

Application of Chemoinformatics to the structural determination of natural compounds

José Luis López-Pérez¹, Roberto Therón², Esther del Olmo Fernández¹,
David Díez², José Fco. Adserias³

¹ Departamento de Química farmacéutica - Facultad de Farmacia

² Departamento de Informática y Automática - Facultad de Ciencias

³ Fundación Universidad de Salamanca

Universidad de Salamanca (Spain)

lopez@usal.es, theron@usal.es

Abstract. This paper describes the characteristics of a free web-based spectral database for the chemical research community, containing ¹³C NMR spectra data from more than 4000 natural compounds, and with a continuous increasing. This database allows flexible searching via chemical structure, substructure, name, and family of compounds, as well as spectral features as chemical shift, allowing the structural elucidation of known and unknown compounds by comparison of ¹³C NMR data.

Key words: structural elucidation, ¹³C NMR spectral database, natural compounds

1 Introduction

Chemoinformatics is the application of informatics methods to solve chemical problems [1]. All areas of chemistry can profit from the use of information technology and management, since both a deep chemical knowledge and the processing of a huge amount of information are needed. Particularly, the area of structure elucidation, that faces very complex problems where the use of spectroscopic information has to be made to elucidate the structure of a reaction product, provides highly interesting challenges for chemoinformatics practitioners. Indeed, this problem will become bigger since, in the future, a closer collaboration between bioinformatics and chemoinformatics specialists will be needed to solve the challenging problems faced in drug design [1].

Natural products are an important source of drugs today. The chemistry of natural products, just as other branches of organic chemistry, has experienced an excellent development in the last decades, as a consequence of the improvement of the different techniques. Natural compounds show great structural diversity, not only in their skeleton, but also in their functional groups of the different parts of the molecule. Therefore the number of described compounds is high, and it increases constantly [2]. In the natural products research, tedious purifications to isolation of constituents are often performed with the main purpose of structure

identification. If the structures of extract constituents were known in advance, the isolation efforts could be focused on truly novel and interesting components, avoiding reisolation of known or trivial constituents and increasing productivity [3].

In order to develop statistical machine learning methods in chemoinformatics, whether supervised or unsupervised, including predictive classification, regression and clustering of molecules and their properties, the need for large and well-annotated datasets has been already pointed out; furthermore, it is crucial to organize these datasets in rapidly searchable databases and to develop computational methods to rapidly extract or predict useful information for each molecule [4]. In this paper we present a web-based spectral database that tries to facilitate the structural identification of the natural compounds even previously to their purification.

^{13}C NMR spectroscopy is the more powerful tool that could be used in the identification and elucidation of natural products. This is largely due to the well-known and exquisite dependence of the ^{13}C chemical shift of each carbon atom on its local chemical environment and its number of attached protons. Furthermore, the highly resolved spectra, afforded by a large chemical shift range and narrow peak width, easily convert to highly reduced lists of chemical shift positions with minimal loss of information's peak intensity and width are features not generally used in dereplication. ^{13}C NMR spectroscopy also can provide the molecular formula. The analysis of spectral data for the unknown compound structure determination remains a usual and laborious task in chemical practice.

Since the advent of computers many efforts have been directed toward facilitating the solution to this problem [1]. Libraries of such spectral lists of data are common for synthetic organic compounds and are an invaluable tool for confirming the identity of known compounds [5]. In the field of natural products, where hundreds of thousands compounds have been reported in the literature, most compounds are absent from commercially available spectral libraries.

When a researcher in natural products isolates and purifies a compound he needs to know, as soon as possible, what skeleton it is, or, even better, its structure, and also if the compound has been described previously. Searching databases allows quick identification of previously registered compounds and can provide insight for the unknown compounds.

2 Structural elucidation trough the web

The structural elucidation of natural products is a great challenge due to its great structural diversity and complexity. For this reason, a database with information on ^{13}C NMR spectra, accessible through a standard browser, is being developed (<http://c13.usal.es>). Currently it contains the structures of several thousands compounds, along with their ^{13}C NMR information with a continuous increasing, because this database has capacity for storing hundreds of thousands compounds. The utility of the database will be conditioned on the number of

entries created and, certainly, on the quality of the spectral information of the entered compounds.

The database has many search facilities and an appearance that allows the comparative study of related compounds. At present, new search tools are being developed and the data input methods are being improved in order to allow researchers from different institutions to enter the information through the Net. The aim of this database is to help identifying or elucidating the structure of a hypothetical new compound, by comparing its ^{13}C NMR data with those related already published. Several tools that will facilitate this task have been developed:

- search by substructure in a graphics environment,
- search by chemical shift,
- both previous search methods combined,
- search refinements,
- results displayed in different layouts, in order to make a comparative study,
- deviation calculus in fixed positions, etc.

This tool was born with the intention to be cooperative, to extend with contributions of researchers, who will enter the information of their own compounds, so that the information can be shared through the Net. Also, we have developed scripts to automatically parse input data, run different tests and populate the database.

2.1 Implementation

To build this database we have used MySQL (<http://www.mysql.com>), one of the leading open-source relational database managers, based in SQL (Structured Query Language). This manager is characterized by being fast, multi-thread, multi-user and robust. MySQL server controls the access to the information, to make sure that several users can work at the same time. We use the open-source Apache Web server and JSP to create Web pages that show contents generated dynamically. Through a correct programming with JSP we achieve the communication between the applets and the database. These tools are those used for systems where the speed and the number of access at the same time is a fundamental feature and the security is not so important.

2.2 Database schema and data format

The basic database schema is relationally organized and the molecular structures are defined and stored in the database with SMILES (Simplified Molecular Input Line Entry Specification) [6][7] code. This format of structural specification, that uses one line notation, is widely used for sharing chemical structure information over Internet [1]. The substructural searches are performed by SMARTS code (SMILES Arbitrary Target Specification), a variation of the SMILES code. While SMILES defines the molecules in the form of alphanumeric chains, that enable an easy manipulation, SMARTS is a more complex code. It uses a Boolean operator that allows choosing all-purpose atoms, groups of alternative atoms, donor and acceptor groups of hydrogen bonds or lyophilise atoms.

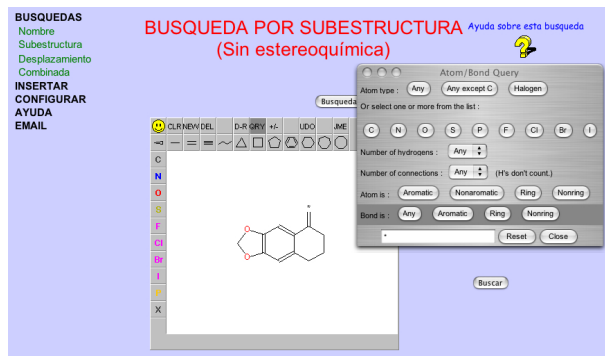


Fig. 1. Substructure searching using the JME editor

2.3 Web interface

The search speed makes possible for the users to enter query molecules through a Web interface. In order to capture the questions to the database graphically, JME⁴, a structure editor which enables to draw several fragments not related to each other (see Figure 1), has been used. This amplifies and makes the search more selective. This editor is an embedded Java applet which allows to match a 2D query against the structures and then display the structures found via the Net.

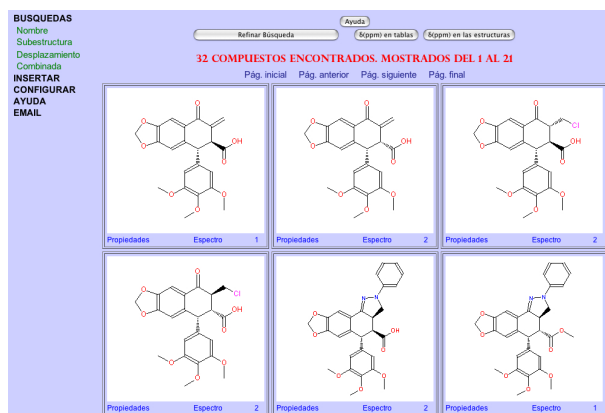


Fig. 2. The search results display

⁴ JME, version 1.2. The authors wish to thank Dr. Peter Ertl the consent the use of this applet as a tool in this database with non-profit purposes.

The JME molecular editor has a palette that speeds up the creation of structures and uses the IUPAC recommendations for depicting the stereochemistry. Using this palette it is possible to add preformed substructures like different size cycles, aromatic rings, simple and multiple bonds, frequently used atoms and also, using the control panel, it is possible to directly enter functional groups, like carboxyl acids, nitro groups and other groups, for example *tert*-butyl, etc. All these facilities enable to generate a new structure rapidly and to speed up the search process.

2.4 Displaying the results

As it can be seen in Figure 2, the search results are displayed graphically. The example in Figure 2 shows the results for the query substructure drawn in Figure 1. These results can be seen individually in more detail, showing the chemical shift of every carbon of the molecule and additional miscellaneous information, such as literature references, molecular formula, family, type, disolvent, etc (the details for the upper-left result of Figure 2 are shown in Figure 3).

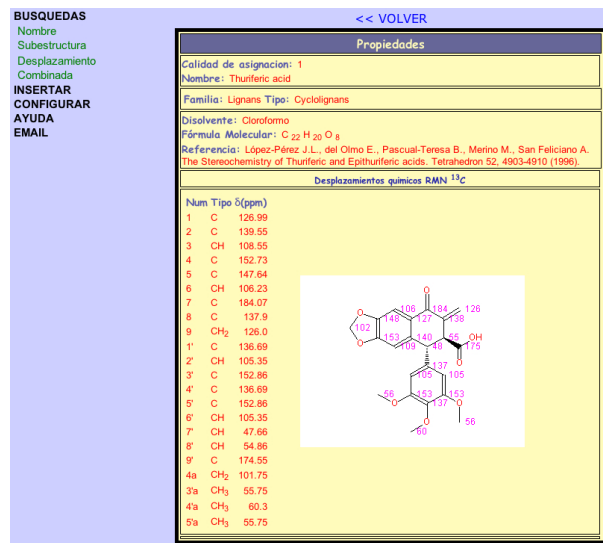


Fig. 3. Detailed chart of a compound found in the search

One script calculates and represents the ¹³C NMR spectra (Figure 4 shows the ¹³C NMR spectrum for the substance of Figure 3) of the compound in a very similar way of the experimental data obtained, that is the proton decoupled (broad band) and the DEPT (Distortionless Enhancement by Polarization Transfer). Another script also calculates and represents the signals corresponding to the deuterated solvent used in the experiment.

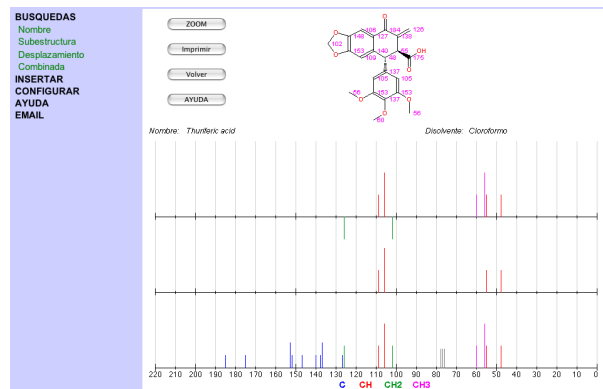


Fig. 4. Chart of the ^{13}C NMR spectra of a substance

3 Substructure search

The best method for locating structures with particular functionality is through substructure searching. Loosely defined, substructure is a particular combination of atoms or functional groups that makes up part of a larger structure. Substructure searching is very useful in locating related compounds reported in the literature that contain the substructures in which we are interested [8]. Substructure searches can be made more or less specific with the addition of the Boolean operators AND, OR, and NOT. In addition, the degree of search can be specified. A substructure search can be so specific as to define a substructure as a single atom or general enough to specify a substructure consisting of several functional groups. The smaller the restricted fragment, the bigger the number of compounds found in the database.

Frequently, from the spectroscopic analysis of several two-dimensional experiments of NMR, like COSY, HMQC (Heteronuclear Multiple Quantum Coherence), HMBC (Heteronuclear Multiple Bond Coherence), we can infer the presence of structural fragments. These fragments can be drawn with the editor and then all the compounds that fulfill simultaneously those requirements can be searched for. We can also choose to specify the stereochemistry.

3.1 Search by chemical shift

The database offers several alternatives in the search process, as it is possible to determine or not determine the carbon position in the molecule (see Figure 5), to carry out iterative search and, finally, a combined search that will allow combining all the mentioned searches.

The researcher can enter the chemical shifts obtained in the different NMR experiments, together with the carbon's hybridization type. The system permits to carry out the enquiry with the required number of carbons, from one carbon

BUSQUEDAS
 Nombre
 Subestructura
 Desplazamiento
 Sin posición
 Con posición
 Iterativa
 Vecindad
 Combinada
INSERTAR
CONFIGURAR
AYUDA
EMAIL

BUSQUEDA POR DESPLAZAMIENTO Ayuda sobre esta búsqueda
sin posición

Busqueda sin tolerancia Buscar

Familia: _____
 Tipo: _____
 Grupo: _____

Desplazamiento	Tipo de Carbono	Tolerancia
128.1	<input type="radio"/> C <input type="radio"/> CH <input type="radio"/> CH ₂ <input type="radio"/> CH ₃	3
175.0	<input type="radio"/> C <input type="radio"/> CH <input type="radio"/> CH ₂ <input type="radio"/> CH ₃	1
184	<input type="radio"/> C <input type="radio"/> CH <input type="radio"/> CH ₂ <input type="radio"/> CH ₃	1
126	<input type="radio"/> C <input type="radio"/> CH <input type="radio"/> CH ₂ <input type="radio"/> CH ₃	1
53	<input type="radio"/> C <input type="radio"/> CH <input type="radio"/> CH ₂ <input type="radio"/> CH ₃	1

Nuevo desplazamiento
 Eliminar desplazamiento

Fig. 5. Chemical shift search without carbon position

to the totality of the compound's carbons. It is possible to specify the required deviation (+/-), by default for all carbons, or to limit in a detailed way each of them. This will limit the search distinctly and, therefore, the researcher will obtain a reasonable and manageable number of compounds. The search based in the most significant carbons of the studied compound will lead to the determination of the family it belongs to.

If the skeleton of the studied substance is known, and if some distinctive chemical shifts of the most important signals are also available, a search by shifts in each particular position of the molecule can be carried out. Therefore the researcher will only obtain the compounds of the family whose shifts, in those positions, match with those of the problem compound.

It is also possible to carry out a combined and simultaneous search by substructure and by chemical shifts, a feature that undoubtedly amplifies the search capacity and increases the possibilities of finding, at least, compounds related with the problem substance.

The iterative search is probably the most particular and specific search of this application. It is possible to include in the search from one only chemical shift, to the totality of the signals of the studied compound. This tool will initially carry out a search by all the entered chemical shifts introduced. If it does not find any compound that does not fulfill the established requirements, it will complete a search by all the shifts except one iteratively. It will perform all the possible combinations until it finds a compound that fulfills some of the requirements.

3.2 Search by proximity

It is a very useful to search by chemical shift of signals that appear to be correlated in HMBC and HMQC experiments, because this type of search enables to limit the number of bounds between the carbon atoms, to which will belong the considered chemical shifts (see Figure 6).

If the number of found compounds is too big and difficult to manage, it can be possible to carry out a search among the compounds obtained in the previous search. We therefore limit the result to a reasonable dimension that will enable

BUSQUEDAS
 Nombre
 Subestructura
 Desplazamiento
 Sin posición
 Con posición
 Iterativa
 Vecindad
 Combinada
INSERTAR
CONFIGURAR
AYUDA
EMAIL

BUSQUEDA POR VECINDAD Ayuda sobre esta búsqueda

Familia:
 Tipo:
 Grupo:
 Distancia máxima: 3.7 amgstron

Desplazamiento	Tipo de Carbono	Tolerancia
128.1	<input type="radio"/> C <input type="radio"/> CH <input type="radio"/> CH2 <input type="radio"/> CH3	3
175.0	<input type="radio"/> C <input type="radio"/> CH <input type="radio"/> CH2 <input type="radio"/> CH3	1

Fig. 6. Proximity search

to find, if not the target compound, a series of molecules closely related to it. By means of them, we will then be able to deduce a structure for the problem compound.

Evidently, the type of patterns and notations that are used, both SMARTS and SMILES, are too complex to be interpreted by organic chemists without specific training in this area. That is the reason why we use a tool able to convert those notations into a graph that represents a substructure that will act as question. This is the task that carries out the applet used in this database.

Conclusions and further work

The development of the presented database, that provides an excellent tool to face complex chemical problems such as structure elucidation, required the joint efforts of information science and chemistry specialists. The excellent results obtained are a good reason to expect success with novel approaches for current research challenges, as the field of chemoinformatics matures and a closer collaboration with bioinformatics is developed.

Currently we are working on the following: increase the number of stored compounds; add more searches such as the *hot spot* search; use information visualization techniques that will give more insight in the analysis process; include supervised and unsupervised machine learning methods that will lead to interesting predictions for the different substructures.

Acknowledgements

We would like to acknowledge Carolina Smith de la Fuente for her help with the translation.

References

1. Gasteiger, J.: Chemoinformatics: a new field with a long tradition. *Analytical and Bioanalytical Chemistry* **384** (2006) 57–64

2. Buckingham, J.: Dictionary of Natural Products. Chapman & Hall/CRC Press (2000)
3. Clarkson, C., Stærk, D., Hansen, S.H., Smith, P.J., Jaroszewski, J.W.: Discovering new natural products directly from crude extracts by hplc-spe-nmr: Chinane diterpenes in harpagophytum procumbens. *Journal of Natural Products* **69** (2006) 527–530
4. Chen, J., Swamidass, S.J., Dou, Y., Bruand, J., Baldi, P.: Chemdb: a public database of small molecules and related chemoinformatics resources. *Bioinformatics* **21** (2005) 4122–4139
5. Robien, W.: Nmr data correlation with chemical structure. In v. R. Schleyer, P., Allinger, N.L., Clark, T., Gasteiger, J., Kollman, P.A., III, H.F.S., Schreiner, P.R., eds.: *Encyclopedia of Computational Chemistry*. Volume 3. John Wiley & Sons, Limited, Chichester, England (1998) 1845–1857
6. Weininger, D.: Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28** (1988) 31–36
7. Weininger, D., Weininger, A., Weininger, J.L.: Smiles. 2. algorithm for generation of unique smiles notation. *J. Chem. Inf. Comput. Sci.* **29** (1989) 97–101
8. Kochev, N., Monev, V., Bangov, I.: Searching Chemical Structures. In: *Chemoinformatics: A textbook*. Wiley-VCH (2003) 291–318