辞書検索で構造検索と同等の回答を得られるか？

2008年11月18日

PLASDOCオンライン研究会

STNグループ

STNグループ・メンバー（2007年度）

▶ MRCテクノリサーチ（株）	杉山　善匡
▶ 旭化成（株）	森　善子
▶ （株）住化技術情報センター	橋本　武彦（**リーダー**）
	本間　文子
▶ ダイセル化学工業（株）	周　興喜
	山崎　登和子
▶ 電気化学工業（株）	渡辺　喜代美
▶ 東ソー（株）	青野　祥博

（五十音順）

今回の内容

- PLASDOCオンライン研究会の紹介
  - 沿革・活動内容、等
- 辞書検索で構造検索と同等の回答を得られるか？
  - 辞書検索（環データ）と構造検索の比較検討
  - 仮説とその検証
- 補足説明
  - 辞書検索と構造検索の使い分け
- おわりに

PLASDOCオンライン研究会の紹介

沿　革

- 1980年、主にポリマーに関する情報検索についての情報収集・情報交換・研鑽を目的として、わが国の化学メーカー（当初18社）の参加のもと、発足
  - *日本PLASDOC協議会とは別の組織です*
- 会員全員が、参加企業間の垣根を越えてノウハウを共有し合い、率直に意見交換
- 会員が希望する検討テーマについて、グループ毎にワーキングを実施し、毎年総会にて発表

総会の様子（2007年6月於軽井沢）

最近のワーキング内容

- 2006年度
  - 『物・製法・用途別無効化資料の調査手法①』～化合物・組成物、製法、用途～
  - 『アジア特許調査の検討（現地出願人の出願状況等）』
- 2007年度
  - **『REGISTRYファイルにおける辞書検索の有効性』**
  - 『物・製法・用途別無効化資料の調査方法②』～塗料・接着剤関係の用途～
  - 『商用データベース（特に中国・韓国特許データベース）の翻訳機能の比較』

最近の臨時勉強会

- データベース・ベンダー等による、当研究会のための臨時勉強会を随時開催
  - 2006年度
    - STN AnaVist の機能および事例紹介
    - パテントマップEXの機能および事例紹介
  - 2007年度
    - クラスター・環データについての説明
    - 構造検索におけるスクリーンの利用についての説明

現在の会員企業

JSR(株),　　MRCテクノリサーチ(株),
旭化成(株) ,　　住友化学(株),
ダイセル化学工業(株),　チッソ(株),
東ソー(株),　　日本ゼオン(株),
(株)パトロ・インフォメーション,
(株)ブリヂストン

（五十音順）

協賛データベース・ベンダー

(株)Crystal Technology,
(株)WIPS,　　**(社)化学情報協会**,
(株)クロスランゲージ,　　(株)ジー・サーチ,
中央光学出版(株) ,　　(株)ティージェイエス,
トムソンコーポレーション(株),
日本アイアール(株),　　日本技術貿易(株) ,
日本パテントデータサービス(株),
(株)パトリス

（五十音順）

連絡先等

- PLASDOCオンライン研究会HP（Yahooグループ）
  - http://groups.yahoo.co.jp/group/plasdoc_online/
- PLASDOCオンライン研究会HP（日本PLASDOC協議会HP内）
  - http://plasdoc.sakura.ne.jp/online.html
- PLASDOCオンライン研究会 事務局
  - dbsstudy@yahoo.co.jp

辞書検索で構造検索と同等の回答を得られるか？

辞書検索（環データ）と
構造検索の比較検討

検討内容

- 化合物の特徴別に辞書検索と構造検索を行った結果を比較検討

- 化合物の特徴
  - 単環ー炭素環化合物
  - 縮合ー炭素環化合物
  - 単環ー複素環化合物
  - 縮合ー複素環化合物

検討内容

- 構造検索
  - ① 部分構造検索（縮環なし）*
  - ② サンプル検索

    * 環系識別子/RIDの検索では、その識別子で特定される環系に**さらに別の環が縮合した環系を持つ化合物の検索ができない**ため
- 辞書検索
  - ① **環データ**（環系識別子/RID，環系の存在数/RID.CNT等）で検索
  - ② **特定元素数**で絞込み

検討内容

- 辞書検索の可否の判断基準

① 検索結果が500件以内、または

② 構造検索の件数（予測値の最大値）と比較して、3倍以内

➔ ***辞書検索可と判断***

検討目的

REGISTRYファイルで化学物質を検索する場合

↓

①構造検索　or　②辞書検索

↓

構造検索は容易に検索できる反面、検索料金が高額

↓

辞書検索で構造検索と同等の回答を得られるか？

辞書検索で構造検索と同等の回答を得られるか？

辞書検索上の注意点

辞書検索上の注意点

- **分子式**や**環データ**
  - ルールに則った明確なデータが収録されているので、検索キーとして安心して利用できる
- **完全名称**や**名称セグメント**
  - 収録間もない物質については、(CA 索引名を含む) **物質名称が収録されていない**こともあり得るため、完全名称や名称セグメントでの検索は常に検索漏れの可能性があることに注意が必要

辞書検索上の注意点

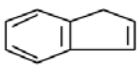
- 環データの調べ方
  - ① 調べたい環系を持つ化合物の名称やCAS登録番号等で検索して
  - ② REGISTRYファイルで構造検索のサンプル検索（またはLREGISTRYファイルで構造検索）をして
  - ③ 環系の元素式（/EA）または元素配列（/ES）で検索して

➔環データを表示する（例えば、**=> D IDE RSD**）

辞書検索上の注意点

- 環系識別子について
  - 環系識別子/RIDは、「骨格」.「元素の位置」.「**結合次数**」を表す3つの数字で構成されている（例：333.70.**45**）
  - 右端の「**結合次数**」は、二重結合、三重結合のみならず、「ノーマライズド結合」をも識別している

例 ：RN 324-85-6



Ring System Data

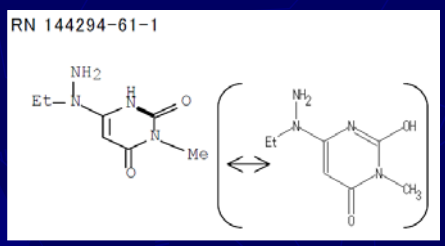
Elemental Analysis EA	Elemental Sequence ES	Size of the Rings SZ	Ring System Formula RF	Ring Identifier RID	RID Occurrence Count
C5-C6	C5-C6	5-6	C9	333.70.45	1

辞書検索上の注意点

- ノーマライズド結合について
  - **ノーマライズド結合**とは、作図上、一見異なる物質に見えるが、実際は同じである物質をシステムに「同じ物質である」と、できる限り認識させるために開発された結合タイプである
  - 作図上、見た目は二重結合でも実はノーマライズド結合になる場合には、**環系識別子の最後の部分**が変化する
  - また、環系に置換基が付くことで環系上の結合状態がノーマライズド結合になる環系構造も存在する

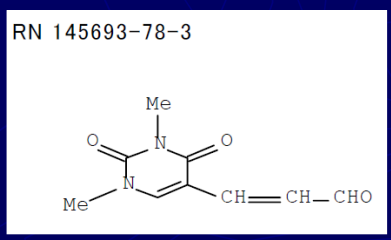
辞書検索上の注意点

- ノーマライズド結合について



46.195.**20**/RID

ノーマライズド結合の数：1



46.195.**18**/RID

ノーマライズド結合の数：0

辞書検索上の注意点

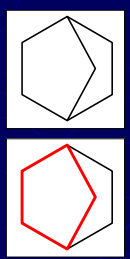
- ノーマライズド結合について
  - このような構造を含む物質を検索する場合、環系識別子の最後の部分（**結合次数**）は変化するので、**左側の2つの部分のみ**で検索する
    （例：=>S **46.195**/RIDで検索する）

➔詳しくは、「REGISTRY ファイル - 検索テクニック」，（社）化学情報協会編，2008年2月，p.38～49参照
http://www.jaici.or.jp/stn/ref-registry.pdf

辞書検索上の注意点

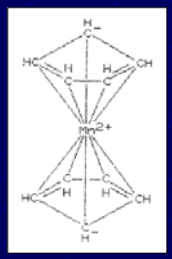
- 最小環の最小集合（SSSR）について
  - REGISTRYファイルに含まれる環データは、ある環系に含まれる全ての環ではなく、それに含まれる**最小環の集合であるSSSR**を元に生成される

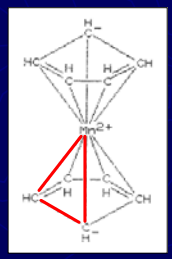
  - 例1　ノルボルナン
    - この場合のSSSRは、シクロヘキサン環の上端・下端および左右それぞれの側の炭素原子と架橋炭素で作られる**C5の5員環の2個**である



辞書検索上の注意点

- 最小環の最小集合（SSSR）について
  - 例2　マンガノセン
    - この物質に対するSSSRは、マンガンイオンとそれぞれのシクロペンタジエン環上の炭素－炭素結合で作られる**C2Mnの3員環10個**の集合である



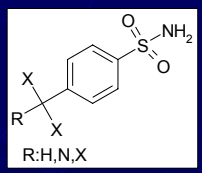


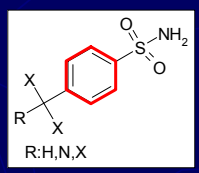
辞書検索で構造検索と同等の回答を得られるか？

検討結果の詳細

単環ー炭素環化合物

- 検討例1
  - 構造検索：**112件**





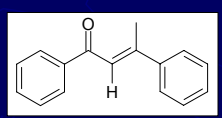
  - 辞書検索

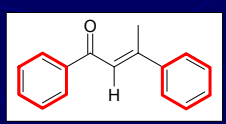
① **46.150.18**/RID：19,522,549件

② ① AND X>=2(P)S>=1(P)N>=1(P)O>=2(P)C>=7：**866,736件**

単環ー炭素環化合物

- 検討例2
  - 構造検索：**720件**





  - 辞書検索

① **RID.CNT>=2** (T) 46.150.18/RID：10,134,058件

② ① AND C>=16：9,652,553件

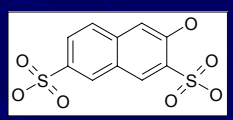
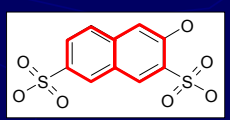
③ ② AND O>=1：**8,921,992件**

単環ー炭素環化合物

- 検討結果のまとめ
  - 辞書検索では置換位置を指定できないため、**環データ**による限定だけでは、相当数のノイズが含まれる
    - ➔構造検索と比較して桁違いの結果が出てしまう
  - 炭素数で限定しても、ヒット数はほとんど減らない。
  - （環外に）NやOを1～2個含んでいても、**窒素数**や**酸素数**の限定ではヒット数はあまり減らない

縮合ー炭素環化合物

- 検討例1
  - 構造検索：**1,664件**

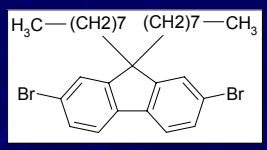

  - 辞書検索

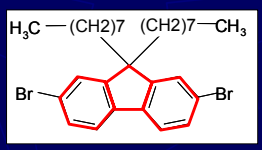
① **591.49.57/RID**：816,430件

② ① AND S>=2 (P) O>=7：**76,555件**

縮合ー炭素環化合物

- 検討例2
  - 構造検索：**772件**





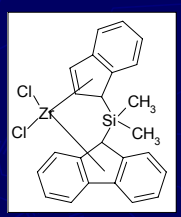
  - 辞書検索

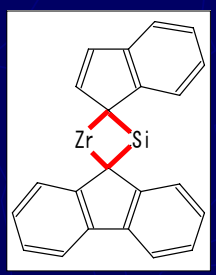
① **1839.6.36/RID**：123,730件

② ① AND BR>=2 (P) C>=29 (P) H>=34：**1,648件**

縮合ー炭素環化合物

- 検討例3
  - 構造検索：**51件**





  - 辞書検索

① **SiCZrC/ESS**：3,056件

② ① AND ESS.CNT>=3 (T) C6/ESS：427件

③ ② AND Si>=1 AND Zr>=1 AND Cl>=2：**269件**

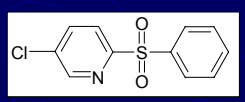
縮合ー炭素環化合物

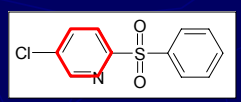
- 検討結果のまとめ
  - 置換基の指定がなく、かつ環の孤立を指定する場合、構造検索と辞書検索（**環データ**）で同等の検索ができる

    ➔検索語料の安い辞書検索を行うほうがコスト的に有利
  - （環外に）NやOを１～２個含んでいても、**窒素数**や**酸素数**の限定ではヒット数はあまり減らない
  - （環外に）Clを含んでいると、**塩素数**で限定すればヒット数を減らすことができる

単環ー複素環化合物

- 検討例1
  - 構造検索：**22件**





  - 辞書検索

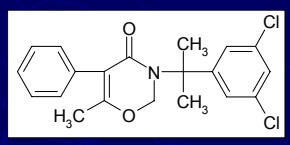
① **46.156.30/RID** AND 46.150.18/RID：1,054,143件

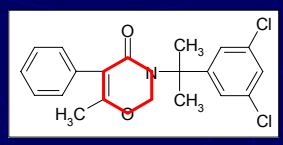
② ② ① AND O>=2 AND CL>=1：203,170件

③ ③ ② AND S>=1：**64,918件**

単環ー複素環化合物

- 検討例2
  - 構造検索：**200件**



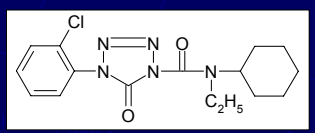


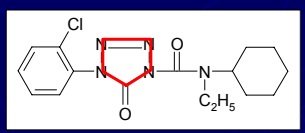
  - 辞書検索

① **46.220.7/RID**：1,575件
② ① (P)>=2 46.150.18/RID：726件
③ ② AND CL>=2：**159件**

単環ー複素環化合物

- 検討例3
  - 構造検索：**163件**



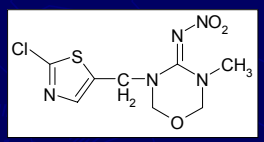


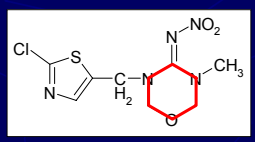
  - 辞書検索

① 46.150.18/RID：20,395,863件
② ① (P) **16.525.2/RID**：4,846件
③ ② (P) 46.150.1/RID：390件
④ ③ AND CL>=1 (P) O>=2：**149件**

単環ー複素環化合物

- 検討例4
  - 構造検索：**504件**





  - 辞書検索

① 16.299.11/RID：291,785件

② ① (P) **46.493.1/RID**：336件

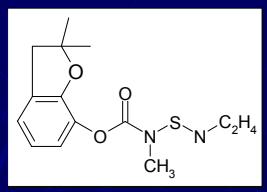
③ ② AND N>=5 (P) CL>=1 (P) O>=3：**301件**

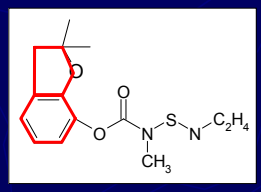
単環ー複素環化合物

- 検討結果のまとめ
  - 環構造に特徴がある場合、辞書検索（**環データ**）で十分な結果を得ることができる
  - **環データ**による検索の時点で**数千件レベル**になっていないと、**特定元素数**等による限定で数百件レベルまで絞りこむことは困難
  - （環外に）塩素や硫黄を含んでいる場合、**塩素数**や**硫黄数**で限定すればヒット数を減らすことができる

縮合ー複素環化合物

- 検討例1
  - 構造検索：**234件**





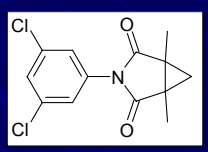
  - 辞書検索

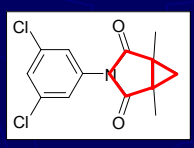
① **333.200.31/RID**：61,041件

② ① AND O>=3 (P) S>=1 (P) N>=2：**6,481件**

縮合ー複素環化合物

- 検討例2
  - 構造検索：**329件**





  - 辞書検索

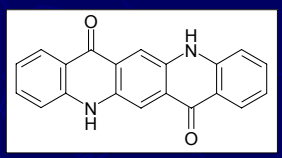
① **42.4.1/RID**：13,888件

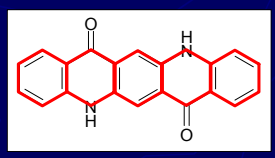
② ① (P) 46.150.18/RID：8,472件

③ ① AND O>=2 (P) CL>=2：**331件**

縮合ー複素環化合物

- 検討例3
  - 構造検索：**685件**





  - 辞書検索

① **8481.17.4/RID**：1,151件

② ① AND O>=2 (P) N>=2：**1,124件**

縮合ー複素環化合物

- 検討結果のまとめ
  - 鎖の部分に特徴がある物質の場合、辞書検索では置換位置を指定できないため、相当数のノイズが含まれてしまい、絞込みは困難
  - 環構造に特徴がある場合、辞書検索（**環データ**）で十分な結果を得ることができる

辞書検索で構造検索と同等の回答を得られるか？

仮説とその検証

まとめと仮説

- 環の構造ではなく、**鎖の構造に特徴**がある場合は、辞書検索（環データ）による絞込みは**困難**
- 環系識別子/RIDのヒット数が**1万件レベル**（以上）の環しか含まない場合、辞書検索による絞込みは**困難**

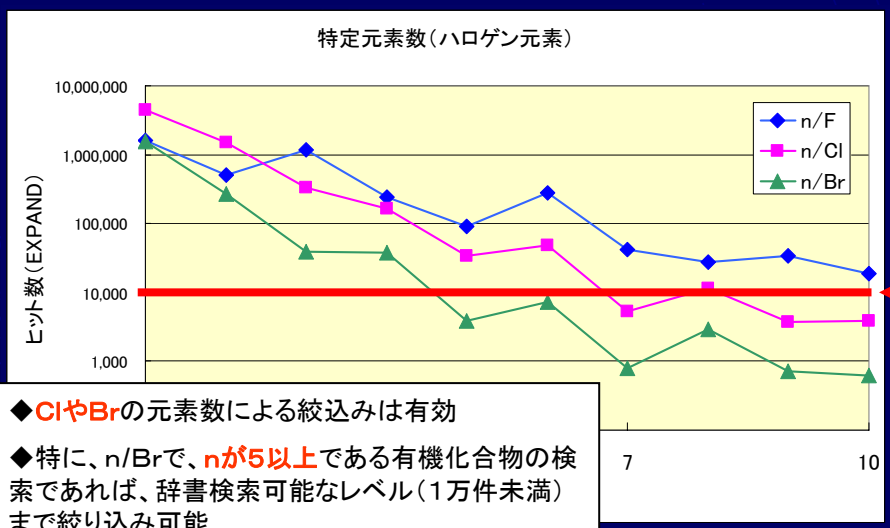
↓

- **ハロゲン元素**、**ヘテロ元素**を多数含む場合、辞書検索（特定元素数）でもある程度の絞込みが**可能？**
- **単環の大きさ**が大きい場合や、**縮合環に含まれる環数**が多い場合、その環データによる絞込みが**有効？**

検証①

- ハロゲン元素数による絞込み
  - F、Cl、Brについて、特定元素数毎のヒット数を調査
- ヘテロ元素数による絞込み
  - N、O、Sについて、特定元素数毎のヒット数を調査

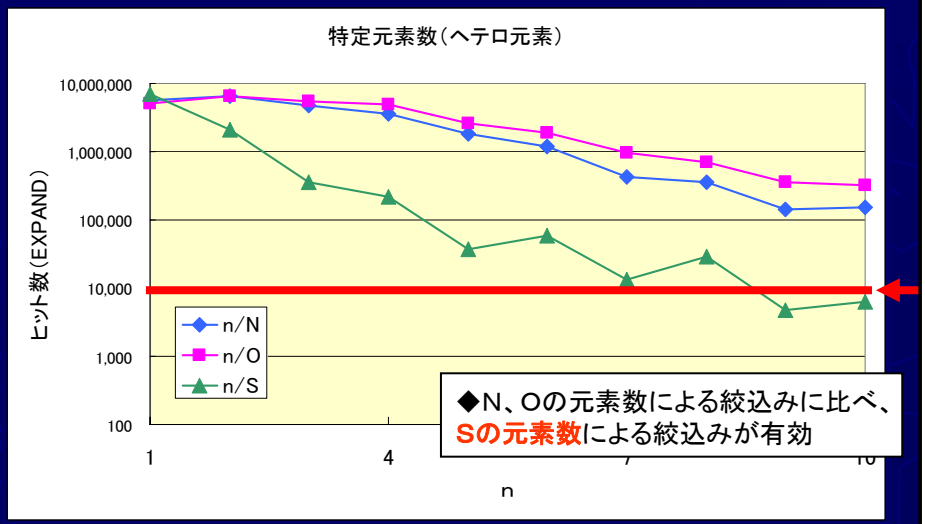
ハロゲン元素数による絞込み



◆ClやBrの元素数による絞込みは有効

◆特に、n/Brで、nが5以上である有機化合物の検索であれば、辞書検索可能なレベル（1万件未満）まで絞り込み可能

ヘテロ元素数による絞込み

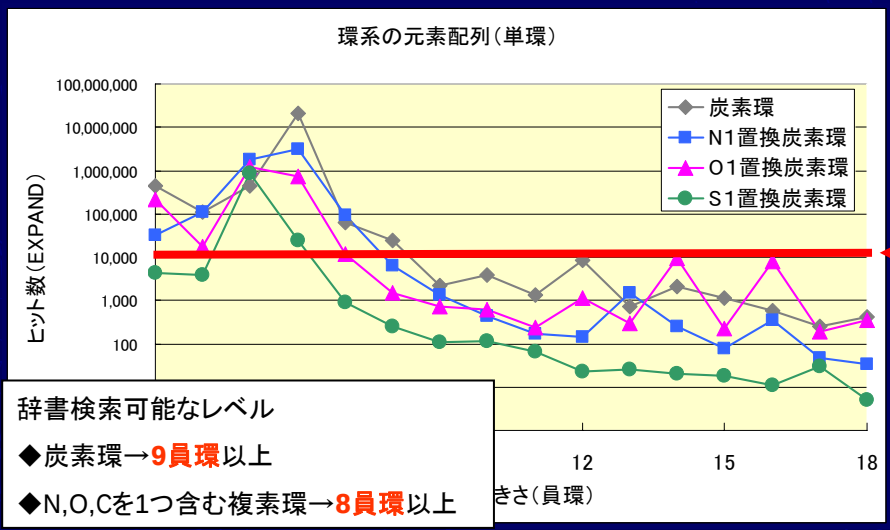


検証②

- 単環の大きさ
  - 以下に示す単環の大きさ毎に、元素配列（/ES）のヒット数を調査

① 炭素環
② 窒素を1つ含む炭素環
③ 酸素を1つ含む炭素環
④ 硫黄を1つ含む炭素環

単環の大きさ



辞書検索可能なレベル

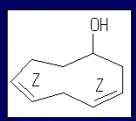
◆炭素環→**9員環**以上

◆N,O,Cを1つ含む複素環→**8員環**以上

検証結果②

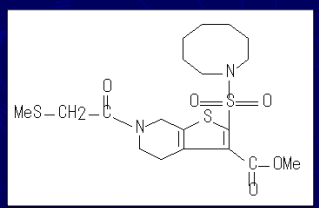
- 単環の大きさ
  - 辞書検索可能な例

① 炭素環 ➔ **9員環**以上



369.14.8/RID

② 窒素を1つ含む複素環 ➔ **8員環**以上

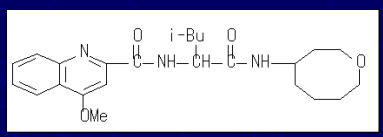


209.65.1/RID

検証結果②

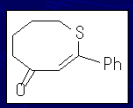
- 単環の大きさ
  - 辞書検索可能な例

③ 酸素を1つ含む複素環 ➔ **8員環**以上



209.66.1/RID

④ 硫黄を1つ含む複素環 ➔ **8員環**以上



209.68.6/RID

検証③

- 縮合環に含まれる環の数
  - 以下に示す縮合環に含まれる環数毎に、元素配列(/ES)のヒット数を調査

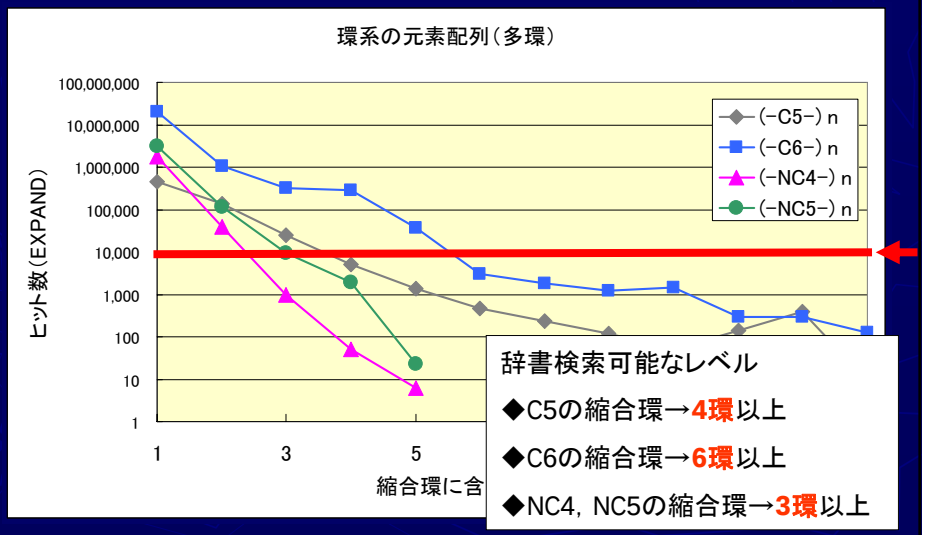
① 炭素5員環[(-C5-)n]

② 炭素6員環[(-C6-)n]

③ 窒素:1、炭素:4含む5員環[(-NC4-)n]

④ 窒素:1、炭素:5含む6員環[(-NC5-)n]

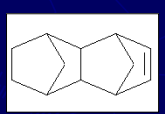
縮合環に含まれる環の数



検証結果③

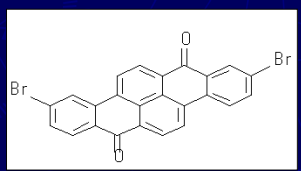
- 縮合環に含まれる環の数
  - 辞書検索可能な例

① 炭素5員環からなる縮合環 ➔ **4環**以上



1284.1.2/RID

② 炭素6員環からなる縮合環 ➔ **6環**以上

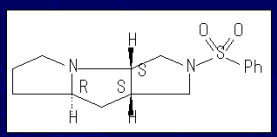


9412.1.3/RID

検証結果③

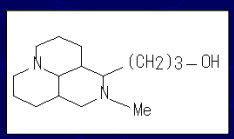
- 縮合環に含まれる環の数
  - 辞書検索可能な例

③ 窒素:1、炭素:4含む5員環からなる縮合環 ➔ **3環**以上



890.124.3/RID

④ 窒素:1、炭素:5含む6員環からなる縮合環 ➔ 3環以上



1784.113.1/RID

補足説明

辞書検索と構造検索の使い分け

辞書検索と構造検索の使い分け

- 料金面での使い分け
  ➔以下の検索語数を越える場合、辞書検索の方が高くなる

完全一致検索(EXA)	7,790 円 ➔	検索語**11**語
ファミリー検索(FAM)	9,060 円 ➔	検索語**13**語
部分構造検索(SSS)	23,200 円 ➔	検索語**34**語

辞書検索と構造検索の使い分け

- 完全一致検索する場合
  - 事前に**分子式 (/MF)** で検索できるかを、接続時間料金が無料のZREGISTRYファイルでEXPANDを用いて確認する
  - 該当件数が少なければ REGISTRY ファイルで分子式検索し、SCAN 表示形式で全件表示して確認した方が経済的かつ効率的

辞書検索と構造検索の使い分け

- 部分構造検索するようなバリエーションのある検索条件の場合
  - 特徴的な環構造を有し、かつ環縮合が不要なときは、**環データ**による検索で検索を完了し得る
  - 環を含まない物質を検索するときは、**RSD/FAをNOT演算**することで、目的物質をある程度までは絞り込みできる

辞書検索と構造検索の使い分け

- 部分構造検索で INCOMPLETE となってしまった場合
  - これを回避する方法として、辞書検索の検索結果に対して**サブセット検索**すると効果的
  - ただし、辞書検索の検索結果に対するサブセット検索においては、構造検索の場合のサブセット検索料金は適用されないため、注意が必要

➔詳しくは、「化学物質 III」,（社）化学情報協会編, 2006年8月, p.34～36参照
http://www.jaici.or.jp/stn/ref-substance.pdf

辞書検索と構造検索の使い分け

- IDS登録物質について
  - IDS 登録物質は収録される構造データがその物質を正しく表現しているわけではないため、**構造検索のみでは検索もれ**となる場合があり得る
  - しかし、IDS登録物質であっても分子式は正しいため、**分子式関連の辞書検索**でヒットできる場合があるので必要に応じて確認する

➔詳しくは、「化学物質 III」,（社）化学情報協会編, 2006年8月, p.13, 90～91参照
http://www.jaici.or.jp/stn/ref-substance.pdf

辞書検索と構造検索の使い分け

- 検索結果を確認する場合の注意点

① REGISTRYファイルで、構造検索または辞書検索を行う
② ヒットした全件（全化合物）について、化学構造式を目視でチェックし、重要化合物をピックアップする
③ ピックアップした重要化合物について、（特許）文献が存在するか、チェックする
  - ②で、**SCAN形式（無料）のみ**表示し検索を終了
    ➔③において、REGISTRYファイルで**化合物名等を再検索**してから、CAファイルにクロスオーバー検索する必要がある
    ➔検索費用が**むしろ高額**になる場合がある

おわりに

今回の検討を進めるにあたり、温かい御指導を賜わり、また、大変有益な情報をご提供下さいました、**社団法人化学情報協会**の皆様に心から感謝申し上げます

ご清聴ありがとうございました

あなたも、当研究会に参加してみませんか？！

dbsstudy@yahoo.co.jp