K04 記述子サンプリング法を用いた能動学習法に基づく創薬スクリーニング

（日本電気（株））○藤原　由希子*、山下　慶子、襲田　勉、麻生川　稔
（田辺製薬（株））朝尾　正昭、島津　秀史、中尾　和也、福島　千晶、清水　良**

はじめに

創薬のリード探索段階における構造展開を手助けするためには、探索スクリーニングの段階でできるだけ多くの構造活性相関情報を取得することが望ましい。そこで我々は、スクリーニング化合物の選抜に能動学習法を用いることを検討しており[1), 12)]、今回は、我々が独自に開発した記述子サンプリング法を用いた能動学習法について報告する。

能動学習法は、学習者（コンピュータ）が訓練データを能動的に選択することで予測精度を向上させる学習法である[2)]。化合物選抜において問題となるのは、一部分の選択データによる構造活性相関ルールに基づいて化合物を選択すると、化合物の多様性が失われる危険性があることである。これは、これまでに発見された薬となるべく異なる化合物を選抜し新たな薬を創るという目的に合致しない。そこで、なるべく多様な化合物を選択するため、記述子サンプリング法を用いた能動学習法を開発した。

スクリーニング化合物選抜法の検証としては、G蛋白質共役型受容体（GPCR）のうち生体アミン受容体に作用する合成リガンドを対象とし、既知リガンド化合物と一般試薬を合わせた化学構造データベースから、能動学習法を用いたリガンド化合物の抽出を検討した。

方法

１）構造情報の取得と分類

治験薬の化学構造が収められているデータベースPharmaprojects（2004.02）[3)]から標的タンパク質の名前で検索し、生体アミン受容体である、アドレナリン、ドパミン、ヒスタミン、ムスカリン、セロトニンの各受容体に作用する1,551化合物を抽出し[4)]、正例（活性あり）とした。

一方、負例（活性なし）は、Pharmaprojectsの正例以外の9,340化合物と、一般試薬データベースAvailable Chemicals Directory（ACD 2002.10）[5)]で次の条件を満たす246,100化合物とした。

- 分子量100～1000
- 重原子数6個以上
- 原子種はC, H, N, O, S, P, F, Cl, Br, Iに限定
- 重複登録された化合物を1個にまとめる
- 同位元素含有、重水素含有化合物は除く
- ペプチドを除く
- 反応性のある化合物を除く

２）パラメータ（構造記述子）の取得

下記、171種類の構造記述子を算出した。

- MDL Molskey[6)]：166種類
- 物理化学定数　：5種類（ClogP[7)], Molecular Weight[8)], Hydrogen-Bond Acceptors[8)], Hydrogen-Bond Donors[8)], Rotatable Bond[9)]）

３）能動学習法

能動学習法は、まずランダムにデータを選択した後、以下の手順を繰り返す：選択データの活性を薬理実験で決定し、構造活性相関ルールを作成し、そのルールを用いて活性未知のデータ中からデータを選択する。この際、全データでなくサンプリングしたデータを用いて複数のルールを生成するBagging[10)]と呼ばれる手法を用いている。しかし、化合物選抜においては化合物の多様性が失われる危険性がある。例えば、活性に関係ない部

*y-fujiwara@db.jp.nec.com,　**ryo@tanabe.co.jp

分構造Sが、たまたま選択データにおいて全正例に含まれ全負例に含まれないと仮定する。この部分構造Sを表現する記述子Dは、訓練データ（選択データ）で最も正例と負例を分類できるので、ほとんどのルールではDが用いられ、Sを含まないデータは負例であると予測される。従って、Sを含まない化合物は選抜されず、選択された化合物は多様性を失ってしまう。この問題を防ぐため、我々は記述子サンプリング法を開発した[11]。これは、個々の分類器生成の際に、サンプリングによる一部の記述子を用いる手法であり、Baggingの手法を記述子に応用したものである。この手法を用いると、記述子Dを用いないルールも生成され、Sを含まない化合物の選抜が可能となる。

検証実験は、10分割交差法を用いて評価した。これは、データを均等に10分割し、1割を予測のためのテストデータとして除いて、残り9割でスクリーニングを行い、その10回の繰り返しの平均で評価するものである。1サイクルに選択する化合物数は5,000、能動学習で生成する分類器数は100とした。能動学習法の有効性を示すため、対照実験として、選択された活性化合物に類似した化合物をTanimoto係数により選択する手法（類縁化合物選択法）の実験も行った。ただし、類縁化合物選択法の記述子はMolskey166種のみである。

結果と考察

1）記述子サンプリング法

記述子サンプリング法の有無によるスクリーニングにおけるヒット率の推移を図1に示す。横軸は選択データ数、縦軸はスクリーニングデータの活性化合物中で選択データに含まれる割合（%）であり、スクリーニング終了時にはどのような手法でも全ての活性化合物が選抜され100%となる。図1より、90 %の活性化合物選抜には、3万データ（全データの約13%）が必要だが、記述子サンプリング法を用いると2.5万（約11%）で十分であった。また、99%の活性化合物選抜には、8万データ（約35%）が必要だが、記述子サンプリング法を用いると7.5万（約32%）で十分であった。

次に、テストデータを予測した際のヒット率の推移を図2に示す。横軸は選択データ数、縦軸はテストデータ中の活性化合物のスコア上位2,000でのヒット率である。全体的に記述子サンプリング法を用いた場合のヒット率が高かった。これらの結果から、記述子サンプリング法は有効であるといえる。以降、記述子サンプリング法を用いた能動学習法を能動学習法DSと略す。

2）能動学習法と他の手法との比較

能動学習法DS、類縁化合物選択法、ランダムスクリーニングに対して、スクリーニングにおけるヒット率の推移を図3に示す。90%の活性化合物選抜には、ランダムスクリーニングは90%のデータ、類縁化合物選択法は5万データ（約22%）が必要であり、99%には、ランダムスクリーニングは99%、類縁化合物選択法は14万（約61%）が必要であった。一方、能動学習法DSは90%には2.5万（約11%）、99%には7.5万（約32%）で十分であり、選択化合物数や実験回数は、ランダムスクリーニングに比べ大きく減り、類縁化合物選択法に比べ半減した。これは、能動学習法DSが実験コストやスクリーニング期間を従来に比べ削減できることを示している。なお、能動学習法DSの結果は、Molskeyの166種の記述子のみを用いた場合でも同様であり[10]、この類縁化合物選択法との差は記述子の差でなく手法の差であることを確認している。

次に、テストデータにおけるヒット率の推移を図4に示す。まず、能動学習法DSとランダムとでの得られる情報の差を示すために、能動学習法DSで得た分類器と、ランダムに得たデータで作成した分類器とでの、テストデータにおけるヒット率を検討した。ランダム選択では、データ数の増加とともにゆるやかにヒット率は向上した。これはデータが多いほど得られる構造活性相関情報が多いことを示している。なお、実際のランダムスクリーニングは、ルールを生成しないので、2,000でのテストデータのヒット率は常に1%未満であり、このランダム選択の高いヒット率は能動学習法DSと同じ学習法を用いたためである。一方、能

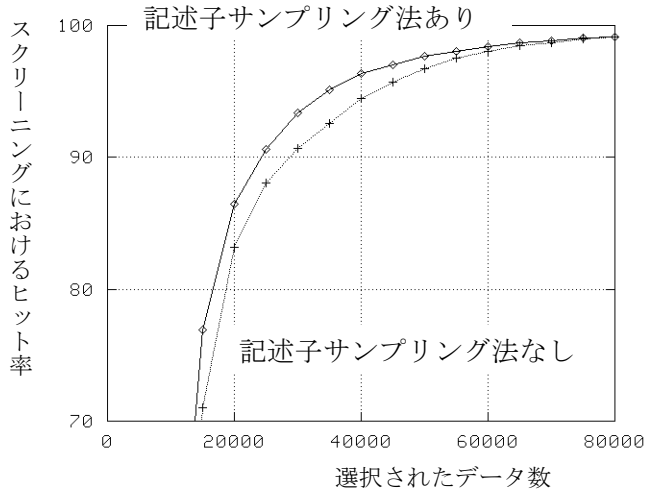

図1. スクリーニングにおけるヒット率

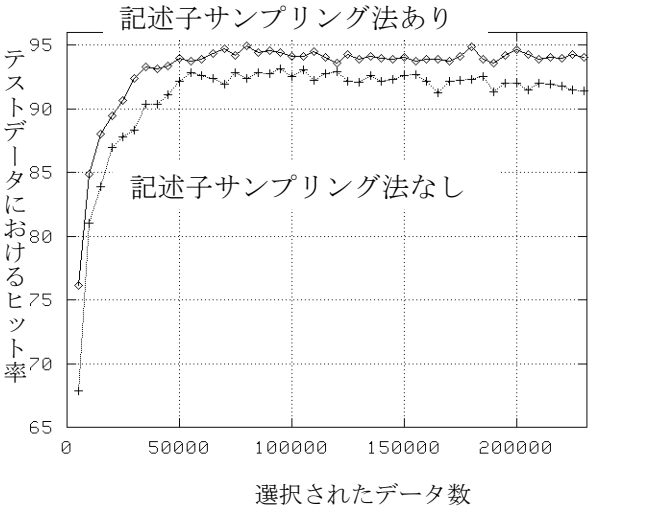

図2. テストデータにおけるヒット率

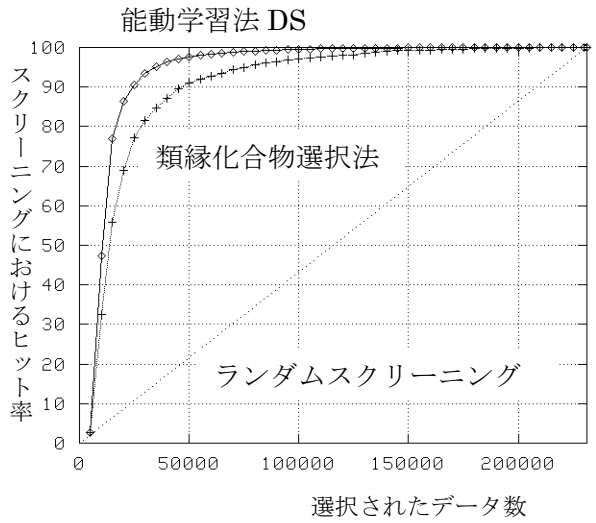

図3. 他の手法とスクリーニングヒット率の比較

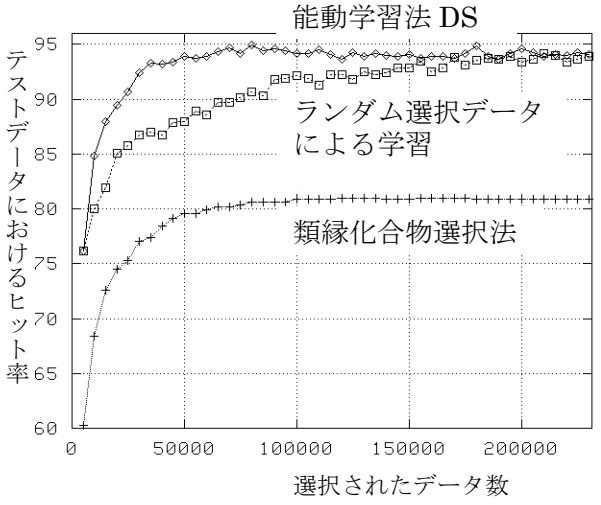

図4. 他の手法とテストデータヒット率の比較

動学習法DSはランダム選択より少ないデータの段階で高いヒット率を得た。例えば、ランダム選択の際の最終的なヒット率94%を、能動学習法DSは約5万件のデータで達成した。これは、能動学習法DSがデータ選択により構造活性相関情報を効率よく取得したことを示す。

能動学習法DSと類縁化合物選択法のヒット率の比較についても図4に示す。類縁化合物選択法は、多くのデータを用いても約8割のヒット率にしかならないのに対して、能動学習法DSは、9割以上のヒット率を達成した。このように能動学習法DSのヒット率が高いのは、類縁化合物選択法は全体的な化合物の構造の類似度に基づいて予測するため全体的に類似しているが活性に関係する部分構造が無い化合物も選択するのに比べ、能動学習法DSでは活性と相関する部分構造についてのルールに基づいて予測するためと考えられる。

3）生成された構造活性相関ルール

記述子サンプリング法によって、生成される構造活性相関ルールがどのように変化するかを図5に示す。これは、最初の5,000データを学習した時の100個の分類器中、最初の4個において、学習には用いなかったアドレナリン受容体リガンド**1**(図6)に対応するルールを示したものである。記述子はfで始まる文字で示しており、Molskeyの場

(A)記述子サンプリング法なしの４ルール

If f101=0 & f86=1 & f19=0 & f48=0 & fHbA>1 & f84=1 & f72=0 *then score:* **0.030**
If f19=0 & f100=1 & f101=0 & f72=0 & f108=0 *then score:* **0.001**
If f19=0 & f17=0 & f86=1 & f101=0 & f34=0 & f105=1 & f114=0 & f66=0 & f79=1 & fHbA>1 & f103=1 & f72=0 *then score:* **0.111**
If f100=1 & f17=0 & f101=0 & f72=0 & f19=0 & f108=0 & f74=0 *then score:* **0.001**

(B)記述子サンプリング法ありの４ルール

If f86=1 & f105=1 & f87=1 & f120=1 & f71=0 & f72=0 *then score:* **0.050**
If f19=0 & f100=1 & f101=0 & f137=1 & f17=0 & f159=1 *then score:* **0.002**
If f19=0 & f86=1 & f101=0 & f17=0 & f34=0 & f105=1 & f114=0 & f66=0 & f94=1 & fHbA>1.5 & f103=1 *then score:* **0.125**
If f100=1 & f113=1 & f115=0 & f84=1 & f74=0 & f134=1 & f8=0 *then score:* **0.800**

図 5. 構造活性相関ルールの例

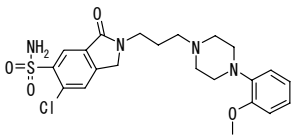

図 6. 化合物 **1**

合、1はそのkeyあり、0はそのkeyなしを示す。また、スコアは0～1の範囲を取り、1に近いほど正例らしさが高いことを示す。図5で示すように、記述子サンプリング法を用いないと、この時点で分類に重要と計算される記述子f72（O-Any-Any-O）、f101（8員環以上の環）、f19（7員環）を含むルールが生成され、これが化合物**1**を正例と予測するのを妨げている。一方、記述子サンプリング法では一部の記述子を無視するため、f72、f101、f19を用いないルール（(B)の4番目）が生成され、これにより化合物**1**を高いスコアで正例らしいと予測することができる。

まとめ

本研究では、記述子サンプリング法を用いることで能動学習法を改良し、大量の化合物群から少数のヒット化合物を効率よく選抜する手法を開発した。また、従来法である類縁化合物選択法より高いヒット率を達成した。今後は、さらに活性化合物が少ない場合について検討し、化合物選抜法を確立していく予定である。

謝辞

本研究の開始に御尽力いただきました土肥俊博士（日本電気（株））、馬見塚拓助教授（京都大学化学研究所）、山西健司博士（日本電気（株））、データベースの利用を許可いただきましたPJB Publications Ltd.様および日本MDLインフォメーションシステムズ（株）様に感謝いたします。

参考文献

[1] 麻生川稔 他，第31回構造活性相関シンポジウム, (2003).
[2] Abe, N and Mamitsuka, H, *Proc. Int. Conf. on Machine Learning (ICML98)* , 1-9, (1998).
[3] PJB Publications Ltd., 18/20 Hill Rise, Richmond, Surrey TW10 6UA, UK. (http://www.pjbpubs.com/)
[4] GPCRDB (http://www.gpcr.org/, Horn, F. and et al., *Nucleic Acids Res*., **26,** 275-279, (1998).)の分類による
[5] MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577, USA. (http://www.mdl.com/)
[6] Durant, J. L. and et al., *J. Chem. Inf. Comput. Sci.*, **42**, 1273-1280, (2002).
[7] Daylight Chemical Information Systems, Inc., 27401 Los Altos, Suite 360, Mission Viejo, CA 92691, USA. (http://www.daylight.com/)
[8] Lipinski, C. A. and et al., *Adv. Drug Delivery Rev.*, **46,** 3-26, (2001).
[9] Accelrys Inc., 9685 Scranton Road, San Diego, CA 92121-3752, USA. (http://www.accelrys.com/)
[10] Breiman, L., *Machine Learning*, **24**, 2, 123-140, (1996).
[11] Fujiwara, Y., et. al., *Genome Infomatics,* **14,** 597-598, (2003).
[12] Asogawa, M., et. al., *Proc. Int. Conf. Intell. Syst. Mol. Biol. (ISMB2004)* , G-7, (2004).