Vol.2016-MPS-108 No.18
Vol.2016-BIO-46 No.18
2016/7/5

フラグメント分割に基づく超高速 化合物プレスクリーニング手法 ESPRESSO

柳澤 渓甫[1,4,a)]　小峰 駿汰[2,4]　鈴木 翔吾[2,4]　大上 雅史[1,3]　石田 貴士[1,3,4]　秋山 泰[1,3,4]

概要：化合物群から薬剤候補化合物を選別するための計算としてタンパク質-化合物ドッキングがよく用いられるがこの手法は計算量が大きく，対象の化合物数が膨大である場合はあらかじめ化合物群を削減するプレスクリーニング手法が用いられる．構造に基づくプレスクリーニングの従来手法は簡易的にドッキング計算を行うもので，通常のドッキング計算に比べれば高速ではあるものの，数千万規模の化合物への適用には速度が大幅に不足している．本研究では，化合物群のフラグメント分割を行い，フラグメント間の計算結果の再利用をすることで超高速にプレスクリーニングを行う ESPRESSO (Extremely Speedy PRE-Screening method with Segmented cOmpounds) を提案する．簡易的なドッキング計算の 1 つである Glide HTVS と比較して，ESPRESSO は約 2,900 万件の化合物を最大約 200 倍高速にプレスクリーニングすることができた．

キーワード：創薬支援，バーチャルスクリーニング，フラグメント分割，化合物プレスクリーニング

ESPRESSO: An ultrafast compound pre-screening method based on compound decomposition

Keisuke Yanagisawa[1,4,a)]　Shunta Komine[2,4]　Shogo D. Suzuki[2,4]　Masahito Ohue[1,3]
Takashi Ishida[1,3,4]　Yutaka Akiyama[1,3,4]

Abstract: Recently, the number of available protein tertiary structures and compounds has increased. However, structure-based virtual screening is computationally expensive due to docking simulations. Thus, methods that filter out obviously unnecessary compounds prior to computationally expensive docking simulations have been proposed. However, the calculation speed of these methods is not fast enough to evaluate more than 10 million compounds. In this study, we proposed a novel, docking-based pre-screening protocol named ESPRESSO (Extremely Speedy PRE-Screening method with Segmented cOmpounds). Partial structures (fragments) are often common among several compounds; therefore, the number of fragment variations needed for evaluation is smaller than that of compounds. Our method increased calculation speeds approximately 200-fold compared to conventional methods.

Keywords: computational drug discovery, virtual screening, compound decomposition, pre-screening

1. はじめに

創薬研究において，大量の化合物から新薬候補の化合物を発見することは「干草の中から針を探す」 (“Finding a needle in a haystack”) ようなものであるとたとえられて

[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
[2] 東京工業大学 大学院情報理工学研究科 計算工学専攻
Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology
[3] 東京工業大学 科学技術創成研究院 スマート創薬研究ユニット
Advanced Computational Drug Discovery Unit (ACDD), Institute of Innovative Research, Tokyo Institute of Technology
[4] 東京工業大学 情報生命博士教育院
Education Academy of Computational Life Sciences (ACLS), Tokyo Institute of Technology
[a)] yanagisawa@bi.cs.titech.ac.jp

Vol.2016-MPS-108 No.18
Vol.2016-BIO-46 No.18
2016/7/5

おり [1]，大量の化合物すべてについて *in vitro* 実験を行う前に計算機を用いて評価を行う virtual screening (VS) が行われている [2]．このうち，タンパク質や化合物の三次元構造情報を用いた手法 (structure-based virtual screening, SBVS) は，物理化学的に相互作用を考慮することができ，既知の薬剤やタンパク質阻害剤を必要としないため注目されている．SBVS ではタンパク質や化合物の 3 次元構造情報が必要であるが，SBVS が広まるに伴ってデータベースも近年拡充されている．例えば，タンパク質の三次元構造情報のデータベースである Protein Data Bank (PDB) には 2015 年末の時点で 11.4 万件余りの構造が登録されており，2014 年から 2015 年にかけて 20%も増加している [3]．一方，化合物のデータベースについても，例えば ZINC データベース [4] には約 3,400 万件の化合物が登録されている．

SBVS では，タンパク質と化合物との結合親和性を評価するためにタンパク質-化合物ドッキング計算（以下，化合物ドッキング）がよく用いられる [5] が，ドッキング計算は最適化問題であり，例えば最も利用されているドッキングツールである AutoDock Vina[6] は 1 つのタンパク質-化合物ペアの評価に約 500 CPU 秒を要するなど，計算量が大きいという問題を持っている [7]．これは，化合物の内部自由度のために探索空間が広がってしまっていることが原因として挙げられる．この計算量では，1 つのタンパク質へ ZINC 全体の化合物約 3,400 万件を評価しようとすると 500 CPU 年以上もの計算時間を要するため，データベース全体をドッキング計算で評価することは事実上不可能である．

このことから，ドッキング計算を行う前に化合物の選別を行うプレスクリーニング (pre-screening) が一般的に行われている [8]．プレスクリーニング手法には大きく分けて化合物情報に基づいた手法 (ligand-based) とタンパク質構造情報に基づいた手法 (structure-based) の 2 つの手法が存在する [7], [9]．化合物情報に基づく手法とは，標的とするタンパク質への既知の薬剤や阻害剤の情報に基づき，機械学習を用いたり，リピンスキーの法則 [10] のように薬剤化合物の特徴を見つけ，対象の化合物に適用することで化合物を選別するものである [11]．この手法は計算量が少なく，速度の観点からプレスクリーニング手法として広く利用されているが，既知化合物の情報に左右され，構造的・物理化学的に新規な化合物を残すことが難しいという問題がある [7]．これに対し，タンパク質構造情報に基づいた手法としては Glide の HTVS (high-throughput virtual screening) モード [12] や Panther[13] などといった，粗く，高速に化合物ドッキングを行うものが挙げられる．これらは既知化合物に依存しないため新規の化合物を残すことができるが，それでもまだ 3,400 万件存在する ZINC 全体を評価するには 1 CPU 年を要するため，速度が不十分である．

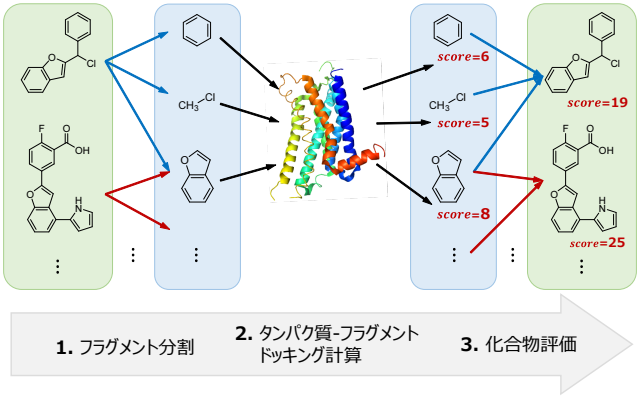

図 1　ESPRESSO の流れ図.

したがって，多少の精度低下を許容して，数千万件もの化合物が登録されているデータベース全体を高速に評価できる，タンパク質構造情報に基づいた手法を開発することが求められている．この際，プレスクリーニング後に Glide SP モード, AutoDock Vina などのツールを利用して化合物ドッキングを行うことから，タンパク質-化合物の結合構造を出力する必要がないということも重要であり，化合物の評価値のみを算出することで高速化を図ることができる．

本研究では，タンパク質構造情報に基づいた超高速なプレスクリーニング手法 ESPRESSO (Extremely Speedy PRE-Screening method with Segmented cOmpounds) を開発した．これは，化合物群を「フラグメント」と呼ぶ内部自由度を持たないような部分構造に分割し，タンパク質-フラグメントドッキング計算（以下，フラグメントドッキング）を行い，得られたフラグメントのスコアから化合物の評価値を算出する，というプロトコルで化合物群を同時に評価するものである．化合物データベースの多くは合成可能な化合物のみを取り扱っているが，そのような化合物の多くが誘導体の関係で成り立っており，結果として類似の部分構造を多く持っている．このことから，化合物群から多数の共通したフラグメントが発生し，超高速なプレスクリーニングを実現する．

2. 提案手法 ESPRESSO

2.1 提案手法のアルゴリズム

ESPRESSO の処理の流れは図 **1** の通りであり，以下の 3 つの手順から成っている．

2.1.1 フラグメント分割

ドッキング計算において，計算量の増大の一因となっているのは化合物の構造変化を引き起こす内部自由度であり，フラグメントに内部自由度を残さなければ計算速度は著しく向上すると考えられる．このことから，フラグメント分割では分割結果であるフラグメントに内部自由度が残らないようにする．フラグメント分割のアルゴリズムは小峰らの手法 [14] を利用する．簡潔にアルゴリズムを示す．

(1) **フラグメント候補の生成**：環構造，二重結合，共役二

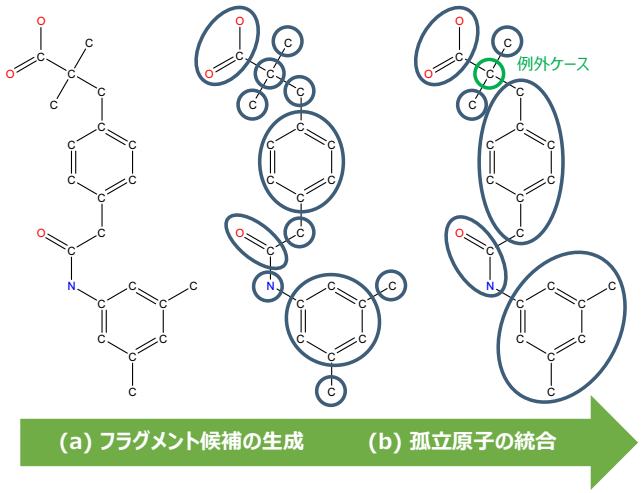

図 2 フラグメント分割例．右図の例外ケースは 4 つのフラグメント候補/孤立原子に隣接しているため統合されない．

重結合（カルボキシ基，アミド基など）でつながった原子同士をひとつのグループにまとめ，フラグメント候補を作成する．

(2) **孤立原子の統合**：フラグメント候補に隣接する孤立原子を統合する．ただし，3 つ以上のフラグメント候補/孤立原子と隣接する孤立原子は例外ケースとして統合せず，そのままとする．

以上のアルゴリズムでまとめられたグループをフラグメントとし，切断面には水素原子を付加する．フラグメント分割例を**図 2** に示す．

化合物データベースには多数の誘導体が含まれており，共通のフラグメントが発生する．共通したフラグメントでは，1 度の計算でスコアが得られるため，ドッキング計算にかかる時間を削減することができる．この計算量削減の効果はデータベースの化合物数に依存するが，一般に化合物数が多いほど計算量の削減効果は大きい．

2.1.2 フラグメントドッキング

各フラグメントについて，タンパク質へのドッキング計算を独立に行い，最良のドッキングスコアをフラグメントのスコアとする．利用するソフトウェアは AutoDock Vina[6] や Glide[12]，GOLD[15] など，スコアを算出するドッキング計算を行うものであれば自由に選ぶことができる．

2.1.3 フラグメントのスコアを利用した化合物評価

得られたフラグメントのスコアから化合物のドッキングスコアを理論的に算出することはできないため，化合物の粗い評価値を算出する．評価値の算出方法は多数の試行を行ったが，本稿では 4 つの計算式について述べる．

(1) フラグメントのスコアの総和 (SUM)

$$\mathrm{SUM} = \sum_f score_f \tag{1}$$

総和は化合物の評価値を算出する最も単純な方法のうちの 1 つである．また，化合物をフラグメントに分割することで結合を切断し，独立にフラグメントドッキングを行うために衝突を考慮しないなど，化合物ドッキングの条件を一部緩和していることに相当することから SUM スコアは化合物のドッキングスコアの緩い上界になると考えられる．

(2) フラグメントのスコアの最良値（MAX）

$$\mathrm{MAX} = \max_f score_f \tag{2}$$

フラグメントのスコアの最良値を取ることも化合物の評価値を算出する最も単純な方法の 1 つである．これは，原子数が化合物よりも大幅に少ないことから化合物のドッキングスコアの下界になることがほとんどである．ただし，化合物がポケットに入りきらないなどの場合にはフラグメントの MAX スコアが化合物ドッキングスコアを上回ることがある．

(3) フラグメントのドッキングスコアの累乗総和（Generalized Sum, GS）

$$\mathrm{GS}_x = \sqrt[x]{\sum_f (score_f)^x} \tag{3}$$

GS_1 は SUM と等価であり，GS_∞ は MAX と漸近的に等価であることから，GS は SUM と MAX の中間をとることができる．本研究では，GS_2 および GS_3 を扱う．

GS は入力されるフラグメントのスコア $score_f$ の値が正であることが必要である．フラグメントドッキングで得られるスコアは一般に負であることから，化合物評価を行うまえにフラグメントのスコアは正負を逆転させ，もし負の値があればその値は 0 とする．

なお，詳しくは 3.2 節で示すが，最も良い予測精度を示したのは GS_3 であったため，ESPRESSO の既定計算式は GS_3 とした．

2.2 データセット

本研究の実験では，DUD-E (Directory of Useful Decoys, Enhanced)[16] を利用した．DUD-E は 102 個の異なるタンパク質とそれらに対応する正例化合物，負例化合物から構成されており，バーチャルスクリーニングやドッキング計算手法のベンチマークデータセットとして広く利用されている．また，ZINC データベース [4] の “all purchasable” サブセットおよび “all boutique” サブセットの計 28,629,602 化合物を利用してプレスクリーニング手法の速度評価を行った．

2.3 実装

ESPRESSO は 2.1.1 節で示したフラグメント分割を C++言語で，2.1.3 節で示した化合物評価を Python

言語で実装しており，GPLv3 ライセンスの元で http://www.bi.cs.titech.ac.jp/espresso/に公開している．

また，フラグメントドッキングには Glide SP および Glide HTVS を利用して実験を行い，比較対象として Glide HTVS を用いた化合物ドッキングを行った．

2.4 計算機環境

本研究では，東京工業大学の TSUBAME 2.5 の Thin ノードを利用した．Thin ノードには 6 コアの Intel Xeon X5670 CPU が 2 個，メモリが 54 GB 搭載されており，ドッキング計算は 12 CPU コアすべてを利用して実行している．

2.5 評価指標

本研究の実験では，計算速度の評価および予測精度の評価を行う．計算速度については，商用ソフトである Glide が 1 ライセンスにつき 1 CPU コアでの計算を許す，という仕様になっているため，CPU 時間での評価を行う．また，予測精度の評価については Enrichment Factor (EF)[17] と呼ばれる指標を用いる．

$$\mathrm{EF}_{x\%} = \frac{\mathrm{Pos}_{x\%}/\mathrm{All}_{x\%}}{\mathrm{Pos}_{100\%}/\mathrm{All}_{100\%}} \tag{4}$$

$\mathrm{All}_{x\%}$ は元の化合物数の x%，$\mathrm{Pos}_{x\%}$ は化合物の評価値が上位 x% である正例化合物の数を示しており，化合物の順位付けによってどれだけ正例が上位 x% に集まったかを示す値となる．本研究では評価指標として $\mathrm{EF}_{1\%}$ および $\mathrm{EF}_{2\%}$ を用いる．

2.6 予測精度の評価方法

本研究で提案する ESPRESSO は，その後に化合物ドッキングが行われることを想定しており，単独で利用することを想定していない．そのため，予測精度は以下に示す手順で評価実験を行う．

(1) 各プレスクリーニング手法を用いて化合物群を 2%, 5%, 10%に削減する
(2) プレスクリーニングで得られた化合物について，Glide SP による化合物ドッキングを行い，化合物群を元の化合物数に対して 1%, 2%まで削減し，$\mathrm{EF}_{1\%}$，$\mathrm{EF}_{2\%}$ を計算する

3. 実験結果

本研究では，ESPRESSO の計算速度と予測精度を評価するため 2 つの実験を行った．それぞれの実験では，従来のプレスクリーニング手法である Glide HTVS による化合物ドッキングも行い，比較対象としている．

3.1 計算速度評価

表 1 に ZINC データベースから取得した 28,629,602 化合物のドッキング計算に要した時間の平均を示す．Glide SP によるフラグメントドッキングを行った ESPRESSO-SP は 2 CPU 日以内に，Glide HTVS によるフラグメントドッキングを行った ESPRESSO-HTVS は 1 CPU 日以内に計算が完了している一方，Glide HTVS による化合物ドッキングには 140 CPU 日以上の時間を要しており，ESPRESSO は最大で約 200 倍の高速化率が得られた．

表 1 ZINC データベースの 28,629,602 化合物を DUD-E の 3 つのターゲットにドッキングしたときの平均計算時間を CPU 時間で示した表．括弧内は Glide HTVS と比較した場合の高速化率を示している．

DUD-E ターゲット	平均計算時間 [CPU hours] ESPRESSO-SP	ESPRESSO-HTVS	Glide HTVS
ACES	42.6(×76.8)	22.8(×143.1)	3268.8
EGFR	38.9(×126.4)	21.5(×229.3)	4925.1
PGH1	41.8(×88.0)	20.9(×175.4)	3674.5

表 2 DUD-E の 102 ターゲットに対する予測精度の平均．太字は ESPRESSO 内で最も予測精度が良かったケースを表す．

プレスクリーニング手法		Enrichment Factors (EF) 2%–1%	5%–1%	10%–1%	5%–2%	10%–2%
ESPRESSO-SP	SUM	4.63	6.73	8.79	3.96	5.43
	MAX	9.09	10.93	11.85	7.49	8.24
	GS_2	7.32	10.12	11.89	6.22	7.71
	GS_3	**9.61**	**12.78**	**15.03**	**8.05**	**9.89**
ESPRESSO-HTVS	SUM	4.53	6.85	8.80	4.05	5.35
	MAX	9.30	9.91	12.25	6.40	8.20
	GS_2	7.08	9.68	11.70	5.77	7.21
	GS_3	9.07	12.10	14.43	7.38	9.26
Glide HTVS		17.85	18.97	19.60	12.50	12.92

(注) a%–b% は，まずプレスクリーニング手法によって化合物を a% まで削減し，次にそれらを Glide SP のドッキング計算で評価した際の $\mathrm{EF}_{b\%}$ を意味する．

3.2 予測精度評価

表 2 に DUD-E 102 ターゲットに対する EF の平均値を示す．化合物の評価値を算出する計算式として GS_3 を用いると最も予測精度が良くなるという結果が得られた．また，SUM, MAX は ESPRESSO-SP と ESPRESSO-HTVS で予測精度がそれほど変わらないが，GS_2 および GS_3 では ESPRESSO-SP を用いたほうが精度が良いという結果も得られている．なお，提案手法 ESPRESSO は Glide HTVS による化合物ドッキングに比べて精度は劣るものの，速度性能のために許容できる範囲である．

4. 考察

4.1 計算量削減の要因

3.1 節より，ESPRESSO は最大約 200 倍の高速性能を達成することがわかった．本実験で用いた ZINC データベースの 28,629,602 化合物はフラグメント分割によって 263,319 フラグメントで表現できることが得られており，

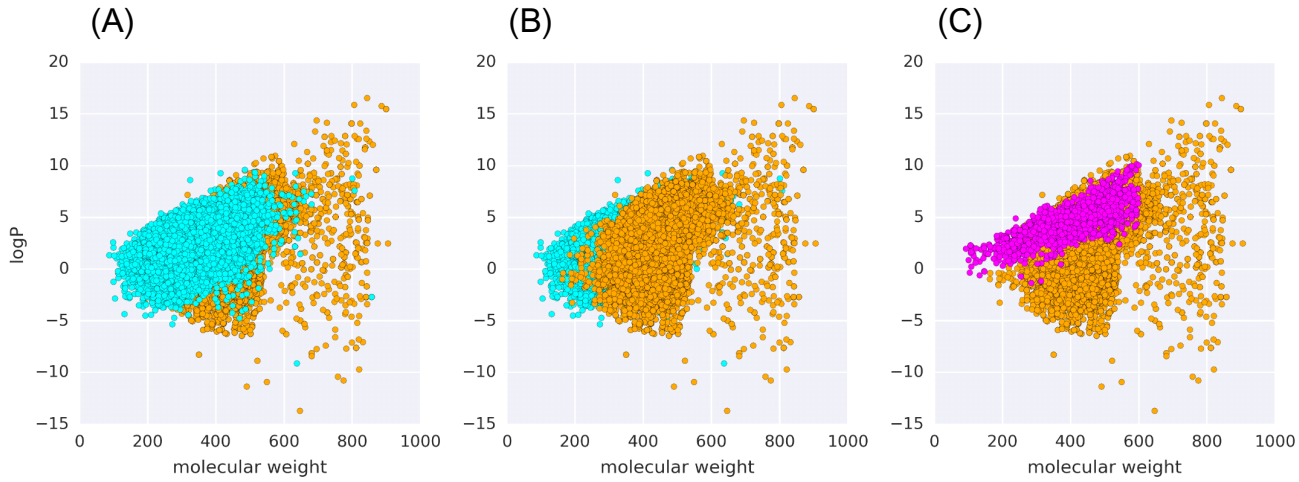

図 4 DUD-E の ACES ターゲットに対する ZINC 化合物のプレスクリーニングを実施した結果を横軸を分子量，縦軸を分配係数 (logP) の散布図で示した図．(A) 水色：ZINC データベースからランダムサンプリングで 0.1%取得した化合物，橙色：ESPRESSO-SP で評価した場合の上位 0.1%の化合物，(B) プロットの重ね方を図 (A) と入れ替えたもの，(C) 紫色：DUD-E に登録されている ACES の正例化合物，橙色：図 (A)，図 (B) と同じ．

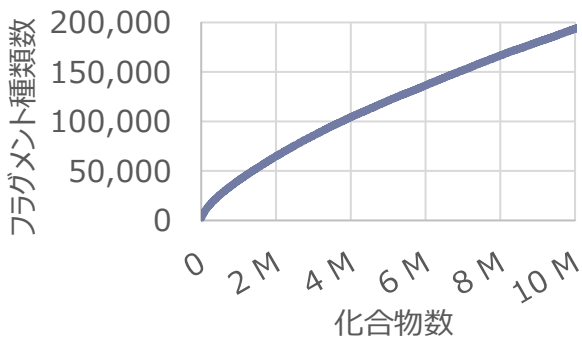

図 3 逐次的にフラグメント分割を行った場合の分割済み化合物数と得られたフラグメント種類数の相関図．ZINC データベースの “drugs now” サブセット 10,639,555 化合物を用いている．

ドッキング計算の回数が 100 分の 1 以下に抑えられている．したがって，内部自由度の削減によるドッキング 1 回の計算時間削減よりもフラグメントの共通化によるドッキング回数の削減の効果が大きい．ドッキング回数の削減は化合物データベースに依存し，例えば ChEMBL (version 21) データベース [18] に登録されている 1,583,897 化合物は 127,360 フラグメントで表現でき，PubChem データベース [19] のうち分子量が 1000 以下の 88,527,810 化合物は 2,082,185 フラグメントで表現することができる．また，**図 3** に示す通り，化合物数が多いほどフラグメント数の増加が鈍るため，一般に 1 つの化合物あたりの計算量はデータベースが大きいほど減少する．

4.2 フラグメントスコアの重み付け

3.2 節より，化合物の評価値の算出には GS_3 が良いことが示された．SUM はすべてのフラグメントを均等に評価するが，この計算式が 4 つの中で最も悪かった．一方で，同様にすべてのフラグメントのスコアを用いるが，より良いスコアに重みをつける GS_3 は良い精度が得られていることから，重要なフラグメントのスコアを重く見ることが重要である．ただし，最良の値のみを利用する MAX よりも GS_3 の精度が高いことから，最も良いフラグメントのスコアのみに着目するだけでは不十分であり，複数のフラグメントのスコアを用いることが必要である．

4.3 プレスクリーニングで得られた化合物の多様性

Drwal ら [7] は既知の化合物情報に基づいたスクリーニングを行うと化合物の構造的多様性が失われることを指摘しており，タンパク質の 3 次元構造情報に基づいた手法のほうが多様性を維持することができることが期待される．しかし，ESPRESSO は構造情報に基づいた手法であるが，フラグメントのスコアから化合物の評価値を求める新しい手法であるため，多様性を保持していることは自明ではない．そこで，DUD-E の ACES ターゲットと ZINC データベースの化合物を使い，ESPRESSO-SP，ガウスカーネルを用いた SVM，Glide HTVS を用いた化合物ドッキングの 3 つの手法を用いて，それぞれ上位 0.1%の化合物を取得した．

ESPRESSO-SP の上位化合物と DUD-E に登録されている正例化合物の物理化学的な値のプロットを**図 4** に，各手法の上位化合物と正例化合物群との Tanimoto 係数の最大値を求め，箱ひげ図で示したものを**図 5** に示す．図 4 より，ESPRESSO は大きな化合物に高いスコアを与える傾向があることがわかる．化合物ドッキングスコアにも同様の傾向があることが指摘されている [20] ため，これはある程度期待に沿った結果といえるが，タンパク質の結合ポケットに入りきらないほど大きな化合物はプレスクリー

Vol.2016-MPS-108 No.18
Vol.2016-BIO-46 No.18
2016/7/5

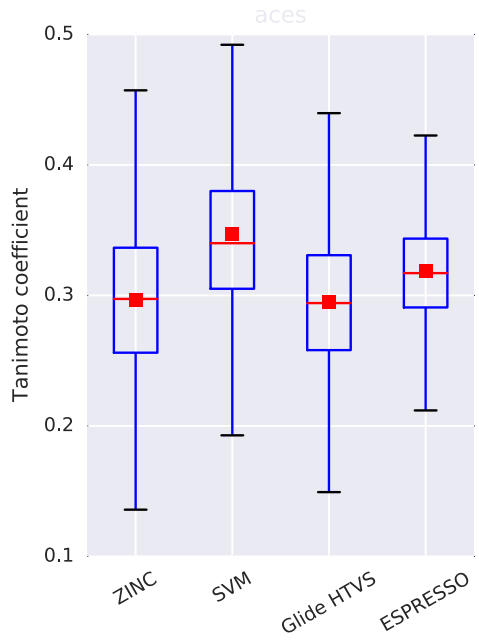

図 5 DUD-E に登録されている ACES の正例化合物に対する Tanimoto 係数の箱ひげ図．四角の点は平均値を表している．左から，ZINC データベースから 0.1%の化合物をランダムサンプリングした場合（ZINC），SVM を用いて 0.1%の化合物をプレスクリーニングした場合（SVM），Glide HTVS による化合物ドッキングで 0.1%の化合物をプレスクリーニングした場合（Glide HTVS），ESPRESSO-SP で 0.1%の化合物をプレスクリーニングした場合（ESPRESSO）を示している．

ニングで除外される必要がある．一方，図 5 より，提案手法は化合物の情報に基づいた SVM よりも既知の薬剤との Tanimoto 係数が低い化合物が多く，多様性が維持できているといえる．

4.4 大きな化合物の除去

図 4 より，ESPRESSO は大きな化合物に高いスコアを与える傾向にあり，タンパク質の結合ポケットに入りきらない化合物を別途除外するなどの考慮が必要であることが判明した．そこで，タンパク質の結合ポケットの体積を計算した上で，体積が一定以上の化合物はそもそも結合部に入りきらないとして除外する方法を実験した．

(1) すべての化合物の体積 V_c を Zhao ら [21] が提案する以下の推定式を用いて計算する．

$$V_c = \sum_{a \in c} V_a - 5.92 N_B - 14.7 R_A - 3.8 R_{NA} \quad (5)$$

V_a は原子のファンデルワールス半径から得られる球体の体積，N_B は化合物の持つ結合本数，R_A は化合物の持つ芳香環の数，R_{NA} は化合物の持つ非芳香環の数である．

(2) タンパク質の結合ポケットの体積 V_p を Sitemap[22] を用いて推定する

(3) $V_c > kV_p$ である化合物を除外する（k は 1 以上の実数値）

k はタンパク質の結合ポケットの柔軟性を示す値である．$k = 1$ のときにはタンパク質の結合ポケットのサイズそのものを閾値として利用するが，タンパク質は構造変化を起こすことから，$k = 1$ は厳しい閾値になることが多い．一般に，k の値が大きいほど偽陽性が発生し，k の値が小さいほど偽陰性が発生するため，ここでは偽陰性を少なくするために $k = 1.5$ を利用して実験を行った．結果として，分子量 750 以上の化合物の多くが除去されたが，既知の薬剤が分子量 600 までに収まっていることを考えると，妥当な結果である．

しかし，大きな化合物に高いスコアを与える ESPRESSO の傾向は小さな薬剤候補化合物の見落としにつながる可能性がある．したがって，体積や分子量を用いた罰則を適用したスコア付け（ligand efficiency[23] など）の考案が必要であると考えられ，今後の課題である．

4.5 ESPRESSO で得られた化合物の例

DUD-E の ACES ターゲットに対して，4.4 節で述べた手法を適用した ESPRESSO-SP を用いたプレスクリーニングを行った場合に，ZINC データベースの化合物中で最も良いスコアを出した化合物は ZINC12181222 であった（**図 6**(A)）．この化合物は分子量 395.5，logP 1.84 であり，薬剤化合物の特徴をまとめたリピンスキーの法則 [10] を満たしているなど，有望な化合物であると言える．

化合物をフラグメント分割した結果を図 6(B) に，それらをタンパク質に独立にフラグメントドッキングした結果を図 6(C) に示す．ESPRESSO は計算量の削減のためにドッキング計算中に衝突を考慮しないので，結果としてフラグメント同士が衝突していることが図 6(C) からわかる．

5. 結論

本研究では，タンパク質構造情報に基づいた超高速なプレスクリーニング手法 ESPRESSO を提案した．この手法はフラグメント分割に基づいており，従来のプレスクリーニング手法である Glide HTVS に比べ，約 2,900 万化合物を最大約 200 倍高速に評価することができることを示した．この速度向上は評価対象の化合物数と正の相関があり，限られた計算資源の中で，登録件数が増え続けている化合物データベース全体を評価することを可能にしている．

化合物の評価値の算出には，予測精度の評価実験の結果に基づき GS_3 を利用しているが，4.3 節で示したような大きな化合物を選択するような偏りが発生している．4.4 節で述べた化合物の体積を利用した方法によって極度に大きな化合物の除去は可能であるが，この手法には精査・改善の余地がある．また，本手法は高速な化合物評価を達成するために予測精度が多少低下しているため，フラグメント

Vol.2016-MPS-108 No.18
Vol.2016-BIO-46 No.18
2016/7/5

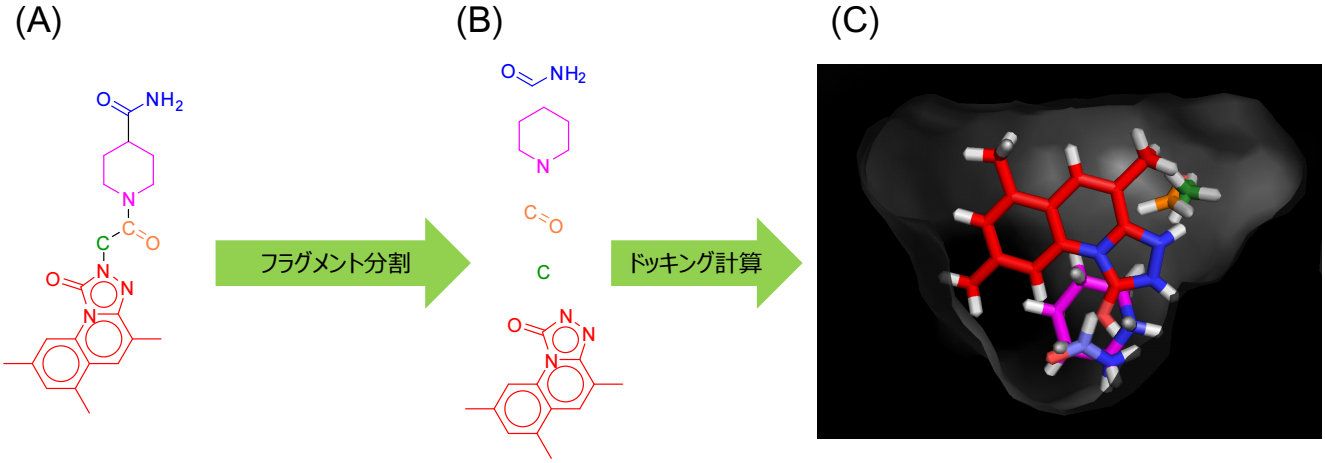

図 6 (A)ESPRESSO-SP を用いた化合物評価で，ACES に対して最も評価値が高かった化合物（ZINC12181222）．(B)ZINC12181222 をフラグメント分割した結果．(C) フラグメントを Glide SP を用いてドッキングした結果の構造． フラグメントの色付けは図 (A)，図 (B)，図 (C) それぞれで対応している．

間の衝突の考慮など，予測精度の改善を検討する必要がある．

謝辞 本研究の一部は文部科学省 博士課程教育リーディングプログラム 東京工業大学「情報生命博士教育院」，JSPS 科研費 基盤研究 (A) (24240044)，および JST CREST 「EBD: 次世代の年ヨッタバイト処理に向けたエクストリームビッグデータの基盤技術」の支援を受けて行われた．

参考文献

[1] Klon A.E. *et al.*: “Finding more needles in the haystack: a simple and efficient method for improving high-throughput docking results”, *J. Med. Chem.* 47, 2743–2749, 2004.
[2] Sliwoski G. *et al.*: “Computational methods in drug discovery”, *Pharmacol. Rev.* 66, 334–395, 2014.
[3] Rose P.W. *et al.*: “The RCSB Protein Data Bank: views of structural biology for basic and applied research and education”, *Nucleic Acids Res.* 43, D345–D356, 2015.
[4] Irwin J.J. *et al.*: “ZINC: a free tool to discover chemistry for biology”, *J. Chem. Inf. Model.* 52, 1757–1768, 2012.
[5] Meng X. *et al.*: “Molecular docking: a powerful approach for structure-based drug discovery”, *Curr. Comput. Aided Drug Des.* 7, 146–157, 2011.
[6] Trott O. *et al.*: “AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient otimization, and multithreading”, *J. Comput. Chem.* 31, 455–461, 2010.
[7] Drwal M.N. *et al.*: “Combination of ligand- and structure-based methods in virtual screening”, *Drug Discov. Today Technol.* 10, e395–e401, 2013.
[8] Kumar A. *et al.*: “Hierarchical virtual screening approaches in small-molecule drug discovery”, *Methods* 71, 26–37, 2015.
[9] Ferreira L.G. *et al.*: “Molecular docking and structure-based drug design strategies”, *Molecules* 20, 13384–13421, 2015.
[10] Lipinski C.A. *et al.*: “Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings”, *Adv. Drug Deliv. Rev.* 23, 3–25, 2001.
[11] Ripphausen P. *et al.*: “State-of-the-art in ligand-based virtual screening”, *Drug Discov. Today* 16, 372–376, 2011.
[12] Friesner R.A. *et al.*: “Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy”, *J. Med. Chem.* 47, 1739–1749, 2004.
[13] Niinivehmas S.P. *et al.*: “Ultrafast protein structure-based virtual screening with Panther”, *J. Comput. Aided Mol. Des.* 29, 989–1006, 2015.
[14] 小峰 駿汰 他: “フラグメント伸長型タンパク質-化合物ドッキングのビームサーチによる高速化”, 情報処理学会研究報告 2015-BIO-42, 1–8, 2015.
[15] Verdonk M.L. *et al.*: “Virtual screening using protein-ligand docking: avoiding artificial enrichment”, *J. Chem. Inf. Comput. Sci.* 44, 793–806, 2003.
[16] Mysinger M.M. *et al.*: “Directory of Useful Decoys, Enhanced (DUD-E): better ligands and decoys for better benchmarking”, *J. Med. Chem.* 55, 6582–6594, 2012.
[17] Hamza A. *et al.*: “Ligand-based virtual screening approach using a new scoring function”, *J. Chem. Inf. Model.* 52, 963–974, 2012.
[18] Bento A.P. *et al.*: “The ChEMBL bioactivity database: an update”, *Nucleic Acids Res.* 42, 1083–1090, 2014.
[19] Kim S. *et al.*: “PubChem substance and compound databases”, *Nucleic Acids Res.* 44, D1202–1213, 2016.
[20] Verdonk M.L. *et al.*: “Virtual screening using protein-ligand docking: avoiding artificial enrichment”, *J. Chem. Inf. Comput. Sci.* 44, 793–806, 2004.
[21] Zhao Y.H. *et al.*: “Fast calculation of van der Waals volume as a sum of atomic and bond contribu-tions and its application to drug compounds”, *J. Org. Chem.* 68, 7368–7373, 2003.
[22] Halgren T.A.: “Identifying and characterizing binding sites and assessing druggability”, *J. Chem. Inf. Model.* 49, 377–389, 2009.
[23] Shultz M.D.: “Setting expectations in molecular optimizations: Strengths and limitations of commonly used composite parameters”, *Bioorg. Med. Chem. Lett.* 23, 5980–5991, 2013.