

Introduction to MOLE DB – on-line Molecular Descriptors Database

Davide Ballabio*, Alberto Manganaro, Viviana Consonni, Andrea Mauri, Roberto Todeschini

Milano Chemometrics and QSAR Research Group, Department of Environmental Sciences – University of Milano-Bicocca, P.za della Scienza, 1 – 20126 Milano, Italy

(Received January 12, 2009)

Abstract

Molecular descriptors are the final result of logic and mathematical procedures which transform chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment. The number of molecular descriptors is tremendously increased in the last decades, due to their fundamental role in modelling and understanding the relationships between chemicals and physico-chemical properties, biological activities, toxicological behaviour, environmental impact, analytical measurements.

A free web-based database, named the MOLE db - Molecular Descriptors Database (http://michem.disat.unimib.it/mole_db/) was recently implemented; it is comprised of 1124 molecular descriptors calculated on 234773 molecules derived from the NCI database. This database is intended as a research and teaching tool, allowing to search for a specific group of molecules and analyse the corresponding values of molecular descriptors.

Basically, the database allows the user to search for a specific group of molecules, view the corresponding values of selected molecular descriptors, and save in an output file the values of a set of molecular descriptors.

* Corresponding author
E-mail: davide.ballabio@unimib.it
Telephone: +39-02-6448.2801
Fax: +39-02-6448.2839

Introduction

In the last decades, several scientific studies have been focused on studying how to convert the information encoded in the molecular structure into one or more numbers, called molecular descriptors. *“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].

The number of molecular descriptors is tremendously increased due to their fundamental role in modelling and understanding the relationships between chemicals and physico-chemical properties, biological activities, toxicological behaviour, environmental impact, analytical measurements. The great interest for molecular descriptors is also documented by published books [1-3] and a huge amount of papers dealing with molecular descriptors in conjunction with several chemometric and chemoinformatic approaches. Moreover, since each molecular descriptor encodes only a small part of the whole chemical information contained into the real molecule, several molecular descriptors are continuously required in order to describe the complexity of the analysed chemical systems.

By now molecular descriptors have become among the most important variables for molecular modelling and, as a consequence of that, they have a strong relationship with statistics, chemometrics and chemoinformatics. It is noteworthy that the application of molecular descriptors provided a big change in the scientific paradigm. While until 30 years ago molecular modelling mainly consisted in searching for mathematical relationships between experimentally measured quantities, now a measured property is usually modelled by several theoretical molecular descriptors able to encompass structural chemical information. Consequently, adequate statistical and informatic tools have been developed in order to handle the huge amount of information given by molecular descriptors and produce predictive mathematical models [4]. In order to explain the complex relationships between molecules and observed quantities, two main streams were developed, one related to the search for relationships between molecular structures and physico-chemical properties (QSPR, Quantitative Structure-Property Relationships) and the other between molecular structures and biological activities (QSAR, Quantitative Structure-Activity Relationships).

Recently, with the aim of collecting information related to different molecular descriptors, we implemented a free web-based database, named the MOLE db - Molecular Descriptors Database (http://michem.disat.unimib.it/mole_db/). The database collects values of 1124 molecular descriptors calculated for 234773 molecules included in the NCI database. The NCI database is the publicly available part of the structure collection assembled by the National Cancer Institute (NCI) and was chosen as it is a valuable free resource providing a diverse and oftentimes unique structure set [5]. By calculating the molecular descriptors on

the NCI molecules and publishing all the results as an accessible web-based database, we wanted to extend and integrate the chemical knowledge connected to the NCI database. All the scientists interested in molecular modelling will be able to easily search the molecular descriptors relative to the NCI molecules and analyze the results of such searches both directly on-line and by exporting the results.

Consequently, the MOLE db - Molecular Descriptors Database is intended as a research and teaching tool, allowing to search for a specific group of molecules and to analyse the corresponding values of molecular descriptors. In the present paper, the MOLE db - Molecular Descriptors Database is introduced. First, it is explained how the database has been built; then, its main features and applications are shown.

Data, software and hardware

The MOLE database was initially built by using the greatest part of the molecules included in the National Cancer Institute (NCI) data set. The Enhanced NCI Database Browser [6] provided molecular structures as SDF files. SDF is a digital format that encloses in a single file information about the structure of a molecule (atom types, 3D atomic coordinates, bond connections, bond types, etc.).

This set of molecules was then processed by DRAGON software [7,8] in order to calculate the corresponding molecular descriptors. DRAGON calculates various molecular descriptors ranging from count descriptors to more complex geometrical descriptors. In Table 1, the 13 groups of DRAGON molecular descriptors included in the MOLE database are listed. The number of descriptors included in each block is reported in the last column of Table 1.

Table 1. DRAGON molecular descriptor blocks included in the database.

DRAGON Block No.	DRAGON Block Name	No. of descriptors
1	constitutional descriptors	48
2	topological descriptors	119
4	connectivity indices	33
5	information indices	47
6	2D autocorrelations	96
8	Burden eigenvalues descriptors	64
10	eigenvalue-based indices	44
12	geometrical descriptors	74
15	WHIM descriptors	99
16	GETAWAY descriptors	197
17	functional group counts	154
18	atom-centred fragments	120
20	molecular properties	29

The selected descriptor blocks cover different approaches in descriptors research and sum up to a total number of calculated descriptors equal to 1124. In particular, in the molecular properties block different logP estimations, drug-like, hydrophilic and refractivity indices are included.

As some molecules can not be processed by DRAGON software, the final number of molecules constituting the database is 234773. Some descriptors could not be calculated on all the molecules and these missing values are reported in the database as "n.a." (and with a standard numerical code equal to -999 when exporting the query results in text files).

Thus, the final data set can be seen as a matrix of 234773 rows (the molecules) and 1124 columns (molecular descriptors), where each element x_{ij} represents the j -th molecular descriptor calculated for the i -th molecule. It has to be highlighted that this data set is static, since it has been calculated once and then inserted into the database; in other words, when a search query is submitted to the database, the selected molecules are found and the corresponding molecular descriptors are shown to the user, but no instant on-line calculations are made. This allows the searching to be relatively fast, so that basic search queries on the database usually require less than 1 second to be solved.

The molecules in the database are numbered with two different counters: the original NCI number and an internal MC number, corresponding to 20000 plus the NCI number. In fact, the first 20000 positions of the database are reserved to other molecules (external to the NCI database), that will be added in the next future.

In order to implement the database, a PHP-MySQL solution was chosen. The MySQL database engine was installed on a dedicated Linux server, and the data were introduced paying particular attention to the data table structure, aiming at optimizing the query response time (which may be a critical issue due to the data size). Whereas, some of the data pre-processing and the calculations of the molecular descriptors were carried out on different platforms (Windows XP and Linux CentOS), the database and its web-based Graphical User Interface are running on an Intel Xeon (3.2 GHz) server with 1 GB of RAM, mounting a Linux (CentOS distribution) operating system.

Database characteristics and capabilities

The MOLE database can be thought of a huge numerical matrix and cannot be processed in its original form, since molecules contained in the NCI database are heterogeneous and consequently the different sources of information should be selected before analysing the database.

In order to search for the relevant information, on the basis of the analysis goal, it is possible to perform on-line search queries on the database. Basically, the query enables the user to

search for a specific portion of information of the whole database. Once the query has been executed and the selected portion of information is available, the user can look at, interpret and save the results directly on-line. Through the database web interface several types of queries are available and subsequently the user can browse the results, export them and examine details of single molecules.

Table 2. Searching criteria provided by the MOLE database query form

Criterion	Note
MC number	Queries on MC number ranges
NCI number	Queries on NCI number ranges
CAS number	Queries on specific CAS numbers
Name	Queries on substrings or exact match of strings
Formula	Queries on substrings or exact match of strings
Molecular Descriptors values	Queries on 5 joined ranges of molecular descriptors

The first form of the web interface is the query form (Figure 1), that enables the user to make a query using different criteria. The available criteria are listed in Table 2.

MOLE db - Molecular Descriptors Data Base
dig into the knowledge! Milano Chemometrics and QSAR Research Group

The MOLE db - Molecular Descriptors Data Base is a free on-line database comprised of 1124 molecular descriptors calculated for 234773 molecules. At the present moment, 1674 queries have been made on the database. This data base is intended as a research and teaching tool; please, read here general details, references, credits, limitations, warranty and conditions. Otherwise, click here for help on how to query the data base.

Search for MC number from to

Search for NCI number from to

Search for CAS number

Search for name Exact match

Search for formula Exact match

Search for descriptor values:

Group name for values and

Group name for values and

Group name for values and

Group name for values and

Group name for values and

Order results by

Developed by Milano Chemometrics and QSAR Research Group :: Info and details on Molecular Descriptors Data Base

Figure 1. MOLE db query form layout

When more than one criterion is used, the query considers them jointly, i.e. concatenating them with an AND logical operator; thus, the selected molecules will be those matching with all the used criteria. With regard to the name and the formula criteria, it is possible to search for any molecule with the name or formula containing the given string, or limit the query only to those molecule that have an exact match with the given string. For the molecular descriptor criterion, it is possible to select up to 5 molecular descriptors and create a particular search range for each of them (between two values, greater/lower than a defined threshold). Moreover, in order to aid the management of results, the maximum number of molecules to be searched for is restricted to 1000 in each query.

After executing a query, all the molecules matching with the selected criteria are shown (Figure 2).

Found 5 molecules, showing 1-5

	Name	MC No	CAS No	Formula	MW	
1.	2-methylbenzo-1,4-quinone	20001	553-97-9	C ₇ H ₆ O ₂	122.13	show details
2.	5-(hydroxy(oxido)amino)-2-imino-2,3-dihydro-1,3-thiazole	20004	121-66-4	C ₃ H ₃ N ₃ O ₂ S	145.16	show details
3.	2-chloro-4,6-bis(hydroxy(oxido)amino)phenol	20003	946-31-6	C ₆ H ₃ ClN ₂ O ₅	218.56	show details
4.	2-aminoantra-9,10-quinone	20005	117-79-3	C ₁₄ H ₉ NO ₂	223.24	show details
5.	2-(1,3-benzothiazol-2-ylidithio)-1,3-benzothiazole	20002	120-78-5	C ₁₄ H ₈ N ₂ S ₄	332.52	show details

Show a descriptor for all the matching molecules:

Block descriptor

Sort by descriptor value

Select descriptor block and fields to be saved:

MC no NCI no Name Formula CAS no MW SMILES

Block

As txt file, fields separated by

[back to query form](#) :: [help](#) :: [new search](#)
 Query executed in 0.213 seconds

Figure 2. MOLE db main form. Results of a search query are shown.

The molecular descriptor values for the resulting molecules can be exported as a plain text file, allowing the user to further analyse them by other applications and softwares. The user can export each molecular descriptors block, one at a time. Furthermore, it is possible to

select other information to be exported, such as MC number, NCI number, CAS number, name, chemical formula, SMILES string and molecular weight of the molecules. Finally, the user can also select the field separator to be used in the output file, so that the exported data can be easily imported in other softwares.

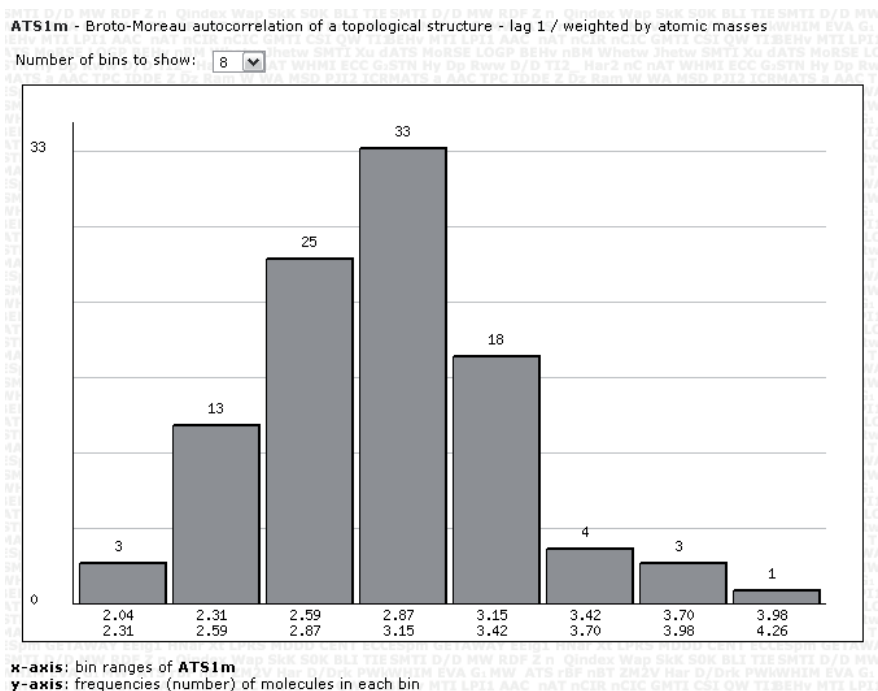


Figure 3. Histogram showing the distribution of a specific molecular descriptor (ATS1m, Broto-Moreau autocorrelation of a topological structure - lag 1 / weighted by atomic masses) on the searched molecules.

In order to facilitate on-line analyses, the web interface allows the user to directly look at the molecular descriptor values for all the selected molecules. It is possible to select a single descriptor and then display its values, sorted or preserving the order of the selected molecules by means of the button “show values”. Furthermore, the user can choose to display a histogram (“show chart”), reporting the frequency distribution of the descriptor values. In this chart (shown in Figure 3), the number of bins to be used to divide the frequency range can be modified.

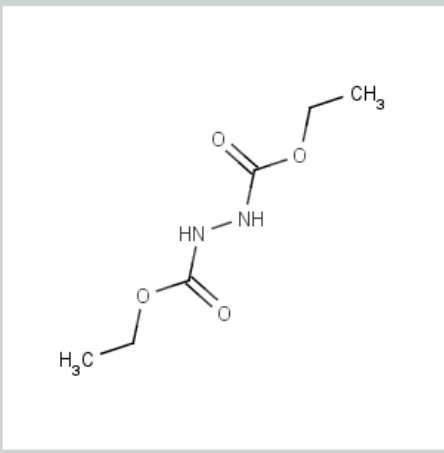
diethyl 1,2-hydrazinedicarboxylate	
Formula: C₆H₁₂N₂O₄ MW: 176.2 CAS number: 4114-28-7 MC number: 23002 NCI number: 3002	Show descriptor values for group: constitutional descriptors topological descriptors connectivity indices information indices 2D autocorrelations Burden eigenvalues descriptors eigenvalue-based indices geometrical descriptors WHIM descriptors GETAWAY descriptors functional group counts atom-centred fragments molecular properties
	
SMILES string: C(NNC(=O)OCC)(OCC)=O	

Figure 4. MOLE db form showing the details of a single selected molecule together with its molecular graph.

It is also possible to look at the details of each specific selected molecule (Figure 4). In the detail page of a single molecule, the information about the selected molecule is shown (Formula, Molecular Weight, CAS number, MC number, NCI number, SMILES string). Moreover, the user can display in a single table all the molecular descriptor values of a single block (see Table 1) for the chosen molecule. Finally, the molecular graph of the molecule (derived from the corresponding SMILES string) is shown in a java applet (MarvinView, developed by ChemAxon Ltd.), allowing the user to perform some basic manipulations and editing on it.

Conclusions

A free web-based database, named the MOLE db - Molecular Descriptors Database (http://michem.disat.unimib.it/mole_db/) was recently released. The database includes 1124 molecular descriptors calculated on 234773 molecules.

The MOLE database allows the user to select a specific group of molecules by means of different query criteria and directly analyse the corresponding molecular descriptors. Thus, the user can directly look at the molecular descriptor values for all the selected molecules or display a histogram with the frequency distribution of the descriptor values. Moreover, the query results can be saved in a plain text file, allowing the user to further analyse the exported molecular descriptors by other applications and softwares.

Furthermore, the database will be enlarged by adding new molecules, with particular attention to chemical sets that could be of particular interest for educational and research purposes (like PCB, CFC, pesticides).

Acknowledgements

Authors would like to thank the authors of the Enhanced NCI Database Browser (<http://129.43.27.140/ncidb2/>) for providing the molecular structures as SDF files.

References

- [1] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Wiley - VCH, Weinheim, 2000.
- [2] M. Karelson, *Molecular Descriptors in QSAR/QSPR*, Wiley-Interscience, New York, 2000.
- [3] M. V. Diudea, *QSPR/QSAR Studies by Molecular Descriptors*, Nova Science Publishers, Huntington, 2001.
- [4] J. Gasteiger, The central role of chemoinformatics, *Chemometrics and Intelligent Laboratory Systems* **82** (2006) 200-209.
- [5] J. H. Voigt, B. Bienfait, S. Wang, M.C. Nicklaus, Comparison of the NCI Open Database with Seven Large Chemical Structural Databases, *Journal of Chemical Information and Computer Science* **41** (2001) 702-712.
- [6] W. Ihlenfeldt, J. H. Voigt, B. Bienfait, F. Oellien, M. C. Nicklaus, Enhanced CACTVS Browser of the Open NCI Database, *Journal of Chemical Information and Computer Science* **42** (2002) 46-57.
- [7] Talete srl, DRAGON for Windows (Software for Molecular Descriptor Calculations), 2008
- [8] A. Mauri, V. Consonni, M. Pavan, R. Todeschini, DRAGON software: an easy approach to molecular descriptor calculations, *MATCH Communications in Mathematical and in Computer Chemistry* **56** (2006) 237-248.