Matching of Chemical Compound Names by Rule Based Name Normalization

Martin Golebiewski[*], Jasmin Šarić[*,#], Henriette Engelken[*], Meik Bittkowski[*], Ulrike Wittig[*], Wolfgang Müller[*], Isabel Rojas[*]

[*] Scientific Databases and Visualization, EML Research gGmbH, Heidelberg, Germany
[#] Present address: Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach, Germany

http://sabiork.villa-bosch.de/normaWeb/

Background

SABIO-RK (http://sabio.villa-bosch.de/SABIORK), a database system that we have developed to provide coherent experimental data, offers information about biochemical reactions and their corresponding kinetics [1,2]. It is populated by merging data derived from public databases like KEGG (Kyoto Encyclopedia of Genes and Genomes [3]) and manually extracted from literature. The consistent integration of the collected data is indispensable to make it comparable. However, the heterogeneity of the data described in the literature causes obstacles for data integration. Standardization methods can be designed to increase the consistency of the data. With this aim, we have developed an application which detects and matches synonymic names of chemical compounds and thereby facilitates the bundling of corresponding data referring to the same compound.

Terminology of Chemical Compounds

Synonymous notations of chemical compounds

Trivial name and systematic chemical description

Valproic acid = 2-Propylpentanoic acid

Different parts of the molecule could be considered as lead structure

Acetylphenol = Phenylacetate

Abberrant order of the substituents of a lead structure (prefixes)

2-Amino-6-methyl-4-pyrimidol = 6-Methyl-2-amino-4-pyrimidol

Description of substituents as prefix (like *amino-*) or suffix (like *–amine*)

2-Aminopropane = Propan-2-amine

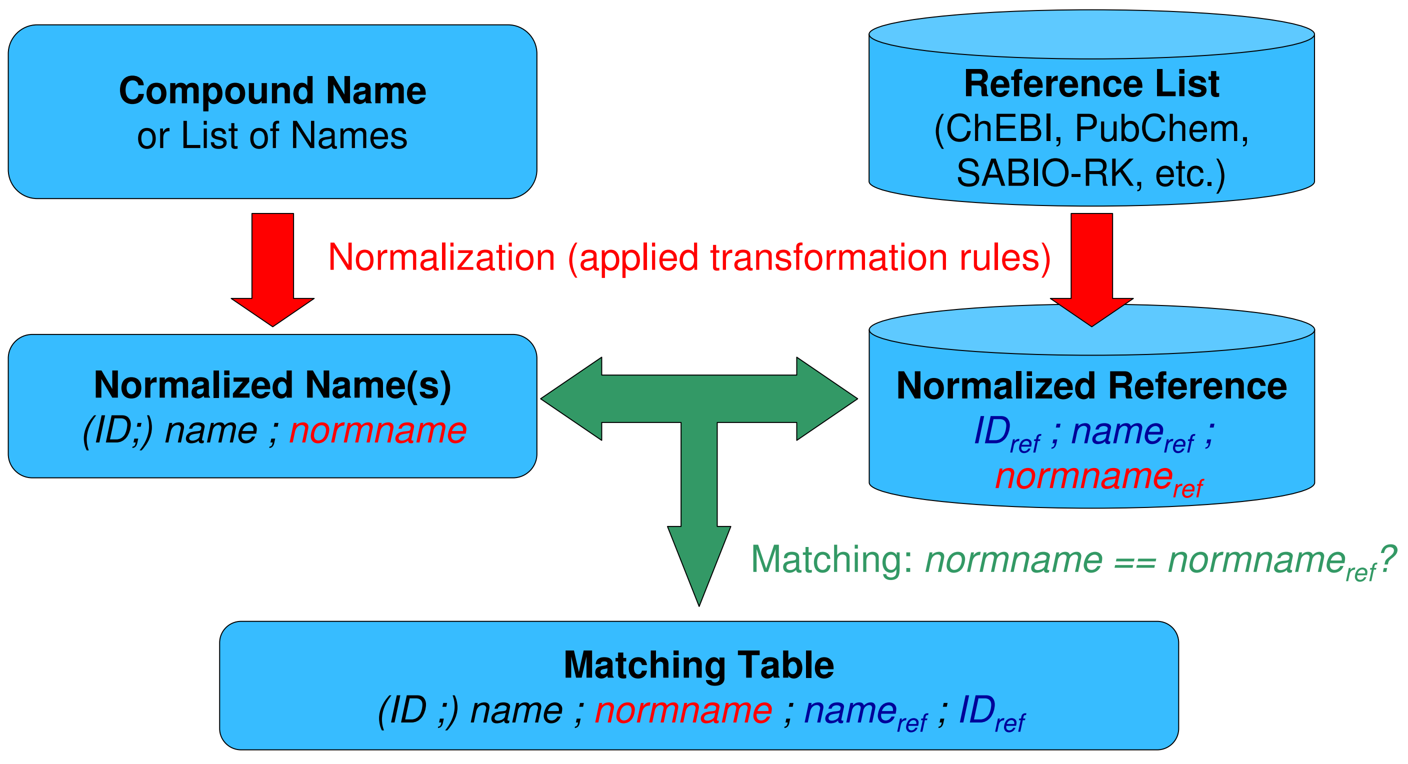
2-Methylpropan-2-ol = 2-Hydroxy-2-methyl-propane

Different nomenclature systems (e.g. abberrant order of the morphems)

2-Amino-6-methyl-4-pyrimidol = 2-Amino-6-methylpyrimidin-4-ol

A chemical compound can have many different synonymous names - trivial, as well as systematic names. Hence, the identification of a chemical compound solely based on its name requires comprehensive chemical knowledge and very often extensive searches in chemical databases. However, this unambiguous identification is crucial for the integration of biochemical data extracted from literature, as many publications exclusively describe a compound by its name.

Normalization of Chemical Compound Names



Compound Name or List of Names

Reference List (ChEBI, PubChem, SABIO-RK, etc.)

Normalization (applied transformation rules)

Normalized Name(s)
(ID;) name ; normname

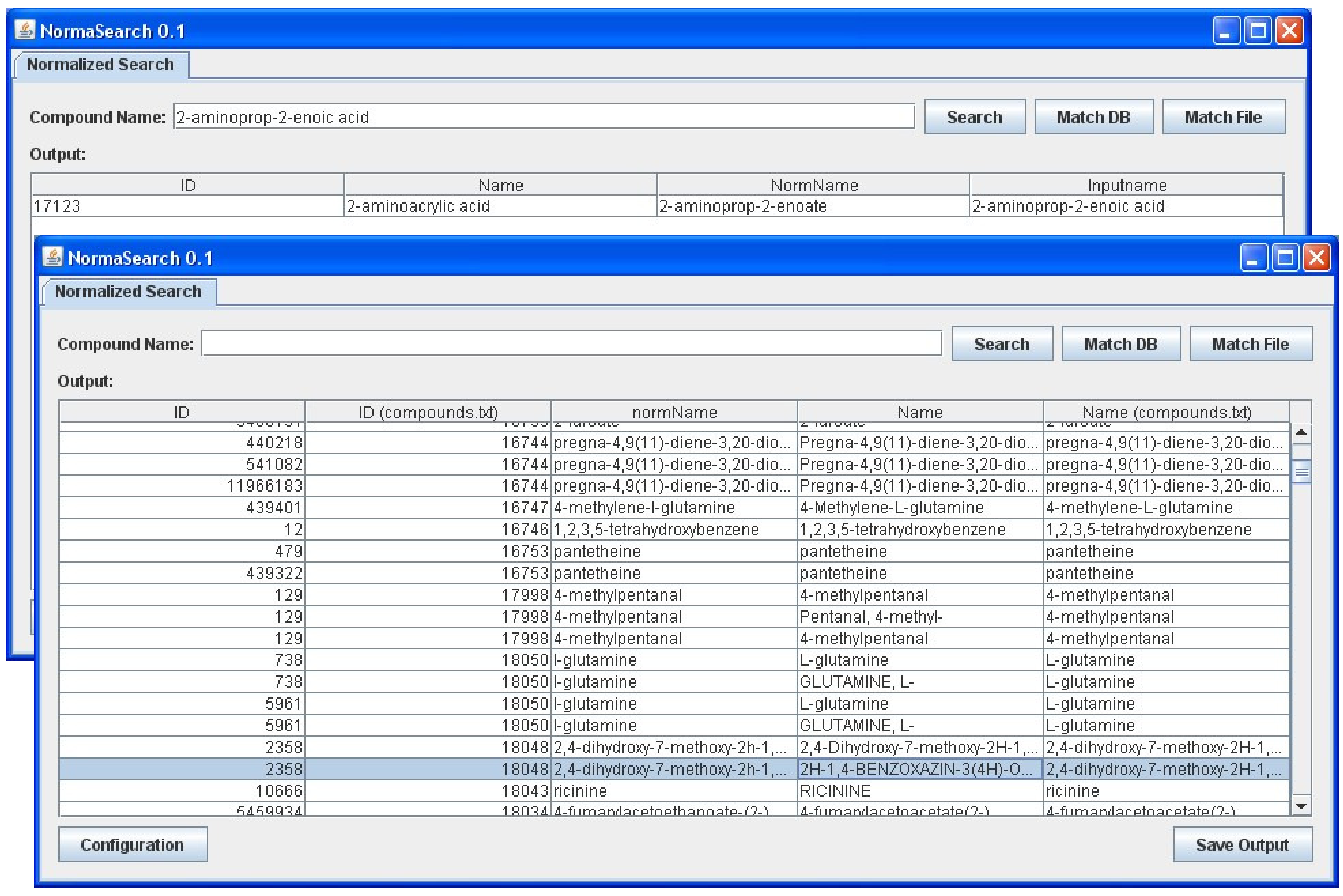
Normalized Reference
ID_{ref} ; $name_{ref}$; $normname_{ref}$

Matching: *normname == $normname_{ref}$?*

Matching Table
(ID ;) name ; normname ; $name_{ref}$; ID_{ref}

The tool that we have developed applies transformation rules to systematically normalize the notation of chemical compound names. Subsequently, matching of synonymous names is achieved by comparison of the normalized name forms. The normalization rules include, among others, reordering of substituent descriptions in the name and replacement of synonymous name constituents (e.g. equivalent trivial names). Matching of conjugated acid-base pairs is optional for biochemicals.

Chemical Compound Name Matching



Compound Name: 2-aminoprop-2-enoic acid

Output:

ID	Name	NormName	Inputname
17123	2-aminoacrylic acid	2-aminoprop-2-enoate	2-aminoprop-2-enoic acid

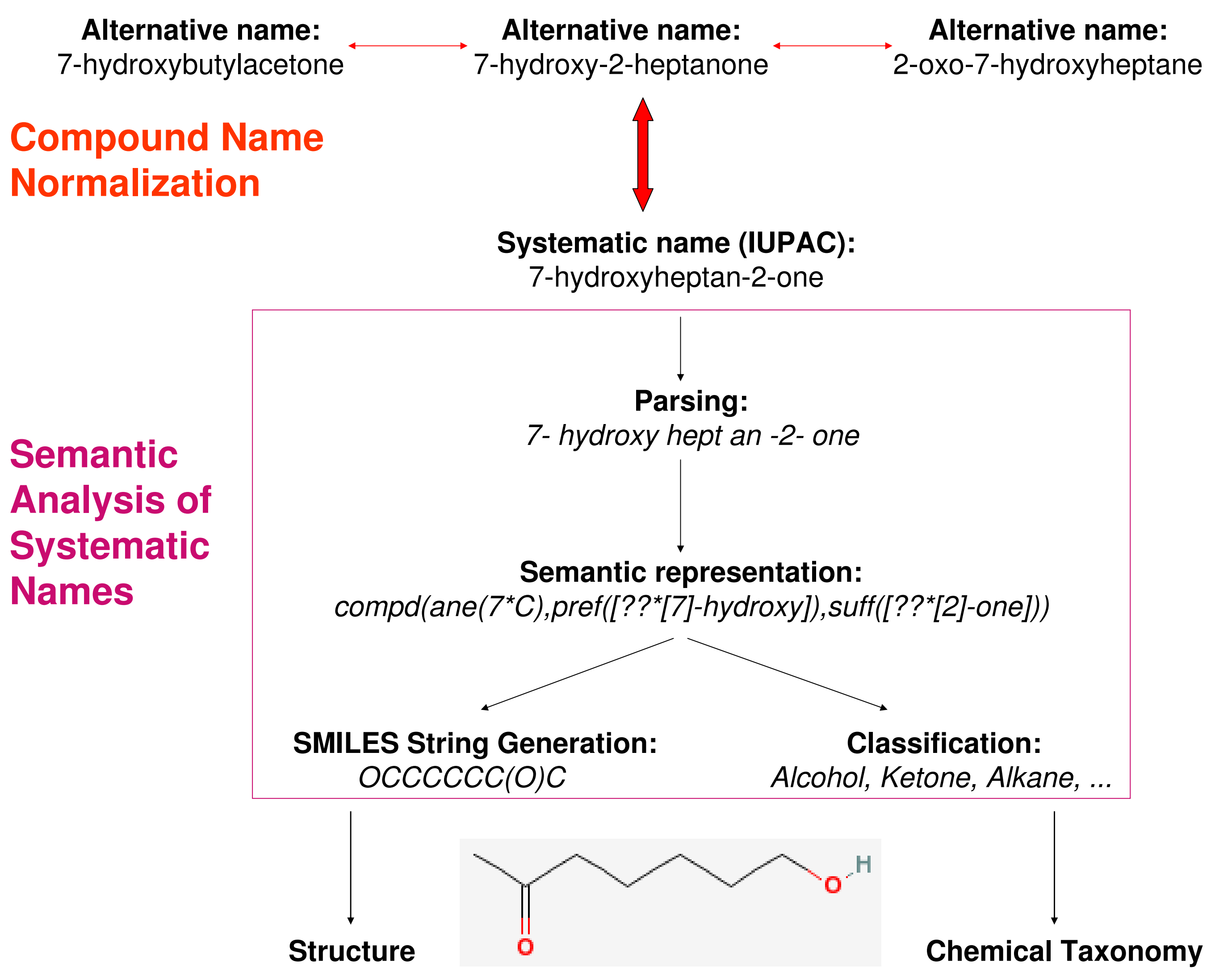
Compound Name:

Output:

ID	ID (compounds.txt)	normName	Name	Name (compounds.txt)
440218	16744	pregna-4,9(11)-diene-3,20-dio...	Pregna-4,9(11)-diene-3,20-dio...	pregna-4,9(11)-diene-3,20-dio...
541082	16744	pregna-4,9(11)-diene-3,20-dio...	Pregna-4,9(11)-diene-3,20-dio...	pregna-4,9(11)-diene-3,20-dio...
11966183	16744	pregna-4,9(11)-diene-3,20-dio...	Pregna-4,9(11)-diene-3,20-dio...	pregna-4,9(11)-diene-3,20-dio...
439401	16747	4-methylene-l-glutamine	4-Methylene-L-glutamine	4-methylene-L-glutamine
12	16746	1,2,3,5-tetrahydroxybenzene	1,2,3,5-tetrahydroxybenzene	1,2,3,5-tetrahydroxybenzene
479	16753	pantetheine	pantetheine	pantetheine
439322	16753	pantetheine	pantetheine	pantetheine
129	17998	4-methylpentanal	4-methylpentanal	4-methylpentanal
129	17998	4-methylpentanal	Pentanal, 4-methyl-	4-methylpentanal
129	17998	4-methylpentanal	4-methylpentanal	4-methylpentanal
738	18050	l-glutamine	L-glutamine	L-glutamine
738	18050	l-glutamine	GLUTAMINE, L-	L-glutamine
5961	18050	l-glutamine	L-glutamine	L-glutamine
5961	18050	l-glutamine	GLUTAMINE, L-	L-glutamine
2358	18048	2,4-dihydroxy-7-methoxy-2h-1,...	2,4-Dihydroxy-7-methoxy-2H-1,...	2,4-dihydroxy-7-methoxy-2H-1,...
2358	18048	2,4-dihydroxy-7-methoxy-2h-1,...	2H-1,4-BENZOXAZIN-3(4H)-O...	2,4-dihydroxy-7-methoxy-2H-1,...
10666	18043	ricinine	RICININE	ricinine
5459934	18034	4-fumarylacetoethanoate-(2-)	4-fumarylacetoacetate(2-)	4-fumarylacetoacetate(2-)

By use of these methods, the tool is capable of normalizing a given name of a chemical compound and matching it against names in (bio-)chemical databases, like KEGG (http://www.genome.jp/kegg), ChEBI (http://www.ebi.ac.uk/chebi), PubChem (http://pubchem.ncbi.nlm.nih.gov), or SABIO-RK, even when there is no exact name-to-name-match (upper screenshot). The tool is also able to match a complete list of compound names against these databases which makes it useful for the automatic cross-annotation of chemical data in databases (lower screenshot).

The Future: From Name to Structure
Combining Name Matching and Semantic Analysis



Alternative name: 7-hydroxybutylacetone

Alternative name: 7-hydroxy-2-heptanone

Alternative name: 2-oxo-7-hydroxyheptane

Compound Name Normalization

Systematic name (IUPAC): 7-hydroxyheptan-2-one

Semantic Analysis of Systematic Names

Parsing: *7- hydroxy hept an -2- one*

Semantic representation: *compd(ane(7*C),pref([??*[7]-hydroxy]),suff([??*[2]-one]))*

SMILES String Generation: *OCCCCCC(O)C*

Classification: *Alcohol, Ketone, Alkane, ...*

Structure

Chemical Taxonomy

After normalization, synonymous notations could potentially be matched to the corresponding systematic name as defined by the International Union of Pure and Applied Chemistry (IUPAC). When combined with our approach to construct chemical structures from systematic names (based on CHEMorph [4]), notations could be translated into a chemical structure (SMILES) and classified by functional groups [5], resulting in the unambiguous identification of these compounds.

References

1) Wittig U, Golebiewski M, Kania R, Krebs O, Mir S, Weidemann A, Anstein S, Saric J, Rojas I: *„SABIO-RK: integration and curation of reaction kinetics data"*, Lecture Notes in Bioinformatics, 4075: 94-103 (2006)

2) Rojas I, Golebiewski M, Kania R, Krebs O, Mir S, Weidemann A, Wittig U: *„Storing and annotating of kinetic data"*, In silico biology, 7: S37-44 (2007)

3) Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: *„From genomics to chemical genomics: new developments in KEGG"*, Nucleic Acids Research, 34: D354-357 (2006)

4) Kremer G, Anstein S, Reyle U: *„Analysing and Classifying Names of Chemical Compounds with CHEMorph"*, in S. Ananiadou and J. Fluck: Proceedings of the Second International Symposium on Semantic Mining in Biomedicine: 37-43 (2006)

5) Wittig U, Weidemann A, Kania R, Peiss C, Rojas I: *„Classification of chemical compounds to support complex queries in a pathway database"*, Comparative and Functional Genomics, 5: 156-162 (2004)

Contact: golebiewski@eml-r.org