Flache und semantische Verarbeitung von Namen biochemischer Verbindungen

Henriette Engelken*[1,2], Martin Golebiewski*[1], Meik Bittkowski[1], Fritz Hamm[2], Jasmin Saric[1,3], Ulrike Wittig[1], Wolfgang Müller[1], Uwe Reyle[2] und Isabel Rojas[1]

[1]EML Research gGmbH, Schloss-Wolfsbrunnenweg 33, 69118 Heidelberg
[2]Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, Azenbergstraße 12, 70174 Stuttgart
[3]jetzt: Boehringer Ingelheim Pharma GmbH & Co. KG, Birkendorferstr. 65, 88397 Biberach / Riß
*E-Mail: engelken@eml-r.org, golebiewski@eml-r.org

Abstract: Termverarbeitung in der Domäne der Biowissenschaften beinhaltet für Information Retrieval, Data Mining, Information Extraction und für die Pflege wissenschaftlicher Datenbanken eine Reihe von Herausforderungen. Wir beschreiben diese Problematik und stellen unsere beiden Lösungsansätze vor. Dabei handelt es sich zum einen um ein normalisiertes Namensmatching und zum anderen um eine semantische Namensverarbeitung.

1 Einleitung

Die Biowissenschaften erleben gegenwärtig ein rasantes Anwachsen von Daten und Literatur. Entsprechend passen sich *Information Retrieval* (IR), *Data Mining* (DM) und *Information Extraction* (IE) an die speziellen Anforderungen in dieser Domäne an. Termverarbeitung ist dabei von großer Bedeutung. Es kann lt. Krauthammer und Nenadic ([KN04]) differenziert werden in Erkennung (*named entity recognition*), Identifizierung (*matching*, *grounding*) und Klassifizierung von Termen. Zu den Termen, welche besonders häufig in wissenschaftlichen Publikationen, Datenbanken und Patenten vorkommen und die wesentlich für das Verständnis und die Interpretation des beschriebenen Inhalts sind, gehören Molekülnamen. Neben Bezeichnungen für makromolekulare Strukturen (Genen, Proteinen, etc.) spielen insbesondere auch die Namen niedermolekularer (bio-)chemischer Verbindungen, der so genannten „small molecules“ eine wichtige Rolle.[1]

Eine eindeutige Bezeichnung einer chemischen Verbindung, die auch die Denkweise eines Chemikers widerspiegelt, ist ihre chemische Struktur. Auch wenn man diese mit

[1] Im vorliegenden Text, ebenso wie bei unseren Projekten, beschränken wir uns auf englischsprachige Namen chemischer Verbindungen. Die Problematik, ebenso wie unsere beiden Lösungsansätze, lassen sich jedoch auch auf Verbindungsnamen in anderen Sprachen wie z. B. dem Deutschen übertragen.

gängigen Strukturrepräsentationsformaten, wie z. B. SMILES ([Wei88]), als String darstellen kann, werden in Publikationen und zum Teil auch in Datenbanken jedoch meist ausschließlich Namen verwendet. Mit dem Ziel, systematische und eindeutige Namen für chemische Verbindungen zu erhalten, hat die *International Union of Pure and Applied Chemistry (IUPAC)* ([Iup93]) eine eigene Nomenklatur entwickelt. Die Nomenklaturregeln beschreiben die Generierung chemischer Verbindungsnamen, indem sie Morpheme (Namensbausteine mit bestimmter Bedeutung) und syntaktische Regeln zu ihrer Zusammensetzung angeben. Die Morpheme können systematisch (*meth, eth, prop, but, hex, ...*) oder trivial (*fructose* statt *arabino-hex-2-ulose*) sein. Die Morpheme eines Namens beschreiben die Bausteine aus denen das bezeichnete Molekül zusammengesetzt ist, d. h. die Atome und Bindungen.

Zwar beschreiben diese Nomenklaturregeln den eindeutigen Gebrauch der Namen und es gibt meist nur eine gültige IUPAC-konforme Bezeichnung eines Moleküls, dennoch ist der stringente Gebrauch dieser Regeln insbesondere in der Literatur wenig verbreitet. Das hat verschiedene Gründe, unter anderem werden bei komplexeren Molekülstrukturen die systematischen Bezeichnungen meist sehr lang und unhandlich, so dass häufig Trivialnamen zur Beschreibung herangezogen werden. Weitere Fälle von Synonymie, d. h. Bedeutungsgleichheit unterschiedlicher Namen, können sich zudem beispielsweise aus unterschiedlicher Reihenfolge bestimmter Namensbestandteile oder durch die Verwendung synonymer Namensbestandteile ergeben. So sind z. B. *butan-1-ol* und *1-butanol* synonym, ebenso wie *glyceraldehyde-3-phosphate* und *3-phosphoglyceraldehyde*. Neben Synonymie ist Unterspezifikation als Charakteristikum zu nennen. Unterspezifikation beschreibt das Problem fehlender Informationen um ein einzelnes bestimmtes Referenzobjekt zu identifizieren. Dies ist bei Klassennamen wie *alcohol* oder *alkene* offensichtlich, da sie nicht eine einzelne Molekülstruktur, sondern eine ganze Menge von Molekülstrukturen bezeichnen. Ebenso können jedoch auch in systematisch zusammengesetzten Namen Bestandteile fehlen. Beim unterspezifizierenden Namen *butanol* z. B. fehlt ein Lokant, weshalb er die Menge {*butan-1-ol*, *butan-2-ol*} bezeichnet.

Die genannten Charakteristika der Verbindungsnamen erschweren ihre manuelle ebenso wie ihre automatische Identifizierung und Klassifizierung. Dadurch führen sie zu einer Reihe von Problemen bei Information Retrieval, Information Extraction und bei der Datenintegration in wissenschaftlichen Datenbanken:

- geringe Vollständigkeit (*recall*) bei IR / IE durch Synonyme und fehlende Klassennamen-Auflösung oder Klassifizierung
- Datenbanken-Pflege: Redundanzen und erschwerte Verlinkung zu anderen Datenbanken durch Synonyme; kein Matching zu existierenden Taxonomien / Ontologien durch fehlende Klassifizierung
- generelle Reaktionsbeschreibungen (IUBMB): ohne Klassennamen-Auflösung und Klassifizierung können Reaktionsteilnehmer nicht identifiziert werden.

Um die genannten Probleme zu beheben, bieten wir zwei verschiedene Lösungsansätze an. Kapitel 2 beschreibt unser normalisiertes Namensmatching und Kapitel 3 stellt unser Projekt zu einer semantischen Verarbeitung der Namen vor.

2 Normalisiertes Namensmatching

Um das Problem der Identifizierung (*matching*, *grounding*) chemischer Verbindungsnamen zu lösen, ist Synonymerkennung ein vielversprechender Ansatz. Liegt eine Referenzliste oder Referenzdatenbank mit Namenseinträgen vor, matcht ein exakter Stringvergleich nur zu einem Teil der tatsächlich bedeutungsgleichen Namen und somit nur zu einem Teil der passenden Einträge. Grund dafür ist, dass viele Einträge nicht den gesuchten Eingabenamen, sondern nur einen synonymen Namen enthalten, welcher in seiner Schreibweise erheblich abweichen kann.

Um diese Synonyme zu erkennen, und somit eine höhere Vollständigkeit (*recall*) beim Namensmatching zu erreichen, haben wir ein Programm zum normalisierten Namensmatching entwickelt [GSE+09]. Mit diesem Programm kann ein gegebener biochemischer Name gegen eine Referenzdatenbank (z. B. SABIO-RK[2], PubChem[3] oder ChEBI[4]) gematcht werden, wenn diese ebenfalls im normalisierten Format vorliegt.[5]

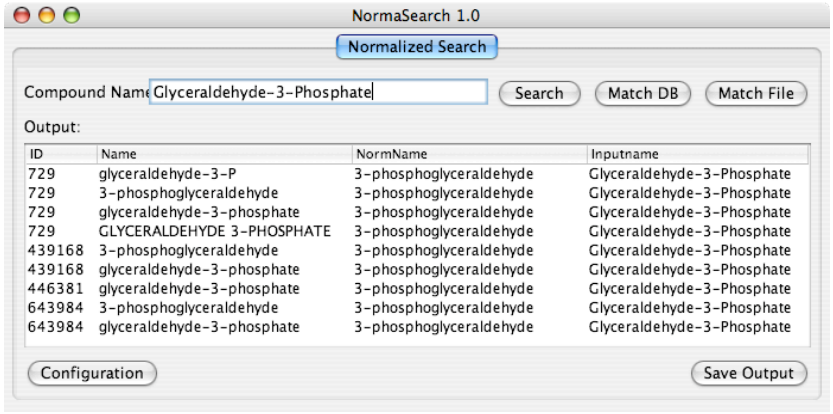

Abbildung 1: Screenshot normalisiertes Namensmatching

Der Kern der Applikation ist die regelbasierte Namensnormalisierung. Die Regeln repräsentieren Expertenwissen über synonyme biochemische Namen und arbeiten mit Mustererkennung (*pattern matching*), d. h. mit regulären Ausdrücken. Die angewandten Regeln normalisieren zunächst Varianten bei der Verwendung von Bindestrichen, Leerzeichen und Ähnlichem. Darüber hinaus beinhalten die Regeln morphosyntaktische Normalisierungen - unter anderem eine Sortierung von Bezeichnungen der Substituenten (Bausteine) eines Moleküls (z. B. *2-hydroxy-1-aminobutane* zu *1-amino-2-hydroxybutane*), eine Normalisierung von Suffixen (Endungen) zu gleichbedeutenden Präfixen (Vorsil-

---

[2] SABIO-RK: `http://sabio.villa-bosch.de/SABIORK/` (12. Mai 2009).
[3] PubChem: `http://pubchem.ncbi.nlm.nih.gov/` (12. Mai 2009).
[4] ChEBI: `http://www.ebi.ac.uk/chebi/init.do` (12. Mai 2009).
[5] In unserem Web-Interface (`http://sabiork.villa-bosch.de/normaWeb`) stellen wir derzeit SABIO-RK, PubChem und ChEBI als normalisierte Referenzdatenbanken zur Verfügung.

ben) (z. B. *–phosphate* zu *phospho-*) und das Ersetzen weiterer synonymer Namensbestandteile.

Zusätzlich zu den manuell aufgestellten Expertenregeln haben wir eine Sammlung von gegenwärtig etwa 8000 synonymen Substring-Paaren generiert. Gewonnen wurden diese Substring-Paare mit einem statistischen Verfahren, welches wir auf eine verfügbare Namensliste (PubChem) von mehreren Millionen Verbindungsnamen angewendet haben. Für die Normalisierung der Namen wird jeweils der erste Eintrag eines solchen Synonympaars durch den zweiten ersetzt. Beispielsweise kann, da *(2-methylphenyl)* und *o-Tolyl* als synonyme Substrings gelistet sind, *(2-methylphenyl)acetonitrile* normalisiert werden zu *o-Tolylacetonitrile* und somit zu einem Datenbankeintrag matchen, welcher nur die letztere Namensform führt.

Normalisiertes Namensmatching in der von uns umgesetzten Form bietet somit die Möglichkeit synonyme Namen zu matchen, welche durch exaktes Stringmatching nicht gefunden worden wären. Die normalisierten Namen stellen dabei lediglich eine interne Repräsentation für ein anschließendes normalisiertes Namensmatching dar, und sind weder eine korrekte Bezeichnung nach IUPAC-Regeln, noch ein systematischer Name für die beschriebene Verbindung. Auch wenn Qualität und Quantität der Matchingergebnisse bei einem solchen Ansatz stets abhängig von Qualität und Quantität der verfügbaren Referenz-Namenslisten sind, stellt unsere Implementierung einen erheblichen Schritt zur Lösung der in der Einleitung genannten Probleme dar, welche sich durch Synonymie von Verbindungsnamen ergeben.

3 Semantische Namensverarbeitung

Eine eindeutige Bezeichnung einer chemischen Verbindung ist ihre chemische Struktur. Ist diese zusätzlich zu einem gegebenen Namen bekannt, lösen sich einige der in der Einleitung beschriebenen Probleme: Synonymie von Namen lässt sich eindeutig erkennen, wenn verschiedene Namen dieselbe chemische Struktur bezeichnen. Zudem lässt sich eine chemische Struktur, im Gegensatz zu einem Namen, anhand ihrer funktionellen Gruppen klassifizieren (z. B. mit [WWK+04]). Funktionelle Gruppen sind bestimmte Atomgruppen, welche die charakteristischen Eigenschaften eines Moleküls festlegen und somit für eine chemische Klassifizierung entscheidend sind. Informationen über die vorhandenen funktionellen Gruppen werden durch die Morpheme eines Namens spezifiziert. So bezeichnet z. B. *hydroxy*, als Präfix vor einem Namen gebraucht, das Vorhandensein einer Hydroxygruppe (OH), während *dehydroxy* die Entfernung einer Hydroxygruppe anzeigt. Gdw. eine biochemische Verbindung eine solche OH-Gruppe enthält, kann sie als ein *alcohol* klassifiziert werden.

Unsere Entwicklung eines Systems zur semantischen Namensverarbeitung (Details vgl. [Eng09]) hat demnach folgende Ziele: Zu einem gegebenen chemischen Verbindungsnamen soll die Molekülstruktur ermittelt werden, welche er bezeichnet. Ebenso soll der Name chemisch klassifiziert werden. Nicht nur für voll spezifizierende Namen sollen diese Ergebnisse erreicht werden, sondern auch für unterspezifizierende Namen und Klassennamen.

Wir folgen Reyle ([Rey06]) mit der Beobachtung, dass jeder chemische Verbindungsname als Beschreibung einer chemischen Struktur angesehen werden kann, da er Bedingungen (*constraints*) vorgibt, welche eine Molekülstruktur spezifizieren. Sogar für unterspezifizierende Namen gilt, dass sie zumindest einige Informationen, und somit Bedingungen (*constraints*), über die Struktur enthalten. Entsprechend ist constraintlogisches Programmieren (CLP) der Kern unseres Systems (Systemarchitektur vgl. Abbildung 2).

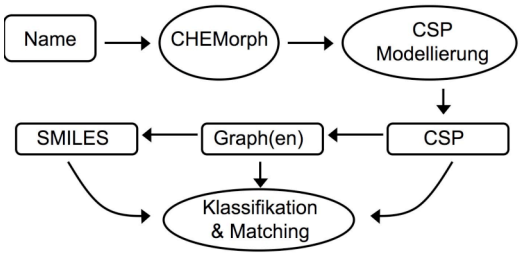

Abbildung 2: Systemarchitektur für semantische Namensverarbeitung

Der erste Schritt zur semantischen Namensverarbeitung ist eine linguistische Analyse durch den Parser CHEMorph ([KAR06]). Ergebnis dieser Analyse ist u. a. eine semantische Repräsentation des Verbindungsnamens, welche in Prädikat-Argument-Notation alle Informationen enthält, die der Name über die Struktur liefert. Dabei werden syntaktische Zusammenhänge, wie die chemische Basisstruktur und ihre zugehörigen Modifikationen, bereits erkannt. Diese semantische Repräsentation dient uns als Basis für die automatische Modellierung eines *constraint satisfaction* Problems (CSPs). Ein solches CSP repräsentiert das Problem, die chemische(n) Struktur(en) zu finden, welche von einem Namen bezeichnet werden. Ein CSP muss dabei auch Standardwissen (*default knowledge*) über die Namen und Strukturen beinhalten, welches in den Namen nicht explizit angegeben ist. Zu diesem Standardwissen gehören beispielsweise die Valenzen von Atomen, wodurch die mögliche Zusammensetzung einer Molekülstruktur eingeschränkt wird. Da eine chemische Struktur als beschrifteter Graph repräsentiert werden kann, modellieren wir CSPs als Sammlung von Constraints über Graphenvariablen. Ein Constraintlöser für Graphen (GRASPER [VA07]) bestimmt als Lösungen im Falle eines voll spezifizierenden Namens exakt einen Graphen. Bei unterspezifizierenden und Klassennamen jedoch kann der Constraintlöser alle Graphen ausgeben, welche mit den durch den Namen vorgegebenen Constraints in Einklang stehen.

Um nun Matching, was eine Synonymerkennung einschließt, und eine Klassifizierung zu erzielen, kann ein ermittelter Graph in ein gängiges Repräsentationsformat für Molekülstrukturen (z. B. SMILES) übersetzt werden. Diese eindeutige Strukturrepräsentation kann direkt mit anderen Strukturen verglichen werden und klassifiziert werden. Indem ein CSP, d. h. eine Constraintsammlung, welches die Bedeutung des Eingabenamens repräsentiert, direkt zum Matching und zur Klassifizierung dienen kann, kann darüber hinaus Matching und eine Klassifizierung auch bei Unterspezifikation und Klassennamen erreicht werden, wenn nicht alle Strukturen (Graphen), die vom Name bezeichnet werden, aufgezählt werden sollen.

Literaturverzeichnis

[Iup93] IUPAC. Commission on the Nomenclature of Organic Chemistry. A Guide to IUPAC Nomenclature of Organic Compounds (Recommendations 1993). Blackwell Scientific Publications, Oxford, 1993.

[Eng09] Henriette Engelken. A System for Semantic Analysis of Chemical Compound Names. In Proceedings of the ACL 2009 Student Research Workshop, Singapore, Malaysia, 2–7 August 2009. Association for Computational Linguistics, 2009.

[GSE+09] Martin Golebiewski, Jasmin Saric, Henriette Engelken, Meik Bittkowski, Ulrike Wittig, Wolfgang Müller, Isabel Rojas. Normalization and Matching of Chemical Compound Names. Available from Nature Precedings (http://dx.doi.org/10.1038/npre.2009.3322.1), 2009.

[KAR06] Gerhard Kremer, Stefanie Anstein und Uwe Reyle. Analysing and Classifying Names of Chemical Compounds with CHEMorph. In Sophia Ananiadou und Juliane Fluck, Hrsg., Proceedings of the Second International Symposium on Semantic Mining in Biomedicine, Friedrich-Schiller-Universität Jena, Germany, 2006, Seiten 37–43, 2006.

[KN04] Michael Krauthammer und Goran Nenadic. Term Identification in the Biomedical Literature. Journal of Biomedical Informatics, 37(6):512–526, 2004.

[Rey06] Uwe Reyle. Understanding Chemical Terminology. Terminology, 12(1):111–136, 2006.

[VA07] Ruben Viegas und Francisco Azevedo. GRASPER: A Framework for Graph CSPs. In Jimmy Lee und Peter Stuckey, Hrsg., Proceedings of the Sixth International Workshop on Constraint Modelling and Reformulation (ModRef ’07), Providence, Rhode Island, USA, 2007.

[Wei88] David Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of Chemical Information and Computer Sciences, 28(1):31–36, 1988.

[WWK+04] Ulrike Wittig, Andreas Weidemann, Renate Kania, Christian Peiss und Isabel Rojas. Classification of chemical compounds to support complex queries in a pathway database. Comparative and Functional Genomics, 5:156–162, 2004.