Efficient Drug Screening Using Active Learning

By Minoru ASOGAWA,* Tsutomu OSODA,† Yukiko FUJIWARA‡ and Yoshiko YAMASHITA‡

ABSTRACT At the lead generation stage for drug discovery, the efficient finding of hit compounds is a key to success. In this paper, we have applied the active learning method as an effective compound screening method and have shown its effectiveness by both computer simulations using known chemical data and actual wet experiments. With regard to the computer simulations, it is shown that one fifth screening is enough for finding ninety percent of all hit compounds. Our method could lessen eighty percent of actual wet experiments. We have performed actual binding experiments, and have also shown that active learning could find almost all 'hits' with a reduced number of actual binding experiments.

KEYWORDS **Drug screening, Active learning, Compounds**

1. INTRODUCTION

At the stage of lead generation for drug discovery, both the combinatorial chemistry method and the high throughput screening method are being successfully utilized. Such methods are able to discover several hit compounds from a huge chemical library that usually consists of from one hundred thousand to one million chemical compounds. However, one of the drawbacks of these methods is that the cost for screening is enormous as the screenings require actual wet experiments.

Developing a hit compound to a lead compound is performed by medicinal chemists and the procedure is carried out based on the knowledge and experience of such chemists. When many hit compounds become available, the development procedure for a hit compound to a lead compound could become much easier. Therefore, it is essential to find many hit compounds for drug discovery.

In this paper, we proposed an active learning method as an efficient chemical screening method and show its effectiveness by both computer simulations using known chemical data and actual wet experiments.

With regard to the computer simulations, it is shown that one fifth of the screenings is enough for finding ninety percent of all hit compounds, compared to the random screening method. Even though, the random screening method is widely used at the present time, our proposed method could reduce the number of the actual wet experiments by eighty percent.

2. CHOICE OF TARGET PROTEIN

In this paper the G-protein coupled receptor (abbreviated as GPCR) is chosen as a target protein. Therefore, the aim of our proposed system is to find chemical compounds that can bind to GPCR. The GPCR family is one of the most well known and important drug target proteins, and continues to be very promising.

Among GPCRs, adrenalin, dopamine, and certonin receptors, which bind to biogenic amines in common, are well studied and a large amount of knowledge has been accumulated about them. Therefore, we have chosen the biogenic amines receptors as a target protein.

3. METHOD

3.1 How to Make a Database for Simulation

To create "hit" chemical compounds, we used the Pharmaprojects database and selected chemical compounds which are annotated as binding to biogenic amine receptors. 1,461 chemical compounds were chosen to be treated as "hit" compounds.

To make "non-hit" compounds, we used the Available Chemicals Directory (abbreviated as ACD) and selected chemical compounds based on several drug-likeliness criteria. 212,914 chemical compounds were

*Fundamental and Environmental Research Laboratories; Solution Development Laboratories
†Solution Development Laboratories
‡Fundamental and Environmental Research Laboratories

chosen and treated as "non-hit" compounds.

Note that, there is no guarantee that these "non-hit" compounds do not bind to biogenic amine receptors. The results of our wet experiments indicated that dozens of chemical compounds actually bind to biogenic amine receptors.

3.2 How to Represent Chemical Compounds in Simulations

Both "hit" and "non-hit" compounds are expressed with 215 descriptors representing their features. These descriptors consist of two groups. One group represents chemical compound substructures, such as number of double bonds, benzene rings and so on. There are 166 descriptors in this group. The other group represents physico-chemical properties of the compounds, such as molecular weight, polar surface area, estimated hydrophobisity and so on. There are 42 descriptors in this group. Consequently, the chemical compounds are expressed with 215 descriptors and their labels, indicating "hit" or "non-hit."

4. ACTIVE LEARNING METHOD

Usually, learning is performed with a prepared learning database. In learning, all contents of the database are used to enable learning all together. The learning system 'passively' obtains learning data, and learns it. In active learning, the learning system 'actively' chooses data that should be learned[1]. The candidate data contains only descriptors and no information about its label. When data is chosen, its label is given based on additional experimentation or from the prepared learning database for computer simulation cases.

The chosen data is accumulated to the database for learning and learning is performed with all of the contents of this database. In the active learning, the system has control over which data should be used for learning to achieve high prediction accuracy with a reduced number of labeled data. Especially for drug screening, the acquisition of labels for compounds requires expensive wet experiments (**Fig. 1**). A reduced number of wet experiment is desirable. For active learning, in our experiment, a Query by Bagging (Qbag) is used, which places no constraints on the learning method. In Qbag, several sub-learning machines are used. Each sub-learning machine is trained by the database. Initially the database for learning consists of random selection from the learning database. This selection is different for each sub-learning machine, so that the learned concepts are also different. Therefore, when data is presented to these sub-learning machines, their predictions may differ. Data that yields the greatest disagreement among sub-learning machines predictions' is chosen as a data to be learned.

In the case of drug screening, there exists only a few dozens of hits, compared to one million of non-hits. Therefore, we have introduced a weighting method, to compensate for this numbers imbalance.

5. COMPUTER SIMULATION

For the computer simulations, a ten-fold procedure is performed to evaluate the efficiency of the active learning. With the random screening method, to obtain ninety percent of hit compounds from 212,914 chemical compounds, about 170,000 labels of

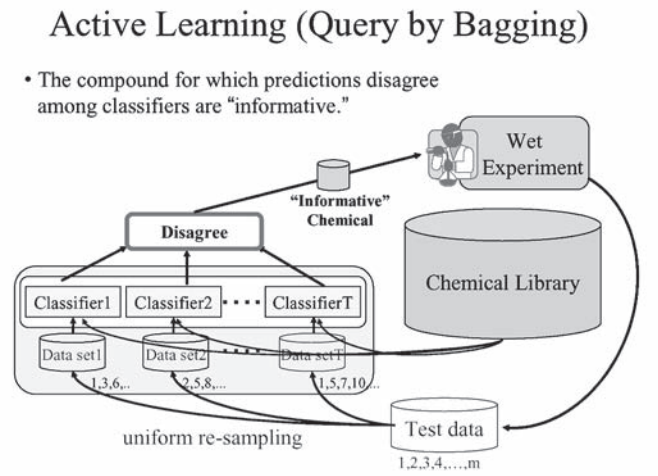

Fig. 1 Basic mechanism of active learning.

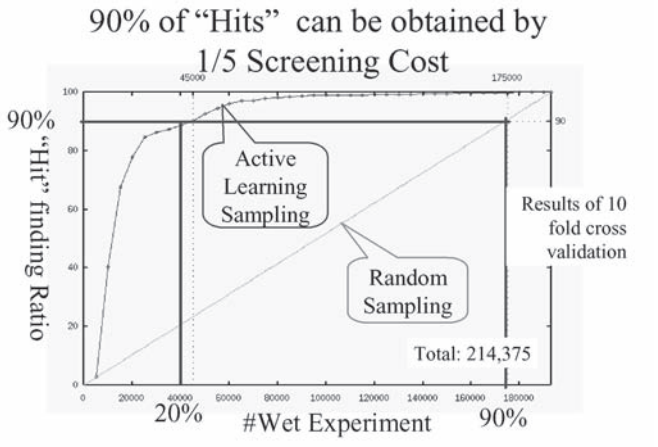

Fig. 2 Result of computer simulation.

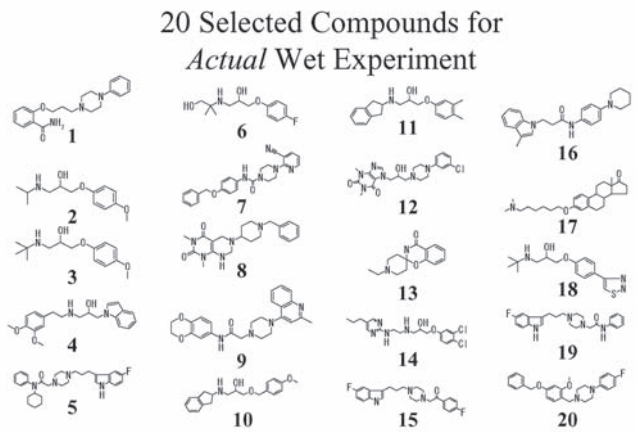

Fig. 3 Selected compounds.

8 "Hits" Compounds

For Biogenic amine receptor binding at 10^{-6}M

No	System Prediction	α1 Adrenaline	α2 Adrenaline	Muscarine	Serotonin
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

Unit of number is a percentage of binding

Fig. 4 Experimental results.

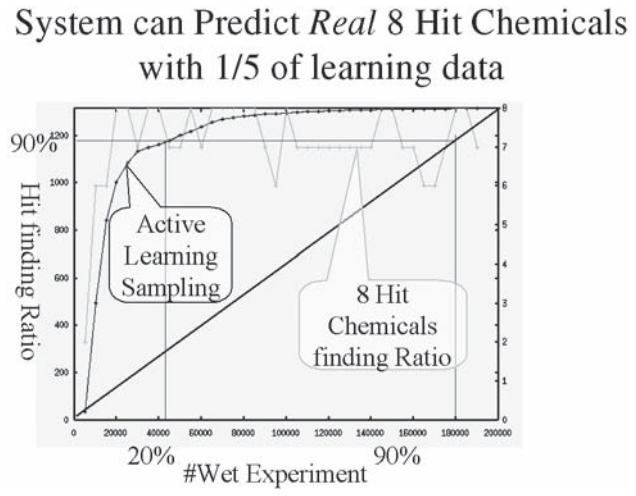

Fig. 5 Experimental results compared with computer simulation results.

chemical compounds from the learning database are necessary (**Fig. 2**). On the other hand, with the active learning method, only about 4,300 labels of chemical compounds suffices. This implies that the active learning method can save up to eighty percent of screening for finding ninety percent of all hit compounds.

6. ACTUAL WET EXPERIMENT

After learning all of the learning data, we have chosen twenty chemical compounds, which show high "hit" predictions, and have measured their binding affinities to biogenic amines receptors (**Fig. 3**). The binding affinities are experimented twice and the mean of two experimental results is used. Eight chemical compounds have shown more than fifty percent binding affinities under ten micro molar concentration conditions (**Fig. 4**). With regard to the computer simulation, it is also shown that the active learning could find almost all of these 'hits,' with about 170,000 labels of chemical compounds information (**Fig. 5**).

7. CONCLUSION

At the stage of lead generation for drug discovery, an efficient finding hit compounds method is a key to success. In this paper, we have applied the active learning method as an effective compound screening method and have shown its effectiveness by both computer simulations using known chemical data and actual wet experiments. For the computer simulations, it is shown that a one fifth screening is enough for finding ninety percent of all hit compounds. Our method could lessen eighty percent of actual wet experiments. We have performed actual binding experiments, and have also shown that the active learning could locate almost all of the 'hits' with a reduced number of actual binding experiments.

ACKNOWLEDGMENTS

This is a partial result of joint research with the Tanabe Seiyaku Co., Ltd., Japan. We would like to thank the collaborators, Masaaki Asao, Emi Kushiyama, Kazuya Nakao, Masataka Kuroda, Kazuteru Wada, Takanori Ogaru, Chiaki Fukushima, and Ryo Shimizu of the Tanabe Seiyaku Co., Ltd., Japan, for their dedicated work and fruitful discussions. We would also like to thank the Japan MDL Information Systems Corp. and PJB Publications Ltd. for a generous allowance in database usage

for this research. We also would like to thank Prof. Hiroshi Mamitsuka, Dr. Kenji Yamanishi and Dr. Hiroki Shirai for fruitful discussions.

REFERENCE

[1] "Active Ensemble Learning - Applications to Data Mining and Bioinformatics," H. Mamitsuka, and N. Abe, IEICE Transactions, J85-DII (5), pp.717-724, 2002.

Received March 24, 2005

* * * * * * * * * * * * * * * * *

Minoru ASOGAWA received his M.E degree in physical electronics from Tokyo Institute of Technology and M.S degree in computer science from the University of Southern California. He joined NEC Corporation in 1986, and was a visiting researcher at Carnegie-Mellon University from 1992 to 1993. He has been engaged in the research and development of bioinformatics at the Central Research Laboratories of the NEC Corporation.

Tsutomu OSODA received his M.E and M.S degrees in information science from University of Tokyo. He joined NEC Corporation in 1995. He has been engaged in the research and development of numerical analysis and machine learning algorithms in the Central Research Laboratories of NEC Corporation.

Yukiko FUJIWARA received her M.S degree in mathematics from Waseda University in 1993. She joined NEC Corporation in 1993, and now belongs to the Fundamental and Environmental Research Laboratories. She has engaged in research on bioinformatics.

Ms. Fujiwara is a member of the Biophysical Society of Japan and the Molecular Biology Society of Japan.

Yoshiko YAMASHITA received her master degree in information and system engineering from Chuo University in 2000, respectively. She joined the Central Research Laboratories of NEC Corporation in 2000. She has engaged in research on bioinformatics in the Central Research Laboratories of NEC Corporation.

She is a member of the Information Processing Society of Japan and the Molecular Biology Society of Japan.

* * * * * * * * * * * * * * * * *