



A conceptual basis to encode and detect organic functional groups in XML



Punnaivanam Sankar^{a,*}, Alain Krief^b, Durairaj Vijayasarithi^a

^a Department of Chemistry, Pondicherry Engineering College, Puducherry 605014, India

^b Department of Chemistry, Faculté N.-D. de la Paix, Namur B 5000, Belgium

ARTICLE INFO

Article history:

Accepted 13 April 2013

Available online 20 April 2013

Keywords:

Chemical structure
Organic functional group
XML
Chemical ontology
OWL

ABSTRACT

A conceptual basis to define and detect organic functional groups is developed. The basic model of a functional group is termed as a primary functional group and is characterized by a group center composed of one or more group center atoms bonded to terminal atoms and skeletal carbon atoms. The generic group center patterns are identified from the structures of known functional groups. Accordingly, a chemical ontology 'Font' is developed to organize the existing functional groups as well as the new ones to be defined by the chemists. The basic model is extended to accommodate various combinations of primary functional groups as functional group assemblies. A concept of skeletal group is proposed to define the characteristic groups composed of only carbon atoms to be regarded as equivalent to functional groups. The combination of primary functional groups with skeletal groups is categorized as skeletal group assembly. In order to make the model suitable for reaction modeling purpose, a Graphical User Interface (GUI) is developed to define the functional groups and to encode in XML format appropriate to detect them in chemical structures. The system is capable of detecting multiple instances of primary functional groups as well as the overlapping *poly*-functional groups as the respective assemblies.

© 2013 Elsevier Inc. All rights reserved.

1. Introduction

Functional groups [1] are a group of atoms that impart some reliable and consistent properties for the molecule, including chemical reactivity. The knowledge of functional groups and their specific property on any molecule is gained by the chemists through their experience and expertise over the years. But the treatment of detecting functional groups [2–7] in a chemical structure needs specific tools, techniques and algorithms applied on a suitable structure representation format. A free and an open source tool, checkmol [2], detect and assign the functional group information on any small molecules with 2D coordinates. The checkmol is a command-line utility program, which reads molecular structure files in different formats and analyzes the input molecule for the presence of various functional groups. The output text can be easily placed into a database table, permitting the creation of chemical databases with a functional group search option. Another output option of checkmol is a set of statistical values viz. the number of atoms, bonds, and rings, the number of differently hybridized carbon, oxygen, and nitrogen atoms, the number of C=O double bonds, the number of rings of different sizes, the number of rings containing nitrogen, oxygen, sulfur, the number of aromatic rings, and the

number of heterocyclic rings, etc. derived from a given molecule to be used for quick retrieval from a database. For a standard molfile input, checkmol produces functional groups as output.

Based on the functional groups assigned by checkmol, a new small molecule chemical ontology (CO) is reported [3] for the application related to biological activity. The CO is developed from the checkmol functional group terms and their relationships and is used to automatically assign the functional groups and to categorize small molecule. Additionally the utility of CO is demonstrated for comparing the molecules as a basic pharmacophore search system on PubChem [4], a resource for chemical structures of small organic molecules along with their biological activities. In another study [5], the functional group detection is carried out by MATLAB tool using molfile derived from PDB files as input and the annotation is achieved through a functional group ontology (FGO). The specific arrangement of atoms and bonds with their coordinate information is extracted to categorize and identify the functional groups. This information is mapped with the concept tree of FGO for annotation. In these studies the functional groups are identified with varying boundaries ranging from a narrow to broader perspective. The main purpose is to bring to focus on the identification of specific part of the molecule and to arrive at possible target molecule through similarity search [3,6,7] with the database of molecules.

Generally the organic chemical groups are viewed by the chemists in different perspectives based on the structural composition. The chemical terms like hydroxy group, formyl group,

* Corresponding author. Tel.: +91 0413 2255952; mobile: +91 9486143908.

E-mail addresses: gapspec@gmail.com, sankar@pec.edu (P. Sankar).

carbonyl group, carboxyl group and amino group are simply representing some characteristic chemical groups. Whereas the terms such as alcohol, aldehyde, ketone, carboxylic acid, and amine represent the functional groups containing the characteristic chemical groups respectively. Further the functional groups are represented in a wider perspective using the terms like *tertiary*-alcohol, aromatic aldehyde, *di*-aryl ketone, aliphatic carboxylic acid, and *primary*-amine based on the skeletal environment on which the functional groups are located. It is important to note the terms used to represent different perspectives of the same chemical group. The terms alcohol and phenol represent two different perspectives of a hydroxyl group depending on the carbon skeleton on which it is located. Thus, the mixing of information related to the group of atoms characterizing the functional group and the structural details in the immediate environment of the group atoms is to be considered with clarity and certainty. Also specific importance is to be given to the atoms common to both the group and the skeleton on which the group is present. For example, the carbon atom bonded to a hydroxyl group in an alcohol is common to the structural part and functional group part of the molecule. Further discrimination as primary alcohol, secondary alcohol, and tertiary alcohol functionality of the group is based on the type of carbon linking the hydroxy group with specific structural environment. In the absence of tangible conceptual basis for defining functional groups, the generic assignment of functional group with unambiguous boundary is difficult. Another issue to be addressed in identifying functional groups is the annotation of functional group with suitable vocabulary. One of the possibilities to provide a controlled vocabulary is by supporting the functional group identification system with suitable chemical ontologies [8] capturing the functional group domain knowledge appropriately.

Accordingly, it is felt that there arises a concern to fix a conceptual basis for functional groups and is especially needed for the treatment of functional group detection from structure representation formats. There is no generic and explicit encoding methodology exclusively for functional group detection in chemical structures in a semantic perspective. Though there are several file formats [9] and structure editing systems [10] to handle structural information, the semantics associated with them are not sufficient to describe properties with deeper semantics. The treatment of the available structure representations for the descriptions of chemical reactions is difficult because of the inadequate or inexplicit semantics on the electronic environment around every atom in a chemical structure. The detail of electrons involving in chemical bonding is particularly important for the development of intelligent applications on reaction modeling. In our earlier work, [11] we reported a semantic representation of chemical structures in XML [12] format detailing the electronic environment around every atom providing the semantics of chemical bonding explicitly. Further a suitable structure editing tool, ChemEd [11], to handle the semantic format is also reported in the study. Now we propose a methodology to encode organic functional groups in XML through a GUI integrated with ChemEd. In this work we report a generic conceptual model of Functional Group enabling to define and detect through a semantic structure markup system [11]. The assignment of individual names of functional groups with a strict vocabulary is achieved through a functional group ontology developed in OWL [13] 2.0 through Protégé [14].

2. Conceptual model of functional group

2.1. Functional group and skeletal group

An organic functional group (FG) identified on an organic structure can be viewed in terms of two components. A group of atoms

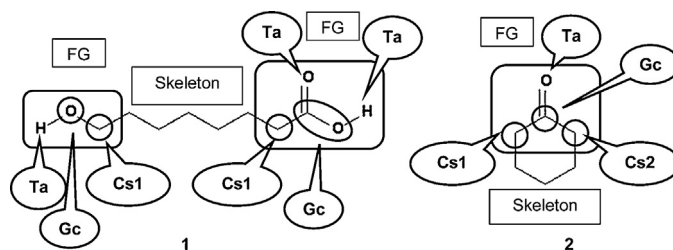


Fig. 1. The conceptual composition of functional group.

generally known as a characteristic chemical group constitutes the first component. The second component is the skeletal carbon (Cs) bridging the group component to a carbon skeleton normally. In general, the group part of any functional group is described with explicit and definite atomic composition. Whereas, a skeletal carbon atom is not represented explicitly because the structural composition of it is variable. A detailed scrutiny of various commonly known functional groups reveals that the group component of the functional group can be viewed as composed of three types of atoms. It is possible to identify a central portion of a group part with minimum and necessary atoms in such a way that this central part is sufficient to describe all the other linked atoms. This central compartment of the group part in a functional group is considered as a group center (Gc) composed of group center atoms. The skeletal carbon atoms bonded to any of the group center atoms form the second type and termed as skeletal carbon. The remaining atoms found as terminals of the group and attached to the group center atoms are termed as terminal atoms (Ta) as shown in Fig. 1.

In this model, the group center is considered as the backbone of functional group connecting the skeletal carbons and the terminal atoms. For example the structure of compound 1 contains two functional groups: namely, alcohol and carboxylic acid. In the alcohol functional group, the single oxygen atom is the group center as it is linked to a hydrogen terminal atom and one skeletal carbon atom (Cs1). So the alcohol functional group belongs to a group center composed of one group center atom. In the carboxylic acid functional group, the group center is identified as composed of two connected (Gc) atoms viz. the carbonyl carbon and the oxygen of the hydroxy group. The carbonyl-oxygen and the hydrogen of the hydroxy group are attached as the terminal atoms (Ta) to this group center. The carbon atom attached to the carbonyl carbon is the skeletal carbon (Cs1). The carboxylic acid functional group therefore belongs to a group center composed of two group center atoms. In this approach the hydrogen terminal atoms need to be explicitly defined in order to identify the functional groups uniquely. In the carboxylic acid functional group the presence of hydrogen on the single-bonded oxygen which defines the oxygen atom as a group center atom. In case of a group center atom with two skeletal carbon atoms, for example in a ketone 2 functional group is characterized by the presence of two skeletal carbons (Cs1 and Cs2) as part of a six member ring making the Gc atom as part of the same ring system.

Extending the concept of Gc further, the group centers can be classified depending upon the number of group center atoms and their specific patterns. Fig. 2 discloses the structure of group centers composed up to four atoms (not limitative) as well as the structure of some related functional groups (not limitative). The group center atoms in the group center pattern are designated as Gc1, Gc2, Gc3, and Gc4 in the Gc-Pattern column and the skeletal carbons in the example structures designated as Cs₁, Cs₂, Cs₃, Cs₄, Cs₅, etc. in Fig. 2.

The commonly known functional groups are categorized basically into three types of Gc-Pattern namely (1) Point-Gc (Fig. 2; 3–5), (2) Linear-Gc (Fig. 2; 6–14), and (3) Branched-Gc (Fig. 2; 15–17). The classification in Linear-Gc is related to the number of group center atoms linked one to the other (two 6–8, three 9–11 and

Gc-Type	Gc-Pattern	Example Structures		
Point-Gc	Gc ₁			
		3	4	5
Linear-Gc	Gc ₁ —Gc ₂			
		6	7	8
		9	10	11
		12	13	14
Branched-Gc				
		15	16	17

Fig. 2. The Gc-Type, Gc-Pattern and some example structures of functional groups.

four **12–14**). The Gc-Patterns shown in Fig. 2 are sufficient enough to describe the commonly known and classical functional groups. However, the Gc-Pattern can be extended appropriately for functional groups with a group center containing more than four group center atoms also if needed.

It is a conventional practice to consider carbon–carbon double bond **18**, and carbon–carbon triple bond **19** as functional groups due to their specific chemical and physical properties. In order to accommodate these structures as equivalent to functional groups in the proposed model, they are treated as skeletal groups (SG) with the notion of functional groups using group centers composed of only carbon atoms without any skeletal carbons and terminal atoms attached to them. The terms representing a skeletal group is associated with a suffix –SG. The present model treats these entities and related structures **20–22** as skeletal groups as they generally form the skeletal part in a chemical structure. The status of skeletal group as equivalent to functional group is especially is needed to define some structural entities as combination of functional group with a skeletal group. Accordingly, the structures of vinyl chloride, vinyl alcohol, etc. can be defined as a combination of a primary functional group with an appropriate skeletal group. This avoids the requirement of defining these structures as individual functional groups. Similarly the fundamental ring skeletons like cyclopentadiene skeleton **23**, benzene ring skeleton **24** and the saturated carbocyclic ring skeletons like **25–29**, are also treated as skeletal groups as equivalent to functional group conceptually in the proposed model. Some of the examples for skeletal groups are shown in Fig. 3 but the model allows extension of the concept of skeletal group to such similar structures further.

2.2. Functional group assembly and skeletal group assembly

Any functional group described in compliance with the fundamental group center pattern (Fig. 2) along with the respective skeletal carbon(s), and terminal atom(s) is considered as a primary functional group (PFG). However, the identity of an individual functional group changes significantly by the presence of more than one similar or different functional group in close proximity. If the skeletal carbons of primary functional groups are separated by at least one methylene group they can be treated as separate functional groups and their individual identity can be retained. When there is a combination of more than one PFG as overlapping structures, they are considered as an assembly of primary functional groups whose characteristics are decided by the constituent functional groups and the way they combine. Such functional groups are distinguished from the primary functional groups and are categorized as functional group assembly (FGA). According to this view three types of FGAs are identified on the basis of combination of skeletal carbons of PFGs. A combination of PFGs with a common skeletal carbon for the constituent functional groups is a fused functional group assembly (FFGA) (Fig. 4, **30–34**). The structures of *alpha*-hydroxy-alcohol-FG (synonymous to 1,1-diol-FG) **30**, *alpha*-hydroxy-nitrile-FG (synonymous to *alpha*-cyanohydrin-FG) **31**, *alpha*-aminoacid-FG **32**, *alpha*-chlorohydrine-FG **33**, *alpha*-chloro-chloride-FG (synonymous to 1,1-dichloro-FG) **34** are examples of FFGA with a single common skeletal carbon atom. If the skeletal carbons of two PFGs combine in such a way that the resultant assembly is composed of two skeletal carbons joined through a chemical bond, then the assembly is a joint functional group assembly (JFGA). The structures **35–38** (Fig. 4) represent this category.












				C=C=C	
18	19	20	21	22	23
					
24	25	26	27	28	29

Fig. 3. The skeletal group examples.

When the skeletal carbons of more than two PFGs are joined as a chain of PFGs then the resultant assembly is detected by the system as a chain functional group assembly (CFGA). The structure **39** is an example for CFGA.

There is another type of assembly possible by the combination of a primary functional group with a skeletal group. When a skeletal carbon of a primary functional group is located as part of a skeletal group or when a group center atom of primary functional group is located as part of a skeletal group, this combination is considered as an assembly of a functional group with a skeletal group and is termed as skeletal group assembly (SGA). In the first case, where the skeletal carbon of primary functional group is part of the skeletal group carbons, the assembly is named as crafted functional group assembly (CFGA) (Fig. 5, **40–47**). If one or more group center atoms of primary functional group become part of the skeletal group, it is detected as an assembly named as embedded functional group assembly (EFGA) (Fig. 5, **48–51**). This approach allows the possibility of combining the properties of the primary functional groups with the characteristics of the skeletal groups on which it is located. Accordingly, the detection of skeletal group assembly is useful to generate descriptors related to the orientation effects in a benzene ring, isomerism in a carbon–carbon double bond, axial or equatorial on cyclohexane ring skeleton, *endo*- or *exo*- orientations in *bi*-cyclic systems, etc.

2.3. Functional group ontology (FOnT)

The conceptual model is converted into a chemical ontology using Protégé [14] tool and is described in OWL [13]. The resultant functional group ontology is named as “FOnT”. The instances of various functional group categories are identified and the FOnT ontology is created in order to provide a strict and common vocabulary to the functional group instances to be defined and detected by the system precisely. The top level concept graph is shown in Fig. 6. The root concept of FOnT is Chemical-Group defined to include the subclasses namely FunctionalGroup and SkeletalGroup. The concept of FunctionalGroup is classified into OneGcAtomFunctionalGroup, TwoGcAtomFunctionalGroup,

ThreeGcAtomFunctionalGroup, and FourGcAtomFunctionalGroup according to the number of group center atoms. Though this sort of classification looks not chemically meaningful, it provides a generic taxonomy suitable for encoding and programming purposes. As the ontologies provide a mechanism of integrating further semantics to any concept defined in it, the context based chemical meanings can be associated through ontological relationships and axioms.

The concept of SkeletalGroup is classified into two subclasses namely AcyclicSkeletalGroup and CyclicSkeletalGroup. The concepts FunctionalGroupAssembly and SkeletalGroupAssembly are also defined as subclasses of ChemicalGroup to include the specific combinations of individual functional groups with themselves as well as with the skeletal groups. The FunctionalGroupAssembly is further classified into FusedFunctionalGroupAssembly, Joint-FunctionalGroupAssembly, and ChainFunctionalGroupAssembly to accommodate respective functional group combinations. Since the functional group assembly is a combination of only functional groups, the concept of FunctionalGroupAssembly is related to the concept of FunctionalGroup with ‘hasFunctionalGroup’ object property relationship. Whereas the SkeletalGroupAssembly is a combination of skeletal group and functional group, so the SkeletalGroupAssembly is associated with two object property relationships viz. ‘hasFunctionalGroup’ and ‘hasSkeletalGroup’ to relate the respective chemical groups. The individual instances for the appropriate concepts are identified and included in the ontology. At present the ontology includes around 200 commonly known functional groups for the subclasses of FunctionalGroup concept. The ontology includes about 15 SkeletalGroup instances and about 30 instances for FunctionalGroupAssembly and SkeletalGroupAssembly concepts.

2.4. Defining and encoding functional groups

The Graphical User Interface (GUI) developed to draw chemical structures in ChemEd [11] system programmed in JAVA [15] is used to define and encode organic functional groups. Accordingly, the system provides the facility to draw the functional group structure on the screen with the graphical tools used to draw chemical






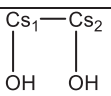
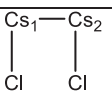
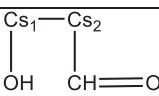
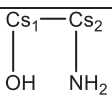
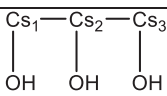
				
30	31	32	33	34
				
35	36	37	38	39

Fig. 4. Example for functional group assemblies.



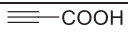

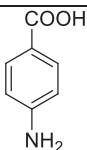
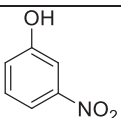
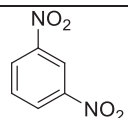
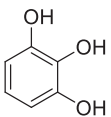




Crafted Skeletal Group Assembly			
			
40	41	42	43
			
44	45	46	47
Embedded Skeletal Group Assembly			
			
48	49	50	51

Fig. 5. The structures of some skeletal group assembly.

structures. In order to define a functional group, the structure of the functional group should be drawn on the screen in such a way that the functional group structure is built with group center atoms and terminated with skeletal carbon(s) and terminal atoms. The group center atoms can be marked with the marking tool as shown in Fig. 7 for carboxylic acid functional group as an example.

Once the structure of functional group is drawn, it is important to mark the group center atoms in the structure. For this purpose, the GUI provides a “Mark Group Center” icon in the top tool bar (Fig. 7). Selection of this tool and a click on the appropriate atom identified as group center atom displays a Group Center DialogBox for the selection of appropriate group center label from a list. The group center can be fixed with the group center atom by selecting the suitable label from the list and confirming through ‘OK’ button. This action displays the group center atom in red color and encircled indicating it as a group center atom. The same procedure can be repeated for fixing more group center atoms in case of group centers with more than one group center atoms. The following

procedure describes the fixing of more than one group center atoms in a functional group structure:

- For a Point-Gc type center, a single atom in the structure should be identified in such a way that all the terminal atoms and skeletal carbon atoms are attached to it and can be marked as a group center atom using the tool.
- For a Linear-Gc type center, a linear chain of atoms with two or three or four atoms should be identified in such a way that all the remaining atoms of the structure are attached to this chain of atoms. The chain of atoms identified can be marked as a Linear-Gc by selecting the atoms in the chain sequentially. A reverse order is also allowed in selecting the atoms sequentially in a chain. For example, in defining a carboxylic acid functional group, the selection of group center atoms like carbon first oxygen next and oxygen first carbon next are equivalent.
- For a Branched-Gc type, in case of a four Gc atom center, a three atom linear chain of atoms should be identified followed by the

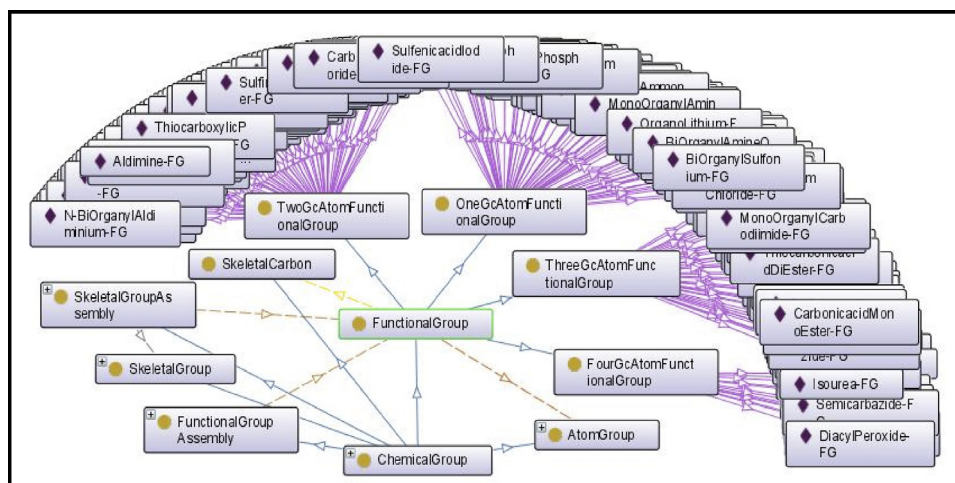


Fig. 6. The top level concept graph of functional group ontology 'Font'.

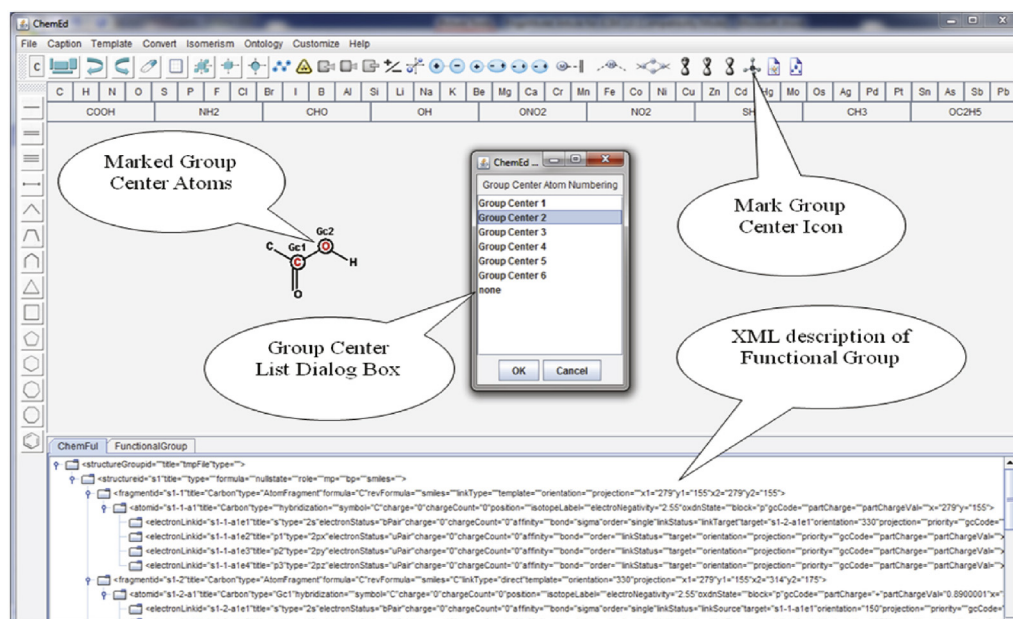


Fig. 7. The graphical user interface (GUI) for defining the carboxylic acid functional group.

fourth one attached to the middle atom of the chain so that all the terminal atoms and skeletal carbon atoms are attached to this group center. Then the group center can be marked in the same sequence they are identified.

Once the assignment of group center atoms is properly fixed, the functional group can be saved as an XML document with the appropriate name defined in the FOnt ontology through an interface. This approach provides the possibility of assigning the functional group names according to an approved vocabulary obtained through chemical ontology. So the title of any functional group can be edited in the FOnt ontology and the same can be assigned to the respective structure through an update algorithm at any time.

The functional group defined on the screen is encoded as a brief XML document and stored in the library. The XML definition of functional group captures the details of functional group in terms of the group center atoms necessary for detection through mapping. The functional group is encoded as an XML document consisting of a description of the FG structure part and FG definition part. The structure part of the XML describes the functional group in terms of structure, fragment, atom and electronLink elements as reported earlier [11]. The complete structural features of functional group are captured in this part of the document. This part allows the functional group structure to be stored and retrieved as stored. Also it enables to modify or edit the functional group structures for necessary changes if needed. The definition part of the structure captures the details of functional group in terms of the group center atoms necessary for detection through mapping. The semantics associated with the definition part of carboxylic acid functional group generated by the system is shown in Fig. 8.

The semantics of functional group definition is provided with suitable attributes associated to the corresponding element. All the elements are provided with 'id', 'title', and 'type' attributes to hold appropriate meanings. In the <definition> element, the 'id' attribute is not provided with any value at present. The 'title' attribute is used to denote the name of the functional group as supplied by the FOnt ontology. The 'type' attribute holds the functional group type like OneGc, TwoGc etc. The 'gcCount' attribute is used to capture the number of group center atoms in the functional group. Whereas, the attribute 'scCount' is to store the number of skeletal carbon atoms

present in the functional group. Two important attributes namely 'pattern' and 'notation' are defined for the detection of functional groups through mapping. The pattern attribute is used to hold the group center pattern such as "Point", "Linear", and "Branched". The 'notation' attribute captures a unique code for the functional group which can be used for detection purpose.

Each <definition> element is the container for group center atoms captured with appropriate number of <atom> elements along with the suitable attributes. The 'id' attribute is needed to hold the actual 'id' of the group center atom in the functional group structure generated by ChemEd. The name of the atom is provided with the 'title' attribute. The 'type' attribute is used to indicate the group center atom's number like "Gc1/Gc2/Gc3/Gc4/Gc5". The details of the group center atom like, symbol, charge, and charge count are added as the semantics of the group center atoms with 'symbol', 'charge', and 'chargeCount' attributes.

Every group center atom is added with a vital attribute 'gcCode' to capture a group center code, to generate the functional group notation by a simple combination of individual gcCodes. The value of this attribute is generated by the system based on the surrounding atoms of the group center atom. For example, in the definition of Carboxylic acid-FG, the gcCode for first group center atom is "db@O;sb@O;sb@C". This code captures the chemical bonding around the group center carbon atom as linked to an oxygen atom through a double bond; to another oxygen atom with a single bond and to a carbon atom with a single bond. The gcCode for the second group center atom oxygen is encoded with the value of "sb@C;sb@H". Accordingly the information of the group center oxygen atom linked to a carbon atom and a hydrogen atom with single bonds is captured. This atom level gcCode is generated from the electronLink level gcCode fragments of the respective atom. The XML code snippet from the structural part of the XML description showing the details of semantics associated with one of the group center atom, the carbonyl carbon generated for carboxylic acid functional group is presented in Fig. 9.

The electronLink level gcCode is generated when the electronLink is linked to an atom. If an electronLink is linked to an oxygen atom through a double bond, the code fragment for the electronLink is generated as "db@O". In the same way for an electronLink describing a single bond carbon the gcCode generated is 'sb@C'. This

```

<definition id="" title="Carboxylicacid-FG" type="TwoGc" gcCount="2" scCount="1"
  pattern="Linear" notation="C[(db@O;sb@O;sb@C)]O[(sb@C;sb@H)]">
  <atom id="" title="Carbon" type="Gc1" symbol="C" charge="0" chargeCount="0"
    gcCode="(db@O;sb@O;sb@C)"/>
  <atom id="" title="Oxygen" type="Gc2" symbol="O" charge="0" chargeCount="0"
    gcCode="(sb@C;sb@H)"/>
</definition>

```

Fig. 8. Definition for Carboxylic acid-FG a TwoGcAtomFunctionalGroup.

value is stored in an attribute named as 'gcCode' for the electron-Link. The complete gcCode is obtained by integrating all the gcCode fragments generated for the individual electronLinks of an atom. During this integration the fragment gcCode of all the electron-Links are merged into a single string of gcCode of an atom through a bond-atom-priority basis in order to make the code as a generic one. The electronLink gcCode are prioritized based on the type of bond and atom by assigning an appropriate number. A numeric value of "1" is assigned for a single bond, "2" for a double bond and "3" for a triple bond. Then the atomic number of the linked atom is added as the decimals to the above bond order values. For example double bonded oxygen is assigned with a value of 2.8 and a value of 3.7 is assigned for a triply bonded nitrogen atom. This priority value of an electronLink is captured with the attribute 'priority'. The complete gcCode for every atom is generated as single string in the descending order of priority values of individual electronLink gcCode fragments. Accordingly, the gcCode string will start with the top priority bond and atom and ends with least priority bond and atom exclusively for the purpose of mapping. Ultimately the gcCodes of all the Gc atoms are combined to generate the unique notation for the functional group keeping the full gcCode within square brackets and attaching the same to the corresponding atom symbol as shown in Fig. 8.

3. Detection of functional group and skeletal group in chemical structures

The algorithm for detecting the functional group in a chemical structure uses the generic definitions of functional groups stored in a library of functional group. This library is a collection of various functional group definitions encoded through the system. For the detection process, all the functional group definitions available

in the library are collected and the group centers of every functional group are mapped in the structure description in XML. This mapping is achieved by identifying a similar Gc-Pattern in the structure description. If a matching Gc-Pattern is identified, then the notation generated for the matching group center of the chemical structure is compared with that of the reference notation in the definition obtained from the library. When both the Gc-Pattern and the notation are mapped for a functional group definition, the mapped group center in the structure is captured as the functional group with reference to the library definition. The semantics of the library definition is transferred to this newly identified functional group on structure. It is found that the algorithm is efficient enough to identify the functional groups as per the definition very precisely. Also the system is capable of identifying multiple instances of same functional group in the structure and provides a list of functional groups detected in the chemical structure drawn on the screen as shown in Fig. 10.

The functional groups detected by the system are displayed in the bottom panel of ChemEd interface in four categories with the titles namely primary functional group, functional group assembly, skeletal group and skeletal group assembly. The functional groups detected based on the group center pattern are categorized as primary functional groups. The system detects a total of six primary functional groups viz. two secondary-carboxamide functional groups, two alcohol functional groups, one sulfide and one tertiary-amine functional groups for the structure displayed on the screen (Fig. 10). These are displayed as a list in the primary functional group panel. Selection of any item in the list displays the corresponding functional group in the structure. In Fig. 10 one of the SecondaryCarboxamide-FG is selected and highlighted in red in the structure. The skeletal groups are detected based on the structural fragments used to draw the chemical structure. The XML

```

<atom id="s1-2-a1" title="Carbon" type="Gc1" symbol="C" charge="0" chargeCount="0" position=""
  isotopeLabel="" electroNegativity="2.55" oxdnState="" block="p"
  gcCode="(db@O;sb@O;sb@C)" partCharge="+" partChargeVal="2.67" x="0" y="0">
  <electronLink id="s1-2-a1e1" title="s" type="2s" electronStatus="bPair" charge="0"
    chargeCount="0" affinity="" bond="sigma" order="single" linkStatus="linkSource"
    target="s1-1-a1e1" orientation="210" projection="" priority="1.6" gcCode="sb@C"
    partCharge="0" partChargeVal="0" x1="0" y1="0" x2="-35" y2="20"/>
  <electronLink id="s1-2-a1e2" title="p1" type="2px" electronStatus="bPair" charge="0"
    chargeCount="0" affinity="" bond="sigma" order="double" linkStatus="linkTarget"
    target="s1-3-a1e1" orientation="90" projection="" priority="2.8" gcCode="db@O"
    partCharge="+" partChargeVal="0.89" x1="0" y1="0" x2="0" y2="-40"/>
  <electronLink id="s1-2-a1e3" title="p2" type="2py" electronStatus="bPair" charge="0"
    chargeCount="0" affinity="" bond="sigma" order="single" linkStatus="linkTarget"
    target="s1-4-a1e3" orientation="330" projection="" priority="1.8" gcCode="sb@O"
    partCharge="+" partChargeVal="0.89" x1="0" y1="0" x2="35" y2="20"/>
  <electronLink id="s1-2-a1e4" title="p3" type="2pz" electronStatus="bPair" charge="0"
    chargeCount="0" affinity="" bond="pi" order="double" linkStatus="linkTarget"
    target="s1-3-a1e4" orientation="90" projection="" priority="" gcCode=""
    partCharge="+" partChargeVal="0.89" x1="-5" y1="0" x2="-5" y2="-40"/>
</atom>

```

Fig. 9. XML code snippet showing the semantics of electronLinks associated with the carbonyl carbon atom of the Carboxylic acid-FG.

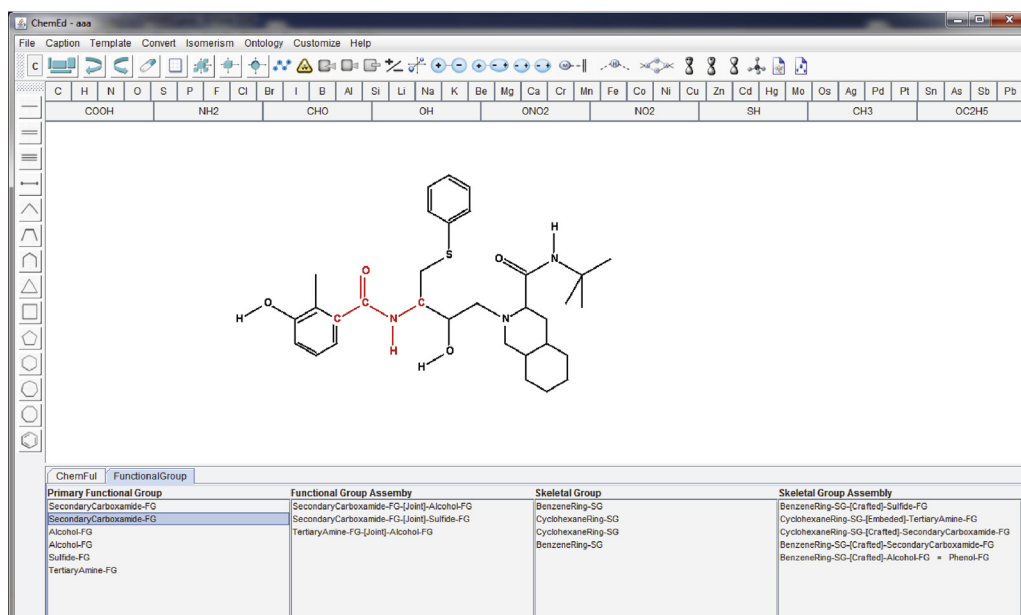


Fig. 10. Screen shot depicting the functional groups detected on a chemical structure.

description of the chemical structure makes this detection straight forward by identifying the corresponding skeleton fragments in the structure. The system detects four skeletal groups, two BenzeneRing-SG, and two CyclohexaneRing-SG. The detected skeletal groups are displayed in the Skeletal Group panel (Fig. 10).

4. Detection of functional group assembly and skeletal group assembly

Suitable algorithms are developed to identify the combinations of primary functional groups among themselves and with the skeletal groups to detect possible assemblies of functional group and skeletal groups described earlier. There are three functional group assemblies detected from the individual primary functional groups and displayed in the respective panel (Fig. 10) viz. SecondaryCarboxamide-FG-[Joint]-Alcohol-FG, SecondaryCarboxamide-FG-[Joint]-Sulfide-FG and TertiaryAmine-FG-[Joint]-Alcohol-FG. Also the panel of Skeletal Group Assembly shows a list of skeletal group assemblies detected on the structure as BenzeneRing-SG-[Crafted]-Sulfide-FG,

CyclohexaneRing-SG-[Embedded]-TertiaryAmine-FG, CyclohexaneRing-SG-[Crafted]-SecondaryCarboxamide-FG, BenzeneRing-SG-[Crafted]-SecondaryCarboxamide-FG, BenzeneRing-SG-[Crafted]-Alcohol-FG. Since there is no nomenclature available for functional group assemblies, the above methodology of description is used. A perfect system of nomenclature may be fixed by refining the system in due course. It has the advantage to show that once the proper nomenclature is fixed, there will be no limitation for the tool we just disclosed, to provide a proper report. Even this tool will help us to fix a functional group nomenclature and we are working toward this end.

Since the functional groups and their assemblies are defined in FOnt ontology with their individual instances, the details captured for the assembly of functional groups and skeletal groups can be processed appropriately to generate some useful descriptors. The specific name of skeletal group assemblies can be inferred through ontological relationships defined between the assembly concepts and the respective individuals in the ontology. It is seen from the report that one of the Alcohol-FG is crafted to a BenzeneRing-SG resulting in a skeletal group assembly. This information can

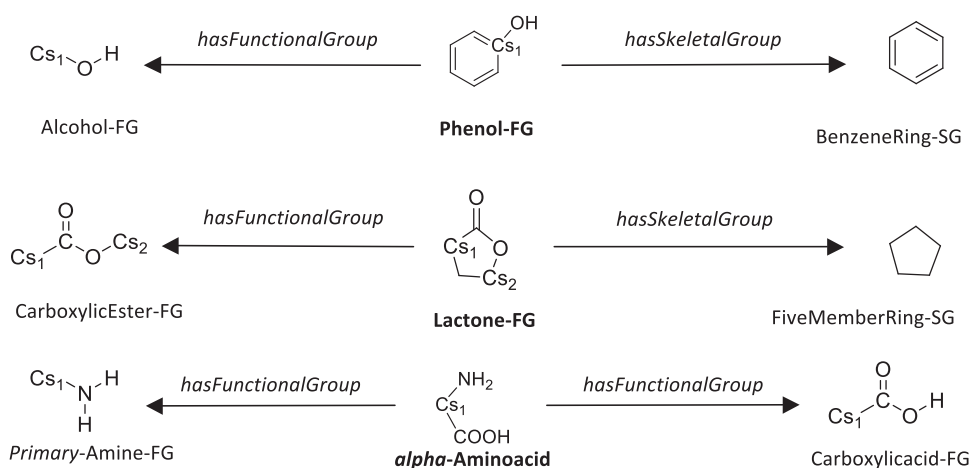


Fig. 11. OWL relationship to infer specific names for functional group and skeletal group assemblies.

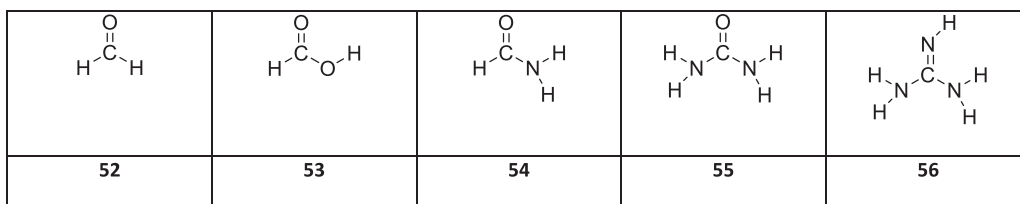


Fig. 12. Functional groups with no skeletal carbons.

be suitably processed through ontological relationship (Fig. 11) into a specific name as “Phenol-FG” normally recognized by the chemists. This inference is added to the functional group report as BenzeneRing-SG-[Crafted]-Alcohol-FG = Phenol-FG in Fig. 10. This approach is extended to other types of assemblies also as illustrated in Fig. 11. A CarboxylicEster-FG detected in a ring as an embedded assembly is inferred as a Lactone-FG through the FOnt ontology. Similarly a PrimaryAmine-FG is detected as fused with a Carboxylicacid-FG, the combination can be inferred as an *alpha*-Aminoacid-FG through FOnt ontology. Since the scope of this article is limited to only detection of functional groups and their assemblies the details of generations of descriptors from the functional group assemblies are not covered in this article.

5. Limitation and validation

In the system there are no template definitions to detect the functional groups present in structures with no skeletal carbon like formaldehyde **52**, formic acid **53** and formamide **54**, urea **55**, and guanidine **56** in Fig. 12. Normally the chemists use to detect the aldehyde, carboxylic acid, carboxamide, urea, guanidine functional groups in these compounds respectively, and therefore the system is designed to formally replace the Cs present in the template by H to detect the functional group present in such a compound and this is what we have achieved.

Validation of the system is carried out by defining around 200 commonly known functional groups in FOnt ontology and in the functional group library. The detection of functional groups and structural groups is checked by drawing more than one thousand chemical structures belonging to different classes in ChemEd [11] interface. It is found that in all the cases the functional groups are detected and located by the system precisely.

6. Conclusion

As a preliminary step to model the chemical reactivity with a semantic approach, we have developed the system to define, describe and detect any organic functional group based on a conceptual model. The generic patterns of group centers identified as backbone of functional groups offer a standard approach of defining organic functional groups through a user friendly GUI. The Integration of this functional group GUI with the structure editor ChemEd enables an easy way of defining the functional groups on computer screen as per the chemist perspective. As the system is supported with a chemical ontology, FOnt, providing a controlled vocabulary of functional group is possible. Automatic generation of functional group definitions along with the semantic structure encoding format in XML facilitates the changes or additions or corrections to be made for any functional group definitions by the experts at any time. This approach is suitable to have the knowledge resources outside the tool and to enhance the resources without affecting the basic tools to which the resources are linked. The semantic markup of functional group description and definition is suitable to develop algorithms for functional group interchange, functional group transformations needed for a meaningful reaction

description. The compatibility between the functional group definition and the structure description in XML will allow the possibility to create virtual material objects with the components of descriptions of structural features as well as the properties of functional group to try reactions simulation using virtual chemical materials. The detection of functional group assemblies and skeletal group assemblies enable the functional groups to be described in a wider perspective and useful to generate valid descriptors in the context of reactivity. Further, combining the structural descriptors of the skeletal carbons enables the possibility of describing the functional groups as functional group classes like alicyclic chloride, tertiary alcohol, phenol, *di*-aryl ether, aryl alkyl ketone, etc. The outcome on this objective is encouraging and will be reported appropriately. The proposed conceptual model is suitable to be adopted by any representation format working with a suitable structure editing tool. Also, the approach could be useful in virtual screening technique [16] and for the evolving proposals of knowledge bases [17]. The approach developed is expected to form the basis for the development of reaction models in a semantics framework rather than working with the reaction databases. Accordingly, intelligent open source applications may be developed with the proposed approach making it suitable for the evolving Semantic Web.

Acknowledgements

The authors P.S. and D.V. acknowledge the financial support by the Department of Science & Technology (DST), New Delhi, India (Project No. SR/S1/OC-86/2009).

The author A.K. acknowledges the financial support of the Fonds National de la Recherche Scientifique (FNRS) Belgium.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.jmgm.2013.04.003>.

References

- [1] (a) WIKIPEDIA, Functional Group, http://en.wikipedia.org/wiki/Functional_group;
(b) WIKIPEDIA, Organic Chemistry/Overview of Functional Groups, http://en.wikibooks.org/wiki/Organic_Chemistry/Overview_of_Functional_Groups;
(c) IUPAC GOLD BOOK, Functional Group, <http://goldbook.iupac.org/F02555.html>;
(d) IUPAC GOLD BOOK, Compendium of Chemical Terminology, <http://goldbook.iupac.org>;
(e) J. March, Advanced Organic Chemistry: Reactions, Mechanisms, and Structure, third ed., Wiley, New York, 1985;
(f) L.L. Thomas, Review of Organic Functional Groups: Introduction to Medicinal Organic Chemistry, fourth ed., Lippincott Williams & Wilkins, Baltimore, MD, Philadelphia, PA, 2003.
- [2] (a) Analysis of Functional Groups in Organic Molecules, <http://merian.pch.univie.ac.at/~nhaider/fga.php>;
(b) N. Haider, Functionality pattern matching as an efficient complementary structure/reaction search tool: an open-source approach, *Molecules* 15 (8) (2010) 5079–5092.
- [3] H.J. Feldmann, M. Dumontier, S. Ling, N. Haider, C.W.V. Hogue, CO: a chemical ontology for identification of functional groups and semantic comparison of small molecules, *FEBS Letters* 579 (2005) 4685–4691.

- [4] PubChem, <http://pubchem.ncbi.nlm.nih.gov/>
- [5] P.K. Varadwaj, T. Lahiri, FGO: a novel ontology for identification of ligand functional group, *Bioinformation* 2 (3) (2007) 113–118.
- [6] R. Benignia, O. Tcheremenskaia, A. Worth, Computational Characterisation of Chemicals and Datasets in terms of Organic Functional Groups – A New ToxTree Rulebase JRC Scientific and Technical Reports Luxembourg, Publications Office of the European Union. 10.2788/3328, EUR 24871 EN ISBN 978-92-79-20643-6, ISSN 1831-9424, 2011.
- [7] P. Resnik, Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language, *Journal of Artificial Intelligence Research* 11 (1999) 95–130.
- [8] (a) M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nature Genetics* 25 (2000) 25–29;
 (b) C. Brooksbank, G. Cameron, J. Thornton, The European bioinformatics institutes data resources: towards systems biology, *Nucleic Acids Research* 33 (2005) D46–D53;
 (c) J. Hastings, D. Magka, C. Batchelor, L. Duan, R. Stevens, M. Ennis, C. Steinbeck, Structure-based classification and ontology in chemistry, *Journal of Cheminformatics* 4 (2012) 8;
 (d) P. Sankar, G. Aghila, Design and development of chemical ontologies for reaction representation, *Journal of Chemical Information and Modeling* 46 (2006) 2355–2368;
 (e) P. Sankar, G. Aghila, Ontology aided modeling of organic reaction mechanisms with flexible and fragment based XML markup procedures, *Journal of Chemical Information and Modeling* 47 (2007) 1747–1762;
 (f) M. Kotera, A.G. McDonald, S. Boyce, K.F. Tipton, Functional group and substructure searching as a tool in metabolomics, *PLoS ONE* 2 (2008) e1537.
- [9] (a) A. Dalby, J.G. Nourse, W.D. Hounshell, A.K.I. Gushurst, D.L. Grier, B.A. Leland, J. Laufer, Description of several chemical structure file formats used by computer programs developed at molecular design limited, *Journal of Chemical Information and Computer Science* 32 (1992) 244–255;
 (b) WIKIPEDIA, Chemical File Format, http://en.wikipedia.org/wiki/Chemical_file_format;
 (c) Accelrys, CTF File Formats, <https://community.accelrys.com/docs/DOC-3451>;
 (d) D. Weininger, SMILES a chemical language and information system. 1. Introduction to methodology and encoding rules, *Journal of Chemical Information and Computer Sciences* 28 (1) (1988) 31–36;
 (e) D. Weininger, A. Weininger, J. Weininger, Algorithm for generation of unique SMILES notation, *Journal of Chemical Information and Computer Science* 29 (2) (1989) 97–101;
 (f) D. Weininger, Graphical depiction of chemical structures, *Journal of Chemical Information and Computer Science* 30 (3) (1990) 237–243;
 (g) DAYLIGHT SMILES, A Simplified Chemical Language, <http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html>;
 (h) P. Murray-Rust, H.S. Rzepa, Chemical markup language and XML Part I. Basic principles, *Journal of Chemical Information and Computer Sciences* 39 (1999) 928–942.
- [10] (a) WIKIPEDIA, Molecule editor, http://en.wikipedia.org/wiki/Molecule_editor;
 (b) Perkin Elmer, Desktop Software, ChemDraw Ultra 12.0 Suite, <http://www.cambridgesoft.com/software/ChemDraw/>;
 (c) ACD/ChemSketch, A complete software package for drawing chemical structures, http://www.acdlabs.com/products/draw_nom/draw/chemsketch/;
 (d) ChemAxon, Advanced Chemical Drawing Software, Marvin Suite Ver 5.10.0, <http://www.chemaxon.com/products/marvin/marvinsketch/>;
 (e) Sourceforge, The Chemistry Development Kit, Ver 1.4.11, <http://sourceforge.net/projects/cdk/files/>;
 (f) S. Krause, E. Willighagen, C. Steinbeck, JChemPaint – Using the collaborative forces of the internet to develop a free editor for 2D chemical structures, *Molecules* 5 (2000) 93–98.
- [11] P. Sankar, A. Krief, G. Aghila, Model tool to describe chemical structures in XML format utilizing structural fragments and chemical ontology, *Journal of Chemical Information and Modeling* 50 (2012) 755–770.
- [12] W3C, Extensible Markup Language (XML), <http://www.w3.org/XML>
- [13] W3C, OWL Working Group, http://www.w3.org/2007/OWL/wiki/OWL_Working_Group
- [14] (a) Protégé, <http://protege.stanford.edu/>;
 (b) WikiHomePage, <http://protege.cim3.net/cgi-bin/wiki.pl/>
- [15] ORACLE, J ava SE 6, <http://www.oracle.com/technetwork/java/javase/overview/index-jsp-136246.html>
- [16] P. Badrinarayan, G.N. Sastry, Virtual high throughput screening in new lead identification, *Combinatorial Chemistry & High Throughput Screening* 14 (2011) 840–860.
- [17] L.L. Chepelev, M. Dumontier, Chemical entity semantic specification: knowledge representation for efficient semantic cheminformatics and facile data integration, *Journal of Cheminformatics* 3 (2011) 20.