データベース構築ソフトウェア設計書

2018 年 2 月 5 日

— 目次 —

１．化合物ライブラリ構築ソフトウェア開発

本設計書では、2 次元 sdf ファイルから物性値を付与した 3 次元 mol2 ファイルを作成する、化合物ライブラリ構築ソフトウェアについて示す。

各章の構成は以下の通りである。

２章. 化合物データベース構築ソフトウェアの処理フロー
３章. 化合物データベース構築ソフトウェアの構成
４章. 各種外部プログラム

２．化合物データベース構築ソフトウェアの処理フロー

化合物データベース構築のパイプラインは、以下の 4 工程で処理を行えるものとする。本設計書では、下記、①、②について記載する。

パイプライン 4 工程：

①2 次元構造の 3 次元化処理

2 次元 sdf から 3 次元 mol2 ファイルを作成する。その際に、光学、配座異性体を生成するとともに、各種物性値を計算する。

②LigandBox ID の付与

データセットのバージョン等を識別する為に、データソース名、データ作成年度等の文字列を付与した LigandBox ID の付与を行う。

③200 万件データセット作成の為のフィルタリング、クラスタリング処理

医薬品として不適切な構造を含む分子のフィルタリング処理を行う。また、数百万～数千万件データから、類似の構造をクラスタリングした上で、クラスタリングされた代表分子を取得して、200 万件データを作成する。200 万件データセットは 1 ディレクトリ 10000 件のデータを保存し、200 ディレクトリを作成する。

④データのランダム化

③で作成されたデータは①の入力ファイルの分子データの順番に依存してディレクトリに配置されるものである。どのディレクトリのデータを使用するとしても、同程度の分子量分布となっている必要がある。そこで、データのランダム化を行い、ディレクトリ毎のデータが分子量分布に偏りが無いように処理を行う。

2-1. 処理フローについて

化合物データベース構築ソフトウェアでは、以下の 3 ステップで処理を行う。
なお、LogP、LogS 計算は、ADMEWORKS/CmdPredictor を用いて行う。

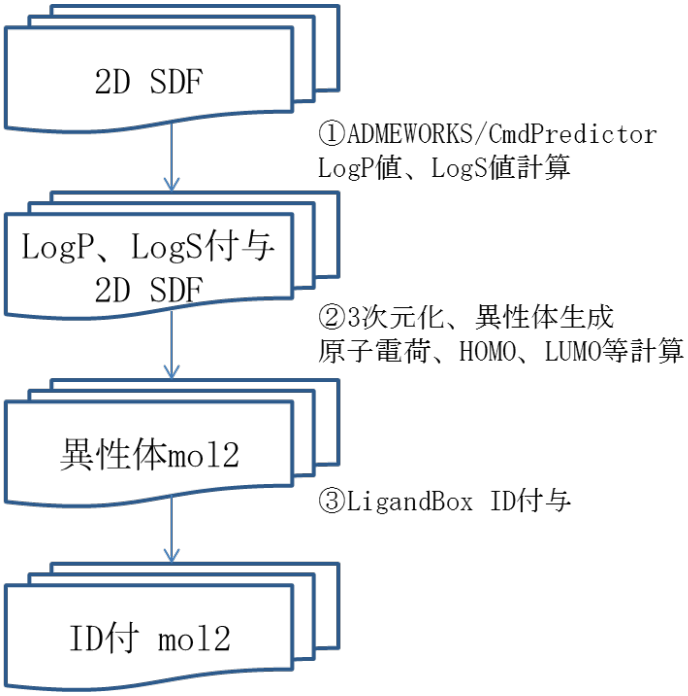


図. 化合物データベース構築ソフトウェア処理手順

３．化合物データベース作成ツールの構成について

化合物データベース作成ツールは、①2 次元構造の 3 次元化処理ツール(create3d)、②LigandBox ID 付与ツール、③フィルタリングツール、⑤クラスタリングツールからなる。本設計書では化合物データベース作成ツールのうち、3 次元化、異性体生成処理(create3d)ツールと、LigandBox ID 付与ツール(setLigandBoxID)について示す。

①2 次元構造の 3 次元化処理ツール(create3d)

2 次元構造の 3 次元化処理、物性値等計算を行う。

②LigandBox ID 付与ツール(setLigandBoxId)

複数のマルチ mol2 を通してユニークな LigandBox ID を付与する。

③フィルタリングツール

医薬品として不適切な構造を持つ分子を除外するフィルタリング処理を行う。

④クラスタリングツール

大量の化合物データのクラスタリングを行い、類似度の低い 200 万データを取得する。

3-1. 2 次元構造の 3 次元化処理(create3d)

従来手法では、LogS 計算は 3 次元化処理とは別ステップで処理を行っていたが、処理手法を変更する事により、本処理内で計算を行う様に変更する。

また、本ツール内で複数の外部プログラムを呼び出し、処理を行っていたが、統合できる部分については、統合し、Hgene で処理する様に変更した。

2 次元 sdf ファイルを入力して、ベース分子の 3 次元化を行った上で、各分子について光学・幾何異性体生成を行うものとする。

プログラム構成：

```
create3d：メインプログラム、2 次元 sdf から 3 次元化を行い、異性体を含むマルチ mol2
    |   ファイルを作成する。
    +-------- split_mols：sdf ファイルをシングル mol ファイルに分割する。
    +-------- ADMEWorks/CMDPredictor：LogS、LogP 計算を行う。（オプション）
    +-------- Hgene：水素付加、簡易 3 次元化を行う。併せて物性値を計算する。
    +-------- tplgeneL：トポロジーファイルを生成する。
    +-------- cosgene_multiPDB：異性体を含む構造を一括でエネルギー最小化
    |                           を行う。
    +-------- confgeneC：光学・幾何異性体の生成を行う。
```

処理手順

①マルチ mol2 ファイルの分割(split_mols)

マルチ mol2 ファイルをシングル mol2 ファイルに分割する。

②LogS、LogP 計算(ADMEWorks/CMDPredictor 処理)

LogS、LosP 値の計算を行う。

③水素付加、簡易 3 次元化、mol2 ファイルフォーマット変換(Hgene 処理)

対象の 1 分子について Hgene を用いて水素付加、簡易 3 次元化、mol2 ファイルフォーマット変換を行う。また、各種物性値の計算、出力を行う。

④トポロジーファイル生成(tplgeneL 処理)

⑤エネルギー最小化(cosgene 処理)

③で作成したベース分子について cosgene を使用してエネルギー最小化を行う。

⑥MOPAC 計算(Hgene 処理 2)

⑤の構造に対して MOPAC AM1 計算を行う。

⑥光学・幾何異性体生成処理(confgeneC 処理)

confgeneC を使用して光学・幾何異性体生成を行う。異性体生成数は 3 とする。

⑦異性体のエネルギー最小化(cosgene 処理)

⑥で作成した座標と④で作成したトポロジーファイルを使用してエネルギー最小化処理を行う。

⑧マルチ mol2 ファイルの出力

物性値付き、異性体を含む分子のマルチ mol2 ファイルを出力する。

出力情報：

入力分子に対応する 3 次元 mol2 データ、及び、その光学・幾何異性体 mol2 データをマルチ mol2 形式で出力する。

3-2. LigandBox ID 付与ツール(setLigandBoxId)

3-1 で作成されたマルチ mol2 ファイルは、LigandBox ID が付与されていない。そこで、複数のマルチ mol2 ファイルの各分子について､ユニークな ID を付与する機能を作成する。ID は、mol2 ファイルのコメント欄に、LIGANDBOX_ID = ZZZZ-XXXXXXXX-YY の形式で出力する。なお、ZZZZ は任意の文字列(主に、そのデータのデータソースや作成年度などを区別する為に付与する)、XXXXXXXX は 8 桁の数字、YY は 2 桁の数字(異性体を区別する)で出力する。

入力情報：

・LigandBox ID を付与するマルチ mol2 ファイルのリストファイル

・LigandBox ID の先頭に付与される任意の文字列

出力情報：

・LigandBox ID が付与されたマルチ mol2 ファイル

3-3. フィルタリングツール

医薬品として不適切な構造を含む分子を除去するプログラムを作成する。本ツールは、医薬品として不適切な構造を含む分子一覧を出力するプログラムと、この一覧ファイルを使用して分子を除去したマルチ mol2 ファイルを出力するプログラムの 2 つから構成されるものとする。

3-3-1. 分子除去条件

分子除去条件は以下とする。

・分子内の sp3 炭素比率が全体の 80%を超える場合

・ヘテロ原子数が 1 以下の場合

・2 つのヘテロ原子が二重結合で結合している場合

・医薬品として不適切とされる構造を含む場合

以下の構造を不適切構造とする。

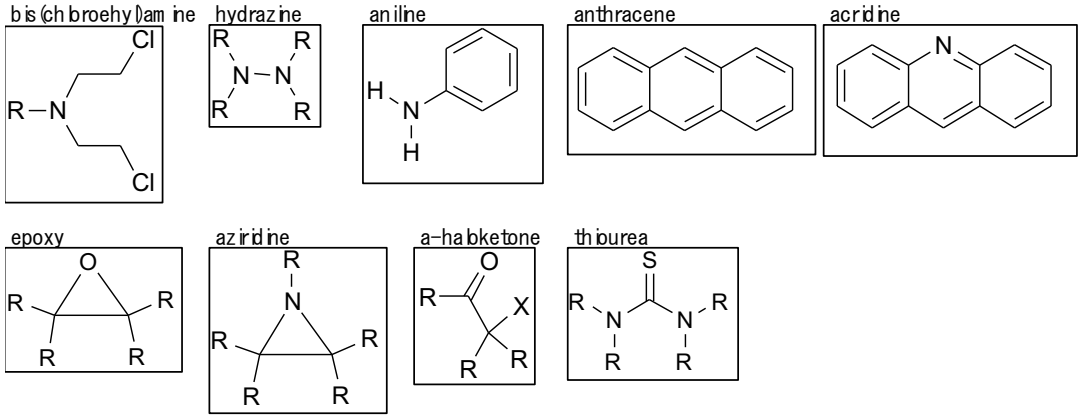


図. 除去構造１：毒性が期待されるもの

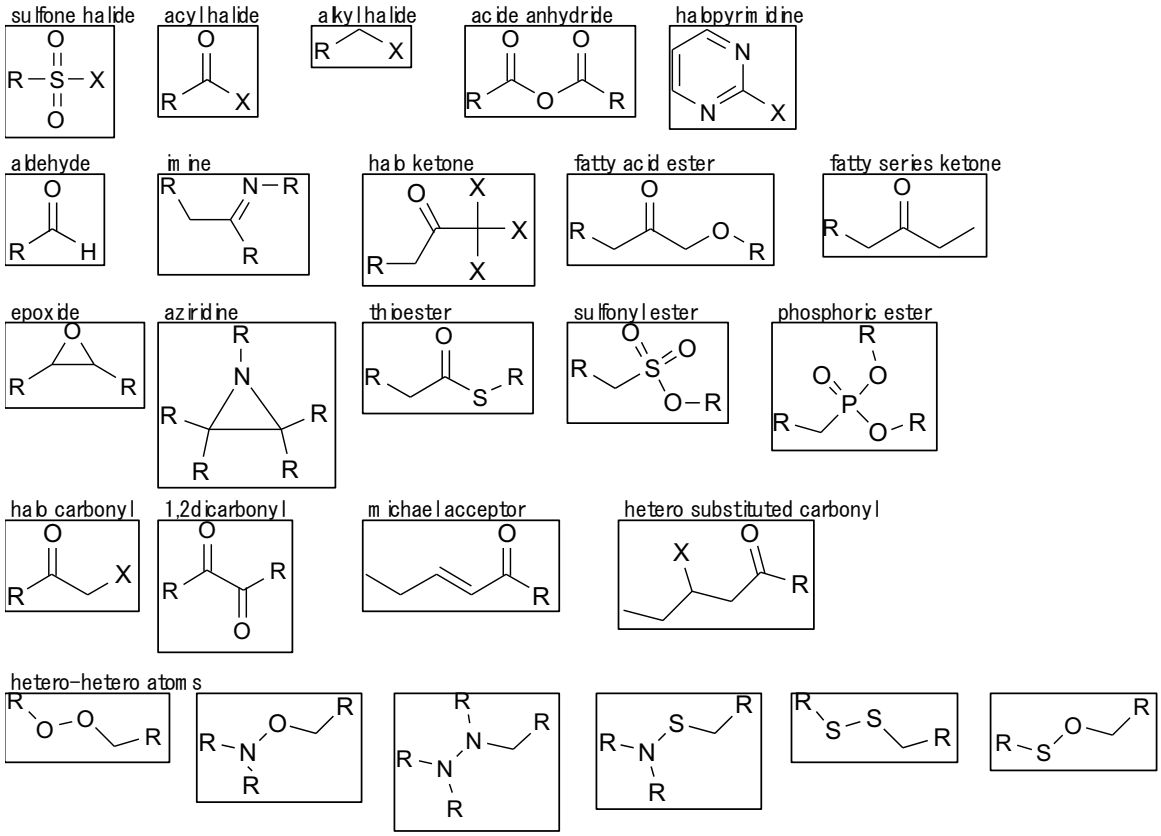


図. 除去構造２：False positive が生じやすいとされるもの

3-3-2. DB 設計

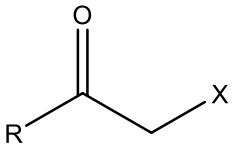
上記 3-3-1 で示した構造は、マルチ mol2 ファイルの形式で登録する。登録ルールは以下の通りとする。

・任意の炭素の原子タイプは R で示す。
・ハロゲンの原子タイプは X で示す。
・その他の原子タイプ、結合次数のルールは通常の Sybyl mol2 ファイルフォーマットに従う。

以下に、DB ファイルの例として、halo carbonyl の例を示す。

```
@<TRIPOS>MOLECULE
a_halo_carbonyl
 5 4 0 0 0
SMALL
NO_CHARGES

@<TRIPOS>ATOM
      1 C1          -1.0717    -0.6187     0.0000 R       1  <1>          0.0000
      2 C2          -0.3572    -0.2062     0.0000 C.2     1  <1>          0.0000
      3 C3           0.3572    -0.6187     0.0000 C.3     1  <1>          0.0000
      4 X            1.0717    -0.2062     0.0000 X       1  <1>          0.0000
      5 O1          -0.3572     0.6187     0.0000 O.2     1  <1>          0.0000
@<TRIPOS>BOND
     1     1     2    1
     2     2     3    1
     3     3     4    1
     4     2     5    2
```



図．ハロカルボニルの DB 記載内容(左)と構造(右)

3-3-3．不適切構造を含む分子リスト

不適切な構造を含む分子リストの出力項目は以下の通りとする。

・number　　：入力マルチ mol2 の分子の通し番号
・finename　：分子名(ID)
・supplier　：サプライア名
・idnumber　：サプライア ID
・info　　：フィルタリング結果

ステータスは、以下の通りである。

MANY_SP3_CARBON　　：sp3 炭素の割合が 80%を超える
ZERO_OR_ONE_HETEROATOM　：ヘテロ原子(N, O, P, S)が 1 個以下である。
NO_CIRC_STRUCTURE ：環構造を持っていない。
CONTAIN_SPECIFIC_STRUCTURE：特定の不適切な分子構造を持っている。
NONE：不適切な分子構造を持っていない。

・mathedmol　：フィルタされた場合、どのような構造でフィルタされたかを示す。

構造フィルタは以下の通りとする。

a_halo_carbonyl	a_halo_ketone	acid_anhydride	acridine
acyl_halide	aldehyde	aniline	anthracene
aziridine	bis_chloro_ethyl_amine		di_carbonyl
epoxide	epoxy	fatty_series_ketone	
per_halo_ketone	halo_pyrimidine	hetero_substituted_carbonyl	

```
hydrazine          imine            michael_acceptor
phosphoric_ester   sulfone_halide   sulfonyl_ester   thio_ester
thiourea           alkyl_halide     long_alkyl
double hetero structure
```

以下にファイル出力イメージを示す。

```
INFORMATION> FindExcludeMol
    number = 1                              ← 入力マルチ mol2 の分子の通し番号
    filename = NS-00000001                  ← 分子名(ID)
    supplier = ACB_Blocks(Fragment)         ← サプライア名
    idnumber = 2IA-0011                     ← サプライア ID
    info = NONE;                            ← フィルタリング結果(NONE はフィルタしない)
    matchedmol =                            ← フィルタされた理由

    number = 89
    filename = NS-00000100
    supplier = Princeton
    idnumber = OSSK_266418
    info = CONTAIN_SPECIFIC_STRUCTURE;      ← 特定の不適切な分子構造を含むことを示す。
    matchedmol = aldehyde                   ← アルデヒド構造を含んでいる。
```

3-3-4. フィルタリングしたマルチ mol2 の出力

入力マルチ mol2 と不適切な構造を含む分子リストを指定して、不適切な構造を除いたマルチ mol2 ファイルを出力する。

3-4. クラスタリングツール

クラスタリングツールは、大量の化合物データ(例えば 500 万件データ)から、ある程度絞った件数のデータセット(例えば100 万件データ)を作成するツールである。本ツールは、化合物データの類似性を計算し、類似度の高いデータは代表 1 件のみ採用し、他のデータは不採用とすることにより、件数を絞ったデータセットを作成するものである。

分子間の類似度の計算は、myPresto/TGS ツールを使用するものとし、本ツールを呼び出すスクリプトを作成するものとする。

作成するスクリプトの機能は以下の通りとする。

・PickUp_data.pl

指定した件数の分子データを TMP ディレクトリにコピーします。閾値を決定する為に少数のデータをピックアップするのに使用します。

・SelectMol2_1st.pl

指定したディレクトリ内の cXXXXX(X は数字)ディレクトリ内の各分子の距離を計算し、閾値以下の分子をデータセット作成対象外とし、リストファイルに出力します。

Judge_SelectMol.pl

少数データでの TGS 処理を行い、その結果をスケールアップした際の予想クラスタリングデータ数を計算し、閾値の調整を行います。予想クラスタリングデータ数が少ない場合は閾値を小さくし、データ数が多い場合は閾値を大きくします。

Adjustmet_clustering.pl

全データを用いて TGS 処理を行った際の最終的なデータセット数が指定した値よりも多い、少ない場合のデータ数の調整を行います。

４．各種外部プログラム

化合物データベース構築ソフトウェアでは、各種の myPresto ツール等を呼び出して処理を行う。本章では、各種ツールについて記載する。

4-1. split_mols

split_mol2 プログラムは、sdf ファイルをシングル mol ファイルに分割する。入力ファイルとなる、2 次元 sdf ファイルは複数件の分子データから構成されており、これらを分子毎に 1 つのファイルに変換するために使用される。

化合物データベース構築ソフトウェアでは、以下のコマンドが実行される。

split_mol2 -i <input sdf> -o <work dir>

4-2. Hgene

Hgene プログラムは、mol ファイルから mol2 ファイルへの変換、水素付加、簡易的な 3 次元化処理、MOPAC mulliken 電荷計算を行う。

化合物データベース構築ソフトウェアでは、以下のコマンドが実行される。

・水素付加

Hgene -imdl <intput mol> -omol2 <output mol2> -p -3d –namiki <source_name> -wc -co -amide

・MOPAC mulliken 電荷計算

Hgene -imol2 <input mol2> -omol2 <output mol2> -mop AM1 –wc –namiki <source_name>

また、Hgene で計算する物性値等は以下の通りとする。

【物性値等の計算項目】

・分子式
・分子量
・分子総電荷
・水素ドナー数
・水素アクセプター数
・HOMO エネルギー値
・LUMO エネルギー値

・キラル原子数

4-3. tplgeneL

tplgeneL プログラムは、cosgene を実行するのに必要なトポロジーファイル、pdb ファイルを作成する。

4-4. cosgene_multiPDB

cosgene_multiPDB プログラムは、mol2 ファイルを入力してエネルギー最小化計算を行う。出力形式は mol2 ファイルである。
化合物データベース構築ソフトウェアでは、以下のコマンドが実行される。

cosgene < “コントロールファイル”

コントロールファイル：

```
EXE> INPUT
     TOPOLO= FORM   NAMETO= XXX.tpl
     COORDI= MOL2   NAMECO= XXX.pdb
     QUIT
EXE> MINI
     METHOD=  STEE  CPUTIM= 60.0
     LOOPLI=  5000  UPDATE= 20
     MONITO=  1000  CONVGR= 0.1D0
     CUTMET=  RESA  CUTLEN= 99.0D0
     DIEFUN=  DIST  DIEVAL= 4.0D0
     QUIT
EXE> END
EOF
```

4-5. confgeneC

confgenC プログラムは、光学・幾何異性体の生成を行う。
化合物データベース構築ソフトウェアでは、以下のコマンドが実行される。

confgene -i <input mol2> -nc a –rp 3 -omol2