Drug Screening of GPCR Using Active Learning

Yukiko Fujiwara[1]
y-fujiwara@db.jp.nec.com
Yoshiko Yamashita[1]
yoshikoy@ct.jp.nec.com
Tsutomu Osoda[1]
osoda@aj.jp.nec.com
Minoru Asogawa[1]
m-asogawa@bq.jp.nec.com
Shun Doi[1]
s-doi@cb.jp.nec.com
Masaaki Asao[2]
asao@tanabe.co.jp
Emi Kushiyama[2]
emi@tanabe.co.jp
Kazuya Nakao[2]
k-nakao@tanabe.co.jp
Masataka Kuroda[2]
m-kuro@tanabe.co.jp
Kazuteru Wada[2]
k-wada@tanabe.co.jp
Takanori Ogaru[2]
ogaru@tanabe.co.jp
Chiaki Fukushima[2]
chiaki@tanabe.co.jp
Ryo Shimizu[2]
ryo@tanabe.co.jp

[1] Bioinformation, Fundamental Research Laboratories, NEC Corporation, 5-7-1 Shiba, Minato-ku, Tokyo 108-8001, Japan
[2] Tanabe Seiyaku Co., LTD., 3-16-89 Kashima, Yodogawa-ku, Osaka 532-8505, Japan

Keywords: active learning, drug screening, GPCR

1 Introduction

In drug screening, active compounds (actives) bound to targets are found in large collections of compounds. Typically, between several hundred thousand to a million compounds are examined. It is highly expensive to test all compounds in random (Random), therefore limited biochemical tests are performed in searching for actives. Consequently, chemists start with some initial tested set and iteratively choose sets of compounds that are closest to previously known actives, where the similarity is usually calculated using the Tanimoto coefficient (Tanimoto). To obtain a high number of actives, we applied active learning to drug screening. One of the active learning approaches is "query by bagging (QBag)" proposed by Abe and Mamitsuka [1].

2 Method and Results

In the computer experiments, the compounds bind to the biogenic amine receptors of G protein-coupled receptor (GPCR) [3] are chosen as actives from Pharmaprojects (2002.03) [5] and the other compounds are chosen as inactives from Available Chemicals Directory (ACD 2002.10) [4]. The total number of compounds was 214,375 and the number of actives was 1,461. The descriptors are 166 MDL Molskey [2].

For drug screening, the modified descriptor selection strategy in QBag (QBagDS) was added. Descriptors were selected by resampling, according to uniform distribution. The obtained active ratios vs. tested compounds of the averages of 10-fold cross validation is shown in Fig. 1(a). In this figure, QBagDS was superior to QBag, Tanimoto and Random. For example, after 20,000 compound tests, QBagDS obtained 1.6% more actives than QBag, and 13.8% more actives than Tanimoto. This experiment showed superiority of active learning and usefulness of the descriptor selection.

For structural development in drug discovery, finding diverse actives was vital. Consequently, though QBag is a boundary selection strategy, we have suggested a different selection strategy (QExpDS). Data was selected with exponentially decreasing probability of active scores. High score data

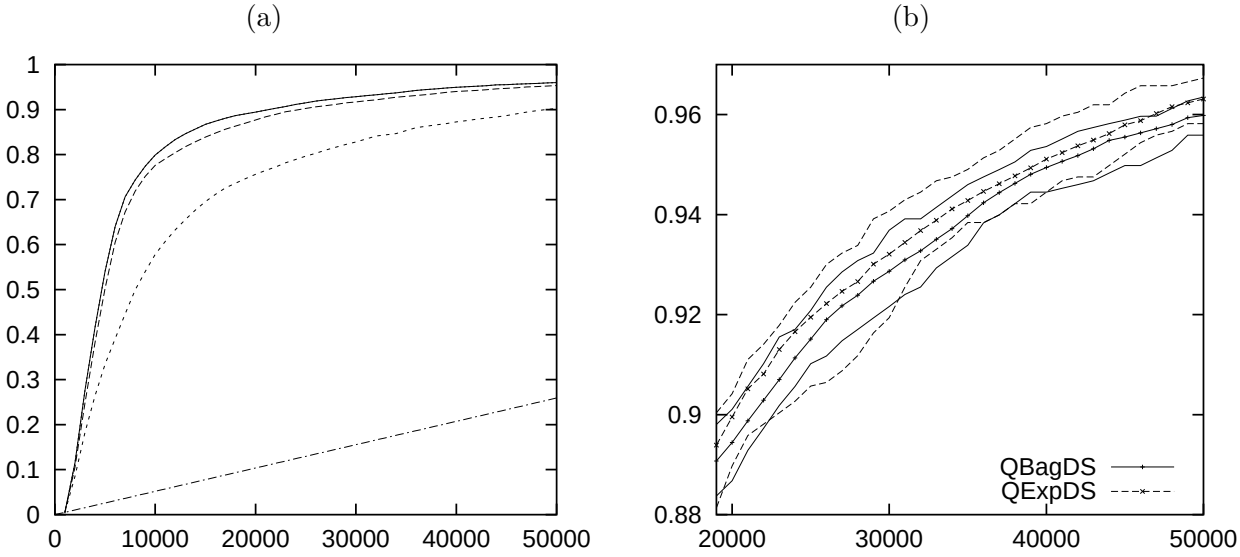

Figure 1: Obtained actives ratio vs. the number of tested compounds. (a) Four methods were plotted: QBagDS (solid, best performance line), QBag (dotted, second best line), Tanimoto (fine dotted, third best line) and Random (worst performance line). (b) Two methods were plotted: QBagDS (solid line), and QExpDS (dotted line).

was mainly selected, however, low score data also appeared. The only difference between QBagDS and QExpDS was the data selection strategies. The results of the averages, and the best and the worst of ten runs are shown in Fig. 1(b). QExpDS was slightly superior to QBagDS. For example, after 20,000 compound tests, it obtained 0.6% more actives than the averages in QBagDS. This indicates that more diverse actives are found by the exponential selection than in the boundary selection.

Acknowledgments

The authors would like to thank to PJB Publications Ltd. and MDL Information Systems Inc. for providing the database. The authors also would like to thank to Dr. Kenji Yamanishi and his group in NEC for providing active learning programs for reference and Dr. Hiroki Shirai in Yamanouchi Pharmaceutical Co., Ltd. for useful advice.

References

[1] Abe, N. and Mamitsuka H., Query learning strategies using boosting and bagging, *Proc. 15th Inter. Conf. on Machine Learning*, 1998.

[2] Durant, J.L. *et al.*, Reoptimization of MDL keys for use in drug discovery, *J. Chem. Inf. Comput. Sci.*, 42(6):1273–1280, 2002.

[3] Horn, F. *et al.*, GPCRDB: an information system for G protein-coupled receptors, *Nucleic Acids Res.*, 26:275–279, 1998.

[4] `http://www.mdli.com/` (MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577, USA.)

[5] `http://www.pjpubs.com/` (PJB Publications Ltd., 18/20 Hill Rise, Richmond, Suvvey TW10 6UA, UK.)