional statistical indices: mean value, dispersion, standard deviation, and intercorrelation coefficients.

Various regression models are available in CODESSA to find the best QSAR model of a property. Linear and

multilinear regression models can be used with small sets of descriptors or when the user knows the structural parameters that determine the investigated property. One of the strong points of CODESSA is the large number of theoretical descriptors that can be computed, more than 450. With such a large set of descriptors, it is almost impossible for the user to select by hand the best QSAR parameters. CODESSA offers two methods to develop QSAR equations with a high predictive power: the best multilinear regression and the heuristic methods. Both methods use some algorithms to develop optimum regression models. I have tested both algorithms with a few QSAR models and both proved to be efficient in selecting good equations. CODESSA offers also principal component and Nonlinear Iterative Partial Least Squares (NIPALS) regressions.

Three multivariate techniques are implemented in CODESSA: Principal Component Analysis (PCA), NIPALS, and target transformation PCA. CODESSA is very flexible in operation: the user can introduce descriptors computed with other software; new descriptors can be added by transforming the older ones with various mathematical functions; cross-terms of selected descriptors can be added to the model; the selection of the descriptors and QSPR models can be performed by the user or by statistical algorithms; the statistical algorithms have numerous parameters which can be set by the user; the results can be plotted, presented in numerical form, or saved for future reference; the chemical structures can be plotted together with some atomic descriptors; and after a model is obtained, CODESSA can offer predictions of the investigated property for new chemical structures.

SOFTWARE DISTRIBUTION

CODESSA Version 2.13 for Windows is distributed by Semichem, 7128 Summit, Shawnee, KS 66216; Tel. (913)-268−3271; Fax: (913)-268−3445; E-mail: aholder@cctr.umkc.edu. The price of CODESSA is $495 for academic users and $2995 for commercial users. CODESSA is available for a large range of workstations and mainframes.

CONCLUSIONS

CODESSA is remarkably easy to use, and very quickly the user can investigate structure−property relations in a wide range of applications, from physical organic chemistry to biochemistry and drug design.

I have tested CODESSA for several small QSAR problems, obtaining good results. As input files I have used HyperChem HIN files and the output files of AMPAC 5 and MOPAC 6. More than 450 descriptors were generated in each case, and by a combined use of the heuristic algorithm and the best multilinear regression their number was reduced to a few significant descriptors. The models were generated in a short time and showed good statistical quality and predictive power.

The main advantage of CODESSA over other statistical packages is the easy generation of more than 450 theoretical molecular descriptors using the widely known AMPAC and MOPAC programs. This is also a possible source of chance effects, if the final model includes descriptors which are not connected with the investigated property, but by chance their values correlate with the values of the property for a particular set of molecules. This danger can be avoided by a careful design of the QSAR/QSPR experiments, by using different heuristics for descriptors selection, by dividing the molecules in a training set (used to develop the QSAR/QSPR models) and a test set (used to estimate the predictive power), and by using cross-validation techniques, like leave-N-patterns-out. A future version of CODESSA shoud consider in greater detail the techniques of avoiding the chance effect.

I recommend CODESSA for all those interested in developing quantitative models between chemical structure and physicochemical and biological properties. It can be of great help in developing QSPR/QSAR models in physical chemistry, organic chemistry, biochemistry, drug design, toxicology, and chemical engineering.[1−3]

REFERENCES AND NOTES

(1) Katritzky, A. R.; Ignatchenko, E. S.; Barcock, R. A.; Lobanov, V. S.; Karelson, M. Prediction of Gas Chromatographic Retention Times and Response Factors Using a General Quantitative Structure−Property Relationship Treatment. *Anal. Chem.* **1994**, *66*, 1799−1807.

(2) Murugan, R.; Grendze, M. P.; Toomey, J. E., Jr.; Katritzky, A. R.; Karelson, M.; Lobanov, V.; Rachwal, P. Predicting Physical Properties from Molecular Structure. *CHEMTECH* **1994**, *24*, 17−23.

(3) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. QSPR: The Correlation and Quantitative Prediction of Chemical and Physical Properties from Structure. *Chem. Soc. Rev.* **1995**, 279−287.

CI950193N