Analysing and Classifying Names of Chemical Compounds with CHEMorph

Gerhard Kremer
Institute for Natural Language Processing, University of Stuttgart

Stefanie Anstein
European Media Laboratory, Heidelberg;
Institute for Natural Language Processing, University of Stuttgart

Uwe Reyle
Institute for Natural Language Processing, University of Stuttgart

{gerhard.kremer stefanie.anstein@ims.uni-stuttgart.de uwe.reyle}

Abstract

We present a prototypical system with a purely linguistic method to analyse organic chemical compound names. It morpho-semantically analyses compound names, generates line-based, machine-readable representations of their corresponding molecular structures (SMILES strings), and triggers a taxonomic classification. CHEMorph is to be used to support manual database curation and as a basis for biochemical text processing. The system is written in Prolog.

1 Introduction and Related Work

Natural Language Processing (NLP) methods are indispensable to exploit the huge and growing amount of biochemical textual data in publications and databases. The information coded in terminological items has to be interpreted, as their identification and understanding is crucial for any NLP application task. In particular, the 'understanding' of chemical compound names, as provided by CHEMorph, is important for the population and curation of biomedical databases.

One such database is SABIO(-RK), a 'System for the Analysis of Biochemical Pathways (-Reaction Kinetics)'. It is being developed by the Scientific Database and Visualization (SDBV) group at European Media Laboratory (EML), Heidelberg, to support researchers (esp. in bioinformatics) with analysing, storing and querying information related to metabolic pathways (Rojas et al., 2002). In SABIO(-RK), data about reactions, their kinetics, compounds and enzymes are collected together with the respective experimental data. The system is soon to be released at `http://projects.villa-bosch.de/sdbv/projects`. Within SABIO(-RK), the collection of data is done by numerous student workers, who read PubMed journals and add database entries manually. Concrete challenges and problems in the process of manual curation are the existing errors, overlaps, and the inconsistency in or between databases. For example, multiple entries for one and the same molecule can occur if abbreviations and full forms of compound names are not matched (e. g. for *NTP* and *Nucleoside triphosphate*). These problems are enforced by the sheer amount of data, which makes manual handling almost impossible. For large-scale database conversion (see also the BioPAX[1] site at `http://www.biopax.org`), it is furthermore necessary to generate missing information on names, structures, or classification at least semi-automatically. In order to facilitate and improve the handling of this huge amount of data, curators need (semi-)automatic support for the database population and integration task.

The (long-term) aim of our project is to provide a tool that supports these tasks with the processing of biochemical compounds, which occur, e. g., in biochemical reactions. CHEMorph 'understands' the chemical terms to be entered into the database by means of a semantic analysis. A line-based, machine-readable molecular structure assignment (SMILES string[2]) and the classification[3] according to functional properties coded in the compound names will serve the automatisation of database population by means of term reference

[1] Biological Pathways Exchange

[2] `http://www.daylight.com/smiles`, Weininger (1988)

[3] see also Spasic et al. (2004) for another approach to classification

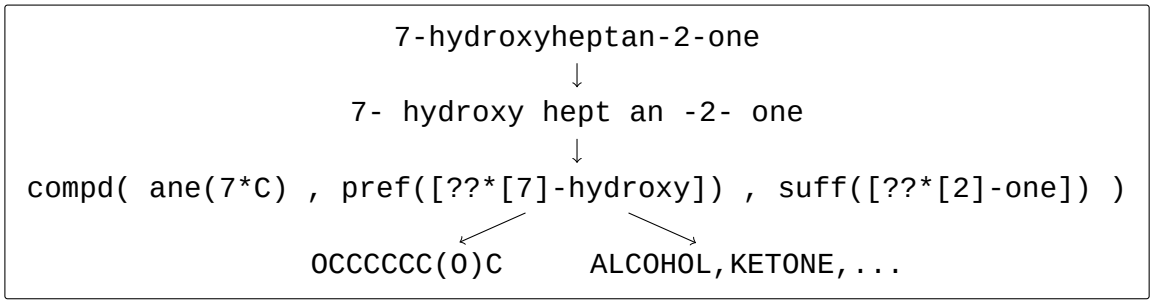

Figure 1: Example analysis of the name *7-hydroxyheptan-2-one*.

and coreference resolution. An example analysis for *7-hydroxyheptan-2-one* is shown in figure 1. The corresponding SMILES string and the list of classes can be seen in the bottom line; the complete molecular structure is shown in figure 2.

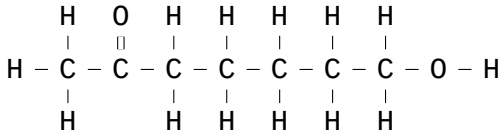

Figure 2: Molecular structure of *7-hydroxyheptan-2-one*

CHEMorph will provide (semi-)automatic support of database curators for efficient and high-quality database population, integration, and curation. Much time-consuming, expensive and error-prone manual work will thus be prevented. The tool will also diminish the problem of data inconsistency. Databases will become more reliable; they can be enriched faster, easier and more consistently.

The salient feature of CHEMorph is that it not only deals with fully specified chemical terms, but also with underspecified terms (i. e. terms lacking information about the molecular structure, which is a very frequent phenomenon in literature), class names and terms that are built up from subterms of any of these. The system, as described in Anstein and Kremer (2005), thus analyses fully specified (e. g. *2-deoxy-beta-D-erythro-pentose*), trivial (e. g. *benzene*) and semi-systematic (e. g. *benzene-1,3,5-triacetic acid*) as well as underspecified (e. g. *deoxysugar*) compound names. These functionalities as well as its depth of analysis distinguishes it from existing systems like 'ChemFinder'[4], 'PubChem'[5], 'ACD/Name'[6], or the 'Chemical Entity Relationship Skill Cartridge'[7] for identification, classification and name-to-structure translation.

Apart from our system, existing SMILES string generators only compute SMILES strings from graphical representations of molecular structures, not from names as they are coded in literature (e. g. 'Accelrys'[8] or the 'ChemAxon'[9] products). Furthermore, existing classifiers work only based on SMILES strings (Wittig et al., 2004) – not on names – or take their information from static databases (such as 'PAREO'[10]). Additional resources for compound name data are, e. g., 'PubChem Compound'[11], 'ChEBI'[12], 'KLOTHO'[13], 'Whatizit'[14], to name a few. Most systems described show no total correspondence to the functionalities of our tool (either because of requiring input other than names or the drawback not to cover new words formed productively according to nomenclature principles); only two systems ('ChemFinder' and Gerstenberger (2001)) conduct a compositional analysis (based on linguistic theory), but none of these handles underspecification. To the best of our knowledge, no comparable tool exists, and CHEMorph clearly extends the functionalities (even though not yet the quantitative coverage) of other systems.

[4] http://chemfinder.cambridgesoft.com

[5] http://pubchem.ncbi.nlm.nih.gov

[6] http://www.scienceserve.com/Software/ACD/ACD_Name.htm

[7] http://www.temis.com/?id=25&selt=14

[8] http://www.accelrys.com

[9] http://www.chemaxon.com

[10] http://genome.jouy.inra.fr/pareo

[11] http://pubchem.ncbi.nlm.nih.gov

[12] http://www.ebi.ac.uk/chebi

[13] http://www.biocheminfo.org/klotho

[14] http://www.ebi.ac.uk/Rebholz-srv/whatizit/form.jsp

2 Our Approach

CHEMorph linguistically analyses names that are based on the IUPAC nomenclature rules for organic compounds in general and the special nomenclature rules for sugar names (IUPAC Commission on Nomenclature of Organic Chemistry, 1993; IUPAC-IUBMB Joint Commission on Biochemical Nomenclature, 1996). As a theoretical basis we used the approach to the analysis of biochemical terminology described by Reyle (2006).

Each given compound name is deconstructed and assigned a semantic representation, which serves as the common basis for generating a representation of the molecular structure in the form of a SMILES string and for classifying the compound.

Underspecified names such as *hydroxyheptan-2-one* are analysed by assigning a partial SMILES string (if possible including information about which part of the name is underspecified) and the classes the compound belongs to.

2.1 Modules and Methods

The system consists of a linguistic parser which splits a term into meaningful subunits (so-called morphemes) and generates the semantic representation from a name. This representation then is used by the SMILES string generator and by the classification module to produce their corresponding output (see figure 3).

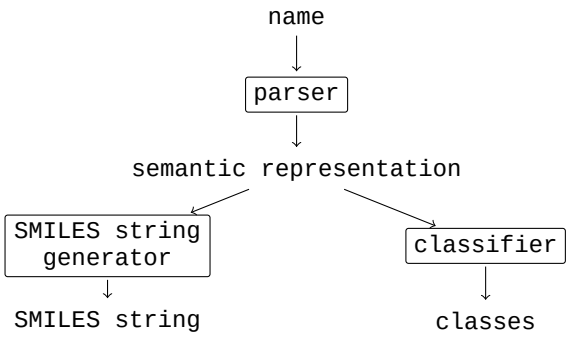

Figure 3: CHEMorph system architecture

Parser

The parser uses a Definite Clause Grammar (DCG, a rule-based Prolog formalism to separate linguistic entities into their segments) which morphologically splits the input name and compositionally constructs its semantic representation (see figure 4 for an example). Therefore, a lexicon is necessary which comprises morphemes, their corresponding syntactic categories and semantic annotations. Furthermore, each DCG rule describes how morphemes and – additionally – how their semantic annotations have to be combined to form a valid name.

The semantic representation for an organic compound name is a triple consisting of a parent term representation, a list of prefix operators, and a list of suffix operators with the following format: *compd(ParentTerm, PrefixList, SuffixList)*. The parent term representation basically consists of a representation of a molecular skeleton structure, typically a number of atoms of the same chemical element. Depending on the input name, this skeleton representation is modified by embedding it as an argument into a predicate-argument expression describing the kind of modifications together with their location(s). This representation may then itself be embedded into descriptions of other modifications, and so on, until the parent term is fully described. The parent term representation is further modified by the prefix and suffix operator list.

SMILES String Generator

The semantic representation describes an order of operations which modify a molecular skeleton structure. This representation is transferred into a Prolog representation of the molecular structure, which describes the properties (locant, bonds, attached elements, etc.) of each chain element. The format of such a chain element representation is shown in (1).

(1)
```
chain_el(Element,Locant,
         Branches_List,Features_List)
```

This intermediate representation is then used for generating the SMILES string. For underspecified structures, a partial SMILES string is generated, and a list of underspecified morphemes is shown (see (2)).

(2) a. format:
```
underspecified(SMILES_String,
     Underspecified_Prefixes_List)
```
b. example:
```
underspecified(C(=O)C([H])(O)...,
     [??*2,3,4-deoxy])
```

In example (2b), the expression *??*{2,3,4}-deoxy* specifies the set of possible locants (2, 3, and 4) for the prefix *deoxy* and describes that no multiplier (*??*) was given in the input name. Thus, the

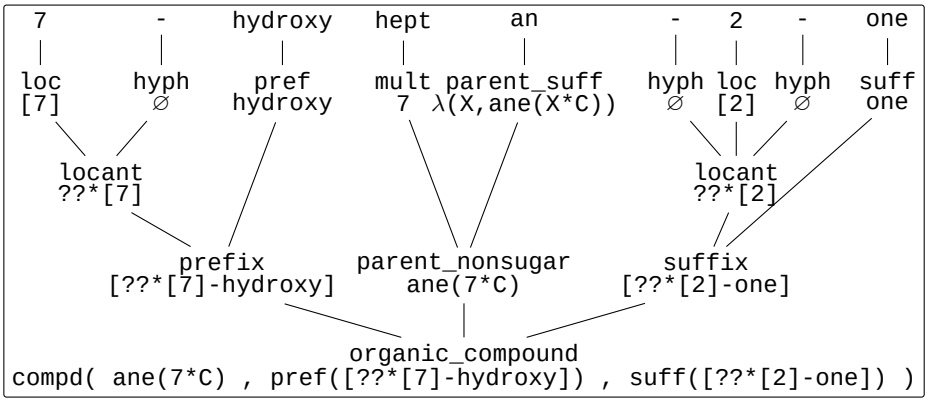

Figure 4: Example analysis for *7-hydroxyheptan-2-one*. The first line shows the name splitted into morphemes. Each one is assigned a morpho-syntactic category and a semantic representation. The grammar rules define how these can be combined to form the semantic output as shown in the bottom line.

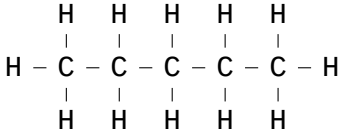

Figure 5: Molecular structure of *2,3-dihydropent-2-ene*

deoxy operation can be applied at most three times to the partial SMILES string, each time removing one oxygen atom attached with a single bond as a side branch to the main chain of the compound.

During the SMILES string generation process, a consistency check for locant-multiplier pairs is conducted to exclude impossible combinations.

Classifier

The classification module extracts and uses the operator names (representing morphemes) to calculate the chemical classes from the semantic representation heuristically. This is achieved by prescribing which combination of operators leads to which class of compounds.

In some cases, the locants associated to operators are used to check for a mutual influence of operators in the determination of the classes. As an example for such an effect, the prefix *2,3-dihydro* in *2,3-dihydropent-2-ene* 'neutralises' the *-ene* (double bond) desaturation: the latter is 'no longer' of the class ALKENE (see figure 5 for an image of the molecule structure).

From the classes generated so far, super-classes are calculated with the help of 'axioms' such as "Primary alcohols are alcohols.".

The output of the classifier is a list of class names as seen in the respective bottom lines of figure 1.

2.2 Results and Evaluation

CHEMorph takes organic chemical compound names as input and provides a semantic representation, a molecular structure representation in form of a SMILES string (if applicable) and a list of classes for a given term. Examples of such analyses are shown in table 1.

For the task of database curation, one application can be to prompt a name and get its SMILES string, its classification, or an error message in case the term cannot be analysed. A different application can be to compare two names and get the output message if they are synonymous or not, together with the reasons such as 'morpho-syntactic variants', 'same SMILES string', '... is a superclass of ...', etc.).

Further applications are the acquisition of an ontology similar to the extract in figure 6 (see also `http://www.geneontology.org`, Stevens et al. (2000), or Bodenreider and Burgun (2002)) and named-entity recognition for text mining applications, see e. g. Fluck et al. (2005).

As CHEMorph is still ongoing work, a large-scale evaluation (see Gaizauskas (1998) for a general overview) on, e. g., data from PubChem, KEGG[16], ChEBI, etc. can only be conducted as soon as the lexicon is (semi-automatically) extended. The system will have to be evaluated ac-

[16]Kyoto Encyclopedia of Genes and Genomes (`http://www.genome.ad.jp/kegg`)

name	results
7-hydroxyheptan-2-one	compd(ane(7*C),pref([??*[7]-hydroxy]),suff([??*[2]-one])) CC(=O)CCCCCO[15] KETONE, ALKANE, ALCOHOL
dipentene	compd(ene(2*[??],ane(5*C)),pref([]),suff([])) underspecified(CCCCC,[2*1,2,3,4-ene])[15] ALKENE
L-threo-tetrodialdose	compd(ose(2*[??],4*C),pref([cfg([L-threo])]),suff([])) C(=O)[C@]([H])(O)[C@@]([H])(O)C(=O) TETROSE, DIALDOSE, TETRODIALDOSE[15]
D-fructose	compd(triv_name(fructose),pref([cfg([D])]),suff([])) C(O)C(=O)[C@@]([H])(O)[C@]([H])(O)[C@]([H])(O)C(O) FRUCTOSE, HEXOSE, ALDOSE[15]
2-pentulose	compd(ulose(??*[2],5*C),pref([]),suff([])) C([H])(O)C([H])(O)C(=O)C([H])(O)C([H])(O) PENTULOSE, KETOSE[15]
pent-2-ulose	compd(ulose(??*[2],5*C),pref([]),suff([])) C([H])(O)C(=O)C([H])(O)C([H])(O)C([H])(O) PENTULOSE, KETOSE[15]

Table 1: Example analyses. The result for each name contains its semantic representation (first line), the corresponding SMILES string (second line), if available, and the value of the class list variable (third line).

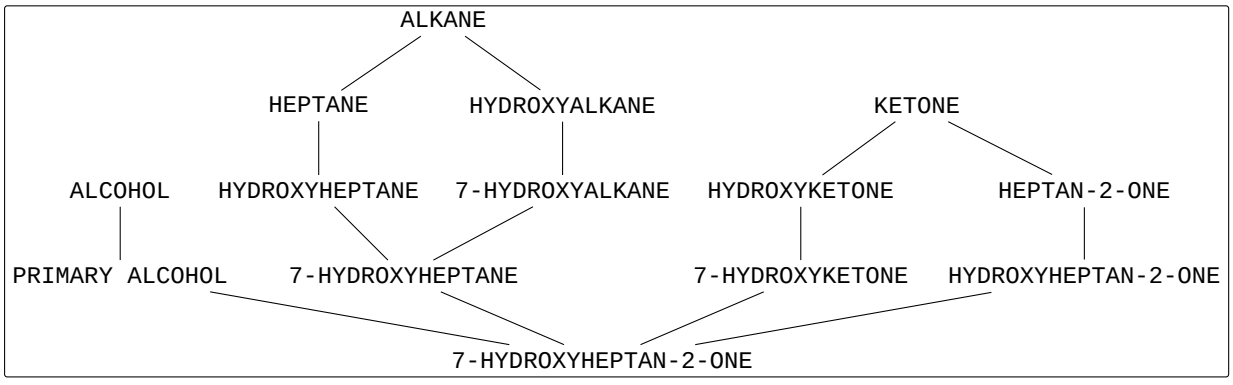

Figure 6: Class hierarchy for *7-hydroxyheptan-2-one* to serve as a knowledge base, generated by a step-by-step abstraction.

cording to a manually annotated reference set of analyses, SMILES strings, and classification for biochemical terms (as test corpora of this kind do not yet exist). At least for fully specified compound names for which SMILES strings exist, our classification result can also be compared to classes calculated by other tools, e. g. by the one based on SMILES strings as described by Wittig et al. (2004). Up to now we concentrated on developing a proper methodology (rather than aiming at a broad coverage of names) and produced a prototype. Nevertheless, we conducted a small evaluation on the system wrt the fragment defined by the nomenclature rules with 100 arbitrarily chosen names appearing there as examples. Semantic analyses were generated for 93 % of these. Failures are mostly due to grammar rules which are missing or still are too restrictive.

3 Conclusion and Outlook

CHEMorph is a system that analyses fully specified – trivial and (semi-)systematic – as well as underspecified compound names to support the curation of biomedical databases. It generates SMILES strings and determines possible classes of the terms.

The system is to be used to detect synonymous entries as well as errors and inconsistencies in or between databases. It can also be used to specify reaction equations which contain names of compound classes, as reaction equations are often expressed with general terms, e. g. 'alcohol' and 'alkyl sulfate' in *3'-phosphoadenylylsulfate + an alcohol = adenosine 3',5'-bisphosphate + an alkyl sulfate*. The generation of more specific forms which are contained in these compound classes presents a serious problem because relations (esp. 'is a', 'part of') between specific and general terms must be identified. Curators need a fully structured ontology that can be used as a knowledge base.

CHEMorph will thus offer more control and reliability for database curators by means of the following: (i) For term reference, it yields a linguistic analysis, a structure assignment and a classification of terms. Underspecified terms, which are frequent in these data, are represented by partial structures and are classified accordingly. The representations of all kinds of terms are used to resolve coreferences by semantic normalisation. (ii) Our system additionally establishes an ontological classification of general and specific terms, including 'is a' and 'part of' relations, but also a relation such as 'derivative of', which holds between two compounds if one is a formal derivative of the other.

To conclude, the system presented takes advantage of NLP methods for consistent, reliable and both time- and money-saving semi-automatical database population and curation.

We created a valuable prototype with many possibilities for its enhancement. On the one hand, this will have to be done by including further nomenclature systems, where one challenge will be to cover all names which people use in the biochemical domain. The focus for this work will be on lexical enrichment (esp. trivial names), as the grammar rules are quite limited. On the other hand, enhancement can also consist in more sophisticated algorithms for more context and domain knowledge, semantic inferences, presupposition resolution, default rules, etc. The unique technique of such a deep linguistic analysis with a 'semantic model' for structural and functional properties of molecules makes it possible to calculate, e. g., molecular features which are wrong or missing in a compound name. CHEMorph can be used as a basis to develop (with interdisciplinary work from computational linguistics and biochemistry) an elaborate system, where fully specified terms are treated in a fully automatic manner and underspecified terminology will be taken care of by an interactive dialogue tool with experts to resolve remaining ambiguities. Such a system would not only be a basis for (semi-)automatic database curation, but also for named-entity recognition for text processing applications (e. g. information extraction or text mining as MedMiner[18] does), where terminology is still the major challenge.

To summarise, this project deals with the semantic processing of terminology in biomedical data in order to advance (semi-)automatic database curation. Its outcome will serve as a basis for further text mining applications and for text understanding, which do need a sophisticated processing and

[16] This part of the results was not generated by CHEMorph, but has been added manually for demonstration. Until now, the system's implementation was divided into two parts: analysis of sugar names including generation of their SMILES strings and analysis of nonsugar names including generation of their corresponding classes. These components are currently being merged.

[18] `http://discover.nci.nih.gov/textmining`

interpretation of the complex biochemical terminology.

Acknowledgements

Thanks to the SDBV group of EML Research for their support. Stefanie Anstein is funded by the Klaus Tschira Foundation gGmbH, Heidelberg (http://www.kts.villa-bosch.de).

References

Stefanie Anstein and Gerhard Kremer. 2005. Analysing Names of Organic Chemical Compounds – From Morpho-Semantics to SMILES Strings and Classes. Master's thesis, University of Stuttgart. Web version available at http://www.ims.uni-stuttgart.de/lehre/studentenarbeiten/fertig/Diplomarbeit_Anstein_Kremer.pdf.

Olivier Bodenreider and Anita Burgun. 2002. What needs to be represented in a biomedical ontology? Ontological Spring: An Introductory Workshop in Ontology, Apr. Institute for Formal Ontology and Medical Information Science, Naumburg, Germany.

Juliane Fluck, Martin Hofmann, Udo Hahn, Joachim Wermter, and Stefan Schulz. 2005. Text Mining in den Life Sciences. *DZKF*, 5/6:20–26.

Rob Gaizauskas. 1998. Evaluation in Language and Speech Technology. *Journal of Computer Speech and Language*, 12(3):249–262.

Ciprian V. Gerstenberger. 2001. Semantische Analyse von Namen Organischer Verbindungen oder Was Bedeutet 3,3'-Ureylen-dibenzamidin? Master's thesis, University of Stuttgart.

IUPAC Commission on Nomenclature of Organic Chemistry. 1993. *A Guide to IUPAC Nomenclature of Organic Compounds (Recommendations 1993)*. Blackwell Scientific publications. Web version retrieved November 2005, from http://www.acdlabs.com/iupac/nomenclature.

IUPAC-IUBMB Joint Commission on Biochemical Nomenclature. 1996. Nomenclature of Carbohydrates (Recommendations 1996). *J. Pure Appl. Chem.*, 68:1919–2008. Web version retrieved November 2005, from http://www.chem.qmul.ac.uk/iupac/2carb/; also available as PDF-File from http://www.iupac.org/publications/pac/1996/pdf/6810x1919.pdf.

Uwe Reyle. 2006. Understanding Chemical Terminology. *Terminology*. In print. Retrieved January 2006, from ftp://ftp.ims.uni-stuttgart.de/pub/papers/reyle/terminology.pdf.

Isabel Rojas, Luca Bernardi, Esther Ratsch, Renate Kania, Ulrike Wittig, and Jasmin Šarić. 2002. A Database System for the Analysis of Biochemical Pathways. *J. In Silico Biol.*, 2,0007(2):75–86.

Irena Spasic, Goran Nenadic, and Sophia Ananiadou. 2004. Learning to Classify Biomedical Terms through Literature Mining and Genetic Algorithms. In Z.R. Yang et al., editor, *Intelligent Data Engineering and Automated Learning - IDEAL 2004*, LNCS 3177, pages 345–351. Springer.

Robert Stevens, Patricia G. Baker, Sean Bechhofer, Gary Ng, Alex Jacoby, Norman W. Paton, Carole A. Goble, and Andy Brass. 2000. TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. *Bioinformatics*, 16(2):184–186.

David Weininger. 1988. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.*, 28:31–36.

Ulrike Wittig, Andreas Weidemann, Renate Kania, Christian Peiss, and Isabel Rojas. 2004. Classification of Chemical Compounds to Support Complex Queries in a Pathway Database. *J. Comp. Funct. Genom.*, 5(2):156–162.