Chem. Listy *91,* 237 - 250 (1997)

POČÍTAČOVÁ REPREZENTACE STRUKTUR CHEMICKÝCH SLOUČENIN

JAROSLAV SILHÁNEK

Vysoká škola chemicko-technologická, Technická 5, 166 28 Praha, E-mail: Jaroslav.Silhanek @vscht.cz

Došlo dne 10.I.1997

Obsah

1. Úvod

Chemie, jako každá vědní disciplína, potřebuje spolehlivý nástroj pro efektivní, spolehlivou a presnou komunikaci o základních objektech svého zkoumání, tj. o chemických sloučeninách. Od počátku historického období slouží k tomuto účelu nomenklatura. Název sloučeniny představuje prakticky jedinou, rozumně použitelnou možnost, jak vyjádřit trojrozměrnou strukturu sloučeniny textovou formou[1]. Je přirozené, že rozvoj počítačových informačních technologií vyvolal zájem o možnosti jejich využití i pro tak specifický úkol, jakým je reprezentace struktury chemické sloučeniny. Je možné konstatovat, že téměř třicetiletý vývoj řešení tohoto problému nabízí dnes již široké aplikační možnosti, včetně vcelku rozumné kompatibility jednotlivých produktů a formátů a v některých případech, jako jsou např. rozsáhlé strukturní báze dat, pak nejautoritativnější odpovědi na otázky o existenci či neexistenci té či oné sloučeniny. Tématem tohoto článku je především stručný přehled principů a možností počítačové reprezentace struktur chemických sloučenin včetně upozornění na způsoby využívání dostupných strukturních bází dat, jaké produkují Chemical Abstracts Service (CAS), ústavy Beilstein a Gmelinne nebo některé další instituce

2. Principy počítačové reprezentace struktur chemických sloučenin

Pomineme-li možnost zobrazovat strukturu chemické sloučeniny v počítači jako obrázek, můžeme hovořit o následujících řešeních uvedeného úkolu:

Reprezentace založené na fragmentaci struktury, způsoby reprezentace cyklických systémů, linearizované vyjádření struktury, topologická reprezentace struktury.

I když je nutné připustit, že v pravém slova smyslu je řešením počítačové reprezentace struktur pouze poslední alternativa, kteráje také dnes zcela dominantní, představují prvé tři způsoby nejenom kroky postupného vývoje, ale i stále užívané principy a v každém případě má smysl o nich stručně pojednat.

2.1. Reprezentace založené na fragmentaci struktury

2.1.1. Wiswesserova lineární notace a podobné systémy

Myšlenkové rozložení struktury chemické sloučeniny na jednotlivé fragmenty, jejich pojmenování a sestavení do výsledného řetězce podle určitých pravidel, je v podstatě principem všech systematických nomenklaturních postupů. Rovněž redukce nazvů fragmentů na všeobecně srozumitelné symboly (Me = methyl, Ph = phenyl a pod.) je běžnou praxí. Využití takové alfanumerické symboliky (např. přiřazení numerickým hodnotám význam fragmentů, 5 = pentyl) se zdálo být lákavé pro vypracování kompaktních lineárních notačních forem zápisů ještě před nástupem počítačových technologií[2]. Tyto aktivity podporoval i IUPAC (systém Dyson/IUPAC), ale prakticky jediného, poměrně značného rozšíření především v USA dosáhla tzv. Wiswesserova lineární notace, známá pod zkratkou WLN. Její popis, včetně souvisejícího vývoje, je v české literatuře dobře dokumentován[3-4], stačí proto shrnout, že se jedná o propracovaný systém umožňující pomocí souboru obsáhlých pravidel zaznamenat strukturu sloučeniny jako lineární řetězec kódů a čísel, a to jediný pro danou strukturu.

Popularita WLN vrcholila v šedesátých a počátkem sedmdesátých let, tedy v počátcích rozvoje počítačových technologií a představovala tak prakticky připravený systém, lákavý především z hlediska úspornosti počítačových pamětí. Této možnosti bylo také využito a jak samotná WLN, tak i podobné systémy, byly převáděny do podoby počítačových programů a bází dat. K nejpropracovanějším patřil zřejmě systém CROSSBOW (Computerized Retrival of StructureS Based On Wiswesser), vybudovaný a používaný v koncernu ICI[5]. V současné době je ale naprostá většina rozsáhlejších souborů sloučenin zpracovaných v těchto systémech převedena do jiných forem a WLN se stává historickou zajímavostí přesto, že umožňovala efektivní vyhledávání strukturních fragmentů. Důvodem je především ještě větší náročnost na dodržování formálních pravidel pro kódování, než je tomu u klasické nomenklatury.

2.1.2. Fragmentační kódy - reprezentace generických vzorců

Kromě přesného popisu struktury sloučenin seřazením strukturních fragmentů podle určitých závazných pravidel, je možné se spokojit buď s úplným nebo i jen částečným výčtem strukturních fragmentů, které se v dané sloučenině vyskytují. I tento jednoduchý požadavek má ale v řadě případů značný význam a byl poměrně často používán i v „předpočítačové" době, např. pro vyjádření přítomnosti strukturních fragmentů v souborech spektrálních dat uložených na děrných štítcích[6]. Podobnou situaci představuje používání obecných vzorců kombinujících přesně definované části struktur s řadou variabilně určených fragmentů nebo substituentů R_1, R_2... R_n. a pod. Zatímco při běžné publikační praxi tak nejenom uspoříme místo, ale dosáhneme i lepší přehlednost, používání takových vzorců v chemických patentech představuje dosud ne zcela uspokojivě vyřešený problém registrace obrovského množství sloučenin, které mohou být tímto způsobem v patentech uvedeny a *de iure* tedy existují. V případě chemických patentů se pro takové obecné strukturní vzorce používá označení Markushovy vzorce[7-8] nebo generické vzorce, od čehož se pak odvozují i názvy některých patentových bází dat, např. báze MARPAT CAS.

Již koncem padesátých let byl zahájen projekt systematického zpracovávání chemických patentů především z hlediska registrace struktur v patentech uváděných sloučenin u firmy Hoechst AG[9], kde byl vytvořen velmi propracovaný systém GREMAS (Generic REtrieval by MAgnetic tape Search), vycházející z praxe obecných strukturních vzorců a používající k jejich popisu tzv. fragmentační kódy. K tomuto systému pak přistoupila řada dalších firem, které v r. 1967 vytvořily konsorcium pod názvem Internationale Dokumentationsgesellschaft für Chemie (IDC)[10]. Zde byl pak vytvořen pravděpodobně dosud nepřekonaný systém registrace struktur chemických sloučenin v patentové literatuře, založený na hierachickém systému fragmentačních kódů, který je ale přístupný pouze firmám tvořících konsorcium. Později byl tento systém modifikován a doplňován o topologickou reprezentaci a objevují se úvahy o jeho zpřístupnění.

Popis a registrace struktur sloučenin obsažených v patentové literatuře pomocí fragmentačních kódů byl rovněž vyvíjen u největší světové patentové informační služby, společnosti DERWENT. Systém vycházel z formátů děrných štítků a byl několikrát přepracován do současné podoby tzv. CPI (Chemical Patent Index) kódů, jinak označovaných jako fragmentační kódy, nové fragmentační kódy, „punch codes" nebo chemické kódy[11]. Současná forma představuje soustavu cca 1000 kódů ve čtyřznakovém formátu písmene a zpravidla trojmístného čísla, s jejichž pomocí je možné zaznamenat přítomnost jednotlivých struk-

turních fragmentů v dané molekule, ale i použití dané sloučeniny, chemickou reakci, resp. její typ. Např. J131 je kód pro karboxyl na aromatickém jádře, M210 kód pro 1 až 6ti uhlíkatý řetězec, N203 značí reakci na benzenovém kruhu, Q333 anorganický pigment atd. Aby bylo možno provádět účinnou selekci struktur, jsou často používány negativní kódy, které zakazují přítomnost toho či onoho strukturního prvku. Možnosti systému jsou opravdu bohaté, ovšem jejich využívání je dosti náročné a je určeno spíše školeným specialistům. S ohledem na postupné změny je nutné brát ohled na starší a novější formy kódů a jejich platnost v určitém období. Jako určitá forma pomoci uživatelům je nabízen počítačový program TOPFRAG[12], který převádí nakreslené obecné (generické) vzorce do systému fragmentačních kódů, které pak mohou být použity pro online přístup k patentovým bázím dat firmy DERWENT.

2.1.3. Chemická nomenklatura v počítačové podobě

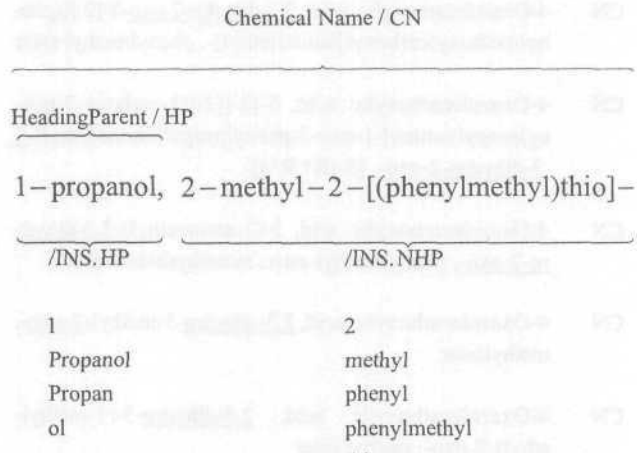
Je-li systematická nomenklatura založena na vhodné hierarchii a seřazení pojmenovaných strukturních fragmentů, můžeme elektronickou formu souboru takových názvů využít i pro vyhledávání látek s požadovanou strukturou. Tato cesta bude tím efektivnější, čím bude struktura názvosloví systematičtější a samozřejmě, čím bude takový soubor systematicky vytvářených strukturních názvů větší. Nejdůležitějšími elektronickými rejstříky nomenklaturních názvů chemických sloučenin jsou báze dat CAS, obsahující dnes cca 20 miliónů názvů, které lze poměrně velmi efektivně využívat.

Je celkem dobře známo, že CAS používá vlastní nomenklaturní pravidla, která se od pravidel IUPAC liší právě větším důrazem na systematičnost a jedinečnost názvu[13]. V tištěných rejstřících je používána tzv. invertovaná podoba názvu, tzn. že základ názvu (tzv. „Parent", resp. „Heading Parent") je uveden vždy na prvém místě a teprve za ním, odděleny čárkou, následují substituenty včetně lokantů v abecedním, resp. numerickém pořadí. Tato forma je zachována i v elektronické verzi a celý indexový název je uložen v poli CN (Chemical Name). Kromě toho jsou v samostatném poli (HP) uloženy samotné názvoslovné základy, neboli Heading Parent.

Hlavním rysem počítačové podoby souboru chemických názvů je hlubší a systematická fragmentace podle určitých závazných pravidel[14]. Prvým stupněm je vytvoření tzv. Natural Segments, které spočívá v odstranění všech závorek a pomlček z názvu. Tak např. z části názvu (2,4-dichloro) jsou vytvořeny přirozené fragmenty „2,4" a „dichloro". Druhým stupněm je vytvoření tzv. Basic Segments, což již představuje hlubší fragmentaci na složky mající význam základních strukturních jednotek, sufixy vyjadřující funkce, násobící prefixy, označení kruhů a další. Fragmentace na tyto jednotky probíhá podle určitého algoritmu a řídí se předem vytvořeným slovníkem fragmentů[14].

Takto je fragmentován celý název, přičemž vytvořené fragmenty jsou uloženy zvlášť v poli fragmentů základního názvu (pole INS.HP neboli Index Name Segment.Heading Parent) a zvlášť v poli fragmentů ostatních částí názvu (pole INS.NHP, neboli Index Name Segment.Non Heading Parent). Kromě toho je k dispozici ještě pole CNS (Chemical Name Segment), kde jsou uloženy pouze Natural Segments bez ohledu zda v Heading nebo Non Heading části a pole ONS (Other Name Segment), obsahující všechny segmenty z názvů, které nejsou indexovými názvy CA[15]. Na obr. 1 je naznačeno strukturování názvu sloučeniny, jednotlivá pole i fragmentace na složky. Celý tento zdánlivě složitý aparát slouží k tomu, aby bylo možné využitím vhodných logických operátorů* při formulaci dotazu v bázi REGISTRY poměrně velmi citlivě vyhledávat sloučeniny obsahující požadované strukturní fragmenty a tím i žádané sloučeniny.

Jako příklad využití je možné uvést případ, kdy hledáme buď všechny nebo jen určité konkrétní substituční deriváty



Obr. 1. **Struktura názvu sloučeniny v bázi REGISTRY**

* Kromě základních logických operátorů AND, OR a NOT, které se vztahují na celý dokument, umožňují moderní databázové vyhledávací systémy využití i tzv. proximitních operátorů, jejichž pomocí požadujeme, aby jimi spojené výrazy se v dokumentu vyskytovaly vedle sebe v uvedeném pořadí (operátor W), vedle sebe v libovolném pořadí (operátor A) nebo v libovolném pořadí i vzdálenosti, ale v jedné informační jednotce (operátor L), kterou je např. právě název sloučeniny.

nebo estery 2-oxo-2,3-dihydrooxazol-4-karboxylové kyseliny. Celkem snadno zjistíme (např. vyhledáním nejjednoduššího esteru v tištěných rejstřících CA), že základ názvu, Heading Parent, je „4-Oxazolecarboxylic acid". Kombinací požadavku, aby se tento fragment vyskytoval v poli /HP spolu s požadavkem na současný výskyt fragmentů „dihydro", „oxo", lokantu „2" a dvojice lokantů „2,3" v poli /INS.NHP nám poskytne informaci, že takových látek je pouze 11. Zobrazením jejich názvů (obr. 2) pak zjistíme, že naprostá většina odpovídá přesně našemu požadavku. Jak je ale vidět, kromě požadovaných sloučenin byla nalezena i sloučenina, která zřejmě do nalezeného souboru nepatří (poslední název na obr. 2). To je způsobeno nevyužitím všech možností dotazovacího jazyka, protože místo požadavku na současný výskyt požadovaných fragmentů v jedné strukturní jednotce (operátor L), mělo být správně požadováno postavení lokantu „2" a fragmentu „oxo" vedle sebe v uvedeném pořadí (operátor W). Pak by se ani tato špatně nalezená struktura neobjevila v souboru výsledků.

I když tato forma „počítačové reprezentace struktur" určitě neřeší všechny problémy spojené s vyhledáváním

FILE 'REGISTRY'
COPYRIGHT (C) 1995 American Chemical Society (ACS)

Formulace dotazu:

L2 11 4-OXAZOLECARBOXYLIC ACID/HP (L) 2,3/INS.NHP (L)DIHYDRO/INS.NHP (L) 2/INS.NHP (L) OXO/INS.NHP

Výběr nalezených sloučenin:

CN 4-Oxazolecarboxylic acid, 2,3-dihydro-2-oxo-5-[2-[[(phenylmethoxy)carbonyl]amino]ethyl]-, phenylmethyl ester

CN 4-Oxazolecarboxylic acid, 5-[2-[[2-[(l-carboxy-3-phenylpropyl)amino]-1 -oxo-3-phenylpropyl]amino]ethyl]-2_ ,3-dihydro-2-oxo-, [S-(R*,R*)]-

CN 4-Oxazolecarboxylic acid, 5-(2-aminoethyl)-2,3-dihydro-2-oxo-, phenylmethyl ester, monohydrobromide

CN 4-Oxazolecarboxylic acid, 2,3-dihydro-5-methyl-2-oxo-, methylester

CN 4-Oxazolecarboxylic acid, 2,3-dihydro-5-(1-methylethyl)-2-oxo-, methyl ester

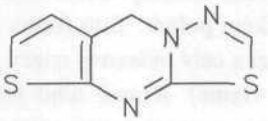
CN 4-Oxazolecarboxylic acid, 2-[4-[3-[[(butylamino)carbonyl]amino]-1 -(3,5-disulfophenyl)-1,5-dihydro-5-oxo-4H-pyrazol-4-ylidene]-2-butenylidene]-2,3-dihydro-3-(2-sulfoethyl)-, 4-methyl

Obr. 2. Ukázka vyhledání derivátů sloučeniny pomocí názvů

struktur chemických sloučenin, představuje velmi silný nástroj pro práci s nomenklaturními názvy látek a může poskytnout řadu zajímavých zjištění. Lze konstatovat, že ve spojení s požadavkem na vyhledání všech sloučenin s určitým sumárním vzorcem (tedy vlastně při počítačovém spojení vzorcového rejstříku s rejstříkem chemických názvů) představuje výše popsaná aplikace dnes zřejmě nejefektivnější (protože relativně lacinou) cestu k zodpovězení otázky, zda sloučenina s určitou strukturou existuje či nikoliv. Vhodnou kombinací sumárního vzorce s odhadnutými fragmenty předpokládaného názvu v příslušných polích lze totiž v naprosté většině případů dojít k relativně malé skupině potenciálně relevantních struktur, jejichž konečné zobrazení pak zodpoví uvedenou otázku.

2.2. Způsoby reprezentace cyklických systémů

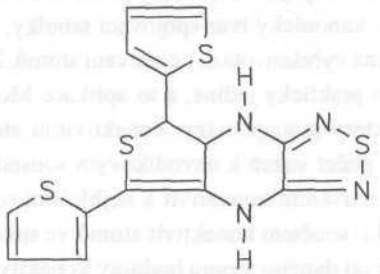
Možnost jednoduché, velmi kompaktní a praktické reprezentace cyklických systémů byla vyřešena již dávno před existencí počítačových technologií a je stále poměrně hojně používána. Svědčí o tom opakované vydávání díla The Ring Index[16], kde byl poprvné uplatněn princip popisu cyklických systémů založený na určení počtu kruhů a na popisu jejich velikosti a elementárního složení. Princip byl převzat v CAS jako tzv. Ring Index Handbook, který představuje velmi užitečný nástroj jak pro pojmenování cyklických systémů, tak i pro možnost vyhledávání odvozených sloučenin. Elektronická forma tohoto způsobu popisu cyklických struktur ale *nabízí* o něco více než forma tištěná. V tištěné podobě popisujeme strukturu pouze na úrovni elementárního složení kruhů a „ručně", v příslušném svazku Ring Index Handbook, dohledáváme konkrétní cyklickou strukturu. Hledáme-li např. v počítačové bázi dat všechny sloučeniny obsahující následující cyklickou strukturu /,



3-rings
5,5,6
C2N2S-C4S-C4S2
(Formulace sumárního vzorce kruhového systému, viz[16])

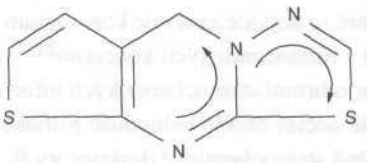
/

požadujeme vyhledání řetězce C2N2S-C4S-C4S2 v poli EA (Elemental Analysis), přičemž ale současně nalezneme i jiné sloučeniny obsahující cyklické struktury se stejným elementárním složením, ale jiným uspořádáním atomů a vazeb, např. //



II

Jednoduchou úpravou lineárního popisu kruhových systémů na základě pravidla, že při popisu elementárního složení každého kruhu vyjdeme od abecedně prvého neuhlíkatého atomu a postupujeme k abecedně dalšímu neuhlíkatému atomu nejkratší cestou, tedy ve výše uvedeném příkladu /:



N2CSC-SC4-NCNC3

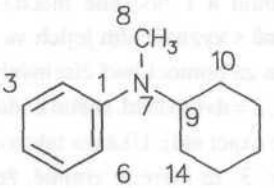
a požadavkem na vyhledání uvedeného řetězce přesněji popisujícího daný cyklický systém tentokrát v poli /ES (Elemental Sequence for Ring Systems), nalezneme s daleko větší pravděpodobností pouze ty látky, které skutečně hledáme.

Připomeňme, že uvedené možnosti reprezentace cyklických systémů demonstrované na příkladu báze REGISTRY CAS, je pochopitelně možné kombinovat s dalšími selekčními podmínkami, např. s požadavkem na přítomnost strukturních fragmentů z různých částí názvů, požadavkem na celkový sumární vzorec a pod.

2.3. Linearizované vyjádření struktury

Strukturu prakticky libovolné molekuly můžeme popsat linearizovaně jednoduše tak, že začneme z jednoho „konce" molekuly a v podstatě zcela mechanicky postupujeme atom po atomu v zápisu struktury s použitím dohodnutých zásad a pravidel ke druhému „konci" struktury. Pravidla se musí týkat situací při větvení řetězců a cyklických struktur. Běžně se lze setkat se dvěma systémy.

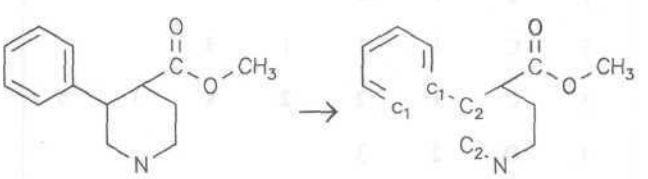
Především je to systém používaný Beilstein Institutem jako transportní formát pro struktury sloučenin označovaný akronymem ROSDAL[17] (Representation of Structure Diagram Arranged Linearly). Vychází z libovolného očíslování všech atomů dané struktury a postupného uvádění jednotlivých atomů/čísel s vazbami naznačenými pomlčkou nebo rovnítkem, přičemž uzavření kruhu vyplývá z uvedení stejných čísel na počátku a konci takového řetězce. Připojené řetězce jsou stejným způsobem uvedeny zvlášť, jsou odděleny čárkou a začínají číslem atomu, ze kterého vychází. Atomy vodíku se neuvádí, uhlíky jsou reprezentovány pouze čísly a ostatní prvky jsou označeny svým symbolem u čísla. Při opakování stejných vazeb je možné střední atomy vynechat, např. ///:



1-2=3-4=5-6=1 -7N-8,
7-8--14-9

III

Další podobné linearizované vyjádření struktury je označované akronymem SMILES[18,19] (Simplified Molecular Input Line Entry System). Je založeno na stejných jednoduchých pravidlech, jednotlivé atomy jsou rovněž očíslovány, ale čísla jsou zde používána jen pokud je to nutné pro označení větvení, uzavření kruhů a substituci a mohou se používat opakovaně. Jinak se přímo používají symboly atomů, velká písmena pro lineární řetězce a malá písmena pro aromatické kruhy. Větvení se uzavírá do závorek tak, jak jsme zvyklí z běžného psaní vzorců, u kruhů se nejdříve provede myšlenkové rozpojení a cyklus je zapsán použitím stejným čísel pro spojené atomy. Formát SMILES je poměrně často používán pro přenos struktur chemických sloučenin mezi různými programovými aplikacemi, jako jsou např. programy pro modelování a pod. Jako příklad je uvedena reprezentace poněkud složitější molekuly jako ukázka možností těchto relativně velmi jednoduchých metod *IV:*



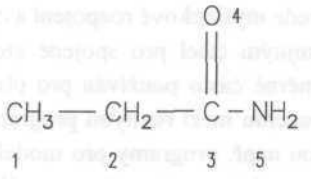
Formát SMILES: c1ccccc1C2C(C(=O)OC)CCNC2

IV

2.4. Topologická reprezentace struktury

Při topologickém popisu se na strukturu chemické sloučeniny díváme jako na orientovaný graf, který je popsán výčtem jednotlivých uzlů (atomů) a hran (vazeb) pomocí tzv. spojovacích tabulek (Connectivity tables, Verknüpfungstafel). Analogie mezi strukturou chemické sloučeniny a topologickým grafem je také východiskem pro řadu operací s touto formou reprezentace. Vytvoření topologického popisu formou spojovací tabulky je velmi jednoduché, spočívá v očíslování všech atomů a v podstatě mechanickém výčtu jednotlivých atomů s vyznačením jejich vazeb ke všem atomům sousedním za pomoci buď číselných označení (1 = jednoduchá vazba, 2 = dvojná atd.) nebo kódů (SE = single exact, DE = double exact atd). Ukázka takové nejjednodušší tabulky je na obr. 3, ze které je zřejmé, že každé spojení je zde uvedeno dvakrát. Tento typ tabulky je označován jako redundantní reprezentace, která sice vyžaduje více paměti, ale protože se s ní lépe pracuje při vyhledávání struktur, je často využívána. Je ovšem možné vytvořit i tabulky neredundantní. Číslování atomů může být principiálně libovolné, konkrétní tvar spojovací tabulky bude pak pochopitelně různý, ale vždy bude zpětně generována stejná struktura.

Bylo věnováno hodně úsilí standardizaci topologické reprezentace, jednak z důvodů aplikace vyhledávacích algoritmů a jednak pro možnost výměny dat. Prvním úkolem



Číslo atomu	Symbol atomu	Druh vazby	Číslo atomu	Druh vazby	Číslo atomu	Druh vazby	Číslo atomu
1	C	1	2				
2	C	1	1	1	3		
3	C	1	2	2	4	1	5
4	O	2	3				
5	N	1	3				

Obr. 3. **Příklad redundantní spojovací tabulky**

je vytvoření jediné spojovací tabulky pro každou sloučeninu, neboli tzv. kanonický tvar spojovací tabulky, což především znamená vyřešení otázky číslování atomů. Z mnoha řešení se ujalo prakticky jediné, a to aplikace Morganova algoritmu[20], který pracuje s tzv. konektivitou atomu vyjádřenou jako počet vazeb k nevodíkovým sousedům. Postupným nahrazováním konektivit k nejbližším sousedům (malá celá čísla) součtem konektivit atomů ve stále větších vzdálenostech od daného atomu hodnoty konektivit rostou a současně se diferencují. Atomu s nejvyšší hodnotou konektivity je pak přiřazeno číslo 1 a na základě podobných pravidel jsou přiřazována další čísla. Rozsah tohoto článku bohužel nedovoluje věnovat této, pro topologickou reprezentaci velmi důležité otázce, podrobnější diskusi a odkazujeme proto na původní literaturu[20].

Dalším krokem jsou pak snahy o vytvoření standardních spojovacích tabulek[21] (Standard Molecular Data Format, SMD Format), o které se nejvíce zasazuje konsorcium evropských chemických a farmaceutických koncernů[22,23]. Jinou logickou extenzí je zahrnutí stereochemických informací. Přesto, že v zásadě nečiní obtíže jednoduše přiřadit ke spojovací tabulce běžné stereochemické deskriptory R, S, E nebo Z a další, pozornost se soustřeďuje spíše na „počítačová" řešení. Ta většinou využívají výše zmíněný Morganův algoritmus a přiřazují pořadí priority skupinám vázaných na chirální centrum[24]. Tedy obdobný postup jako Cahn-Ingold-Prelogova pravidla. Řešení této problematiky bylo motivováno už požadavky počítačových syntéz a již v r. 1974 navrhují Wipke a Dyott[25] systém SEMA (Stereochemically Extended Morgan Algorithm) začleněný do neredundantních spojovacích tabulek. Další přechod od topologie k topografii vede jednak ke krystalografickým databázím, ale i ke koncepci reprezentace prostorových poměrů doplněním spojovacích tabulek o další parametry atomů i vazeb charakterizujících prostorové poměry. Pro generování trojrozměrných dat na tomto principu je k dispozici program CONCORD vyvinutý na texaské univerzitě[26,27], s jehož pomocí byly získány asi 4 milióny 3-D strukturních reprezentací ze spojovacích tabulek převzatých z báze REGISTRY. V této bázi pak nalezneme upozornění, zda pro danou sloučeninu 3-D data existují a pokud ano, je možné je z této báze převzít a použít např. pro molekulární modelování.

Bez ohledu na konečné vyřešení otázky standardního formátu je možné konstatovat, že topologická reprezentace je dnes bezpochyby nejrozšířenější formou počítačové reprezentace struktury chemických sloučenin a je na ní založena řada praktických aplikací.

3. Využití počítačové reprezentace struktur chemických sloučenin

Možnost pracovat se strukturou chemických sloučenin pomocí počítačových technologií může být využita v řadě aplikací, od vkládání struktur sloučenin do textů v textových editorech, až po studium prostorových interakcí. Často je ale hlavním zájmem možnost vytvářet seznamy sloučenin včetně jejich graficky vyjádřených struktur, neboli budovat strukturní báze dat. Na rozdíl od fragmentačních principů kódování struktur, kde je vztah mezi ukládáním struktur a jejich vyhledáváním vcelku průhledný, je situace v případě topologické reprezentace struktury zásadně odlišná. Záznam struktury v tabelárním tvaru je sice velmi jednoduchý, ale zpětné vyhledání určité struktury nebo její části představoval (a stále ještě představuje) problém, jehož uspokojivé řešení si vyžádalo nemalé úsilí a v rozsahu tohoto článku může být zmíněn jen velmi všeobecně. Vývoj topologické reprezentace do podoby široce použitelného programového nástroje přitom probíhal v podstatě dvojím směrem.

Především byl topologický popis vyvíjen pro budování strukturních bází dat v pravém slova smyslu, tj. soupisů chemických sloučenin obsahujících jak jejich textový, nomenklaturní popis a samozřejmě neomezené množství dalších údajů, tak i graficky vyjádřenou strukturu s možností danou látku vyhledat nejen podle jejího názvu, ale rovněž podle nakreslené struktury, resp. jejích částí. Pionýrská práce v tomto směru byla provedena v CAS při vytváření báze dat REGISTRY[28], která dnes představuje v podstatě autoritativní celosvětový registr principiálně všech existujících chemických sloučenin zmíněných v informačních zdrojích od r. 1967.

Jiný směr vývoje topologického popisu vedl k vytvoření co nejuniverzálnějších „prázdných" databázových systémů umožňujících samostatné efektivní ukládání vybraných látek a vytváření vlastních strukturních bází dat. Takových programů je dnes již nabízena celá řada, jako ilustraci je možno uvést produkty společnosti Molecular Design Ldt.[41], MACCS, REACCS nebo ISIS. Prostřednictvím těchto „prázdných" databázových systémů je samozřejmě možné využívat i řadu komerčních bází dat, ovšem oddělení databázového systému od konkrétní báze je naprosto zřetelné (je možné je např. samostatně nakupovat).

Zaměříme se dále především na volně dostupné strukturní báze dat, které dnes představují jeden z nejdůležitějších informačních zdrojů pro všechny chemické obory.

3.1. Strukturní báze dat

3.1.1. Chemical Abstracts - báze dat REGISTRY

Jako jedno z nejprozíravějších rozhodnutí v oblasti zpracovávání chemických informací je možno považovat jednoznačnou orientaci na topologickou reprezentaci, pro kterou se rozhodli v CAS v r. 1967, kdy ještě vše hovořilo pro počítačově méně náročné fragmentační metody. To, že CAS zahájila v tomto roce systematické ukládání všech sloučenin, se kterými se při sekundárním zpracovávání primárních zdrojů setkává, v podobě spojovacích tabulek, vyústila dnes ve vybudování největší světové databanky chemických sloučenin zahrnující i jejich grafické strukturní reprezentace. Tato databanka byla nejdříve pouze interní záležitostí CAS, kam se sloučeniny pouze ukládaly a nebylo možno je vyhledávat, ale záhy byly informace zpřístupňovány prostřednictvím sítí a v současné době představuje tato báze dat pod označením REGISTRY volně přístupný soupis cca 15 miliónů chemických sloučenin (včetně jejich struktur a dalších údajů), o kterých byla zmínka v primárních zdrojích od r. 1967 (později rozšířeno od r. 1957) do současnosti.

Grafická reprezentace struktury každé ukládané sloučeniny umožnila jednoduchou a účinnou kontrolu, zda daná sloučenina je již v systému uložena nebo zda je zcela nová. Tím, že CAS při zápisu struktury přiřadila každé sloučenině poprvé ukládané do této báze dat pořadové číslo (Registry Number), které sloužilo především jako počítačová adresa, na které se shromažďovaly odkazy na všechny další informace o dané látce, vytvořila unikátní registrační systém chemických látek a nabídla tak maximálně jednoduchý způsob jednoznačné identifikace všech chemických sloučenin[28]. Označování chemických sloučenin registračními čísly CAS se velmi rychle vžilo a rozšířilo se jako univerzální identifikace látek i do komerční a legislativní oblasti. Připomeňme, že toto číslo nemá naprosto žádnou souvislost se strukturou dané látky a je číslem čistě pořadovým. Jeho formální struktura sestává z maximálně šestimístného čísla, po pomlčce následuje dvoumístné číslo a po další pomlčce pak jednomístné, počítačem generované kontrolní číslo ověřující správnost předchozí sekvence, tedy NNNNNN-NN-N. Kategorická zásada, že každé sebemenší strukturní obměně nebo i nepřesnosti v jednoznačném určení struktury musí být přiřazeno jiné, v sekvenčním pořadí právě platné číslo, vede např. k tomu, že není-li v primárním zdroji jasně uvedeno, že se jedná o racemát (i když tomu všechny okolnosti jednoznačně nasvědčují), má taková lát-

ka jiné registrační číslo, než je-li údaj o racemické směsi explicitně konstatován.

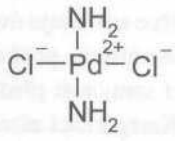
Na obr. 4 je uvedena ukázka záznamu z báze dat REGISTRY, tedy jakási registrační karta chemické sloučeniny. Karta obsahuje kromě vlastního registračního čísla (rejstříkové, tak i všechny ostatní zaznamenané názvy dané látky), molekulový vzorec, strukturní vzorec a některé další informace, jakož i odkazy najiné báze dat, kde se informace o dané sloučenině vyskytují (dostupné v databázovém středisku STN International[42], tak i v jiných databázových střediscích). Zajímavý je i údaj o počtu odkazů v různých bázích CAS, což dává představu o míře zájmu o danou sloučeninu. Přístup do této báze je zatím možný pouze prostřednictvím sítě (dnes nejvýhodněji Internetem), cena jednoho uvedeného záznamu je v současné době **1,17** USD. Práce s bází REGISTRY je podrobně popsána v řadě materiálů vydaných databázovým střediskem STN International[29].

3.1.2. Další strukturní báze dat Významnější je

Další rozsáhlou strukturní bází dat je elektronická verze Beilsteinova kompendia s cca 7 milióny organických sloučenin, přístupná buď prostřednictvím sítě nebo dnes nabízená i pro lokální instalaci pod označením „CrossFire“[30]. Analogická anorganická báze GMELIN obsahuje zatím kolem 1 miliónu struktur, ale je velmi rychle doplňována a pro r. 1996 je kromě síťového přístupu ohlášena i verze pro lokální instalaci ve stejném programovém prostředí jako BEILSTEIN, tedy v systému CrossFire. Je nutné konstatovat, že počítačová strukturní reprezentace anorganických sloučenin představuje poněkud složitější problém a je zatím méně propracována než reprezentace organických sloučenin. Ani báze REGISTRY neuvádí u všech registrovaných anorganických sloučenin struktury a elektronická verze Gmelina v prostředí CrossFire je proto očekávána s velkým zájmem.

Dále je možné uvést několik menších strukturních bází dat určených pro práci pod systémem ISIS nebo MACCS (viz kap. 3.3.), nebo bázi Available Chemical Directory (dříve Fine Chemical Directory), což jsou v podstatě spojené katalogy chemikálií, které je možné prohlížet pomocí graficky vyjádřených struktur nebo jejich částí[31]. ale možnost vyhledávat struktury látek v tzv. reakčních bázích dat, které představují nadstavbu strukturních bází dat. O těchto bázích se zmíníme poněkud podrobněji v samostatné kapitole, zde pouze konstatujme, že stejně jako ve strukturních bázích individuálních slou-

ANSWER 1 REGISTRY COPYRIGHT 1996 ACS
RN 14323-43-4 REGISTRY Regis <—— čn íčíslo
CN Palladium, diamminedichloro- (8CI, 9CI) (CA INDEX NAME)
OTHER CA INDEX NAMES:
CN Diamminedichloropalladium (6CI, 7CI)
OTHER NAMES:
CN Chlorpalladosamine
MF Cl2 H6 N2 Pd
CI CCS <——CCS=Coordination Compounds
LC STN Files: CA, CAOLD, CAPLUS, CAPREVIEWS, CHEMCATS, CHEMLIST, CSCHEM, GMELIN*, IFICDB, IFIUDB, MSDS-OHS, MSDS-SUM, RTECS*, TOXLINE, TOXLIT, USPATFULL
(*File contains numerically searchable property data)
Other Sources: EINECS**, NDSL**, TSCA**
(**Enter CHEMLIST File for up-to-date regulatory information)



2 REFERENCES IN FILE CAPREVIEWS
68 REFERENCES IN FILE CA (1967 TO DATE)
68 REFERENCES IN FILE CAPLUS (1967 TO DATE)
26 REFERENCES IN FILE CAOLD (PRIOR TO 1967)

Obr. 4. **Příklad záznamu z báze dat REGISTRY**

čenin, můžeme v těchto zdrojích hledat jak struktury výchozích látek nebo produktů, tak i strukturní fragmenty v kontextu jejich přeměn.

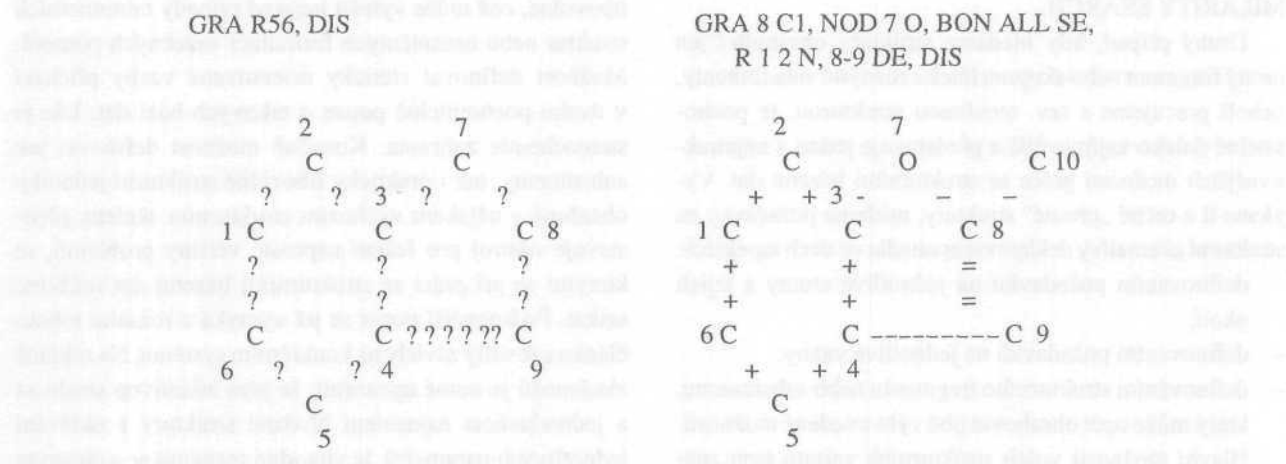
Strukturní báze dat byla budována i pro rozsahem jedinou konkurenci Chemical Abstracts, Referativnyj Žurnal, vytvářený moskevským ústavem VINITI. Na této práci se podílel i ústav Zentrale Informationsverarbeitung Chemie (ZIC) bývalé NDR, což vedlo zřejmě k tomu, že tato báze, která obsahuje údajně 2,5 miliónů organických a organokovových sloučenin za období 1975-1991, přešla do správy společnosti InfoChem GmbH[32]. Informace o přípravě sloučenin byly převedeny do reakční báze InfoChem, z níž jsou zpřístupňovány části na mediu CD-ROM pod označením ChemReact10 nebo nejnověji ChemReact41. O tom, zda producent referátového časopisu Referativnyj Žurnal pokračuje v budování strukturní báze dat bez účasti východoněmeckého ústavu nejsou zprávy.

3.2. Práce se strukturními bázemi dat

Prvým předpokladem práce se strukturní bází dat je principiální možnost zobrazit strukturu na obrazovce počítače. Tento problém je dnes vyřešen všeobecně rozšířeným grafickým rozhraním současných počítačů a na trhu je celá řada komunikačních programů umožňujících velmi pohodlné kreslení struktur. Použití takových programů ale není nezbytnou podmínkou a je docela dobře možné pracovat pouze s řádkovým monitorem, což byla dlouhou dobu standardní situace. Pro ilustraci je možné uvést jednoduchý příklad vytváření struktury pomocí programovacího jazyka používaného v CAS, který současně demonstruje hlavní aspekty budování strukturního dotazu platné prakticky ve všech strukturních bázích dat.

V zásadě jakoukoliv strukturu lze vytvořit pomocí tří příkazů: GRA (Graph) - kreslící, resp. vytvářející struktury, NOD - definující atomy (uzly) a BON - pro definice vazeb. Tak např. příkazem GRA R56 na obr. 5 jsou vytvořeny dva kondenzované kruhy, pětičlenný a šestičlenný, které jsou automaticky očíslovány a příkazem DIS zobrazeny. Dále můžeme s využitím zobrazeného číslování „přikreslit" další substitutenty, např. GRA 8 C1 připojí jednouhlíkatý řetězec na atom č. 8 (a dostane číslo 10), příkaz NOD 7 O vymění implicitní uhlík na místě atomu č. 7 za kyslík, příkazem BON ALL SE jsou nejdříve všechny vazby definovány jako jednoduché (SE = single exact), načež vazby v prvém kruhu (R 1 2) jsou určeny jako aromatické (podle uzance jsou zobrazeny křížkem), neboli v terminologii strukturního jazyka jsou „normalizované", což je vyjádřeno parametrem N a konečně mezi atomy 8 a 9 je požadována dvojná vazba (DE = double exact). Příkaz DIS (display) vždy zobrazí vzniklou strukturu, kterou pak můžeme korigovat. Vodíkové atomy se nezobrazují a jsou doplňovány implicitně. Tento způsob „kreslení" struktur má samozřejmě řadu dalších příkazů umožňujících definovat např. náboj na atomu, požadovat v určitém místě struktury alternativní atomy nebo vazby, definovat obecně pojmenované substituenty, otevírat konkrétní místa struktury pro další substituci a řadu dalších zpřesnění strukturního zadání, vždy s odkazem na číslo konkrétního atomu (uzlu).

Grafické komunikační programy umožňují zadávání struktury pochopitelně mnohem elegantněji s následným exportem do formátu, který je vyžadován konkrétní struk-



Obr. 5. Formulace dotazu na strukturní bázi dat na textovém terminálu

turní bází dat, se kterou chceme pracovat, nejčastěji tedy REGISTRY nebo BEILSTEIN. Ovšem v ukázce na obr. 5 naznačený způsob vytváření strukturního zadání, tj. přesná definice atomů (uzlů) nejenom co se jejich kvality týká, ale také určení jejich okolí a všech vazeb, musí být dodržen a představuje současně také jeden z nejmocnějších nástrojů využití strukturních bází dat k vyhledávání požadovaných sloučenin.

V zásadě je možno pracovat se strukturními bázemi dat na dvou úrovních:

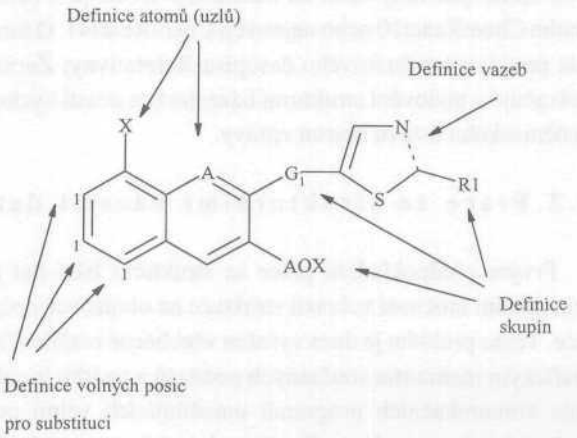
- buď hledáme určitou konkrétní sloučeninu, jejíž strukturu jednoznačně nakreslíme a definujeme její části. Hledaná sloučenina (struktura) tedy v dané bázi dat buď existuje nebo ne. Tento úkol je zpravidla označován jako EXACT SEARCH,
- nebo hledáme množinu sloučenin, které vyhovují určitým alternativním strukturním požadavkům, např. obsahují další substituenty nebo alternativní atomy a pod. V takovém případě je přesně zadána jen část struktury nebo jen její fragment a jsou definovány požadavky na možné rozšíření. Tento přístup se označuje jako SUBSTRUCTURE SEARCH.

Pojem „přesná struktura" je zpravidla možné blíže specifikovat, např. je možné zadání omezit nebo naopak rozšířit o geometrické či optické izomery, izotopické záměny, o ionizované nebo radikálové formy a pod. Hledaná struktura může být také součástí komplexů nebo solí. Poslední jmenovaná možnost je v bázi REGISTRY realizována zvláštním algoritmem hledání, označovaném jako FAMILY SEARCH. Problematika vyhledávání alternativních struktur má blízko k problémům podobnosti struktur, na což reagují některé systémy (ISIS) zavedením funkce SIMILARITY SEARCH.

Druhý případ, kdy hledáme struktury obsahující jen určitý fragment nebo skupinu látek s různými substitutenty, neboli pracujeme s tzv. otevřenou strukturou, je pochopitelně daleko zajímavější a představuje jednu z nejatraktivnějších možností práce se strukturními bázemi dat. Vyjdeme-li z určité „přesné" struktury, můžeme požadavky na strukturní alternativy deklarovat zpravidla ve třech aspektech:

- definováním požadavků na jednotlivé atomy a jejich okolí,
- definováním požadavků na jednotlivé vazby,
- definováním strukturního fragmentu nebo substituentu, který může opět obsahovat obě výše uvedené možnosti.

Hlavní možnosti voleb strukturních variatn jsou znázorněny na obr. 6. Grafické komunikační programy obsahují někdy až marnotratnou nabídku možností. Tak je např. možné požadovat kterýkoliv prvek z Mendělejevovy tabulky. Pro každý atom je možné určit nejenom jeho kvalitu, náboj, izotop, radikál, ale především jej otevřít pro substituci jedním, dvěma nebo maximálním počtem substituentů. Jinak lze požadavky na substituci deklarovat pomocí tzv. „H-count", neboli určením počtu vodíkových atomů, které musí na daném atomu zůstat. Velmi důležitým nástrojem deklarace substruktur je dovolení nebo zakázání možnosti, aby z daného atomu vycházely při substituci cyklické vazby. Tímto požadavkem zpravidla velice podstatně rozšiřujeme nebo naopak zužujeme velikost nalezeného souboru látek.



Obr. 6. Možnosti definování strukturních variant

V případě vazeb je samozřejmě možné požadovat buď jednoznačně jednoduché nebo násobné vazby, nebo jejich různé alternativy, přičemž často je užitečné definovat vazby libovolné, což může vyřešit nejasné případy tautomerních struktur nebo neustálených formulací vazebných poměrů. Možnost definovat stericky orientované vazby přichází v úvahu pochopitelně pouze u takových bází dat, kde je stereochemie zahrnuta. Konečně možnost definovat jak substituenty, tak i prakticky libovolné strukturní jednotky obsažené v nějakém výchozím strukturním skeletu, představuje nástroj pro řešení naprosté většiny problémů, se kterými se při práci se strukturními bázemi dat můžeme setkat. Podrobnější popis se již vymyká z rozsahu tohoto článku a je vždy závislý nakonkrétním systému. Na základě zkušenosti je nutné upozornit, že přes zdánlivou snadnost a jednoduchost nakreslení hledané struktury i zadávání jednotlivých parametrů, je vhnodné seznámit se s přesným významem deklarovaných požadavků, protože ty zásadním způsobem ovlivňují výsledek jakéhokoliv vyhledávání.

3.2.2. *Vlastní vyhledávání ve strukturních bázích dat*

Běžného uživatele strukturních bází dat v zásadě nemusí vůbec zajímat, jakým způsobem je jeho požadavek na vyhledání určité struktury počítačem zpracováván, vyjma otázky, jak je vyhledání rychlé nebo pomalé. Ale i tato otázka je důležitá především při práci s bázemi dat uloženými na vzdálených databázových střediscích, se kterými pracujeme prostřednictvím sítí a většinou platíme za tzv. „connect time". Pracujeme-li s lokálně instalovanou strukturní bází nebo s bází na mediu CD-ROM, není ani doba prohledávání kritická. Pro ilustraci je možno uvést, že běžné doby zpracování strukturního dotazu se pohybují řádově v minutách a jednotlivé systémy mají různé formy zabezpečení proti riziku příliš dlouhých vyhledávacích dob v případě nevhodně formulovaného dotazu.

Ve skutečnosti představují ale vyhledávací algoritmy nejdůležitější složku každého systému a bylo třeba jak rozsáhlé práce programátorské, tak i dosažení určité úrovně počítačových technologií, aby mohl být splněn úkol, nalézt v několika miliónech topograficky zaznamenaných struktur chemických sloučenin konkrétní sloučeninu.

Teoreticky se jedná o to, nalézt ve strukturní bázi graf, který je izomorfní s grafem hledané sloučeniny, což je v případě grafů majících větší počet uzlů a hran obecně obtížný problém. V terminologii strukturních bází dat se používá termín „atom-by-atom-search"[33]. Je zřejmé, že tento úkol bude tím snažší, čím menší bude množina grafů, které pro splnění zadaných strukturních parametrů přicházejí vůbec v úvahu. Obecné řešení vyhledávacích algoritmů tedy spočívá ve dvoustupňovém procesu, kdy je nejdříve vhodnou preselekcí (screening) vybrána vhodná podmnožina z celé báze dat, na kterou jsou pak aplikovány časově náročné postupy „atom-by-atom-search". Problém se proto soustřeďuje na volbu takových preselekčních kriterií, které vedou rychle a spolehlivě k dostatečně malému souboru potenciálně relevantních struktur.

Poměrně jednoduchá je tato otázka, pokud hledáme konkrétní uzavřenou strukturu, neboli tzv. „EXACT SEARCH", kde je přirozeným omezujícím kriteriem sumární vzorec. Podobně jakýkoliv výskyt méně obvyklých heteroatomů v sumárním vzorci je rovněž účinným selekčním faktorem. Obecně je problém preselekce nejčastěji řešen prostřednictvím strukturních fragmentů, označovaných někdy jako fragmentační kódy, jindy jako topologické indexy[34], screens[35], hash codes[36] a pod. Na rozdíl od fragmentačních kódů používaných při popisu generických struktur, o kterých jsme hovořili v kap. 2.2.2., se zde jedná o kódy (nebo indexy), vyjadřující přítomnost strukturních stavebních prvků, jako např. přítomnost řetězce O-C-C-C-N nebo přítomnost terciárního uhlíku s další specifikací vycházejících vazeb (acyklické nebo cyklické) a pod. Tyto kódy jsou generovány automaticky ze spojovacích tabulek jak na straně hledané struktury, tak i báze dat, přičemž nepřítomnost strukturního fragmentu v popisu určité struktury v bázi je dostatečným důvodem pro vyřazení této struktury z dalšího prohledávání. Velmi důležitou součástí použití fragmentačních kódů je jejich statistické frekvenční zpracování na dostatečně velkém souboru, které bylo provedeno ve spolupráci s CAS skupinou největších švýcarských farmaceutických koncernů a představovalo důležitou pomůcku pro práci se strukturními bázemi dat v počátcích jejich zavádění[37]. Fragmentační kódy, kterým mohou být přiřazeny různé váhy, představují pak také hlavní nástroj pro vytváření programů typu SIMILARITY SEARCH.

Je samozřejmé, že velmi důležitou roli hraje i počítačový hardware. Tak např. v CAS je pro práci s bází REGISTRY používán systém paralelního zpracování dotazů na 11 minipočítačích[38]. Podobná multiprocesorová architektura byla vyvinuta pro Beilstein Institut. Objektivní porovnávání jednotlivých vyhledávacích systémů je poměrně složité a závisí do značné míry na tom, jaký typ struktury je vyhledáván[39].

3.3. Reakční báze dat

Pod označením reakční báze dat jsou chápány strukturní báze dat s implementovanou a zdůrazněnou možností ptát se přímo na vzájemné přeměny zadaných struktur. Dotazovací program musí proto umožňovat definovat pro každou zadanou strukturu její roli, kterou v požadované přeměně hraje, tj. možnost označit, zda se jedná o výchozí látky nebo produkty. Topologická reprezentace v zásadě umožňuje, aby byla provedena projekce atomů jedné struktury do druhé (mapování struktury) s následnými informacemi o tom, na kterých atomech dochází k přeměnám (identifikace reakčních center), které vazby zanikají a které vznikají, popřípadě, pokud báze obsahuje stereochemické atributy, zda je či není zachována chiralita reakčního centra. Dále nás samozřejmě zajímají experimentální podmínky, použitá činidla nebo katalyzátory a rozpouštědla i kvantitativní vyjádření výsledků přeměn v podobě výtěžků. Reakční báze musí být kromě toho koncipována s ohledem na celou řadu možných aspektů, pro které budeme hledat informace. Kromě vyhledání přípravy určité látky, nás bude např. zajímat příprava celé skupiny sloučenin, způsob pro-

vedení konkrétní strukturní změny nebo záměny funkční skupiny, nebo budeme hledat způsob použití určitého činidla nebo katalyzátoru a v neposlední řadě pak se můžeme zajímat o reakční mechanismus. To vše činí z problematiky reakčních bází dat daleko obsáhlejší a komplikovanější problém, než jen možnost určit, co je výchozí látka a co produkt. Nicméně reakční bázi dat můžeme zadat konkrétní strukturu jako každé jiné strukturní bázi dat a očekávat odpověď, v jakých reakčních situacích se tato sloučenina vyskytuje. Je proto na místě se alespoň výčtem o reakčních bázích dat zmínit.

Tak především Beilstein Institut rozšířil svou strukturní bázi CrossFire[30] o možnost definovat co je výchozí látka a co produkt a zpřístupnil tak vyhledávání cca 10 miliónů reakcí, které jsou v tomto díle obsaženy. Tato rozšířená verze je označována jako CrossFire Reaction Plus a byla uvedena na trh v r. 1996.

Chemical Abstracts Service produkuje reakční bázi dat CASREACT[40], která zpracovává informace obsažené v sekci organické chemie v Chemical Abstracts od r. 1985 (patenty od r. 1991). Báze obsahuje cca 1,3 miliónů jednostupňových reakcí a více než 1,8 miliónů vícestupňových reakcí a každý týden je doplňována o 800 až 1.300 nových reakcí. Těsná vazba na strukturní bázi REGISTRY není překvapující, např. substrukturní vyhledávání jednotlivých reakčních partnerů je prováděno v této bázi a pak převáděno do báze CASREACT, kde jsou teprve přiřazeny odpovídající role výchozích látek nebo produktů. Pro nalezené odkazy na primární zdroje jsou opět k dispozici abstrakty v „hlavní" bázi Chemical Abstracts, CA File. muláře a

Celá řada reakčních bází dat je k dispozici pro vyhledávání v systému ISIS[41]. Kromě elektronické verze známého reakčního kompendia Theilheimer a jeho pokračování jako Journal of Synthetic Methods (báze REACCS-JSM), je to soubor syntetických reakcí ze 71 svazků Organic Synthesis (ORGSYN), obdobný soubor celkem osmi svazků Comprehensive Heterocyclic Chemistry (CHC), ale dnes pravděpodobně nejdůkladněji zpracovávaná báze dat ChemlnformRX, která se zaměřuje na její nové reakce nebo syntetické postupy, nové využití známých reakcí, nová činidla a jejich použití či jakékoliv jiné neobvyklé reakční cesty.

O další reakční bázi jsme se již zmínili v kap. 3.1.1. jako o produktu spolupráce bývalého SSSR a NDR v oblasti počítačové strukturní reprezentace a komercializované společnostíInfoChempodnázvemChernReactl0aChemReact41. Tyto zdroje jsou dodávány jako CD-ROM a mohou být provozovány na počítačích kategorie PC[32].

3.4. Formy přístupu a využívání strukturních bází dat

Vývoj rozsáhlých a pro nejširší chemickou veřejnost proto jedině zajímavých strukturních bází dat byl od počátku spojen s vysokými požadavky na výpočetní techniku, které mohly splňovat jen velké sálové počítače se vzdáleným přístupem. Proto stále nejširší nabídku strukturních bází nalezneme v databázových střediscích operujících na globální bázi. Patří sem např. Knight-Ridder (dříve DIALOG) a především STN International[42], které bylo z počátku založeno právě pro zpřístupnění databázových zdrojů CAS, včetně strukturních bází. Dnes nabízí toto středisko celkem 11 strukturních bází dat, vedle bází Chemical Abstracts REGISTRY, CASREACT, MARPAT a MARPAT-Preview, jsou to především báze BEILSTEIN a GMELIN, dále reakční báze ChemInformRX a ChemReact a báze společnosti DERWENT DJSMDS a DJSMONLINE, které jsou vlastně pokračováním Theiheimerova kompendia. Další báze LREGISTRY, LCASREACT, LMARPAT a LBEILSTEIN jsou cvičné („learning") a jsou zpřístupňovány za relativně malý poplatek.

Přístup do jednoho ze dvou počítačových center STN International v USA (Columbus) nebo v Německu[42] (Karlsruhe) je dnes především díky Internetu podstatně jednodušší a není již komplikován problémy s telekomunikačním připojením. Přístup je ovšem nutné administrativně vyřídit, protože každý vstup je placen. Stačí poslat např. e-mail na adresu help@cas.org se žádostí o zaslání forpo zaplacení celkem malého jednorázového vstupního poplatku a přidělení účtovacího čísla a přístupového hesla je možné začít pracovat. I když je možné pracovat s textovým terminálem (viz kap. 3.2.), vážnější zájemce o práci se strukturními bázemi dat pravděpodobně použije dnes komunikační program typu „front-end" STN Express, umožňující grafickou komunikaci i přípravu dotazů předem a mající řadu dalších, práci usnadňujících funkcí.

Je ovšem skutečností, že tato forma, kdy každé připojení ke strukturní bázi znamená spuštění počítadla jak doby připojení, tak i převzatých informací, není nejvhodnější a zřejmě způsobuje, že přes velmi lákavé možnosti získáváníinformacínejsoustrukturníbázedatvyužívány tak, jak by bylo možné očekávat. Je nutné upozornit, že vzhledem k vysokým nákladům na přípravu a udržování strukturních bází dat, byly ceny za jejich využívání nasazeny poměrně značně vysoko a přes různá systémová opatření umožňující předem ověřovat rozsah i správnost očekávaných výsledků za velmi malou cenu před spuštěním

vlastního prohledávání za plnou cenu, stojí např. substrukturní rešerše jedné otevřené struktury cca 100 USD. Převažujícími uživateli byly proto do nedávna především velké chemické a farmaceutické koncerny a vlastní práci s bází prováděli školení profesionální rešeršéři.

Zdá se, že zásadní průlom do této situace přinesl Beilstein Institut, který v r.1995 jednak radikálně změnil zpoplatňování báze dat BEILSTEIN v síti tím, že zrušil veškeré poplatky za dobu přístupu i za prohledávání a účtuje pouze skutečně převzaté informace a jednak nabídl tutéž bázi pod označením „CrossFire" k lokální instalaci na výpočetní technice uživatele. Taková instalace je realizována na základě roční subskripce bez jakýchkoliv dalších poplatků a její využívání je pak pro pracovníky subskribující instituce již bezplatné. I když náklady na nutnou výpočetní techniku jsou dosti vysoké, především vzhledem k nutnosti pořízení diskových polí s kapacitou cca 18 Gbyte, přece jen jsou již dnes v relativním dosahu i menších institucí nebo chemických fakult. Nejdůležitější výhodou této formy je ovšem skutečnost, že uživatelé v dané insitutici, tedy např. i studenti chemických fakult, mohou s touto bází pracovat bez jakéhokoliv stresu vyplývajícího z průběžného utrácení peněz.

CAS nastoupila podobnou cestu, ale zatím daleko opatrněji. V r. 1995 byl uveden na trh systém SciFinder, který v podobě uživatelsky velmi přátelského klientského programu umožňuje síťový přístup do všech bází Chemical Abstracts, tedy i strukturních, kde je samozřejmě možná plně grafická komunikace[43]. Tento systém rovněž předpokládá jednorázové roční předplatné na institucionální bázi, ovšem CAS zatím umožňuje takový neomezený přístup především podnikové sféře, a to za 65.000 USD ročně. Zpřístupnění pro univerzity a cena za takový přístup je ve stadiu příprav.

Podobná situace je v případě reakčních bází dat, kde je daleko výhodnější lokální instalace na základě ročního předplatného, na jejímž základě je pak možné bez stresu studovat různé syntetické strategie a plně tak využít potenciál dané báze. Typickým případem je již zmíněná báze ChemInformRX[41], pracující v prostředí systému ISIS, zatímco báze CHEMREACT je zatím přístupná pouze prostřednictvím sítě.

4. Perspektivy dalšího vývoje

Pravděpodobně nejdůležitějším aspektem ovlivňujícím současný i budoucí vývoj v oblasti počítačové reprezentace struktur chemických sloučenin, je prudký rozvoj počítačových technologií, kdy stoupající výkony široce dostupné kategorie osobních počítačů na straně jedné a klesající ceny výkonných serverů na straně druhé, umožňují přenést tyto aplikace z prostředí sálových počítačů a velkého chemického a farmaceutického průmyslu, do běžné chemické praxe a logicky pak i do výuky. Podobný vývoj mají i síťové komunikace, což vede k podstatnému zlepšení přístupu ke vzdáleným rozsáhlým strukturním bázím dat. Důsledkem je pak stále stoupající tlak na co největší zjednodušování formalit spojených s přístupem ke strukturním nebo reakčním bázím dat. Již uskutečněné kroky v tomto směru nasvědčují tomu, že v pravém slova smyslu rutinní využívání možností, které počítačové reprezentace struktur chemických sloučenin nabízejí, na sebe nenechá dlouho čekat.

LITERATURA

1. Bláha K., Ferles M., Staněk J.: *Nomenklatura organické chemie.* Academia, Praha 1985.
2. Hanč O., Hlavica B., Hummel V., Jelínek J.: *Chemická literatura a její využití v praxi.* SNTL, Praha 1961.
3. Kohnová Z.: Chem. Listy *69,* 248 (1975).
4. Urbánková I., Bočková H.: Chem. Listy *70,* 40 (1976).
5. Eakin D. R., Hyde E., Palmer G.: Pestic. Sci. *5,* 319 (1974).
6. Hanč O. a kol.: *Chemické a farmaceutické informační systémy,* str. 116. SNTL, Praha, Alfa, Bratislava 1982.
7. Rosa M. C: J. Pat. Off. Soc. *34,* 324 (1952).
8. Valance E. H.: J. Chem. Doc. *1,* 87 (1961).
9. Fugmann R., Braun W., Vaupel W.: Nachr. Dok. *14,* 179 (1963).
10. Rossler S., Koll A. G.: J. Chem. Doc. *10,* 128 (1970).
11. Simmons E.: J. Chem. Inf. Comput. Sci. *24,* 10 (1984).
12. Derwent Information Ltd, Derwent House, 14 Great Queen Street, London WC2B 5DF.
13. Index Guide 1987-1991, Appendix IV, CAS Columbus.
14. Registry File, Basic Name Segment Dictionary. STN International, June 1993.
15. Registry File, Dictionary Searching, STN International, June 1992.
16. Patterson A. M., Cappell L. T., Walker D. F.: *The Ring Index,* 2nd Ed. American Chemical Society, Washington 1960.
17. Barnard J. M., Jochum C. J., Welford S. M.: *Chemical structure information: interfaces, communication and standards* (Warr W. A., ed.), str. 76. ACS Symposium Series Nr.400. American Chemical Society, Washington 1989.

18. Weininger D.: J. Chem. Inf. Comput. Sci. *28,* 31 (1988).
19. Weininger D., Weininger A., Weininger J. L.: J. Chem. Inf. Comput. Sci. *29,* 97 (1989).
20. Morgan H. L.: J. Chem. Doc. 5, 2 (1965).
21. Garavelli J. S.: *Chemical structure Information: interfaces, communication and standards* (Warr W. A., ed.), str. 118. ACS Symposium Series Nr.400. American Chemical Society, Washington 1989.
22. Bebak H. et al.: J. Chem. Inf. Comput. Sci. *29,* 1 (1989).
23. Barnard J. M.: J. Chem. Inf. Comput. Sci. *30,* 81 (1990).
24. Petrarca A. E., Lynch M. F., Rush J. E.: J. Chem. Doc. 9,32(1969).
25. Wipke W. T., Dyott T. M.: J. Am. Chem. Soc. 96,4834 (1974).
26. Rusinko A. III., Skell J. M., Balducci R., McGarity C. M., Pearlman R. S.: CONCORD, a program for the rapid generation of high quality approximate 3-dimensional molecular struktures. University of Texas, Austin, and Tripos Associates, St.Louis, USA.
27. Rusinko A. III., Sheridan R. P., Nilakantan R., Haraki K. S., Bauman N., Venkataraghavan R.: J. Chem. Inf. Comput. Sci. *29,* 251 (1989).
28. Dittmar P. G., Stobaugh R. E., Watson C. E.: J. Chem. Inf. Comput. Sci. *16,* 111 (1976).
29. Using the CAS Registry File on STN, Student Manual (I, II, III), Instructor Package (I, II, III), 1996, STN International, c/o CAS, Columbus, Ohio 43202-0228 USA.
30. Lawson A. J., Swienty-Bush J.: *Proceedings of the 17th International Online Meeting,* Learned Information, Oxford 1993.
31. ACD, The Available Chemical Directory. Molecular Design MDL AG, Mühlebachweg 9, CH-4123 Allschwill 2, Switzerland.
32. InfoChem GmbH, Landsbergstrasse 408, D-81241 Miinchen, Germany.
33. Ray L. C, Kirsch R. A.: Science *126,* 814 (1957).
34. Balaban A. T.: J. Chem. Inf. Comput. Sci. *25,* 334 (1985).
35. *Adding Screens in Structure Searching.* Chemical Abstracts Service, Columbus, April 1983.
36. Wipke W. T., Krishnan S. K., Ouchi G. I.: J. Chem. Inf. Comput. Sci. *18,* 32 (1978).
37. Graf W., Kaindl H. K., Kniess H., Warszawski R.: J. Chem. Inf. Comput. Sci. *22,* 177(1982).
38. Farmer N. et al: *Chemical Structures. The International Language of Chemistry* (Warr W. A., ed.), str. 283. Springer Verlag, Berlin 1988.
39. Hicks M. G., Jochum C: J. Chem. Inf. Comput. Sci. *30,* 191(1990).
40. CASREACT, User Guide, STN International, c/o FIZ Karlsruhe, P.O.Box 2465, W-7500 Karlsruhe 1, Germany, February 1993.
41. The MDL Reaction Library Product Line, MDL Information Systems AG, Mühlebachweg 9, CH-4123 Allschwil 2, Switzerland.
42. STN International c/o FIZ Karlsruhe, P.O.Box 2465, D-76012 Karlsruhe, Germany;hlpdeskk@fiz-karlsruhe.de
43. Cain R., Schwall K.: CHEMTECH Aug. 1995, 8.

J. Šilhánek *(Department of Organic Technology, Institute of Chemical Technology, Prague):* Computer Representation of **Chemical Structures**

The review summarizes basic problems of the computer representation ofgraphic images of the structures of organic compounds. All principal methods, from linear notations (WLN) to topological representation, are covered. Main features of the methodology of usage are described on examples from the most important structure information, sources, REGISTRY of Chemical Abstracts and Beilstein's „CrossFire". Some other structure and reaction databases are also listed. The problems and methodology of access to such information sources in local-network environment or of connection to distant database vendors are discussed in details. It is stressed that so called „flat-fee" payment policy which gives an unlimited access for staff members of subscribing institution or students at universities, is the best way for the really efficient and productive exploitation of structure and reaction databases.