# Property-based Design of Ion-Channel-Targeted Library

**Ji Young Ahn[1] Ky-youb Nam[1] Byung Ha Chang[2] Jeong Hyeok Yoon[2] Seung Joo Cho[3] Hun Yeong Koh[3] Kyoung Tai No[4]**

[1]*Division of Drug Discovery, Research Institute of Bioinformatics & Molecular Design, Seoul, Korea*
[2]*Department of Chemical Genomics, Information & Data Revolution Technology Inc., Gyounggido, Korea*
[3]*Biochemicals Research Center, Division of Life Science, Korea Institute of Science and Technology, Seoul, Korea*
[4]*Department of Biotechnology, Yonsei University, Seoul, Korea*
*Email : olive@bmdrc.org*

**ABSTRACT**: The design of ion channel targeted library is a valuable methodology that can aid in the selection and prioritization of potential ion channel-likeness for ion-channel-targeted bio-screening from large commercial available chemical pool. The differences of property profiling between the 93 ion-channel active compounds from MDDR and CMC database and the ACDSC compounds were classified by suitable descriptors calculated with preADME software. Through the PCA, clustering, and similarity analysis, the compounds capable of ion channel activity were defined in ACDSC compounds pool. The designed library showed a tendency to follow the property profile of ion-channel active compounds and can be implemented with great time and economical efficiencies of ligand-based drug design or virtual high throughput screening from an enormous small molecule space.

## 1 INTRODUCION

There are many chemicals with the characteristics of drug -like that are possible to synthesize $10^{60}$ chemicals theoretically [1]. But all compounds could not test the high throughput screening to identify the interactions between the ligands and the target protein.

Several papers have discussed the thesis that drugs have distinct properties differentiating them from other chemicals. Lipinski's 'rule of five' [2] is not directly used to classify compound into a dug or a non-drug but provides a heuristic guide for determining if a compound will be orally bioavailiable. It was found that in a high percentage of World Drug Index (WDI) database [3] in terms of molecular weight, number of hydrogen-bond acceptor and donor, and calculated logP. Ajay *et al.* described a solution in the design of a central nervous system (CNS)-active library based on a neural network classification procedure [4]. The neural network method correctly classified 90% of the compounds from Comprehensive Medicinal Chemistry (CMC) database, while it misclassified only 10% of the compounds in the Available Chemical Directory (ACD) database. Konstantin *et al.* classified the molecules into 'G-protein- coupled receptor (GPCR)-ligand-like' and 'non-GPCR-ligand-like' from the Ensemble Database [5], [6]. The method employed a set of descriptors for encoding the molecular structures and by training of a neural network for classifying the molecules. Manallack *et al.* tried to identify the screening candidates for kinase and GPCR using BCUT descriptors [7] and neural networks. Andrew and Sophie *et al.* showed that chemical structures form different target spaces appear to occupy certain area. They say that if the intended target is well characterized, potential compounds can be compare with compounds have been developed successfully into drugs or those known activity [1].

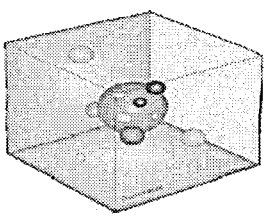

**Figure 1.** Exploring chemical space [1].

In this paper, we present a methodology of focused library design of an ion channel active compounds collected from CMC and MDDR database through the profiling of physicochemical properties. The fundamental study such as ion channel is essential to cerebral disease, epilepsy, neuropatic pain, and so on [7]. With increasing throughput of the drug discovery process and the need for cost efficiency, a design for a library must also strive to be higt hit rate and cost-effective. In this study, the goal of focused library design is to provide high structural diversity seed for ion channel screening related to cerebral disease while constraining pertinent physicochemical properties to suitable range for small molecule drugs.

## 2 METHODS

### 2.1 Data and Descriptor

The 93 known ion channel compounds were selected from the MDL Drug Data Report (MDDR, version 2003.2) which contains over 140000 compounds and the Comprehensive Medicinal Chemistry (CMC) database (version 2002.1) which contains 8225 compounds. Structures were extracted according to the assigned activity class. We assumed that a molecule is ion channel ligand related to cerebral disease if it contains the "ion channel" and 17 keywords for CNS activity in the activity class field [4]. The CMC database contains drug-like molecules, but MDDR database contains over 90% of early discovery stage compounds that is biological testing and may not be drug-like. To design the library that is useful from the step of lead compounds selection, not only launched drug but also above phase I trial compounds were considered a range of known compounds. Chemical pool was from Available Chemical Directory Screening (ACDSC) database that consists of

about 2.2 million commercial available compounds.

A set of 188 constitutional, electrostatic, geometrical, and physicochemical descriptors that was calculated from 2D representations of molecules was explored. All descriptors were calculated at pH 7.4 using preADME$^{TM}$ software tool [9]. It is a program for prediction of *in silico* ADME (absorption, distribution, metabolism and excretion) properties and developed in-house at the BMD with various physicochemical properties [10]. It is easy to deal with large number of dataset fast.

## 2.2 Analyses

To handling the data with ease, file format of compound set was transferred from SD file to binary data file format. Figure 2 shows a schematic illustration of designing a focused library. The Principle component Analysis (PCA), Clustering, and Similarity analysis were accomplished by Cerius$^{2\,TM}$ program [11].

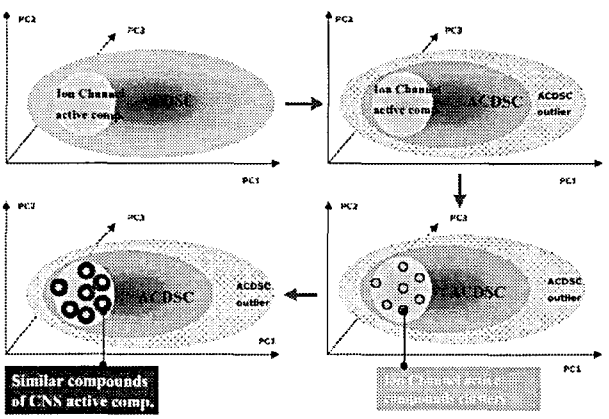


**Figure 2.** Schematic illustration of designing an ion channel focused library.

### 2.2.1 Profile analysis

To discriminate between the ACDSC data and the ion channel ligands, two methods were used. First, the physicochemical distributions of ACDSC and ion channel compounds were shown using histogram. This analysis can be made of an intuitive thinking about showing the difference between two sets. Figure 3 shows that one descriptor, 2D VSA-HA, represent a broad and non focused distribution so that it is not good character to present the ion channel. The other descriptor, SKlogP, is well distributed and focused so that it was selected as descriptor to present the property of ion channel.

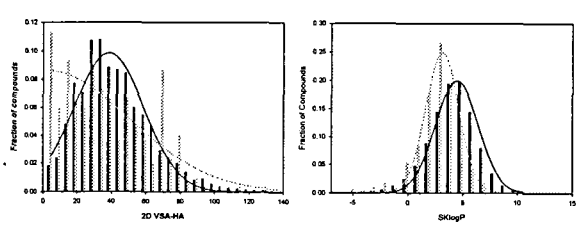


**Figure 3.** The comparison of two distribution charts. 2D VSA-HA is sum of 2D *van der Waals* surface area of hydrogen bond acceptor and SKlogP is logP calculate by preADME software. Black bar and line is the ACDSC data and gray bar and line is ion channel ligands.

t'-statistics can help with the quantitative comparison to complement the comparison using histogram. t'-test were calculated in accordance with the formula

$$t' = (X_1 - X_2)/\sqrt{\sigma_1^2/n + \sigma_2^2/n} \qquad (1)$$

where for each of the two compounds sets, ACDSC and known ligands, $X$ is the mean, $\sigma^2$ is the variance, $n$ is the sample size, and 1 and 2 denote the corresponding set [5].

For the compounds sets studied here, t'-test showed the significance of the difference between the two distributions for topological polar surface area, number of aromatic rings, partial negative surface area 1$^{st}$, hydrogen bonding acceptor and molecular weight in order value. In table 1, PPSA1 shows negative t' values. It means that the distribution of ion channel ligands shifts more high value than ACDSC compounds.

| Descriptor | t' | Δ |
|---|---|---|
| TPSA | 8.31 | 25.65 |
| A.R. | 7.46 | 0.545 |
| PNSA1 | 6.08 | 26.9 |
| HA | 5.83 | 0.95 |
| MW | 4.64 | 45.1 |
| LDI | 3.90 | 0.04 |
| PPSA1 | -2.37 | -3.92 |

Δ is difference mean value between ACDSC and ion channel ligands; TPSA is topological polar surface area; A.R. is number of aromatic rings; PNSA1 is sum of partial negative surface area; HA is the number of hydrogen bond acceptor; MW is the molecular weight; LDI is the local dipole index PPSA1 is sum of partial positive surface area

**Table 1.** Five descriptors having the high t' values and 1 descriptors having a negative t' value.

t'-statistics are applicable for testing the significance of the difference between the two distributions for each descriptor, that is, a descriptor having large value of t'-test shows large difference between two data set and can be regarded as a distinguishing trait.

### 2.2.2 Data boundary

Whereas ACDSC data show symmetrical bell-shaped property curve, ion channel ligands present leaned curve with tail. For exclusion of far from considered range, the standard deviation usually was used if data were supposed to have a normal distribution, but cannot be applicable to leaned data. While standard deviation is less sensitive method to outliers, the squared dependence on the deviation from the mean still weights the outlier values rather heavily. To exclude a slight distribution of curve tail and focus the range on trait of ion channel property having Gaussian density, tail range was removed by median absolute deviation. The median absolute deviation (MAD) is the median of the distances between each data point and the overall median for the dataset. And it is removes the squared dependence on the deviation from the mean, so the method is much less sensitive to outliers [12]. It is defined as

$$MAD = |x - M|/D \qquad (2)$$

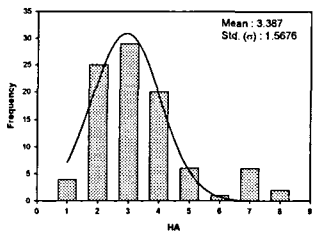


**Figure 4.** The distribution profile of hydrogen bond acceptor of ion channel data. It shows unsymmetrical distribution about mean.

where $x$ is the stands for descriptor value in a population, M is the median value of the population, and D is the median of $|x-M|$.

For example, in figure 4, the range of Mean±2σ is from 0.2518 to 6.5222. But include range by MAD is from 0 to 5. As a result of boundary by MAD, ACDSC data is 1414078 compounds and ion channel ligand data is 83 compounds.

### 2.2.3 PCA analysis

Principle component analysis (PCA) [13] performs a linear transformation of the n-vectors, reduces the dimension of the descriptor space, removes the highly correlated descriptors, and estimates a normalize set of $p$ principal components. Five principle components accounting for 88% of descriptor variance were determined. ACDSC compounds and ion channel ligands were assembled by representing molecules in the five dimensional space spanned by the orthogonal PCs scaled to unit variance.

### 2.2.4 Cluster analysis

Hierarchical clustering is the process of subdividing a group of compounds into clusters of compounds that exhibit a high degree of both intracluster similarity and intercluster dissimilarity [14]. Figure 5 shows the output as a dendrogram represented by an average linkage method of hierarchical clustering. Figuer 6 shows the 14 clusters that are formed through the cluster analysis and represent the physicochemical area of ion channel ligands.

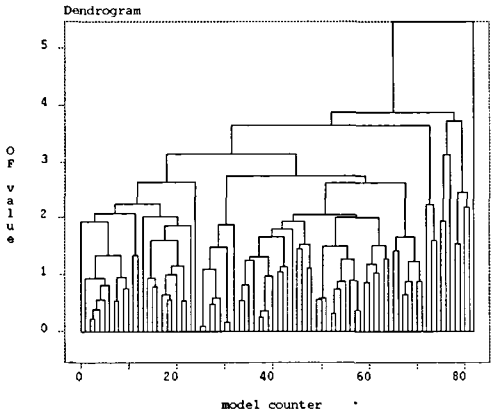


**Figure 5.** Dendrogram of hierarchical clustering of ion channel ligands.

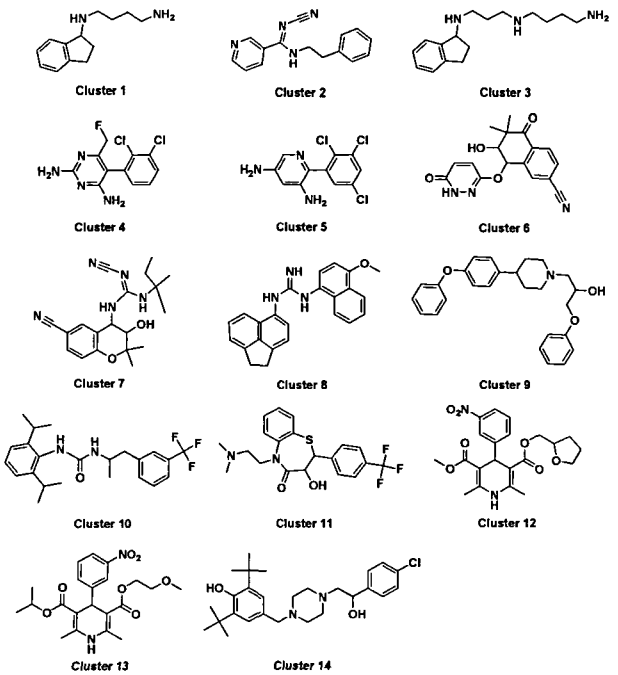


**Figure 6.** 14 clusters of ion channel ligands.

### 2.2.5 Similarity analysis

14 clusters were made use of seeds to collect the ion channel ligand-like compounds. The half of the smallest distance among clusters, 0.641, was criterion for the distance based similarity analysis. The distance-based similarity was calculated using the elucidean distance matrix, and the tanimoto coefficient.

The ion channel-focused library consisted of total 14450 compounds from ACDSC database.

## 3 RESULTS

In this work, 188 descriptors were calculated but seven descriptors clearly reveal the specific character of ion channel ligand distinguishable from chemical pool by histogram and t'-statistics. The six descriptors, TPSA, A.R., PNSA1, HA, MW and LDI shift toward lower value but the PPSA1 shift toward higher than distribution of ACDSC compounds. From the opposite result between the PNSA1 and PPSA1, it can be supposed that a compound needs more positive surface such as piperazine fragment in order to have an ion channel activity.

The outlier ranges of ion channel properties were excluded using MAD from two data sets and then the ion channel data set seems to have Gaussian density. If the tail of upper range is not removed, the designed library will show broad deviation. After PCA analysis, two compounds set were assembled by five PCs. The selected 14 clusters by an average linkage method of hierarchical clustering represent only physicochemical space of ion channel ligands. The compounds collected from 14 ion channel clusters within same distance, 0.641, on ACDSC chemical pool through the distance-based similarity analysis.

The ion channel-focused library had a tendency to resemble the property distribution of known ion channel ligands than ACDSC dataset. Table 2 shows that ion channel focused library have smaller standard deviation than

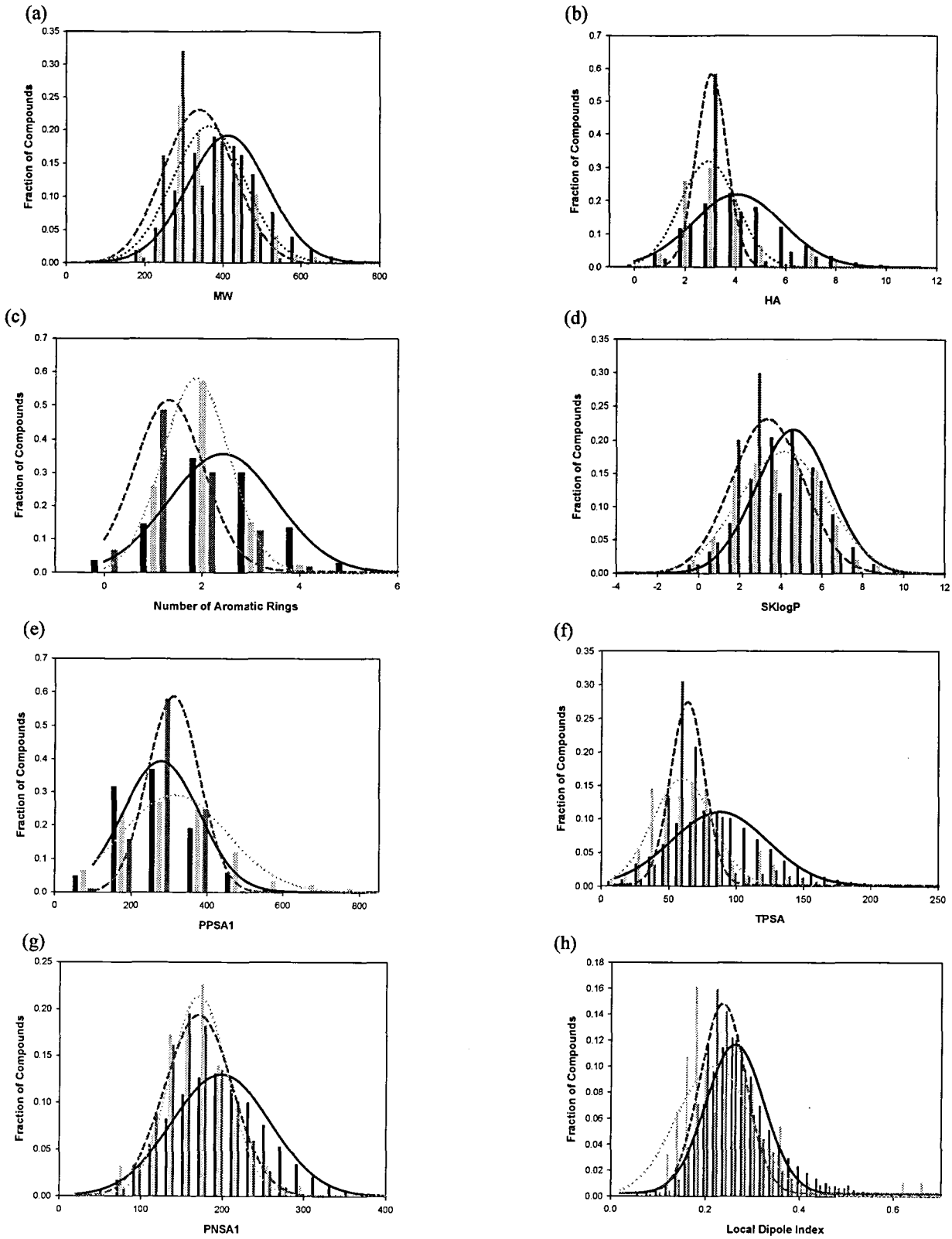


**Figure 7.** Molecular descriptor distributions of ACDSC compounds (black bar and solid line), known ion channel compounds (light gray and doted line), and designed ion channel library (gray and dash line). (a) Molecular weight; (b) Number of hydrogen bond acceptor; (c) Number of aromatic rings; (d) SKlogP; (e) Sum of partial polar surface area 1$^{st}$ ; (f) Topological polar surface are; (g) Sum of partial negative surface area 1$^{st}$ ; (h) Local dipole index.

ACDSC or known ligands. The four descriptors, HA, PPSA1, PNSA1 and TPSA didn't reveal a difference of peak distribution, but MW, number of aromatic rings and SKlogP have lower peak distribution comparing with one of known ion channel ligands (figure 7). This tendency means that the designed library has many compounds which have possibilities of passive transport through the BBB tight junction.

| Descriptor | ACDSC | | Known ligands | | Designed library | |
|---|---|---|---|---|---|---|
| | Mean | Std. | Mean | Std. | Mean | Std. |
| MW | 397 | 109 | 351.9 | 93.54 | 342.6 | 76.02 |
| HA | 4.33 | 1.94 | 3.38 | 1.57 | 3.28 | 1.14 |
| A.R. | 2.48 | 1.16 | 1.935 | 0.70 | 2.08 | 0.68 |
| SKlogP | 4.09 | 1.99 | 3.67 | 1.84 | 3.12 | 1.60 |
| PPSA1 | 244.4 | 102.8 | 283.6 | 138.7 | 206.3 | 84.46 |
| TPSA | 87.68 | 38.63 | 62.03 | 29.75 | 67.02 | 23.32 |
| PNSA1 | 192.6 | 63.83 | 165.7 | 42.64 | 164.3 | 39.94 |
| LDI | 0.27 | 0.089 | 0.23 | 0.093 | 0.25 | 0.075 |

Std. is a standard deviation, MW is the molecular weight; HA is the number of hydrogen bond acceptor; A.R. is the number of aromatic rings; PPSA1 is sum of partial positive surface area; TPSA is topological polar surface area; PNSA1 is sum of partial negative surface area; LDI is the local dipole index.

**Table 2.** The statistics of ACDSC compounds, known ion channel ligands, and designed ion channel library

# 4 CONCLUSIONS

The cerebral targeted drug must penetrate the blood brain barrier (BBB). The BBB consists of a continuous layer of endothelial cells joined by tight junctions and has efflux transporter such as P-glycoprotein at the cerebral vasculature. It is reason why another criterion will be generated for CNS-targeted compounds.

The area of known ion channel ligands related cerebral disease occupied certain part that could be distinguished from the chemical pool. After identifying the characteristic descriptors of known ion channel ligands, for example topological polar surface area, number of aromatic rings, partial negative surface area 1st, and so forth, ion channel focused library to research new or novel lead compounds could be designed with commercial available compounds having the same property range as known ones.

Designed library showed the property distribution profile that had similar distribution of known ligands with more narrow deviation value without any penalty function.

It could be made use of virtual screening library or bio-screening seed as wanted number through the diversity analysis relating to ion channel-targeted cerebral disease while constraining pertinent physicochemical properties to suitable range and diverse structure. These methodologies help select molecules with a significantly enhanced hit rate from vast chemical pool without lots of the time and efforts of synthesizing and testing them in the early stage of research and discovery of drug development.

## ACKNOWLEDGEMENT

This research was supported by a grant from the Center for Biological Modulators of the 21st Century Frontier R&D Program and Vision 21 Program from Korea Institute of Science and Technology, the Ministry of Science and Technology, Korea.

## REFERENCES

[1] S. P. Zeman, Exploring biological space, url:http://www.nature.com/

[2] C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug Delivery Rev. 23:3-25, 1997.

[3] World Drug Index, Derwnt Information, London, 1997

[4] Ajay, G. W. Bemis, M. A. Mrucko, Designing libraries with CNS activity, J. Med. Chem., 42:4942-4951, 1999

[5] V.B. Konstantin, E. T. Sergey, A. L. Stanley, I. O. Lang, A. I. Andrey, and P. S. Nikolay, Property-based design of GPCR-Targeted Library. J. Chem. Inf. Comput. Sci. 42:1332-1342, 2002.

[6] V.B. Konstantin, A. L. Stanley, V. S. Andrey, E. T. Sergey, A. I. Andrey, and P. S. Nikolay, Structure -based verus property-based approaches in the design of G-protein-coupled receptor-targeted libraries. J. Chem. Inf. Comput. Sci. 42:1332-1342, 2002..

[7] F. R. Burden, Molecular identification number for sub-structure searches. J. Chem. Inf. Comput. Sci. 29: 225-227, 1989.

[8] S. Jo, K. H. Lee, S. Song S, Y. K. Jung and C. S. Park, Identification and functional characterization of cereblon as a binding protein for large-conductance calcium-activated potassium channel in rat brain. J. Neurochem. In press, 2005.

[9] preADME, v 1.0.1, Research institute of bioinformatics and molecular design (BMD), Korea. url: http://preadme.bmdrc.org

[10] R. Todeschini and V. Consonni. Handbook of molecular descriptors, Wiley-VCH, Weinheim, Germany, 2000.

[11] Cerius 2, v 4.8 - Accerly Inc. : 9685 Scranton Rd., San Diego, CA 9212-3752, U.S.A. url: http://www.accerlys.com

[12] B. A. Tounge, L. B. Pfahler, and C. H. Reynolds, Chemical information based scaling of molecular descriptors: a universal chemical scale for library design and analysis, 42:879-884, 2002.

[13] W. G. Glen, W. J. Dunn, D. R. Scott, Principal component analysis and partial least squares regression. Tetrahedron Comput. Methdol. 2:349-376, 1989.

[14] V. Schoonjans, D. L. Massart, Combining spectroscopic data (MS, IR): exploratory chemometic analysis for charactering similarity/diversity of chemical structures. J. Pharm. Biomed. Anal. 26:225-239, 2001.