

#### 1.はじめに

近年、フラグメントに基づくリード創出が注目されている<sup>1,2</sup>。背景には、従来 HTS で使用されてきたライブラリの問題点がある。たとえば、取りこぼしなく化合物探索を行うためには、従来考えられてきたよりも細かい網の目で、化合物空間を網羅しなければならないことが指摘されている<sup>3,4</sup>。また、HTS から得られるヒット化合物が、最適化の余地の少ない分子量の大きな化合物であることが多いことも度々論じられる。これらのことから、*drug-like* な化合物を網羅的に探索しようとするものの限界が認識されるようになり、最適化の余地を残した *Lead-like* な化合物<sup>5</sup>や、フラグメントを探索する手法が求められている。

一方、分子ドッキングに基づくバーチャルスクリーニングは、方法論の進歩により既に実用段階に入っている<sup>6,7</sup>。しかし、結合様式や化合物の順位付けに使用するスコア関数はまだ改善の余地が多く、特に小さな化合物フラグメントの結合様式や活性の有無を正確に予測することは現段階では困難である<sup>8</sup>。また、一般にドッキングスコア関数は大きな化合物ほど上位にランクする傾向があり、これを補正するための方法が幾つか提案されている<sup>9,10</sup>。我々は、化合物フラグメントに基づくバーチャルスクリーニングを実効あるものにするため、スコア関数の規格化とパラメータ最適化の方法を開発した。

#### 2.理論及び方法

##### スコア関数の分子サイズに対する依存性

最初に、ドッキングスコア関数の分子の大きさに

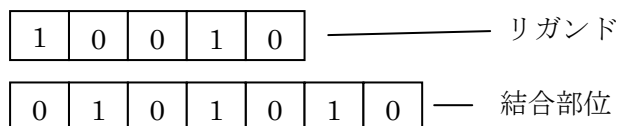


図 1. リガンド-蛋白質相互作用のモデル

対する依存性の本質を理解するために、単純化したモデルを用いる。リガンドを長さ  $n$ 、蛋白質のリガンド結合部位を長さ  $m$  とし、0 又は 1 を要素とする配列で表す(ただし  $m \geq n$ )。これらの間のアライメントが結合様式に対応すると考える(図 1)。リガンドと結合部位の間で、1 と 1 又は 0 と 0 が対応するとき、スコアに -1 が加算されるものとする。このとき、長さ  $n$  のリガンドの種類(配列の数)は  $2^n$  であり、このうち特定の結合部位に対して特定の様式で結合したときにスコアが  $-x$  になるリガンドの数は、二項分布

$$\frac{n!}{x!(n-x)!} \quad (1)$$

で与えられる。この分布は、 $n$  が無限大の極限と相互作用の単位を細かくする極限で正規分布に帰着し、平均、分散が共に  $n$  に比例する。これらの性質は、相互作用のパターンが複雑になったり、強さがまちまちな相互作用が混在する場合でも共通である。そこで、化合物データベースをドッキングして得られるスコアの分布を、平均と分散がそれぞれ化合物の大きさ( $n$ )に対して線形となるような正規分布でモデル化することが、第一段階としては妥当であると考えられる。ここで、 $n$  に対応する化合物の大きさの尺度として、単に大きさを表す分子量や原子数よりも、蛋白質と相互作用しうる要素の数と強さを考慮した指標を用いるのが適当である。そこで、今回はこの指標と

して Andrews のスコアを用いる<sup>11</sup>。Andrews のスコアは、ターゲット横断的に QSAR 解析を行い、化合物が持ちうる活性値を各化学基や原子の数の重み付の和で表したものである。

#### パラメータ最適化手法

従来、経験的なドッキングスコア関数のパラメータ最適化の手法として、既知のタンパク質-リガンド複合体の活性データにスコア関数を一致させる、最小二乗法が多く用いられてきた。

$$\sigma^2 = \frac{1}{N} \sum_i^N (S_i - a_i)^2 \quad (2)$$

ここで  $S_i$  は  $i$  番目の蛋白質-リガンド複合体データに対するスコア値、 $a_i$  は  $i$  番目の複合体のリガンドの活性、 $N$  は用いる複合体データの数。最小二乗法では  $\sigma^2$  を最小化する。しかし、通常用いられるデータには多くの誤差や外れ値が含まれている。例えば、リガンドの結合に伴って蛋白質の 4 次構造が変化する例(ストレプトアビジン-ビオチン複合体<sup>12</sup>等)での構造変化に伴う自由エネルギー変化や、タンパク質のプロトン化状態がリガンドによって異なる例(Hiv-1 プロテアーゼ<sup>13,14</sup>等)でのプロトン化・脱プロトン化に伴う自由エネルギー変化は、現在のドッキングスコア関数には反映されていない。このようなことが実験的に詳細に検討されている複合体はむしろ少数であり、使用するデータからこのような例を除外したり、補正したりすることは難しい。また、異なるアッセイ系で測定された活性値を同等に扱うことから大きな誤差が生じる。このように多くの誤差を含むデータに対してスコア関数を忠実にフィッティングすることは、本来避けるべきである。そのために、ロバスト推定と呼ばれる、外れ値を許す最適化手法が用いられることも多い。ここでは、前節のスコア分布のモデルを用い、フラグメントの活性を予測するのに適した新しいスコア関数最適化手法を導入する。

$$\frac{1}{N} \sum_i^N (S_i - a_i)^2 = \sigma_0^2 \quad (3)$$

$$\bar{Z} = \frac{1}{N_{Active}} \sum_{i, Active}^N \frac{S_i - (au_i + b)}{\sqrt{cu_i + d}} \quad (4)$$

ここで、 $\sigma_0$  は定数とし、(3)をパラメータ最適化の束縛条件とする。 $\bar{Z}$  は活性化合物の、スコア分布中における位置を表す  $Z$  スコアの平均値で、今回の最適化手法ではこれを最小化する。 $N_{Active}$  は  $Z$  スコア最適化に用いる活性化合物の数、 $u_i$  は  $i$  番目の活性化合物の Andrews のスコア。 $au_i + b$  と  $cu_i + d$  はそれぞれ、前節の分布モデルに従ったスコア分布の平均と分散で、パラメータ  $a, b, c, d$  はデータベース中の化合物のドッキング結果に基づいて、以下のような条件から決める。

$$P = \prod_{i, DB}^{N_{DB}} \frac{1}{\sqrt{2\pi(cu_i + d)}} \exp \left[ -\frac{\{S_i - (au_i + b)\}^2}{2(cu_i + d)} \right] \quad (5)$$

$$\frac{\partial P}{\partial a} = \frac{\partial P}{\partial b} = \frac{\partial P}{\partial c} = \frac{\partial P}{\partial d} = 0 \quad (6)$$

$P$  は分布パラメータ  $a, b, c, d$  を仮定したときの、データベース中の化合物群のスコアの出現確率を表し、式(6)は  $P$  を最大化するように  $a, b, c, d$  を求める最尤法の条件である。式(3)~(5)に現れる  $S_i$  がそれぞれスコアパラメータを含んでいるので、これら全ての式を連立させて最適化を行う。計算手法としては、ラグランジュの未定乗数法を用い、式(3)と式(6)を束縛条件とし、式(4)の停留値を求める。

#### 最適化に使用した複合体データ

スコア-活性相関の条件式(3)には、以下の 4 つのグループに属する 40 複合体を用いた。

- ・ HIV プロテアーゼ
- ・ トリプシン、トロンビン
- ・ ノイラミニダーゼ
- ・ サーモライシン

$Z$  スコア最適化には、以下の 4 つのタンパク質と既知リガンドとの複合体と、市販の drug-like な化合物から多様性を考慮して抽出した 1000 化合物をドッキングした結果を用いた。

- ・ プロテインチロシンフォスファターゼ 1B
- ・ ウロキナーゼ
- ・ メチル基転移酵素 ErmC'
- ・ ストロメライシン (MMP-3)

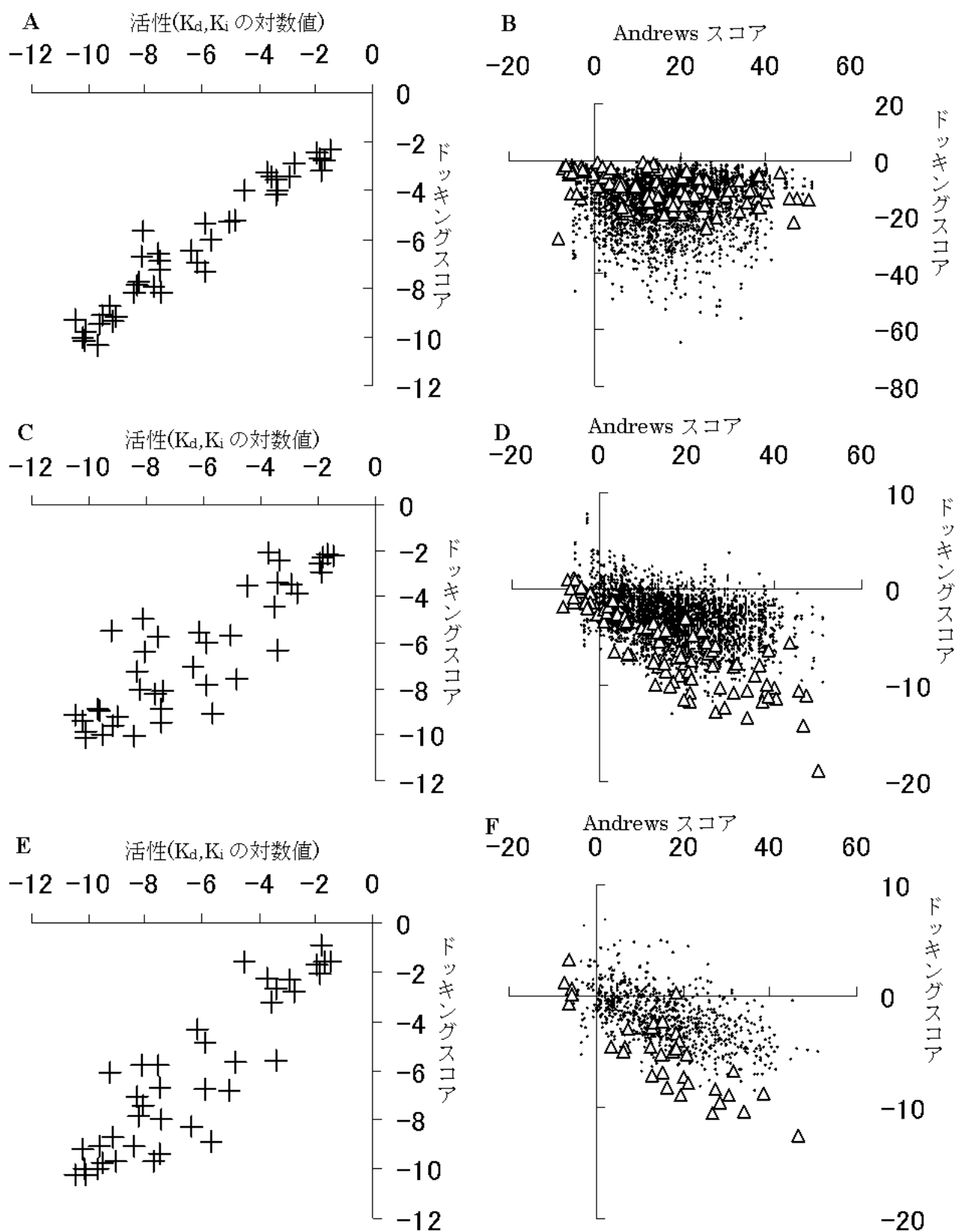


図 2 パラメータ最適化

+ : スコア-活性相関の最適化に使用した化合物。Δ : Z スコア最適化に使用した活性化合物。• : Z スコア最適化に使用したデータベース化合物。

A, B : 活性値に対して最小二乗法でフィッティングした結果。C, D : ラグランジュの未定乗数法でスコア活性相関と Z スコアを最適化した結果。E, F : ウロキナーゼのデータを除いて行ったバリデーションの結果。

活性化化合物としては、既知の阻害剤に加えて、NMR 又は X-線結晶解析によって同定された弱い活性を持つフラグメントを用いた<sup>15-21</sup>。

### スコア関数形

ドッキングスコアには、ファン・デル・ワールス項、水素結合項、簡略化した GB/SA 項<sup>22</sup>、回転結合数を含む関数形を用いた。水素結合項には、一般的な強い水素結合の他に、ハロゲンや $\pi$ 電子が関わる弱い水素結合も含め、水素結合距離などのパラメータは文献<sup>23,24</sup>に記載された実験データから決めた。

### 3. 結果及び考察

図 2 に、パラメータ最適化の結果を示す。A,B は、40 個の既知複合体データを用いて、最小二乗法でスコア活性相関だけを最適化したもので、相関係数 0.97 の高い相関が得られた(図 2A)。しかし、データベース化合物と活性化化合物との分離は悪かった(図 2B)。特に、比較的小さい分子について、False-Positive が多数現れている。これに対して C,D は、ラグランジュの未定乗数法を用いて、スコア-活性相関と Z-スコアの両方を最適化した結果を示している。相関は悪くなっているが、データベース化合物と活性化化合物との分離は大幅に改善されている。また、E,F は Z-スコア最適化の際にウロキナーゼのデータを除いて行ったバリデーションの結果を示している。F はこの結果得られたパラメータを用いて、ウロキナーゼに対して活性を持つ化合物とデータベース化合物との分離の様子を示したもので、D と同様の傾向を持つことからこの方法の安定性が示された。この方法で求められる各化合物の Z スコア

$$Z_i = \frac{S_i - (au_i + b)}{\sqrt{cu_i + d}}$$

を化合物選択の基準にすると、同程度の Andrews のスコアを持つ化合物群の中で、特にあるターゲットに対して特異的に結合する化合物を選ぶことができる。そこでこの Z スコアを特異性スコアと呼ぶことにする。

この方法の利点として、以下のような点が挙げられる。

- ・ 分子量が小さく最適化の余地がある Lead-like な化合物や、フラグメントの探索に適している。
- ・ 活性値に対してだけフィッティングする従来の手法に比べて、活性化化合物と不活性化化合物の分離というバーチャルスクリーニングの目的に即した関数となり、False-Positive が減少する。
- ・ 誤差を含む活性データに対して過度にフィッティングすることが避けられる。

フラグメントのドッキングと、それに基づいたリード設計の検証例についても報告する。

### 4. 参考文献

- [1] D.A.Erlanson, *et al. J. Med. Chem.* **47**:3463-3482(2004)
- [2] D.C.Ress, *et al. Nat. Rev. Drug. Discov.* **3**:660-672(2004)
- [3] Y.C.Martin, *et al. J. Med. Chem.* **45**:4350-4358(2002)
- [4] R.Nilakantan, D.S.Nunn *Drug Discov. Today* **8**:668-672(2003)
- [5] T.I.Oprea, *et al. J. Chem. Info. Comp. Sci.* **41**:1308-(2001)
- [6] J.C.Alvarez *Curr. Opin. Chem. Biol.* **8**:365-370(2004)
- [7] 高橋 ファルマシア 2005 年 2 月号
- [8] F.Sirockin, *et al. J. Am. Chem. Soc.* **124**:11073-11084(2002)
- [9] Y.Pan, *et al. J. Chem. Info. Comput. Sci.* **43**:267-272(2003)
- [10] G.P.A.Vigers, J.P.Rizzi *J. Med. Chem.* **47**:80-89(2004)
- [11] P.R.Andrews, *et al. J. Med. Chem.* **27**:1648-1657(1984)
- [12] P.C.Weber, *et al. Science* **243**:85-88(1989)
- [13] T.Yamazaki, *et al. J. Am. Chem. Soc.* **116**:10791-10792(1994)
- [14] Y-X.Wang, *et al. Biochemistry* **35**:9945-9950(1996)
- [15] B.G.Szczepankiewicz, *et al. J. Am. Chem. Soc.* **125**:4087-4096(2003)
- [16] G.Liu, Z, *et al. J. Med. Chem.* **46**:3437-3440(2003)
- [17] G.Liu, *et al. J. Med. Chem.* **46**:4232-4235(2003)
- [18] P.J.Hajduk, *et al. J. Am. Chem. Soc.* **119**:5818-5827(1997)
- [19] E.T.Olejniczak, *et al. J. Am. Chem. Soc.* **119**:5828-5832(1997)
- [20] P.J.Hajduk, *et al. J. Med. Chem.* **42**:3852-3859(1999)
- [21] V.L.Nienaber, *et al. Nat. Biotech.* **18**:1105-1108(2000)
- [22] 高橋、米田、古谷 第 30 回構造活性相関シンポジウム(2002)
- [23] G.C.Pimentel, A.L.McClellan "The Hydrogen Bond"(1960)
- [24] G.R.Desiraju, T.Steiner "The Weak Hydrogen Bond"(1999)