Resolution of Redundant Semantic Type Assignments for Organic Chemicals in the UMLS

C. Paul Morrey[a,*], **Ling Chen**[b], **Michael Halper**[c], and **Yehoshua Perl**[d]

[a] Information Systems, New York-Presbyterian Hospital, New York, NY 10030

[b] Science Department, Borough of Manhattan Community College, City University of New York, New York, NY 10001

[c] Information Technology Department, New Jersey Institute of Technology, Newark, NJ 07102

[d] Structural Analysis of Biomedical Ontologies Center, Computer Science Department, New Jersey Institute of Technology, Newark, NJ 07102

Abstract

Objective—The Unified Medical Language System (UMLS) integrates terms from different sources into concepts and supplements these with the assignment of one or more high-level semantic types (STs) from its Semantic Network (SN). For a composite organic chemical concept, multiple assignments of organic chemical STs often serve to enumerate the types of the composite's underlying chemical constituents. This practice sometimes leads to the introduction of a forbidden *redundant* ST assignment, where both an ST and one of its descendants are assigned to the same concept. A methodology for resolving redundant ST assignments for organic chemicals, better capturing the essence of such composite chemicals than the typical omission of the more general ST, is presented.

Methods and Material—The typical SN resolution of a redundant ST assignment is to retain only the more specific ST assignment and omit the more general one. However, with organic chemicals, that is not always the correct strategy. A methodology for properly dealing with the redundancy based on the relative sizes of the chemical components is presented. It is more accurate to use the ST of the larger chemical component for capturing the category of the concept, even if that means using the more general ST.

Results—A sample of 254 chemical concepts having redundant ST assignments in older UMLS releases was audited to analyze the accuracy of current ST assignments. For 81 (32%) of them, our chemical analysis-based approach yielded a different recommendation from the UMLS (2009AA). New UMLS usage notes capturing rules of this methodology are proffered.

Conclusions—Redundant ST assignments have typically arisen for organic composite chemical concepts. A methodology for dealing with this kind of erroneous configuration, capturing the proper category for a composite chemical, is presented and demonstrated.

*Corresponding author: Tel: (201) 657-0690, Fax: (973) 642-7029, chm9111@nyp.org.

Keywords

1 Introduction

The Unified Medical Language System (UMLS) [1, 2] has been created through the integration of a collection of about 150 source vocabularies from the biomedical domain. These sources are varied in their scope and purpose, and their integration provides a vehicle for expanding their utility beyond their original applications [3]. The integrated terms and relationships are housed in the Metathesaurus (META) [4, 5], where they have been mapped into concepts and links between them.

The Semantic Network (SN) supports the integration by providing a collection of 133 broad categories, called semantic types (STs), that enable high-level grouping of the META's concepts without regard to their sources [6–9]. In particular, each concept is assigned one or more of these STs in order to elaborate its overarching semantics. This arrangement has helped enhance applications in areas such as knowledge retrieval [10], inter-terminology mapping [11, 12], and natural language processing [13, 14], among others.

In this paper, we deal with a specific kind of error, called *redundant assignment* [15], that can occur in the assignment of STs. This error occurs when a given concept has been assigned multiple STs and one of them is more general than another in the context of the SN's tree-structured hierarchy. For example, the assignment of Organic Chemical* to a concept also assigned Lipid (a child of Organic Chemical) is redundant. A natural way to resolve this error is to remove the assignment of the more general ST since its assignment is implied by the assignment of its descendant, the more specific ST [16].

This resolution of a redundant ST assignment is suitable when the semantics of the multiple ST assignment is that of a conjunction, that is, the concept fits multiple categories being both "a this and a that." However, when the two STs assigned a concept are from the subtree of the SN rooted at Organic Chemical, the semantics of a multiple ST assignment is different. Such an assignment is typically found for a concept that represents a *composite chemical*, which is obtained by combining other chemicals. Such composite chemical concepts are common in the UMLS with ST assignments from the subtree rooted at Organic Chemical.

The composite chemical represented by the concept could be a *conjugate* created by a chemical reaction of multiple chemicals, or it could be a *complex* formed from a mixture of chemicals. In each case, the composite chemical concept is collectively assigned all the STs assigned to its individual component chemicals. Hence, the logic that a more general ST assignment is redundant when a more specific ST assignment is also given has no basis in the case of a composite chemical concept, which is simply enumerating the types of the components. However, such redundant assignments are forbidden by the NLM in all cases, with no exception for these organic chemical composites.

A rule is needed for handling a redundant ST assignment from the Organic Chemical subtree that best reflects the essence of a composite chemical, similar to the solution when the more

---

*Semantic Types appear in bold font; concepts appear in italics

specific ST accurately captures the essence of a concept that does not denote a composite chemical. When reviewing the ST assignment choices made by the NLM in resolving redundant ST assignments to organic chemicals in earlier releases, no clear rule is detected. Sometimes, the more general ST was removed, and sometimes the more specific one was removed.

In this paper, we present a systematic methodology for properly resolving a redundant ST assignment in line with principles of chemistry. Our approach is based on a chemical analysis at the molecular level. The relative sizes of the respective constituents are the driving factors. In this way, the ST assignment better captures the nature of the composite chemical. The methodology is applied to a sample of organic chemicals for which a redundant assignment appeared in earlier releases of the UMLS and was resolved in later releases of the UMLS—allowing for simple comparisons.

Our methodology is suggested for use by editors when they are categorizing new composite chemical concepts that are being added to the UMLS. New usage notes are provided to guide the editors in this endeavor. Furthermore, the methodology should be used for revisiting organic chemical concepts that were identified to have redundant ST assignments in earlier releases of the UMLS. In a study of a sample of 254 such concepts, it was found that for 32% of them the current ST assignment does not accurately capture the essence of the concept.

2 Background

The SN efficiently expresses type information by utilizing inheritance along the IS-A path between types. Inheritance makes the explicit specification of certain information at lower-level descendant STs unnecessary when that same information already appears in higher-level ancestor STs [16].

Let *C* be a concept assigned both STs B and A such that B is a descendant of A. Then the assignment of A to *C* is called *redundant* [15] because it can be inferred from the assignment of B to *C* and the fact that there is an IS-A path from B to A. As an example, the concept *Dinprost* (C0012471), had four ST assignments in 2007AA: Eicosanoid, Pharmacologic Substance, Biologically Active Substance, and Hormone. Because Hormone IS-A Biologically Active Substance, the assignment of Biologically Active Substance to *Dinprost* is redundant. Note that the assignments of Eicosanoid and Pharmacologic Substance are not redundant.

We previously developed an algorithm [15] for the detection of all redundant ST assignments in the UMLS. For the past several years, we have been monitoring the UMLS and compiling data about redundant ST assignments which is presented in Table 1. For example, in the 2006AB release, there were 1,747 concepts with redundant ST assignments, e.g., Carbohydrate and Organic Chemical. (The "n/a" for 1998 indicates that the value was not recorded.)

These data have been periodically supplied to the curators of the UMLS at the National Library of Medicine (NLM). Some of the redundant assignments persisted through more than one release and are counted in multiple rows in Table 1. In general, the reported redundant ST assignments were removed within one or two releases afterward. But while existing redundant assignments were removed, others were created for new concepts integrated into the UMLS.

We note that there were no cases of redundant ST assignment detected in seven of the last eight UMLS releases. Only three such errors were detected in 2008AA, none of which were

concepts representing organic chemicals. These have since been corrected. Table 2 lists these concepts with their multiple STs and abbreviated source vocabularies. For all three concepts, the assignment of the more general ST was properly removed as indicated by the name of the concept. The NLM has implemented a program to detect redundant ST assignments as part of the quality assurance regimen before a new release (S. Srinivasan, personal communication).

We previously performed research into the modeling of conjugate and complex types in the framework of the Refined Semantic Network (RSN) [17]. In that research, concepts with a redundant ST assignment were identified as erroneous but the resolution of those errors was not discussed.

There have been examples of combination ST assignments involving multiple STs from the Organic Chemical subtree (included in Figure 1 as a reference), where one is Organic Chemical and another is its descendant. In such a case, the assignment of Organic Chemical is redundant, and the prescribed course of action for resolving the redundancy is to remove it [16]. In fact, between 2007AA and 2007AB, such a redundant Organic Chemical assignment was handled for 119 concepts, and afterward no redundant assignments remained.

3 Methods

3.1 Chemistry Based Analysis

The standard means of resolving a redundant assignment [16] may be inappropriate when dealing with composite chemical concepts, and we present a systematic methodology for proper resolution in line with principles of chemistry and chemical analysis. Before getting to our methodology, let us note that the combination of multiple "organic chemical" STs is meant to convey the types of the constituent chemicals in the case of a composite chemical. For example, an assignment of Organic Chemical and Lipid is not meant to indicate that the chemical compound is both an organic chemical and a lipid—which would be redundant since any lipid is an organic chemical—but rather that the composite chemical, call it C, is composed of two other chemicals: the first, an organic chemical, and the second, a lipid. Let us denote the first as C_1 and the second as C_2, and analyze this situation further. Note that C_1 is not a lipid because otherwise it would have been assigned Lipid rather than Organic Chemical, with the most specific relevant ST being used [9]. Likewise, it is none of the other types that are descendants of Organic Chemical, namely, Carbohydrate; Amino Acid, Peptide, or Protein; Organophosphorus Compound; Nucleic Acid, Nucleoside, or Nucleotide; Steroid; and Eicosanoid (see Figure 1).

In the field of chemistry, there are various families of organic chemicals. For each major family, an ST exists in the SN, e.g., Lipid and Carbohydrate. Some major families are grouped together into a single ST, e.g., Amino Acid, Peptide, or Protein. However, some minor families of organic chemicals do not have STs named for them. An example is "organometallic compounds," with concepts such as *manganese ethylenebis(dithiocarbamate)* (C0029252) and *Copper 3-phenyl salicylate* (C0301127). For lack of a better name, we will call these minor organic chemical families *auxiliary organic chemicals*. Since no descendant of Organic Chemical is suitable for assignment to auxiliary organic chemical concepts, they are assigned Organic Chemical as the most specific available ST. However, these chemical concepts are not necessarily more general than other organic chemical concepts. They are categorized at a higher level due to a lack of granularity in the SN's Organic Chemical subtree. In our example, C_1 represents an auxiliary organic chemical, while C_2 represents the lipid. Their combination creates the chemical denoted by C, which could be a conjugate or complex.

In the case of a chemical formed from two lipids, the composed chemical's concept will be assigned Lipid. If the combination is of two auxiliary organic chemicals, then the concept will be assigned Organic Chemical. Suppose a composite chemical is formed from a combination of two chemicals, each of which is assigned a different child ST of Organic Chemical, say Lipid and Carbohydrate. Then the composite chemical is assigned both STs, Lipid and Carbohydrate. All of these cases fit a legitimate pattern of ST assignment.

But the above pattern of using an enumeration of the constituents' types for assignment to the composite chemical concept breaks down when we are faced with a chemical composed of a lipid and an auxiliary organic chemical. In such a situation, it entails the redundant assignment of Organic Chemical, the parent of the other assigned ST Lipid. Assigning only Lipid would cause the loss of the type-level knowledge concerning the contribution of the auxiliary organic chemical. The same can be said for using only Organic Chemical.

3.2 Resolution Methodology

We now present a resolution strategy for this problem based on the analysis of the underlying chemical compositions at the molecular level. The basis of the method is a comparison of the relative sizes of the component moieties of the concept having a redundant ST assignment. The ST assignment representing the larger, more dominant component is retained—even if it is the more general ST. Let us point out that in some rare cases involving components assigned the ST pairs "Organophosphorous Compound and Organic Chemical," "Steroid and Lipid," and "Eicosanoid and Lipid," their resolution does not follow the larger component due to conventions of chemistry. These cases are not covered by our methodology, but are handled by the rules described in the last three usage notes of Table 6. The following steps formally describe the methodology:

STEP 1)   Identify distinct Organic Chemical STs of all components of the composite organic chemical represented by the concept.

STEP 2)   IF there is no redundancy among the STs,

THEN return.

STEP 3)   IF exactly two STs are involved in a redundancy,

THEN determine the relative sizes of the two components involved in that redundancy. Only the ST assigned to the component of larger size will be assigned to the composite chemical concept.

ELSE

IF there are exactly three STs involved in the redundancy

THEN determine the relative sizes of the three components.

IF the ST of the largest-sized component is the parent of the STs of the smaller-sized components,

THEN assign only this largest-sized component ST to the composite chemical concept;

ELSE assign the largest-sized component ST and its sibling ST to the composite chemical concept, but do not assign their parent ST[†].

[†]Note that if three STs are involved in the redundancy then one is the parent and the other two must be children. Hence in this case the largest-sized component ST has only one sibling ST.

Note that this methodology is applicable to both kinds of composite chemicals, conjugates and complexes. For the conjugates, we are measuring the sizes of the moieties—the components of the molecule of the conjugate concepts. For complexes, there are separate molecules for each of the components, which are not connected by covalent bonds. Thus, we compare the sizes of the molecule of each component involved in the mixture.

3.3 Illustrative Examples

As an illustration, consider the conjugate concept *vicenistatin* (C0660734), which was assigned Organic Chemical, Carbohydrate, and Pharmacologic Substance in 2007AA, with the Organic Chemical assignment being redundant. Figure 2 gives the structure of the *vicenistatin* molecule. The right side of the figure shows a structural component (an amino sugar with a total of seven carbons) that is a carbohydrate and causes the assignment of Carbohydrate. *Vicenistatin* also has another structural component (left side of the figure) that is a 20-member cyclic amide consisting of 23 carbons (including the side chains). It is an auxiliary organic chemical, and leads to the assignment of the general Organic Chemical. (The additional assignment of Pharmacologic Substance is from the functional perspective.)

In this case, the auxiliary organic chemical component is the larger of the structures of the chemical. Therefore, being forced to choose only one ST to avoid the redundancy, the correct ST assignment to describe the structure of this chemical is Organic Chemical— keeping the more general type rather than the more specialized type. This is different from the current ST assignment in the UMLS, where we find Carbohydrate. The assignment of Pharmacologic Substance is, of course, retained.

For a complex chemical concept, the relative sizes of the component molecules are considered. For example, in 2007AA, *bis(glutathionato)platinum(II)* (C0661297), representing a complex chemical, was assigned the three STs Organic Chemical; Amino Acid, Peptide, or Protein; and Pharmacologic Substance. The assignment of Organic Chemical is redundant. Because the peptide component *Glutathione* (C0017817) is larger than the auxiliary organic chemical (organometallic) component, containing platinum, our methodology assigns Amino Acid, Peptide, or Protein and Pharmacologic Substance. The NLM agrees with this result and dropped the assignment to Organic Chemical in 2009AA.

Our methodology implies a rule for the initial ST assignment when conjugate or complex concepts are entered into the UMLS. Corresponding UMLS usage notes should be added to Organic Chemical and its descendants to clarify the process for non-redundant ST assignments involving conjugate or complex organic chemicals. For example, a usage note for Carbohydrate would be: "Conjugates or complexes in which the larger component is a carbohydrate and the smaller component cannot be categorized more specifically than ST Organic Chemical should be assigned only ST Carbohydrate." In the sample of redundant ST assignments chosen for the application of our methodology, there were no cases of redundancy among three STs. We will address the potential for such a situation in Section 5.

4 Results

Table 3 shows the number of concepts with redundant ST assignments involving Organic Chemical for three UMLS versions. For example, in 2006AB, there were 1,626 such concepts. No such redundancies were encountered in the versions more recent than 2007AA. Some concepts have been counted more than once in Table 3 (in consecutive versions). The total number of distinct concepts is 1,668.

We selected a sample of 254 from these concepts for review. The sample contained all 127 concepts from 2007AA. Of these 127, 84 were from 2006AC, and six were from both the

2006AC and the 2006AB and had not yet had their redundant ST assignments fixed. The remaining 37 were newly added in the 2007AA release. The sample contained an additional 127 concepts that were selected randomly from among the 1,626 concepts with redundant ST assignments involving Organic Chemical found in the 2006AB (see Table 3).

The analysis prescribed by our methodology shows that 54% (138 of 254) of the concepts should be assigned the more general Organic Chemical. The more specific ST should be assigned in 42% (107 of 254) of the cases. In 4% (9 of 254), we found both of the assigned chemical-viewed-structurally STs were invalid.

We compared our analysis with UMLS 2009AA and found that for 68% (173 of 254) of the concepts, the NLM changed the structural ST assignments in the way that we suggest, following the larger component. In the other 32% (81 of 254), the change of the structural ST assignment was different from our recommendation. In Table 4, the distribution of the 254 concepts is shown comparing our recommendations of assigned chemical-viewed-structurally STs with the assignments in 2009AA. From this distribution, it is clear that from the perspective of the relative sizes of the moieties we cannot identify any systematic approach used by the UMLS editors for such concepts.

Further applications of our methodology are illustrated by the following three examples. The review was performed by one of the authors (LC) who is a chemistry professor.

Example 1

The concept *spongistatin 1* (C0248118) was assigned the STs Organic Chemical, Carbohydrate, and Pharmacologic Substance in the 2007AA and identified for review by an algorithm [15]. It was determined that *spongistatin 1* is composed of a larger auxiliary organic chemical moiety and a smaller carbohydrate moiety. Since the auxiliary organic chemical moiety is larger, Organic Chemical is retained, along with the functional ST Pharmacologic Substance, while the assignment to Carbohydrate is dropped. The NLM also chose to retain Organic Chemical while dropping Carbohydrate.

Example 2

The concept *1a-docosahexaenoyl mitomycin C* (C0756517) was assigned the STs Organic Chemical, Lipid, and Pharmacologic Substance in 2006AB. It was determined that *1a-docosahexaenoyl mitomycin C* is composed of a larger auxiliary organic chemical moiety and a smaller lipid moiety. Since the auxiliary organic chemical moiety is larger, Organic Chemical is retained, along with the functional ST Pharmacologic Substance, while the assignment of Lipid is dropped. However, in resolving the redundancy, the NLM chose instead to retain Lipid while dropping Organic Chemical.

Example 3

The concept *leucine betaine* (C0391154) was assigned the STs Organic Chemical and Amino Acid, Peptide, or Protein in 2006AB. It was determined that *leucine betaine* is composed of a larger amino acid moiety and a smaller auxiliary organic chemical moiety. Since the amino acid moiety is larger, Amino Acid, Peptide, or Protein is retained while the assignment to Organic Chemical is dropped. However, in resolving the redundancy, the NLM chose instead to retain Organic Chemical while dropping Amino Acid, Peptide, or Protein.

Table 5 shows 14 cases in which the analysis using our methodology results in a different ST assignment from that chosen by the NLM to resolve the redundancy. Appendix A

contains the entire list of 81 such cases. In Table 6, we include our recommended usage notes for all STs of the SN beneath and including Organic Chemical.

The last three usage notes of Table 6 do not follow the general methodology rule of this paper retaining the ST of the largest component of the molecule, but other rules of chemistry. The rule for assigning Organophosphorous Compound follows the convention of the NLM in designating this ST (see http://semanticnetwork.nlm.nih.gov/Download/RelationalFiles/SRDEF). A chemical that has a combination of an organic chemical and an organophosphorous compound is assigned only the Organophosphorous Compound ST, regardless of the sizes of the components, with the exceptions of phospholipids, sugar phosphates, and phosphoproteins. A steroid has a unique four-fused nucleus of 27 carbon atoms that is always a major structural component; even in combination with a lipid component [18]. Hence this case is actually in line with our methodology, implicitly, and the Steroid ST is assigned. Eicosanoid is a child of Lipid without a unique structure and with a moderate size [19]. A chemical with both an eicosanoid component and a lipid component is thus a lipid and is thus assigned only the Lipid ST.

5 Discussion

This paper handles an anomaly in the semantics of the assignment of multiple STs to the concepts of the META. The typical semantics is that of a conjunction, meaning a concept shares the semantics of both STs, e.g., in being both a Disease or Syndrome and an Anatomical Abnormality. However, when both STs are coming from the subtree of the SN rooted at Organic Chemical (Figure 1), the semantics is of a chemical obtained by a reaction or mixture of two chemicals, each of which has been assigned a different chemical ST, e.g., Lipid and Carbohydrate. The situation becomes critical where such multiple ST assignments cause a redundant ST assignment, forbidden according to UMLS policy. As shown, such a situation is common in the Organic Chemical subtree, and the standard resolution rule of deleting the assignment of the more general ST does not work in that context.

The analysis of concepts that were previously assigned certain redundant organic chemical STs has revealed that resolution of such redundancy was performed inconsistently. Our analysis has yielded the insights, offered in this paper, with respect to how such concepts' categorizations, using the STs of the SN, should be done more systematically and consistently. The value of our categorizations is that they better reflect the semantics of the composition of organic chemical concepts obtained by a chemical reaction or mixture of multiple chemicals.

The importance of this paper lies in its consideration of analysis of the chemical's structure in determining the categorization of concepts representing conjugate or complex chemicals. For such chemicals, their nature is in general determined by the larger component from the molecular perspective. Our categorization methodology has an eye toward the rules of chemistry.

A limitation of the methodology within the framework of the SN is the loss of the knowledge of the ST of the smaller component. Redundant ST assignments are not permitted in the UMLS—a rule enforced by the quality assurance procedures that are performed prior to the release of a new version. Allowing redundant ST assignments to exist for organic chemicals in the UMLS is one way of retaining knowledge about all of the components of a chemical, but would require a change in the policy. We are not suggesting this change for the current SN framework unless a driving need can be established. However, for the alternative network of the "Refined Semantic Network" [20, 21], this

problem can be dealt with by adding special types called *intersection types*, similar to the treatment in [17].

Another limitation of this work is that it relies on the previous assignment of a redundant ST to identify a concept for analysis. As we previously noted, the curators of the UMLS have already implemented a method for identifying concepts with redundant ST assignments and eliminating those redundancies before releasing the META. We certainly encourage a review of concepts that previously had redundant assignments. The results of our study suggest that for a meaningful percentage of these concepts, a review may change their assignment when following the chemical analysis-based technique suggested here. Over the years, thousands of such redundancies were resolved by UMLS editors. Fortunately, the NLM has records of those concepts and can retrieve them for review.

However, our major purpose is to help the UMLS editors in improving the process of categorization of such new organic chemical concepts as they are added to the META, and in properly eliminating redundancies when they are found. The methodology and new usage notes for the STs in the SN subtree rooted at ST Organic Chemical will hopefully serve this purpose. Overall, the result should be better modeling of conjugate and complex chemical concepts in the UMLS.

In our methodology, we consider a case of redundancy involving three different STs. We did not encounter such a conjugate or complex concept, but according to the rules of chemistry, one is possible. Thus, we included such a case for the time when it might be needed. We did not include any rule to deal with a case of a redundancy involving more than three STs because we consider the future existence of such a concept in the UMLS very unlikely since review of the UMLS did not find any concept assigned four or more STs which are all children of Chemical Viewed Structurally. Due to the structure of the SN's Entity subtree and the rules of chemistry, the only possible configuration for such occurrence is with Organic Chemical and two of its children, or one child, which is not Lipid, and one grandchild from Lipid. According to our methodology, if the organic chemical component is the largest, it will be the only ST assigned. Otherwise, both descendants of Organic Chemical, but not Organic Chemical itself, will be assigned.

There are three usage notes in Table 6 that do not follow the methodology about assigning the ST of the largest component. Those usage notes are for Organophosphorous Compound, Steroid, and Eicosanoid. Those usage notes are derived from the rules of chemistry.

As explained in the definition of the ST Chemical, a chemical can be categorized independently from the structural aspects and from the functional aspects (see http://semanticnetwork.nlm.nih.gov/Download/RelationalFiles/SRDEF). Thus, a chemical concept is typically assigned more than one ST. One ST is from the subtree of SN rooted at Chemical Viewed Structurally, and at least one ST is from the subtree rooted at Chemical Viewed Functionally. The problem of non-conjunctive semantics for an assignment of multiple STs is limited to structural STs, and thus this paper concentrated on the methodology of resolution of redundant assignments of structural chemical STs. For functional STs, the disjunctive semantics of multiple STs such as Pharmacologic Substance and Indicator, Reagent, or Diagnostic Aid is valid—since the same chemical may have multiple functional aspects. Thus, we see many chemicals with both structural and functional chemical STs. Sometimes the functional ST may be missing. For example, in our review of the concept *Dox-D-penetratin* (see Table 5), our domain expert (LC) recommended the addition of the ST Pharmacologic Substance as well as the replacement of Organic Chemical with Amino Acid, Peptide, or Protein.

6 Conclusion

The review and analysis of concepts that previously had redundant ST assignments in the UMLS has demonstrated that organic chemical concepts present a unique challenge in categorization. When an organic conjugate or complex chemical is being assigned a semantic type, the type for each of its components is determined. Except for a few rare cases (described by the last three usage notes of Table 6), we recommend that a combination of the STs of the components of an organic chemical that form a redundancy be resolved by assigning the ST of the larger molecular component—even when this ST is more general than an ST of another component. Such an assignment better reflects the nature of the concept's denoted chemical. Suggested additional corresponding UMLS usage notes to regulate the categorization of complex or conjugate organic chemicals are provided in Table 6.

The effect on the categorization of conjugate or complex organic chemical concepts in the UMLS was analyzed with respect to the assigned types of the SN for a sample of such concepts. A disciplined methodology is presented that systematically uses an ST assignment to convey the larger molecular component.

Acknowledgments

References

1. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004; 32:D267–70. [PubMed: 14681409]

2. Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO. The Unified Medical Language System: an informatics research collaboration. J Am Med Inform Assoc. 1998; 5:1–11. [PubMed: 9452981]

3. Campbell KE, Oliver DE, Shortliffe EH. The Unified Medical Language System: toward a collaborative approach for solving terminologic problems. J Am Med Inform Assoc. 1998; 5:12–6. [PubMed: 9452982]

4. Schuyler PL, Hole WT, Tuttle MS, Sherertz DD. The UMLS Metathesaurus: representing different views of biomedical concepts. Bull Med Libr Assoc. 1993; 81:217–22. [PubMed: 8472007]

5. Tuttle, MS.; Sherertz, DD.; Olson, NE.; Erlbaum, MS.; Sperzel, WD.; Fuller, LF., et al. Using META-1, the first version of the UMLS Metathesaurus. 14th Annual Symposium on Computer Applications in Medical Care; Los Alamitos, CA. 1990. p. 131-5.

6. McCray, AT. UMLS Semantic Network. 13th Annual Symposium on Computer Applications in Medical Care; Washington, DC. 1989. p. 503-7.

7. McCray, AT. Representing biomedical knowledge in the UMLS Semantic Network. High-Performance Medical Libraries: Advances in Information Management for the Virtual Era; Westport, CT. 1993. p. 45-55.

8. McCray AT. An upper-level ontology for the biomedical domain. Comp Funct Genomics. 2003; 4:80–4. [PubMed: 18629109]

9. McCray, AT.; Hole, WT. The Scope and Structure of the First Version of the UMLS Semantic Network. 14th Annual SCAMC; Los Alamitos, CA. 1990. p. 126-30.

10. Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. J Am Med Inform Assoc. 2008; 15:87–98. [PubMed: 17947625]

11. Bodenreider O, Nelson SJ, Hole WT, Chang HF. Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. Proc AMIA Symp. 1998:815–9. [PubMed: 9929332]

12. Kumar A, Ciccarese P, Quaglini S, Stefanelli M, Caffi E, Boiocchi L. Relating UMLS semantic types and task-based ontology to computer-interpretable clinical practice guidelines. Stud Health Technol Inform. 2003; 95:469–74. [PubMed: 14664031]

13. Leroy G, Rindflesch TC. Effects of information and machine learning algorithms on word sense disambiguation with small datasets. Int J Med Inform. 2005; 74:573–85. [PubMed: 15897005]

14. Yamamoto Y, Takagi T. Biomedical knowledge navigation by literature clustering. J Biomed Inform. 2007; 40:114–30. [PubMed: 16996316]

15. Peng, Y.; Halper, M.; Perl, Y.; Geller, J. Auditing the UMLS for redundant classifications. In: Kohane, IS., editor. Proc AMIA Symp. 2002/12/05. San Antonio, TX: 2002. p. 612-6.

16. McCray AT, Nelson SJ. The representation of meaning in the UMLS. Methods Inf Med. 1995; 34:193–201. [PubMed: 9082131]

17. Chen L, Morrey CP, Gu H, Halper M, Perl Y. Modeling multi-typed structurally viewed chemicals with the UMLS Refined Semantic Network. J Am Med Inform Assoc. 2009; 16:116–31. [PubMed: 18952946]

18. Boyer, RF. Concepts in Biochemistry. 3. Hoboken, NJ: John Wiley & Sons; 2006. p. 245-252.

19. Campbell, MK.; Farrell, SO. Biochemistry. 5. Belmont, CA: Brooks/Cole; 2005. p. 209-211.

20. Geller J, Gu H, Perl Y, Halper M. Semantic refinement and error correction in large terminological knowledge bases. Data Knowledge Eng. 2003; 45:1–32.

21. Gu H, Perl Y, Geller J, Halper M, Liu LM, Cimino JJ. Representing the UMLS as an object-oriented database: modeling issues and advantages. J Am Med Inform Assoc. 2000; 7:66–80. Selected for reprint in Haux R, Kulikowski C, eds.: Yearbook of Medical Informatics, International Medical Informatics Association, Rotterdam, 2001:271–285. [PubMed: 10641964]
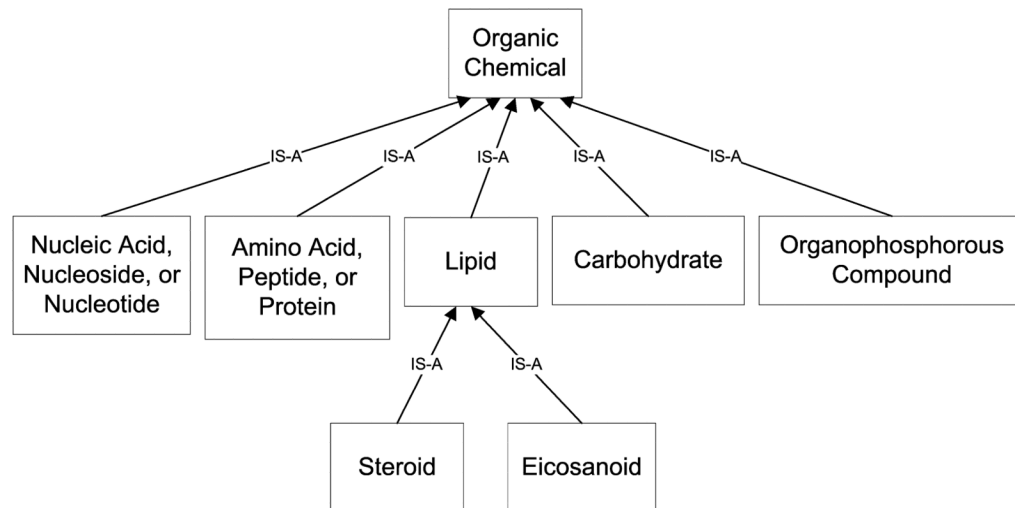
Figure 1.
The subtree of the SN rooted at the ST **Organic Chemical**.
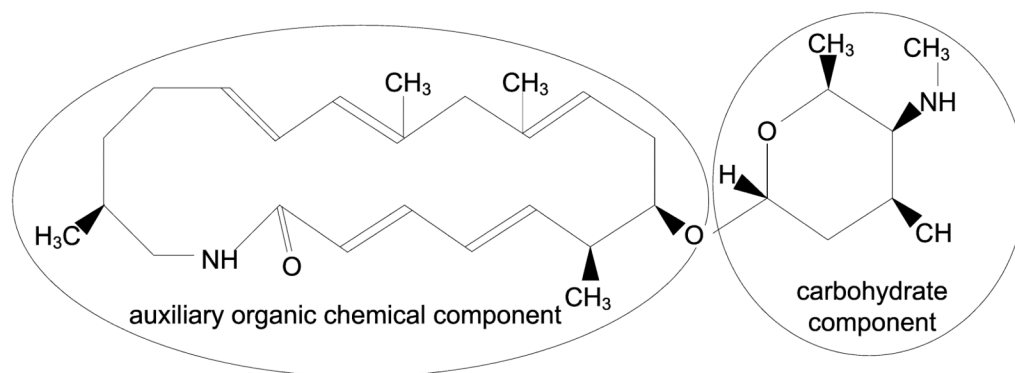
Figure 2.
The vicenistatin molecule.

Table 1

Redundant semantic type assignments across UMLS releases

UMLS Version	# Concepts with Redundant Semantic Type Assignments	# Redundant Semantic Type Combinations
1998	8,622	n/a
2001	12,161	40
2004	3,035	3
2006AB	1,747	19
2006AC	91	7
2007AA	598	11
2007AB	0	0
2007AC	0	0
2008AA	3	2
2008AB	0	0
2009AA	0	0
2009AB	0	0
2010AA	0	0
2010AB	0	0

Table 2

Concepts with redundant semantic type assignments in UMLS 2008AA

CUI	Preferred Term	Assigned STs in 2008AA	Source Vocabularies (abbreviated)	Assigned ST in 2008AB
C0266239	Congenital anomaly of bile ducts	Congenital Abnormality, Anatomical Abnormality	ICPC2ICD10ENG, RCD, SNOMEDCT, CST, SNMI, MDR	Congenital Abnormality
C2004426	Congenital cataract and lens anomalies (& 46)	Congenital Abnormality, Anatomical Abnormality	SNOMEDCT	Congenital Abnormality
C0349265	Severe mental and behavioral disorders associated with the puerperium, not elsewhere classified	Disease or Syndrome, Mental or Behavioral Dysfunction	RCD, SNOMEDCT, MTH	Mental or Behavioral Dysfunction

Table 3

Concepts with redundant semantic type assignments involving **Organic Chemical**

UMLS version	# concepts
2006AB	1,626
2006AC	90
2007AA	127

Table 4

Comparison of recommended versus actual semantic type assignments

Semantic Type	Our Recommendation (# concepts)	UMLS 2009AA (# concepts)
Organic Chemical	81	74
Nucleic Acid, Nucleoside, or Nucleotide	21	21
Organophosphorous Compound	12	11
Amino Acid, Peptide, or Protein	72	65
Carbohydrate	27	48
Lipid	29	25
Steroid	11	10
Eicosanoid	1	0

Table 5

Fourteen sample cases in which our analysis differs from the resolution of the NLM

halipeptin C (C1743613) 2009AA: Amino Acid, Peptide, or Protein Recommended: Organic Chemical	*diphenyl glycine* (C0912734) 2009AA: Amino Acid, Peptide, or Protein Recommended: Organic Chemical
calceolarioside A (C0661036) 2009AA: Carbohydrate Recommended: Organic Chemical; Pharmacologic Substance	*FK 506-dextran conjugate* (C0676168) 2009AA: Carbohydrate Recommended: Organic Chemical; Pharmacologic Substance
Dox-D-penetratin (C0913752) 2009AA: Organic Chemical Recommended: Amino Acid, Peptide, or Protein; Pharmacologic Substance	*leucine betaine* (C0391154) 2009AA: Organic Chemical Recommended: Amino Acid, Peptide, or Protein
iminoglutaric acid (C0957525) 2009AA: Organic Chemical Recommended: Amino Acid, Peptide, or Protein	*maltodapoh* (C0763377) 2009AA: Organic Chemical Recommended: Carbohydrate; Pharmacologic Substance
3-nitro-2-pyridyl glycopyranoside (C1137050) 2009AA: Organic Chemical Recommended: Carbohydrate	*guanofosfocin* (C1313468) 2009AA: Organic Chemical Recommended: Nucleic Acid, Nucleoside, or Nucleotide
callipeltose (C1098831) 2009AA: Carbohydrate Recommended: Organic Chemical	*(2S)-2-(1-oxo-1H-2,3-dihydroisoindol-2-yl)pentanoic acid* (C0915989) 2009AA: Lipid Recommended: Organic Chemical
naphthol AS-MX phosphate, sodium salt (C0959383) 2009AA: Organic Chemical Indicator, Reagent, or Diagnostic Aid Recommended: Organophosphorus Compound; Indicator, Reagent, or Diagnostic Agent	*N-(5-(dimethylamino)naphthylsulfonyl)phosphotyrosine* (C0764026) 2009AA: Organic Chemical Recommended: Organophosphorus Compound; Indicator, Reagent, or Diagnostic Agent

Table 6

Recommended additions to usage notes of semantic types

UMLS Semantic Type	Recommended Addition to Usage Note
Organic Chemical	Conjugates or complexes in which the larger component cannot be categorized more specifically than ST **Organic Chemical** and another component is assigned a descendant of **Organic Chemical** (except for ST **Organophosphorous Compound**) should be assigned only ST **Organic Chemical**.
Nucleic Acid, Nucleoside, or Nucleotide	Conjugates or complexes in which the larger component is a nucleic acid, nucleoside, or nucleotide and the smaller component cannot be categorized more specifically than ST **Organic Chemical** should be assigned only ST **Nucleic Acid, Nucleoside, or Nucleotide**.
Amino Acid, Peptide, or Protein	Conjugates or complexes in which the larger component is an amino acid, peptide, or protein and the smaller component cannot be categorized more specifically than ST **Organic Chemical** should be assigned only ST **Amino Acid, Peptide, or Protein**.
Carbohydrate	Conjugates or complexes in which the larger component is a carbohydrate and the smaller component cannot be categorized more specifically than ST **Organic Chemical** should be assigned only ST **Carbohydrate**.
Lipid	Conjugates or complexes in which the larger component is a lipid and the smaller component cannot be categorized more specifically than ST **Organic Chemical** should be assigned only ST **Lipid**.
Organophosphorous Compound	Conjugates or complexes in which one component cannot be categorized more specifically than ST **Organic Chemical** and the other component contains phosphorous should be assigned ST **Organophosphorous Compound**. This rule does not depend on the sizes of the components.
Steroid	Conjugates or complexes in which one component is assigned ST **Steroid** and the other component is assigned ST **Lipid** should be assigned only ST **Steroid**. This rule does not depend on the sizes of the components.
Eicosanoid	Conjugates or complexes in which one component is assigned ST **Eicosanoid** and the other component is assigned ST **Lipid** should be assigned only ST **Lipid**. This rule does not depend on the sizes of the components.

APPENDIX A

The list of the concepts sampled for which our recommended chemical-viewed-structurally Semantic Type assignment differs from the UMLS 2009AA.

---

folate monoglutamate (C0527846)
09AA: Amino Acid, Peptide, or Protein
Recmd: Organic Chemical

halipeptin C (C1743613)
09AA: Amino Acid, Peptide, or Protein
Recmd: Organic Chemical

trans-4-(aminomethyl)cyclohexanecarbonyl-O-(2-bromobenzyloxycarbonyl)tyrosine 4-acetylanilide (C0661082)
09AA: Amino Acid, Peptide, or Protein; Pharmacologic Substance
Recmd: Organic Chemical; Pharmacologic Substance

4-aminomethylcyclohexanecarbonyl-O-2-bromobenzyloxycarbonyltyrosine 4-acetylanilide (C0661083)
09AA: Amino Acid, Peptide, or Protein; Pharmacologic Substance
Recmd: Organic Chemical; Pharmacologic Substance

suprofen acyl glucuronide (C0660234)
09AA: Carbohydrate
Recmd: Organic Chemical

5'-O-fructofuranosylpyridoxine (C0660392)
09AA: Carbohydrate
Recmd: Organic Chemical

neoandrographolide (C0660410)
09AA: Carbohydrate
Recmd: Organic Chemical

phenyl 6HADPT-lactose (C0660430)
09AA: Carbohydrate
Recmd: Organic Chemical

carboxymefenamic acid glucuronide (C0662387)
09AA: Carbohydrate
Recmd: Organic Chemical

1-O-(2-(3-carboxy-2-methylphenyl)aminobenzoyl)glucopyranuronic acid (C0662388)
09AA: Carbohydrate
Recmd: Organic Chemical

mefenamic acid 1-O-acylglucuronide (C0662389)
09AA: Carbohydrate
Recmd: Organic Chemical

mefenamic acid glucuronide (C0662390)
09AA: Carbohydrate
Recmd: Organic Chemical

1-O-(2-(2,3-dimethylphenyl)aminobenzoyl)glucopyranuronic acid (C0662391)
09AA: Carbohydrate
Recmd: Organic Chemical

2-bromoethyl-2,3,4,6-tetra-O-acetyl-beta-D-glucopyranoside (C1722964)
09AA: Carbohydrate
Recmd: Organic Chemical

cucurbitane (C1740157)
09AA: Carbohydrate
Recmd: Organic Chemical

O-palmitoylmannose (C1742863)
09AA: Carbohydrate
Recmd: Lipid

vicenistatin (C0660734)
09AA: Carbohydrate; Pharmacologic Substance
Recmd: Organic Chemical; Pharmacologic Substance

calonyctin A-2b (C0661042)
09AA: Organic Chemical; Pharmacologic Substance
Recmd: Carbohydrate; Pharmacologic Substance

poly(2-methacryloyloxyethyl phosphorylcholine-co-n-butyl methacrylate) (C0212461)
09AA: Organophosphorus Compound; Pharmacologic Substance

Recmd: Lipid; Pharmacologic Substance

deoxyhaemoglobin-2,3-diphosphoglycerate complex (C0661309)
09AA: Organic Chemical
Recmd: Amino Acid, Peptide, or Protein; Biologically Active Substance

gadolinium phosphatidylethanolamine-DTPA (C0660441)
09AA: Organic Chemical; Pharmacologic Substance
Recmd: Lipid; Indicator, Reagent, or Diagnostic Aid

GP 1-668 (C0381639)
09AA: Organic Chemical
Recmd: Nucleic Acid, Nucleoside, or Nucleotide; Pharmacologic Substance

5-chloro-1-(2,3-dideoxy-3-fluoro-glycero-hex-2-enopyranose-4-ulosyl)uracil (C1100225)
09AA: Nucleic Acid, Nucleoside, or Nucleotide
Recmd: Organic Chemical; Pharmacologic Substance

naphthol AS-MX phosphate, sodium salt (C0959383)
09AA: Organic Chemical Indicator, Reagent, or Diagnostic Aid
Recmd: Organophosphorus Compound; Indicator, Reagent, or Diagnostic Agent

BIM 23197 (C0673881)
09AA: Organic Chemical
Recmd: Amino Acid, Peptide, or Protein; Pharmacologic Substance

iminoglutaric acid (C0957525)
09AA: Organic Chemical
Recmd: Amino Acid, Peptide, or Protein

S-(4-nitrobenzyl)glutathione-iodo-4-azidosalicyclic acid (C0538001)
09AA: Organic Chemical
Recmd: Amino Acid, Peptide, or Protein; Indicator, Reagent, or Diagnostic Agent

diamminechloro(glutathionato-S)platinum(II), (SP-4-3)-isomer (C0959804)
09AA: Organic Chemical
Recmd: Amino Acid, Peptide, or Protein

11-deoxyrhodomycinone 2,3,6-trideoxy-4-oxohexopyranosyl-(1-4)-2,6-dideoxyhexopyranosyl-(1-4)-2,3,6-trideoxy-3-dimethylaminohexopyranoside (C0258879)
09AA: Carbohydrate; Pharmacologic Substance
Recmd: Organic Chemical; Antibiotic

O-hydrazinocarbonylpentyl galactoside (C0295370)
09AA: Organic Chemical; Pharmacologic Substance
Recmd: Carbohydrate

5`-O-fructofuranosylpyridoxine (C0660392)
09AA: Carbohydrate
Recmd: Organic Chemical

calceolarioside A (C0661036)
09AA: Carbohydrate
Recmd: Organic Chemical; Pharmacologic Substance

3`-O-acetylfrangulin A (C1098533)
09AA: Organic Chemical
Recmd: Carbohydrate

callipeltose (C1098831)
09AA: Carbohydrate
Recmd: Organic Chemical

3-O-beta-D-glucopyranosyl-(1-3)-(beta-D-galactopyranosyl-(1-2))-beta-D-glucopyranosyl oleanolic acid 28-O-beta-D-glucopyranosyl-(1-6)-beta-D-glucopyranoside (C1120534)
09AA: Organic Chemical
Recmd: Lipid

(-)-7-O-methyleucomol 5-O-beta-D-glucopyranoside (C1121794)
09AA: Carbohydrate
Recmd: Organic Chemical

3-nitro-2-pyridyl glycopyranoside (C1137050)
09AA: Organic Chemical
Recmd: Carbohydrate

7-O-6'-O-malonylcachinesidic acid, triacetyl derivative (C1137197)
09AA: Organic Chemical
Recmd: Carbohydrate

TDP-3-amino-3,4,6-trideoxy-xylo-hexopyranose (C1171965)

09AA: Carbohydrate
Recmd: Nucleic Acid, Nucleoside, or Nucleotide

guanofosfocin (C1313468)
09AA: Organic Chemical
Recmd: Nucleic Acid, Nucleoside, or Nucleotide

phenyl-O-(2,3,4,6-tetra-O-acetylgalactopyranosyl)-1-4-3,6-di-O-acetyl-2-deoxy-2-phthalimido-1-thio-beta-glucopyranoside (C0660431)
09AA: Carbohydrate
Recmd: Organic Chemical

phenyl 3,6,2',3',4',6'-hexa-O-acetyl-2-deoxy-2-phthalimido-1-thiolactopyranoside (C0660432)
09AA: Carbohydrate
Recmd: Organic Chemical

glucose phenylosazone (C0660981)
09AA:Carbohydrate
Recmd: Organic Chemical

calceolarioside A (C0661036)
09AA: Carbohydrate
Recmd: Organic Chemical

1',2'-(3,4-dihydroxyphenyl-alpha,beta-dioxoethanol)-4'-O-caffeoyl-O-rhamnopyranosyl-1-3-O-glucopyranoside (C0661037)
09AA: Carbohydrate
Recmd: Organic Chemical

crenatoside (C0661038)
09AA: Carbohydrate
Recmd: Organic Chemical

1-O-(2-(3-hydroxymethyl-2-methylphenyl)aminobenzoyl)glucopyranuronic acid (C0662385)
09AA: Carbohydrate
Recmd: Organic Chemical

CM-glucuronide (C0662386)
09AA: Carbohydrate
Recmd: Organic Chemical

moracin M-3'-O-glucopyranoside (C0661369)
09AA: Carbohydrate; Pharmacologic Substance
Recmd: Organic Chemical; Pharmacologic Substance

CE 1037 (C0660665)
09AA: Lipid; Pharmacologic Substance
Recmd: Organic Chemical; Pharmacologic Substance

MDL 201,404YA (C0660670)
09AA: Lipid; Pharmacologic Substance
Recmd: Organic Chemical; Pharmacologic Substance

WAY 100252 (C0660702)
09AA: Lipid; Pharmacologic Substance
Recmd: Organic Chemical; Pharmacologic Substance

NSC 363223 (C0661397)
09AA: Nucleic Acid, Nucleoside, or Nucleotide
Recmd: Organic Chemical

isotiazofurin (C0661398)
09AA: Nucleic Acid, Nucleoside, or Nucleotide
Recmd: Organic Chemical

5-amino-1-(5'-phosphoribofuranosyl)-4-nitroimidazole (C0248635)
09AA: Nucleic Acid, Nucleoside, or Nucleotide; Pharmacologic Substance
Recmd: Organic Chemical; Pharmacologic Substance

desferri-ferricrocin (C0057528)
09AA: Organic Chemical
Recmd: Amino Acid, Peptide, or Protein; Biologically Active Substance

beta-(isoxazolin-5-on-4-yl)alanine (C0535028)
09AA: Amino Acid, Peptide, or Protein
Recmd: Organic Chemical; Antibiotic

R 820 (C0668665)
09AA: Organic Chemical
Recmd: Amino Acid, Peptide, or Protein

N-(2-(carboxymethyl)amino-2-oxo-1-((((phenylmethyl)seleno)thio)methyl)ethyl)glutamine (C0755834)
09AA: Organic Chemical

Recmd: Amino Acid, Peptide, or Protein; Pharmacologic Substance

N-(1-phenylalanine)-4-(1-pyrene)butyramide (C0757369)
09AA: Organic Chemical
Recmd: Amino Acid, Peptide, or Protein; Indicator, Reagent, or Diagnostic Agent

N-(5-(dimethylamino)naphthylsulfonyl)phosphotyrosine (C0764026)
09AA: Organic Chemical
Recmd: Organophosphorus Compound; Indicator, Reagent, or Diagnostic Agent

PF1070 A (C0908372)
09AA: Organic Chemical
Recmd: Amino Acid, Peptide, or Protein; Antibiotic

diphenyl glycine (C0912734)
09AA: Amino Acid, Peptide, or Protein
Recmd: Organic Chemical

Dox-D-penetratin (C0913752)
09AA: Organic Chemical
Recmd: Amino Acid, Peptide, or Protein; Pharmacologic Substance

leucine betaine (C0391154)
09AA: Organic Chemical
Recmd: Amino Acid, Peptide, or Protein

1-O-(2-(3-carboxy-2-methylphenyl)aminobenzoyl)glucopyranuronic acid (C0662388)
09AA: Carbohydrate
Recmd: Organic Chemical

isocytisoside (C0673185)
09AA: Carbohydrate
Recmd: Organic Chemical

FK 506-dextran conjugate (C0676168)
09AA: Carbohydrate
Recmd: Organic Chemical; Pharmacologic Substance

NaChito-EDTA (C0757131)
09AA: Organic Chemical
Recmd: Carbohydrate; Biomedical or Dental Material

maltodapoh (C0763377)
09AA: Organic Chemical
Recmd: Carbohydrate; Pharmacologic Substance

3,5-dihydroxypiperidine-4-yl O-glucopyranosyl-1-4-glucopyranoside (C0766530)
09AA: Organic Chemical
Recmd: Carbohydrate; Pharmacologic Substance

adenosine diphosphate-mannose (C0913639)
09AA: Organic Chemical
Recmd: Nucleic Acid, Nucleoside, or Nucleotide

ribonolactone, (D)-isomer (C0957712)
09AA: Organic Chemical
Recmd: Carbohydrate

N-(galactopyranosyl)pyridinium bromide (C0967529)
09AA: Organic Chemical
Recmd: Carbohydrate

NFAT 133 (C0299154)
09AA: Organic Chemical
Recmd: Lipid

YM 47522 (C0390043)
09AA: Organic Chemical; Antibiotic
Recmd: Lipid; Antibiotic

1a-docosahexaenoyl mitomycin C (C0756517)
09AA: Organic Chemical; Pharmacologic Substance
Recmd: Eicosanoid; Pharmacologic Substantce

N-methacryloyl-n-butyric acid (C0759992)
09AA: Organic Chemical
Recmd: Lipid

(2S)-2-(1-oxo-1H-2,3-dihydroisoindol-2-yl)pentanoic acid (C0915989)
09AA: Lipid
Recmd: Organic Chemical

G1268267X (C1100994)
09AA: Organic Chemical
Recmd: Steroid

(2S)-1,2-O-6,9,12,15-dioctadecatetraenoyl-3-O-(alpha-D-galactopyranosyl-(1""-6‴)-O-beta-D-galactopyranosyl)-glycerol (C1610810)
09AA: Organic Chemical
Recmd: Lipid