MATCH
Communications in Mathematical
and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **56** (2006) 237-248

ISSN 0340 - 6253

DRAGON SOFTWARE: AN EASY APPROACH TO MOLECULAR DESCRIPTOR CALCULATIONS

Andrea Mauri[*]*, Viviana Consonni, Manuela Pavan, and Roberto Todeschini*

Milano Chemometrics and QSAR Research Group, Dept. of Environmental Sciences,
University of Milano-Bicocca, P.za della Scienza 1 – 20126 Milano (Italy)
Web site: www.disat.unimib.it/chm/

(Received April 12, 2006)

Abstract

Due to the relevance that molecular descriptors are constantly gaining in several scientific fields, software for the calculation of molecular descriptors have become very important tools for the scientists. In this paper, the main characteristics of DRAGON software for the calculation of molecular descriptors are shortly illustrated.

1. Introduction

Molecular descriptors play a fundamental role in chemistry, pharmaceutical sciences, environmental protection policy, health research and quality control, being used to predict biological and physico-chemical properties of molecules (QSAR/QSPR) and for virtual screening of molecule libraries. They are obtained when molecules, thought as real objects, are transformed into a molecular representation enabling mathematical treatments. In the last few years, a lot of molecular descriptors have been proposed derived from different theories and approaches.

"The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment" [1]. Obviously, it follows that the information content of a molecular descriptor depends on the kind of molecular representation and algorithm used for its calculation.

There are simple molecular descriptors derived by counting some atom-types or structural fragments in the molecule, other derived from algorithms applied to a topological representation (molecular graph) and usually called topological or 2D-descriptors, and there are molecular descriptors derived from a geometrical representation called geometrical or 3D-descriptors.

All the molecular descriptors should contain chemical information, should satisfy some basic invariance properties and general requirements, and should be derived from well-established procedures which enable molecular descriptors to be calculated for any set of molecules. It is obvious – almost trivial - that a single descriptor or a small number of descriptors cannot wholly represent the molecular complexity or model all the physico-chemical properties and biological interactions. As a consequence, although we are getting used to deal with approximate models (nothing is perfect!), we should be aware that "approximate" is not a synonym of "useless".

The field of molecular descriptors is strongly interdisciplinary and involves a number of different theories. For the definition of molecular descriptors, knowledge of algebra, graph theory, information theory, computational chemistry, theories of organic reactivity and physical chemistry is usually required, although at different levels. For the use of the molecular descriptors, knowledge of statistics, chemometrics, and the principles of the QSAR/QSPR approaches is necessary in addition to the specific knowledge of the problem.

DRAGON software has been conceived to provide the user with a variety of molecular descriptors derived from different molecular representations, allowing the user to choose those molecular descriptors which are more suitable for his/her specific research.

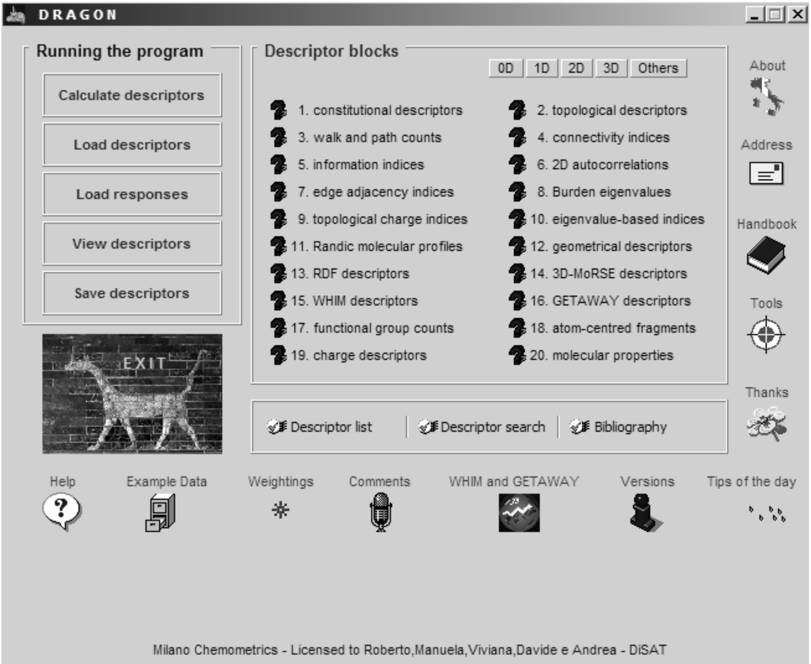
The first release of DRAGON was developed in 1994 by Milano Chemometrics and QSAR Research Group with the name "WHIM/3D QSAR", being specific for the calculation of the WHIM descriptors [2]. Successively, a lot of other descriptors have been implemented leading to a new software, which in 1997 provided about 600 descriptors and was released with the name DRAGON. Since 1997 DRAGON has been regularly updated by inclusions of new molecular descriptors in order to advance research in QSAR and new algorithms for optimising precision and time performances as well as its capability to read different molecular file formats. Actually, DRAGON [3] allows the calculation of 1,664 molecular descriptors and it is designed to work both for Windows and Linux systems. There are two versions for Windows, DRAGON professional, which can only work in stand-alone mode and DRAGON plus, which can work both in stand-alone and background mode. For Linux there only is one version, called dragonX, which only works in background mode by a command line.

DRAGON was not designed as QSAR software; it provides only molecular descriptors and does not perform QSAR analysis nor geometry optimization. However, by DRAGON it is possible to merge calculated molecular descriptors and user-defined properties for a set of molecules, providing a complete output file which is easily loaded by any correlation analysis application. Moreover, a menu has been included in DRAGON which allows the calculation of pair correlation between molecular descriptors and experimental properties and the graphical analysis of the distribution of molecules in the descriptor and response space.

2. DRAGON structure

DRAGON was designed as a user-friendly software which performs descriptor calculations according to a simple logical sequence (Figure 1):

1. loading of the molecular files
2. selection of the descriptors
3. calculation of the descriptors
4. saving of the calculated descriptors

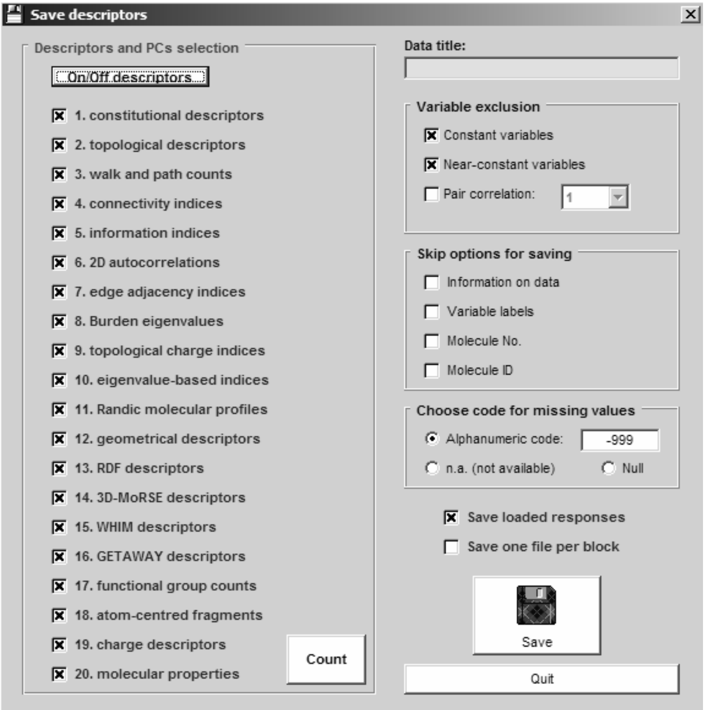


Milano Chemometrics - Licensed to Roberto,Manuela,Viviana,Davide e Andrea - DiSAT

Figure 1: Main form of the DRAGON software.

Together with the molecular descriptors, DRAGON also allows the calculation of the Principal Components (PCs) of descriptors included in the selected block in order to compress into a small number of variables the descriptor block information. Moreover, non-informative and redundant descriptors can be excluded from the output file by three different options available in the "Save descriptors" window (Figure 2):

- Constant variables: if checked, descriptors with standard deviation lower than 0.0001 or all values missing will not be saved;
- Near-constant variables: if checked, allows the exclusion of the descriptors with only one value different from the remaining ones;
- Pair correlation: if checked, allows the exclusion of one of the two descriptors with a correlation coefficient equal or greater than the selected threshold value. The allowed threshold values are from 0.9 to 1. For each pair of correlated descriptors, the one showing the highest pair correlation with the other descriptors will be automatically excluded.



Descriptors and PCs selection

☒ 1. constitutional descriptors
☒ 2. topological descriptors
☒ 3. walk and path counts
☒ 4. connectivity indices
☒ 5. information indices
☒ 6. 2D autocorrelations
☒ 7. edge adjacency indices
☒ 8. Burden eigenvalues
☒ 9. topological charge indices
☒ 10. eigenvalue-based indices
☒ 11. Randic molecular profiles
☒ 12. geometrical descriptors
☒ 13. RDF descriptors
☒ 14. 3D-MoRSE descriptors
☒ 15. WHIM descriptors
☒ 16. GETAWAY descriptors
☒ 17. functional group counts
☒ 18. atom-centred fragments
☒ 19. charge descriptors
☒ 20. molecular properties

Data title:

Variable exclusion

☒ Constant variables
☒ Near-constant variables
☐ Pair correlation: 1

Skip options for saving

☐ Information on data
☐ Variable labels
☐ Molecule No.
☐ Molecule ID

Choose code for missing values

● Alphanumeric code: -999
○ n.a. (not available) ○ Null

☒ Save loaded responses
☐ Save one file per block

Figure 2: Form of the saving options for the calculated descriptors.

Once the descriptors have been calculated, DRAGON allows performing some simple correlation and graphical analysis. Moreover, up to 100 user-defined variables can be loaded as text file and added to the calculated molecular descriptors. This option allows a simple link with other packages. DRAGON gives also the possibility to load descriptors previously calculated in order to perform descriptor analysis and/or adding user-defined variables and/or saving the descriptors using different options (Figure 2).

3. Molecule file format and pre-treatment

To run DRAGON the user needs molecular structure files previously obtained by other specific molecular modelling software. The most common molecular file formats are accepted; in particular:

1. Sybyl © MOL2 files (.mol, .ml2, mol2) by Tripos, Inc.
2. Sybyl © Molfiles (.sm2) as provided by ChemOffice, CambridgeSoft Corp.
3. Sybyl © multiple Molfiles (.mol, .ml2) by Tripos, Inc.
4. Molfiles (.mol) by Molecular Design Ltd. (MDL)
5. Multiple SD files (.sdf) by Molecular Design Ltd. (MDL)
6. HyperChem © files (.hin) by Hypercube, Inc.
7. SMILES notations (.smi) by Daylight
8. MacroModel © files (.dat, .out) by Schrodinger

To make full use of DRAGON calculations, 3D optimised structures with hydrogens should be used. However, DRAGON can also deal with H-depleted molecules and 2D-structures; in this case, it's apparent that some restrictions to descriptor calculation are applied. DRAGON does not perform geometry optimisation, nor transform topological structures into the corresponding 3D geometrical structures. If SMILES notations are used, hydrogen atoms are automatically added to the molecule structure; however, as no information on three-dimensional arrangement of atoms is represented in a SMILES notation, 3D descriptors cannot be calculated.

DRAGON 5.4 for Windows allows calculations for batches up to 50,000 molecules with a maximum of 1,000 atoms per molecule. The Linux version of DRAGON, dragonX, has no constraints.

Most of the common organic and inorganic compounds, both charged and uncharged, are correctly processed. DRAGON cannot process molecules containing atoms for which some physicochemical properties are undefined, disconnected structures such as salts, and some molecules with radicals.

Once a molecule structure has been read, DRAGON treats it in order to obtain a standard internal representation regardless of the format of the input files. The aim is to obtain unique descriptor values for a given molecule starting from different representations.

In DRAGON molecule representation, atoms are characterized by their atomic number, connectivity (i.e. the number of bonded atoms), valence (i.e. the sum of conventional bond orders of the incident bonds), charge (if specified in the input file), and some atomic properties encoding chemical information. The main atomic parameters used as the atomic weightings for molecular descriptor calculations are mass, van der Waals volume, Sanderson electronegativity and polarizability. All the weights are scaled with respect to the carbon atom. Moreover, all bonds belonging to aromatic rings are assigned a conventional bond order of 1.5 while bonds belonging to non-aromatic conjugated systems are alternating single and double bonds. Some delocalised bonds are represented according to specific rules; for example the nitro group is represented by two double bonds N = O and the N-oxide by one double bond N = O. Aromaticity detection is performed by an internal algorithm and thus it is independent of the conventional bond orders assigned by the user.

4. Molecular descriptor blocks

DRAGON provides more than 1,600 molecular descriptors that are divided into 20 logical blocks (Table 1) in order to help the user in managing so many descriptors. The user can calculate not only the simplest atom type, functional group and fragment counts, but also several topological and geometrical descriptors. Some molecular properties such as logP, molar refractivity, number of rotatable bonds, H-donors, H-acceptors, and topological surface area (TPSA) are also calculated by using some common models taken from the literature. Moreover, the Lipinski's alert (also known as "the rule of 5") together with some drug-like indices are provided to allow the selection of compounds for biological screening and/or the design of combinatorial libraries.

Constitutional descriptors (block 1) are the most simple and commonly used descriptors, reflecting the composition of a molecule without any geometrical information. Examples of these descriptors are the number of atoms, bonds, rings, specific atom types, rotatable bonds, etc. Enumerative descriptors are also counts of functional groups (block 17) and Ghose-Crippen atom-centred fragments (block 18).

The descriptor blocks 2 – 10 contain topological and topographic descriptors. Topological descriptors are based on a graph representation of the molecule. They are numerical quantifiers of molecular topology obtained by the application of algebraic operators to matrices representing molecular graphs and whose values are independent of vertex numbering or labelling. They can be sensitive to one or more structural features of the molecule such as size, shape, symmetry, branching and cyclicity and can also encode chemical information concerning atom type and bond multiplicity. Topographic indices are derived from the graph representation of molecules in the same way as the topological indices, but using the geometric distances between atoms instead of the topological distances.

The blocks 11 – 16 include descriptors derived from the knowledge of the 3D structure of the molecule, and the block 20 some molecular properties derived from literature models, such as Moriguchi logP, Ghose-Crippen logP, Lipinski rule-of-five, etc.

General details on the molecular descriptor theory, together with bibliographic references are provided in the extended User Help.

Table 1. DRAGON descriptor blocks with the number of descriptors calculated for each block.

ID	Block description	No. of descriptors
1	constitutional descriptors	48
2	topological descriptors	119
3	walk and path counts	47
4	connectivity indices	33
5	information indices	47
6	2D autocorrelations	96
7	edge adjacency indices	107
8	Burden eigenvalues	64
9	topological charge indices	21
10	eigenvalue-based indices	44
11	Randic molecular profiles	41
12	geometrical descriptors	74
13	RDF descriptors	150
14	3D-MoRSE descriptors	160
15	WHIM descriptors	99
16	GETAWAY descriptors	197
17	functional group counts	154
18	atom-centred fragments	120
19	charge descriptors	14
20	molecular properties	29

5. Data analysis

After the descriptor calculation, the user can analyse descriptor values, statistics and graphics by interactive explorative tools available by the "View descriptors" menu:

a) View block descriptor values for a selected molecule

b) View descriptor values and univariate statistics

c) View molecules in the descriptor/response space

d) Descriptor pair correlations

e) View the most and least correlated descriptors with a selected one

Option a) allows the user to view for each molecule the values of all the descriptors belonging to a selected block, while option b) allows the user to view the values of a selected descriptor for all the processed molecules, together with the histogram graph relative to the selected descriptor (Figure 3) and some simple statistics such as standard deviation, average, minimum and maximum values.

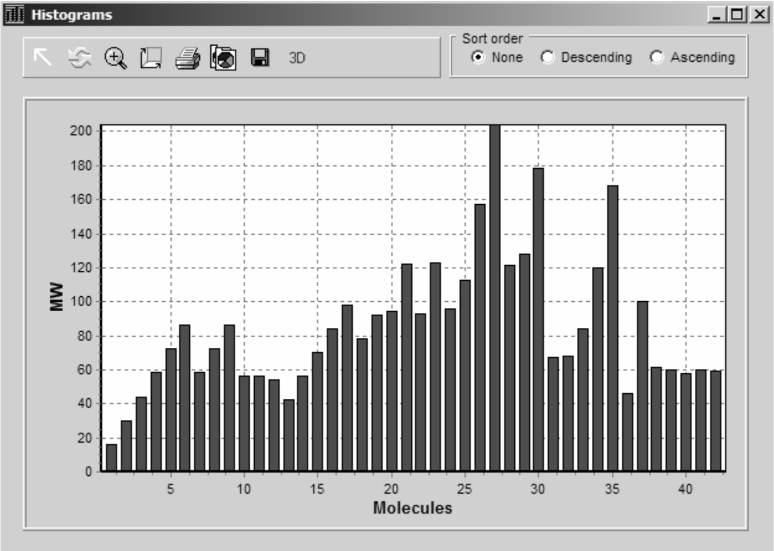
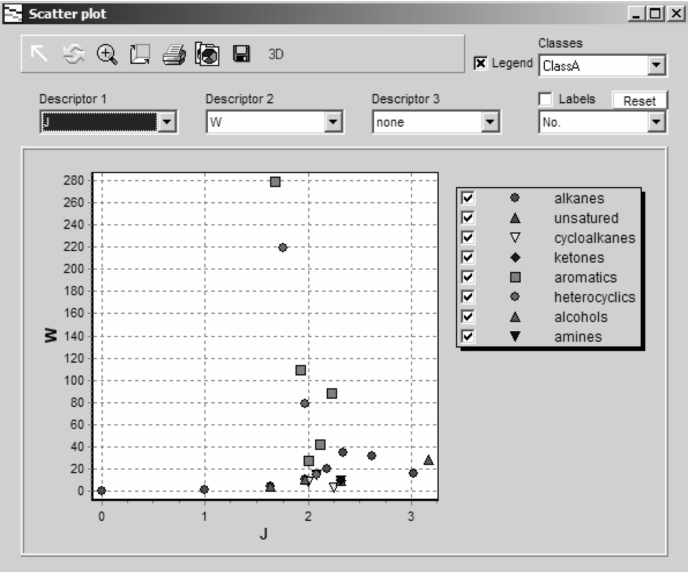


Figure 3: Histrogram of the descriptor (MW) values for the 42 loaded molecules.

Option c) opens the menu for graphics. This menu has been conceived to analyse the molecule distribution in the space defined by some selected molecular descriptors or PCs (Figure 4). Moreover, if quantitative user-defined variables like experimental properties are available, a graphical correlation analysis can be performed (Figure 5). This helps the user to identify which descriptors have some dependence on a given molecular property and could be useful in modelling it.

Option d) allows the user to calculate the correlation coefficient between two selected molecular descriptors or between one descriptor and one user-defined property. Finally, option e) allows the user to obtain three lists of descriptors and user-defined variables with different correlation levels with a selected descriptor or response (Figure 6).



Scatter plot

☒ Legend

Classes: ClassA

Descriptor 1: J

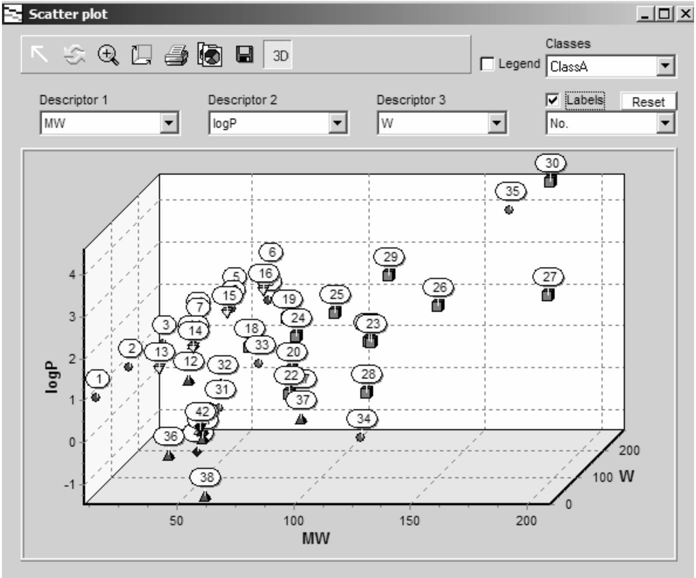
Descriptor 2: W

Descriptor 3: none

□ Labels

No.

Figure 4: Scatter plot of two descriptors (W and J) for the loaded molecules coloured according to the imported categorical variable encoding the molecule classes.



Scatter plot

□ Legend

Classes: ClassA

Descriptor 1: MW

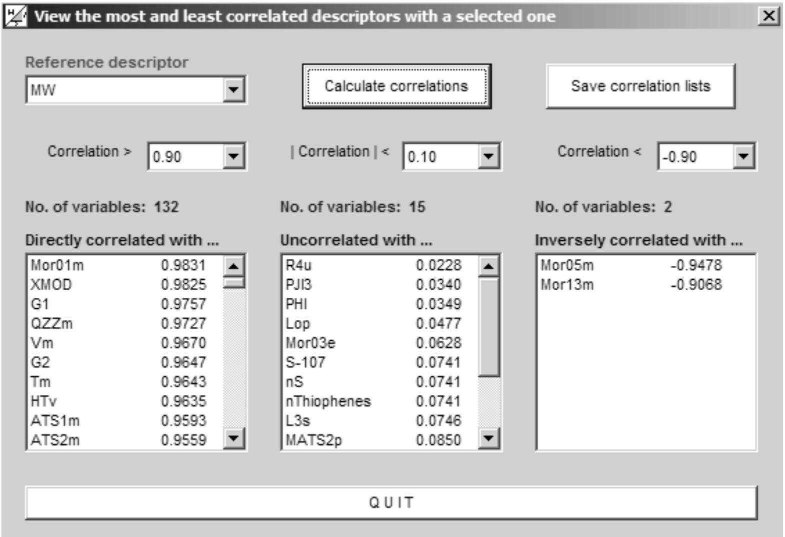
Descriptor 2: logP

Descriptor 3: W

☑ Labels

No.

Figure 5: 3D-scatter plot of two descriptors (MW and W) and a loaded property (logP) for the loaded molecules, labelled according to molecule sequential number.


View the most and least correlated descriptors with a selected one
Calculate correlations
Save correlation lists
QUIT

Reference descriptor: MW

Correlation > 0.90 | Correlation | < 0.10 | Correlation < -0.90

No. of variables: 132 | No. of variables: 15 | No. of variables: 2

Directly correlated with ...

Mor01m	0.9831
XMOD	0.9825
G1	0.9757
QZZm	0.9727
Vm	0.9670
G2	0.9647
Tm	0.9643
HTv	0.9635
ATS1m	0.9593
ATS2m	0.9559

Uncorrelated with ...

R4u	0.0228
PJI3	0.0340
PHI	0.0349
Lop	0.0477
Mor03e	0.0628
S-107	0.0741
nS	0.0741
nThiophenes	0.0741
L3s	0.0746
MATS2p	0.0850

Inversely correlated with ...

Mor05m	-0.9478
Mor13m	-0.9068

Figure 6: Lists of the most directly correlated, uncorrelated and inversely correlated descriptors, with the selected one (molecular weight), for user-defined correlation levels.

6. Operative modes, system requirements and performances

DRAGON has been designed to work both under Windows and Linux systems. DRAGON for Windows is provided in two different versions, DRAGON PROFESSIONAL, which can only work with the graphical user interface (GUI) and DRAGON PLUS, which can work both with the graphical interface and using the command line. DRAGON version for Linux, called dragonX, has not a graphical interface and it works only as a command line tool. Therefore, both Windows DRAGON PLUS and Linux dragonX allow calculations in command line mode, i.e. by reading a script file containing all the required information for loading files, calculating and saving the molecular descriptors. By this option, DRAGON can be used as client software by other applications.

A special version – called E-DRAGON – has been developed under the project Virtual Computational Chemistry Laboratory (VCC-LAB). E-DRAGON [4] uses a Java applet which employs DRAGON in order to calculate the molecular descriptors directly within an HTML page. The unique limitations are the maximum number of molecules per batch (150) and the maximum number of atoms per molecule (100). E-DRAGON can load three molecular file formats (SYBYL Mol2, MDL sdf and SMILES).

The system requirements for Windows platforms are Microsoft Windows 95/98/ME/2000/XP, Windows NT 4.0 or above, with a Pentium processor and at least 128 MB of memory RAM. For

Linux systems, dragonX has been tested on kernel 2.4 or above, on Fedora Core 1, 2, 3, 4, 5; Red Hat Linux, SUSE and Debian.

DRAGON performances have been evaluated by using two data sets from the NCI Open DataBase [5]: data set A (3,576 molecules with known experimental logP) and data set B (the first 10,000 molecules) whose characteristics are shown in Tables 2 e 3.

Table 2. Some characteristics of the data set A used for the DRAGON benchmark.

Variable	*Average*	*Minimum*	*Maximum*
molecular weight	179.61	30.03	823.04
number of atoms	23.19	4	120
number of cycles	1.28	0	7
number of circuits	1.99	0	94

Table 3. Some characteristics of the data set B used for the DRAGON benchmark.

Variable	*Average*	*Minimum*	*Maximum*
molecular weight	239.58	32.06	1701.27
number of atoms	31.63	3	176
number of cycles	1.46	0	11
number of circuits	1.99	0	124

The calculation times on the two different data sets are collected in Table 4. All the calculations have been performed on an Intel® Pentium® 4 M (1.7 GHz, 512 MB RAM). Times include also saving time.

Table 4. DRAGON benchmarks for the two considered data sets. Time is given in minutes; in brackets the number of processed molecules per minute.

	data set A		*data set B*	
Operative system	*SMILES*	*SDF*	*SMILES*	*SDF*
Windows XP Home SP2	9 (397)	15 (238)	42 (238)	77 (130)
Linux Fedora Core 3 (kernel 2.6)	4 (894)	10 (358)	21 (476)	52 (256)

The MlogP and AlogP models, both implemented in DRAGON, has been evaluated on the data set A, giving correlations with the experimental logP of $R^2 = 0.935$ and $R^2 = 0.931$, respectively.

7. Conclusions

DRAGON is user-friendly software, which, in the last few years, has been used in more and more different scientific fields. In addition to the easy use, DRAGON can provide a huge number of molecular descriptors, can calculate them from the most common molecule file formats, gives different options in output saving, and, lastly, assures reliable descriptor values. In effect, a work is continuously in progress to check descriptor values, fix bugs, update algorithms and add new descriptors. DRAGON is widely regarded as the major software for the calculation of molecular descriptors. Even since DRAGON was first released, its user base has grown steadily. DRAGON has become a standard in many organizations, and it is used today in many of the biggest pharmaceutical companies, in several major research departments and numerous of the largest universities in the world.

References

[1] R.Todeschini and V.Consonni (2000). *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim (GER), 667 pp.

[2] R. Todeschini, M. Lasagni, and E. Marengo (1994). New Molecular Descriptors for 2D- and 3D-structures. Theory. *J. Chemometrics*, **8**, 263-273.

[3] Talete srl, *Dragon* (ver. 5.4), Milano, Italy. Web site: www.talete.mi.it/products/software.htm

[4] I.V. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V.A. Palyulin, E.V. Radchenko, N.S. Zefirov, A.S. Makarenko, V.Y. Tanchuk, and V.V. Prokopenkov (2005). Virtual Computational Chemistry Laboratory - Design and Description. *J. Computer-Aided Mol. Des.*, **19**, 453-463.

[5] W.-D. Ihlenfeldt, J.H. Voigt, B. Bienfait, F.Oellien, and M.C. Nicklaus (2002). Enhanced CACTVS Browser of the Open NCI Database. *J. Chem. Inf. Comput. Sci.*, **42**, 46-57.