A Java Tool for the Management of Chemical Databases and Similarity Analysis Based on Molecular Graphs Isomorphism

Irene Luque Ruiz and Miguel Ángel Gómez-Nieto

University of Córdoba. Department of Computing and Numerical Analysis.
Campus de Rabanales. Albert Einstein Building
E-14071 Córdoba (Spain)
{iluque,mangel}@uco.es

Abstract. This paper describes a computational chemistry solution for the management of large chemical databases of molecules and the performing of isomorphism calculation for the analysis of database similarity and diversity. The system has been fully developed using Java language and it uses other free and standard Java library. The system allows to the user the building of databases of molecules, store information about the molecules, the matching among molecules using different isomorphism paradigms and the similarity/diversity analysis of databases through a wide number of similarity indices.

Keywords: Computational chemistry, isomorphism, matching, similarity and diversity analysis, chemical database management, Java.

1 Introduction

The investigations in Computational Chemistry [1] use computers to assist in solving chemical problems, they incorporate the results of theoretical chemistry into efficient computer programs, aimed to calculate the structures and properties of molecules.

While its results normally complement the information obtained by chemical experiments, computational chemistry find predict hitherto unobserved chemical phenomena. Several major areas may be distinguished within computational chemistry [1-3]:

- The prediction of the molecular structure of substances.
- Storing and searching for data on chemical entities.
- Identifying correlations between chemical structures and properties (QSPR and QSAR).
- Computational approaches to help in the efficient synthesis of compounds.
- Computational approaches to design molecules that interact in specific ways with other molecules (e.g. drug design).

Computational Science plays an important role in modern molecular discovery systems [4-5]. Nowadays, a large number of compounds are tested and optimized in

M. Bubak et al. (Eds.): ICCS 2008, Part II, LNCS 5102, pp. 369 – 378, 2008.

order to identify a molecule or sets of molecules that act favorably in some desired circumstances. One of the factors limiting in identifying new candidates is the availability of proper collections of chemical compounds in order to propose a prediction model of some molecular property needed.

Similarity analysis engines [6-7] are often used to compute resemblances between molecules in order to analyze and organize databases of molecules and to develop new QSAR models. The similarity between two molecules can be measured as the differences between molecular graphs representing their chemical structures, if structurally similar molecules are more likely to have resembling properties.

In this paper, we present a tool devoted to the database management of molecules in order to the development of chemoinformatic solutions. Chemoinformatics or chemical informatics is the use of computers and informational techniques, applied to a range of problems in the field of chemistry. These in silico techniques are used in pharmaceutical companies in the process of drug discovery. Applications of chemoinformatics include the storage of information related to chemical compounds, the in silico representation of chemical structures using specialized formats, the virtual screening of large in silico virtual libraries (usually built using combinatorial chemistry) of compounds, QSAR used to predict the activity of compounds from their structures, etc.

The tool described in this paper embraces some of these chemoinformatic applications and is open to include new ones in a future. We describe the database management component aims at the building of molecules databases, permitting the storing, retrieving, visualizing, and similarity analysis of sets of molecules organized in different projects.

The paper is organized as follows: first of all we describe the tool functionality using the main use case. In the next section the database model is described, later the system architecture and tool characteristics are described using some representative screenshots. Finally, we discuss the applications and future works.

2 Background

Although there is a wide set of databases [8] in Internet aimed to offer researchers a tool to access to chemical information, these databases do not permit the user its own information management. They are government, company or industry owner databases and researchers can only access and visualize the information stored in these databases. There are other proprietary systems aimed to the management of chemical information. They usually are modular systems, where the researchers have to pay a high cost for obtaining each one of the functionalities needed (database management, similarity analysis, molecular descriptor calculation, etc.).

In the last year, different free solutions have been developed. CDK [9] is a free Java library of chemical-informatics packages aimed to QSAR, and 2D and 3D chemical solutions. Other examples are Bioeclipse [10] and Instant JChem [11]. Both software products, currently, allow to the researcher the management of chemical information, but they do not include other functionality. However, Bioeclipse, a free

Java tool, allows the expert researchers to develop new functionality and to include it in the user environment of the tool.

We describe in this paper a fully developed tool in Java aimed to both, management of user chemical databases and similarity analysis. This system includes different Java free components developed by other researchers and algorithms developed by the authors in order to offer to the scientific community a tool for the management of databases of chemical compounds, and the similarity/diversity database analysis using similarity principles based on isomorphism calculation.

The investigations in Computational Chemistry use computers to assist in solving chemical problems, they incorporate the results of theoretical chemistry into efficient computer programs, aimed to calculate the structures and properties of molecules. While its results normally complement the information obtained by chemical experiments, computational chemistry find predict hitherto unobserved chemical phenomena.

3 System Description

The system functionality is shown in the use case diagram of Fig. 1. Three main functionalities are represented:

- *Database management*: the user can build new databases from existing databases or files storing molecules in any of the standard file formats (SIMILE. MOL, SDF, etc.) [12]. Databases can be built using both Oracle 10g and MySQL 5.0 database systems, using the relational database model.
- *Project management*: the user can build new projects incorporating molecules from the existing databases and/or external files of molecules. The project management is quite similar to the database management; however; here the user can modify, delete or insert new molecules characteristics (e.g. properties) in order to tailor specific groups of molecules for a later study.
- *Isomorphism calculation*: the user can execute different isomorphism algorithms on the groups of molecules stored in the projects. The results of the isomorphism calculation are managed by the user who can carry out different similarity analysis in order to study database diversity, molecular resemblance or to developed different QSAR models.

3.1 Database and Project Management

The user can create as many databases of molecules as required. Databases are created empty using a creation script or importing molecules from both other databases, and files containing molecules in some of the standard formats.

For existing databases new molecules can be added, extracted or transferred to other database or even saved in external files. Using item menus, control or drag and drop tips (like in standards Windows applications) the user can manage the database content easily.

When the molecules are imported into the database, properties of the molecules are stored. These properties are allocated in a referenced table storing the property value and its meaning. New properties can be imported from external files or added using specific functionality of the systems.

Panels distributed along the windows application allow the user to visualize the information stored in the database. Molecules are visualized using *Structure* [13] paint tool; a free tool aimed to the representation of molecular graphs. Remaining information about molecules as: properties values, number and type of nodes and edges, molecular weight, etc., is visualized with tables distributed along the windows environment. The management of the project (see Fig. 1) is quite similar to the previously described database management. Indeed, the system considers a project as a tailored and owner user database.

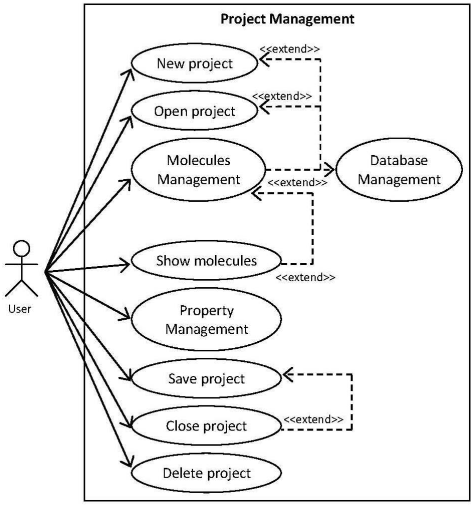


Fig. 1. Use case diagram of Project Management

User can create as many projects as necessary, importing the information from existing databases or external files, and afterwards modify or extend the database content in order to prepare a set of molecules from a chemoinformatic analysis.

3.2 Isomorphism Calculation

Several isomorphism criteria are included in the system described in this paper.

- *Exact isomorphism*: calculates whether two molecular graphs are exactly equals. This calculation is performed using CDK class [9], and the results just inform if two molecular graphs are equals or not.
- *Subgraph isomorphism*: calculates whether a molecular graph is part (is included) of another one. CDK class is used for this calculation. Results are obtained if one of the matched molecules is an entire subgraph of another one.
- *All Common Fragments* (CF): calculates the set of all common fragments between the two matched molecular graphs. In this set some fragments may be included in other larger fragments. As result a list of the common fragments between the two

matched molecules is obtained. An owner method also allows us to obtain all the non-common fragment of both matched molecules.

- *Maximum common subgraph* (MCS): calculates the maximum common subgraph between two given molecular graphs. This isomorphic method is performed using an algorithm developed by the authors [14].
- *Maximum overlapping set* (MOS): also called Maximum Common Edges Subgraphs (MCES) is the set of common fragments without repetition between two molecules.

The calculation of MCS and MOS isomorphism also retrieves the non isomorphic fragments of the both matched graphs; that is, those fragments belonging to each matched molecular graph not included in the maximum common fragment (in the matched graph). Non isomorphic fragments are widely applied in QSAR models for activity predictions of drugs.

In all the isomorphism calculation methods, the user can include all molecules of a project or only a selected group to perform the calculation.

The results of the isomorphism calculation are represented in an appropriate canvas as tables. The user can interact with the tables in order to visualize the graphs resulting of the isomorphism calculation (common and non-common) and the corresponding information related with the result. Moreover, these results can be stored in external text files and the isomorphic and non-isomorphic subgraphs in standards molecule formats.

3.3 Similarity Analysis

Similarity analysis is carried out with any of the results obtained with the isomorphism calculation utility. Several similarity indices can be considered for this analysis (Tanimoto, cosine, Raymond, Dice, Kulczynski, Simpson, Forbes, Hamman, Pearson, Yule, etc.) [5-6] and the results are shown in symmetrical tables in a specific windows canvas. Moreover, the user can save these results in external text files in order to carry out, for instance, a further statistical analysis.

4 Database Structure

Figure 2 shows the resumed class diagram of the database structure [15]. Projects are stored in a table with information related with the project identification and management. Each project maintains information about a set of molecules.

For each molecule information on its identification, name, formula, smile format description, references to the database where was extracted, between other are stored.

The class Properties maintains information about different properties which may be related with the molecules. This is a master class. *PropertyMolecule* class is related with both Molecule and Properties classes, so, keeping information about the property value of each molecule stored in the database.

Isomorphism results are represented by different classes depending on the isomorphism type. Hence, *GraphIsomReslt* class stores information about the exact and subgraph isomorphism. Attributes exact and subgraph inform and store information relative to these isomorphism types.

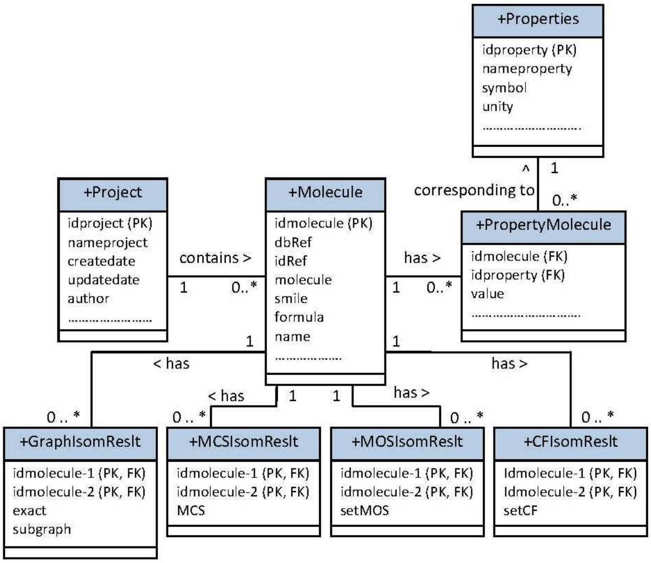


Fig. 2. Main classes of the database

Furthermore, *MCSIsomReslt*, *MOSIsomReslt* and *CFIsomReslt* classes store information about the types of isomorphism MCS, MOS and CF respectively. For all these classes the isomorphism subgraphs are stored as a set of subgraphs; the non-isomorphic subgraph can be later obtained through the knowledge of both the matched molecules and isomorphic subgraphs (both class attributes).

All classes related with the representation of isomorphism information have defined some constraints in order to not store information; that is, pairs (idmolecule-1, idmolecule-2) are unique without order consideration.

5 System Architecture

The system has been designed using UML 2.0 method [15], and as we have previously commented, fully codified using Java, free source Java code and Eclipse software environment [16].

Figure 3 shows some of the main system classes. Any remarkable functionality of the system is implemented using Eclipse Perspectives model. A Perspective is a set or windows or Panels with related functionality and that the user usually uses in a combined way. The user can change between different Perspectives as desired; however, only a Perspective (a set of Panels) can be active at the same time.

In an execution session different database connections can be open at the same time. In this way, the user can manage information (molecules) of different databases in order to insert new molecules in a project. Besides, different projects can be open at a time. However, the user can only manipulate the molecules of the selected or active

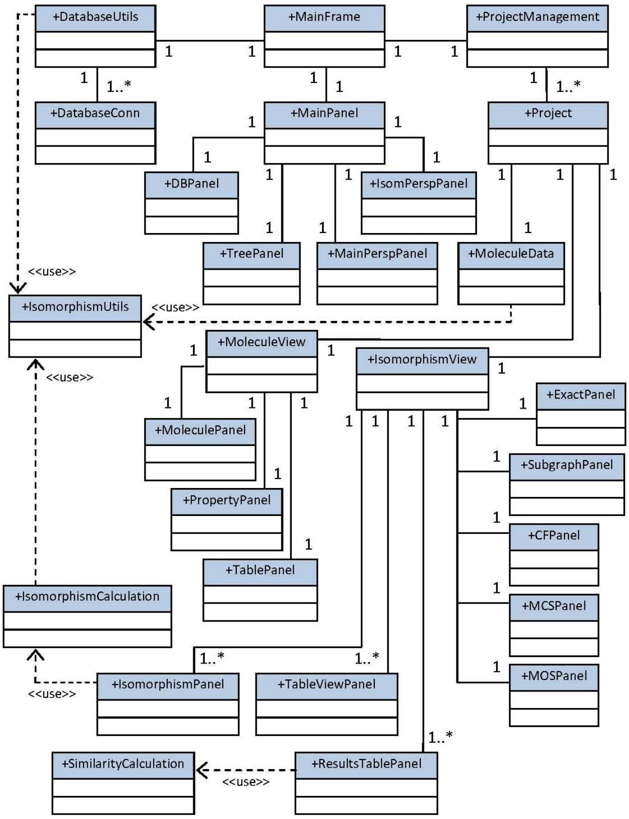


Fig. 3. Some of fthe main system classes

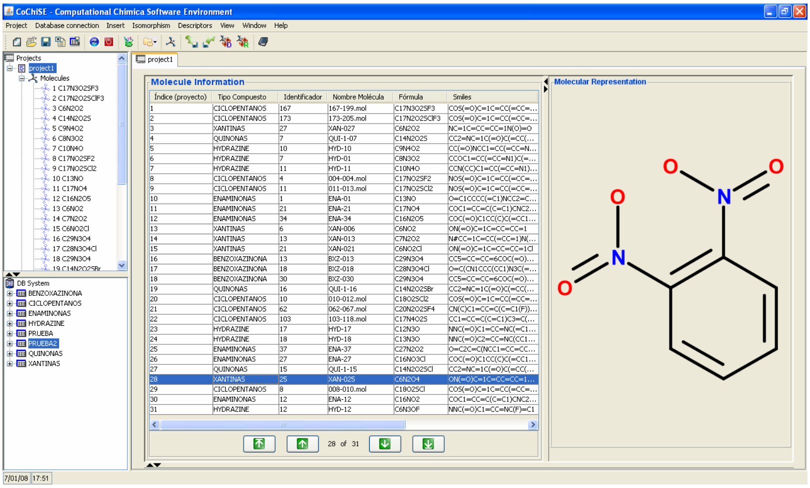
project. For each one of the specific functionalities a Panel class in charge of the manipulation of the corresponding information is defined. So, the user can manage and visualize huge information at a time in the same Perspective.

As can be observed in Fig. 3, the user navigates among the projects and databases through a hierarchical menu or tree where database and project objects information can be accessed and managed (*TreePanel* class).

Similarity calculation is performed using the *SimilarityCalculation* class. This class implements methods for the calculation of all the similarity indices. Results are presented in different tables using *ResultsTablePanels* class.

5.1 Molecule Management Interface

Figure 4 shows the screenshots corresponding to the molecule management perspective. As observed in the windows left-hand two panels are devoted to the project and database management respectively.



Índice (proyecto)	Tipo Compuesto	Identificador	Nombre Molécula	Fórmula	Smiles
1	CICLOPENTANOS	167	167-199.mol	C17N3O2SF3	COS(=O)C=1C=CC(=CC=...
2	CICLOPENTANOS	173	173-205.mol	C17N2O2SClF3	COS(=O)C=1C=CC(=CC=...
3	XANTINAS	27	XAN-027	C6N2O2	NC=1C=CC=CC=1N(O)=O
4	QUINONAS	7	QUI-1-07	C14N2O2S	CC2=NC=1C(=O)C(=CC(...
5	HYDRAZINE	10	HYD-10	C9N4O2	CC(=O)NCC1=CC(=CC=N...
6	HYDRAZINE	7	HYD-01	C8N3O2	CCOC1=CC(=CC=N1)C(=...
7	HYDRAZINE	11	HYD-11	C10N4O	CCN(CC)C1=CC(=CC=N1)...
8	CICLOPENTANOS	4	004-004.mol	C17NO2SF2	NOS(=O)C=1C=CC(=CC=...
9	CICLOPENTANOS	11	011-013.mol	C17NO2SCl2	NOS(=O)C=1C=CC(=CC=...
10	ENAMINONAS	1	ENA-01	C13NO	O=C1CCCC(=C1)NCC2=C...
11	ENAMINONAS	21	ENA-21	C17NO4	COC1=CC=C(C=C1)CNC2...
12	ENAMINONAS	34	ENA-34	C16N2O5	COC(=O)C1CC(C)C(=CC1...
13	XANTINAS	6	XAN-006	C6NO2	ON(=O)C=1C=CC=CC=1
14	XANTINAS	13	XAN-013	C7N2O2	N#CC=1C=CC(=CC=1)N(...
15	XANTINAS	21	XAN-021	C6NO2Cl	ON(=O)C=1C=CC=CC=1Cl
16	BENZOXAZINONA	13	BXZ-013	C29N3O4	CC5=CC=CC=6COC(=O)...
17	BENZOXAZINONA	18	BXZ-018	C28N3O4Cl	O=C(CN1CCC(CC1)N3C(=...
18	BENZOXAZINONA	30	BXZ-030	C29N3O4	CC5=CC=CC=6COC(=O)...
19	QUINONAS	16	QUI-1-16	C14N2O2SBr	CC2=NC=1C(=O)C(=CC(...
20	CICLOPENTANOS	10	010-012.mol	C18O2SCl2	COS(=O)C=1C=CC(=CC=...
21	CICLOPENTANOS	62	062-067.mol	C20N2O2SF4	CN(C)C1=CC=C(C=C1(F))...
22	CICLOPENTANOS	103	103-118.mol	C17N4O2S	CC1=CC=C(C=C1)C3=C(...
23	HYDRAZINE	17	HYD-17	C12N3O	NNC(=O)C1=CC=NC(=C1...
24	HYDRAZINE	18	HYD-18	C13N3O	NNC(=O)C2=CC=NC(CC1...
25	ENAMINONAS	37	ENA-37	C27N2O2	O=C2C=C(NCC1=CC=CC...
26	ENAMINONAS	27	ENA-27	C16NO3Cl	COC(=O)C1CC(C)C(=CC1...
27	QUINONAS	15	QUI-1-15	C14N2O2SCl	CC2=NC=1C(=O)C(=CC(...
28	XANTINAS	25	XAN-025	C6N2O4	ON(=O)C=1C=CC=CC=1...
29	CICLOPENTANOS	8	008-010.mol	C18O2SCl	COS(=O)C=1C=CC(=CC=...
30	ENAMINONAS	12	ENA-12	C16NO2	COC1=CC=C(C=C1)CNC2...
31	HYDRAZINE	12	HYD-12	C6N3OF	NNC(=O)C1=CC=NC(F)=C1

Fig. 4. Project and database management perspective

Using a tree menu, different projects and the corresponding molecules can be accessed. Furthermore, in the windows right-hand a table is used to show all the molecule data. When a molecule is set both in the tree menu or molecule table, the corresponding molecular graph is also shown. The user can manage different projects and tables of molecules in the database. So, the user can copy molecules between projects or between database tables and projects. Moreover, the user can import new molecules to a project or database table at any time.

5.2 Isomorphism Management Interface

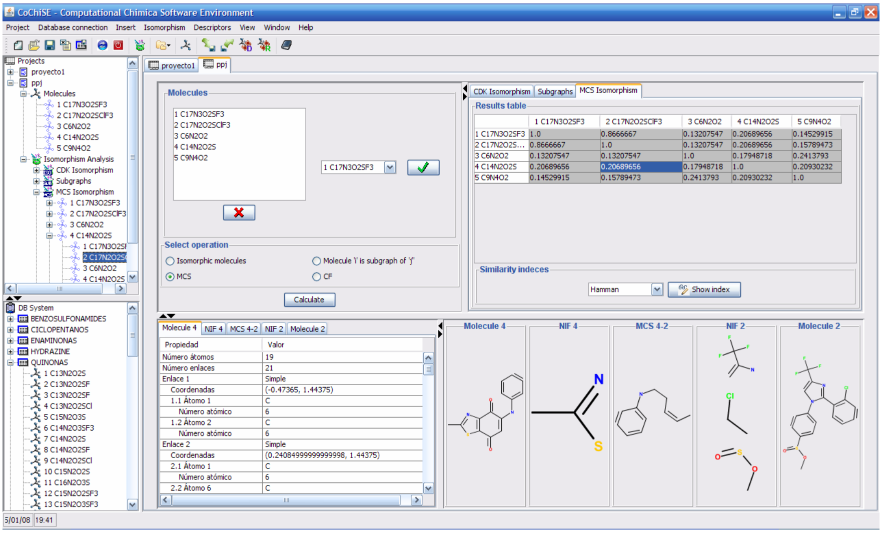
Figure 5 shows the screenshots corresponding to the isomorphism management perspective. The left-hand of the window is the molecule management perspective. However, these panels are dynamic, and the user can minimize them if they are not necessary any more.

As Fig. 5 shows, a panel is devoted to the user selection of the isomorphism type calculation. User can select all project molecules or a subset and the type of isomorphism to be calculated.

Once the isomorphism is calculated, the isomorphism results are shown through two panel: a) a panel shows the molecular graphs corresponding to the common and non-common fragments corresponding to the matched molecules which are selected in the left-hand tree project menu (or in the similarity results table), and b) a table shows the information corresponding to the selected isomorphism, namely, number and type of nodes and edges of the molecules, common and non-common fragment.

Isomorphism results are automatically stored in the project database (see tree project menu) as we described in section 4.

Another panel is used for the visualization of the similarity analysis results. As is observed in Fig. 5 (upper right-hand panel), user can select any of the similarity indices for the similarity calculation. Results are shown in a linkage table to the other



○ Isomorphic molecules ○ Molecule 'i' is subgraph of 'j"
● MCS ○ CF

1 C17N3O2SF3

Results table

	1 C17N3O2SF3	2 C17N2O2SClF3	3 C6N2O2	4 C14N2O2S	5 C9N4O2
1 C17N3O2SF3	1.0	0.8666667	0.13207547	0.20689656	0.14529915
2 C17N2O2S...	0.8666667	1.0	0.13207547	0.20689656	0.15789473
3 C6N2O2	0.13207547	0.13207547	1.0	0.17948718	0.2413793
4 C14N2O2S	0.20689656	0.20689656	0.17948718	1.0	0.20930232
5 C9N4O2	0.14529915	0.15789473	0.2413793	0.20930232	1.0

Propiedad	Valor
Número átomos	19
Número enlaces	21
Enlace 1	Simple
Coordenadas	(-0.47365, 1.44375)
1.1 Átomo 1	C
Número atómico	6
1.2 Átomo 2	C
Número atómico	6
Enlace 2	Simple
Coordenadas	(0.24084999999999998, 1.44375)
2.1 Átomo 1	C
Número atómico	6
2.2 Átomo 6	C

Fig. 5. Isomorphism and Similarity perspective

panels (tree project menu, molecular graphs visualization and isomorphism data table). Results of similarity analysis can be saved by the user in external files using the corresponding option of the main menu or right-hand mouse button.

6 Discussion

Applications of Computer Science to Chemistry are playing an important role in the research advances. Much of these advances are based on the building and manipulation of large databases of chemicals compounds and the development of new algorithms for the classification, screening and analysis of the chemicals databases.

In this paper we have described a useful system for the management of chemical databases and the development of computational chemistry applications based on the similarity among chemical compounds using different isomorphism paradigms.

The systems described in this paper is able to work with some of the most popular relational database systems and it allows to the user the building of different projects using molecules from other projects, databases or external files. Manipulation of projects and molecules is guided by a windows environment in which different Perspectives (panel's sets with a related functionality) easy to the user the management and visualization of information.

The system includes the matching calculation among molecules using different isomorphism paradigms. Matching results can be visualized, stored and used for similarity analysis of projects, databases or molecules sets. The similarity calculation can be carried out using a wide set of similarity indices.

The modular characteristic of the system allows us the development of new computational chemistry solutions and to enlarge of the system with new Perspectives.

Nowadays, we are working for to include a new perspective for the calculation of molecular descriptors.

Acknowledgements. This work was supported by the Comisión Interministerial de Ciencia y Tecnología (CiCyT) and FEDER (Project: TIN2006-02071).

References

1. Young, D.: Computational Chemistry. A Practical Guide for Applying Techniques to Real World. John Wiley & Sons, Chichester (2004)
2. Cramer, C.J.: Essentials in Computational Chemistry. John Wiley & Sons, Chichester (2002)
3. Jensen, F.: Introduction to Computational Chemistry. John Wiley & Sons, Chichester (2007)
4. Van de Waterbeemd, H. (ed.): Structure-Property Correlations in Drug Research. Academic Press, Austin (1996)
5. Johnson, M.A., Maggiora, G.M. (eds.): Concepts and Applications of Molecular Similarity. John Wiley & Sons Ltd, Chichester (1990)
6. Willett, P.: Chemical Similarity Searching. J. Chem. Inf. Comput. Sci. 38, 983–996 (1998)
7. Hansch, C., Leo, A.: Exploring QSAR: Fundamentals and Applications in Chemistry and Biology. ACS Professional Reference Book, Washington DC (1995)
8. Irwin, J.J., Shoichet, B.K.: ZINC- a free database of commercially available compounds for virtual screening. J. Chem. Inf. Model. 45, 177–182 (2005)
9. Chemistry Development Kit, http://www.sourceforge.net/projects/cdk
10. Bioeclipse, http://www.bioclipse.net
11. ChemAxon Kft., http://www.chemaxon.com
12. Dalby, A., Nourse, J.G., Hounshell, W.D., Gushurst, A., Grier, D.L., Leland, B.A., Laufer, J.: Description of Several Chemical-Structure File Formats Used by Computer-Programs. J. Chem. Inf. Comput. Sci. 32, 244–255 (1992)
13. Structure, http://sourceforge.net/projects/structure
14. Cerruela García, G., Luque Ruiz, I., Gómez-Nieto, M.A.: Step-by-Step Calculation of All Maximum Common Substructures through a Constraint Satisfaction Based Algorithm. J. Chem. Inf. Comput. Sci. 44, 30–41 (2004)
15. Booch, G., Rumbaugh, J., Jacobson, I.: Unified Modeling Language User Guide, 2nd edn. Addison-Wesley Reading, Reading (2005)
16. Eclipse Software, http://www.eclipse.org