Chem. Listy *111*, 710−715 (2017) Referát

LINEÁRNÍ REPREZENTACE CHEMICKÝCH STRUKTUR

JIŘÍ JIRÁT[a,b] a DANIEL SVOZIL[a,b]

[a] CZ-OPENSCREEN: Národní infrastruktura pro chemickou biologii, Ústav molekulární genetiky AV ČR v.v.i. Vídeňská 1083, 142 20 Praha 4, [b] CZ-OPENSCREEN: Národní infrastruktura pro chemickou biologii, Laboratoř informatiky a chemie, Fakulta chemické technologie, Vysoká škola chemicko-technologická v Praze, Technická 5, 166 28 Praha 6
Jiri.Jirat@vscht.cz, Daniel.Svozil@vscht.cz

Došlo 24.7.17, přijato 5.10.17.

Klíčová slova: lineární reprezentace, SMILES, InChI, InChIKey, SMARTS

Obsah

1. Úvod – co jsou linearizované zápisy chemických struktur?
2. SMILES
3. InChI (a InChIKey)
4. InChIKey
5. SMARTS
6. Závěr

1. Úvod – co jsou linearizované zápisy chemických struktur?

Ke každodenní rutině chemika patří práce s chemickými strukturami uloženými v elektronické podobě – hledá ve strukturních nebo reakčních databázích[1], kreslí strukturní vzorce do publikace/dokumentace nebo se jen dívá na zobrazení chemických struktur na obrazovce, byť by to bylo např. jen v katalogu dodavatele chemikálií.

Málokdo se zabývá tím, jak jsou vlastně strukturní data ukládána a přenášena („to je přece věc vývojářů databází a programátorů, proč by mne to mělo zajímat“). Ale každý chemik by rozhodně měl vědět, že prakticky vždy je pro tyto účely využívána aplikace teorie grafů – každá molekula či i jen fragment jsou zaznamenány jako graf s množinou vrcholů/uzlů (atomy) a hran, které je spojují (vazby), spolu se všemi jejich vlastnostmi. Nejtypičtější způsob zápisu, který je používán, jsou tzv. spojovací tabulky, umožňující prakticky neomezené zaznamenání molekuly. Ty mají však také jednu nevýhodu, a tou je jejich relativní velikost (každý atom bývá zapsán na jedné řádce a taktéž každá vazba na jedné řádce), která při současných počtech známých (tj. publikovaných) struktur – šplhajících se někam ke 100 miliónům, začíná hrát významnou roli. Nemluvě o počítačově generovaných „chemických prostorech“[2], kde jsou počty molekul ještě řádově větší.

Již dlouhou dobu tak byl řešen problém, jak úsporně ukládat údaje o topologii molekuly, kdy na souřadnicích atomů nezáleží (nebo je ani neznáme) a jediné, co nás zajímá, je spojení jednotlivých atomů. Minimalizace množství zaznamenávaných informací hrála vždy velkou roli – jako obvykle vývoj techniky nikdy nestačil požadavkům uživatelů a přes ohromný rozvoj výpočetní, úložní a konektivitní kapacity jsou úlohy na hraně možností. Čím výkonnější prostředky, tím náročnější a rozsáhlejší úlohy jsou řešeny – viz např. virtuální screening[3], hledání vztahu mezi strukturou a biologickou aktivitou[4], odhadování syntetické dostupnosti[5], správa knihoven chemických látek[6] nebo různé systémy pro zaznamenávání dat z testování s vysokou propustností[7]. Takže stále vyvstávala otázka, jak molekuly nejen efektivně ukládat do databáze, ale jak je také přenášet mezi aplikacemi.

Jedním z elegantních řešení jsou tzv. linearizované zápisy. Jejich princip se dá obvykle shrnout do několika bodů:

- z cyklického grafu je potřeba udělat acyklický (zaznamenáme pouze tzv. *kostru* grafu), protože acyklickou strukturu (byť větvenou), lze snadno zapsat do řádky textu,
- nějakým způsobem zaznamenat cykly/chybějící cyklické vazby,
- vytvořit kódování pro atomy/funkční skupiny a vazby.

Během dekád byly vyvinuty různé linearizované zápisy[8], v tomto článku však představíme jen ty nejpoužívanější formáty a standardy.

2. SMILES

Název formátu SMILES je zkratkou pro Simplified Molecular Input Line Entry Specification[9]. Formát samotný byl navržen pro použití lidmi, zápis sloučeniny se podobá „normálnímu“ zápisu chemických struktur. Byl vytvořen v 80. letech 20. století, přesto nejeví žádné známky zastarávání a je stále velmi populární a využívaný. Nově je dostupná i aktualizovaná formální specifikace, dostupná pod svobodnou licencí[10]. Umožňuje – ale nevyžaduje – tzv. kanonickou formu[9].

Graf sloučeniny je zapsán jako *strom* (acyklický graf), resp. *les* (více *stromů* – pokud se látka skládá z více nesouvislých fragmentů). Chybějící *tětivy* (vazby, které jsme odstranili ze zápisu, abychom získali acyklický graf) jsou zaznamenány spojovacími čísly u příslušných atomů.

Symboly atomů se píší do hranatých závorek spolu s počtem vodíků, specifikací náboje a isotopu: [CH4],

[CH3]-[CH3], [C], [Pb], [Zn++], [Zn2+], [14CH3-], [2H+]

Prvky běžné v organické chemii patří do tzv. „organic subset“ – „organické podmnožiny“. U těch není třeba používat hranaté závorky a uvádět počet vodíků explicitně – mají implicitní vaznost. Patří sem prvky B, C, N, O, P, S, F, Cl, Br a I.

Vazby jsou reprezentovány následovně: jednoduchá „-“, nebo se neuvádí; dvojná „=“; trojná „#“; čtverná „$“; aromatická „:“; nulová „.“.

Jestliže není vazba uvedena, předpokládá se jednoduchá vazba: „CCC“ je to samé co „C-C-C“.

Aromatická vazba mezi atomy z „organic subset“ je implikovaná, pokud je symbol zapsán malými písmeny: „C:1:C:C:C:C:C1“ je to samé co „c1ccccc1“, oba zápisy reprezentují benzen.

Nulová vazba se používá např. pro vyjádření solí, komplexů, apod.: [Na+].[Cl-] místo [Na]Cl, CC(=O)[O-].[NH4+], apod. Zápis bez tečky by implikoval jednoduchou kovalentní vazbu mezi atomy, což by byl z chemického hlediska nesmysl.

Větvení se zapisuje pomocí závorek: CC(C)(C)CCl je 1-chlor-2,2-dimethylpropan. Větvení lze rekurzivně vkládat do sebe, ve větvích lze používat čísla pro vytváření kruhů: CC(C(C(C)C)C)CC, viz obr. 1a-b.

Kruhy se vytvářejí pomocí čísel za symbolem atomu. Odpovídající čísla se spojí vazbou. Čísla, která se jednou spojí, lze recyklovat: „C1CC1C2CC2“ = „C1CC1C1CC1“. Čísla jsou pouze jednociferná „C12“ neznamená jeden kruh s číslem 12, ale dva s čísly 1 a 2, viz obr. 1c.

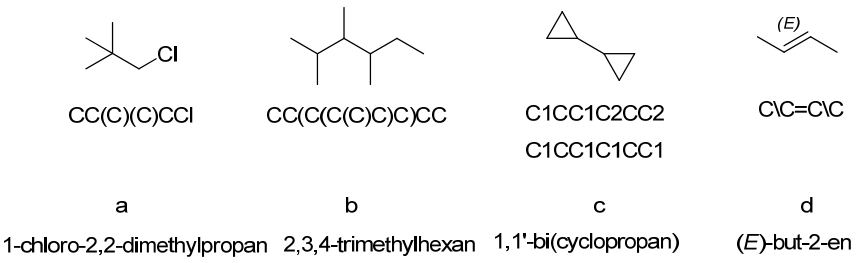
Pro vyjádření isomerie na dvojných vazbách se používají odlišné symboly pro jednoduché vazby - „\“ a „/“.„\“ značí směr dolů, „/“ směr nahoru. C\C=C\C je zápis pro *trans*-but-2-en, neboli (*E*)-but-2-en, viz obr. 1d.

Stereochemie na sp^3 uhlíku je zaznamenávána v podobném duchu, jako je vytvářen celý zápis, tj. „kráčíme“ molekulou (grafem) a na chirálním centru zaznamenáme pořadí a smysl rotace substituentů tak, jak je „vidíme“: pořadí substituentů je určeno zápisem, smysl rotace substituentů symboly @ (proti směru hodinových ručiček), nebo @@ (rotace po směru hodinových ručiček). Příklad viz obr. 2.

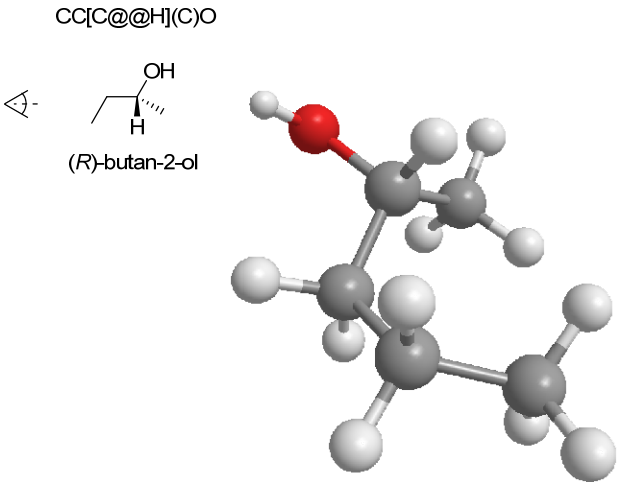
SMILES je skvělý a intuitivní pro zápis topologie molekuly do textového řetězce – snadno můžeme přenášet strukturní data tak, že na každé řádce textového souboru je právě jedna „struktura“, což je z hlediska počítačového zpracování velmi elegantní záležitost. Nevýhodou je však fakt, že pro danou strukturu můžeme vytvořit více správných zápisů SMILES. Např. i tak jednoduchou molekulu, jako je kyanovodík, můžeme zapsat těmito způsoby: C#N, N#C, [CH]#N, N#[CH]. To zásadním způsobem znemožňuje přímočaré použití zápisu SMILES jako identifikátoru a např. porovnáním dvou zápisů SMILES jednoduše říci, zda jsou dvě struktury shodné, nebo ne. Proto byl vyvinut tzv. kanonický SMILES[9]. Vytvoření kanonického zápisu spočívá v tzv. kanonickém očíslování atomů v molekule, kdy atomy v dané molekule jsou očíslovány vždy stejným způsobem, a následném vytvoření kanonického zápisu. Tento postup zajišťuje, že pro danou molekulu je vytvořen právě jen jeden, vždy stejný, zápis. V případě kyanovodíku je to varianta C#N. Kanonické zápisy pak už lze použít pro porovnávání, protože máme jistotu, že ať už byla struktura nakreslena/vypočítána/vygenerována jakkoli, bude ve výsledku reprezentována vždy tím samým zápisem SMILES. Přes dostupnost této možnosti se však SMILES nikdy neujal jako identifikátor (použitelný např. pro indexaci v databázi), zůstal na pozici skvělého formátu pro výměnu strukturních dat, který je srozumitelný i člověku. Pozici identifikátoru založeného na topologii molekuly zaujal identifikátor popsaný v následující části.

3. InChI (a InChIKey)

Jedná se o relativně velice mladý formát (vznik kolem r. 2005), je produktem spolupráce organizací IUPAC a NIST (IUPAC International Chemical Identifier)[11–13]. Jednou z hlavních ideí bylo vytvoření neutrálního standardního identifikátoru chemických struktur založeného na struktuře sloučeniny – např. CAS RN je vázané na přidělení službou Chemical Abstracts Service, a není dostupné pro látky, které jsou např. jen hypoteticky předpovězené při generování chemického prostoru[2]. Standardy InChI se stále intenzivně vyvíjejí a jsou přidávána nová rozšíření a vlastnosti, např. standard RInChI pro zaznamenávání chemických reakcí[14].



Obr. 1. **Příklady zápisů SMILES: (a) větvení zapsané pomocí závorek; (b) možnost větvení rekurzivně vnořovat; (c) zápis cyklů pomocí čísel u atomů a možnost čísla recyklovat; (d) zápis stereochemie na sp^2 atomu**



Obr. 2. **Způsob zápisu stereochemie na sp^3 atomu ve formátu SMILES.** Zápis CC[C@@H](C)O je pro (*R*)-butan-2-ol. Pořadí substituentů je Et, H, C, O, smysl rotace po směru hodinových ručiček (dva symboly @), pohled od Et k centru. Zápis (jeden z možných) opačné konfigurace je CC[C@@H](O)C

Hlavní vlastnosti bychom mohli shrnout do následujících bodů:

- je to kanonický formát,
- s vysokým stupněm normalizace,
- jeho vrstevnatá struktura umožňuje porovnávat struktury na různých úrovních detailu.

V r. 2009 vzniklo tzv. Standardní InChI, které neumožňuje nijak nastavovat parametry výstupu, jako to uměla předchozí verze. Identifikátor InChI, který je vytvořen podle algoritmu pro Standardní InChI, má na začátku řetězec „1S/“ (cit.[15]).

Na rozdíl od SMILES bylo předpokládáno jen jednosměrné zakódování struktury do identifikátoru, tj. vůbec nebylo pomýšleno na to, že by člověk nebo stroj převáděl InChI zpátky na strukturu – přestože to lze dělat.

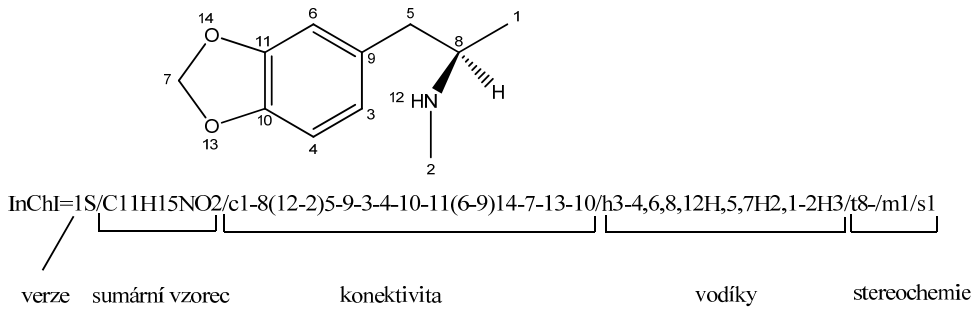
Celé InChI se skládá z jednotlivých vrstev, tzv. „layers“. Vrstvy jsou odděleny lomítkem, viz obr. 3. Povinné vrstvy jsou vrstva sumárního vzorce, konektivity a vodíková vrstva. Čísla atomů jsou přiřazena v pořadí sumárního vzorce. Na obr. 3 je přiřazení následovné: 1-11 = C, 12 = N, 13,14 = O (vodíky jsou přeskočeny).

Tautomery mají stejné InChI, protože InChI ve vrstvě konektivity neukládá řády vazeb a tautomery se tak liší pouze umístěním atomů vodíku, které je zaznamenané ve vodíkově vrstvě (je třeba si stále připomínat, že InChI není zápis struktury, ale jednoznačný identifikátor založený na struktuře, proto zápis řádu vazeb není podstatný, jelikož tato informace je již implicitně uložena ve vodíkové vrstvě). Vrstva popisující vodíky obsahuje u tautomerů informaci typu „2 atomy vodíku rozprostřené po 3 atomech kyslíku“ (což je daleko blíže realitě). Standardní InChI implicitně obsahuje tzv. FixedH („Fixed hydrogens“, tedy „zafixované vodíky“), což způsobí přidání další vrstvy, která popisuje jedno konkrétní umístění vodíků, viz obr. 4.

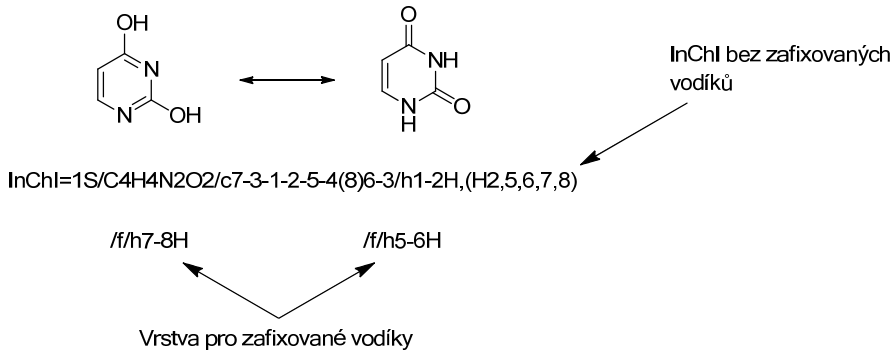
4. InChIKey

Identifikátor InChI má z hlediska tvorby databází jednu velkou nevýhodu – má proměnlivou délku, může být velmi dlouhý (pro molekuly se stovkami atomů může snadno dosáhnout délky stovek znaků) a obsahuje relativně pestrou směs znaků. Typickým požadavkem na položku typu ID v databázích je její rozumně omezená délka a ideálně jednoduchá sada znaků (číslice, písmena z ASCII sady, jejich kombinace, event. znak typu pomlčka, podtržítko aj.), což InChI se svou shora teoreticky neohraničenou délkou nesplňuje.

Proto byl pro potřeby rychlé indexace navržen tzv. InChIKey, což je „hash“ konstantní délky založený na InChI. Tzv. hašovací funkce je funkce nebo algoritmus, který převádí dlouhá data (např. textový řetězec) na relativně krátký řetězec nazývaný „hash“. Typické vlastnosti hašování jsou: různě dlouhá data jsou kódována do stejně dlouhého řetězce; malá změna ve vstupních datech vyvolá velkou změnu ve výstupu; je prakticky nemožné převést „hash“ zpět na vstupní data (jednosměrné zobrazení); pro praktické aplikace lze předpokládat, že různá data budou mít různé „hashe“ (ale tzv. kolize – různá data mohou mít



Obr. 3. **Struktura identifikátoru InChI, hlavní vrstvy identifikátoru**



Obr. 4. **Zápis tautomerů v identifikátoru InChI. InChI bez zafixovaných vodíků popisuje oba tautomery.** Přidáním vrstvy pro zafixované vodíky explicitně zvolíme jednu tautomerní formu – buď jsou vodíky umístěny na atomech 7 a 8, nebo 5 a 6

stejný „hash" – není teoreticky vyloučena). Princip hašování je široce využíván ve všech oblastech vývoje softwaru a ukládání dat, nejznámějšími příklady jsou ukládání hesel v šifrované podobě nebo přenos šifrované komunikace.

Algoritmus pro tvorbu InChIKey používá standard SHA-2 256 (publikovaný NIST), a to k hašování hlavních a vedlejších řetězců[15].

InChIKey se skládá ze tří částí:

- 14 znaků pro strukturu (tzv. „major"),
- 10 znaků pro vrstvy stereo, fixedH a další (tzv. „minor"),
- 1 znak pro protonaci/deprotonaci (N = 0, M = -1, O = +1).

Příklad – sloučenina (1*E*)-1-fluoro-1-iodoprop-1-en bude mít InChI=1S/C3H4FI/c1-2-3(4)5/h2H,1H3/b3-2+ , její InChIKey bude: DKEPSVLMHRNWRG-NSCUHMNNSA-N.

Jednotlivé části InChIKey pak budou vytvořeny takto:

- Major Block (C3H4FI/c1-2-3(4)5/h2H,1H3 → DKEPSVLMHRNWRG)
- Separator (-)
- Minor Block – stereochemie (/b3-2+ → NSCUHMNN)
- Standard Flag (v příkladu: S)
- Version (natvrdo: A)
- Separator (-)
- Protonation Flag (N - struktura v příkladu není protonována)

Hašování má však jednu nevýhodu/omezení – může dojít k tzv. *kolizi*, tj. dva rozdílné původní textové řetězce/data mohou po hašování mít shodný „hash". Při výpočtu tzv. teoretické kolizní odolnosti byla vypočtena 50% šance pro single kolizi v 1. bloku na $6{,}1 \cdot 10^9$, tj. pro dataset 6,1 miliardy molekul je očekávaný počet kolizí ½. Obdobný odhad pro 2. blok je $3{,}7 \cdot 10^5$. Celkově je odhad kolizní odolnosti celého InChIKey $6{,}1 \cdot 10^9$ molekulárních skeletů krát $3{,}7 \cdot 10^5$ stereo/proton/isotop isomerů $\sim 2 \cdot 10^{15}$ (cit.[16]).

Ještě před spuštěním v r. 2007 byla kolizní odolnost testována na sadě reálných (databáze ZINC, PubChem) i generovaných struktur (databáze GDB, FP42), celkem $\sim 77 \cdot 10^6$ záznamů (počet struktur po sloučení všech sad a následné deduplikaci) a žádná kolize nebyla nalezena[17]. Od té doby však již byly nalezeny (u generovaných struktur) kolize, a to jak ve vedlejším bloku (stereochemie)[18], tak už i v hlavním bloku (konektivita)[19]. Přes to si však InChIKey již úspěšně našlo cestu do většiny velkých databází (např. Reaxys, PubChem).

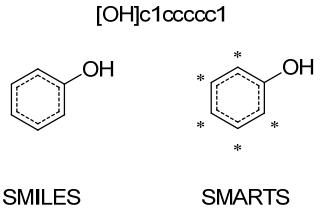
5. SMARTS

Zápisy SMILES nebo identifikátor InChI (resp. InChIKey) slouží velmi dobře pro kódování/přenos konkrétních struktur, tj. jeden zápis SMILES odpovídá právě jedné struktuře. Co ale v případech, kdy potřebujeme zaznamenat obecnou skupinu sloučenin, nebo jen substrukturu nebo funkční skupinu?

Pro tyto potřeby byl vyvinut linearizovaný zápis pro popis substruktur – SMARTS. Jedná se o elegantní a přímočaré rozšíření zápisu SMILES – „SMiles ARbitrary Target Specification" [20]. Obdobný zápis založený na InChI neexistuje, je to dáno principiálním rozdílem mezi SMILES a InChI – InChI bylo navrženo jako identifikátor pro konkrétní strukturu, nikoli pro nekompletní struktury nebo struktury obsahující generické skupiny.

Jako SMARTS výraz můžeme použít libovolný zápis SMILES, neboť platí, že kterýkoli zápis SMILES je zároveň výrazem SMARTS (ale neplatí to naopak!). Kromě toho ovšem můžeme použít různé rozšiřující symboly, obvykle nazývané „wildcards", a to pro atomy i vazby; dále jsou k dispozici dokonce i logické operátory nebo vnořovaná prostředí.

Ilustrační příklad: SMILES řetězec [OH]c1ccccc1 reprezentuje fenol. Pokud ovšem tento řetězec interpretujeme jako výraz SMARTS, pak nám (je-li použit např. jako dotaz) najde všechny struktury obsahující fenol, a to tak, že mohou být substituovány kterékoli z uhlíkových atomů, které nejsou v poloze 1 (kyslík má explicitně uveden vo-



Obr. 5. **Ukázka rozdílné interpretace téhož lineárního zápisu: je-li považován za zápis SMILES (vlevo), identifikuje právě jednu molekulu (fenol); je-li považován za zápis SMARTS (vpravo), pak umožňuje jakékoli substituce na všech atomech do jejich implicitní valence**

dík; uhlík v poloze 1 už nemá volnou valenci, takže zde žádné substituce nemohou být), viz obr. 5.

Navíc můžeme použít řadu dalších generických symbolů, např. „*“ znamená „jakýkoli atom“, „a“ je „aromatický atom“, „A“ alifatický, H<n> (za <n> dosazujeme číslo) „<n> připojených vodíků“ a mnoho dalších. Podobně symboly vazeb jsou oproti SMILES rozšířeny o symboly „~“ (jakákoli vazba), „/?“ (stereochemická vazba "nahoru nebo nespecifikovaná"), „\?“ (stereochemická vazba "dolu nebo nespecifikovaná"), @ (jakákoli vazba v cyklu) a další[20].

V zápisech SMARTS můžeme též použít logické operátory pro kombinování základních vlastností.

Příklady:

Mějme molekulu bifenylu, která v zápisu SMILES bude zapsána např. takto: c1ccccc1c1ccccc1. Použijeme-li následující vzory SMARTS (C, cc, c:c, c-c), budou jim odpovídat vždy pouze určité části molekuly – viz tab. I.

Význam SMARTS výrazu „[n;H1]“ je „aromatický dusík a zároveň musí mít právě jeden vodík“ (např. jako v 1*H*-pyrrolu).

Použití formátu SMARTS je tak již patrně zřejmé – můžeme velmi jednoduchým a srozumitelným způsobem definovat různé strukturní motivy a použít je pro screening a hledání potenciálních biologicky aktivních látek s žádaným účinkem, jako např. v práci Steinmetze a spol. [21], kde hledali agonisty pro kyselinu retinovou (RAR). A to vše ve formátu textového řetězce, který lze jednoduše interpretovat a přenášet mezi různými systémy. Pro vizualizaci a porozumění zápisům ve formátu SMARTS je velmi užitečný nástroj SMARTSviewer (http://smartsview.zbh.uni-hamburg.de/)[22] .

6. Závěr

Linearizované zápisy nebo identifikátory odvozené ze struktury hrají velmi významnou úlohu v procesu ukládání a přenášení chemických struktur, zejména v oblastech, kde je zpracováváno několik desítek či více miliónu záznamů a kde jejich velká část zatím nebyla experimentálně nikdy potvrzena (jako např. generovaných hypotetických struktur).

Pro snadný a úsporný přenos struktur (v případě, že nás zajímá pouze topologie molekuly, nikoli její geometrie) je vynikající formát SMILES, pro jednoznačnou identifikaci identifikátor InChI, event. z něj odvozený InChI-Key. Formát SMARTS nám pak umožňuje definovat obecnější (sub)struktury, které mohou být použity pro selekci vybraných molekul.

Tento článek vznikl za podpory MŠMT v rámci Národního programu udržitelnosti I projekt LO1220 (CZ-OPENSCREEN).

LITERATURA

1. Jindřich J.: Chem. Listy *111*, 731 (2017).
2. Čmelo I., Svozil D.: Chem. Listy *111*, 724 (2017).
3. Svozil D.: Chem. Listy *111*, 738 (2017).
4. Škuta C., Svozil D.: Chem. Listy *111*, 747 (2017).
5. Voršilák M., Svozil D.: Chem. Listy *111*, 760 (2017).
6. Popr M., Sedlák D., Bartůněk P.: Chem. Listy *111*, 772 (2017).
7. Muller T., Sedlák D., Bartůněk P.: Chem. Listy *111*, 766 (2017).
8. Engel T., v knize: *Chemoinformatics* (Gasteiger, J.;

Tabulka I
Ukázka použití výrazů SMARTS pro molekulu bifenylu

Výraz SMARTS	Význam	Odpovídající části v molekule bifenylu
C	jakýkoli alifatický uhlík	žádná (všechny uhlíky jsou aromatické)
cc	jakýkoli pár spojených aromatických uhlíků	všechny dvojice spojených uhlíkových atomů (celkem 13)
c:c	aromatické uhlíky spojené aromatickou vazbou	všechny dvojice spojených uhlíkových atomů **uvnitř** cyklů (celkem 12)
c-c	aromatické uhlíky spojené jednoduchou vazbou	pouze dvojice uhlíkových atomů, mezi kterými je vazba spojující dvě benzenová jádra (není aromatická)

Engel, T., ed.), Wiley-VCH, Weinheim 2004.
9. Weininger D.: J. Chem. Inf. Comput. Sci. *28*, 31 (1988).
10. *OpenSMILES specification*. http://opensmiles.org/opensmiles.html, staženo 3. 10. 2017.
11. Heller S. R., McNaught A., Pletnev I., Stein S., Tchekhovskoi D.: J. Cheminform. *7*, 23 (2015).
12. *The IUPAC International Chemical Identifier (InChI)*. https://iupac.org/who-we-are/divisions/division-details/inchi/, staženo 2. 10. 2017.
13. *About the InChI Standard*. http://www.inchi-trust.org/about-the-inchi-standard/, staženo 3. 10. 2017.
14. *The RInChI Project*. http://www-rinchi.ch.cam.ac.uk/, staženo 3. 10. 2017.
15. *IUPAC International Chemical Identifier (InChI) Programs InChI version 1, software version 1.05 (January 2017)*. http://www.inchi-trust.org/downloads/, staženo 31. 5. 2017.
16. Pletnev I., Erin A., McNaught A., Blinov K., Tchekhovskoi D., Heller S.: J. Cheminform. *4*, 39 (2012).
17. *InChITrust Technical FAQ* http://www.inchi-trust.org/technical-faq/, staženo 31. 5. 2017.
18. *InChIKey Collision.* http://www-jmg.ch.cam.ac.uk/data/inchi/, staženo 30. 5. 2017.
19. *An InChIkey Collision is Discovered and NOT Based on Stereochemistry.* http://www.chemconnector.com/2011/09/01/an-inchikey-collision-is-discovered-and-not-based-on-stereochemistry/, staženo 30. 5. 2017.
20. *SMARTS - A Language for Describing Molecular Patterns*. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html, staženo 31. 5. 2017.
21. Steinmetz F. P., Mellor C. L., Meinl T., Cronin M. T. D.: Molecular Informatics *34*, 171 (2015).
22. Schomburg K., Ehrlich H.-C., Stierand K., Rarey M.: J. Chem. Inf. Model. *50*, 1529 (2010).

J. Jirát and D. Svozil *(Laboratory of Informatics and Chemistry, University of Chemistry and Technology, Prague):* **Linear Representation of Chemical Structures**

An overview and description of the most used linear structure representations and identifiers based on 2D representation of molecule structures (SMILES, InChI, InChIKey) is given. For each type the following is described: algorithm for their generation, notation, basic principles of stereochemistry encoding, suitability for data transfer of chemical structures and for use as unique identifier. Format SMARTS (description of substructures and generic atoms/groups/bonds) is also briefly introduced.