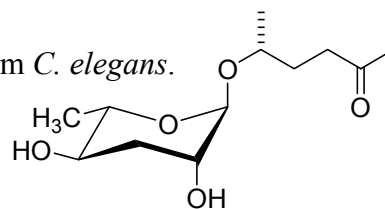


SMMID's: Unique identifiers for small molecule metabolites.

1. The issue: small molecules/secondary metabolites are referred to by a plethora of names and abbreviations. Some compounds are referred to by more than 10 different names, and in certain cases different compounds are referred to by the same name. Significantly, there is no established system for naming newly identified metabolites that would permit *database searching for small molecules metabolites in the same manner as for genes*.

Example I: new signaling molecule (dauer, sex pheromone) from *C. elegans*.



Incomplete list of names for this compound:

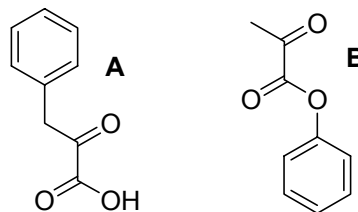
5-O-ascarylosyl-5R-hydroxy-2-hexanone
5-O-(3'S,5'R-dihydroxy-2',3',5',6'-tetrahydro-4'H-pyranyl)-5R-hydroxy-2-hexanone
1-O-(1'-methyl-5'-oxopentyl)-ascarylose
1-O-(1'-methyl-5'-oxopentyl)-2S,3S,5R-trihydroxy-2,3,5,6-tetrahydro-4H-pyran
1-O-(5-oxo-2-hexyl)-(2S,3S,5R-trihydroxy-2,3,5,6-tetrahydro-4H-pyran
“ascaroside C6”
“C6”

Example II: ambiguity of existing nomenclature systems.

Case: phenyl pyruvate. Structure **A** shows “phenylpyruvic acid”, frequently and somewhat misleadingly referred to as “phenyl pyruvate”, which based on IUPAC nomenclature would denote compound **B**. However, referring to the sodium salt of **A** as “sodium phenylpyruvate” would be considered correct: “phenylpyruvate” denotes a salt or ester of phenylpyruvic acid, whereas “phenyl pyruvate” would refer to the phenyl ester of pyruvic acid.

Incomplete list of names used for structure **A**:

phenylpyruvic acid
phenyl-pyruvic acid
phenyl pyruvate
phenylpyruvate
3-phenylpyruvic acid
2-oxo-3-phenylpropanoic acid
3-phenyl-2-oxopropanoic acid
2-oxo-3-phenylpropionic acid
3-phenyl-2-oxopropionic acid
 α -oxo-benzenepropanoic acid
 α -oxo-benzenepropionic acid
 α -oxo-benzenepropionic acid
 α -keto-benzenepropanoic acid
 α -keto-benzenepropionic acid



alpha-keto-benzenepropionic acid
2-oxo-3-phenylpropanoate
3-phenyl-2-oxopropanoate
2-oxo-3-phenylpropionate
3-phenyl-2-oxopropionate
(...)

Ambiguities and parallel usage of different names prevents development of effective text mining tools for small molecules/secondary metabolites. Effective processing and sharing of small molecules/secondary metabolite data in chemical biology and metabolomics is virtually impossible.

2. Problems with existing naming schemes.

CAS#: proprietary (American Chemical Society); many biological journals are not indexed by CAS; frequently derivatives not fully characterized chemically will not get indexed by CAS; cumbersome numbers, no recognition value

IUPAC names: very complicated to derive; complicated syntax; not always unambiguous; IUPAC names are often exceedingly long and complicated; thus unsuitable as identifier.

SMILES: very technical though precise; useful for as descriptor in databases but not as in-text small molecule identifier.

3. A new proposal for naming small molecules/secondary metabolites in biology: SMMID's

I. *Small molecules newly identified from an organism or a biological environment are assigned a unique **Small Molecule Metabolite Identifier (SMMID)** consisting of **four lower case non-italicized letters** that refer to the general structural class of the compound, followed by a pound sign and a number.*

Examples:

ascr#1 (formerly "daumone" or "C7" or "(6R-(tetrahydro-3'R,5'R-dihydroxy-6'S-methyl-2H-pyran-2'R-yloxy)-heptanoic acid")

ascr#2 (formerly "C6" or "5R-(tetrahydro-3'R,5'R-dihydroxy-6'S-methyl-2H-pyran-2'R-yloxy)-2-hexanone")

ascr#3 (formerly "C9", or "8R-(tetrahydro-3'R,5'R-dihydroxy-6'S-methyl-2H-pyran-2'R-yloxy)-2E-nonenic acid")

ascr#4 (formerly "nematone-1" or "5R-(3'-O-beta-D-glucosyl-tetrahydro-3'R,5'R-dihydroxy-6'S-methyl-2H-pyran-2'R-yloxy)-2-hexanone")

ascr#5 (formerly "C9-glycoside" or "name too long to spell out")

roid#1 (formerly "delta4-dafachronic acid" or "3-keto-4-cholestenoic acid")

roid#2 (formerly "delta7-dafachronic acid" or "3-keto-7,(5a)-cholestenoic acid")

iprs#12 (formerly 15-E_{2t}-Isoprostane)

cael#1 (formerly caeliferinA16:1)

Compare: ***daf-22*** (three-letters, italicized, lower case) or **DAF-22** (non-italicized, upper case)

All newly identified metabolites will be named by a central authority (for starters: Frank Schroeder and Lukas Mueller, METACYC SGN databases, Cornell University) by request (web page will soon be up) of the authors of any paper describing the compounds. Metabolites already in the literature could be named retroactively.

II. Stereoisomers are distinguished by the addition of a second numeral. Examples:

plac#1.1 (R-3-phenyllactic acid)

plac#1.2 (S-3-phenyllactic acid)

The first-discovered stereoisomer of any compound will be named with the ending ".1". i.e. "**xxxx#x.1**".

III. A database (maintained initially by Frank Schroeder and Lukas Mueller, Cornell University) will provide online support for the SMMID's, including

- Structure (structural drawing, SMILES)
- Compound ID (common names, CAS, Beilstein, IUPAC)
- Original reference

Databases such as **wormbase** and **MetaCyc** will feature entries for each SMMID, featuring:

- Structure (structural drawing, SMILES)
- Compound ID (common names, CAS, Beilstein, IUPAC)
- Links (pubchem, other databases)
- Original reference
- List of references that mention this compound, context information