2005－BIO－1（7）
2005／7／25

能動学習法を用いた創薬スクリーニング方法

麻生川　稔*、襲田　勉**、藤原　由希子*、山下　慶子*
日本電気(株)
*基礎・環境研究所,　**バイオ IT 事業推進センター

創薬の探索で、標的とするタンパク質と結合する化合物スクリーニングする過程に於いて、膨大な化合物に対して全ての化合物をハイスループットでスクリーニングする方法(HTS)が、近年良く使われている。本発表では、能動学習法を用いることによって、化合物を効率良く選択し実験を行うことにより、HTS に比べて効率的スクリーニングが可能であることを示した。手法の検証に、創薬研究で重要な G 蛋白質共役型受容体(G-protein coupled receptor:GPCR)を標的とする化合物群を利用した。ヒット化合物が0.6%含まれる約 20 万種類の化合物を用いたシミュレーション実験では、全体の 20%の化合物の実験から、90%のヒット化合物を選抜することができた。また、本手法で新規に発見した 8 化合物(1μM 濃度)についても、同様に得ることが可能であることを示した。

Efficient Drug Screening using Active Learning

Minoru Asogawa*, Tsutomu Osoda**, Yukiko Fujiwara*, Yoshiko Yamashita*

* Bioinformation, Fundamental Research Laboratories, NEC Corporation
** Bio-IT Business Promotion Center, NEC Corporation

At the phase of lead chemical compounds search for drug discovery, both the combinatorial chemistry method and the high throughput screening (HTS) method are successfully utilized and discover several hit chemical compounds from huge chemical library.　In this paper, we proposed an active learning method as an efficient chemical screening method and shown its effectiveness. To demonstrate system performance G-protein coupled receptor is chosen as a target protein. According to the computer simulation results, it is shown that one fifth of screening is enough for finding ninety percent of all hit compounds, from 200,000 compound library.　By utilizing this method, we have found eight novel chemical compounds, and found that we could have found those compounds same efficiency as the computer simulation.

１．はじめに

創薬のリード探索段階において、コンビナトリアルケミストリーとハイスループットスクリーニングによって複数のヒット化合物（一次スクリーニングで見出される活性化合物）が比較的短期間に取得できるようになった。見出されたヒット化合物から多様な構造修飾による最適化の対象となるリード化合物へ展開していくのはメディシナルケミストの役割であり、これまで主として経験と知識に頼って実施されてきた。現在もまだ、ヒット化合物からリード化合物を効率的に見出すための構造展開手法は確立

されてはいない。そのような構造展開を進めるためには、一次スクリーニングの段階でできる限り多くの構造活性相関情報を取得する必要がある。すなわち、ある時点までに得られた活性化合物の構造情報を利用して、次の期間にどのような化合物群の活性データを取得すれば無駄のない構造展開ができるかを考えて、被験化合物をデザイン・選抜することが重要となる。我々は、構造展開に必要な構造活性相関情報を効率的に取得できるスクリーニング化合物選抜法として、能動学習法[1)]を用いることを検討した。

今回、スクリーニング化合物選抜法の検証にG蛋白質共役型受容体（G-protein coupled receptor：GPCR）を標的とする化合物群を利用した。GPCRファミリーは医薬品の代表的な標的タンパク質であり[2)]、今後も創薬研究の対象として重要と考えられる。GPCRの中でも特に、アドレナリン受容体、ドパミン受容体、セロトニン受容体など、いわゆるbiogenic amineを天然のリガンドとする一群の受容体は薬理学の分野で詳しく研究されてきた。今回、構造情報を最も豊富に取得できる一連の化合物群として、これらのbiogenic amine受容体に作用する合成リガンド（アゴニスト・アンタゴニスト）を利用することにした。すなわち、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、効率的なリガンド化合物の抽出を、能動学習法を用いて検討した。

２．方法

１）構造情報の取得と分類

治験薬の化学構造および開発状況が収められているデータベースPharmaprojects（2002.03）[3)]から標的タンパク質の名前で検索し、biogenic amine受容体であるアセチルコリン、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1461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）は一般試薬データベースAvailable Chemicals Directory （ACD 2002.10）[5)]から次の条件で化合物を選抜し、212914化合物を負例として得た。

- 分子量100～1000に限定
- 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
- 重複して登録されている化合物を1個にまとめる
- 同位元素含有化合物、重水素含有化合物は除く
- 信頼性の低い推算パラメータ値を含む化合物を除く

今回、構造が多様な大量の化合物群から、正例とした化合物群を効率よく選抜できる手法を構築する目的で、ACDの化合物（およびbiogenic amine受容体以外のGPCR基質）を負例として利用した。そのため、負例は実際に生理活性を測定して活性がないと確認された化合物群ではなく、後述の通り、負例としていた化合物群の中にも活性を持つ化合物が含まれていた。

２）パラメータの取得

下記、215種類のパラメータ（構造記述子）を算出した。

- MDL Molskey[6)]：166種類
- 物理化学定数：7種類（ClogP[7)] , CMR[7)], Topological Polar Surface Area[8)], Molecular Weight[9)], Hydrogen-Bond Acceptors[9)], Hydrogen-Bond Donors[9)], OHとNHとNH_2の合計数）

・ Cerius2/Diversity[10]で求めたパラメータ：42 種類

3）能動学習法

能動学習とは、学習者（コンピュータ）が学習データを能動的に選択する学習方法である。データを選択するときには、最も学習の効率を向上させると期待できるデータを選ぶようにプログラムされている。学習アルゴリズムへの制約を課さない Qbag（Query by Bagging）という手法が安倍、馬見塚によって提案されている[11]。

一次スクリーニングでヒット化合物は数万～数十万化合物のうち数十個程度の割合しか存在せず、従来の能動学習法を適用すると、本来リード化合物の候補となる化合物さえも不適当であると判断してしまう。そこで、リサンプリングするときに上記のような判定の誤りを減らすような方式を考案した。

4）選抜された化合物の評価

負例に含まれる化合物すべてが実験的に活性なしと判定されたものではない。負例としていた化合物の中で、“正例らしさ”の数値が高いものについて生物活性試験を行い、実際に活性が確認できれば効率的に活性化合物を選抜できたと考えることができる。そこで、負例化合物のうち“正例らしさ”の数値が高い上位 500 個の中から入手容易な Maybridge 社[12]の 20 化合物を選び、Cerep 社[13]にてアドレナリン受容体、ムスカリン受容体、セロトニン受容体の拮抗作用を測定した。化合物濃度は 10^{-6}M で二回測定し、%阻害率の平均値を求めた。

3．結果と考察

1）学習精度（ROC 曲線）

図 1 に ROC 曲線（Receiver Operating characteristic curve）を示した。データをすべて使ったときの ROC 曲線はランダムサンプリング法と能動学習法は一致する。図 1 にはデータ数が 48000 件のときの ROC 曲線を示した。図 1 では、能動学習法は ROC 曲線がランダムサンプリング法より上側に存在している。すなわち、データ数が少ない段階において、能動学習法がより高い学習精度を達成していることがわかる。

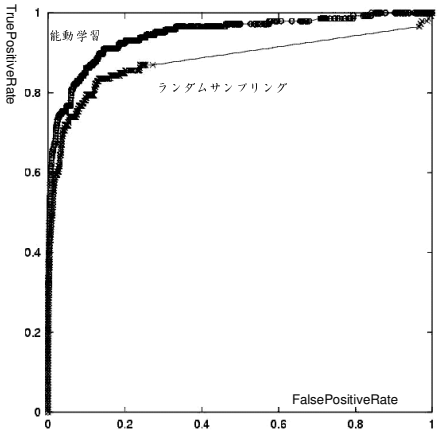


図 1．ROC 曲線

横軸は偽正率（正例と判断された負例数／全負例数）、縦軸は真正率（正例と判断された正例数／全正例数）

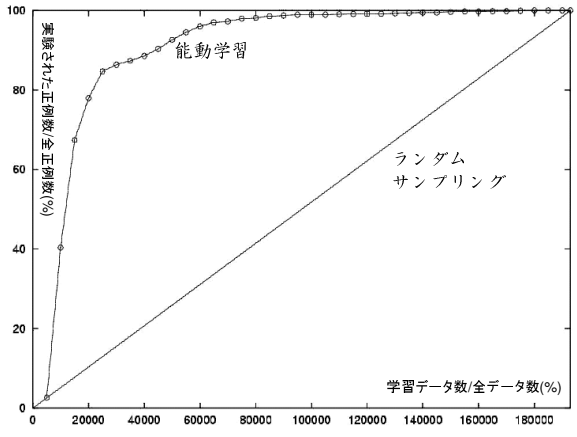


図 2．カバー率

横軸は全データ中の選択されたデータの割合、縦軸は全正例中の選択された正例の割合。10 fold cross validation の結果

２）カバー率

図 2 にカバー率の推移を示した。図 2 からランダムサンプリング法によって正例の 90%のデータを取り出すためには、全部のデータの 90%を学習に利用しなければならないことがわかる。一方、能動学習法では、90%の正例を取り出すのに、全体の 20%のデータを利用すればいいことがわかる。このことは、活性化合物をなるべく早く見つける目的の為には、能動学習によって学習を行った方がランダムサンプリングより効率的であることを示している。

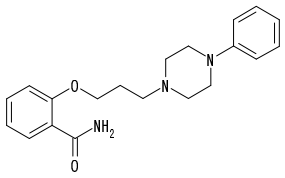
３）選抜された化合物の活性評価

負例の中で“正例らしさ”の数値が高い 20 化合物の生物活性を測定した結果、8 化合物が biogenic amine をリガンドとする GPCR に対して 1μM で 50%以上の拮抗作用を示した（表 1）。これは、能動学習法によって上位にランクされる化合物を選ぶことで、高い確率で活性化合物を見出すことができることを示唆する結果である。

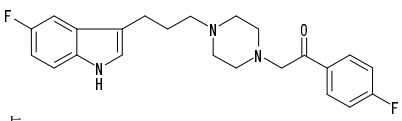
表 1．選抜した 20 化合物の生物活性結果

No	能動学習法	α１アドレナリン	α２アドレナリン	ムスカリン	セロトニン
1	86.9	**90**	42	13	38
4	75.0	**70**	0	0	26
5	73.2	47	3	**64**	**55**
12	67.3	**66**	0	0	0
15	66.1	**96**	21	23	**60**
17	65.5	0	0	**86**	20
19	64.3	**52**	0	21	**52**
20	60.7	**96**	**79**	0	13

10^{-6}M での各受容体に対する拮抗作用（%）を測定した。能動学習法の数値は 0～100(%)の得点で、大きい方が正例らしい、すなわち biogenic amine 受容体のリガンドらしい化合物である。



Compound 1



Compound 15

４．まとめ

本研究では、能動学習法を用いることにより、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確立できた。これは、創薬のリード探索段階における「少数のヒット化合物を取得後、次にどのような化合物群をスクリーニングすれば短期間でリード化合物を見出せるか」という課題に対する解決法のひとつと期待される。今後、パラメータセットを種々選択して再解析を行い、化合物とタンパク質との相互作用に関連する“意味のある”パラメータセットを見出し、化合物選抜法の改良をすすめる予定である。

謝辞

本研究は、田辺製薬(株)と共同で行ったものであり、共同研究者の田辺製薬(株)の朝尾 氏、櫛山 氏、中尾 氏、黒田 氏、和田 氏、大軽 氏、福島 氏、清水 氏に深く感謝いたします。また、本研究の基本的アイデアを着想された土肥 氏(日本電気)、ご助言を頂いた馬見塚 教授(京都大学化学研究所)、参考資料をご提供頂きました山西 氏(日本電気)、有益なご助言を頂きました白井 氏(アステラス製薬)、共同研究者にデータベースの利用を許可頂いた PJB Publications Ltd.、および日本 MDL インフォメーションシステムズ(株)に感謝いたします。

参考文献

1)N.Abe and H.Mamitsuka, Query Learning Strategies Using Boosting and Bagging, Proceedings of the 15th International Conference on Machine Learning(ICML98), pp. 1-9, 1998.

2) Fauman, E. B., Hopkins, A. L., Groom, C. R., *Methods Biochem. Anal.*, **2003**, *44*, 477-497.

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*., **1998**, *26*, 275-279.）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.*, **2002**, *42*, 1273-1280.

7) Daylight Chemical Information Systems, Inc., 27401 Los Altos, Suite 360, Mission Viejo, CA 92691, USA.（http://www.daylight.com/）

8) Ertl, P., Rohde, B., Selzer, P., *J. Med. Chem.*, **2000**, *43*, 3714-3717.

9) Lipinski, C. A. and et al., *Adv. Drug Delivery Rev.*, **2001**, *46*, 3-26.

10)Accelrys Inc., 9685 Scranton Road, San Diego, CA 92121-3752, USA.（http://www.accelrys.com/）

11) Abe, N., Mamitsuka, H., *Proceedings of the 15th International Conference on Machine Learning (ICML98)*, **1998**, 1-9.

12) Maybridge plc, Trevillett, Tintagel, Cornwall PL34 OHW, UK.（http://www.maybridge.com/）

13) Cerep, 128, Rue Danton, BP 50601, 92506 Rueil-Malmaison, France（http://www.cerep.fr/）