

化合物データベース作成用 cosgene の作製

2018年2月5日

Copyright (C) 2006-2018 Next Generation Natural Product Chemistry (N²PC)

1. 概要

化合物データベースの作成では2次元 sdf ファイルから3次元 mol2 ファイルを作成するが、その際、異性体の構造を生成する部分などでの膨大なファイル IO が、システムのネットワークを混雑させシステムの機能を損なうと同時に煩雑なスクリプト処理が発生するので、この工程の簡素化を行う。今までの流れは以下の通り。

(1) `confgeneC/myPresto` による構造の発生

↓

(2) `tplgeneL/myPresto` にかけて、PDB ファイルとトポロジーファイルを作成

↓

(3) それぞれの PDB ファイルを `cosgene/myPresto` でエネルギー最小化

↓

(4) 立体構造のひずみをチェックし、化学的に間違っただ構造は、初期座標を入れ替えてやりなおし

↓

(5) エネルギー最小化された構造同士が、同じと見なされる構造に収束していないかを判別し、PDB ファイル形式を mol2 ファイルに変換して異なる構造をデータベースに収録

ここで、(1)で発生した複数の構造に対して、(2)で生成するトポロジーは共通である。なお、(1)までのステップは、別途、新しい手法で簡素化・高速化を行うので、当案件に含めない。

当案件では、`cosgene/myPresto` をデータベース生成用の専用ソフトに改変し、1つのトポロジーファイルに対して、複数の初期座標が与えられた場合、一度に、それぞれの分子を共通のトポロジーファイルで構造を最適化し（ステップ 3）、構造の誤りのチェックを行い、誤りがあれば再実行し（ステップ 4）、エネルギー最小化された構造の同一性を判断し（ステップ 5）、異なる構造に名前をつけて、PDB 形式及び `tpl2mol2/myPresto` の機能を用いて mol2 ファイル形式で出力するようにする。

GBSA の計算をするための `mkGBSAin.pl` の機能を内蔵すること（既に ELIE 用に作成したもので内蔵している）。

入力ファイル形式は、複数の分子が ENDMDL で区切られた形式とする。

2. 処理方式

2. 1. confgeneC と minimize の実行と処理方式

分子構造のひずみを検討するため、回転可能な角度を 4 個持つ化合物 2515296-02.mol2 について confgeneC と minimize を実行し、minimize 前と minimize 後の構造の目視確認を行った。

この結果、以下の 2 点が確認された。

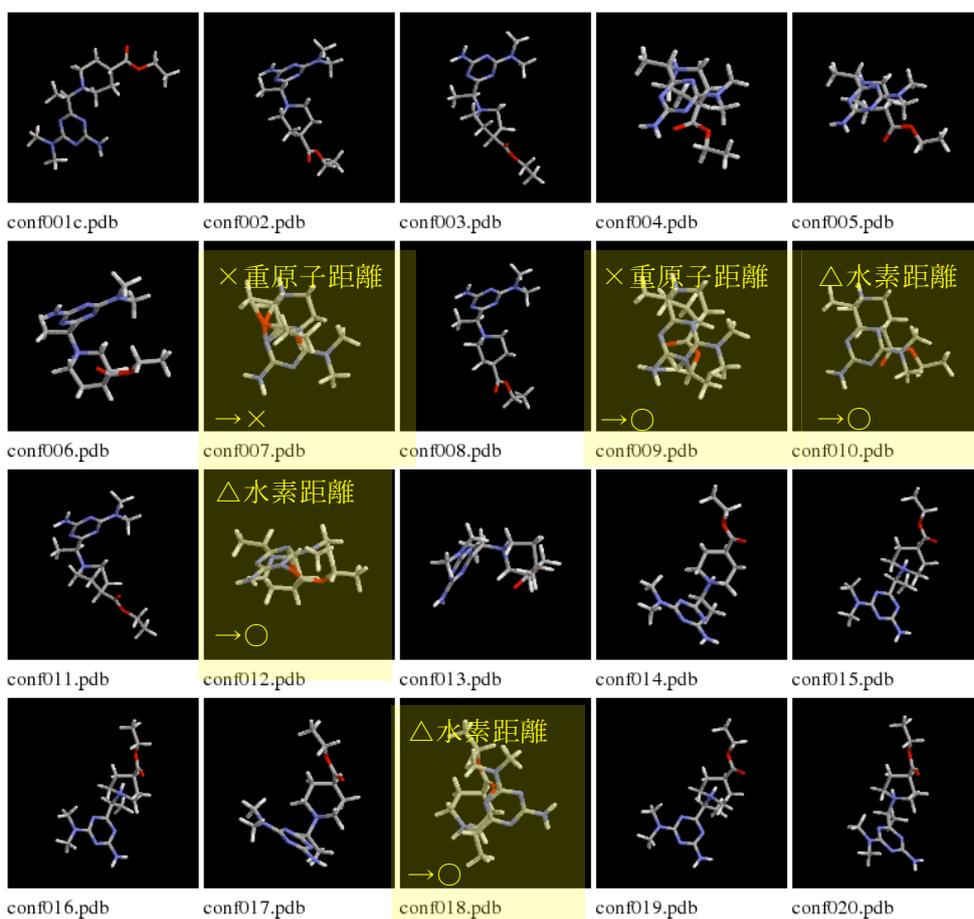
(1) confgeneC では環をまたがる位置に bond を配置し、これが不安定な構造のもとになる。

→本構造を検出対象とする必要がある

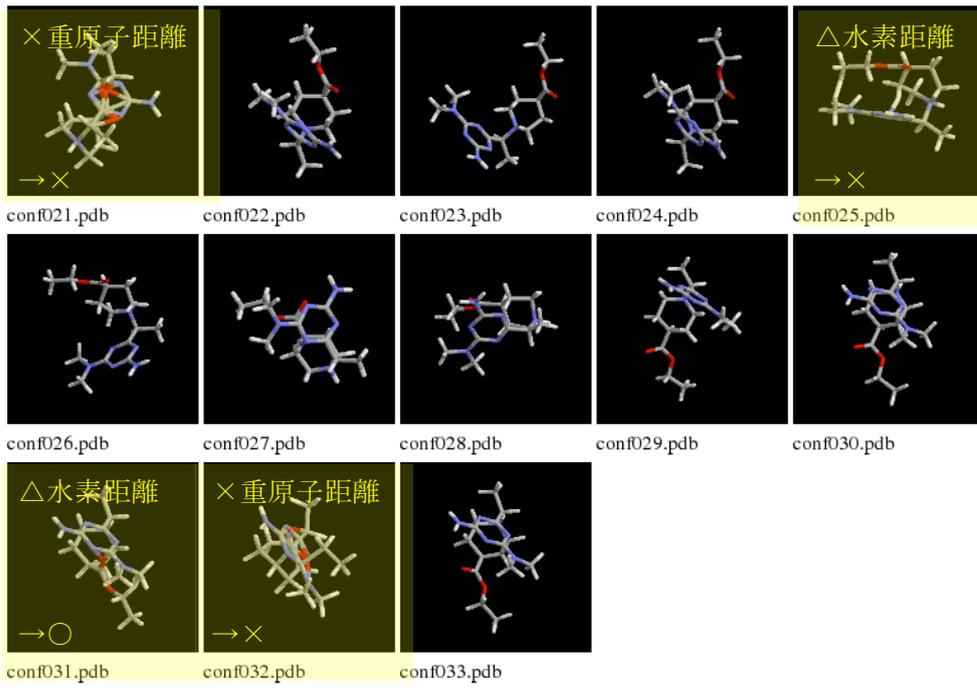
(2) confgeneC で作成した座標は規準となる部分構造が見当たらない

→minimize 後の直接の直交座標系比較では同一構造が判定できない。

また、主軸を重ねることが困難な分子(対称性大)も想定される。



minimize 前構造と minimize 後の目視結果(1/2)



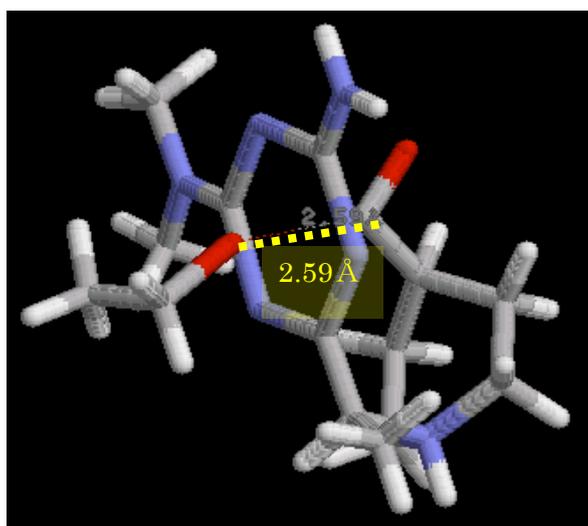
minimize 前構造と minimize 後の目視結果(1/2)

(1)立体構造のひずみについて

立体構造のひずみとしては下記の二つの構造が考えられる。

(1-1)環の構造をまたがる BOND

conf007 のように環をまたがった結合がある場合 2 分子の bond 距離は平衡距離の 1.35 Å と大きく異なる 2.59 Å となった。このような不正な BOND を立体構造のひずみとして検出する。



conf007 の minimize 後構造(環を挟む bond が長い)

このひずみを検出する方法として、以下 2 つの方法があり、(1-1-b)は基準があいまいなため、(1-1-a)の bond 距離のチェックを採用する。

(1-1-a)bond 距離が平衡距離*0.5Å以上の場合、構造を不正とみなす

全 bond に対し、平衡距離との差を求め、差分が 0.5 Å 以上ある場合、構造を不正とみなす。

(1-1-b)bond ポテンシャルが高い場合、構造を不正とみなす

bond ポテンシャル(E_{bond})は $k*(r-r_0)**2$ で、 k は 300.0~500.0 付近の値であるため、 $r-r_0$ の差が 1.0 Å の場合、 E_{bond} は 300.0~500.0 の値となる。

E_{bond} / bond 数が 100.0KCAL/MOL を超えた場合に構造を不正とみなす。

(1-2)1-4 間原子距離、1-5 間原子距離

OH の H は電荷のみ持ち、vdW 半径を持たないため、付近に-の電荷を持つ原子が存在する場合、O-H の距離が伸び、H と-電荷の原子の距離がほぼ 0.0 に近づく現象が発生する。

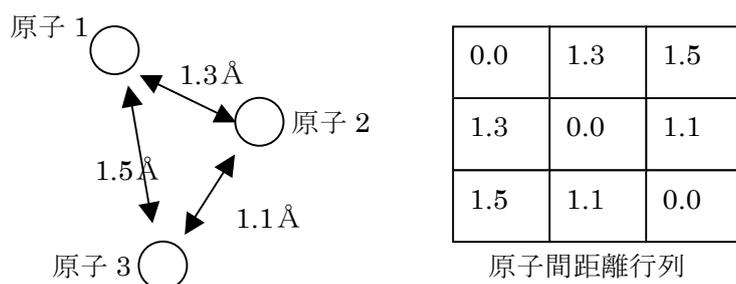
1-4, 1-5 の原子間距離が 1.0 Å 以下の場合、この構造を不正とみなす。

(2)構造の同一性判定

分子配座の同一性判定を行う方法として、以下の 2 つが考えられ、(2-1)の原子間距離行列比較を採用する。

(2-1)原子間距離行列の比較

n 原子で構成される分子の配座が m 個存在するとし、 i 番目配座での原子間距離行列 $S_i(i=1,m)$ を作成する。(S は $n \times n$ の対称行列)



各配座で距離行列 S_i を作成した後に、既に登録された配座の距離行列群 S_1, S_2, \dots, S_{i-1} の各行列要素の 2 乗距離の和 ΔS を計算し、 ΔS が閾値以内の場合に配座として登録する処理を繰り返す。

0.0	1.3	1.5
1.3	0.0	1.1
1.5	1.1	0.0

S1

0.0	1.2	1.6
1.2	0.0	1.2
1.6	1.2	0.0

S2

$$\Delta S = (1.3-1.2)**2 + (1.5-1.6)**2 + (1.1-1.2)**2 = 0.03$$

(2-2)torsion 角の比較

n 個の torsion で構成される分子の配座が m 個存在するとし、各配座での torsion 角行列 $T_i(i=1,m)$ を作成する。(S は $1 \times n$ の行列)

→torsion は相対的な指標であるが、同じ角度で分子の中心の torsion が回転した場合と末端の torsion が回転した場合に、同一の評価値となる。

torsion 比較場合、この配座の違いを評価できない問題がある。

以上