An Ontological Approach to Represent Molecular Structure Information

Eva Armengol and Enric Plaza

IIIA - Artificial Intelligence Research Institute,
CSIC - Spanish Council for Scientific Research,
Campus UAB, 08193 Bellaterra, Catalonia (Spain).
{eva, enric}@iiia.csic.es,

Abstract. Current approaches using Artificial Intelligence techniques applied to chemistry use representations inherited from existing tools. These tools describe chemical compounds with a set of structure-activity relationship (SAR) descriptors because they were developed mainly for the task of drug design. We propose an ontology based on the chemical nomenclature as a way to capture the concepts commonly used by chemists in describing molecular structure of the compounds. In this paper we formally specify the concepts and relationships of the chemical nomenclature in a comprehensive ontology using a form of relational representation called *feature terms*. We also provide several examples of describing chemical compounds using this ontology and compare our proposal with other SAR based approaches.

1 Introduction

The IUPAC (www.chem.qmul.ac.uk/iupac/) chemical nomenclature is a standard form to describe the (organic and inorganic) molecules from their chemical structure. In Artificial Intelligence some proposed representations of the molecules describe them atom by atom obtaining cumbersome descriptions that may be not easily understandable by chemists. From our point of view, a formal representation using the IUPAC nomenclature could be very useful since allows a direct description of the chemical structure, in a way very familiar to the chemist. For instance, chemists commonly describe the *anthracene* as a molecule formed by a group of three benzenes and they know some of its properties and the relative position of each atom. However, representations describing the molecules atom by atom do not take into account expert knowledge; therefore they need explicitly represent the 14 atoms of the *anthracene*, their bindings, interactions, etc. We propose and ontological approach to represent information about the molecular structure of a chemical compound. In this approach, a compound can be described as *anthracene* without any reference to individual atoms.

In the next section we briefly explain the chemical nomenclature and how the representation we propose capture this nomenclature. Then, in section 3 we explain a formal representation called *feature terms* and how the chemical compounds can be described using them. Finally, we discuss trade offs our proposal by comparing it with other SAR based approaches.

2 Chemical nomenclature concepts

Following the recommendations of the IUPAC (1994) the organic compounds can be classified in four groups: 1) Based on Carbon, Hydrogen and Oxygen, 2) Based on other elements, 3) Natural products (antibiotics, lipids, nucleic acids, etc), and 4) Others.

We focus on compounds belonging to the groups 1 and 2 above, because we consider them as most elemental than the compounds included in the groups 3 and 4 in the sense that often compounds belonging to the two last groups are either extensions (lipids are chains of hydrocarbons) or particular cases (ions may be parts of functional groups) of compounds in groups 1 and 2.

Compounds included in the first group are the hydrocarbons, ring systems, alcohols, ethers, phenols and derivatives, aldehydes, ketones, quinones and derivatives, and carboxilic acids and derivatives. The second group includes compounds that are based on elements such as nitrogen, phosphorus, silicon, sulfur, halogens, metals, etc. Notice that these compounds could also be regrouped in two different classes taking into account whether they can be found alone (such as the hydrocarbons or the ring systems) or not (alcohols, ethers, aldehydes, etc). Therefore, we consider the following alternative classification of the first two groups of compounds above: a) Hydrocarbons, b) Ring systems, and c)Functional groups. In the following subsections we will analyze these groups separately.

2.1 Hydrocarbons

The *hydrocarbons* (also called *alkanes*) are chains of atoms that only contain carbon (C) and hydrogen (H). According to the number of C atoms of the chain they take different names (Fig. 1 shows some hydrocarbons). The left part of the figure shows hydrocarbons called *saturated* since all the bonds are single, i.e. all the C atoms (except those in the extremes) are bonded to two H and to two other C atoms. The right part of Fig. 1 shows *unsaturated* hydrocarbons (also called *alkenes* and *alkynes*) since some of the bonds are either double (*alkenes*) or triple (*alkynes*). For instance, the *1,3-butadiene* is an unsaturated hydrocarbon with two double bonds, one in position 1 and another in position 3. Notice that in the nomenclature of unsaturated hydrocarbons the position of the double and

methane: CH_4	ethyne: $CH - CH$
ethane: $CH_3 - CH_3$	ethene: $CH_3 = CH_3$
propane: $CH_3 - CH_2 - CH_3$	propyne: $CH \equiv C - CH_3$
butane: $CH_3 - CH_2 - CH_2 - CH_3$	1,3- butadiene: $CH = CH - CH = CH_2$
pentane: $CH_3 - CH_2 - CH_2 - CH_2 - CH_3$	4-hexen, 1-yne: $CH \equiv C - CH_2 - CH = CH - CH_3$
saturated HC	unsaturated HC

Fig. 1. Examples of acyclic saturated and unsaturated hydrocarbons.

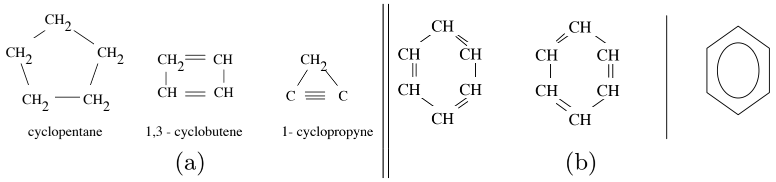


Fig. 2. (a) Three examples of cyclic saturated and unsaturated hydrocarbons. (b) The two resonant forms of the benzene are shown on the left, while the usual representation of the benzene is shown on the right.

triple bonds is part of the compound name. Also, the suffix *-ene* (or *-en*) denotes hydrocarbons with double bonds whereas those with triple bonds have the suffix *-yne*. The name of compounds with double and triple bonds have both suffixes (as the *4-hexen, 1-yne* shown on the bottom right part of Fig. 1).

Moreover, both saturated and unsaturated hydrocarbons may be cyclic (i.e. *cycloalkanes*). Figure 2a shows some examples of saturated and unsaturated cyclic hydrocarbons. Concerning to the nomenclature, the name of cyclic hydrocarbons is the same that the name of the acyclic hydrocarbons preceded by the prefix *cyclo-* (i.e. cyclopentane; 1,3-cyclobutene).

2.2 Ring systems

Ring systems include the cyclic hydrocarbons called *aromatic* or *arenes*. Aromatic rings are defined as those rings where the electrons are free to cycle around circular arrangements of atoms, which are alternately simply and doubly bonded. A typical example of aromatic rings is the *benzene*, that is a cyclohexane that can take the two forms show in the left part of Fig. 2b. In fact, since both forms of the benzene are equivalent (this is called *resonance*) and none of them accurately represents the benzene structure, the most common representation for

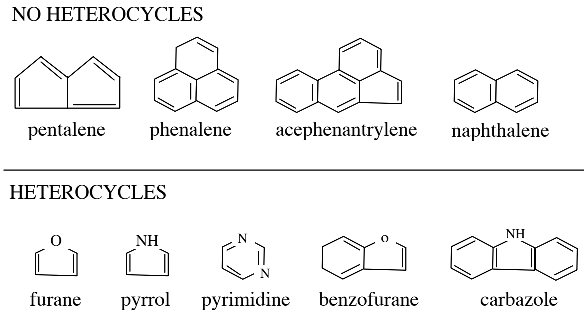


Fig. 3. Examples of both heterocyclic and no heterocyclic ring systems.

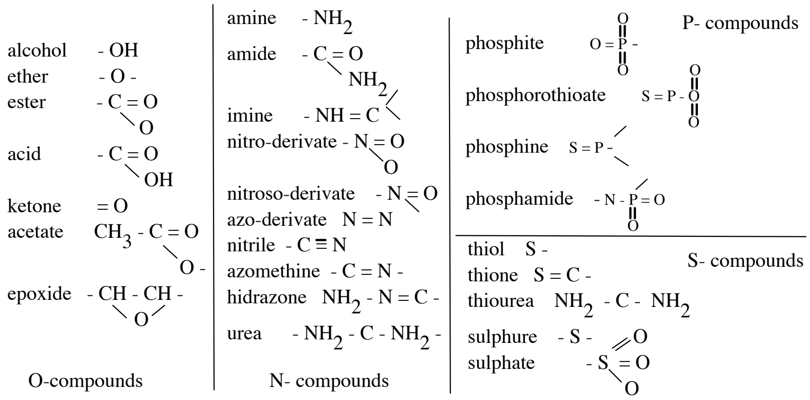


Fig. 4. Functional groups.

the benzene is that shown at the right part of Fig. 2b. In this representation the double bonds have been replaced by a central circle meaning that the electrons have a free circulation among the atoms.

Ring systems could be *monocycles* or *polycycles*. Monocycles are those systems formed by only one ring; polycycles are those formed from the association of several cyclic hydrocarbons. The upper part of Fig. 3 shows a sample of monocycles and polycycles. The most common ring systems are based on the benzene. Both monocycles and polycycles can be classified as *heterocycles* (when all the atoms are C and H) or as *no heterocycles* (when there is some atom different of C). The bottom part of Fig. 3 shows some examples of heterocycles.

2.3 Functional groups

A *functional group* is an atom or group of atoms that replaces one H atom in an organic compound and that defines the structure of a family of compounds and determines the chemical properties of that family. Based on the atoms they contain, we propose to classify the functional groups as follows: 1) *O-compounds* based on the oxygen, 2) *N-compounds* based on the nitrogen, 3) *P-compounds* based on the phosphorus, and 4) *S-compounds* based on the sulphur.

Some functional groups could be classified as belonging to more than one class since they may contain more than one atom different from H (for instance, oxygen and nitrogen). In such situations, we considered them as belonging to one class depending on which atom is considered as the most important of the functional group. Figure 4 shows some of the functional groups that we considered. Notice that the amide, the nitro-derivate and the nitroso-derivate could be considered both O-compounds or N-compounds.

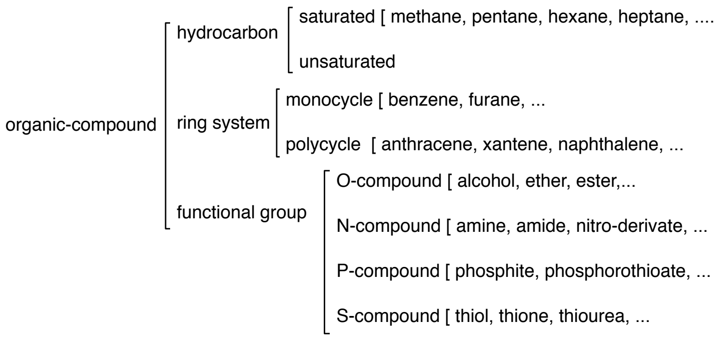

organic-compound
hydrocarbon
saturated [methane, pentane, hexane, heptane, ...
unsaturated
ring system
monocycle [benzene, furane, ...
polycycle [anthracene, xantene, naphthalene, ...
functional group
O-compound [alcohol, ether, ester,...
N-compound [amine, amide, nitro-derivate, ...
P-compound [phosphite, phosphorothioate, ...
S-compound [thiol, thione, thiourea, ...

Fig. 5. The hierarchy of sorts for organic chemistry concepts.

3 Representation of the chemical compounds

In this section we introduce a formal specification of the chemical ontology based on feature terms. Several Artificial Intelligence approaches describe complex objects using a *relational* representation. In this kind of representation, an object is described by its parts and the relations among these parts. In particular, we use a relational representation called *feature terms*. Using feature terms, the concepts explained in the previous section have been specified by a hierarchy of *sorts* (Fig. 5). Moreover, a sort is described by a set of features where each feature represents a relation of this sort with another sort. In the next section feature terms are briefly introduced and then, in section 3.2, we explain how the chemical compounds can be described using feature terms.

3.1 Feature Terms

Feature Terms (also called feature structures or ψ-terms) are a generalization of first order terms. The difference between feature terms and first order terms is the following: a first order term, e.g. $f(x, y, g(x, y))$ can be formally described as a tree and a fixed tree-traversal order. In other words, parameters are identified by position. The intuition behind a feature term is that it can be described as a labelled graph, i.e. parameters are identified by name. For instance, the definition of a particular object using feature terms is the following:

```
(define (sort object-name)
    (feature-1 obj-1)
    ....
    (feature-N obj-N))
```

where feature-1,..., feature-N are the names of the features that describe the object *object-name*. The object and also the values of the features belong to a *sort* and sorts are related among them by a hierarchy of sort/subsorts. Figure 5 shows the hierarchy of sorts we define to capture the chemical concepts that will be introduced later. The definition of a sort is as follows:

```
(define-sort sort
    (feature-1 sort-1)
    (.... )
    (feature-N sort-N))
```

where *sort* is the name of the sort that we are defining and feature-1,..., feature-N are the names of the features that describe the objects belonging to *sort*. When a *sort-i* is a subsort of another sort *sort-j* this is defined as follows: (define (*sort-j* *sort-i*) ...), and *sort-i* inherits all the features of *sort-j*. For instance, Fig. 5 shows that *benzene* is a subsort of *monocycle* that, in turn, is a subsort of *ring system* that, in turn, is a subsort of *organic-compound*. The values of the features (e.g. feature-1) are restricted to the sort that is declared (e.g. sort-1) .

A more detailed explanation about the feature terms and the subsumption relation can be found in [1]. In the next section we explain how feature terms are used to represent chemical compounds. Also, we detail the sort hierarchy that represents the chemical concepts introduced in the previous section.

3.2 Chemical compounds described as feature terms

A chemical compound is described by a feature term of sort *chemical-compound* with features characterizing the compound. The definition of the sort *chemical-compound* is the following:

```
(define-sort chemical-compound
    (molecular-structure compound)
    (tests test-results))
```

Feature terms of sort chemical-compound are described by two features: the molecular-structure of the compound and the tests features that contains the results of some tests done on the compound. Notice that the value of molecular-structure has to be an object of sort *compound*. In this section we focus on the explanation of the representation of the molecular structure of the compounds.

Our ontology proposal is based on the chemical nomenclature but we also want to describe the molecular structure as accurately as possible. Nevertheless the nomenclature has some ambiguities since some compounds may have several synonym names. This means that in our ontology a compound can be described in several ways. To handle the synonyms of a compound we use the notion of *multi-instance* [2]. When a compound has synonym descriptions the only difference is that the feature molecular-structure from the sort *chemical-compound* is a set that contains all the possible synonym descriptions of that compound.

To describe the molecular structure of a compound, we defined the sort *compound* which has, in turn, two subsorts: *organic-compound* and *inorganic-compound*. The specification of the *organic-compound* sort is the following:

```
(define-sort (compound organic-compound)
      (main-group compound)
      (radical-set compound))
```

Organic compounds can be described as composed by two parts: the main group and the radical-set both with values of sort *compound*. The main group of a molecule is often the part of the molecule that is either the largest or the part located in a central position. Radicals are groups that are usually smaller than the main group (commonly they are functional groups). A main group can contain several radicals and a radical can, in turn, have a set of radicals. Both main group and radicals are the same kind of molecules, i.e. a molecule may appear as the main group in a compound and also as a radical in another compound.

Let us analyze now how to represent the different kinds of chemical compounds following the classification introduced in section 2.

Hydrocarbons Although there are saturated and unsaturated hydrocarbons, their nomenclature follows the same idea: the basic name of the hydrocarbon is the number of C atoms. The name of unsaturated hydrocarbons has the suffix *-ene* when there are double bonds and *-yne* when there are triple bonds. When (saturated or unsaturated) hydrocarbons are cyclic the prefix (*cyclo* is added to the basic name. Using feature terms we define the sort *hydrocarbon* as a subsort of *compound* as follows:

```
(define-sort (organic-compound hydrocarbon)
      (cyclic? boolean)
      (p-radicals position-radicals))
```

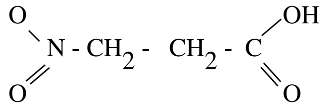
Since *hydrocarbon* is a subsort of *organic-compound*, it inherits the features main-group and radical-set. When cyclic? is *true* means that the hydrocarbon is cyclic otherwise it is acyclic. The sorts *saturated-hydrocarbon* and *unsaturated-hydrocarbon* are subsorts of *hydrocarbon* so they inherit the features cyclic? and p-radicals.

The radicals of a compound are situated in a determined position with respect to the main group. This is represented with the sort *position-radical* as follows:

```
(define-sort position-radical
      (position numeric)
      (radicals compound))
```

Figure 6 shows the saturated hydrocarbon called *3-nitropropionic acid* that is a compound that has a propane (i.e. an hydrocarbon with three C atoms) and two radicals: a *nitro-derivate* in position 3 and an *acid* in position 1. The figure also shows the description of the *3-nitropropionic acid* using feature terms.

To represent unsaturated hydrocarbons the *unsaturated-hydrocarbon* sort is defined as follows:



```
(define (saturated-hydrocarbon 3-nitropropionic-acid)
    (main-group propane)
    (radical-set acid nitro-derivate)
    (p-radicals (define (position-radical)
                    (position one)
                    (radicals acid))
                (define (position-radical)
                    (position three)
                    (radicals nitro-derivate))))
```

Fig. 6. Molecular structure of the *3-nitropropionic acid* and its representation using feature terms.

```
(define-sort (hydrocarbon unsaturated-hydrocarbon)
    (main-group saturated-hydrocarbon)
    (p-bonds p-bond))
```

Notice that in this definition the sort of main-group is a *saturated-hydrocarbon*. Also, the *unsaturated-hydrocarbon* sort has a feature called p-bonds which values are of the sort *p-bond* defined as follows:

```
(define-sort p-bond
    (bond kind-of-bond)
    (position numeric))
```

i.e. by means of this sort we can define the kind of bonds of an unsaturated hydrocarbon and its position. For instance, the representation of the *4-hexen, 1-yne* shown in Fig. 1 is the following:

```
(define (unsaturated-hydrocarbon 4-hexen-1-yne)
    (cyclic? false))
    (main-group hexane))
    (p-bonds (define (p-bond)
                 (bond triple)
                 (position one))
             (define (p-bond)
                 (bond double)
                 (position four))))
```

This description comes directly from the chemical name that states that the double bond is in position 4 and the triple bond in position 1, that is to say, the C atoms are numerated from left to right.

Ring systems are also defined as composed of a main group and radicals. When compounds have only one ring system then this ring system is the main-group. The problem, however, is how to determine the position of the radicals, since

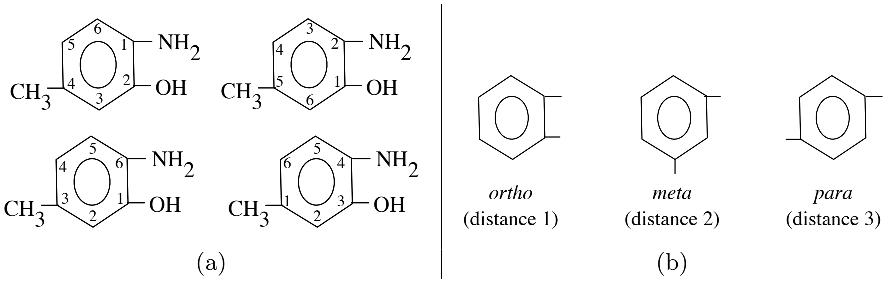


Fig. 7. a) Several possible numerings of the radicals in a molecule. b) Names used in chemistry for the relative positions of the radicals of a benzene ring.

(without taking into account the nomenclature rules) the position 1 could be any of the radicals; once that position is fixed, the position 2 could be determined clockwise or counter clockwise. Figure 7a shows an example on how the positions of the radicals can change depending on which radical is considered to be in position 1.

In chemistry, when the main group is a benzene, some positions of the radicals (Fig. 7b) have particular names (*ortho, meta, para*). We take this idea for defining the positions of the radicals of a ring system. Thus, the *ring-system* sort is defined as a subsort of *organic-compound* as follows:

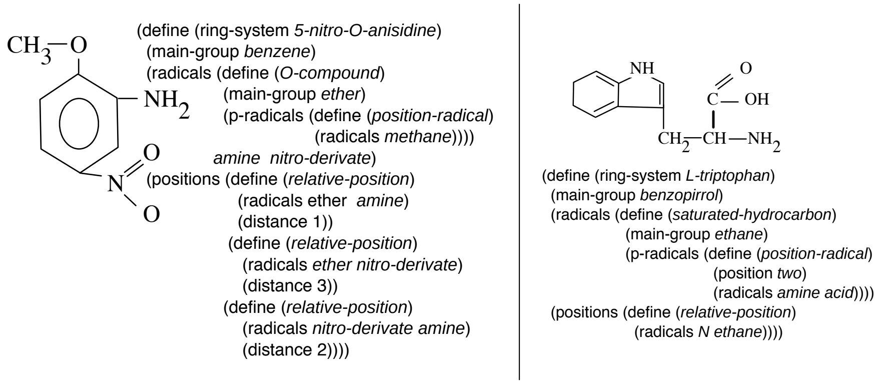
```
(define-sort (organic-compound ring-system)
    (radicals compound)
    (positions position))
```

where the sort *position* represents the positions of the radicals. We defined three subsorts of *Position*: 1) *absolute-position*, 2) *relative-position* and 3) *atom-position*. The sort *absolute-position* will be used in compounds where the positions of the radicals are straightforward (as in the hydrocarbons). The sort *relative-position* is used when the position of the radicals are defined by their distance (as in the positions *ortho, meta* and *para*). The sort *relative-position* is defined as follows:

```
(define-sort (position relative-position)
    (radicals compound)
    (distance number))
```

The sort *atom-position* is used when a radical is placed in a particular atom different of the C. The description of a ring system can contain the three kinds of positions. Figure 8 shows two examples of ring systems with radicals and their descriptions using feature terms.

Functional groups The sort *functional-group* is a subsort of *organic-compound* defined as follows:



```
(define (ring-system 5-nitro-O-anisidine)
  (main-group benzene)
  (radicals (define (O-compound)
                (main-group ether)
                (p-radicals (define (position-radical)
                                (radicals methane))))
              amine  nitro-derivate)
  (positions (define (relative-position)
                 (radicals ether  amine)
                 (distance 1))
               (define (relative-position)
                  (radicals ether nitro-derivate)
                  (distance 3))
               (define (relative-position)
                  (radicals nitro-derivate amine)
                  (distance 2))))
```

```
(define (ring-system L-triptophan)
  (main-group benzopirrol)
  (radicals (define (saturated-hydrocarbon)
                (main-group ethane)
                (p-radicals (define (position-radical)
                                (position two)
                                (radicals amine acid)))))
  (positions (define (relative-position)
                 (radicals N ethane))))
```

Fig. 8. Two examples of ring systems and their descriptions in our ontology.

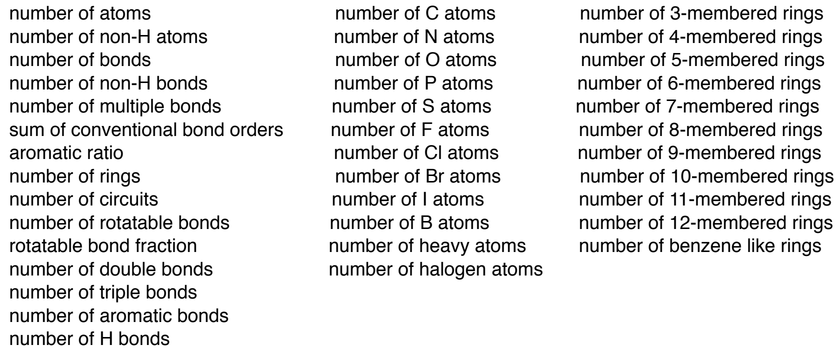
```
(define-sort (organic-compound functional-group)
      (radical-set compound)
      (p-radicals position-radical))
```

In turn, the sort *functional-group* has four subsorts: *O-compound, N-compound, S-compound* and *P-compound*. These sorts inherit the features of the *functional-group* sort, i.e. main-group and radical-set.

4 Discussion

Most Machine Learning (ML) tools used to build models of Toxicology are based on the *Structure-Activity Relationship (SAR)* descriptors. These descriptors represent the chemical compounds from several points of view (structural, physical properties, etc) and they are the basis to build equational models that relate the structure of a chemical compound with its physical-chemical properties. There is a number of commercial tools allowing the generation of these descriptors (CODESSA [5], TSAR (www.accelrys.com/products/tsar/), DRAGON [6], etc) and each one gives their own set of descriptors. Thus, methods that build toxicity models have to select a subset of these descriptors. As a consequence, the final model and, therefore its performance, will depend on which descriptors have been considered as the most important.

The main difference among the representations based on SAR and our ontological approach is that the former describe the molecular structure of the chemical compounds in an exhaustive way. SAR representations consist on a set of descriptors that can be grouped in several subsets according to the characteristics they describe. Thus, there are constitutional descriptions that capture structural features (Fig. 9 shows such descriptors), topological descriptors that capture 2D features, connectivity indices, WHIM descriptors, etc. Therefore, the description of a compound using SAR descriptors consists on giving a value for



number of atoms
number of non-H atoms
number of bonds
number of non-H bonds
number of multiple bonds
sum of conventional bond orders
aromatic ratio
number of rings
number of circuits
number of rotatable bonds
rotatable bond fraction
number of double bonds
number of triple bonds
number of aromatic bonds
number of H bonds

number of C atoms
number of N atoms
number of O atoms
number of P atoms
number of S atoms
number of F atoms
number of Cl atoms
number of Br atoms
number of I atoms
number of B atoms
number of heavy atoms
number of halogen atoms

number of 3-membered rings
number of 4-membered rings
number of 5-membered rings
number of 6-membered rings
number of 7-membered rings
number of 8-membered rings
number of 9-membered rings
number of 10-membered rings
number of 11-membered rings
number of 12-membered rings
number of benzene like rings

Fig. 9. Constitutional descriptors used in representations based on SAR.

each descriptor. Notice that the descriptions of compounds based on SAR are vectors of attribute values, a very simple representation from which a comprehensive chemical ontology cannot be directly derived.

The representation we propose is more conceptual than SAR in the sense that directly uses the concepts understood by the chemists. We defined a chemical ontology with the chemical concepts in such a way that the molecular structure of a compound can be described using those concepts (Fig. 5). Thus, when our ontology describes a compound as formed by a benzene and an acid, chemists clearly understand the underlying structure of this compound, whereas using constitutional descriptors this compound should be described as composed of 22 atoms, 9 non-H atoms, 1 ring, 2 O atoms, etc. Therefore the molecular description using the ontology we propose is more understandable than descriptions based on SAR. Moreover, experimental evidence [3, 4] shows that the predictive performance of our approach is comparable to that of the approaches using representations based on SAR (although our ontology only incorporates structural information).

Some authors use approaches that are not centered on the representation of specific atoms but on molecular structures. For instance, González et al [7] and Deshpande and Karypis [8] represent chemical compounds as labeled graphs, using graph techniques to detect the set of molecular substructures (subgraphs) more frequently occurring in the chemical compounds of the data set. Conceptually, these two approaches are related to ours in that we describe a chemical compound in terms of its radicals (i.e. substructures of the main group).

Currently, we defined an ontology that only takes into account the structural aspects described by the constitutional descriptors of the SAR representations. In the future, we plan to extend this ontology with some other aspects of the chemical compounds that could be useful for predictive toxicology. Thus, our goal is not simply incorporating all SAR descriptors into the ontology, but rather developing a chemical ontology that captures the necessary molecular information.

Acknowledgements This work has been supported by the MCYT-FEDER Project SAMAP (TIC2002-04146-C05-01). The authors thank Dr. Lluis Bonamusa for his assistance in developing the representation of chemical compounds.

References

1. Armengol, E., Plaza, E.: Bottom-up induction of feature terms. Machine Learning **41** (2000) 259–294
2. Dietterich, T., Lathrop, R., Lozano-Perez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence Journal **89** (1997) 31–71
3. Armengol, E., Plaza, E.: Relational case-based reasoning for carcinogenic activity prediction. Artificial Intelligence Review **20** (2003) 121–141
4. Armengol, E., Plaza, E.: Lazy learning for predictive toxicology based on a chemical ontology. In Dubitzky, W., Azuaje, F., eds.: Artificial Intelligence Methods and Tools for Systems Biology, In Press. Kluwer Academic Press (2004)
5. Katritzky, A., Petrukhin, R., Yang, H., Karelson, M.: CODESSA PRO. User's Manual. University of Florida (2002)
6. Todeschini, R., Consonni, V.: Handbook of Molecular Descriptors. Methods and principles in Medicinal Chemistry. Wiley-VCH, Weinheim (2000)
7. Gonzalez, J., Holder, L., Cook, D.: Application of graph-based concept learning to the predictive toxicology domain. In: Proceedings of the Predictive Toxicology Challenge Workshop, Freiburg, Germany. (2001)
8. Deshpande, M., Karypis, G.: Automated approaches for classifying structures. In: Proc. of the 2nd Workshop on Data Mining in Bioinformatics. (2002)