CFam: a chemical families database based on iterative selection of functional seeds and seed-directed compound clustering

Cheng Zhang¹,²,³, Lin Tao¹,⁴,*, Chu Qin¹,⁴, Peng Zhang¹, Shangying Chen¹, Xian Zeng¹, Feng Xu⁵,⁶, Zhe Chen⁶, Sheng Yong Yang² and Yu Zong Chen¹,²,*

¹Bioinformatics and Drug Design Group, Department of Pharmacy, and Center for Computational Science and Engineering, National University of Singapore, Singapore 117543, ²State Key Laboratory of Biotherapy and Cancer Center, West China Hospital, West China Medical School, Sichuan University, Chengdu 610041, China, ³Computational and Systems Biology, Singapore-MIT Alliance, National University of Singapore, Singapore, ⁴NUS Graduate School for Integrative Sciences and Engineering, Singapore 117456, ⁵College of Pharmacy and Tianjin Key Laboratory of Molecular Drug Research, Nankai University, Tianjin 300071, China, ⁶State Key Laboratory of Medicinal Chemistry & Biology, Tianjin International Joint Academy of Biotechnology & Medicine, Tianjin 300457, China and ⁷Zhejiang Key Laboratory of Gastro-intestinal Pathophysiology, Zhejiang Hospital of Traditional Chinese Medicine, Zhejiang Chinese Medical University, No. 54 Youdian Road, Hangzhou 310006, China

Received August 29, 2014; Accepted November 06, 2014

ABSTRACT

Similarity-based clustering and classification of compounds enable the search of drug leads and the structural and chemogenomic studies for facilitating chemical, biomedical, agricultural, material and other industrial applications. A database that organizes compounds into similarity-based as well as scaffold-based and property-based families is useful for facilitating these tasks. CFam Chemical Family database http://bidd2.cse.nus.edu.sg/cfam was developed to hierarchically cluster drugs, bioactive molecules, human metabolites, natural products, patented agents and other molecules into functional families, superfamilies and classes of structurally similar compounds based on the literature-reported high, intermediate and remote similarity measures. The compounds were represented by molecular fingerprint and molecular similarity was measured by Tanimoto coefficient. The functional seeds of CFam families were from hierarchically clustered drugs, bioactive molecules, human metabolites, natural products, patented agents, respectively, which were used to characterize families and cluster compounds into families, superfamilies and classes. CFam currently contains 11 643 classes, 34 880 superfamilies and 87 136 families of 490 279 compounds (1691 approved drugs, 1228 clinical trial drugs, 12 386 investigative drugs, 262 881 highly active molecules, 15 055 human metabolites, 80 255 ZINC-processed natural products and 116 783 patented agents). Efforts will be made to further expand CFam database and add more functional categories and families based on other types of molecular representations.

INTRODUCTION

Similarity-based clustering and classification of compounds have been extensively used in diverse tasks ranging from the search of bioactive agents for drug discovery (1–4) to the molecular and chemogenomic studies in such applications as chemspace navigation and analysis (5,6), structure-target relationship investigation (7–12), cross-pharmacology profiling of intra-family and cross-family targets (13,14) and receptor de-orphanization (15). For facilitating these and other tasks and for the orderly management of known compounds and the study of new compounds, it would be advantageous to organize the known compounds into chemical families based on structural similarity (16,17) as well as molecular scaffold classification (5,18,19) and molecular descriptor projection (19,20). This requires a method and resource for defining, generating and maintaining a comprehensive set of chemical families. To the best of our knowledge, such a resource is not yet publically available. We therefore developed the CFam Chemical Family database (http://bidd2.cse.nus.edu.sg/cfam) both as a database of function-based chemical families and as a re-
Generating a chemical family database would rely heavily on automated algorithms for classifying large number of known compounds that exceed 30 million compounds, 1.4 million bioactive molecules and 760,000 patented agents in the Pubchem (21) and ChEMBL (22) databases, which evokes two problems. One is the difficulty to strictly use hierarchical clustering algorithm for grouping such a large number of known compounds, even though k-means hierarchical clustering algorithm is capable of clustering 800,000 compounds (2,16) and non-hierarchical ones can cluster millions of compounds (23). The second is the difficulty to systematically define chemical families and select family members relevant to both structural and chemical studies and applications in pharmaceutical, biomedical, agricultural and industrial research and development. These problems also arise in generating protein domain families, which have been resolved by selecting subsets of proteins of known functions as the seeds of protein domain families to both define each family’s functional and structural characteristics and select family members by multiple sequence alignment against the seed proteins (24). We employed a similar strategy for generating the CFam chemical families.

To make CFam chemical families more relevant to the applications in pharmaceutical, biomedical, agricultural, material and other industrial applications as well as to the research in chemistry and related scientific disciplines, the seeds of the CFam families were or are to be iteratively selected from hierarchically clustered approved drugs, clinical trial drugs, investigative drugs, bioactive molecules, human metabolites, food ingredients and additives, flavors and scents, agrochemicals, natural products, patented agents, toxic substances, purchasable compounds and other known compounds based on the literature-reported high-similarity measures (25-28). These families were further clustered into CFam superfamilies and classes by hierarchically clustering the seeds based on the literature-reported intermediate similarity (11,29,30) and remote similarity (3,13,30) measures. Although this iterative hierarchical clustering procedure seems similar to the incremental clustering algorithm used in selecting representative proteins for clustering proteins (31) and representative compounds for clustering large compound libraries (23), there are two significant differences. One is that the seed selection and clustering processes are based on hierarchical clustering algorithms. The second is the preferential selection of compounds of higher functional importance as the seeds in the order of drugs, bioactive molecules, human metabolites, etc.

Currently, CFam database includes the seeds, members and names of families, superfamilies and classes functionally characterized by the approved drugs, clinical trial drugs,
### Figure 2.
A CFam molecule page resulting from the name search by inputting ‘aspirin’ and selecting ‘molecule’.

<table>
<thead>
<tr>
<th>CFAM Mol ID</th>
<th>CPAM00072836</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family</td>
<td>CFA0534 Cycloxygenase inhibitor salicylate derivative Aspirin family</td>
</tr>
<tr>
<td>Superfamily</td>
<td>CFSAD463 Cycloxygenase inhibitor salicylate derivative Aspirin Superfamily</td>
</tr>
<tr>
<td>Class</td>
<td>CPCAD426 Class 426</td>
</tr>
<tr>
<td>External ID</td>
<td>DAP000843</td>
</tr>
<tr>
<td>External Source</td>
<td>TTD</td>
</tr>
<tr>
<td>PubChem CID</td>
<td>2244</td>
</tr>
<tr>
<td>Functional Type</td>
<td>Approved Drug</td>
</tr>
<tr>
<td>Molecule Name</td>
<td>Aspirin</td>
</tr>
<tr>
<td>IUPAC Name</td>
<td>2-acetyloxybenzoic acid</td>
</tr>
<tr>
<td>Synonyms</td>
<td>ACETYL SALICYLIC ACID; 2-Acetoxybenzoic acid; Ecotrin; Acetaminophen</td>
</tr>
<tr>
<td>InChI</td>
<td>InChI=1S/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H, 1H3, (H, 11, 12)</td>
</tr>
<tr>
<td>InChIKey</td>
<td>BSYNRMU5XBSQ-UHFFFAOYSA-N</td>
</tr>
<tr>
<td>Formula</td>
<td>C9H8O4</td>
</tr>
<tr>
<td>Molecular Weight</td>
<td>180.16</td>
</tr>
<tr>
<td>Cross Link</td>
<td>CHEMBL25 DB00945</td>
</tr>
</tbody>
</table>

### Figure 3.
The CFam approved drug families browsing page resulting from the clicking of ‘Family’ in the section header titled ‘Browse CFam Family/Superfamily/Class by Functional Category’ and ‘Approved Drug Families’ in the section.
investigative drugs, highly active molecules (IC50 or Ki < 1 μM against molecular target), human metabolites, zinc-processed natural products and patented agents. Table 1 provides the statistics of CFam seeds, compounds, families, superfamilies and classes with respect to the seven functional categories of compounds.

**DATA COLLECTION AND PROCESSING**

Because of the high computational cost of clustering large number of compounds, the first version of CFam primarily focuses on the following seven categories of compounds of functional significance: 1691 approved drugs from TTD (32) and Drugbank (33), 1228 clinical trial drugs and 12386 investigative drugs from TTD (32), 262 881 highly active molecules (IC50 or Ki < 1 μM against molecular target) from Chembl version 18 (22), 15 055 human metabolites from HMDB (34), 80 255 ZINC-processed natural products from ZINC (35) and 116 783 patented agents from PubChem (21) databases, respectively. For database entries with multiple non-linked components, only the largest component was selected. Hydrogens were added and salt ions were removed by using Open Babel (36), duplicates were identified and removed by comparative analysis of their InChIKeys, which is a hashed version of InChI (37) designed to be nearly unique for each individual compound with a collision resistance of $2.2 \times 10^{15}$ (38).

**GENERATION OF CFam FAMILIES OF HIGH SIMILARITY COMPOUNDS**

Molecular similarity and analysis may be conducted from different structural, physicochemical and functional
perspectives by using different types of molecular representations. These include molecular descriptors (19,20,39), molecular scaffolds (5,16,19), molecular fingerprints (3,16,17) and other molecular representations, such as chemical graphs, pharmacophore patterns and molecular fields (40–43). Multiple forms of chemical families can thus be generated from these molecular representations in a similar manner as the multiple forms of protein families generated from multiple-sequence alignment of protein domains (24,44), conserved signature profiling of selected sequence segments (45), structure classification (46,47) and combined analysis of these and other features (48).

Due to the high computational cost in clustering large number of compounds, in the first version of CFam, we only used one type of molecular representation, the 2D molecular fingerprints (specifically, the 881-bit PubChem substructure fingerprints computed by using PaDEL (49)), for representing molecules, which was selected because of its computational efficiency, demonstrated effectiveness in similarity searching and extensive applications in drug discovery (3,50–54). The other types of molecular representations will be used in the future version of CFam for generating other forms of chemical families.

The seeds of CFam families were assigned and used to assemble compounds into CFam families by the following iterative hierarchical clustering procedure. In the first iteration, 1691 approved drugs were clustered by hierarchical clustering algorithm with the 2D fingerprint Tanimoto coefficient (2DF-TC) as the similarity metric and the complete linkage as the linkage criterion. Tanimoto coefficient was used because it is the most popular similarity metric for measuring compound similarity (3). Complete linkage was used because of its relatively good performance in clustering bioactive compounds in a recent comparative study (55). The criterion for grouping compounds into a cluster of high-similarity compounds is 2DF-TC >0.85, which was adopted because it is a widely used criterion for avoiding structural redundancy in selecting compound libraries for screening bioactive compounds (25,26).

High-similarity compounds grouped by this criterion typically have 30–81% chance of having the same activity in the same bioassay (26–28). The drug/drugs in each cluster was/were assigned as the seed/seeds of a CFam-approved drug family with the family name systematically characterized by the target/targets, activity type (e.g. inhibitor), molecular class/classes (e.g. benzisoxazole derivative) and drug name/names of the seed/seeds.

In the second iteration, the 2DF-TCs of the 1228 clinical trial drugs against the seed/seeds of the existing CFam families were first computed. If the 2DF-TC of a drug is >0.85 with respect to all the seeds/seed of a family, the drug was assigned as a seed of that family. If the 2DF-TC of a drug is >0.85 to some but not all of the seeds of a family, the drug was assigned as a member of that family. If the 2DF-TC of a drug is >0.85 to the seeds of more than one family, the drug was tentatively assigned to the family/families with the largest 2DF-TC and the remaining/families was/were marked as a cousin family to the assigned family/families and these cousins are indicated in the CFam database (e.g. CFFAD942 Prostaglandin G/H synthase 2 inhibitor diarylsubstituted isoxazole derivative valdecoxib family is a cousin family of CFFAD3 D2 dopamine receptor ligand benzisoxazole derivative risperidone family) so that the cousin families can be subsequently evaluated for possible merger into a combined family. The remaining unassigned clinical trial drugs were subject to the same procedure as that of the first iteration to assign them as the seed/seeds of CFam clinical trial drug families for assembling compounds into the respective families.

In the subsequent iterations, each set of 12 386 investigatory drugs, 262 881 highly active molecules, 15 055 human metabolites, 80 255 ZINC-processed natural products and 116 783 patented agents were in turn subject to the same procedure as that of the second iteration to assign compounds into the existing CFam families or as the seed/seeds of the new CFam investigative drug families, bioactive molecule families, human metabolite families, natural product families and patented agent families for assembling compounds into the corresponding families, respectively. If the 2DF-TC of a compound is >0.85 to the seeds of more than one family, it was preferentially assigned in order of priority to approved drug, clinical trial drug, bioactive molecule (currently highly active molecule), human metabolite, natural product and patented agent family, respectively. Certain functional categories, such as human metabolites and natural products, are of special interests beyond one scientific discipline. Therefore, if a compound from these categories (e.g. a natural product) was preferentially assigned to a family of a different category (e.g. approved drug), that family was marked and is displayed as a
family containing compound/compounds from this special category (e.g. approved drug family with natural product).

While possible, the names of these families were systematically determined in a similar manner as those of approved drugs. Many clinical trial and investigative drugs have little molecular class information and large number of bioactive compounds and natural products are without a common name, which make it difficult to automatically search for their molecular class names. Therefore, while possible, the IUPAC systematic names were used to extract common substructure names as putative molecular class names. Efforts will be made to determine the molecular classes of these families from the structure information of their seed/seeds. For the remaining families that we were unable to obtain molecular class information, their family names were tentatively characterized by the name/names or ID/IDs of their seed/seeds.

**GENERATION OF CFam SUPERFAMILIES OF INTERMEDIATE TO HIGH SIMILARITY COMPOUNDS, AND CFam CLASSES OF REMOTE TO HIGH SIMILARITY COMPOUNDS**

The centroid seeds of the CFam families were further clustered by hierarchical clustering algorithm with the 2DF-TC as the similarity metric and the complete linkage as the linkage criterion, so that the CFam families can be assembled into CFam superfamilies and classes. The criterion for assembling CFam family/families into a superfamily of intermediate to high similarity compounds is 2DF-TC > 0.70, which was applied because compounds satisfying this criterion have been regarded as similar to one other (30,56), and those with slightly lower similarity typically have remote similarity (29). Compounds grouped by this intermediate-similarity criterion may have up to 30% chance of having the same activity in the same bioassay (11). These superfamilies were systematically named from the common target classes, chemical classes and individual family names of the constituent family names. A superfamily is typically composed of compounds of the same or highly similar molecular scaffolds targeting the same target, members of the same target subfamilies or target sites accommodating similar molecular scaffolds. For instance, the CFSAD2 cAMP-specific 3′, 5′-cyclic phosphodiesterase, TNF inhibitor xanthine derivative superfamily includes two families of xanthine derivatives against the two targets and three families of structurally similar purine derivatives, N-alkylguanine acyclonucleosides and theobromines.

The criterion for further assembling CFam superfamily/superfamilies into CFam classes of remote to intermediate similarity compounds is 2DF-TC > 0.57, which was used because it can reasonably capture similarity compounds with cross-pharmacology relationships but not necessarily have the same activity (13). A CFam class typically consists of a large number of compounds that bind to multiple members of a target family/subfamily and/or target families/subfamilies with binding-sites accommodating similar molecular scaffolds, which makes it difficult to systematically name it. Therefore, CFam classes were tentatively named by their CFam class IDs only. Efforts will be made to manually determine their names. An example of a CFam class is CFCAD3, which is composed of the binders of GPCR Class A subfamilies A1 (C-C chemokine receptors), A9 (neuropeptide Y receptors), A13 (cannabinoid receptors), A17 (dopamine receptors), A18 (muscarinic acetylcholine receptors) and A19 (3-HT receptors), cholinesterases, tryptases, dopamine transporters and sodium channel proteins, etc.

**DATABASE STRUCTURE AND ACCESS**

CFam can be searched by three different modes (Figure 1). The first mode enables the search of CFam by inputting a compound name or ID (currently support CFam, Pubchem, Chembli, Zinc and TTD compound IDs), a CFam family name or ID, a CFam superfamily name or ID and a CFam Class ID, respectively. The relevant information may be obtained by clicking the buttons of ‘Molecule’, ‘Family’, ‘Superfamily’ and ‘Class’, respectively. For instance, inputting ‘aspirin’ and then clicking ‘Molecule’ leads to the CFam molecule CFAMM00072836 page which shows that aspirin belongs to the CFam CFFAD534 cyclooxygenase inhibitor salicylate derivative aspirin family (Figure 2). The second mode enables the browsing of CFam families, superfamilies and classes of any functional category, respectively, which can be proceeded by first clicking the ‘Family’, ‘Superfamily’ or ‘Class’ word in the section header titled ‘Browse CFam Family/Superfamily/Class by Functional Category’, and then clicking a specific functional category below the header. For instance, clicking ‘Family’ and then ‘Approved Drug Families’ leads to the page of CFam approved drug families list (Figure 3). The third mode facilitates the alignment of an input compound in SMILES or molecular fingerprint format against CFam seeds to identify CFam families with high, intermediate and remote similarity to the input compound. The list of up to 30 CFam families with at least one seed having 2DF-TC > 0.85 (high similarity family), 0.85 ≥ 2DF-TC > 0.7 (intermediate similarity family) and 0.7 ≥ 2DF-TC > 0.57 (remote similarity) to the input compound is provided. Figure 4 shows the result page of the alignment of aspirin with CFam seeds. To facilitate the development of chemical family databases and the structural and functional analysis of molecules, CFam seeds can be downloaded from the CFam main page (Figure 1).

**REMARKS**

Specialized chemical information resources, such as the chemical family databases, complement the general chemical databases for facilitating focused studies on the navigation, classification and the structural and functional characterization of molecules. The chemical family databases that comprehensively cover the known chemspace and characterize molecules from different molecular representations are increasingly needed given the rapidly expanding pools of molecules from synthetic and natural sources (57–59) and the increasing need to analyze higher number and more variety of compounds for diverse applications (13–15,19). To meet such a need, CFam will be further updated to expand existing functional families and add new families of moderately active molecules (IC50 or Ki 1–10 μM against...
molecular target), food ingredients and additives, flavors and scents, agrochemicals, natural products beyond ZINC processed ones, toxic substances, purchasable compounds and other compounds. Although some of the CFam families are currently composed of seeds only, these seeds are nonetheless useful for facilitating further development of chemical families and function-based classification of compounds.

**FUNDING**

Funding for open access charge: Major State Basic Research Development Program of China [2013CB967204]. The authors would also like to thank the Singapore Academic Research Fund (R-148-000-181-112).

**Conflict of interest statement.** None declared.

**REFERENCES**


