Vol.2022-BIO-69 No.15
2022/3/11

フラグメント化された化合物立体構造データベースの構築

稲垣 雅也[1]　柳澤 渓甫[1]　大上 雅史[1]　秋山 泰[1,a)]

概要：創薬における Fragment-Based Virtual Screening (FBVS) では，まず代表的なフラグメント（化合物の部分構造）に関して標的タンパク質との結合親和性を計算した後に，それらをなるべく多く満たすような化合物をデータベースから検索，発見することを目指す．本研究では，このうち後者の過程を助けることを目的として，薬剤候補化合物について，(1) 化合物が持つフラグメントの種類と，(2) 化合物の取り得る立体配座中での 2 つのフラグメント間の相対位置関係をあらかじめ整理したデータベースを構築することを提案する．このとき，(1)(2) の情報をビット列で表現し，(2) については当該化合物がとり得る立体配座を最大 200 種類計算して，それらの状態に対応するビット表現を論理和として重ね合わせることとした．検証実験では，PDBbind データベースの既知薬剤化合物 6,858 件から上述のデータベースを構築し，検索条件に合う化合物が抽出できることを確認した．

キーワード：Fragment-Based Virtual Screening，化合物データベース，化合物表現

Database of Drug Candidates Represented by 3D Positional Relationships between Fragments

Inagaki Masaya[1]　Yanagisawa Keisuke[1]　Ohue Masahito[1]　Akiyama Yutaka[1,a)]

Abstract: Fragment-Based Virtual Screening (FBVS) in drug discovery performs binding score calculation for many representative chemical fragments on the target protein, and then retrieves candidate compounds that satisfy the detected fragment conditions. To support the latter process, we propose a new database in which each registered chemical compound is associated with the pre-calculated information of (1) the types of fragments it has, and (2) the relative positions of all fragment pairs it has considering the possible 3D conformation of the compound. We use a bit-sequence representation for (1) and (2), and overlay (logical-OR) the bit representations of up to 200 possible conformations in (2). In a verification experiment, we built a database including 6,858 drug candidates in the PDBbind database, and showed that it is able to retrieve compounds that meet the search conditions.

Keywords: Fragment-Based Virtual Screening，compounds database，representation of a compound

1. はじめに

近年の分子標的創薬において，標的となるタンパク質と結合するリガンドを発見することは主要な課題である．計算機上で薬剤候補となる化合物を絞り込む手法として，Structure-Based Virtual Screening (SBVS) がある．この方法ではタンパク質と化合物の構造に着目して探索が行われる．タンパク質と化合物との間における結合親和性を予測することで，化合物が薬剤候補となりうるかについて選抜が行われる．Compound-Based Virtual Screening (CBVS) では，化合物 1 つ 1 つと各タンパク質との間において結合親和性を模したスコアを計算することで結合に必要な条件を満たすと予測される化合物を絞り込む．しかし，

[1] 東京工業大学 情報理工学院 情報工学系
Department of Computer Science, School of Computing, Tokyo Institute of Technology
[a)] akiyama@c.titech.ac.jp

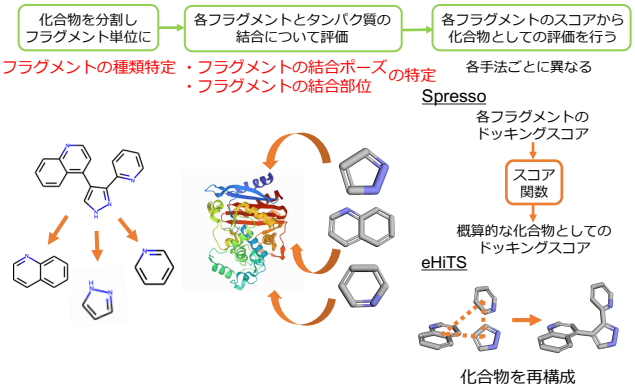

図 1 フラグメントの立体構造に基づくドッキング計算手法の先行研究の手順

drug-like な化合物空間の大きさは 10^{30}〜10^{60} と推定されている [1]．さらに，各化合物には結合の回転などの影響により原子位置が異なる複数の立体構造（配座）が存在するため，CBVS での計算量はより多くなる．そのため，タンパク質と全ての化合物の組み合わせについて結合親和性を模したスコアを網羅的に計算することは現状では難しい．

そこで，化合物を小さな部分構造であるフラグメント単位に分割した上で，薬剤候補化合物を探す Fragment-Based Virtual Screening (FBVS) と呼ばれる手法がある．CBVS では化合物を丸ごと用いて直接タンパク質に結合する際のスコアを計算する一方，FBVS では化合物をフラグメントと呼ばれる部分構造に分解した後，そのフラグメントをタンパク質に結合させた際のスコアを取得する．この方法では，Compound-Based Virtual Screening (CBVS) と比較して，計算効率が良いという利点がある．フラグメントは化合物を分割した小さな部分構造であり，多くの化合物に共通する部分構造である．そのため，同一の標的に対するフラグメントの予測結果を他の化合物評価の際に使い回すことが可能になり，計算量を削減できる．また，フラグメントは化合物と比較して効率的に化学空間をサンプリングしやすいことも FBVS の利点の一つである [2]．例えば，ZINC12 データベース [3]（ZINC データベースの 2012 年版）の“all purchasable”サブセットと“all boutique”サブセットを合わせた 28,629,602 化合物をフラグメント分割すると，化合物数よりも少ない 263,319 フラグメントが得られる [4]．つまり，実際の化合物数よりも少ないフラグメントで化合物空間を表現できる可能性がある．

フラグメントの立体構造に基づくドッキング計算手法の先行研究として，Spresso[4]，FlexX[5]，eHiTS[6]，久保田らの手法 [7] 等が挙げられる．フラグメントの立体構造に基づくドッキング計算手法の先行研究の手順をまとめると図 1 のようになる．図 1 を見ると，化合物のフラグメントへの分解，各フラグメントとタンパク質の結合の評価はいずれの先行研究でも経ている過程である．つまり，それらの過程はドッキング計算手法において，必要かつ重要な過程であると言える．最終段階である化合物としての評価は各手法によって方法が異なる．例えば，Spresso ではタンパク質に結合するフラグメントの種類から化合物全体としての結合親和性をあらかじめ定義した計算式を用いて概算として評価していること，eHiTS では各フラグメントの結合ポーズと結合部位からフラグメント間の立体的な相対関係に基づいて化合物としての結合ポーズを評価していることがわかる．

以上のように，先行研究の手順からフラグメントに基づくドッキング計算手法において重要な要素は以下の 3 要素であると考えられる．

(1) 化合物が含むフラグメントの種類
(2) 化合物内におけるフラグメント間の相対的な位置関係
(3) 化合物内における各フラグメントの結合ポーズ

そこで，本研究では，化合物における上記の 3 要素をあらかじめ整理したデータベースを作成する．このデータベースの活用により，標的に対して効果的なフラグメントや，フラグメントの組み合わせに関する部分的な情報が得られた際にそれらの情報を条件として満たす薬剤候補化合物を素早く抽出できるようになると考えられる．

2. データベース構築手法

これまでに述べたように，先行研究から，フラグメントに基づくドッキング計算における薬剤候補化合物の絞り込みに重要な要素は，1）化合物が含むフラグメントの種類，2）化合物内におけるフラグメント間の相対的な位置関係，3）標的タンパク質に対する各フラグメントの結合ポーズの 3 点である．そこで，化合物が含むフラグメントの種類と複数の配座におけるフラグメント間の立体的な相対関係をあらかじめ記録したデータベースを構築することを提案する．要素 3）標的タンパク質に対する各フラグメントの結合ポーズについても，結合ポーズは各フラグメントの位置と回転の合成で表現できることから，提案するデータベースは上記の薬剤候補化合物の絞り込みに重要な 3 要素を踏まえたデータベースになる．また，化合物がとりうる複数の配座情報をあらかじめ計算しデータベース上で表現する．これによって，このデータベースはフラグメントに基づくドッキング計算結果に適合する条件を持った化合物の検索を素早く行うことを可能にする．構築は大きく分けて 3 つの手順からなる（図 2）．

データベース構築手順 1 結合親和性の高いフラグメントの抽出

データベース構築手順 2 化合物内のフラグメントの有無に基づくフラグメント有無ビット列作成

データベース構築手順 3 化合物内のフラグメント対の立体位置関係（配座）に基づくフラグメント相対関係ビット列作成

以下ではこれらの手順について説明を行う．

2.1 結合親和性の高いフラグメントの抽出（データベース構築手順 1）

2.1.1 既存化合物データベース内の全フラグメントを取得

既存化合物データベース内の化合物をフラグメントに分解することで，化合物データセット中の全フラグメントを取得する．用いたフラグメントへの分割方法は，各フラグメントが内部自由度をもたないように化合物を分割する方法である（図 3）．これは化合物の形状をフラグメント間の立体的位置関係（配座）だけで表現するためである．

2.1.2 全フラグメントから結合親和性の高いフラグメント群を抽出

予備実験として，ZINC15[8]（ZINC データベースの 2015 年版）の in-vitro データセットを用いて各フラグメントの出現回数を数える実験を行った．結果，出現回数上位 500 件のフラグメントの出現回数は図 4 のようになった．これより，データセット内でのフラグメントの出現回数に大きな差があることがわかる．そこで，タンパク質への結合親和性が高いフラグメントのみに着目することを目的として，Term Frequency-Inverse Document Frequency (TF-IDF) を用いたフラグメントの選定を行う．これによって，タンパク質への結合親和性が高い薬剤候補となりうる化合物の取りこぼしを少なくした上で，1 化合物あたりのデータサイズを制限することができると考えられる．タンパク質に結合することが確認されている既存の化合物データセットを D_A，一般的な化合物のデータセットを D_G，フラグメ

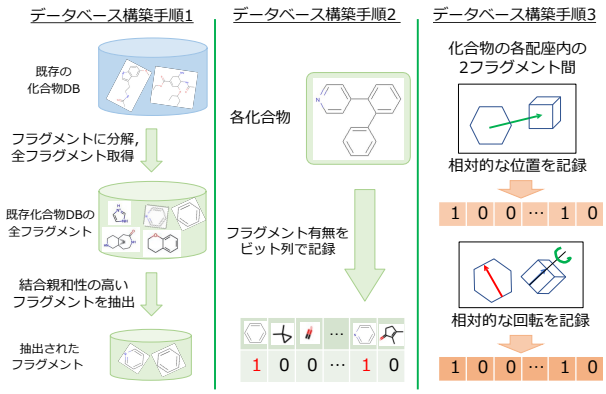

図 2　データベース構築手順

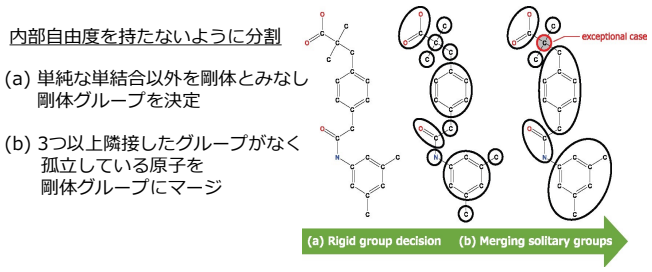

図 3　化合物のフラグメントへの分割方法 [4]

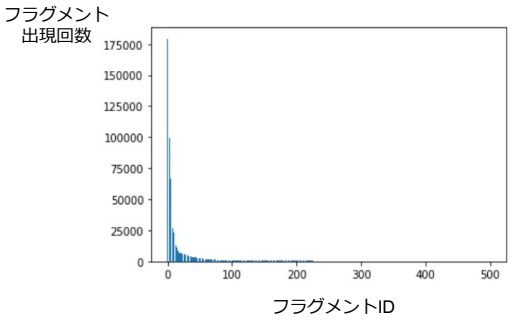

図 4　ZINC15 の in-vitro データセット内における出現回数上位 500 フラグメントの出現回数

ントを f とした時，本研究における各フラグメント f に関する TF，IDF 及び TF-IDF を式 1，2，3 と定める．

$$TF(f) = D_A\text{の化合物内に } f \text{ がある確率} \tag{1}$$

$$IDF(f) = \ln\left(\frac{1}{(D_G + D_A)\text{ の化合物内に } f \text{ がある確率}}\right)+1 \tag{2}$$

$$TF\text{-}IDF(f) = TF \times IDF \tag{3}$$

TF を式 1 のように定めた理由はタンパク質に結合することが確認されている既存の化合物データセットに多くの化合物に含まれるフラグメントはタンパク質への結合に関与する可能性が高いと考えられるためである．また，IDF を式 2 のように定めた理由は，一般的な化合物のデータセットに含まれる確率が高いほど非特異的なフラグメントである可能性が高いと考えられるためである．片方のデータセットのみに含まれるフラグメントが存在することが考えられるため，IDF を求める際には両データセットの化合物を合算した後にフラグメントの出現回数を調査する．式 3 により各フラグメントの TF-IDF を計算し，値の高い順にソートを行う．その後，PAINS[9] とフラグメントの原子数によりフラグメントのフィルタリングを行う．これにより，構造によらず陽性を示す（偽陽性）フラグメントと結合親和性が低いと考えられる原子数が 1 個であるフラグメントの抽出を避ける．

2.2 化合物内のフラグメントの有無に基づくフラグメント有無ビット列作成（データベース構築手順 2）

データベース構築手順 1 で抽出されたフラグメント群の各フラグメントをビットと対応づけることで，抽出されたフラグメントの個数と同等の長さのビット列を作成し，各ビットの初期値を 0 とする．各化合物において化合物がデータベース構築手順 1 で抽出されたフラグメントを含む時，そのフラグメントに対応するビットに 1 を代入する．このようにして化合物が含むフラグメントに関するビット

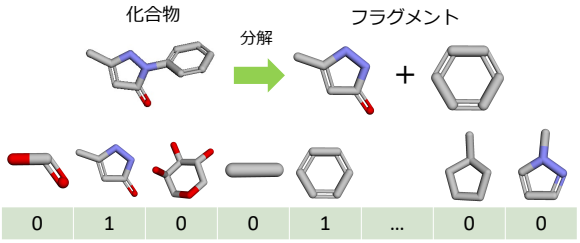

図 5　フラグメント有無ビット列の作成

列を作成し，フラグメント有無ビット列として定める（図 5）．

2.3　化合物内のフラグメント対の立体位置関係（配座）に基づくフラグメント相対関係ビット列作成（データベース構築手順 3）

配座生成ツールを用いて各化合物の配座を複数生成する．その上で，それぞれの配座内における 2 フラグメント間の立体位置関係（相対位置と相対回転）を記録する．

フラグメントの回転状態を数値で表現するためには，あらかじめ基準となる回転状態を定め，配座内におけるフラグメントの向きと比較する必要がある．そこで，回転の基準となるフラグメントの位置として各フラグメントに回転基準状態を定義する．

2.3.1　配座内におけるフラグメント間相対位置の取得

フラグメントペアの一方のフラグメントが基準回転状態に置かれるように分子全体を回転させた後，2 フラグメント間の重心間ベクトルを測定することで，フラグメント間相対位置関係を取得する（図 6）．

2.3.2　配座内におけるフラグメント間相対回転の取得

各フラグメントが基準回転状態に置かれるようにフラグメントを回転させ，その際の回転行列を取得することで，配座内における各フラグメントの回転状態を定義する．そして，配座内の 2 フラグメント間において回転行列の積 (AB^{-1}) から回転の差を計算することで 2 フラグメント間相対回転を取得する（図 7）．

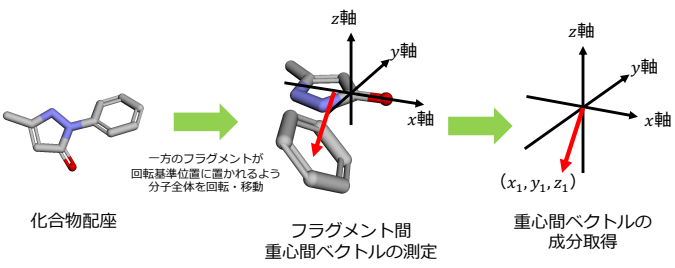

図 6　フラグメント間相対位置関係の取得

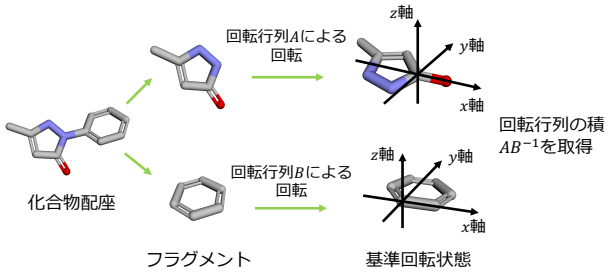

図 7　フラグメント間相対回転の取得

2.3.3　複数配座を表現した化合物ビット列の作成（図 8）

配座における全フラグメントペアの相対関係情報を 1 つのビット列に埋め込む．そして，配座から得たビット列の論理和を取ることで 1 つのビット列に統合を行う．その後，データベース構築手順 2 で作成したフラグメント有無ビット列と合体させる．これによって，化合物ビット列は複数の配座情報を含んだ表現となる．完成する化合物ビット列は図 9 のようになる．化合物ビット列を化合物 ID と smiles 記法に結び付けてファイルに書き込むことでデータベースを構築する．

3.　データベース検索手法

本データベース内の化合物情報はビット列で表現されている．データベースへの入力をデータベース作成時と同様の操作でビット化することにより，論理和と論理積を用いた検索を実現した．

入力されたビット列とデータベース内の各ビット列の論理積をとる．この時，論理積から得られたビット列と入力されたビット列が完全一致であるならば，データベース内のビット列が，入力ビット列のフラグメント状態を包含することになる．そのため，データベース内のビット列に対応する化合物が入力条件を満たすことになる．この方法を

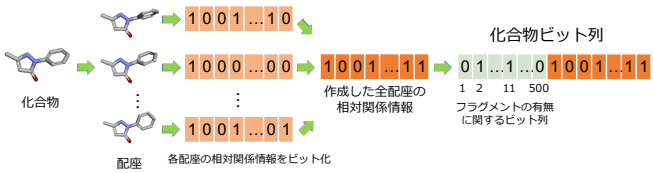

図 8　化合物のビット列表現の手順

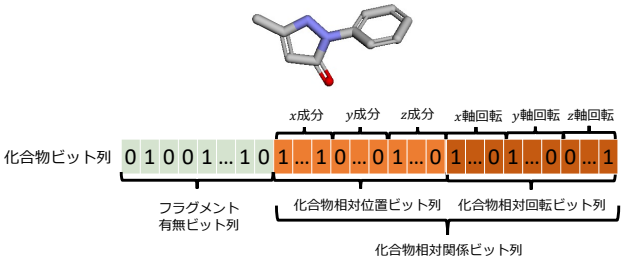

図 9　化合物立体構造に基づく化合物ビット列

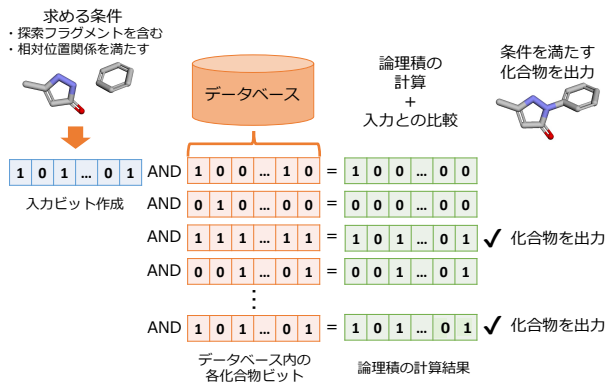

図 10　データベース内化合物の検索

データベース内の全化合物に用いて化合物検索を行う（図 10）．出力は化合物 ID と smiles 記法で行われる．

4. 実験

本実験では，RDKit[10] を用いた．RDKit 上において，ホウ素がなす結合が特殊であることが多く処理できないことがあったため，全ての実験においてホウ素原子を含む化合物及びフラグメントは調査対象から除外した．

4.1 TF-IDF によるフラグメント抽出

本実験では，本提案手法のデータベース構築手順 1 に従い，データベース構築に必要なフラグメント群を定める．タンパク質に結合することが確認されている既存の化合物データセット (D_A) として，ZINC15 の in-vitro データセットを使用する．in-vitro データセットは 10 μM 以下の濃度で実験的にタンパク質との結合が確認された化合物のデータセットである．また，一般的な既存の化合物データセット (D_G) として，ZINC15 の in-stock データセットを使用する．in-stock データセットは業者から直接買うことができる化合物のデータセットである．2 つのデータセットの概要とデータセットが含む化合物数，フラグメント数をまとめて表 1 に示す．各フラグメント f の TF-IDF の式 1，式 2 は用いるデータセットを定めたことで式 4，式 5 のように定義できる．これらの式から両データセット内の全てのフラグメントに対して TF-IDF を計算し，500 種類のフラグメントを抽出する．

その後，抽出された 500 種類のフラグメントに関して，FBVS で用いるフラグメントとして適切であるかの評価を行うために，フラグメント選択の目安である Rule of 3[11] で吟味される特徴量（分子量，水素ドナー，水素アクセプター，LogP，PSA，回転結合数）を求める．今回は PSA に代わって TPSA を用いることとする．また，フラグメントが平面的でなく 3 次元的に広がりを持った形状かを調べるために PMI 分析 [12] を用いる．PMI 分析では，水素が付加している際のフラグメントの形状を調査する．さらに，500 種類のフラグメントの多様性の評価を行うために，1024 ビットの ECFP4 フィンガープリントと MACCS フィンガープリントに基づいた化合物間の類似度を Tanimoto 係数で測定する．

$$TF(f) = in\text{-}vitro \text{ の化合物内に } f \text{ がある確率} \quad (4)$$

$$IDF(f) = \ln\left(\frac{1}{(in\text{-}stock + in\text{-}vitro) \text{ の化合物内に } f \text{ がある確率}}\right)+1 \quad (5)$$

表 1　実験 4 – 1 で用いるデータセットの説明

データセット名	化合物数	フラグメント数	データセットの概要
ZINC15[8] in-stock	7,575,898	188,526	業者から直接買うことができる化合物
ZINC15[8] in-vitro	306,347	66,207	実験的にタンパク質との結合が確認された化合物（≦10 μM）

4.2 既知リガンドを用いたデータベース検索

本実験ではデータベース内の化合物が標的と結合する際の配座情報を含むかを調べるために，PDBbind[13] から得た既知リガンドに関する SDF ファイルの情報を用いてデータベースを構築する．PDBbind から得た既知リガンドに関する SDF ファイルを内部自由度を持たないフラグメントに分解する．そして，OMEGA[14] を用いてそれらの配座を生成し，それらの情報を基にデータベースを構築する．OMEGA は各化合物につき最大 200 配座を生成し，各配座間の平均二乗偏差 (RMSD) が 0.5 Å 以下にならないように設定する（デフォルト設定）．OMEGA により生成された配座は元データとなる PDBbind の座標に依らず，独立的に生成される．

データベースの構築にあたり，PDBbind から取得した化合物数が 9,178 件であり，OMEGA による配座生成時に 1,159 件の化合物で配座生成プロセスが失敗した．この失敗は容易に解消できないものであると判断したため，これらの化合物をデータベース対象化合物から除外した．その後，本データベース構成時にフラグメント座標を取得できない化合物が 1,159 件，PDBbind から取得したオリジナルデータ SDF ファイルにおいて結合が切断されていた化合物が 2 件あった．この過程から，化合物を含むデータベースが含む化合物数は 6,858 件となった．

検索のための入力ビット列は PDBbind から得た既知リガンドに関する SDF ファイルの情報をビット列化する．データベースがその入力によりどのような出力を行うかについて調べる．

相対位置ビットと相対回転ビット共に，ゆとり幅を設ける．これは実数値のビット列化により，実数値上で近い値がビット上で区別されることを防ぐためである．ゆとり幅については各相対位置及び相対回転の刻み幅に倍数 c をかけることで定義を行う（$C_r = c \times w_r$，$C_\theta = c \times w_\theta$）．

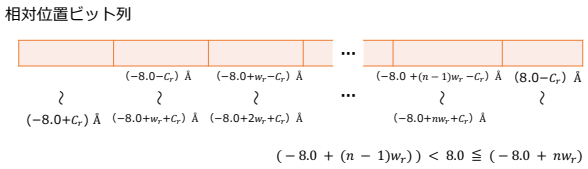

図 11 実験 4.2 における実数値と相対位置ビット列の対応

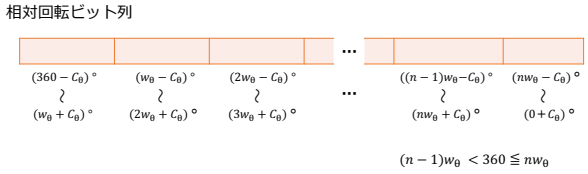

図 12 実験 4.2 における実数値と相対回転ビット列と実数値の対応

なお，相対位置ビット列と相対回転ビット列について，各ビットと各成分，及び各ビットと各軸周りの回転角が図 11，図 12 になるように定める．

5. 実験結果

5.1 TF-IDF によるフラグメント抽出結果

抽出されたフラグメントに関する分子量，水素アクセプター，水素ドナー，LogP，TPSA に関する特徴量の分布を図 13 に示す．なお，回転結合数はフラグメントの定義より，全てのフラグメントにおいて 0 となる．これらに基づいて，Rule of 3 の条件を満たす抽出されたフラグメントの割合を求めると表 2 のようになる．

また，PMI 分析を用いて水素付加後のフラグメントの形状に関する分布を図 14 に記載する．PMI 分析における立体的なフラグメントの明確な定義はなく，$(NPR1 + NPR2 \geqq 1.07$[15]$)$ は立体的にみなされる場合がある [16]．そこで，今回は $NPR1 + NPR2 \geqq 1.07$ を満たすフラグメントを立体的なフラグメントとして扱うとする．この時，抽出されたフラグメントの 50.0 %が 3 次元的に広がりを持った形状をしている．

フラグメント間の類似度を調査するために，抽出されたフラグメント群に属する 2 フラグメントペアの 1024 ビットの ECFP4 フィンガープリントに基づく Tanimoto 係数と，MACCS フィンガープリントに基づく Tanimoto 係数を算出した．各 Tanimoto 係数の分布を図 15 に示す．図 15 から，ともに値が 0.0〜0.1 となる部分に分布の山を持ち，値が大きくなるにつれて分布が少なくなっていることがわかる．

5.2 既知リガンドを用いたデータベース検索結果

ゆとり幅倍率 $c_r = 1/3$，$c_\theta = 1/3$ における相対位置と相対回転のビット刻み幅による出力が検索条件元となった化合物を含む割合の変化を図 16 に，相対回転のビット刻み幅が 60 度の時のデータベースからの出力数に関する箱ひげ図を図 17 に示す．相対位置ビット刻み幅が 1.5 Å，相対回転ビット刻みが 60 度，ゆとり幅倍率 $c_r = 1/3$，$c_\theta = 1/3$ の時，出力された化合物群に検索条件の元となった化合物が含まれる割合は 9 割以上であった．また，1 つの入力に対する出力数はデータベースの大きさ（化合物 6,858 件）と比較して 2.0 % 前後と小さかった．

表 2 Rule of 3 の各指標における抽出フラグメント群の評価

各指標と基準値	条件を満たす抽出フラグメントの割合
分子量 300 以下	99.0 %
H ドナー 3 以下	91.4 %
H アクセプター 3 以下	74.4 %
LogP 3 以下	98.2 %
PSA 60 以下	76.8 %
回転結合数 3 以下	100 %

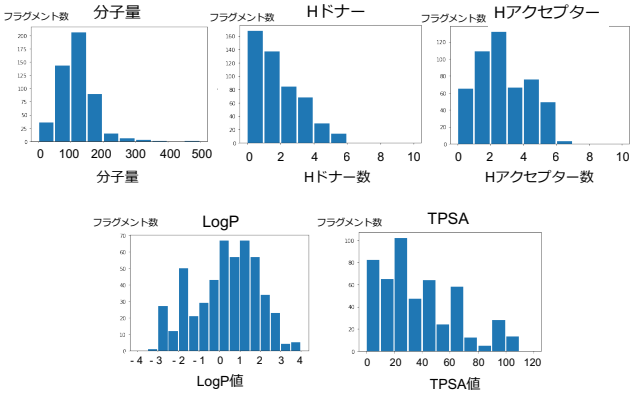

図 13 抽出されたフラグメントの特徴量の分布

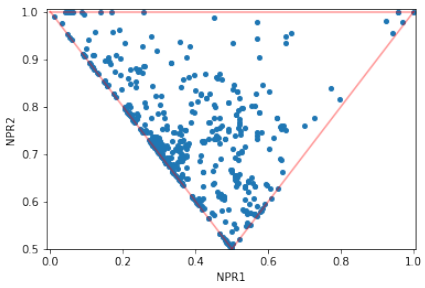

図 14 抽出されたフラグメントの形状（左上：直線形，中央下：平面形，右上：立体形）

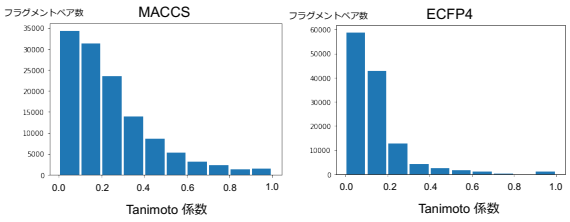

図 15 抽出されたフラグメント間の Tanimoto 係数

6. 考察

6.1 TF-IDF によるフラグメント抽出に関する考察

実験 5.1 で行った TF-IDF によるフラグメントの抽出に

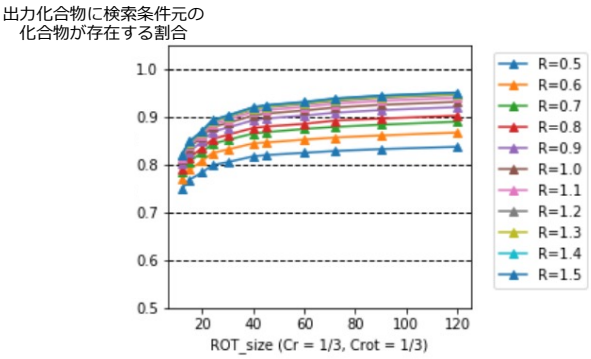

図 16 ゆとり幅倍率 $c_r = 1/3$，$c_\theta = 1/3$ におけるビット刻み幅による出力が検索条件元となった化合物を含む割合の変化

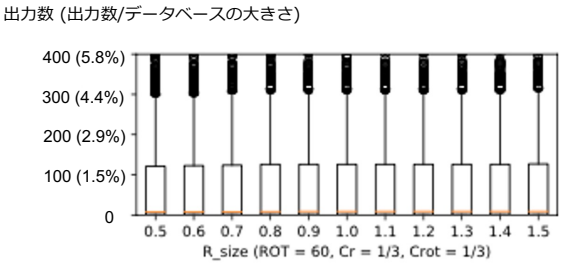

図 17 相対回転のビット刻み幅が 60 度の時のデータベースからの出力数に関する箱ひげ図

ついて述べる．

初めに，抽出されたフラグメントが FBVS に使われるフラグメントとして適しているかについて考察をする．抽出されたフラグメントの多くはフラグメント選択の目安である Rule of 3 を満たし，立体的に広がったフラグメントによる親和性予測にも適応が可能であると考えられることから，SBVS で用いられるフラグメントとして，抽出されたフラグメント群は適切であるといえる．

次にフラグメント群の多様性について考える．実験 4.1 の Tanimoto 係数の分布から抽出されたフラグメント群に所属するフラグメントは類似性が低いと考えられ，多様なフラグメントから構成されていることから幅広い化学空間の探索が可能になると言える．

最後に，TF-IDF を用いた効果について考察する．本研究の実験 4.1 では既存化合物データセットである ZINC15 の in-vitro データセットをフラグメント抽出元として用いた．抽出されたフラグメント 500 件で抽出元の in-vitro データセットの上位 500 件に含まれないフラグメントは 42 件あった．この 42 件のうち，PAINS による除外対象は 3 件，原子数制限による除外対象は 2 件であった．つまり，37 種類のフラグメントは TF-IDF の順位づけの効果により抽出されなかったこととなる．以上のことから，フラグメント抽出の過程において，単に結合親和性の高いデータセットである in-vitro データセット内における出現回数が多い順に取得したのではなく TF-IDF に一定の効果があったことが伺える．

6.2 既知リガンドを用いたデータベース検索に関する考察

実験 4.2 の結果から，OMEGA の配座生成により構築されたデータベースにおいて，化合物がタンパク質と結合する配座でのフラグメントの座標情報を入力として該当化合物を抽出可能であることが確認できた．

この結果から，OMEGA の配座生成により構築されたデータベースは化合物が標的に結合する際の配座情報を含んでいることが伺える．つまり，あらかじめ配座が計算された本データベースを活用することで，FBVS における化合物の再構築手順などが簡略化でき，化合物評価の計算が従来よりも早くなることが考えられる．

出力された化合物群に検索条件の元となった化合物が含まれない場合については，OMEGA により生成される配座間の RMSD が 0.5 Å より大きくなるように設定されていること（デフォルト設定）が原因であると考察できる．原子数が多い化合物である場合，1 原子が RMSD に及ぼす影響は小さくなる．つまり，仮に 2 つの配座間において数原子の座標が大きくずれていたとしても化合物内の多くの原子座標が同等の値に近ければ，配座間の RMSD は 0.5 Å 以下になり，OMEGA はどちらか一方の配座を生成しないことが想定される．この時の座標が大きくずれた数原子により構成される部分構造が，抽出されたフラグメントに該当する場合，本来はそのフラグメントの原子は複数の原子位置をとるにも関わらず，OMEGA が生成する配座には反映されないことになる．このようなことが相対位置や相対回転の取得に影響を及ぼし，データベース検索の際に，出力された化合物群に検索条件の元となった化合物が含まれないことがあると考えられる．

7. 結論

本研究では，フラグメントに基づく SBVS での重要な要素として（1）化合物が含むフラグメントの種類，（2）化合物内におけるフラグメント間の相対的な位置関係，（3）化合物内における各フラグメントの結合ポーズ，の 3 要素を考えた．そして，それらの要素から構成されるビット列の化合物表現を作成し，それを用いた化合物立体構造データベースを構築した．

各化合物表現は化合物がとりうる配座情報を含んだ情報を持つ．化合物表現の作成時には，既存化合物データセット内でのフラグメントの出現回数の大きな差があることから，全てのフラグメントを扱うのではなく TF-IDF をもとに PAINS と原子数による制限を加えた．抽出されたフラグメントについて，SBVS におけるフラグメント単体としての評価，多様性の評価，抽出元となる既存化合物データセット内での出現回数について調査した．調査の結果，抽出されたフラグメントは Rule of 3 や分子形状の観点から SBVS で用いられるフラグメントとして適切であり，類似性が低く，TF-IDF に一定の効果があったことが伺えた．

さらに，これらのフラグメントに基づき作成される化合物表現により PDBbind 中の既知薬剤化合物 6,858 件から提案データベースを構築し，検索条件に合う化合物の抽出をする実験を行った．各化合物表現には化合物がとりうる最大 200 配座の情報を含む設定とした．結果として，多くの化合物で出力数をデータベースサイズの 2.0 % 前後に抑えることができ，出力された化合物群に検索条件の元となった化合物が含まれる割合は 9 割以上であった．この実験結果から，本研究における提案データベースが標的との結合時の配座情報を持つ化合物を抽出可能かつ調査対象化合物を削減可能であると結論づけた．

今後の課題として，出力された化合物群に検索条件の元となった化合物が含まれない場合の改善が挙げられる．現在は配座生成時に配座間の RMSD が 0.5 Å 以下にならないようにする制限があるが，それを取り除く，またはより RMSD の閾値を小さくする（より似ている配座の出力を許容する）ことで配座間における数原子のずれによる相対位置や相対回転の取得への影響を抑えることができると考えられる．また，検索条件に対する出力数の削減を目指す．現在は全ての相対位置や相対回転情報を同一ビット列に埋め込んでいるが，データセット内で出現回数が多いフラグメントペアに関するビット列と出現回数が少ないフラグメントペアに関するビット列を分けて保存する．これにより，各ビットに 1 が立つ確率を低くすることが可能になり，条件に合致する化合物を取りこぼしを少なくしたまま，出力数の削減が見込まれる．さらには，タンパク質と結合する際のフラグメント条件から検索を行うための検索機能だけではなく，他のドッキングツールとの組み合わせにより化合物のドッキング評価の効率化を目指すことが今後の課題として挙げられる．

参考文献

[1] Gisbert Schneider. Automating drug discovery. *Nature Reviews Drug Discovery*, Vol. 17, pp. 97–113, 2018.

[2] Richard J. Hall, Paul N. Mortenson, and Christopher W. Murray. Efficient exploration of chemical space by fragment-based screening. *Progress in Biophysics and Molecular Biology*, Vol. 116, pp. 82–91, 2014.

[3] John J. Irwin, Teague Sterling, Michael M. Mysinger, Erin S. Bolstad, and Ryan G. Coleman. ZINC: A Free Tool to Discover Chemistry for Biology. *Journal of Chemical Information and Modeling*, Vol. 52, pp. 1757–1768, 2012.

[4] Keisuke Yanagisawa, Shunta Komine, Shogo D. Suzuki, Masahito Ohue, Takashi Ishida, and Yutaka Akiyama. Spresso: an ultrafast compound pre-screening method based on compound decomposition, *Bioinformatics*, Vol. 33, pp. 3836–3843, 2017.

[5] Matthias Rarey, Bernd Kramer, Thomas Lengauer, and Gerhard Klebe. A fast flexible docking method using an incremental construction algorithm. *Journal of Molecular Biology*, Vol. 261, pp. 470–489, 1996.

[6] Zsolt Zsoldos, Darryl Reid, Aniko Simon, Sayyed Bashir Sadjad, and A. Peter Johnson. eHiTS: A new fast, exhaustive flexible ligand docking system. *Journal of Molecular Graphics and Modelling*, Vol. 26, pp. 198–212, 2007.

[7] 久保田 陸人, 柳澤 渓甫, 吉川 寧, 大上 雅史, 秋山 泰. 共通な部分構造の再利用による高速なタンパク質リガンドドッキング手法の開発. 情報処理学会研究報告 バイオ情報学 (BIO) , Vol. 2020-BIO-61(4), pp. 1–8, 2020.

[8] Teague Sterling, and John J. Irwin. ZINC 15 – Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling*, Vol. 55, pp. 2324–2337, 2015.

[9] Jonathan Baell, and Michael A. Walters. Chemistry: Chemical con artists foil drug discovery. *Nature*, Vol. 513, pp. 481–483, 2014.

[10] RDKit: Open-source cheminformatics; http://www.rdkit.org.

[11] Miles Congreve, Robin Carr, Chris Murray, HarrenJhoti. A ‘Rule of Three’ for fragment-based lead discovery?. *Drug Discovery Today*, Vol. 8, pp. 876–877, 2003.

[12] Wolfgang H. B. Sauer, and Matthias K. Schwarz. Molecular shape diversity of combinatorial libraries: a prerequisite for broad bioactivity. *Journal of Chemical Information and Modeling*, Vol. 43, pp. 987–1003, 2003.

[13] Zhihai Liu, Minyi Su, Li Han, Jie Liu, Qifan Yang, Yan Li, and Renxiao Wang. Forging the Basis for Developing Protein–Ligand Interaction Scoring Functions. *Accounts of Chemical Research*, Vol. 50, pp. 302–309, 2017.

[14] OMEGA 4.1.2.0: OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com.

[15] Nicholas C. Firth, Nathan Brown, and Julian Blagg. Plane of Best Fit: A Novel Method to Characterize the Three-Dimensionality of Molecules. *Journal of Chemical Information and Modeling*, Vol. 52, pp. 2516–2525, 2012.

[16] David J. Hamilton, Tom Dekker, Hanna F. Klein, Guido V. Janssen, Maikel Wijtmans, Peter O’Brien, and Iwan J.P. de Esch. Escape from planarity in fragment-based drug discovery: A physicochemical and 3D property analysis of synthetic 3D fragment libraries. *Drug Discovery Today: Technologies*, Vol. 38, pp. 77–90, 2020.