Vol.2011-BIO-26 No.8
2011/9/13

機械学習による EP2 受容体を標的とした化合物の探索手法の開発

小原祥平† 遠里由佳子† 伊藤將弘†

EP2 は，細胞内シグナル伝達にかかわる膜タンパク質のプロスタグランジン E2 の受容体であり，重要な創薬標的である．本研究では，EP2 のヒット化合物の探索法の構築を目指した．具体的には，PubChem BioAssay の化合物のスクリーニングデータに対し，サポートベクターマシン(SVM) とおよび *k*-近傍法を用いて 2 つの予測システムの開発を行った．化合物の構造を表すビット列の特徴選択法として，Filter 法と Kullback-Leibler (KL) 情報量を考慮した．これら 2 つの特徴選択を折衷した結果，SVM と KL 情報量との組み合わせにおいて，77.2%と高い探索精度を示した．

Development of searching method for hit compounds binding EP2 receptor by using machine learning

Shouhei Kohara†, Yukako Tohsato† and Masahiro Ito†

The Prostaglandin E2 receptor as known as EP2, a membrane G protein-coupled receptor, plays an important role in multicellular organisms, and it is there highly desirable to have model that can predict whether a compound interacts for the EP2. Therefore, we developed two machine leaning methods; support vector machines (SVM) and *k* nearest neighbor (*k*-NN), for the computational screening of the EP2, and adapted the techniques for the PubChem BioAssay data of the EP2 receptor and their molecular structures, which are represented by bit strings by MACCS key conversion. Moreover, our method was added two feature extraction procedures, Filter method and Kullback-Leibler (KL) divergence analysis. Finally, our method showed a high accuracy, 77.2%, by using the SVM with the KL divergence.

1. はじめに

G タンパク質共役型受容体の一種である EP2 は，脂質メディエーターのプロスタグランジン E2（PGE2）をリガンドとして神経細胞の刺激の増強や cAMP の上昇などの生理作用を及ぼす [1]．新薬開発の初期段階では，このような生理作用に関わる受容体に対し，十分な活性を有することが知られている化合物の構造や知見を用いて，膨大な化合物データから，活性がある可能性が高い化合物を探索することが重要となる．近年，*in vitro* 実験で得られた活性情報は，PubChem[2]をはじめとする化合物データベースに蓄積され，一般に公開されるようになった．そこで，これらのデータを用いて，高精度な標的受容体と化合物間の活性の予測を目指し，様々な予測法が提案されている [3]．特に，機械学習の手法として，画像認識の技術にも応用され高い認識性能を持つことが知られているサポートベクターマシン（SVM）[4]や，データ間の距離を元にした多数決の原理に基づく *k* 近傍（*k*-NN）法は，膨大なデータの中から注目する特徴や規則を発見し，未知のデータに対して予測を行うことに適している．そこで本研究では，SVM と *k*-NN を用いて，EP2 に結合し，阻害活性を持つ可能性のある化合物を探索する手法の開発を行った．

2. 実験データと提案手法

2.1 EP2 の阻害活性データと構造データの取得

EP2 の阻害活性データを PubChem BioAssay から取得した(AID1422). このデータは，ラットの細胞にヒトの EP2 を発現させるプロモータを注入し，エネルギー共鳴移動法により EP2 の結合から生成する cAMP の濃度を測定して，化合物の EP2 に対する阻害活性を調べたものである．PubChem BioAssay には，約 250.000 もの化合物の EP2 との相互作用の強さが 0~100 と定量的にスコア付けされている．そして，スコアで 0~39 までが「活性なし」，40~100 までが「活性あり」とされ，それぞれの化合物数は 1,253，約 248,000 となっている．そこで，スコア 39 以下の化合物を負例，スコア 40 以上の化合物を正例として用いるために，各 1,253 (計 2,506) 化合物を取得した．ここで負例のデータは，スコア分布に従ってランダムに抽出している．加えて，化合物の構造式のデータを PubChem Compound から取得し，MACCS Key[5]を用いて 166 種類の部分構造の有無を 0 と 1 で表す 166 桁のビット列に変換した．

2.2 予測システムの構成

本研究では，予測システムとして，SVM と *k*-NN の 2 種類を構築する．SVM は，学習データを線形に分類することで，予測したい化合物がどちらに属するのか判定す

†立命館大学 生命科学部 生命情報学科
Department of Bioinformatics, College of Life Sciences, Ritsumeikan University, 1-1-1 Nojihigashi, Kusatsu, Shiga 525-8577, Japan

る．ソフトウェア R の e1071 のパッケージを使用し，C-SVM（別名 Soft Margin SVM）の RBF (Radial Basis Function) カーネルを使用した[4]．一方，*k*-NN では，既知のデータの中から予測する化合物の構造によく似た *k* 個の代表となるデータを選び，それらのデータに付加されている属性（活性あり，もしくは活性なし）の数の多数決で予測データの属性を決定する．ビット列 *x*, *y* 間の距離は式(1)の Tonimoto 係数で求めた．

$$T(x,y)=\frac{N_z}{N_x+N_y-N_z} \qquad (1)$$

ここで，N_x, N_y はビット列 *x*, *y* における 1 のビット数であり，N_z は *x* と *y* がともに 1 のビット数である．$T(x,y)$ は 0 から 1 までの値をとり，1 に近づくほど二つのビット列間の類似度は高い．

2.3 ビット列の特徴選択

入力に使用するビット列のうち，活性の有無の判別に影響を与える重要なビットを選択することによる予測精度の向上を目的として，Filter 法[6)6)]と Kullback-Leibler (KL) 情報量[3)]の 2 種類の特徴選択法を試みた．

Filter 法は，ビットの重要性を式(2)で算出する．

$$\omega_i=\frac{\mu_i(+)-\mu_i(-)}{\sigma_i(+)+\sigma_i(-)} \qquad (2)$$

μ_i と σ_i は，正例(+)もしくは負例(-)に属する化合物のうち，*i* 番目のビットの 1 の数の平均と標準偏差にあたる($1 \leqq i \leqq 166$)．ω_i が大きな正の値を取れば，*i* 番目のビットは正例の判定に重要なビットとなり，大きな負の値を取れば，負例の判定に重要なビットとなる．

一方，KL 情報量は，ビットの重要性を式(3)で算出する．

$$\hat{p}_i^+=\frac{mp_i^+ + p_i^-}{m+1},\quad \hat{p}_i^-=\frac{np_i^- + p_i^+}{n+1},\quad D_i=\hat{p}_i^+\log\frac{\hat{p}_i^+}{\hat{p}_i^-}+\hat{q}_i^+\log\frac{\hat{q}_i^+}{\hat{q}_i^-} \qquad (3)$$

p_i と q_i は，正例(+)もしくは負例(-)に属する化合物のうち，*i* 番目のビットが 1 の確率，および *i* 番目ビットが 0 になる確率 q_i=(1-p_i)であり，*m* と *n* は正例と負例のデータ数にあたる．p_i と q_i は，2 つのクラスのデータ数による影響を考慮し補正されている．D_i 値が大きければ，正例と負例の分類に重要なビットとなる．

3. 実験と結果

機械学習である SVM と *k*-NN，特徴選択法である Filter 法と KL 情報量のすべての組み合わせに対して，10-fold cross validation で予測の評価を行った(表 1)．その結果，SVM で，ビット列の特徴選択法による予測精度の向上がすべての評価指標においてみられ，なかでも KL 情報量との組み合わせが最も良い予測精度を示した．実際に，KL 情報量では 11 のビットが，Filter 法では 14 のビットが予測に重要でないビットとして削られている．一方，*k*-NN では，ビット列の特徴選択の明らかな効果がみられなかった．

表 1 分類器とビット列の特徴選択法の組み合わせ評価結果

分類器		特徴選択法	Accuracy	Sensitivity	Specificity	Precision
SVM		-	0.754	0.730	0.778	0.767
		Filter 法	0.769	0.743	0.782	0.777
		KL 情報量	0.772	0.748	0.896	0.792
k-NN	*k*=5	-	0.735	0.690	0.770	0.756
	k =7	Filter 法	0.738	0.712	0.763	0.751
	k =9	KL 情報量	0.740	0.712	0.768	0.754

4. おわりに

本研究では，予測システムとビット列の特徴選択法の組み合わせを，PubChem BioAssay に登録された EP2 の活性データに適用することで，その有効性を検証した．その結果, SVM と KL 情報量での組み合わせで最も良い予測精度が示された. 今後は，構造のどの部分が EP2 の阻害活性に影響を与えているかを調査し，手法のさらなる改良を行う予定である．

参考文献

1) 清水孝雄編: 脂質生物学がわかる 脂質メディエーターの機能からシグナル伝達まで, 羊土社, (2004).
2) Eric W. Sayers, *et al.*: Database Resources of the National Center for Biotechnology Information, *Nucleic Acids Research*, Vol. 38, pp. D5–D16 (2010).
3) Britta Nisius, and Juragen Bajorath: Reduction and Recombination of Fingerprints of Different Design Increase Compound Recall and the Structural Diversity of hits: *Chem Biol Drug Des*, Vol. 75, pp. 152–160 (2010).
4) Alexandros Karatzogolou, David Meyer, and Kurt Hornik: Support Vector Machines in R: *Journal of Statistical Software*, Vol. 15, No. 9, pp. 1–28 (2006).
5) Chun Wei Yap: Software News and Update PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptor and Fingerprints, *J Comput Chem*, Vol. 32, pp. 1466–1474 (2011).
6) Derick C. Wis, Donald P. Visco Jr, and Jean-Loup Faulon: Data mining PubChem Using a Support Vector Machine with the Signature Molecular Descriptor: Classification of Factor XIa Inhibitors, *Journal of Molecular Graphics and Modling*, Vol. 27, pp. 466–475 (2008).