POPIS A URČOVÁNÍ PODOBNOSTI MOLEKUL S POMOCÍ MOLEKULÁRNÍCH DESKRIPTORŮ

JIŘÍ NOVOTNÝ[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] *Laboratoř informatiky a chemie, Fakulta chemické technologie, Vysoká škola chemicko-technologická v Praze, Technická 5, 166 28 Praha 6*
novotnym@vscht.cz, Daniel.Svozil@vscht.cz

Došlo 4.8.17, přijato 4.10.17.

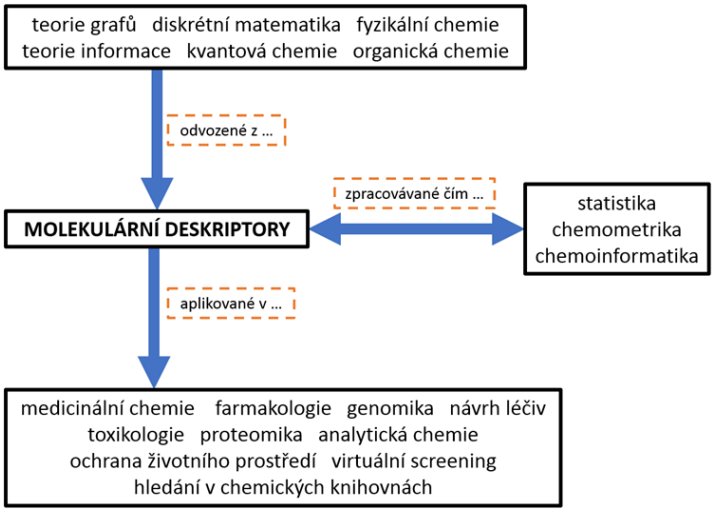
Klíčová slova: molekulární deskriptory, fingerprinty, podobnost molekul

Obsah

1. Úvod

Pokud chceme molekuly studovat matematickými a statistickými metodami, je nezbytně nutné numericky charakterizovat jejich vlastnosti, k čemuž slouží právě molekulární deskriptory (dále jen deskriptory). Formálně jsou deskriptory popsány takto[1]: „*Molekulární deskriptor je konečným výsledkem logické a matematické procedury, která transformuje chemickou informaci zakódovanou v symbolické reprezentaci molekuly do užitečného čísla či výsledku nějakého standardizovaného experimentu.*“ Zjednodušeně řečeno, deskriptory se snaží extrahovat a sumarizovat informace zakódované ve struktuře molekul do podoby, která je matematicky uchopitelná, tj. typicky do čísla, vektoru či matice. Příkladem deskriptoru je např. molekulová hmotnost či počet těžkých (tj. nevodíkových) atomů. Tím, že molekuly numericky popíšeme, jsme poté schopni na ně aplikovat matematické a statistické metody a mimo jiné i kvantitativně vyjádřit podobnost mezi nimi. Platí zde dva principy: za prvé, fyzikálně-chemické a biologické vlastnosti molekuly souvisí s její strukturou. Za druhé, strukturně podobné molekuly budou mít podobné fyzikálně-chemické vlastnosti.

Deskriptory hrají fundamentální roli v chemoinformatice a jsou odvozeny např. pomocí teorie grafů, teorie informace či fyzikální, kvantové a organické chemie[1] (obr. 1). Následně jsou deskriptory využity dvěma skupinami metod: QSAR (Quantitative Structure-Activity Relationship)[2] zjišťující kvantitativní vztahy struktura-aktivita a QSPR (Quantitative Structure-Property Relationship) zaměřené na kvantitativní vztahy struktura-vlastnosti. Tyto



Obr. 1. **Schéma toho, odkud jsou odvozeny molekulární deskriptory, čím jsou zpracovávány a co je oblastí jejich aplikace**[1]

metody jsou široce využívány v mnoha oblastech – můžeme zmínit medicinální chemii, počítačový návrh léčiv (zde konkrétně virtuální screening[3] a prohledávání chemického prostoru[4]), toxikologii, analytickou chemii či environmentální studie.

V současné době jsou definovány řádově tisíce deskriptorů. V jednom z největších sborníků[1] jich je popsáno přes 3300 a většinu z nich je možné spočítat v nějakém specializovaném softwaru. Každý z těchto deskriptorů přináší kousek informace o molekule, resp. její struktuře. Pokud se vhodně zvolí jejich množina, je pak možné s vysokou přesností predikovat vlastnosti látek, ať už se jedná o fyzikálně-chemické vlastnosti či biologickou aktivitu.

Deskriptory se dají rozdělit do dvou hlavních skupin[1]. První z nich jsou založeny na experimentálním měření (např. *logP*, molární refraktivita apod.) a obecně se dají označit jako fyzikálně-chemické deskriptory. Druhá skupina jsou tzv. teoretické deskriptory, které jsou odvozeny ze symbolické reprezentace molekuly a dále se rozdělují podle toho, kolika dimenzionální je daná reprezentace. Zde se pracuje s tzv. 0D, 1D, 2D, 3D a 4D deskriptory a tyto jednotlivé typy budou představeny v samostatných kapitolách. Některé typy deskriptorů se však mohou v rámci tohoto dimenzionálního rozřazení překrývat nebo nemají jasně definovanou třídu, tudíž je toto rozřazení nutné brát spíše orientačně. Obecně však platí, že čím vyšší „D“ deskriptor má, tím je jeho výpočet náročnější, ale zároveň obsahuje více chemických informací. Některé teoretické deskriptory jsou odvozeny z fyzikálně-chemických teorií a mají určité přirozené překrytí s experimentálními metodami. Fundamentální rozdíl mezi experimentálními a teoretickými deskriptory je ten, že ty teoretické, na rozdíl od experimentálních, neobsahují statistickou chybu způsobenou šumem při experimentálním měření. Na první pohled je zřejmé, že teoretické deskriptory jsou výhodnější co se týče ceny, času a dostupnosti, na druhou stranu však experimentální měření mohou poskytnout údaje obtížně dostupné pro teoretické výpočty.

2. Vlastnosti deskriptorů

Návrh nových deskriptorů, které jsou schopny zachytit nové aspekty molekulární struktury, je vlastně nikdy nekončící proces. V tomto druhu výzkumu jsou kromě kreativity a představivosti důležité i pevné teoretické základy. Nicméně deskriptory nelze vymýšlet jen tak bez omezení – ve všech případech musí splňovat následující čtyři podmínky[5], z nichž je především důležitá podmínka invariance, což obecně znamená, že algoritmus jejich výpočtu nezávisí na konkrétních vlastnostech molekulární reprezentace:

a) Invariance vůči značení a číslování atomů molekuly. Deskriptory využívající číslování atomů proto musí používat vlastní kanonické číslování.
b) Invariance vůči rotaci a translaci ve zvolené vztažné soustavě. Deskriptor např. nesmí nabývat různých hodnot v závislosti na poloze molekuly vzhledem k určité pevné referenční ose. Tyto invariance jsou především nutné pro 3D deskriptory.
c) Jednoznačná, algoritmicky spočitatelná definice čistě vycházející z molekulární struktury.
d) Hodnoty deskriptoru musí ležet ve vhodném číselném intervalu. Např. deskriptory, které ve své definici obsahují součin nějaké vlastnosti atomu, mohou u velkých molekul velmi rychle nabývat vysokých hodnot, což musí být patřičně ošetřeno.

Další z vlastností deskriptorů je i konformační invariance, která se dělí do čtyř tříd podle stupně závislosti deskriptoru na konformaci molekuly[1]:

a) Bez konformační závislosti (NCD deskriptory). Všechny deskriptory, které neberou v potaz 3D geometrii molekuly. Příklad: molekulární hmotnost.
b) Nízká konformační závislost (LCD deskriptory). Tyto deskriptory mají malou varianci pouze u relevantních konformačních změn. Příklad: *cis*/*trans* a nábojové deskriptory.
c) Střední konformační závislost (ICD deskriptory). Deskriptory vykazující malou varianci u všech konformačních změn. Příklad: deskriptory založené na hmotném středu molekuly.
d) Vysoká konformační závislost (HCD deskriptory). U těchto deskriptorů se silně projevuje jakákoliv změna konformace. Příklad: deskriptory popisující interakční energii.

Deskriptory mohou dále vykazovat tzv. degeneraci[1]. Degenerovaný deskriptor nabývá pro různé molekuly stejných nebo podobných hodnot. Degenerace má čtyři úrovně: žádná (N), nízká (L), střední (I) a vysoká (H).

3. Stručný přehled deskriptorů

V této kapitole si popíšeme deskriptory podle jejich dimenzionality. I v rámci jedné dimenze je možno deskriptory dělit na podtřídy, které jsou uvedeny v tab. I. Jelikož deskriptorů existuje opravdu hodně a pochopitelně je tu nelze všechny popsat, lze se s mnoha z nich seznámit v knize *Molecular Descriptors for Chemoinformatics*[1], která je v současné době zřejmě největším sborníkem deskriptorů. U většiny deskriptorů popsaných v tomto článku je odkaz na související literaturu, nicméně všechny zde popsané deskriptory lze také nalézt ve výše uvedené knize.

3.1. 0D deskriptory

K výpočtu těchto deskriptorů je použit sumární vzorec molekuly. Jelikož tato molekulární reprezentace neobsahuje žádné informace o topologii a geometrii molekuly, tak se z ní dají získat pouze deskriptory vycházející z počtu atomů, resp. sčítající příspěvky od každého atomu. Tyto deskriptory nejčastěji spadají do skupiny tzv. konstitučních deskriptorů. Příklady: molekulární hmotnost, počet dusíkových atomů, van der Waalsův poloměr, atomová

Tabulka I
Základní typy deskriptorů a jejich „dimenzionalita“ a vlastnosti. N-, L-, I-, H-CD u invariantních vlastností, resp. N, L, I, H u degenerace, značí žádnou, nízkou, střední a vysokou konformační invarianci, resp. degeneraci[1]

Deskriptor	Molekulární reprezentace	Matematická reprezentace	Konformační invariance	Degenerace
Molekulární hmotnost	0D	skalár	NCD	H
Počty atomů	0D	skalár	NCD	H
Počty fragmentů	1D	skalár	NCD	H
Topologické indexy	2D	skalár	NCD	L/I
Molekulární profily	2D	vektor	NCD	N
2D autokorelační deskriptory	2D	vektor	NCD	N/L
3D autokorelační deskriptory	3D	vektor	MCD	N
Konstanty substituentů	3D	skalár	NCD/LCD	L/I
WHIM deskriptory	3D	vektor	HCD	N
3D-MoRSE deskriptory	3D	vektor	LCD/MCD	N
GETAWAY deskriptory	3D	vektor	MCD	N
Povrchové a objemové deskriptory	3D	skalár	HCD/MCD	L
Kvantově-chemické deskriptory	3D	skalár	MCD/HCD	N/L
Compass deskriptory	3D	vektor	HCD	N
Interakční energie	4D	matice	HCD/RD	N
GRIND deskriptory	4D	vektor	HCD	N

polarizabilita, elektronegativita.

3.2. 1D deskriptory

Jako 1D reprezentace molekuly je brán seznam fragmentů, které tato molekula obsahuje. Tento seznam nemusí obsahovat všechny fragmenty, ze kterých se daná molekula skládá – může být pouze částečný a obsahovat jen např. funkční skupiny či substituenty, které jsou pro nás zajímavé. Z principu lze 1D reprezentaci vytvořit i z 2D a 3D reprezentace pomocí fragmentace molekuly podle určitých pravidel. 1D deskriptory jsou nejčastěji používány v podstrukturní analýze a podstrukturním hledání. Dále jsou mezi nimi deskriptory, které jsou počítány z příspěvků jednotlivých fragmentů.

LogP. Velmi často používaný deskriptor, protože ukazuje hydrofobicitu molekuly, což je velmi důležitá vlastnost u léčiv, která obecně souvisí s biologickou aktivitou léčiva (tj. schopnost vázat se na molekulární cíl – protein) a taktéž jeho schopností procházet buněčnými membránami. Experimentální měření *logP* může být složité, zejména u amfoterních iontů a velmi lipofilních či polárních látek, tudíž predikce této veličiny je nesmírně užitečná. Pro výpočet *logP* existuje několik algoritmů, které nejčastěji používají příspěvky jednotlivých fragmentů, např. *AlogP*[6,7], *XlogP*[8], *MlogP*[9] či *CLogP*[7] (obr. 2). Výsledkem těchto algoritmů je vždy aproximace reálného *logP*.

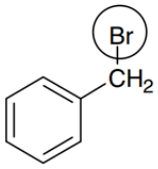
Topological Polar Surface Area (TPSA)[11]. Udává sumu ploch polárních atomů (obvykle kyslíků, dusíků a na ně připojených vodíků). Je to důležitý deskriptor při určování schopnosti transportu léčiva.

Molární refraktivita[1]. Je to míra totální polarizability jednoho molu látky, přičemž závisí na refraktivním indexu (n), hustotě (d) a molekulární hmotnosti (MW). Vzhledem k tomu, jak je tento deskriptor definován, je často používán jako míra sterického objemu molekuly a taktéž jako míra polarizability molekuly:

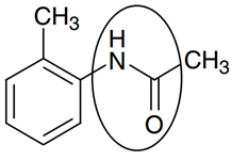
$$MR = \frac{n^2 - 1}{n^2 + 2}\frac{MW}{d} \qquad (1)$$

3.3. 2D deskriptory

K výpočtu je použita 2D reprezentace molekuly ve formě tabulky konektivity, grafu či lineární notace jako je SMILES, SMARTS, InChI apod. Jelikož 2D molekulární reprezentace již obsahuje topologii molekuly (tj. jak jsou atomy mezi sebou propojeny) a v omezené míře i stereochemii, je z ní možné získat užitečné informace o struktuře. Z grafové reprezentace molekuly vychází topologické deskriptory, což jsou většinou nejrůznější indexy postavené na tzv. teorii grafů. V molekulárním grafu jsou vrcholy tvořeny atomy a hrany reprezentují vazby mezi nimi (obr. 3). Další významnou skupinou 2D deskriptorů jsou



Bromide fragment	0.480
1 aliphatic isolating carbon	0.195
6 aromatic isolating carbons	0.780
7 hydrogens on isolating carbons	1.589
1 chain bond	-0.120
Total	2.924



NH-amide fragment	-1.510
2 aliphatic isolating carbons	0.390
6 aromatic isolating carbons	0.780
10 hydrogens on isolating carbons	2.270
1 chain bond	-0.120
1 benzyl bond	-0.150
ortho substituent	-0.760
Total	0.900

Obr. 2. **Ukázka výpočtu *CLogP* pomocí příspěvků jednotlivých fragmentů**[10]

2D „fingerprinty“, což jsou nejčastěji bitové vektory, kde pod každým bitem je zaznamenána přítomnost či nepřítomnost určitého strukturního rysu v molekule. Některé 2D „fingerprinty“ však používají i reálná čísla či kategorické hodnoty a kromě strukturních rysů jsou v nich zaznamenány i topologické informace.

Topologické indexy charakterizují molekulu podle její velikosti, míry větvení a celkového tvaru. Topologický index musí splňovat následující vlastnosti: jeho hodnoty by měly mít strukturní interpretaci, měl by korelovat s nějakou molekulární vlastností nebo významně zlepšit tuto korelaci v kombinaci s jinými deskriptory a měl by mít dostatečnou diskriminační sílu. Diskriminační sílu udává velikost grafu (měřeno počtem vrcholů), kdy se poprvé objeví degenerace indexu. Nyní si představíme několik nejznámějších topologických indexů.

Wienerův index[12]**.** Vymyšlen již v roce 1947 pro korelaci s bodem varu alkanů. Počítá se jako suma všech vzdáleností (měřeno topologicky počtem vazeb) mezi všemi dvojicemi uhlíkových atomů v molekule.

Indexy větvení (branching indices). První takový index byl navržen v roce 1976 panem Randićem[13], a to tak, aby jeho velikost rostla paralelně s vybranými vlastnostmi alkanů (např. bod varu). Je počítán z grafu neobsahujícího atomy vodíku a je založen na stupni vrcholu každého atomu.

$$I = \sum_{vazby} \frac{1}{\sqrt{\delta_i \delta_j}} \qquad (2)$$

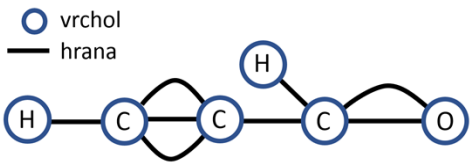
Suma ve vzorci jde přes všechny dvojice vazeb a δ_i a δ_j je stupeň *i*-tého a *j*-tého atomu (vrcholu) v dané vazbě. Každá vazba má tedy specifický příspěvek.

Tento index byl později zobecněn do tzv. chí-indexů konektivity[14]. Zaprvé, tyto indexy mají několik stupňů podle toho, kolik je do sumy zahrnuto vazeb (proto o nich mluvíme v množném čísle). Nultý chí-index (${}^0\chi$) má tedy sumu pouze přes samotné atomy, ${}^1\chi$ je stejný jako Randićův index, kde jde suma přes všechny dvojice atomů, a vyšší chí-indexy pak mají sumy přes všechny trojice, čtveřice atd. vzájemně propojených atomů. Za druhé, δ vyskytující se ve vzorci je ještě speciálně upravena jedním nebo druhým způsobem. První je tzv. jednoduchá delta, která je definována jako $\delta_i=\sigma_i-h_i$, kde σ_i je počet sigma elektronů *i*-tého atomu a h_i je počet vodíkových atomů k němu připojených. Druhá je tzv. valenční delta, definovaná jako:

$$\delta_i^v = Z_i^v - h_i \qquad (3)$$

kde Z_i^v je celkový počet valenčních elektronů (sigma, pi a volné elektronové páry) *i*-tého atomu. Díky zahrnutí dodatečných informací jsou chí-indexy schopny rozlišit např. $–CH_3$ od $–CH_2–$ při použití jednoduché delty. Ta již nestačí na rozlišení $–CH_3$ a $–NH_2$, nicméně valenční delta je již schopna toto rozlišit. Srovnání delta hodnot ukazuje tab. II a v tab. III jsou pak uvedeny chí-indexy pro různé isomery hexanu.

Topologických indexů existuje opravdu hodně, ale vzhledem k omezené délce tohoto článku není možné je zde popsat všechny. Alespoň však zmíníme názvy některých dalších indexů: dominantní vlastní číslo (Leading Eigenvalue)[15], Balabanův J index[16], Randićovo ID číslo[17], Kappa tvarové indexy (Kappa Shape Indices)[14], Hosoyův Z index[18].



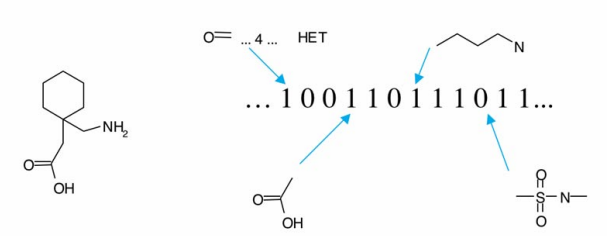
Obr. 3. **Ukázka molekulárního grafu. Grafy umožňují abstraktně vyjádřit nejrůznější problémy, přičemž jsou zanedbány geometrické vlastnosti a zapsána je pouze topologie. Grafy se proto hodí i pro popis molekul**

Tabulka II
Hodnoty jednoduché a valenční delty pro několik běžných typů atomů[10]

	Fragment			
Typ delty	$–CH_3$	$–CH_2–$	$=CH_2$	$–NH_2$
δ_i	1	2	1	1
δ_i^v	1	2	2	3

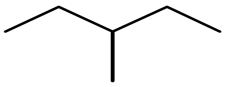
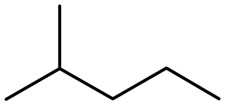
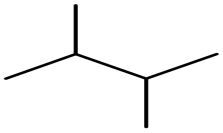
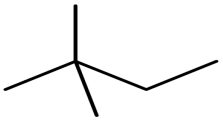
Dalšími významnými 2D deskriptory jsou 2D „fingerprinty“. Ty byly původně vyvinuty pro účely rychlého podstrukturního a podobnostního hledání. Mimo jiné je z nich však možné počítat deskriptory založené na příspěvcích fragmentů. „Fingerprinty“ se dělí na dvě hlavní skupiny: první je založena na předdefinovaných slovnících fragmentů (tzv. strukturní klíče) a druhá na hašovacích metodách. Jak již bylo zmíněno, nejčastěji je „fingerprint“ zaznamenán ve formě bitového vektoru. Jedná se tedy vektor, jehož hodnoty na jednotlivých pozicích jsou buď 0 nebo 1. Každá pozice vyjadřuje přítomnost (1) či nepřítomnost (0) určitého strukturního rysu.

Strukturní klíče. Zde má každá pozice ve „fingerprintu“ předdefinovaný určitý strukturní rys. Při tvorbě „fingerprintu“ se tedy postupně prochází všechny rysy, a pokud daná molekula rys obsahuje, je na této pozici nastavena hodnota 1 (obr. 4). Samotný pojem „strukturní rys“ je však velmi široký – může se jednat o pouhou přítomnost určitého strukturního fragmentu (např. „atom dusíku“, „karboxylová skupina v ortho poloze na benzenu“ či „alespoň tři atomy kyslíku“), ale také třeba o složitější rys popisující elektronovou konfiguraci (např. „uhlík v sp^2 hybridizaci“ nebo „dusík s trojnou vazbou“). Vytváření strukturních klíčů je časově náročné, protože se postupně zkouší všechny rysy. Následné využití už je ale rychlé – pokud např. chceme v databázi hledat určitou molekulu a každá molekula v ní má svůj strukturní klíč, tak se pro náš dotaz pouze vytvoří jeden strukturní klíč a ten je poté bitově porovnán se všemi klíči v databázi. Velmi známým strukturním klíčem je 166-bitový MDL[19] (někdy též zvaný MACCS či ISIS), ve kterém je definováno 166 strukturních rysů, které jsou považovány za důležité v medicinální chemii. Strukturní klíče mohou být



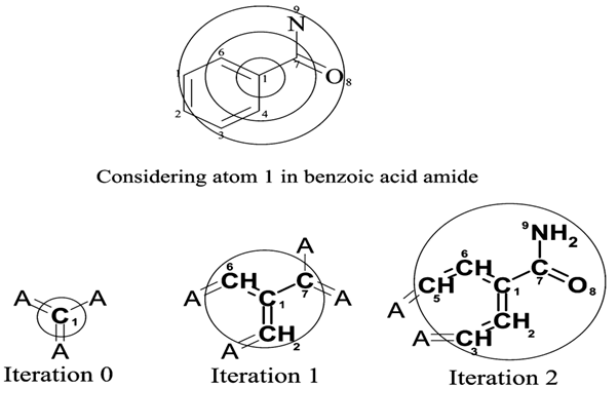
Obr. 4. **Názorná ukázka strukturního klíče[20]. Můžeme si povšimnout, že fragment se sírou molekula neobsahuje, a proto má její „fingerprint“ na této pozici nastavenou hodnotu 0**

Tabulka III
Chí-indexy pro různé isomery hexanu[10]

Fragment	Cesty délky 2	Cesty délky 3	Cesty délky 4	Cesty délky 5	$^0\chi$	$^1\chi$	$^2\chi$
	4	3	2	1	4,828	2,914	1,707
	5	4	1	0	4,992	2,808	1,922
	5	3	2	0	4,992	2,770	2,183
	6	4	0	0	5,155	2,643	2,488
	7	3	0	0	5,207	2,561	2,914

z principu vymýšleny pro konkrétní aplikace či oblasti výzkumu. Již zmíněný MDL klíč je např. v softwaru RDKit definován pomocí SMARTS výrazů. Není tedy problém si pomocí těchto výrazů definovat vlastní strukturní rysy.

Hašované „fingerprinty“. Tyto „fingerprinty“ nemají, na rozdíl od strukturních klíčů, předdefinované strukturní rysy a místo toho kódují informaci o všech strukturních fragmentech v molekule. Tyto fragmenty jsou vybírány různými způsoby, ale následný postup je společný všem hašovaným „fingerprintům“: fragmentům jsou přiřazeny unikátní identifikátory (čísla), která nějakým způsobem popisují strukturu fragmentu (např. za sebe zřetězená čísla atomů ve fragmentu), a následně jsou tato čísla tzv. zahašována do zvoleného intervalu. Velikost tohoto intervalu udává délku „fingerprintu“, tj. počet jeho bitů (např. 1024). Hašování převede vstupní identifikátor do čísla z intervalu (např. z intervalu 1 až 1024) a bit na této pozici se nastaví na jedničku. Hašování je operace, která zajistí, že jeden konkrétní fragment nastavuje vždy stejný bit. Občas však mohou vzniknout kolize, kdy jsou různé identifikátory zahašovány do stejné pozice ve „fingerprintu“. Typickým představitelem hašovaných „fingerprintů“ jsou



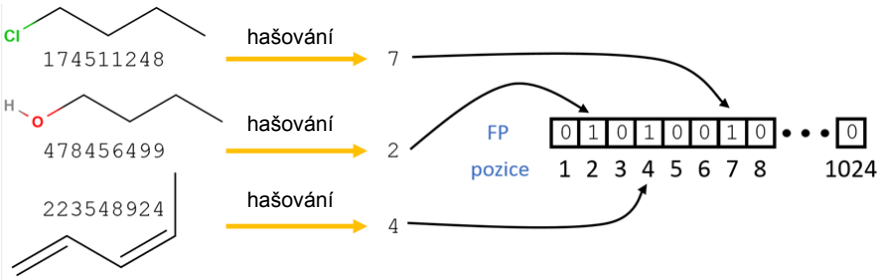
Obr. 5. **Ukázka iterativní tvorby konečného fragmentu o poloměru dvou vazeb v *ECFP*.[21] Tento fragment (viz. „Iteration 2“) byl postupně vytvořen z jediného atomu („C“ s číslem 1). Nakonec mu bude přiřazen unikátní identifikátor, který bude následně zahašován do čísla ve zvoleném intervalu**

tzv. **Extended Connectivity Fingerprints** (ECFP)[21], které jsou založeny na iterativním cirkulárním mapování fragmentů. U těchto „fingerprintů“ si můžeme zvolit, jak velké fragmenty (měřeno topologicky) budou zakódovány. Velmi zjednodušeně popsáno, nejprve se unikátně očíslují atomy v molekule a poté se v každé iteraci z každého atomu/fragmentu vytvoří nový fragment větší o jednu vazbu (každým možným směrem), kterému je následně přiřazen unikátní identifikátor na základě struktury fragmentu a jeho okolí. Iterace je zastavena, jakmile jsou unikátně očíslovány všechny fragmenty zadané velikosti (obr. 5). Nakonec jsou identifikátory fragmentů z poslední iterace zahašovány do zvoleného intervalu a je z nich vytvořen „fingerprint“ (obr. 6).

3.4. 3D a 4D deskriptory

3D reprezentace struktury již dovoluje do deskriptorů zahrnout i prostorové uspořádání molekuly. Není-li však 3D reprezentace dostupná z např. krystalografického experimentu, je třeba počítačově nalézt konformaci s nejnižší energií, což je pro velkou množinu molekul časově náročná procedura. 3D deskriptory se dají zařadit do několika skupin: kvantově-chemické deskriptory, objemové deskriptory, povrchové deskriptory, interakční energie a další. Opět jsou zde zastoupeny obdoby topologických indexů, ale jelikož se nyní pohybujeme v 3D prostoru, jsou zde tyto nazvány topografickými indexy a místo topologických vzdáleností používají vzdálenosti geometrické. Vzhledem ke složitosti 3D deskriptorů si pouze uvedeme názvy některých nejznámějších z nich: WHIM (Weighted Holistic Invariant Molecular) deskriptory[22], 3D-MoRSE (Molecule Representation of Structure based on Electron diffraction) deskriptory[23], GETAWAY (GEometry, Topology, and Atom-Weights AssemblY) deskriptory[24,25].

4D deskriptory taktéž vycházejí z 3D reprezentace molekulární struktury, ale obsahují „ještě něco navíc“. Některé popisují tzv. stereoelektronickou reprezentaci, která ukazuje vlastnosti molekuly spojené s elektronovou hustotou a interakcemi s okolím. Tyto deskriptory mají formu skalárního pole a typicky se používají v QSAR metodách založených na mřížce (Grid-Based QSAR).



Obr. 6. **Ukázka poslední fáze tvorby „fingerprintu“. Fragmenty z poslední iterace jsou „zahašovány“ do zvoleného intervalu (zde 1 až 1024) a postupně se tak nastavují bity ve „fingerprintu“ na hodnotu 1. Zde konkrétně mají fragmenty z poslední iterace poloměr dvou vazeb a jedná se tedy o „fingerprint“ *ECFP_4***

Další 4D deskriptory jsou založeny na tzv. stereodynamické reprezentaci molekuly, což je časově závislá reprezentace, která kromě 3D geometrie popisuje i flexibilitu, konformace, transportní vlastnosti apod. Tyto deskriptory se uplatňují v metodách dynamického QSAR, 4D podobnostní analýze molekul či 4D-QSAR (cit.[1]).

4. Podobnost molekul

Princip molekulární podobnosti (Molecular Similarity Principle)[26] říká, že strukturně podobné molekuly vykazují podobné fyzikálně-chemické či biologické vlastnosti. Zjišťování podobnosti molekul má proto velký význam v mnoha oblastech, kde je třeba nacházet látky s podobnými vlastnostmi. Příkladem může být vývoj nových léčiv, kdy je známa biologicky aktivní látka s léčebným účinkem, avšak některé její vlastnosti jsou nevyhovující (rozpustnost, toxicita), a proto je třeba hledat strukturně podobnou látku, která může mít lepší vlastnosti. Kvantitativně se podobnost dvou molekul vyjadřuje pomocí tzv. podobnostních koeficientů. Získané hodnoty párových podobností je pak možné použít v metodách shlukování a vizuálně tak prozkoumat podobnost molekul v určité množině.

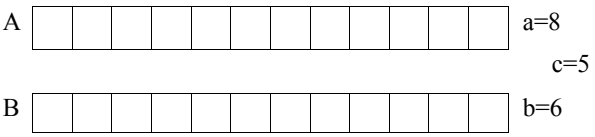
Jen na okraj zmíníme, že zde popsané podobnostní metody jsou založeny na molekulárních deskriptorech. Nicméně existují i metody založené na zarovnání molekul, kde se porovnávají jejich konformace či pole, která je obklopují. Tyto metody jsou obecně výpočetně velmi náročné, ale přinášejí přesné výsledky[10].

4.1. Podobnost založená na „fingerprintech“

Jelikož jsou „fingerprinty“ tvořeny bitovými vektory, tak je práce s nimi velmi rychlá. Jedná se asi o nejpoužívanější podobnostní metodu, která nachází uplatnění zejména v chemických databázích. Nejznámějším podobnostním koeficientem sloužícím pro porovnání binárních „fingerprintů“ je Tanimotův koeficient S_{AB} definovaný jako:

$$S_{AB} = \frac{c}{a+b-c} \qquad (4)$$

kde a je počet bitů s hodnotou 1 u molekuly A, b je počet bitů s hodnotou 1 u molekuly B a c je počet společných bitů s hodnotou 1 u molekuly A i B. Hodnota koeficientu leží v intervalu á0;1ñ, kdy hodnota 1 značí identické „fingerprinty“ (což však nemusí značit identické molekuly!) a hodnota 0 značí nulovou podobnost (tj. „fingerprinty“ nemají žádné společné „jedničky“). Příklad výpočtu Tanimotova koeficientu:



$$S_{AB} = \frac{5}{8+6-5} = 0,56$$

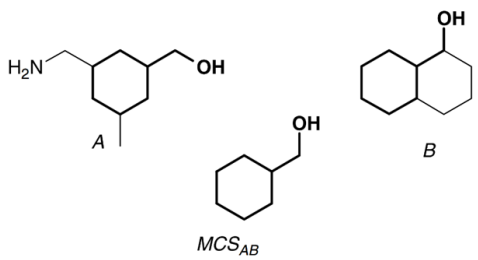
Kromě Tanimotova koeficientu můžeme ještě zmínit např. Dice koeficient, kosinovou podobnost či Tverskyho index[10].

4.2. Podobnost založená na vzdálenosti

Popíšeme-li molekuly deskriptory jinými než „fingerprinty“ (např. fyzikálně-chemickými deskriptory, topologickými indexy apod.), můžeme pak na ně pohlížet jako na body ve vícedimenzionálním prostoru a měřit mezi nimi vzdálenost. Zde vzdálenost představuje míru rozdílnosti molekul – čím je tedy vzdálenost větší, tím jsou molekuly rozdílnější. Jedná se tedy o opak podobnosti, kde větší hodnota příslušného koeficientu značila větší podobnost molekul. Mezi nejznámější míry vzdálenosti patří Euklidovská či Hammingova (Manhattanská) vzdálenost.

4.3. Podobnost založená na grafech

Nevýhodou výše uvedených podobnostních metod je, že nám neumožňují identifikovat lišící se podstruktury. Existují však metody založené na molekulárních grafech, které toto umožňují, a kromě jiného umí spočítat i podobnost molekul podobně jako to dělá např. Tanimotův koeficient. Tyto metody jsou většinou svou náročností tzv. NP-kompletní, tedy výpočetně velmi náročné, což je vylučuje z použití pro velké množiny molekul. Z podobnostních metod založených na grafech si uvedeme jednu velmi známou, která se nazývá Maximum Common Subgraph (MCS). Tato metoda vrací největší množinu atomů a vazeb, které jsou společné dvěma molekulárním grafům (obr. 7). Počty atomů a vazeb v MCS lze použít pro výpočet podobnostního koeficientu, který je velmi podobný Tanimotovu koeficientu, tj. leží mezi 0 a 1 a kvantifikuje míru podobnosti mezi dvěma molekulami[10].



Obr. 7. **Ukázka metody Maximum Common Subgraph. MCS_{AB} je podstruktura, kterou mají molekuly A a B společnou[10]**

5. Závěr

Deskriptory umožňují matematicky popsat molekuly, což je uplatňováno v mnoha oborech, kde je třeba hledat nové látky se specifickými vlastnostmi či predikovat neznámé vlastnosti látek. V tomto článku jsme si stručně představili význam deskriptorů, jejich základní vlastnosti a typy a také jakým způsobem je pomocí nich možné zjišťovat míru podobnosti či rozdílnosti molekul.
Deskriptorů existují řádově tisíce – naštěstí existují přehledné sborníky, které jsou čas od času doplněny novými deskriptory. Asi největším je již dříve zmíněný sborník *Molecular Descriptors for Chemoinformatics*, který obsahuje přes 3300 deskriptorů. Aby však deskriptory mohly být použity v praxi, je třeba jejich výpočet implementovat v počítači. Na to naštěstí existuje spousta „open-source" i komerčního softwaru, ať už ve formě nástroje s uživatelským rozhraním či knihovny pro nejrůznější programovací jazyky. Mezi často používaný „open-source" software pro výpočet deskriptorů patří např. RDKit[27], CDK[28] či PaDEL-Descriptor[29]. Dále dobrý přehled softwaru pro výpočet deskriptorů lze nalézt na webu Milano Chemometrics & QSAR Research Group[30].

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

LITERATURA

1. Todeschini R., Consonni V. (ed.): *Molecular Descriptors for Chemoinformatics*. Wiley, Weinheim 2010.
2. Škuta C., Svozil D.: Chem. Listy *111*, 747 (2017).
3. Svozil D.: Chem. Listy *111*, 738 (2017).
4. Čmelo I., Svozil D.: Chem. Listy *111*, 724 (2017).
5. Consonni V., Todeschini R.: http://www.moleculardescriptors.eu/tutorials/T3_moleculardescriptors_requirements.pdf, staženo 12.5.2017.
6. Ghose A. K., Crippen G. M.: J. Chem. Inf. Comput. Sci. *27*, 21 (1987).
7. Ghose A. K., Viswanadhan V. N., Wendoloski J. J.: J. Phys. Chem. A *102*, 3762 (1998).
8. Cheng T., Zhao Y., Li X., Lin F., Xu Y., Zhang X., Li Y., Wang R., Lai L.: J. Chem. Inf. Model. *47*, 2140 (2007).
9. Moriguchi I., Hirono S., Liu Q., Nakagome I., Matsushita Y.: Chem. Pharm. Bull. *40*, 127 (1992).
10. Leach A. R., Gillet V. J.: *Introduction to Chemoinformatics.* Springer Netherlands, Dordrecht 2007.
11. Prasanna S., Doerksen R. J.: Curr. Med. Chem. *16*, 21 (2009).
12. Wiener H.: J. Am. Chem. Soc. *69*, 17 (1947).
13. Randic M.: J. Am. Chem. Soc. *97*, 6609 (1975).
14. Hall L. H., Kier L. B., v knize: *The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling* (Lipkowitz K. B., Boyd D. B., ed.), sv. 2. J. Wiley, Hoboken 2007.
15. Lovász L., Pelikán J.: Periodica Mathematica Hungarica *3*, 175 (1973).
16. Balaban A.: Chem. Phys. Lett. *89*, 399 (1982).
17. Randi M.: J. Chem. Inf. Comput. Sci. *26*, 134 (1986).
18. Haruo H.: Bull. Chem. Soc. Jpn. *44*, 2332 (1971).
19. http://www.dalkescientific.com/writings/diary/archive/2014/10/17/maccs_key_44.html, staženo 17.5.2017.
20. https://icep.wikispaces.com/Characterizing+2D+structures+with+descriptors+and+fingerprints, staženo 15.5.2017.
21. Rogers D., Hahn M.: J. Chem. Inf. Model. *50*, 742 (2010).
22. Todeschini R., Gramatica P.: SAR QSAR Environ. Res. *7*, 89 (1997).
23. Devinyak O., Havrylyuk D., Lesyk R.: J. Mol. Graph. Model. *54*, 194 (2014).
24. Consonni V., Todeschini R., Pavan M.: J. Chem. Inf. Comput. Sci. *42*, 682 (2002).
25. Consonni V., Todeschini R., Pavan M., Gramatica P.: J. Chem. Inf. Comput. Sci. *42*, 693 (2002).
26. Klopmand G.: J. Comput. Chem. *13*, 539 (1992).
27. Landrum G.: http://www.rdkit.org/ , staženo 15.5.2017.
28. https://cdk.github.io/ , staženo 27.9.2017.
29. http://www.yapcwsoft.com/dd/padeldescriptor/, staženo 27.9.2017.
30. Group M. C. Q. R.: http://www.moleculardescriptors.eu/softwares/softwares.htm, staženo 19.7.2017.

J. Novotný[a,b] and D. Svozil[a,b] *([a] CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Institute of Molecular Genetics of the Academy of Sciences of the Czech Republic, Prague, [b] Laboratory of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology, Prague):* **Characterization and Similarity Measurement of Molecules Using Molecular Descriptors**

When working with molecules as data, you need to describe their structures and properties in a numerical way. And that is what molecular descriptors are used for. Molecules can thus be seen as points in multidimensional space and one can measure similarity and distance between them using various coefficients and metrics and also apply machine learning methods to predict not yet known physicochemical properties and biological activity of the system. In this article, principles and examples of molecular descriptors, as well as their use to measure similarity and distance between molecules, are shown.