Databases and ontologies

GLASS: a comprehensive database for experimentally validated GPCR-ligand associations

Wallace K. B. Chan[1],[†], Hongjiu Zhang[2],[†], Jianyi Yang[2], Jeffrey R. Brender[2], Junguk Hur[3], Arzucan Özgür[4] and Yang Zhang[1],[2],*

[1]Department of Biological Chemistry, [2]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA, [3]Department of Basic Sciences, University of North Dakota, School of Medicine and Health Sciences, Grand Forks, ND 58203, USA and [4]Department of Computer Engineering, Bogazici University, Istanbul, Turkey

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors

Associate Editor: Jonathan Wren

Abstract

Motivation: G protein-coupled receptors (GPCRs) are probably the most attractive drug target membrane proteins, which constitute nearly half of drug targets in the contemporary drug discovery industry. While the majority of drug discovery studies employ existing GPCR and ligand interactions to identify new compounds, there remains a shortage of specific databases with precisely annotated GPCR-ligand associations.

Results: We have developed a new database, GLASS, which aims to provide a comprehensive, manually curated resource for experimentally validated GPCR-ligand associations. A new text-mining algorithm was proposed to collect GPCR-ligand interactions from the biomedical literature, which is then crosschecked with five primary pharmacological datasets, to enhance the coverage and accuracy of GPCR-ligand association data identifications. A special architecture has been designed to allow users for making homologous ligand search with flexible bioactivity parameters. The current database contains ∼500 000 unique entries, of which the vast majority stems from ligand associations with rhodopsin- and secretin-like receptors. The GLASS database should find its most useful application in various *in silico* GPCR screening and functional annotation studies.

Availability and implementation: The website of GLASS database is freely available at http://zhanglab.ccmb.med.umich.edu/GLASS/.

Contact: zhng@umich.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

G protein-coupled receptors (GPCR) represent one of the largest families of transmembrane proteins that bind extracellular molecules and activate intracellular signal transduction pathways, which mediate many physiological functions through their interaction with heterotrimeric G proteins. Many human diseases, including cancer and diabetes, have been found to be associated with the malfunction of the biological roles of GPCRs (Dorsam and Gutkind, 2007). Currently, ∼30–50% of drugs on the market target GPCRs, making them one of the most attractive membrane receptors for drug development (Klabunde and Hessler, 2002; Overington *et al.*, 2006). While experiment-based assays for novel

chemical compounds remain the standard procedure for drug discovery, *in silico* screening is gaining increasing acceptance as an important complementary method to narrow down the drug searching scope and to guide experimental design. Another advantage of the computational approach is due to its high speed and low cost, which enables high-throughput and large-scale database screening (Lipinski *et al.*, 2001).

Both the experimental and computational drug discovery approaches rely on existing GPCR-ligand experimental data to provide insight for screening and selecting new drugs. A variety of GPCR-orientated databases, such as GPCRDB (Horn *et al.*, 1998), TinyGRAP (Beukers *et al.*, 1999), GPCR-OKB (Khelashvili *et al.*, 2010), GDD (Gatica and Cavasotto, 2012) and GPCR-RD (Zhang and Zhang, 2010), have been developed, which generated important impacts on various molecule-level studies on the elucidation of GPCR structure and function.

There are however very few databases that can provide comprehensive resources for GPCR-ligand interactions that are essential in assisting GPCR virtual screening studies (van Laarhoven *et al.*, 2011; Weill and Rognan, 2009; Zhou and Skolnick, 2012). One difficulty in developing such databases is that the GPCRs can be associated with a large number of ligands in various binding affinities, and the GPCR-ligand association data in many chemical libraries are often mixed with various false-positives. A collection of GPCR-ligand associations with stringent experimental validations and careful human curation is essential to ensure the quality of the datasets. Second, with the success of the sequencing and structural genomics projects, the number of available GPCR and ligand interactions increase rapidly. But most of the new studies are scattered in a wide spread of publications and archives, which makes it difficult to keep the databases up to date. For example, GLIDA (Okuno *et al.*, 2008) was a useful GPCR-ligand binding database designed for chemical genomic drug discovery; but it has ceased updates to its server since October 2010. The current GLIDA library contains around 39 000 GPCR-ligand entries, whereas the amount of unique GPCR-ligand interactions available in the literature in our estimation is above 500 000 by the combination of the pharmacological database and literature search. The missing of such a substantial amount of new data significantly degrades the usefulness of the databases to the experimental and computational drug discovery studies.

In this study, we have developed a new *GPCR-ligand association* (GLASS) database for use as a general platform in assisting GPCR-related drug screening studies. Drawing from multiple primary data sources, GLASS focuses on a comprehensive and yet precise collection of the experimentally validated GPCR-ligand interactions with strong affinities. To ensure the completeness of the database and to keep it up to date, we incorporate a newly developed text-mining pipeline to search through PubMed literature to discover new GPCR-ligand interactions, which are then crosschecked with the primary pharmacological datasets to ensure the quality of data collection. All the GPCR-ligand association data are manually curated and made freely available to the community.

2 Data and Methods

The GPCR-ligand association data in GLASS consist of two major resources. The first resource consists of five primary pharmacological datasets from ChEMBL (Gaulton *et al.*, 2012), BindingDB (Liu *et al.*, 2007), IUPHAR (Sharman *et al.*, 2011), DrugBank (Knox *et al.*, 2011) and PDSP (http://pdsp.med.unc.edu/pdsp.php), which contain various bioactive ligand and protein interaction data. The second is the GPCR-specific text mining from PubMed

abstracts. A flowchart of the GLASS construction is depicted in Figure 1.

2.1 Database recombination pipeline

A list of all reviewed UniProt IDs pertaining to GPCRs was first collected from UniProtKB (Magrane and Consortium, 2011). Data relevant to each GPCR, such as species, gene name and primary sequence, were simultaneously extracted. We used a combination of synonymous GPCR names from IUPHAR and UniProtKB.

In the second step, flat line databases were downloaded from the pharmacological databases of ChEMBL, BindingDB, IUPHAR, DrugBank and PDSP. Data entries were filtered only for GPCR-related ones using UniProt ID and compiled together. The ligands without chemical identifiers were eliminated. Meanwhile, the statistical analysis of the distributions among the Ki, Kd, IC50 and EC50 values revealed that the majority (>95%) of the experimental ligand-GPCR associations have the activity values below $10 \mu M$ (Supplementary Fig. S1). Thus, an activity filter was implemented, i.e. the entries with a Ki, Kd, IC50 and EC50 higher than $10 \mu M$ were excluded, in order to sieve out weak and suspicious GPCR-ligand associations. Once an entry passes all criteria, records on the pharmacological data (e.g. ligand activities), the references to the original literature of study, and the chemical identifiers such as SMILES or InChI, are collected from the original pharmacological databases.

2.2 Text mining pipeline

The abstracts of all literatures were collected through NCBI Entrez system (Maglott *et al.*, 2011) using BioPython (Cock *et al.*, 2009). The mining process of the GPCR-ligand associations from the retrieved texts contains two steps: identification of GPCR names, ligand names and binding triggers and recognition of the GPCR-ligand interactions (Fig. 1).

Step-1: Identification of GPCR names, ligand names and binding triggers. Two named entity recognition tools were applied to extract GPCR and ligand names from the abstracts. First, SciMiner (Hur *et al.*, 2009) was used to identify the names of genes and proteins. These names are then matched and associated with HGNC [HUGO (Human Genome Organization) Gene Nomenclature Committee] official symbols (Gray *et al.*, 2014). Because SciMiner can simultaneously recognize both GPCR proteins and small peptides that bind to the GPCRs, names returned by SciMiner are split into two categories. Entries that are present in the UniProt GPCR list were collected as receptors, and those that are not present were collected as possible ligands. Since the same protein acronym can be shared by multiple distinct receptors, a scoring scheme based on the co-occurrence of abbreviated symbols and longer descriptions in the same document was employed in SciMiner to overcome the ambiguity of the acronyms.

Second, we exploited ChemSpot 2.0 (Rocktaschel *et al.*, 2012) to extract the chemical names of small molecules and short peptides, which is often considered to be more challenging than protein name recognition due to the variation of naming systems used by different literature. To increase the sensitivity of chemical recognition, ChemSpot uses a hybrid pipeline integrating a condition random field (CRF) model trained for IUPAC entry identification and a dictionary-based approach built from ChemIDplus for extracting drugs, abbreviations, molecular formulas and trivial names. The recognized chemical names are then matched and connected to an InChI string (Rocktaschel *et al.*, 2012), where common solvents including water molecules and the named entities that were not associated with identifiers in the program output were discarded.
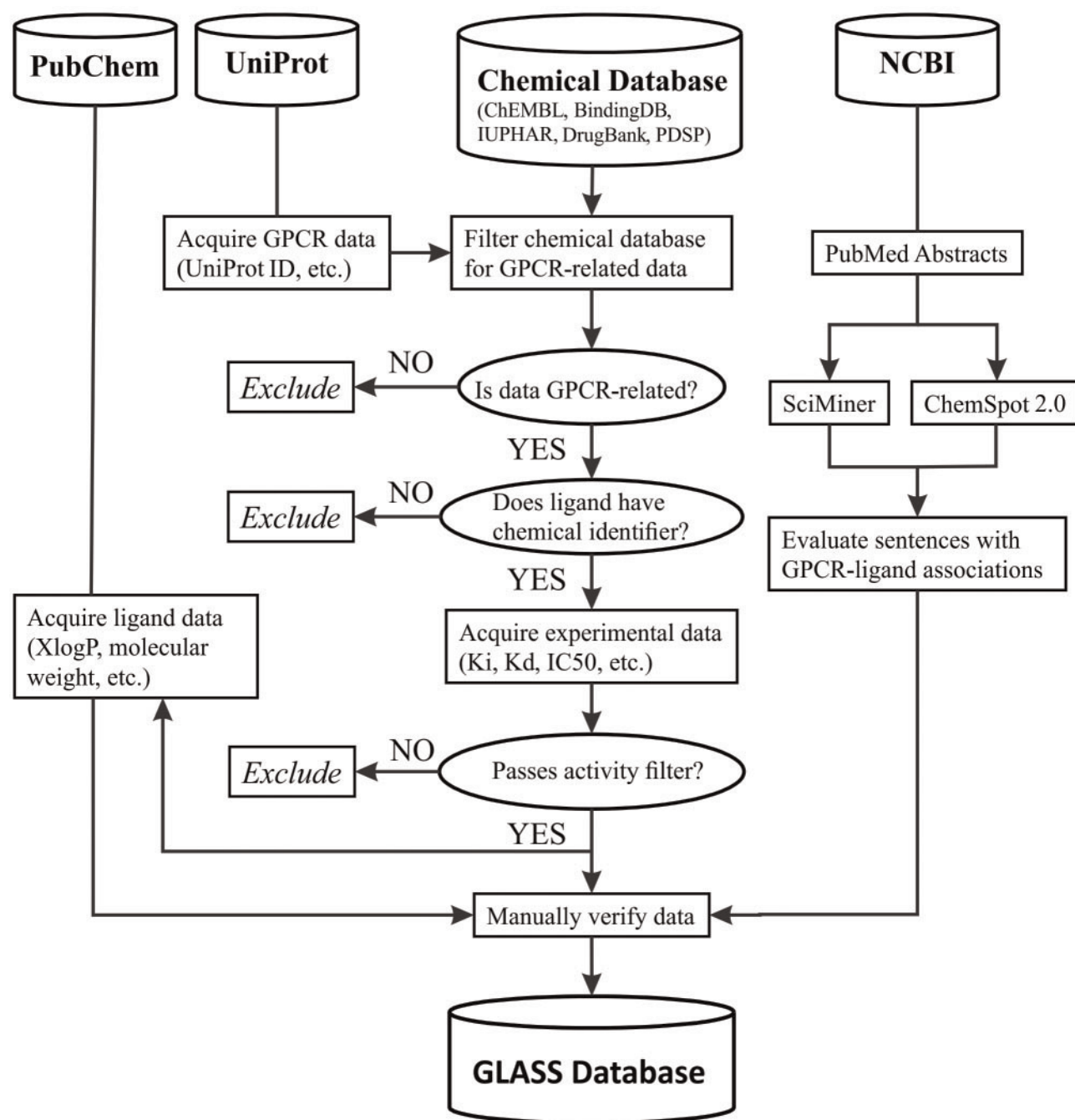
Fig. 1. Flowchart for the construction of the GLASS database

With both groups of GPCR and ligand entities, the frequency of occurrences was calculated for each abstract, which was subsequently stored in the database together with raw abstract texts and named entity positions. The name frequency data were later used to sort text mining results in the protein and ligand query pages. All ChemSpot identifications are treated as possible ligands and are tested in *Step-2*.

Binding triggers are the key words that explicitly describe biological relationship of the ligand and GPCR names, which are recognized from the target sentences using regular expression match. Three types of triggers are covered in our regular expression dictionary: (i) verbs that describe the biological phenomena, such as 'bind', 'activate', 'antagonize'; (ii) nouns and adjectives that describe

properties of ligands, such as 'agonist', 'antagonist', 'inhibitory'; (iii) nouns that describe properties of the interactions, including 'Ki', 'Kd', 'affinity', 'EC50', 'IC50'. A benchmark on 200 manually labeled abstracts showed that the regular expression can retrieve the correct ligand-GPCR binding triggers with a precision of 78.8% and a recall of 86.0%.

Step-2: Recognition of ligand-GPCR associations. The association of the GPCR and ligand names is recognized using a dependency tree based machine-learning model that is trained on a set of abstracts with known ligand-GPCR associations. The dependency tree based classification was previously proposed by Ozgur and coworkers to extract protein-protein interactions from text mining
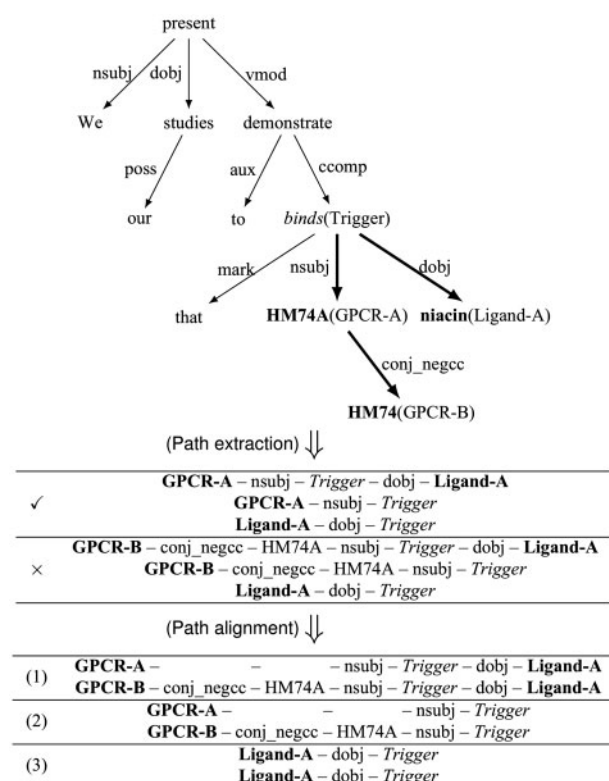
Fig. 2. An illustrative example of the dependency syntax tree parsed from the sentence 'We present our studies to demonstrate that HM74A, but not HM74, binds niacin at high affinities'. Here HM74A and niacin are associated where HM74 and niacin are not. An arrow points from an origin (or a parent node) to a word (or a child node) that is syntactically dependent on the origin. Arrows are assigned with tags, describing what attributes child nodes contribute to their parent nodes. After pruning the dependency tree, only the bold edges are left, which connect a receptor name on one side and a ligand name on the other side, and are represented as a chain of words and tags, or a path. Paths are generated for every possible pairs of GPCRs and possible ligands in a sentence

(Erkan *et al.*, 2007; Ozgur and Radev, 2009). Here we extend the idea for GPCR-ligand association recognition.

First, we collect all sentences that contain at least one GPCR name, one ligand name and one binding trigger. These sentences are then parsed into a tree of structured words with semantic predicate-argument relationships, called a typed dependency parse tree (see Fig. 2), using the Stanford CoreNLP parser (Manning *et al.*, 2014). For each sentence and every combination of a GPCR name, a ligand name and a binding trigger, the dependency parse trees returned by Stanford CoreNLP were pruned to retain the simplest tree that connects all three words.

As illustrated in Figure 2, three paths were extracted from the pruned tree: (i) connecting the GPCR name on one end and the ligand name on the other end; (ii) connecting the GPCR name and the binding trigger; (iii) connecting the ligand name and the binding trigger. These three paths were grouped together and 3310 groups of paths, extracted from 100 abstracts in the training set, were randomly selected from literature and manually labeled based on whether a true interaction is present in the paths. Following Ozgur *et al* (Erkan *et al.*, 2007; Ozgur and Radev, 2009), given every two groups of paths, each path from one group was aligned against its corresponding one from the other group using the edit distance metric. This alignment can enhance the consensus of the sentence structure and therefore improve the efficiency of classification.

Three similarity scores were calculated as the edit distances between all three pairs of aligned paths. A $N \times 3N$ matrix ($N = 3310$ being the number of paths in the training set) is constructed using these pair-wise similarities, which is passed as a $3N$-dimensional feature space to train a Gradient Boosting Decision Tree (GBDT) model. The training starts with an initial decision tree classifier trained on the initial data, branching of which stops at the third level. Based on this weak classifier, GBDT then iteratively calculates the prediction errors against the training set, constructs new decision tree classifiers trained on the prediction errors and adds them into the model (Friedman, 2001).

The GPCR-ligand associations recognized by the GBDT model in text mining are crosschecked with the data in the pharmacological datasets. Any conflicts between the two sources will be verified by manual checking before the data are integrated into GLASS. Manual reading of original articles and comparison with original databases are often needed at this step for ensuring the quality of the data. Nearly 20% of GPCR-ligand associations, which are not included in the current pharmacological datasets, have been extracted from literature after manual cross validation.

2.3 Architecture of the GLASS library

The GLASS database was built using MySQL, while the Internet webpage was augmented with a combination of Perl and Python CGI scripts to facilitate the communication of the interfaces with the MySQL database.

For each GPCR-ligand association, relevant chemical information, such as XlogP, molecular weight, hydrogen bond acceptor and donor, 2D structure image, synonyms and IUPAC name, were extracted from PubChem using the compound identifier (CID) of each ligand via their Chemical Identifier Exchange service. The 3D SDF files were generated from respective canonical SMILES strings using Open Babel (O'Boyle *et al.*, 2011).

For the GPCRs from the human genome, the associated conditions and diseases from experiments were compiled from TTD (Qin *et al.*, 2014) when available. The 3D structure information is provided for each GPCR by cross-linking to the PDB if the experimental 3D structures are available. A JSmol image is created for each GPCR to allow users to view the 3D structure of the receptor.

To facilitate comparative interaction studies, GLASS provides an interactive search engine to allow users to collect homology ligand/compounds through either substructure or chemical similarity from the experimentally validated data. Using the JSME molecular editor (Bienfait and Ertl, 2013), users are allowed to draw a chemical structure of the compounds, which is then converted into a SMILES string. Subsequently, it is transferred to Open Babel for either a substructure or similarity search against the indexed ligands. An SDF file is pre-created containing all ligand indexes in order to expedite the searching process. For the chemical similarity search, users are able to select the Tanimoto coefficient cutoffs. The resultant ligands are returned as SMILES strings. Finally, the SMILES strings are used as probes to search against the database in order to collect homologous ligands, which are returned as images of the chemical structure and their names. Tanimoto coefficients are returned, as well, if the similarity search was selected.

3 Results

3.1 GLASS in numbers

As of the time of submission of this article, GLASS contains 913 908 GPCR-ligand entries, collected from multiple sources of

Table 1. Summary of the GLASS database

Type of entry	Number of entries
All GPCRs	3048
With ligand association	681
Without ligand association	2367
Unique ligands	277 651
Drug-like ligands	190 145
All GPCR-ligand associations	913 908
Unique associations	513 246
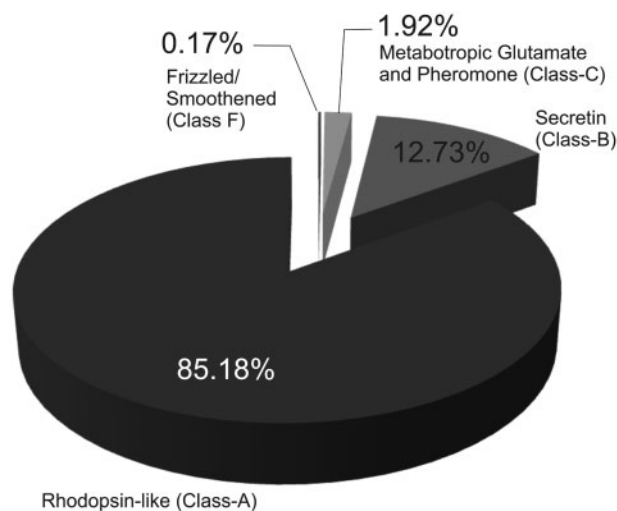
Fig. 3. Distribution of GPCR-ligand data in GLASS by family. All values presented as percentage of total. Fungal, cyclic AMP, slime mold, OA and T2R receptors, which have insufficient (<10 entries) or no data, were excluded from the plot

experiments. Some associations appear more than once in different experiments. After removing the redundant entries, there are 513 246 unique associations each containing a species-specific GPCR paired with an interacting ligand (460 439 unique associations remain if removing the redundancy across species and accounting for orthologues).

A total of 3048 GPCR entries in GLASS were extracted from UniProt (Magrane and Consortium, 2011), where 681 GPCRs have at least one ligand associations. The other 2367 GPCR entries have no ligand associate data in the experiment literature as of the present time. Among the GPCR's with ligand associations, there are ~754 different types of ligand/compound associations per receptor on average; but the median value is only 77 due to the fact that several receptor families have a dominantly high number of ligand associations (see below). The total number of unique ligands in GLASS is 277 651. A summary of the current GLASS database is presented in Table 1.

Most of the ligand associations in GLASS are skewed towards the Class-A rhodopsin-like family of GPCRs, which makes up ~85% of the association data (Fig. 3). The top five receptors in the rhodopsin-like family, all of which have more than 65 000 ligand associations, are from serotonin, adenine and adenosine nucleotide, adrenergic, opioid and dopamine receptors. These receptors also represent the set of the most popularly studied GPCRs in literature due to their importance in pharmaceutical applications and research. A histogram of the ligand associations for the entire Class-A family is shown in Supplementary Figure S2.
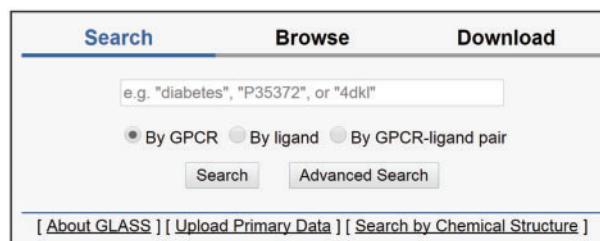
Fig. 4. A screen shot of the GLASS homepage showing options for searching, browsing and downloading of database-related data

The non-rhodopsin-like families of GPCRs constitute a far lesser proportion of ligand associations. Nevertheless, the human glucagon-like peptide 1 receptor from the Class-B secretin family contains the most abundant GPCR-ligand associations among all the human GPCRs, containing over 100 000 entries. The other non-rhodopsin-like GPCRs with more than 2 000 GPCR-ligand associations are the metabotropic glutamate and pheromone family of receptors, both from the Class-C metabotropic glutamate/pheromone family. There are only two members (UniProt ID: Q88935 and P56726) from the Class-F family that have associated experimental data, while little to no GPCR-ligand associations are found for the GPCRs from the fungal mating pheromone (Class-D), cyclic AMP (Class-E), slime mold, ocular albinism (OA) and taste receptor (T2R) families. Supplementary Figures S3–S5 list the detailed data distributions of ligand associations for Class-B, C and F families. This highly uneven ligand association distribution explains the reason that the median number of ligands per receptor is much lower than the average.

3.2 Database features
The GLASS database is updated every month, and all data are made freely available at: http://zhanglab.ccmb.med.umich.edu/GLASS/. Three features have been developed for searching, browsing, or downloading of the GPCR-ligand association data in GLASS, as shown in Figure 4, which are outlined in the following.

3.2.1 Searching GLASS
An efficient search function is essential to the development of biomedical databases. GLASS provides three options on the home page for searching the database based on three types of queries: (i) GPCR-based, (ii) ligand-based and (iii) GPCR-ligand-based. Users can choose these options by selecting the radio button of interest before or after typing the desired input (Fig. 4).

Using the GPCR-based search, users can search for a GPCR of interest using a variety of inputs, including UniProt ID, gene name, or associated medical conditions. Clicking on the 'Search' button takes the user to a page listing all GPCRs that match the query; clicking and following the link of the GPCR of interest will bring the user to a detailed page with GPCR-related information, including GPCR name, species, gene name, synonyms, associated diseases, primary sequence and its length, atomic structural model and database identifiers. All ligands that are associated with the GPCR are listed at the bottom of the page. Figure 5 presents an example of output of the GPCR-based search from the human β2 adrenergic receptor.

The ligand-based search requires knowledge of the name, chemical identifier, or PubChem ID of the ligand of interest. Clicking on the 'Search' button will bring the user to a page of results of all ligands matching the query. Clicking and following the link of the ligand of interest will bring up a detailed page with the ligand name,

Fig. 5. Illustration of the output of GPCR-based search from GLASS. This example is from the human β2 adrenergic receptor, where ligands associated with the receptor are listed at the bottom of the page. The 3D structure shown was from the PDB (ID: 2RH1) solved by Cherezov *et al.* (2007)
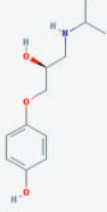

Fig 6. Illustration of the output page for the ligand-based search on GLASS. The ligand shown is prenalterol, one of the associationed ligands for the human β2 adrenergic receptor in Figure 5. GPCRs bound with prenalterol are listed at the bottom of the page

molecular formula, IUPAC name, synonyms, physico-chemical properties, chemical identifiers, database identifiers, 2D chemical structure and a list of GPCR targets with experimental data. An example output involving the ligand, prenalterol, is shown in Figure 6, where all GPCRs that bind with the ligand are listed at the bottom of the page.

Although the GPCR-ligand association information can be retrieved from the GPCR- and ligand-based searches, GLASS provides a third GPCR-ligand-based search option if the respective GLASS ID of the interaction is known. In the above example, the GLASS ID of the human β2 adrenergic receptor and prenalterol association is '8792'. By searching on '8792', the users will be brought to a page containing GPCR and ligand information, as well as experimental binding affinity data. In this example, the free energy of binding was reported to be 9.76 kcal/mol from the reference with the PubMed ID 24063433.

In addition to the ligand-, GPCR- and ligand-GPCR-based searching options, GLASS provides a target-based search for users who wish to locate a particular ligand by either chemical similarity or match of substructure (Fig. 7). Using the JSME chemical editor, the user can manually draw a ligand of interest or import a MOL or SDF file. Substructure search queries should be for the ligands of sufficient chemical complexity, as it would otherwise match too many ligands and result in an unreasonably long search. Searching by chemical similarity, there are options to select for a percentage cutoff. Results are returned with respective ligands and 2D chemical structure images; Tanimoto coefficients are also provided for similarity searches. All ligands found can be downloaded in SDF file format. An example to search homologies of morphine is illustrated in Figure 7.

3.2.2 Browsing GLASS

A comprehensive list of GPCRs and ligands from GLASS is provided on the home page to enable browsing of all entries in bulk.
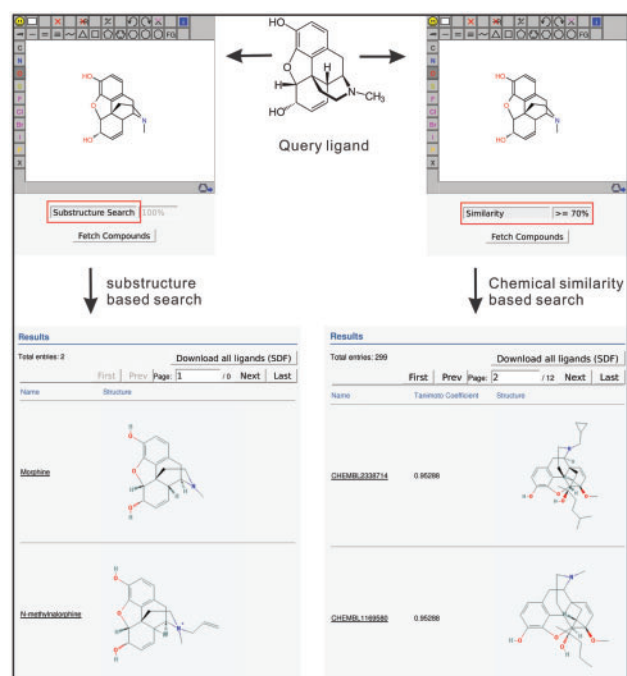
Fig. 7. Searching GLASS database for ligands using either the substructure similarity (Left Panel) or chemical similarity (Right Panel). The users first specify the ligand by importing a MOL or SDF file or draw the molecule into the JSME molecular editor. In this example, morphine is the query molecule. By clicking on the 'Fetch Compounds' button, the substructure search pipeline will look for ligands that have the molecule that the users specified as part of its chemical structure; the chemical similarity search pipeline will return all ligands that are at least 70% chemically identical to the query (the cutoff is adjustable). While the ligand name and chemical structure are provided for the users in both searches, Tanimoto coefficients are seen only with the chemical similarity search

Additionally, the user can also browse all GPCRs as sorted by their respective families as designated by UniProt (Magrane and Consortium, 2011). According to this schema, the rhodopsin-like family GPCR entries are further divided into the level of sub-families due to the high volume of entries, while the rest of the families remain in one level.

3.2.3 Downloading GLASS
Tables of GPCR, GPCR-ligand and ligand data are all made available for download in TSV file format. A zipped SDF file of all GLASS ligands in 3D format is available and ready for use in molecular docking experiments; physicochemical properties and molecular descriptors are included within the property tags for the user's convenience.

4 Summary
We have developed a new database, GLASS, which encompasses a wide breadth of GPCR-related pharmacological data, gathered from a multitude of data sources and PubMed literature mining. GLASS contains over ten times more ligand and GPCR-ligand interaction data than the leading databases, which makes GLASS the most comprehensive and up-to-date GPCR-ligand association repository in the field. It is however the novel sets of data collection and feature setting, rather than the sheer amount of data, which makes GLASS database unique.

First, the data extraction procedure was augmented with a novel text-mining pipeline, which makes it possible for automated database updating. More importantly, careful manual crosschecking of the existing datasets with the text-mining data increase both the accuracy and the coverage of the GPCR-ligand data collections. Nearly 20% of ligand association data are collected from the literature mining after manual literature validation following the automated text mining process. We have noticed that many of the bioactive ligand associations have been described in literature but were missed in most of the pharmacological databases. To address the issue, users are given the option to browse through text-mining results in order to discover missed GPCR-ligand interactions. This will prove invaluable in cases where the missing data yields bioactive ligands for a certain GPCR of interest that have chemical structures distinct from those currently in databases.

Second, the current structure of GLASS database has been made to retain the majority of GPCR-ligand pharmacological data after some definitive filters to rule out false positives; this gives users options to choose proper cutoff values for certain experimental parameters, such as binding constants. This will avoid any subjective pre-cutoffs that limit user's flexibility. Certain GPCR-ligand databases, such as GLIDA (Okuno *et al.*, 2008), only give a list of ligands with biological activities as opposed to experimental parameters. For example, a ligand could be designated as an agonist for a GPCR, but we are left unaware of how it came to be as such. The pre-cutoff setting makes it difficult to customize ligand datasets by experimental values for analysis. GLASS database was designed to ensure all of its extracted data available for user manipulation. The presence of this option means that analyses can be performed on individual GPCRs to elucidate their ligand preferences based on various cutoff values.

One of the focuses of GLASS is to provide references to various experimental and computational virtual screening studies. For instance, an important approach to GPCR virtual screening is to collect ligand profiles from homologous ligand-GPCR interactions (Zhou and Skolnick, 2012), where the completeness of the ligand-GPCR associations in GLASS will be essential to increase the sensitivity and recognition power of the ligand profiles. With its comprehensive coverage of datasets and consistent updates of data, we expect that GLASS become an important primary GPCR resource and impart its usefulness in many other biomedical studies, including *in silico* GPCR drug discovery, GPCR de-orphanization and functional annotation.

References
Beukers,M.W. *et al.* (1999) TinyGRAP database: a bioinformatics tool to mine G-protein-coupled receptor mutant data. *Trends Pharmacol. Sci.*, 20, 475–477.

Bienfait,B. and Ertl,P. (2013) JSME: a free molecule editor in JavaScript. *J. Cheminf.*, 5, 24.

Cherezov,V. *et al.* (2007) High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. *Science*, **318**, 1258–1265.

Cock,P.J. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.

Dorsam,R.T. and Gutkind,J.S. (2007) G-protein-coupled receptors and cancer. *Nat. Rev. Cancer*, **7**, 79–94.

Erkan,G. *et al.* (2007) Semi-supervised classification for extracting protein interaction sentences using dependency parsing. *EMNLP-CoNLL*, 228–237.

Friedman,J.H. (2001) Greedy function approximation: a gradient boosting machine. *Ann. Statist.*, **29**, 1189–1232.

Gatica,E.A. and Cavasotto,C.N. (2012) Ligand and decoy sets for docking to G protein-coupled receptors. *J. Chem. Inf. Model.*, **52**, 1–6.

Gaulton,A. *et al.* (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, **40**, D1100–1107.

Gray,K.A. *et al.* (2014) Genenames. org: the HGNC resources in 2015. *Nucleic Acids Res.*, gku1071.

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

Hur,J. *et al.* (2009) SciMiner: web-based literature mining tool for target identification and functional enrichment analysis. *Bioinformatics*, **25**, 838–840.

Khelashvili,G. *et al.* (2010) GPCR-OKB: the G protein coupled receptor oligomer knowledge base. *Bioinformatics*, **26**, 1804–1805.

Klabunde,T. and Hessler,G. (2002) Drug design strategies for targeting G protein coupled receptors. *Chembiochem*, **3**, 928–944.

Knox,C. *et al.* (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.*, **39**, D1035–1041.

Lipinski,C.A. *et al.* (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.*, **46**, 3–26.

Liu,T. *et al.* (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.*, **35**, D198–D201.

Maglott,D. *et al.* (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **39**, D52–D57.

Magrane,M. and Consortium,U. (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database*, **2011**, bar009.

Manning,C.D. *et al.* (2014) The Stanford CoreNLP natural language processing toolkit. In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60.

O'Boyle,N.M. *et al.* (2011) Open Babel: an open chemical toolbox. *J. Cheminf.*, **3**, 33.

Okuno,Y. *et al.* (2008) GLIDA: GPCR–ligand database for chemical genomics drug discovery—database and tools update. *Nucleic Acids Res.*, **36**, D907–D912.

Overington,J.P. *et al.* (2006) How many drug targets are there? *Nat. Rev. Drug Discov.*, **5**, 993–996.

Ozgur,A. and Radev,D.R. (2009) Supervised classification for extracting biomedical events. In: *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*. Association for Computational Linguistics, pp. 111–114.

Qin,C. *et al.* (2014) Therapeutic target database update 2014: a resource for targeted therapeutics. *Nucleic Acids Res.*, **42**, D1118–1123.

Rocktaschel,T. *et al.* (2012) ChemSpot: a hybrid system for chemical named entity recognition. *Bioinformatics*, **28**, 1633–1640.

Sharman,J.L. *et al.* (2011) IUPHAR-DB: new receptors and tools for easy searching and visualization of pharmacological data. *Nucleic Acids Res.*, **39**, D534–D538.

van Laarhoven,T. *et al.* (2011) Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics*, **27**, 3036–3043.

Weill,N. and Rognan,D. (2009) Development and validation of a novel protein-ligand fingerprint to mine chemogenomic space: application to G protein-coupled receptors and their ligands. *J. Chem. Inf. Model.*, **49**, 1049–1062.

Zhang,J. and Zhang,Y. (2010) GPCRRD: G protein-coupled receptor spatial restraint database for 3D structure modeling and function annotation. *Bioinformatics*, **26**, 3004–3005.

Zhou,H. and Skolnick,J. (2012) FINDSITE(X): a structure-based, small molecule virtual screening approach with application to all identified human GPCRs. *Mol. Pharm.*, **9**, 1775–1784.