創薬の効率を飛躍的に高めた化合物スクリーニング計算

―― 3次元構造の化合物データベースの開発 ――

福西 快文*[1]、杉原 裕介[2]、三上 義明[3]、酒井 広太[2]、楠戸 寛[3]、中村 春木[4]

毎年、医薬品探索向けに数百万種類の化合物が、それらの構造式のカタログとともに販売される。我々はこれら構造式から3次元構造化合物データベースを作成するソフトウェアを開発し、2004年以降、化合物データベースの構築・配布を行ってきた。また、多数の蛋白質と、これら化合物をドッキングさせた結果もデータベース化して配布している。これらのデータベースをバーチャルスクリーニングに用いることで、我々は複数の標的蛋白質で高い確率で活性化合物を発見してきた。

キーワード：化合物データベース、myPresto、バーチャルスクリーニング、in-silico drug screening、化合物ライブラリー

Advanced in-silico drug screening to achieve high hit ratio

– Development of 3D-compound database –

Yoshifumi Fukunishi *[1], Yuusuke Sugihara[2], Yoshiaki Mikami[3],
Kohta Sakai[2], Hiroshi Kusudo[3]and Haruki Nakamura[4]

Every year, several millions compounds for drug screening have been released by many vendors in the world, however, the structural information released on these compounds is limited to 2D. We have developed a software system to generate a database of 3D structures of these compounds and have distributed our database. We have also developed a database of protein-compound docking scores for 180 proteins for these millions compounds. Based on these databases, we have found new active compounds for many drug targets.

Keywords：Chemical compound database, myPresto, virtual screening, in-silico drug screening, compound library

1　はじめに

ポストゲノム時代の主たる目標の1つは創薬の革新であるが、遺伝子解析の爆発的な進歩に比べて創薬プロセスには困難がつきまとい、期待された成果が上がっていない。その中で、計算機薬物スクリーニング（in-silico ないしバーチャルスクリーニング（以後 VS という））は、創薬プロセス効率化の1つの道と考えられている。VS は、医薬品の種となる分子を、既にある分子の中から選ぶ計算である。したがって、VS には、計算で扱うことのできる、分子構造の3次元化された化合物のデータベース（以後化合物 DB という）が必須である。化合物 DB には、既に海外製品があるが、価格、品質、成果物の扱いに問題があるため自作した。手法は第4章に述べるが、化学情報学を用いたデータ重複の除去、分子力場法による3次元構造の作成と量子化学計算による原子電荷計算により化合物 DB を作成した。また、あらかじめ用意した多数の蛋白質と化合物との結合エネルギーを推算した新規な DB を開発した。これらの利用により創薬標的蛋白質に対して高い確率で活性化合物を予測できるようになった。

1 産業技術総合研究所　バイオメディシナル情報研究センター　〒135-0064 江東区青海 2-41-6、2 株式会社 富士通九州システムエンジニアリング PLM ソリューション統括部ライフ・サイエンスシステム部　〒261-8588 千葉市美浜区中瀬 1-9-3 富士通幕張システムラボラトリ、3 株式会社 日立東日本ソリューションズ 計算科学ソリューション部　〒210-0007 川崎市川崎区駅前本町 12-1、4 大阪大学 蛋白質研究所 〒565-0871 吹田市山田丘 3-2
1. Biomedicinal Information Research Center, AIST　Aomi 2-41-6, Koto-ku 135-0064, Japan ＊E-mail : y-fukunishi@aist.go.jp, 2. FUJITSU KYUSHU SYSTEM ENGINEERING LIMITED Life Science System Dept. PLM Solution Div.,　Nakase 1-9-3, Mihama-ku, Chiba 261-8588, Japan, 3. Hitachi East Japan Solutions, Ltd.,　Ekimaehonchou 12-1, Kawasaki-ku, Kawasaki 210-0007, Japan, 4. Institute for Protein Research, Osaka University,　Yamadaoka 3-2, Suita 565-0871, Japan

Received original manuscript October 28,2008, Revisions received December 9,2008, Accepted December 9, 2008

2　目標

目標は、毎年市販される数百万種類の化合物をもとにして、VS で用いることができる化合物 DB を短期間に作成し、すみやかに医薬品産業界で利用可能にすることである。世界の主たる数十の試薬ベンダーは 2 次元の分子構造を記載した電子ファイルを試薬カタログとして配布しているが、VS では、2 次元の化学構造式ではなく、3 次元の立体的な分子構造が必要である。したがって、カタログ上の数百万種類の 2 次元構造式から 3 次元構造を作成し、それらをデータベース化して配布することとした。

3　価値

医薬品の開発は、いかなる手法であれ、最初は化合物データベースから標的蛋白質に結合しうるヒット化合物を探索することから始まる。現代的な創薬では、計算機を用いて化合物データベースを探索することは必須である。ここにはいくつかの問題があった。

（1）VS 用化合物 DB は、海外ソフトウェア会社により 1980 年ごろより医薬品メーカー向けに開発・市販されてきたが、年間 1 ライセンス当たり 400 - 600 万円と高価である [1]。

（2）化合物 DB を作成するためのソフト類も海外ソフトウェア会社より市販されている [2]。我々も VS を行うため、高価な化合物 DB 作成ソフトなどを使ってみたが、品質に問題があり、しばしば誤った 3 次元分子構造を提示する、あるいは水素原子の付加に間違い・存在確率が低い構造の生成があるなどの問題があった。

（3）化合物 DB を作成するための市販ソフトには、使用許諾条件上、生成したデータ を他者に配布できない。

後に述べるように、我々は化合物 DB を元にして、蛋白質－化合物相互作用行列という新しい VS 用 DB を作成し、配布することが使命である。しかし、市販ソフトを用いると、この目的が達成できない。もし、化合物データ生成ソフトを自作し、さらに化合物 DB も自前で作成すれば、上記の問題が解決される。これらを配布すれば、高価なライセンス料を支払えない中小企業・アカデミック研究者に VS の利用を促し、大企業に対しても新しい高度な VS 手法を普及できるなどの経済的・技術的効果が見込まれる。

4　プロセス

4.1　全体像

全体の開発プロセスを以下のように設定した。約 10 ステップある（図 1）。試薬ベンダーから提供される 2 次元 SD ファイル[用語 1] には化合物の重複があるため、まず、これを除外する（例えばどのベンダーでもメタノールは売っている）。2 次元構造式では通常、水素原子（H 原子）が省略されているため H 原子を付加する。全ての原子に、原子間の距離や結合角などのパラメータを割り当てる。この情報をもとに 2 次元構造座標から 3 次元構造座標を生成し、光学異性体があれば光学異性体も発生させる。量子化学計算により原子電荷を計算し、等価な原子が等価な電荷を持つようにする。こうして生成した 3 次元データはリレーショナルデータベースに収録する。各ステップにつき市販ソフトが存在する場合が多く、それらの特許を回避するようにソフト開発を進めた。以下に各ステップについて述べる。

4.2　大量のデータの扱い

データは数が多いと取り扱いに困難をきたす。数百万の化合物情報は、1 ファイルに格納するとファイルサイズが大きくなりすぎて計算機で扱えなくなる。1 ファイルに 1 化合物を記載すると、計算機システムの制約上、数百万のファイルを 1 つのフォルダーに置くことができない。そこで、1 ファイルに 1 化合物を記載し、1 フォルダーに 1 万化合物程度を置き、このフォルダーを数百用意するという階層構造で数百万の化合物情報を扱うことにした。できあがりの化合物 DB は 64bit アーキテクチャー上で 1 つのリレーショナルデータベースとして保存できた。

4.3　分子の重複を除く：化合物同一性の判断

2 つの分子が同一か異なるかを判定する必要がある。400 万個の分子の同一性を判断する作業は 400 万 x 400 万回に達するため、我々は下記に述べる高速な数段階での同一性判断法を開発した。また、速度を優先するために判断の精度を一定程度落とすことを許容した。なぜなら実際に購入した市販化合物には、構造式の同定の誤りや、品質管理の落ち度により実際の構造と異なる場合が数%ある。したがって、数学的厳密さを過度に追求しても意味がないからである。

4.3.1　擬似分子量による化学組成式の同一性判断

化学組成式は、分子に含まれる元素の種類と数を表記したものであり、メタノール（CH_3-OH）ならば $C_1O_1H_4$ と

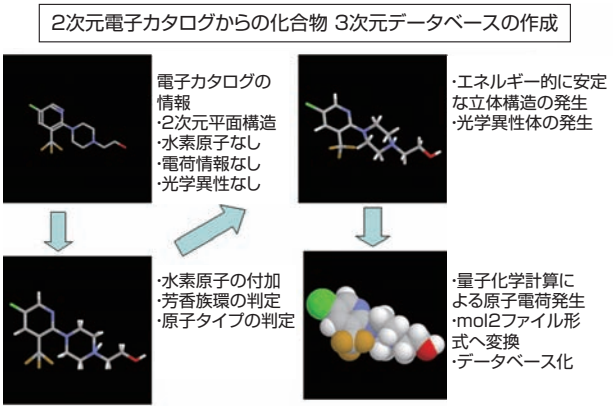

図1　化合物立体構造の作成手順

なる。化学組成式の比較は速い方法である。組成が違えば、それ以上の同一性判断をする必要はない。組成式も、文字列比較では時間がかかりすぎる。 我々は、各元素について原子量を小数点以下 3 桁まで、近似的に末尾がゼロでない数字を入れて分子量を計算し、1 分子あたり 6 桁の数字を割り振る手法を開発した。化学組成式の文字列比較をしなくても、分子量を 1 回の計算で比較することにより実用上、組成式の比較をほぼ間違いなく済ますことができた。

4.3.2　グラフ不変量による分子トポロジーの同一性判断

化学組成式が一致しても、構造式が異なる場合がある。分子のグラフを比較するにはグラフを重ね合わせればいいわけだが、分子グラフの重ね合わせ計算は計算時間が原子の数の多項式で記述できない NP（Non-deterministic Polynomial）完全問題である。一般に、計算時間が多項式時間の場合は高速なアルゴリズムが存在するが、NP 完全の場合は効率的アルゴリズムが存在せず、非常に時間がかかる [3]。そこで我々は、分子の結合行列 M（原子 i と j が結合しているなら、M（i,j） = 1、結合がなければ M（i,j） = 0）を用いることで分子のトポロジーを比較する方法を開発した（図 2）。

分子のグラフにおいて、原子番号の順番は意味がないのでグラフの不変量を計算すればいいわけだが、結合行列はエルミート行列なので、行列固有値を求めればこれが不変量となる。グラフ不変量としては細谷インデックスなどが知られているが計算しにくい [4]。固有値計算は原子数Nに対してN^3の計算量ですむので実用的である。行列の次元を半分程度に減らすために H 原子を除き、原子の種類を反映させるため、対角項には原子の原子番号を代入した。

4.3.3　幾何異性体の判別

4.3.2 でグラフのトポロジーをほぼ正確に判別できるようになったが、シス／トランスなどの幾何異性体を判別できない。そこで我々は、幾何異性体を判別できるグラフ不変量を開発した。2 重結合している原子 i、j について、4 本の結合の先にある部分グラフ断片の部分グラフ行列最大固有値の順に、1、2（1’、2’）と番号を打つ（図 3）。そして 1 → 2 ベクトルと、1’ → 2’ ベクトルが平行なら i-j 行列要素は -2、逆平行なら +2 とすることにより、全体のグラフ行列の固有値から幾何異性体を判別できるようにした。

4.4　水素原子の付加

2 次元構造式の C、N、O、S などの原子に対し、結合次数から不足する H 原子の数を予測し、H 原子を付加すべき原子と、それに隣接する原子との位置関係から、もっともらしい H 原子の座標を計算して分子に加えることにした。H 原子を付加するソフトウェアは、babel [5]/openbabel[6] など各種あるが、H 付加が必ずしも正確でない。我々は、様々な官能基の H 化状態を調査し、真空中及び水中（pH 7.0 近傍）で支配的イオン形をとる H 付加状態を生成するようにした。分子全体の正確なイオン形の予測は困難なので、各官能基ごとに代表的イオン形を適用した。2 次元化学構造は、あくまで模式図であるので、原子間距離は 1 Åであったり 10 Åであったりする。そこで、化学結合の平均距離が 1.5 Åになるようにスケールした。

4.5　力場パラメータの付加

分子の 2 次元構造式から 3 次元構造を生成することは、分子力場によって行うことにした。我々の化合物 DB は、立体構造作成のために General AM BER force field（GAFF）を用いている [7]。開発当初（現在もだが）、ほとんど全ての分子に対して、GAFF のパラメータが存在せず、分子構造が決められなかった。そこで我々は、結晶構造データベース CSD[8] と、手作業で作成した 660 種類の分子を第一原理量子力学計算で構造最適化計算して正しい分子構造情報を得て、原子タイプ、力場パラメータ、パラメータのないとき全ての原子にパラメータが割り振られる仕掛けの追加を行い、99.9 %以上の分子を扱えるようにした。また、力場パラメータの整備に加え、一般の化合物に力場パラメータを割り当てるソフトウェア tplgeneL を開発した。なお、tplgeneL は、酵素の研究のために、化学反応

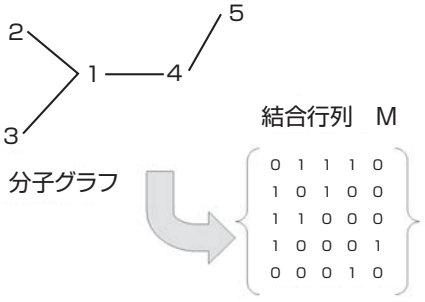

図2　分子グラフから結合行列Mの作成

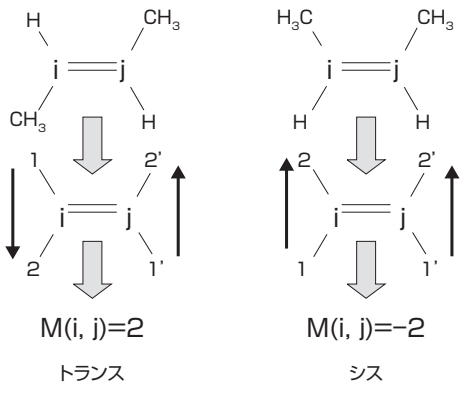

図3　幾何異性体の判別
→は、部分グラフの固有値の順を示す

遷移状態にもパラメータを割り当てる能力がある。

4.6　3次元構造の生成

分子に力場パラメータが与えられると、3 次元分子構造を生成することができる。我々は、分子動力学シミュレーションソフトウェア cosgene[9] を既に開発しており、cosgene によるエネルギー最適化で 3 次元構造を生成した。初期座標にランダムな変位を加えないと、（X、Y）座標データのみの 2 次元構造式のままでは Z 軸方向の力が発生せず、3 次元構造は発生できない。生成した 3 次元構造は構造の妥当性（原子間距離や結合角度など）をチェックするソフトにかけ、歪んだ構造を生成した場合は初期座標を作成しなおし 3 次元構造の生成を再試行した。

4.7　光学異性体の判定と、異性体の発生

炭素原子など結合が 4 本ある原子のそれぞれの化学結合に異なる分子断片が結合している場合、その中心原子が光学活性中心となる。したがって、中心原子に結合する 4 つの結合の先の分子断片の同一性判断が必要になる。我々の開発した手法では、中心原子の結合を切断し、その先の分子断片の同一性比較を 3.2 節と同様の手法で行う。中心原子が環に含まれる場合はやや複雑になるが、ほぼ同様の手法で行った。光学活性中心が 1 つならば、各原子の座標（X, Y, Z）を（X, Y, -Z）とすることで鏡像体を生成できる。光学活性中心が 2 つ以上ある場合は、結合の付け替えをする必要があり、新たに開発したソフトウェア confgeneC でこれを行った。

4.8　量子化学計算による原子電荷の計算

量子化学計算では、分子構造に加えて、分子の電子スピンと電荷が必要である。創薬に用いる分子はラジカルであってはならず、磁性をもつ分子もまれなので、分子のスピンは 0 の閉殻系とした。分子の電荷は、系が安定となる電荷を化学結合から自動計算する手法を開発した。分子全体の電荷は各原子の形式電荷の和とした。たとえば、炭素原子の形式電荷は、化学結合の本数の和が 4 であれば 0、3 本であれば +1 とする。窒素の電荷は、化学結合の本数の和が 4 であれば +1、3 本であれば 0 とする。酸素の電荷は、化学結合の本数の和が 2 であれば 0、1 本であれば -1 とする。こうして得られた形式電荷を分子全体で和をとって、分子の電荷とした。

原子電荷の計算方法にはいろいろある。Gasteiger 法 [10] では、原子に電気陰性度を振って、有機電子論にのっとって、原子同士が互いの電子を引っ張り合い、平衡状態となる電子分布を求める。荒っぽい見積もり方法で、たいていの分子で 1 秒未満で計算できる。半経験的量子化学計算では MOPAC [11] の AM1 モデルと PM3 モデル（最近では PM7）が有名である。PM3 モデルは分子の生成熱を再現するように有効ハミルトニアンをパラメータフィットした優れた手法だが、アミド結合など医薬品に普通に見られる構造を正しく計算できない。AM1 モデルは同じく有効ハミルトニアンをパラメータフィットした手法で、生成熱の予測は不正確だが、アミドなどの構造は正しく計算される。ただし、窒素を含む環構造では、場合によって不正確な場合がある。計算時間は、分子構造を固定した場合、通常、数秒－数十秒で済み、原子の大きさ N に対して計算量は N^3 に比例する程度である。電荷の精度はかなり高い。第一原理量子化学計算では、一般には RHF/6-31G* による波動関数の計算と RESP による部分原子電荷計算法が用いられる [12]。相当正しい電荷を与えるが、分子構造を固定した場合、通常、計算時間は数分－数十分かかり、原子の大きさ N に対し、計算量は N^4 に比例する。

原子電荷は、蛋白質－化合物ドッキングで正しい計算ができなければ意味がない。そこで、132 種類の蛋白質－化合物複合体のドッキング計算を、我々の蛋白質－化合物ドッキングソフト[用語2] sievgene で行った [13]。その結果、RHF/6-31G* では正しい構造が 56 %の確率で得られ（精度 2 Å）、MOPAC AM1 電荷では 2－3 %劣る程度、Gasteiger 電荷でも 5 %劣る程度であった。1 万化合物程度の小規模な薬物スクリーニング実験もシクロオキシゲナーゼ 2、サーモライシンなど数標的で行ってみたが、ヒット率は電荷が正確なほど良いが、Gasteiger 電荷でも数 %劣る程度であった。

数百万分子の原子電荷計算をしなければならないので、高速で計算できるに越したことはないし、RHF/6-31G* ほどの精度にこだわる必要がないことは分かったが、化合物 DB は全ての基礎となるため、MOPAC AM1 電荷を採用することにした。なお、通常 MOPAC は MOPAC 専用の入力形式を必要とするが、我々は MOPAC を改良して、創薬分野において化合物を表現する標準的な mol2 ファイル形式で入出力できるようにした。このための MOPAC の改良用パッチファイルも無償で提供している。

4.9　等価原子の判定

メチル基における 3 つの H 原子は、化学的に等価で同じ電荷を持つようにしなければならない。原子の等価性の判断は原子電荷の計算に必要である。

任意の原子 i と原子 j とが等価であるという意味を次のように考える。原子 i と原子 j が、i=j のとき等価であることは自明である。そうでないとき、原子 i と原子 j が直接結合していないならば、原子 i に結合する全ての原子が原子 j に結合する全ての原子と等価であること、原子 i と原子 j が結合しているときは、それ以外の全ての結合する原子が互いに等価であることである。

等価原子の判定の方法は、以下の通りである。任意の原

子 i と j を選び、i と j に「既に訪問した」という印をつける。原子が「等価」なら、「既に訪問した」と印を残す。

i=j なら「等価」である。等価である場合、それ以上の検証を行わない。

i、j の結合が 1 本で、i、j がともに既に訪問した原子に結合し、元素記号が同じなら「等価」である。

i に結合する原子 m_i と j に結合する原子 m_j の全てに対し、m_i、m_j に「既に訪問した」という印を仮に付して上記の判定を行う。m_i, m_j が同じでなければ、m_i、m_j の「訪問」の印を解除する。すべての m_i/m_j が「等価」と判定されれば、i と j は「等価」だとする。

原子 i、j から出発して等価性の判定が行われる原子を灰色の○で示し、判定が終了する原子を●で示した（図 4）。i と j を出発した探索がぶつかるところ（黒丸）まで調べれば良く、グラフ全体を検索しなくても良い。

4.10　データベースへの格納とファイルのダウンロード

化合物 DB の構造はリレーショナルデータベースとなっており、スキーマには、化合物 mol2 ファイルの情報（原子名、3 次元座標、原子電荷、化学結合次数など）に加え、分子量、MOPAC AM1 モデルでの HOMO/LUMO エネルギー、GBSA モデルで計算した溶媒和自由エネルギー、1 原子当たりの溶媒和自由エネルギーなどを記載した。1 原子当たりの溶媒和自由エネルギーは化合物のケミカルスペース（化合物空間）での位置を示すのに有効な情報であり、DB としては、その多様性（収集された化合物がどれだけ多様であるか）を示す指標として用いられる [14]。

化合物 DB からは、化合物を mol2 ファイル形式でダウンロードすることができる。

4.11　蛋白質－化合物相互作用行列の計算

化合物 DB に収録した化合物ライブラリーに対し、標的蛋白質以外に多数の蛋白質を用意し、総当り式にドッキング計算を行い、蛋白質-化合物相互作用行列を作成してデータベース化した。この DB は以下に述べる我々の開発した薬物

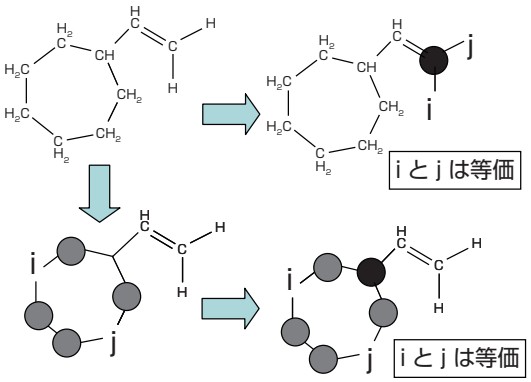

図4　等価原子の判定
●印は、訪問した原子を表す。

スクリーニング手法:multiple target screening（MTS）法 [15]、docking score index（DSI）法 [14] の基礎 DB となっており、我々の VS に欠かせない資源である（図 5）。

通常の VS で、標的蛋白質に結合する化合物をドッキングスコア[用語 3]（スコア）の強い順に選択しても、そのヒット率は低い。標的蛋白質に対し強いスコアを示す化合物を選択すると、その化合物は、他の蛋白質に対しても強いスコアを示して標的蛋白質に対して選択的に強い結合性を示しているわけではないことがしばしば見られる。MTS 法では、逆に、1 つの化合物に着目し、それがどの蛋白質に結合するのかを調べ、標的蛋白質に最も強く結合する化合物をヒット化合物の候補とする。

蛋白質-化合物相互作用行列を利用すると、スコアの精度を改善することもできる。類似性の高い蛋白質に対して同一化合物の結合自由エネルギーは、近い値をとると考えられる。詳細は文献に譲る [16] が、蛋白質の類似度に応じて、重みつきのスコアの平均を取ることで、スコアの誤差を低減することができ、具体的には、下記の式でスコアを補正した。

$$s^{new\,i}{}_a = \frac{\sum_b s_b^{\,i} R_a^{\,b}}{\sum_b R_a^{\,b}} \qquad (1)$$

ここで $s^{new\,i}{}_a$、$s_b^{\,i}$、$R_a^{\,b}$ は、それぞれ新しく定義された蛋白質 a と化合物 i のスコア、蛋白質 b と化合物 i のスコア、蛋白質 a と b の相関係数である。

また、既知の活性化合物が存在する場合、既知活性化合物が優先的に予測されるようにスコアを補正することもできる。下記の式のように、補正後のスコアをスコアの線形結合で記述し、モンテカルロ計算によってデータベースエンリッチメント[用語 4]が最大化されるように係数 $M_a^{\,b}$ を決定した。

$$s^{new\,i}{}_a = \sum_b s_b^{\,i} M_a^{\,b} \qquad (2)$$

COX-2、HIV プロテアーゼ 1 など 12 種類の標的蛋白質に対して MTS 法を適用した結果、化合物ライブラリーから予測上位 1 %の化合物を採択したとき、ランダムスクリーニングに比べて約 40 倍の発見率の向上を得ることが示された [16]。

DSI 法は、蛋白質－化合物相互作用行列を用いて、既知の活性化合物と類似の化合物を検索する方法である。異なる化合物であっても、同一の蛋白質に強く結合する化合物は類似の化合物とみなすことができる（図 5）。DSI 法では標的蛋白質の立体構造は必要でないことから、G-protein coupled receptor（GPCR）のように立体構造未知の標的蛋

白質に対しても適用することができる。また MTS 法と同様に、DSI 法は既知化合物の発見率を最大化するようにスコアを修正する手法と組み合わせることができる。前述した蛋白質に GPCR を加えた計 14 種の標的蛋白質に対して DSI 法を適用した結果、化合物ライブラリーから予測上位 1 %の化合物を採択したとき、ランダムスクリーニングに比べて平均約 70 倍の発見率の向上を得ることが示された [17]。

5　目標にどれだけ近づいたか

我々は現時点で当初の目標の 90 %以上を達成したと言える。我々の最初の化合物 DB は 2004 年にリリースされ、ただちに TNF－α converting enzyme という標的蛋白質のスクリーニングに投入された。182 種類の蛋白質と 100 万化合物の蛋白質－化合物相互作用行列を用いて、MTS 法と DSI 法が適用され、900 種類の化合物を購入し、そこから 35 種類の活性化合物を得ることができた。これは、先に行われた 10 万化合物のランダムスクリーニングで 7 種類の活性化合物を得た結果に対して約 500 倍効率が高く、また別に行われた市販ソフト Glide によるスクリーニングで 700 種類の化合物を購入し、活性化合物が一つも得られなかったのに比較して、格段の発見率の向上が得られた。その後、化合物 DB は毎年更新され、現在は 2007 年度版となっている。6 年間に、我々は 10 種類近くの標的蛋白質について直接スクリーニングを行ったが、いずれも数 %～ 20 %という確率で活性化合物が得られた。これはランダムな実験よりも数百倍から千倍以上の高い確率である。また、化合物 DB 及び蛋白質－化合物相互作用行列は、毎年製薬メーカーを中心に、国内外の 10 ないし 20 の機関に配布が続けられている。ソフトウェア類は myPresto として [18]、化合物 DB は LigandBox として一部が公開されている [19]。

化合物データベース

タンパク質集団	1	2	3	4	5	6	7	8	9	Known active compound
Target protein	-0.9	-1.2	-9.1	-3.8	-3.2	-9.9	-0.9	-3.8	-9.6	-9.2
1	-4.3	-4.4	-3.5	-2.1	-2.8	-2.8	-6.1	-2.1	-3.8	-3.3
2	-8.4	-8.1	-2.1	-7.5	-6.6	-6.6	-5.1	-7.5	-2.1	-2.5
3	-5.4	-0.2	-5.5	-0.9	-0.4	-5.4	-3.2	-0.9	-7.5	-7.2
4	-4.4	-7.5	-0.1	-8.4	-5.1	-4.4	-2.8	-4.3	-0.9	-0.2
5	-8.1	-0.9	-6.1	-8.1	-3.8	-0.4	-6.6	-8.4	-6.1	-6.6
6	-8.2	-3.3	-5.5	-5.4	-2.1	-5.5	-0.4	-8.1	-5.1	-5.4
7	-2.1	-3.2	-4.3	-4.4	-5.4	-4.3	-7.5	-5.4	-4.3	-4.4
8	-7.2	-2.8	-0.5	-7.2	-4.4	-0.6	-5.4	-4.4	-8.9	-8.1
9	-0.2	-6.6	-0.4	-0.2	-8.1	-0.7	-4.4	-0.2	-8.8	-8.2
10	-6.6	-0.4	-2.2	-6.6	-9.3	-2.2	-8.1	-8.2	-2.9	-2.1

MTS法でのヒット化合物
DSI法でのヒット化合物
類似化合物

図5　MTS法とDSI法の概念図
表の数字は、スコア。強いスコアは濃い色で表示した。

6　やり残したこと

第一に、我々の化合物 DB は、亜鉛などの金属を含むメタロプロテアーゼの阻害剤に向いていない。分子のイオン形は水中での支配的イオン形になっているが、金属に配位するときのイオン形は異なる。水中ではチオール（-SH）は通常 -SH だが、金属に配位すると、脱 H 化した $-S^-$ である。このような金属配位におけるイオン形の変化はいろいろな官能基で見られる。メタロプロテアーゼの VS において、発見率はイオン形の良し悪しに強く依存することを我々は突きとめた。そこで、メタロプロテアーゼ用の化合物 DB を開発しようとしている。

第二に、我々の化合物 DB は無機化合物を含んでいない。金属錯体のような無機化合物は一般に薬になりにくいとされ、通常、化合物 DB から除外する。しかし、近年、ペプチド以外の活性化合物が一切知られていないインシュリン受容体蛋白質の活性化合物として亜鉛錯体が発見されるなど、無機化合物の新たな可能性が知られるようになってきた。無機化合物の可能性を検討するためにも、無機化合物の DB 化が必要だと感じている。

第三に、我々の化合物 DB の配布は口コミにのみ依存し、論文、ホームページなどで認知されていないことである。これは、我々の化合物 DB が、商社から提供されるカタログデータに依存しているためである。カタログの配布は試薬の販売目的に限られ、試薬ベンダーの広告を貼る必要がある。ZINC[20] は、試薬ベンダーとの直接交渉により、大学のホームページ上に試薬ベンダーの広告を張ることで、無償ダウンロードを実現している。産総研は、私企業の広告ができないことから無償ダウンロードはできず、ユーザーが独自に入手したカタログに対し、産総研がデータベース化サービスを行ったとして、化合物 DB を配布することになっている。研究助成を受ける共同研究先の社団法人から配布する手段はあるが、試薬ベンダーと交渉する力はない。産官学連携の中では、私企業との関わりは避けられない。こういった問題も将来の課題と思える。

謝辞

本研究は新エネルギー・産業技術総合開発機構（NEDO）及び経済産業省の援助によって行われた。

用語説明

用語 1：2 次元 SD ファイル：分子の元素名、XYZ 座標、結合次数、光学活性中心、整数化した原子電荷を記載する分子を記述するファイル。

用語 2：蛋白質－化合物ドッキングソフト：蛋白質の立体構造に対して化合物を蛋白質表面付近に配置し、エネルギー的に安定と思われるもっともらしい蛋白質－化合物複合

体構造を計算により作成することを言う。薬物スクリーニングでは1化合物のドッキングを数秒～1分程度で行う。DOCK、AutoDock、myPresto などがある。

用語3：ドッキングスコア：ドッキングソフトによって見積もられる蛋白質 - 化合物の相互作用の強さの値で、通常は結合自由エネルギーに相当する。

用語4：エンリッチメント：薬物スクリーニング計算において予測化合物数に占める真のヒット化合物数の割合。通常、ランダムな実験では1万化合物に1化合物ヒットするので、もし計算で予測したヒット化合物候補100化合物中にヒットが1件あれば、ランダム実験に対するエンリッチメントは100倍となる。

参考文献

[1] http://www.mdl.com/jp/products/experiment/cims/index.jsp
[2] http://www.molecular-networks.com/software/corina/index.html
[3] M. Hattori, Y. Okuno, S. Goto and M. Kanahisa: Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc.*, 125 (39), 11853-65 (2003).
[4] J. Gasteiger and T. Engel: *Chemoinformatics: A textbook.* WILEY-VCH: Weinheim (2003).
[5] http://www.lmcp.jussieu.fr/sincris-top/logiciel/prg-babel.html
[6] http://openbabel.org/wiki/Main_Page
[7] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case: Development and testing of a general amber force field. *J. Compt. Chem.*, 25 (9), 1157-1174 (2004).
[8] http://www.ccdc.cam.ac.uk/products/csd/
[9] Y. Fukunishi, Y. Mikami and H. Nakamura: The filling potential method: A method for estimating the free energy surface for protein-ligand docking. *J. Phys. Chem. B.* 107 (47), 13201-13210 (2003).
[10] J. Gasteiger and M. Marsili: A new model for calculating atomic charges in molecules. *Tetrahedron Lett*., 3181-3184 (1978).
[11] http://openmopac.net/index.html
[12] J. Wang, P. Cieplak and P.A. Kollman: How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.*, 21 (12), 1049-1074 (2000).
[13] Y. Fukunishi, Y. Mikami and H. Nakamura: Similarities among receptor pockets and among compounds: Analysis and application to *in silico* ligand screening, *J. Mol. Graph. and Model.*, 24 (1), 34-45 (2005).
[14] Y. Fukunishi, Y. Mikami, K. Takedomi, M. Yamanouchi, H. Shima and H. Nakamura: Classification of chemical compounds by protein-compound docking for use in designing a focused library. *J. Med. Chem.*, 49 (2), 523-533 (2006).
[15] Y. Fukunishi, Y. Mikami, S. Kubota and H. Nakamura: Multiple target screening method for robust and accurate in silico ligand screening. *J. Mol. Graph. and Model.*, 25 (1), 61-70 (2005).
[16] Y. Fukunishi, S. Kubota and H. Nakamura: Noise reduction method for molecular interaction energy: application to in silico drug screening and in silico target protein screening. *J. Chem. Info. Mod.*, 46 (5), 2071-2084 (2006).
[17] Y. Fukunishi, S. Hojo and H. Nakamura: An efficient in silico screening method based on the protein-compound affinity matrix and its application to the design of a focused library for cytochrome P450 (CYP) ligands. *J. Chem. Info. Mod.*, 46 (6), 2610-2622 (2006).
[18] http://presto.protein.osaka-u.ac.jp/myPresto4/index_e.html
[19] http://presto.protein.osaka-u.ac.jp/LigandBox/web_search.cgi
[20] J. J. Irwin and B. K. Shoichet: ZINC – a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.*, 45 (1), 177-82 (2005).

執筆者略歴

福西 快文（ふくにし よしふみ）

1994年京都大学工学研究科博士課程修了。通商産業省工業技術院融合領域研究所非常勤職員、HFSP フェロー、Rutgers 大学ポスドク、（独）理化学研究所（JST ポスドク）、（株）日立製作所などを経て、2000年より産業技術総合研究所　バイオメディシナル情報研究センター 主任研究員。専門：計算化学。本論文では、試作品の作成、各種アルゴリズムの考案、全体の設計を担当した。

杉原 裕介（すぎはら ゆうすけ）

1996年広島大学大学院理学研究科高分子化学講座修士課程修了。1996年荒川化学工業（株）入社、2000年同退社、2001年（株）富士通九州システムエンジニアリング入社。本論文では、主にカタログからの化合物の3次元化を担当した。

三上 義明（みかみ よしあき）

1987年（株）日立東日本ソリューションズ入社、計算科学ソリューション部所属。現在、バーチャルスクリーニングなどのシステム開発やコンサルテーション業務に従事。情報処理学会会員。本論文では、主に蛋白質－化合物相互作用行列の作成を担当した。

酒井 広太（さかい こうた）

1989年九州大学大学院理学研究科高分子化学講座修士課程修了。1989年（株）富士通九州システムエンジニアリング入社。本論文では、主にカタログからの化合物の3次元化を担当した。

楠戸 寛（くすど ひろし）

2002年（株）日立東日本ソリューションズ入社、計算科学ソリューション部所属。現在、並列計算システムの構築や研究支援業務に従事。情報処理学会会員。本論文では、主に蛋白質－化合物相互作用行列の作成を担当した。

中村 春木（なかむら はるき）

1980年東京大学理学研究科博士課程修了。東京大学工学部助手、蛋白工学研究所、生物分子工学研究所を経て、1999年より大阪大学蛋白質研究所教授。専門:生物物理学。本論文では、主に公開データの取り込み、全体の統括を担当した。

査読者との議論

議論1　化合物データベースの作成の意義

コメント・質問（小野 晃）

研究目標を明確に設定し、そのための要素技術の選択のシナリオを図1のように明快に描出し、そして現実にオペレーションできるデータベースとして統合していったことは、典型的な第2種基礎研究の手法

であり、また製品化研究としても同時に、非常に優れた研究と思います。標的蛋白質が今後次第に明確にされていく中で、本データベースの価値は一層増すものと期待します。

回答（福西 快文）
化合物データベースは、人類が過去にどのような分子を合成したか・合成することができたか、という知識のプールです。直接、医薬品探索に用いるだけでなく、どのような分子が合成しやすいのか、どのような分子を人類は思いつかなかったのか、を考えるときの基礎となって新しい研究領域が開けることも期待しています。

議論2　未知化合物の予測

コメント・質問（小野 晃）
この化合物データベースは、特定の標的蛋白質と、用意された多数の化合物の間の化学的結合の強さを効率的に調べるもので、医薬品の発見率が飛躍的に向上しました。一方、この化合物データベースを使って、特定の標的蛋白質に対してより強く結合するような未知の化合物をユーザーが予測するといった使い方も可能でしょうか。

回答（福西 快文）
未知の化合物を予測できる可能性はあります。特定の標的に対し薬物探索を行った場合、そのヒット化合物群が、共通の特徴を有する化合物集団に分類される傾向がありました。これらの集団の特徴を兼ね備える未知の物質を合成デザインするという研究の方向はありうると思います。

議論3　データベースの改良可能性

コメント・質問（小野 晃）
4.3 節で述べられていますが、データの正確さを過度に追求することの無意味なことは重要な指摘と思います。その意味で、このデータベースの作成におけるいろいろなプロセスをさらに見直して、目的に対してより合理的にする余地は今後もありうると考えてよいでしょうか。

回答（福西 快文）
データベースを目的に対してより合理的にする余地はあると考えます。現在は、分子が水中においてとりやすい分子形（水素原子のつき方など。カルボン酸なら -COOH → -COO^-）を作成していますが、蛋白質と結合する場合は分子形が変化することが知られています。最近では、強く電荷を帯びた蛋白質ポケットを標的とする薬物ドッキング計算が難しいので話題になっています。負電荷を帯びたポケットの中ではカルボン酸が、-COOH になる場合があります。標的蛋白質に対して、可能性の高い分子形を準備することは重要になると思います。

議論4　先行する海外データベースとの比較

コメント・質問（小野 晃）
4.11 節で、本データベースを使ったスクリーニングはランダムスクリーニングに対して 40 倍、あるいは 70 倍の発見率の向上が図られたとしています。一方、化合物データベースとして海外で先行的に開発されてきたものに対して本データベースは、どの程度発見率の優位性を持っていると評価されますか。

回答（福西 快文）
通常、計算による予測化合物が外れた場合、文献として公表されないため、データベースの優劣の詳細な比較は難しいです。我々の場合、計算による予測化合物を 100-300 種類購入した場合の活性化合物の発見率は 3 % -30 %になります。今までに、発見率が 0 %であった標的は 5 標的中 1 標的程度です。論文で見かける発見率は高々 10 %で、計算による予測は 2 標的当たり 1 標的程度で有効とされているようですので、海外での事例に比べて有効性が高いと考えられます。

議論5　互変異性体とイオン化状態の考慮

コメント・質問（広川 貴次）
本分野では、海外の市販データベースやソフトウェアが大きなシェアを占めつつありますが、ここまで高品質の化合物データベースおよび他に類を見ない独創性の高い蛋白質－化合物相互作用行列データベースが日本発で公開されていることは大変素晴らしいです。論文中で記載されている各プロセスは、化合物を電子的に取り扱う際に問題となる様々な問題について十分に対応されており、研究者が安心してデータベースを利用できる優れた内容になっています。製品化研究としても高く評価できます。
化合物の水素付加について互変異性体（Tautomer）や全体としてのイオン化状態（pH7.0 を前提としているのか等）の取り扱いは、4.4 節の「我々は、様々な官能基の…正確に生成するようにした。」という説明の範囲で考慮されているのでしょうか。

回答（福西 快文）
イオン化状態は、pH7.0 を前提にしています。しかし、正確な pKa 予測は困難なため、分子全体での pKa 予測を行うのではなく、分子に含まれる各官能基の pH7.0 での代表的な構造を適用するようにしています。互変異性体も同様な取り扱いになっています。よって、科学的にはまだ厳密性は追求されていないのですが、babel/openbabel といったオープンソフトが高エネルギーな状態を頻繁に生成するのに比べると、精度は上がっています。

議論6　選択性の悪い化合物群の予測

コメント・質問（広川 貴次）
4.11 節のタンパク質 - 化合物相互作用行列の計算の活用事例として、例えば、多くの標的蛋白質に結合しやすい選択性の悪い（良い場合もあるかも）化合物群がこのデータベースを利用して予測できるのであれば、in silico frequent hitter や in silico chemical alert for target selectivity といった独自性の高いアノテーションが考案できるのではと思います。既に実施しているかも知れませんが、その可能性について議論いただくことは可能でしょうか。

回答（福西 快文）
frequent hitter は、VS において数十%も出現し、経費の無駄とスクリーニング実験の障害となっています。我々のスクリーニングでも、予測化合物の約 20 %が frequent hitter と報告されています。現在、文献から数十の frequent hitter とされる分子を収集し、立体構造の作成をしています（*J. Med. Chem*. 2003, vol 46, page 4477-4486, *J. Med. Chem*. 2002, vol45, page 137-142）。データが揃い、タンパク質－化合物相互作用行列の計算に混ぜることができれば、これらがどのタンパク質にも強いスコアを示すような特徴が現れるかも知れません。しかし、frequent hitter は電子顕微鏡で観察すると、多くがミセルコロイドを形成しており、ミセルがタンパク質に張り付くことで frequent hitter となるのでは、という報告もあります（*J. Med. Chem*. 2002, vol 45, page 1712-1722）。もし、単分子で存在する化合物がタンパク質に対して非選択性を示すのが frequent hitter の本質ならドッキング計算で解析が可能と思いますが、ミセル形成が frequent hitter の原因なら、無限希釈状態に相当するドッキング計算では frequent hitter を識別できない可能性があります。今のところ分子記述子を用いた溶解度予測を frequent hitter に適用してみたところ、疎水性が強い分子ほど frequent hitter であり、frequent hitter でない医薬分子と区別される傾向が見えてきました（CBI 学会 2008 年大会 P1-06）。水への溶解性が frequent hitter を決めているのなら、ミセル形成が主たる原因となります。ドッキング計算のほうが、より鮮明に frequent hitter を区別できる可能性は残っています。少し時間はかかりますが、本件の検討を進めたいと思います。
化合物の非選択性からくる副作用の解析については、MTS法及びDSI法での検討を行ったことがあります。今のところ、余り明確に現象が見えていません。非ステロイド系消炎鎮痛剤（NSAID）の代表的な

標的であるCOX2と、胃粘膜保護作用を持つCOX1は、アミノ酸相同性60 %の酵素であり、NSAIDがCOX2だけでなくCOX1を阻害することで胃潰瘍を引き起こすことは問題となっていました。近年、コキシブ系COX2選択性NSAIDが開発されました。そこで、COX2選択性NSAIDと非選択性NSAIDが、タンパク質-化合物相互作用行列の利用で区別できるかを検討しましたが、区別できませんでした。実は、COX2選択性は非常に微妙なもので、COX2を80 %阻害する濃度でのCOX1の阻害は選択的NSAIDで20 %、非選択的NSAIDで80 %程度でした（*Proc. Natl. Acad. Sci*. USA. 1999, vol 96, page 7563-7568）。したがって、薬物選択性は程度の問題だと考えられます。強い選択性・非選択性は、タンパク質－化合物相互作用行列の利用で、区別できると考え、現在、約1500種類のタンパク質構造をドッキング計算に利用可能な形で準備しています。計算能力の問題で実際の解析はできませんが、近い将来には解析可能になると期待しています。