Module 5: Identifying Chemical Entities

Learning Objectives

By the end of this module, students will:
1. review various chemical identifiers used for representing small molecules.
2. explain what common names, systematic names, and INN names are.
3. explain what SMILES, SMARTS and SMIRKS are.
4. explain what InChI and InChIKey are.
5. review SMILES specification rules.
6. compare and contrast SMILES and InChI.
7. demonstrate how to interpret SMILES, SMARTS, InChI strings into their corresponding chemical structures.

1. Chemical Names

Chemical names are the chemical identifiers that are most widely used in our everyday life. Some of them (such as water, iron, salt and so on) have been traditionally used for a very long time, while others have been introduced quite recently. Below are chemical names commonly used for organic and pharmaceutical compounds in Cheminformatics.

1.1 IUPAC names and related names (for organic compounds)

- **Systematic name**[1]: A chemical name generated according to the IUPAC nomenclature, which is a systematic method of naming chemical compounds, recommended by the International Union of Pure and Applied Chemistry (IUPAC) (http://www.iupac.org).
- **Common or trivial name**[2]**:** a non-systematic, traditional name that is widely used in industry as well as in academia.
- **Retained name**: a trivial name that may be used in the IUPAC nomenclature.
- **IUPAC name**: a systematic name that meets the recommended IUPAC rules. Strictly speaking, IUPAC names include both systematic and retained names, but the terms "IUPAC name" and "systematic name" are often used interchangeably in many publications.
- **Preferred IUPAC name**[3] **(PIN)**: a name that is preferred among two or more IUPAC names. In 2013, the IUPAC revised the rules for determining preferred IUPAC names, detailed in the so-called "IUPAC blue book".[4] However, this book is not freely available due to copyright. A previous version of the blue book is available here.

1.2 Trade and generic names (for pharmaceuticals)

- **Trade name (or brand name)**

  A "proprietary" name that a business uses for trading commercial products or services. Most prescription drugs placed on the market are given brand names to distinguish them as being produced and marketed exclusively by a particular manufacturer. In the United States, these names are usually registered as trademarks with the Patent Office. Registration gives the registrant certain legal rights with respect to the use of the name.

- **Generic name**

  A "non-proprietary" name that is a shorthand version of the drug's ingredient(s), used for generic drugs. Read this page[5] in the online Merck Manual about drug naming and the difference between generic and trade names.

- **International Nonproprietary Name (INN)**

  An official generic and non-proprietary name given to a pharmaceutical drug. INNs make communication more precise by providing a unique standard name for each active ingredient, to avoid prescribing errors. The INNs are available on this World Health Organization (WHO) web page[6]. More detailed information on the INNs is available on this Wikipedia page[7] as well as on the WHO guidance on INN.

  Some countries have their own set of unique nonproprietary names assigned to pharmaceuticals marketed in their territories. Examples are United States Adopted Names (USANs), British Approved Names (BANs), Japanese Accepted Names (JANs), Australian Approved Names (AANs), and French Approved non-proprietary names (Dénominations Communes Françaises, DCF). These names are not necessarily the same as INNs. For example, the USAN of Tylenol is "acetaminophen", while its INN is "paracetamol".

2. Registry Numbers and Database Identifiers

Chemical Abstracts Service (CAS)[8], a division of the American Chemical Society, collects publicly disclosed information on chemical substances (primarily from scientific literature) and organizes them into a database called CAS Registry. The CAS Registry can be searched using CAS products such as SciFinder[9] and STN[10]. In this database, unique identifiers are assigned to each chemical substance, providing an unambiguous way to identify a chemical substance when there are many possible systematic, generic, proprietary or trivial names. These unique identifiers are referred to as CAS Registry Numbers[11,12], CAS RNs, or CAS Numbers.

CAS Numbers are numeric identifiers that can contain up to ten digits, divided by hyphens into three parts: the first consisting from two up to seven digits, the second consisting of two digits, and the third consisting of a single digit. These numbers themselves have no inherent chemical meaning (e.g., information on chemical structure), but are assigned in sequential order to new substances found by CAS scientists for inclusion in the database.

While CAS Registry Numbers have been widely used, the CAS Registry is a proprietary resource that comes with a non-trivial fee. These days many databases provide their contents free of charge to the general public, and unique identifiers used in these databases are getting popular. Examples are PubChem CIDs (Compound IDs), ChemSpider IDs, and ChEMBL IDs. Details about these databases will be discussed in Module 6.

3. Line Notations

Line notations represent structures as a linear string of characters. They are widely used in Cheminformatics because computers can more easily process linear strings of data. Examples of line notations include the Wiswesser Line-Formula Notation (WLN)[13], Sybyl Line Notation (SLN)[14,15] and Representation of structure diagram arranged linearly (ROSDAL)[16,17]. Currently, the most widely used linear notations are the Simplified Molecular-Input Line-Entry System (SMILES)[18-21] and the IUPAC Chemical Identifier (InChI)[22-25], which are described below.

3.1. SMILES and related notation

SMILES

The Simplified Molecular-Input Line-Entry System (SMILES)[18-21] is a line notation for describing chemical structures using short ASCII strings. SMILES was developed in the late 1980s and implemented by Daylight Chemical Information Systems (Santa Fe, NM), but it is still widely used today. A detailed information on SMILES can be found in Chapter 3[26] of the Daylight Theory Manual as well as the SMILES tutorial[27].

SMILES Specification Rules

In SMILES, atoms are represented by their atomic symbols. The second letter of two-character atomic symbols must be entered in lower case. Each non-hydrogen atom is specified independently by its atomic symbol enclosed in square brackets, [ ] (for example, [Au] or [Fe]). Square brackets may be omitted for elements in the "organic subset" (B, C, N, O, P, S, F, Cl, Br, and I) if the proper number of "implicit" hydrogen atoms is assumed. "Explicitly" attached hydrogens and formal charges are always specified inside brackets. A formal charge is represented by one of the symbols + or -. Single, double, triple, and aromatic bonds are represented by the

symbols, -, =, #, and :, respectively. Single and aromatic bonds may be, and usually are, omitted. Here are some examples of SMILES strings.

C	Methane (CH_4)	COC	Dimethyl ether (CH3OCH3)
CC	Ethane (CH_3CH_3)	CCO	Ethanol (CH3CH2OH)
C=C	Ethene (CH_2CH_2)	CC=O	Acetaldehyde (CH3-CH=O)
C#C	Ethyne (CHCH)	C#N	Hydrogen Cyanide (HCN)
		[C-]#N	Cyanide anion

Branches are specified by enclosures in parentheses and can be nested or stacked, as shown in these examples.

CC(C)CO	Isobutyl alcohol (CH_3-CH(CH_3)-CH_2-OH)
CC(CCC(=O)N)CN	5-amino-4-methylpentanamide

Rings are represented by breaking one single or aromatic bond in each ring, and designating this ring-closure point with a digit immediately following the atoms connected through the broken bond. Atoms in aromatic rings are specified by lower cases letters. Therefore, cyclohexane and benzene can be represented by the following SMILES.

C1CCCCC1	Cyclohexane (C_6H_{12})
c1ccccc1	Benzene (C_6H_6)

Although the carbon-carbon bonds in these two SMILES are omitted, it is possible to deduce that the omitted bonds are single bonds (for cyclohexane) and aromatic bonds (for benzene). One can also represent an aromatic compound as a non-aromatic, KeKulé structure. For example, the following is a valid SMILES string for benzene.

C1=CC=CC=C1	Benzene (C_6H_6)

Note that aromaticity is not a measurable physical quantity, but a concept without a unanimous mathematical definition. As a result, different aromaticity detection algorithms often disagree with each other on whether a given molecule is aromatic or not, making it difficult to interchange information between databases that use different aromaticity detection algorithms for SMILES generation.

Also note that a ring structure can have multiple potential ring-closure points. For example, a six-membered ring has six bonds, each of which can be a ring-closure point. As a result, a ring compound may be represented by many different but equally valid SMILES strings. Actually, it is very common that there are a lot of SMILES strings that represent the same structure, whether it has a ring or not, because one can start with any atom in a molecule to derive a SMILES string. Therefore, it is necessary to select a "unique SMILES" for a molecule among many possibilities. Because this is done through a process called "canonicalization", this unique SMILES string is also called the "canonical SMILES".

Isomeric SMILES

Isomeric SMILES allows for specifying isotopism and stereochemistry of a molecule. Information on isotopism is indicated by the integral atomic mass preceding the atomic symbol. The atomic mass must be specified inside square brackets. For example, C-13 methane can be represented by "[13CH4]". Configuration around double bonds is specified by "directional bonds" (characters / and \). For example, E- and Z-1,2-difluoroethene can be represented by the following isomeric SMILES:

F/C=C/F or F\C=C\F (E)-1,2-difluoroethene (trans isomer)
F/C=C\F or F\C=C/F (Z)-1,2-difluoroethene (cis isomer)

Configuration around tetrahedral centers are indicated by the symbols "@" or "@@"

C[C@@H](C(=O)O)N L-Alanine
C[C@H](C(=O)O)N D-Alanine

More detailed information on chirality specification can be found in Chapter 3[26] of the Daylight Theory Manual.

Limitations of SMILES

SMILES is proprietary and it is not an open project. This has led different chemical software developers to use different SMILES generation algorithms, resulting in different SMILES versions for the same compound. Therefore, SMILES strings obtained from different databases or research groups are not interchangeable unless they used the same software to generate the SMILES strings. With an aim to address this interchangeability issue of SMILES, an open-source project has launched to develop an open, standard version of the SMILES language called OpenSMILES.[28] However, the most noticeable community effort in this area is development of InChI, which is described in next section.

3.2. International Chemical Identifier (InChI) and InChIKey

InChI

The IUPAC International Chemical Identifier (InChI)[22-25] was originally developed by the IUPAC and continuing development efforts have been made by the InChI Trust[25]. InChI is non-proprietary, open-source, and freely available to the scientific community. Especially, because the software for generating InChI strings is also freely available, it avoids the interoperability issue that different implementations of SMILES language have.

InChI encodes a chemical structure into "layers". Each layer holds a distinct and separable class of structural information, with the layers ordered to provide successive structural refinement. There are currently six InChI layer types, each different class of structural information: the main layer, a charge layer, a stereochemical layer, an isotopic layer, a fixed-H layer and a reconnected layer. The main layer, which specifies chemical formula, atoms, and bonds between them, is required for all InChIs. However, the other layers appear only when corresponding input information is provided. Layers and sublayers start with "/" (forward slash) followed by a letter denoting the identity of the layer (except for the chemical formula layer). Below are some examples of InChI.

InChI=1S/CH4/h1H4 (methane)

InChI=1S/C2H6/c1-2/h1-2H3 (ethane)

InChI=1S/C2H6O/c1-2-3/h3H,2H2,1H3 (ethanol)

InChI=1S/C3H7NO2/c1-2(4)3(5)6/h2H,4H2,1H3,(H,5,6)/t2-/m0/s1 (L-alanine)

These InChI strings are not easy for a human to understand (especially compared to SMILES strings). It is because InChI was developed as a "machine-readable" chemical identifier, with an aim to enable a computer to regenerate the corresponding chemical structure from the InChI string generated by another computer. For this reason, InChI is often called as the bar code for chemical structures.

Because the layered structure of InChI allows one to represent a chemical structure with a desired level of details, InChI software may generate different InChI strings for the same molecule. This flexibility may be regarded as an obstacle to standardization and interoperability. In response to this concern, the standard InChI was introduced which contains the same level of structural details and the same conventions for drawing perception, by using standard option settings in InChI software. The standard InChI representations begin with "InChI=1S/", while the non-standard InChI begins with "InChI=1/". The digit "1" following "InChI=" is the current InChI version number.

InChIKey

The length of an InChI string increases with the size of the corresponding chemical structure, and it is very common that molecules with more than 100 atoms result in very long InChI strings, which are not appropriate to use in internet search engines (such as Google, Yahoo, Bing, and so on). In addition, these search engines do not care about case sensitivity nor special characters used in InChI. To address this issue, the InChIKey was introduced for Internet and database searching/indexing. It is a 27-character string derived from InChI, using a hashing algorithm. Hashing is a one-way mathematical transformation typically used to calculate a compact fixed length digital representation of a much longer string of arbitrary length.

The InChIKey consists of three blocks, separated by hyphens, for example:

BSYNRYMUTXBXSQ-UHFFFAOYSA-N (aspirin)

HEFNNWSXXWATRW-UHFFFAOYSA-N (ibuprofen)

RZVAJINKPMORJF-UHFFFAOYSA-N (acetaminophen)

The first block of 14 characters (out of 27 characters in total) encodes core molecular constitution, described by the InChI main layer. The other structural features (such as stereochemistry, isotopic substitution, exact position of hydrogens, and metal ligation data) are encoded into the second block. The protonation or deprotonation state is encoded in the last InChIKey character.

Many databases such as PubChem[29], ChemSpider[30], ChEBI[31], and NIST Chemistry Webbook [32] accept InChI and InChIKey strings as queries to search for chemical structures. InChIs and InChIKeys can also be used as queries in UniChem[33] to produce cross-references between chemical structure identifiers from different databases.

4. Generic Structures

A generic structure indicates a group of structurally similar compounds, using a symbol such as "R" (as in R-CH_2-OH, where R = H, CH_3, CH_2CH_3, $CH(CH_3)_2$, $C(CH_3)_3$, and so on). Generic structures are commonly used in chemistry texts as well as in chemical patents in which the inventor claims a whole class of related compounds. Generic structures are more often called "Markush" structures after Dr. Eugene A. Markush, who involved in a legal case which set a precedent in the USA for generic chemical structure patent filing.

An early example of research projects on Markush structure storage and retrieval is the Sheffield Generic Structures Project, which led to a text-based language for generic structure description called GENSAL (GENeric Structure LAnguage)[34] as well as an extended connection table representation for generic structures[35]. The Sheffield generic structures system was never implemented commercially, but influenced two commercial systems: MARPAT[36] (developed by CAS) and Markush DARC (currently Thomson Reuters' Merged Markush Service[37]).

Some public databases, such as PubChem, allow one to search for generic structures, using SMARTS (SMiles ARbitrary Target Specification). It is a language used for describing molecular patterns. SMARTS is useful for substructure searching, which finds a particular pattern (subgraph) in a molecule. SMARTS are straightforward extensions of SMILES. All SMILES symbols and properties are legal in SMARTS. SMARTS includes logical operators and additional molecular descriptors. Detailed information on SMARTS is given in the SMARTS specification document[38] in the Daylight theory manual and SMARTS tutorial.[39]

Another extension of SMILES is SMIRKS[40,41], which is a line notation for generic reactions. A generic reaction represents a group of reactions that undergo the same set of atom and bond changes. Note that SMILES and SMARTS can be used to represent reactions, using the ">"

symbol between the reactants, products, and agents, as described in the SMILES and SMARTS specification documents. (Therefore, these SMILES and SMARTS that describe reactions are often called reaction SMILES and reaction SMARTS, respectively.) On the other hand, SMIRKS is used to represent *types* of reactions (e.g., S_N2 reaction). More detailed information on SMIRKS is given in the SMIRKS specification document[40] and SMIRKS tutorial[41].

References

(1) Systematic Name (from Wikipedia) (https://en.wikipedia.org/wiki/Systematic_name) (Accessed on 6/30/2015).
(2) Trivial Name (from Wikipedia) (https://en.wikipedia.org/wiki/Trivial_name) (Accessed on 6/30/2015).
(3) Preferred IUPAC name (from Wikipedia) (https://en.wikipedia.org/wiki/Preferred_IUPAC_name) (Accessed on 6/30/2015).
(4) Favre, H. A.; Powell, W. H. *Nomenclature of Organic Chemistry: IUPAC Recommendations and Preferred Names 2013*; Royal Society of Chemistry, 2013.
(5) Overview of Generic Drugs and Drug Naming (from Merck Manual - Consumer Version) (https://www.merckmanuals.com/home/drugs/brand-name-and-generic-drugs/overview-of-generic-drugs-and-drug-naming) (Accessed on 6/30/2015).
(6) International Nonproprietary Names (http://www.who.int/medicines/services/inn/en/) (Accessed on 6/30/2015).
(7) International Nonproprietary Name (from Wikipedia) (https://en.wikipedia.org/wiki/International_Nonproprietary_Name) (Accessed on 6/30/2015).
(8) Chemical Abstracts Service (CAS) (https://www.cas.org/) (Accessed on 6/30/2015).
(9) SciFinder (https://www.cas.org/products/scifinder) (Accessed on 6/30/2015).
(10) STN (https://www.cas.org/products/stn) (Accessed on 6/30/2015).
(11) CAS REGISTRY and CAS Registry Number FAQs (https://www.cas.org/content/chemical-substances/faqs) (Accessed on 6/30/2015).
(12) About CAS - FAQs: What is a CAS Registry Number? (https://www.cas.org/about-cas/faqs#casrn) (Accessed on 6/30/2015).
(13) Wiswesser, W. J. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 88.
(14) Ash, S.; Cline, M. A.; Homer, R. W.; Hurst, T.; Smith, G. B. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 71.
(15) Homer, R. W.; Swanson, J.; Jilek, R. J.; Hurst, T.; Clark, R. D. *J. Chem Inf. Model.* **2008**, *48*, 2294.
(16) Barnard, J. M.; Jochum, C. J.; Welford, S. M. *Acs Symposium Series* **1989**, *400*, 76.
(17) Rohbeck, H. G. In *Software Development in Chemistry 5*; Gmehling, J., Ed.; Springer Berlin Heidelberg: 1991, p 49.
(18) Weininger, D. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31.
(19) Weininger, D.; Weininger, A.; Weininger, J. L. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97.
(20) Weininger, D. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 237.
(21) SMILES: Simplified Molecular Input Line Entry System (http://www.daylight.com/smiles/) (Accessed on 6/30/2015).

(22) Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. *J. Cheminform.* **2013**, *5*, 7.
(23) Heller, S.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. *J. Cheminform.* **2015**, *7*, 23.
(24) The IUPAC International Chemical Identifier (InChI) (http://www.iupac.org/home/publications/e-resources/inchi.html) (Accessed on 6/29/2015).
(25) InChI Trust (http://www.inchi-trust.org/) (Accessed on 6/29/2015).
(26) Daylight Theory Manual, Chapter 3: SMILES - A Simplified Chemical Language (http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html) (Accessed on 6/23/2015).
(27) Daylight SMILES Tutorial (http://www.daylight.com/dayhtml_tutorials/languages/smiles/index.html) (Accessed on 6/23/2015).
(28) OpenSMILES Home Page (http://www.opensmiles.org/) (Accessed on 6/23/2015).
(29) PubChem (https://pubchem.ncbi.nlm.nih.gov) (Accessed on 6/29/2015).
(30) ChemSpider (http://www.chemspider.com) (Accessed on 6/29/2015).
(31) ChEBI (https://www.ebi.ac.uk/chebi/) (Accessed on 6/29/2015).
(32) NIST Chemistry Webbook (http://webbook.nist.gov/chemistry/) (Accessed on 6/29/2015).
(33) UniChem (https://www.ebi.ac.uk/unichem/) (Accessed on 6/29/2015).
(34) Barnard, J. M.; Lynch, M. F.; Welford, S. M. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 151.
(35) Barnard, J. M.; Lynch, M. F.; Welford, S. M. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 160.
(36) MARPAT (https://www.cas.org/content/markush) (Accessed on 6/30/2015).
(37) Merged Markush Service (http://ip-science.thomsonreuters.com/support/patents/dwpiref/reftools/classification/markush/) (Accessed on 6/30/2015).
(38) Daylight Theory Manual, Chapter 4: SMARTS - A Language for Describing Molecular Patterns (http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html) (Accessed on 6/23/2015).
(39) Daylight SMARTS Tutorial (http://www.daylight.com/dayhtml_tutorials/languages/smarts/index.html) (Accessed on 6/23/2015).
(40) Daylight Theory Manual, Chapter 5: SMIRKS - A Reaction Transform Language (http://www.daylight.com/dayhtml/doc/theory/theory.smirks.html) (Accessed on 10/8/2015).
(41) Daylight SMIRKS Tutorial (http://www.daylight.com/dayhtml_tutorials/languages/smirks/index.html) (Accessed on 10/8/2015).

Questions

1. Go to the PubChem database (http://pubchem.ncbi.nlm.nih.gov) and search for omeprazole and esomeprazole. Fill in the table below with appropriate chemical representations for the two molecules and answer the following questions.

(1) What is the structural difference between omeprazole and esomeprazole?

(2) Do omeprazole and esomeprazole have the same InChI and InChIKeys as each other?

(3) Do omeprazole and esomeprazole have the same canonical SMILES? Explain why.

Omeprazol (from PubChem)	
CID	
IUPAC name	
Canonical SMILES	
Isomeric SMILES	
InChI	
InChIKey	
Esomeprazole (from PubChem)	
CID	
IUPAC name	
Canonical SMILES	
Isomeric SMILES	
InChI	
InChIKey	

2. Go to ChemSpider (http://www.chemspider.com) and search for omeprazole and esomeprazole. Fill in the table below with appropriate chemical representations for the two molecules and answer the following questions.

(1) Are the systematic names from ChemSpider the same as those from PubChem?

(2) Are the canonical SMILES from ChemSpider the same as those from PubChem?

(3) Are the InChI and InChIKeys from ChemSpider the same as those from PubChem?

Omeprazol (from ChemSpider)	
ChemSpider ID	
IUPAC name	
Canonical SMILES	
Isomeric SMILES	
InChI	
InChIKey	
Esomeprazole (from ChemSpider)	
ChemSpider ID	
IUPAC name	
Canonical SMILES	
Isomeric SMILES	
InChI	
InChIKey	

3. Compare the SMILES strings from PubChem with those from ChemSpider for the following compounds, in terms of how the two databases deal with perceived aromaticity of the molecules. Explain an advantage and a disadvantage of the SMILES strings used in each database.

	SMILES from PubChem	SMILES from ChemSpider
Benzene		
pyridine		
Pyrrole		
Furan		
Thiophene		
Selenophene		
Tellurophene		

4. Suppose that you are a project manager at Google, who are in charge of implementing a chemical search algorithm to the Google search. This algorithm accepts a chemical structure as an input through the search box on the Google homepage (http://www.google.com), but the input needs to be a text string that represents a chemical structure. Therefore, you need to choose a line notation that is most appropriate for this search system, among the canonical SMILES, InChI, and InChIKey. Choose only one and justify your choice over the others, based on what you have learned from this module and from Questions 1, 2 and 3).