Assessing the Chemical and Biological Diversity of an Ion Channels Knowledge Database

Ismail Ijjaali1,∗
Elodie Dubus1
Emmanuel Bourinet2–5
François Petitet1

1Aureus Pharma; Paris, France
2Département de physiologie, Institut de Génomique Fonctionnelle; 3CINS UMR5203; 4INSERM U661; 5IFR3 Universités Montpellier I & II; Montpellier, France
∗Correspondence to: Ismaïl Ijjaali; Aureus Pharma; 174 Quai de Jemmapes; Paris 75010 France; Tel.: 33.1.40.18.57.57; Fax: 33.1.40.18.57.58; Email: ismail.ijjaali@aureus-pharma.com

ABSTRACT

The aim of the present work is to assess the chemical and biological diversity of ligands reported in scientific articles or patents to be active against ion channels targets. A specific query of the AurSCOPE Ion Channel knowledge database was constructed to retrieve a set of the most active non-peptide ligands tested in binding or electrophysiology experiments against all ion channel families. A biological activity threshold cutoff expressed by $K_i$, IC$_{50}$, or EC$_{50}$ was set to 300 nM. This activity cutoff was selected such that we would retrieve a set of compounds, which contain the most active ligands for all target families, but is a reasonable number to analyze. To encode the chemical space for the entire active dataset (9897 molecules), ChemAxon’s chemical fingerprints were computed and optimized and then employed to cluster the dataset at a variety of different similarity thresholds. Concurrently, the exploration of the biological space was performed by associating with each chemical cluster the corresponding target or target family. Tri-dimensional visualization of different voltage- and ligand-gated ion channel families projected into the active chemical space was obtained after a principal components analysis performed using selected molecular descriptors. The findings presented herein give a global picture of the realm of ion channels active ligands and link the knowledge on chemical structures with their respective biological activities.

INTRODUCTION

Due to their crucial physiological role and their involvement in a broad range of acquired pathologies (hypertension, diabetes, neuropathies, epilepsy, Parkinson, etc.) as well as in congenital genetic forms of these diseases such as cystic fibrosis, cardiac arrhythmia, Liddle syndrome, periodic paralysis, ataxia, autism,1–3 ion channels have been a target of choice for the pharmacological industry. A number of drugs have been developed against these diseases. Local and general anesthetics, muscle relaxants, cardiac antiarrhythmics, antihypertensives and oral hypoglycemics4–7 are among the most popular ion channel based drugs currently on the market. Understanding structural, functional and pharmacological aspects of ion channels is of great importance for drug discovery. This improved understanding of the ion channel heterogeneity combined with increased throughput of electrophysiological screening technologies explains the high interest of industrial research scientists for ion channel targets.

Knowledge management systems that gather, organize and structure biological and chemical data are valuable tools which help analyze and better comprehend pharmacological space by visualizing and bringing into focus the main features and the complex relationships that exist between chemical structures and bio-activities. Analysis of such knowledge databases using diverse computational approaches leads to a better understanding of biologically relevant chemical space (defined by the combination of multidimensional molecular descriptors including their molecular mass, their lipophilicity, or their topological features for example) and increases the ability to identify promising therapeutic compounds. Similar systems have been applied to other therapeutic target areas, primarily in the area of G protein-coupled receptors (GPCRs).8,9 In working towards this goal, we have generated databases on various topics including GPCRs, ion channels, kinases, ADME/drug-drug interactions by using a meta-analysis of a large number of published scientific articles and patents.10 In this paper, we present the exploration of active pharmacological space of the ion channel database created.
The intention of the present work is to bring a general overview on the active chemical space of ion channels, evaluate its diversity and link it to biological targets. For this purpose, we have used two methods: chemical fingerprint-based clustering and molecular descriptor-based principal component analysis.

**MATERIALS AND METHODS**

**Molecular dataset.** The chemical structures were extracted from the Aureus Pharma AurSCOPE Ion Channel knowledge database. This knowledge database focuses on ligands described as ion channel blockers, openers, or activators and covers all ion channel superfamilies including calcium, chloride, potassium, sodium channels as well as transmitter-gated ion channels. Chemical structures of ligands (>52,000) as well as a precise description of their associated target and biological activity (>205,000) have been collected by the analysis of more than 5,700 published articles and patents. A biological activity threshold cutoff expressed by $K_i$, $IC_{50}$, and $EC_{50}$ was set to retrieving affinities or activities less than 300 nM measured in binding or electrophysiological protocols on all ion channel families (254 wild target entries in the knowledge database). This resulted in 11,519 molecules. As our analysis is a 2D approach, duplicate structures with different stereochemistry as well as molecules with different counter ions were eliminated (1,548 molecules). In addition, molecules with a molecular weight greater than 700 were disregarded (74 molecules) leading to a final set of 9,897 unique molecules. To define active ligands, the 300 nM threshold was chosen in order to keep the global number of molecules manageable for further analysis and retain the most active chemical ligands available for all ion channels families. Although this criteria is largely valid for ligand gated channels, in the case of voltage gated channels it may exclude some molecules described with poor apparent affinities but that become very effective. For example, chemical clustering is used to prioritize and select query dataset for virtual screening. 20

**Biological data mining.** In the ion channel knowledge database, biological activities have been classified into various biological protocols such as binding, in vivo, electrophysiology, isolated organ, and flux (uptake and release). Within each protocol, 50–60 descriptors such as binding, in vivo, electrophysiology, isolated organ, and biological activities have been classified into various biological protocols including calcium, chloride, potassium, sodium channels as well as transmitter-gated ion channels. Chemical structures of ligands (>52,000) as well as a precise description of their associated target and biological activity (>205,000) have been collected by the analysis of more than 5,700 published articles and patents. A biological activity threshold cutoff expressed by $K_i$, $IC_{50}$, and $EC_{50}$ was set to retrieving affinities or activities less than 300 nM measured in binding or electrophysiological protocols on all ion channel families (254 wild target entries in the knowledge database). This resulted in 11,519 molecules. As our analysis is a 2D approach, duplicate structures with different stereochemistry as well as molecules with different counter ions were eliminated (1,548 molecules). In addition, molecules with a molecular weight greater than 700 were disregarded (74 molecules) leading to a final set of 9,897 unique molecules. To define active ligands, the 300 nM threshold was chosen in order to keep the global number of molecules manageable for further analysis and retain the most active chemical ligands available for all ion channels families. Although this criteria is largely valid for ligand gated channels, in the case of voltage gated channels it may exclude some molecules described with poor apparent affinities but that become very effective as function of the stimulation rate (frequency dependent block). We therefore included in the molecules analyzed all the local anesthetics such as lidocain that fall into this condition.

**Chemical clustering.** Chemical clustering is a widely used technique that can be applied at many stages of the drug discovery process. For example, chemical clustering is used to prioritize and choose chemical representatives to submit to virtual or experimental screening campaigns. 17–19 We have used this method in a calcium T-type channel study to select query dataset for virtual screening. 20 For the clustering process, we employed a version of the Jarvis-Patrick algorithm based on CF fingerprints as implemented in Jarp algorithm of ChemAxon's module JKlustor. 14,21 It has been shown that this method performs well for chemical clustering and is computationally efficient for large databases. 17,22

**RESULTS AND DISCUSSION**

Global biological diversity in the ion channel “active” subset. Initially, we analyzed the entire subset in order to understand the global distribution of the active molecules dataset vs corresponding ion
channel targets. As shown in Figure 1, γ-aminobutyric acid receptors (GABA_A), nicotinic acetylcholine receptor (nAChR), N-methyl D-aspartate (NMDA), voltage-gated calcium, serotonin 5-HT_3 and voltage-gated potassium channels are the most represented, with more than 8835 active molecules, representing 89% of the whole set. Less represented target families include inositol phosphate IP3, adenine nucleotide P2X, glycine, acid-sensing ion channel, and chloride channel. The threshold used in this study (300 nM) may partially explain the discrepancies found in the size of the molecular datasets. Indeed, a large number of compounds active in the nanomolar range are known for GABA, NMDA or 5-HT_3 receptors while currently only a relatively smaller number has been discovered for targets such as P2X channels. In practice, a more precise analysis per target family should be made with an activity threshold selected in accordance with the global knowledge related to the each target under study as in some cases a compound active at 1 μM will be being considered as an active lead. This is particularly important in the case of use- and state-dependent blockers where the apparent affinity and efficacy of the compounds depends largely upon the experimental conditions.

The prototypical molecules here are local anesthetics with their effects on voltage-gated sodium channels. Another obvious reason for the unbalanced dataset obtained relies on the vast amount of historical pharmacological knowledge available for some well-known targets while for some other targets where research is more recent there is a relative paucity of ligands published or patented for these targets.

Subsequently, a principal component analysis (PCA) study was performed using 12 GCUT bidimensional molecular descriptors implemented in MOE to represent the chemical space covered by the ligand of a given target family. Among these descriptors, three were found to be correlated, and were subsequently excluded. The first three components are generally defined by partial charge descriptors, size, and polarizability. They represent the axes of the 3D plots where different datasets have been projected.

This distribution was mapped into a 3D representation of the chemical space (Fig. 2), according to previously defined PCA analysis. Using this representation, the chemical diversity of the dataset and the repartition of each target family within this space can easily be visualized. For clarity, only GABA, nAChR, and NMDA families are visualized in Figure 2. A large domain is delimited by both GABA and nAChR ligands and there is clear overlap between some families such as 5-HT_3 and nAChR (not shown) and NMDA and GABA. This overlap follows somehow the global structural homology of these ligand gated channels all made of an hetero- or homo-pentameric assembly of constitutive transmembrane subunits.

**Biological activities linked to main clusters.** At an 85% similarity threshold, clustering identified 1663 singletons and 938 clusters. The most populated cluster has 334 individuals (Fig. 3). Singletons represent 16.8% of the dataset. Under these conditions,
a visual examination showed that large clusters were reasonably homogeneous. The average cluster size includes 8.8 structures. The top 10 clusters contain about 17.5% (1732 molecules) of the dataset, and the top 100 clusters represent about 47.6% (4711 molecules).

Figure 4 shows chemical representatives of the top 15 most populated clusters. The larger cluster is primarily composed of triazolo-phtalazine derivatives. The second largest cluster contains 286 molecules which include a pyrolidin series, while cluster 3 contains pyrido-benzimidazole derivatives. Other major classes of structures among the top 15 clusters include diazabicyclic derivatives (160), 5-substituted-3-oxadia zoyl-1,6-naphthyridin-2(1H)-one derivatives (152), pyrazoloquinoliones (140), imidazo-pyrimidine derivatives (121). As we considered only active molecules, each compound exhibits at least one affinity or activity expressed with $IC_{50}$, $EC_{50}$ or $K_i$ below 300 nM as measured in binding and/or electrophysiological experiments on wild type targets. Targets associated with molecules corresponding to the top 12 most populated clusters (representing 1890 compounds and 19% of the whole “active” ion channel set) were analyzed (Table 1).

Clusters appear to be linked to a restricted list of target families with 10 of 12 being associated to only one target family. Not surprisingly, the majority of the clusters among the 12 most populated contains GABA$_A$ receptors ligands. Indeed, the GABA$_A$ receptor is the target with the largest number of molecules associated in our dataset (Fig. 1). In clusters 1 and 9 two targets are found: GABA$_A$ receptors and calcium channels. An in depth analysis of the molecules belonging to these clusters shows that, excepted for a few 5$\alpha$ reduced neuroactive steroids, all the compounds exhibiting activities on GABA$_A$ receptors were not tested on calcium channels, or inversely molecules active on calcium channels were not tested on GABA$_A$ receptors. For example, in cluster 1, 313 molecules are GABA$_A$ receptor ligands and 21 calcium channel inhibitors and in cluster 9, 93 target the GABA$_A$ receptor and 1 calcium channel. None of these molecules were tested on both targets. Bipotent ligands acting on both GABA$_A$ receptor and calcium channel targets are not present in these two clusters as well as in the full ion channel database even when the activity threshold is readjusted to 700 nM. Nevertheless, the similarity of the chemical structures active on GABA$_A$ and calcium channel indicates a certain degree of pharmacological promiscuity between these two targets. This is the case despite the fact that the sequences and/or structural features are clearly different. The bipotent concept ligand can be exploited in drug discovery in two ways. First, by considering the molecule as a starting lead for the second target. Secondly, by combining two different target activities with rationale for defined therapeutic application. An example of such bipotent ligand is exemplified by the work of Pathirathna et al. showing that 5$\alpha$ reduced neuroactive steroids can partially block T-type channels with an apparent affinity ($EC_{50}$~1 $\mu$M) just above our cut-off, and activate in the same time GABA$_A$ receptors. The conclusion of this study is that the synergy of these effects contributes to the enhanced analgesic effect of the molecule. This bipotency concept has been recently illustrated by Paolini et al in a review on pharmacological spaces and is also demonstrated in our joint paper where fluspirilene, a known ligand for D$_2$ GPCR receptors and also active on sodium channels is identified as a ligand of T-type calcium channels.
Figure 4. Chemical representatives of the top 15 populated chemical families.
Chemical spaces associated with target families. For a given target family, molecules exhibiting an affinity or an activity measured in binding or electrophysiology protocols were projected to a 3D representation of the chemical space according to the three first principal components of the PCA analysis performed with bidimensional descriptors. Each target family was associated with a specific 3D space as illustrated in Figures 5 and 6 for potassium and P2X channels, respectively.

Families such as nAChR and 5-HT₃ are limited to a restricted homogeneous space which is well defined and partly overlapping. Both target families are members of the ligand-gated ion channel superfamily of neurotransmitter receptors responsible for rapid transmission of nerve impulses at the synapse and have, therefore, been the subject of intensive research for many years. The cys-loop family, of which the 5-HT3 receptor is a member, includes the nAChR as well as the GABA₃ receptor and the glycine receptor. A diverse range of endogenous and artificial ligands activate these receptors, but the family shares many similarities of structure and function. For example, sequence homologies have been reported for 5-HT₃ and α7 subunit of nAChR. In addition, chemical homology can be observed as the overlapping space occupied by molecules having affinities or activities for nAChR or 5-HT₃ or both. Molecules exhibiting activities below 300 nM on the two target types have been described and can be exemplified by tropisetron, d-tubocurarine and some quinuclidine derivatives. The interest of such a bipotent mechanism of action is still unclear and further studies are required to better understand the role of the nicotinic impact of tropisetron. Nevertheless, tropisetron has been shown to improve cognitive impairments. Such rationale based on dual mechanism of action may be successfully used in drug discovery programs either to increase therapeutic efficiency or decrease side effects.

The GABA₃ receptor is also a ligand-gated ion channel. It is the most populated target in our dataset with homogenous chemical space even if some sub-ensembles of this channel can be identified. As shown above, some overlap has been described with voltage-gated calcium channels.

Chemical spaces covered by calcium, potassium or sodium channel ligands are relatively dispersed and such diversity can be explained by the heterogeneity of each target family. These target families can be subdivided in several sub-families with each having distinct structures and functions (see the IUPHAR compendium for classification and nomenclature).
For example, calcium channels can be divided into at least three main subgroups: Ca\textsubscript{V}1 (L-type), Ca\textsubscript{V}2 (N-, P/Q-, R-type) and Ca\textsubscript{V}3 (T-type) and potassium channels subdivided into voltage-gated; calcium-activated, inwardly rectifying and background. In each case, the subgroups are well populated in the knowledge database and may cover different portions of the chemical space. In addition for each subgroup, several molecular types are now known i.e. Ca\textsubscript{V}3.1, Ca\textsubscript{V}3.2 and Ca\textsubscript{V}3.3 for T-type calcium channels. As our knowledge database is comprised of published papers or patents in the field beginning with publications from the 1950’s, the target nomenclature has evolved with the understanding of molecular diversity of channel families. In our companion paper, the query dataset used for virtual screening was selected on “T-type” in order to get a sufficient number of molecules and further experimentally tested on transfected Ca\textsubscript{V}3.2 channels.

Two other families show limited but very specific chemical spaces. IP3 and P2X channels are primarily targeted by heavy molecules with inositol phosphate- or nucleotide-like structures. This particular behavior may be explained by the very specific nature of the ligand which contains phosphate, sulfate groups and/or oses. Drug discovery programs on these targets are relatively new and it is particularly challenging to identify new druggable chemical scaffolds.

Chemical spaces associated with sub-families of nAChR. As a specific example, the nAChR ligands were considered more closely and the corresponding dataset (1654 molecules) was divided in ligands specific for receptor sub-families based on subunit composition of this ion channel. In this study, ligands exhibiting binding or blocking (as measured in electrophysiological assays) affinities below 300 nM and targeting receptors containing \(\alpha_1, \alpha_2\beta_4, \alpha_3\beta_4, \alpha_4\beta_2\) or \(\alpha_7\) subunits were selected. The number of \(\alpha_1, \alpha_2\beta_4\) and \(\alpha_3\beta_4\) specific active ligands was relatively small in the knowledge database extract but \(\alpha_4\beta_2\) or \(\alpha_7\) active compounds were the most represented (533 and 208, respectively). The associated chemical space is displayed in Figure 7.

The two chemical spaces delimited by \(\alpha_4\beta_2\) and \(\alpha_7\) ligands are distinct but overlapping. Consequently, some areas may be associated with selective \(\alpha_4\beta_2\) or \(\alpha_7\) compounds and others with non selective compounds able to recognize both receptors sub-types. Of particular notice, the \(\alpha_7\) chemical space seems less diverse and this may be in relation to the homo-pentameric nature of the channel. In our dataset, cluster 112 is comprised of 14 molecules active at 100% on \(\alpha_4\beta_2\), clusters 2 and 16 (286 and 63 molecules, respectively) correspond to ligands that act selectively on \(\alpha_4\beta_2\) for 20 (cluster 2) to 27% (cluster 16), others being tested on undefined nicotinic receptors. Nevertheless, most of the compounds were tested on brain tissues using \[^{3}\text{H}]-cytisine and are associated with \(\alpha_4\beta_2\) receptors as these experimental conditions are in general used to characterize \(\alpha_4\beta_2\) receptors. On the other hand, clusters 160 and 177 (10 and 9 molecules, respectively) are comprised of 100% selective \(\alpha_7\) compounds.

Cluster 4 (160 molecules) contains 28% of \(\alpha_4\beta_2\) receptor active ligands and 3% of \(\alpha_7\) ligands. Interestingly in this cluster, one molecule is a bipotent ligand for both \(\alpha_4\beta_2\) and \(\alpha_7\) receptor subunits.

CONCLUSIONS

This study attempts to achieve a global exploration of pharmacological spaces of active ligands extracted from AurSCOPE Ion Channel knowledge database. The chemical diversity of this dataset was assessed using fingerprint-based clustering in order to identify the most represented chemical families. These were subsequently associated with the corresponding ion channel family. Furthermore, molecular descriptor-based principal component analysis was performed to visualize and project different clusters into the chemical space taking into account the related biological data. Most clusters were found to be linked to only one target family with GABA\textsubscript{A} receptors being the most represented ion channel. Next, principal components were used to project all ligands associated with a given ion channel family. This clearly showed the chemical space coverage of each target. Possible overlaps within these spaces can be related to biological specificity and selectivity issues. More detailed biological knowledge (target subunit combination, pharmacological protocol used, tissues organs or cell lines, parameters…) has permitted us to precisely delineate relevant biological areas as illustrated by the nicotinic receptor subunit classification. These boundaries are obviously based on the information reported in the knowledge database and the descriptors used to encode the corresponding chemical space. Such detailed knowledge can be maximally utilized at various steps in drug discovery projects. This has been successfully applied to a ligand-based virtual screening analysis that allowed us to identify new T-type calcium channel blockers.

Future directions for the present work include the further investigation of active and inactive chemical spaces. Other studies could include the deciphering of chemical and biological complex relationships for a specific chemical family, and the analysis of multipotent ligands.
55. Yan D, White MM. Interaction of D-tubocurarine analogs with mutant 5-HT3 receptors. Neuropharmacology 2002; 43:367-73.

56. Baker SR, Boots J, Brunsvik M, Dobson D, Green R, Hayhurst L, Keenan M, Wallace L. High affinity ligands for the alpha7 nicotinic receptor that show no cross-reactivity with the 5-HT3 receptor. Bioorg Med Chem Lett 2005; 15:4727-30.

57. Hashimoto K, Fujita Y, Hagiwara H, Iyo M. Phencyclidine-induced cognitive deficits in mice are improved by subsequent subchronic administration of tropisetron: Role of alpha7 nicotinic receptors. Eur J Pharmacol 2006; 553:191-5.

58. IUPHAR compendium of voltage-gated ion channels. http://www.iuphar-db.org/iuphar-ic/index.html.

59. Carroll FL, Ware R, Brieaddy LE, Navarro HA, Damaj MI, Martin BR. Synthesis, nicotinic acetylcholine receptor binding, and antinociceptive properties of 2’-fluoro-3’-(substituted phenyl)deschloroepibatidine analogues: Novel nicotinic antagonist. J Med Chem 2004; 47:4588-94.

60. Lee J, Davis CB, Riviero RA, Reitz AB, Shank RP. Synthesis and structure-activity relationship of novel pyridyl ethers for the nicotinic acetylcholine receptor. Bioorg Med Chem Lett 2000; 10:1063-6.