An Ontology for Pharmaceutical Ligands and Its Application for in Silico Screening and Library Design

Ansgar Schuffenhauer,*,† Jürg Zimmermann,‡ Ruedi Stoop,§ Jan-Jan van der Vyver,§
Steffano Lecchini,§ and Edgar Jacoby*,†

Novartis Pharma AG, Drug Discovery Center, Compound Management and Computation Unit,
CH-4002 Basel, Switzerland, Novartis Pharma AG, Central Technologies, Combinatorial Chemistry Unit,
CH-4002 Basel, Switzerland, and Institute of Neuroinformatics, University/ETH Zürich,
Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

Received November 9, 2001

Annotation efforts in biosciences have focused in past years mainly on the annotation of genomic sequences. Only very limited effort has been put into annotation schemes for pharmaceutical ligands. Here we propose annotation schemes for the ligands of four major target classes, enzymes, G protein-coupled receptors (GPCRs), nuclear receptors (NRs), and ligand-gated ion channels (LGICs), and outline their usage for in silico screening and combinatorial library design. The proposed schemes cover ligand functionality and hierarchical levels of target classification. The classification schemes are based on those established by the EC, GPCRDB, NuclearDB, and LGICDB. The ligands of the MDL Drug Data Report (MDDR) database serve as a reference data set of known pharmacologically active compounds. All ligands were annotated according to the schemes when attribution was possible based on the activity classification provided by the reference database. The purpose of the ligand-target classification schemes is to allow annotation-based searching of the ligand database. In addition, the biological sequence information of the target is directly linkable to the ligand, hereby allowing sequence similarity-based identification of ligands of next homologous receptors. Ligands of specified levels can easily be retrieved to serve as comprehensive reference sets for cheminformatics-based similarity searches and for design of target class focused compound libraries. Retrospective in silico screening experiments within the MDDR01.1 database, searching for structures binding to dopamine D2, all dopamine receptors and all amine-binding class A GPCRs using known dopamine D2 binding compounds as a reference set, have shown that such reference sets are in particular useful for the identification of ligands binding to receptors closely related to the reference system. The potential for ligand identification drops with increasing phylogenetic distance. The analysis of the focus of a tertiary amine based combinatorial library compared to known amine binding class A GPCRs, peptide binding class A GPCRs, and LGIC ligands constitutes a second application scenario which illustrates how the focus of a combinatorial library can be treated quantitatively. The provided annotation schemes, which bridge chem- and bioinformatics by linking ligands to sequences, are expected to be of key utility for further systematic chemogenomics exploration of previously well explored target families.

INTRODUCTION

The immediate impact of the completion of the human genome project to the drug discovery process is its further systematization. All targets of a particular gene family are now visible, and systematic exploration of selected target families without a priori restriction to a specific therapeutic area appears to be a promising way to speed up the lead finding process. Beyond target validation, the challenge reverts to medicinal chemistry to find ligands for the sequences and to provide the molecules with which their novel biology and pharmacology can be studied. The newly identified macromolecular receptors may belong in part to established therapeutically important target classes such as enzymes, GPCRs, NRs, and LGICs, which are the most successful drug target families and which are early examples of the systematization approach. Correspondingly, every newly discovered orphan receptor of these classes can be considered as a potential drug target.[1] Because of the broad knowledge existing about the previously investigated members of these families, including the structural classes of pharmaceutically active compounds and sequence information, it is a logical expectation that the pharmacological investigation of the new targets should benefit from knowledge-based compound selection and design strategies which try to extract relevant characteristics from the established knowledge. To realize this expectation, given that the chem- and bioinformatics worlds have evolved more or less independently, it is necessary to establish necessary cross references by appropriate annotation schemes. Annotation efforts in biosciences have focused in the past years mainly

* Corresponding author phone: +41 61 32 45385; fax: +41 61 3242395; e-mail: ansgar.schuffenhauer@pharma.novartis.com (Schuffenhauer); phone: +41 61 32 46186; fax: +41 61 3242395; e-mail: edgar.jacoby@pharma.novartis.com (Jacoby).

† Novartis Pharma AG, Drug Discovery Center, Compound Management and Computation Unit.

‡ Novartis Pharma AG, Central Technologies, Combinatorial Chemistry Unit.

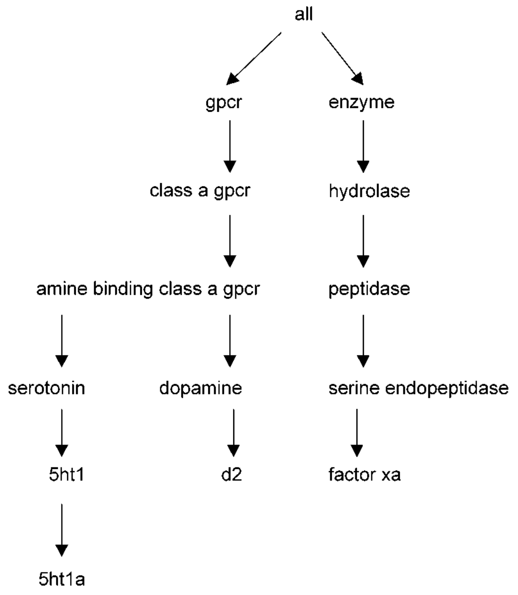
§ University/ETH Zürich.

10.1021/ci010385k CCC: $22.00
Published on Web 05/23/2002

on the annotation of genomic sequences and comprehensive gene ontologies such as GO[2]—annotating the biological process, the molecular function, and the cellular component of gene products—are the ultimate goals of this research.[3] More specifically, several nomenclature and classification committees have organized comprehensive class-specific molecular information systems for enzymes,[4] GPCRs,[5,6] NRs,[7] and LGICs.[8] Compared to this, only a very limited effort has been put into annotation schemes for pharmaceutical ligands. Ligand molecular information systems have mainly evolved from the need to track literature and patent information. Catalogues such as MDDR,[9] WDI,[10] and CMC[11] are typical databases which provide structural information about pharmaceutical ligands together with molecular target or therapeutic class information. Because the molecular target information provided within the ligand systems contains only the target name, if at all, and does not provide any further relationship among molecular targets, the potential of these systems remains limited. Ligands of close homologous receptors are for instance generally accepted as a starting point in lead finding programs for receptors for which no specific ligands are yet known. Therefore ligand classification schemes which reflect phylogenetic or other relationships of conserved molecular recognition should be expected to be useful for lead finding. Correspondingly, we will herein describe the adaptation of annotation schemes for pharmaceutical ligands of the four major target families. The MDDR01.1[9] database, which includes target information for a large number of its ligands, constitutes the underlying ligand data set. The ligand-target classification for each of the four considered target families is based on the references established by the EC,[4] GPCRDB,[5] NuclearDB,[7] and LIGCDB.[8] The resulting ligand ontologies will be demonstrated to be useful for in silico screening and library design.

METHODS

Storage of Ligand-Target Classification Schemes in a Relational Database. The classification information we used was collected from four different sources: EC-IUBMB enzyme database[4] for enzymes; GPCRDB[5] for G protein-coupled receptors; NucleaRDB[7] for nuclear receptors; and LGIC database[8] for ligand-gated ion channels. These databases use different classification criteria and allow different numbers of classification levels. Thus, a scheme to store the classification information contained in these databases in a uniform way has to restrict itself to the most general way for storing classification information. This was achieved by keeping the data in an Oracle database table (named *target_class*) in which each record represents an edge in the classification tree and contains as central information the ID of the node (fieldname: *target*) and the ID of its parent node (fieldname: *parent*). For example, the statement: "A D2 receptor is a dopamine receptor" translates into a record with the value "D2 receptor" for *target* and "dopamine receptor" for *parent* (Figure 1). Additional fields contain references to the source of this classification information. Oracle SQL provides a query syntax for hierarchical queries, which allows one to search the classification tree recursively for all nodes of the tree downward from a starting node at an arbitrary depth. For example, the command "SELECT *target* FROM *target_class* START WITH *target* = 'amine binding class a gpcr' CONNECT BY PRIOR *target* = *parent*" retrieves all nodes which are amine binding class A GPCRs. Indexes for the columns *target* and *parent* make these searches fast and efficient.



TARGET	PARENT
d2	dopamine
dopamine	amine binding class a gpcr
amine binding class a gpcr	class a gpcr
class a gpcr	gpcr
gpcr	all
5ht1a	5ht1
5ht1	serotonin
serotonin	amine binding class a gpcr
factor xa	serine endopeptidase
serine endopeptidase	peptidase
peptidase	hydrolase
hydrolase	enzyme
enzyme	all

Figure 1. Representation of the classification tree in a relational database table.

The classification data was extracted from the HTML code of the source database web pages using Perl scripts. During this process spelling was standardized. The relation between the classified targets and the structures in MDDR01.1[9] was established by a table linking the MDDR activity keys with a node in the classification tree. In addition, this table also contains for each activity the description of the intrinsic activity. Here a controlled set of terms taken from the textual description of the activity keys in MDDR was adapted; the terms include the following: activator, agonist, analogue, antagonist, inhibitor, inverse agonist, modulator, and partial inverse agonist. Although each of the MDDR activity keys can point to only one node in the target classification tree, multiple activities of structures can be stored in two ways. If the description of an activity key names a family of targets, it points to a node in the target tree having several targets as child nodes. A compound with such an activity key is thus considered to be unspecifically active on all these targets. Second, within MDDR a compound can have more than one activity key pointing to different targets on which the compound is active.

Joining the MDDR activity key table with the target classification table allows one to extract all activity keys which are associated with the target equally or at a level below the starting node of the query. The columns *target* and *act_key* were each indexed. The activity keys can then be used to search the MDDR01.1 database itself. The data in this table were compiled manually using the descriptions provided by MDDR and are contained in the Supporting Information.

In Silico Screening − Similarity Searching. The molecules in the MDDR01.1 database were split randomly into two sets. The first half served as a candidate data set for similarity searching and the second half was used to form reference sets. The similarity to the most similar molecule in the reference set (the nearest neighbor) was used as the similarity criterion. All molecules of the candidate data set were ranked by their similarity, and each molecule was examined if it was described as active on the same target as the reference set. It was also examined, if the candidate compounds belonged to the activity classes one and two steps further up in the classification tree than the reference target. The number of hits for each target class was plotted against the similarity rank. In this study the D2 dopamine receptor was chosen as target of the reference compounds. Within the compounds similar to this reference set, we studied the retrieval of ligands of the dopamine D3 and the serotonin 5HT1A receptor, which were considered in our retrospective analysis as "new" targets homologous to the reference target. In this way we can investigate the possibility to identify ligands for a new target by similarity searching, without knowing ligands for this target itself, but only for a target similar to it.

The Tanimoto coefficient based on standard Unity 2D fingerprints (Sybyl 6.7 software, Tripos Inc.) was used as the similarity measure. Unity fingerprints were stored as hexadecimal coded strings ("VARCHAR2") in Oracle tables, and the similarity searches were performed with a proprietary searching program written in C^{++} able to deal with multistructure reference sets and to access descriptor and fingerprint data stored in an Oracle database.

Library Design − Analysis of Library Focus. The focus of a tertiary amine combinatorial library obtained by reductive amination of commercially available building blocks (182 secondary amines × 170 aldehydes) was examined by computing the distributions of the maximum Tanimoto similarity based on Unity 2D fingerprints between each molecule of the library and tertiary/secondary amine reference sets obtained from the MDDR01.1 database, including 7176 amine binding class A GPCR ligands, 2166 peptide binding class A GPCR ligands, and 1165 LGIC ligands. All computations were performed with Selector Compare Databases software (Sybyl 6.7, Tripos Inc.).

RESULTS

Ligand-Target Annotation. Of the 799 activity keys used in MDDR01.1, 309 could be linked to a target in our classification scheme. This allowed us to annotate 53 211 of the total 113 821 compounds within MDDR01.1. Although there still remained activity keys with targets not covered by our annotation scheme, most of the activity keys which could not be linked to a target described only the therapeutic use of the compound and did not name a molecular target at all. Examples of such keys are "analgetic", "antibiotic", or "cognition enhancer". As can be seen from the overviews of the classification trees depicted in Figures 2 and 3, which summarize for each class the previously most intensively investigated targets, most of the classifiable compounds are active on enzymes (28 418) and GPCRs (20 961); substantially fewer LGIC (2941) and NR (1443) ligands were classifiable. Within the enzymes, hydrolases, especially the peptidases, were previously most intensively investigated, followed by the oxidoreductases and transferases, the latter class contains the kinases. However, the EC naming system, which is based on considerations of chemical catalysis, is restricted to exactly four levels of hierarchy and for instance does not discriminate between the different types of protein tyrosine kinases. In the MDDR01.1 database itself there is no distinction made between the different protein kinase inhibitors as well. In the case of cyclooxygenases (COXs), the limitation to four levels of hierarchy in the EC system again does not allow for distinguishing between the types COX-1 and COX-2, although pharmacologically important differences exist.

The second important group of targets are the GPCRs, where most structures are either active on the peptide binding or on the amine binding class A GPCRs. The GPCRDB uses an unlimited number of hierarchy levels and the scheme distinguishes between subtypes of receptors. The same is true for the scheme of NR database, which was built by the same researchers following the same guidelines and based on the results of sequence analyses.

The LGICs are classified in three different superfamilies without evolutionary relationship. Each of the LGICs consists of several subunits: Five in the case of the nicotinic (cys-loop) superfamily (nicotinic acetylcholine receptor, $GABA_A$ and $GABA_C$ receptors, glycine receptors, 5-HT_3 receptors and some glutamate activated anionic channels), four in the glutamate activated cationic channels, and three in the ATP gated channels (ATP2x and ATP2z receptors). This leads to a special problem as each subunit is a separate protein recorded with its own ID in sequence databases. Different types of subunits of the same receptor family can be aggregated combinatorially to form an ion channel with distinct functionality. Ligands can interact with binding sites on the subunits controlling the opening and closing of the channel or with the whole ion channel for example by blocking the pore. The classification starts with the superclasses and classes and continues with the subunits, shifting the meaning of a record in the classification table from "is an instance of" to "is part of".

In Silico Screening. As a first application example, we report a retrospective in silico screening experiment with the reference set of 270 dopamine D2 receptor binding compounds (Figure 4). Hereafter "binding" means agonists as well as antagonists. All compounds in the candidate set were ranked by their similarity and examined if they belonged to the following classes: dopamine D2 binding compounds (248 possible hits); dopamine receptor binding compounds (one hierarchy level up, 752 possible hits); and compounds binding to amine binding class A GPCRs (two hierarchy levels up, 4026 possible hits). The number of compounds retrieved in each of these groups versus their similarity rank is plotted in Figures 5−7. Almost all dopamine D2 ligands were in the 10% of the database most similar to the reference set (Table 1). These 10% cover 48% of compounds binding to other dopamine receptors but not to D2. Of the ligands binding to any other amine binding class A GPCR but not to dopamine receptors, 30% were found in the 10% most

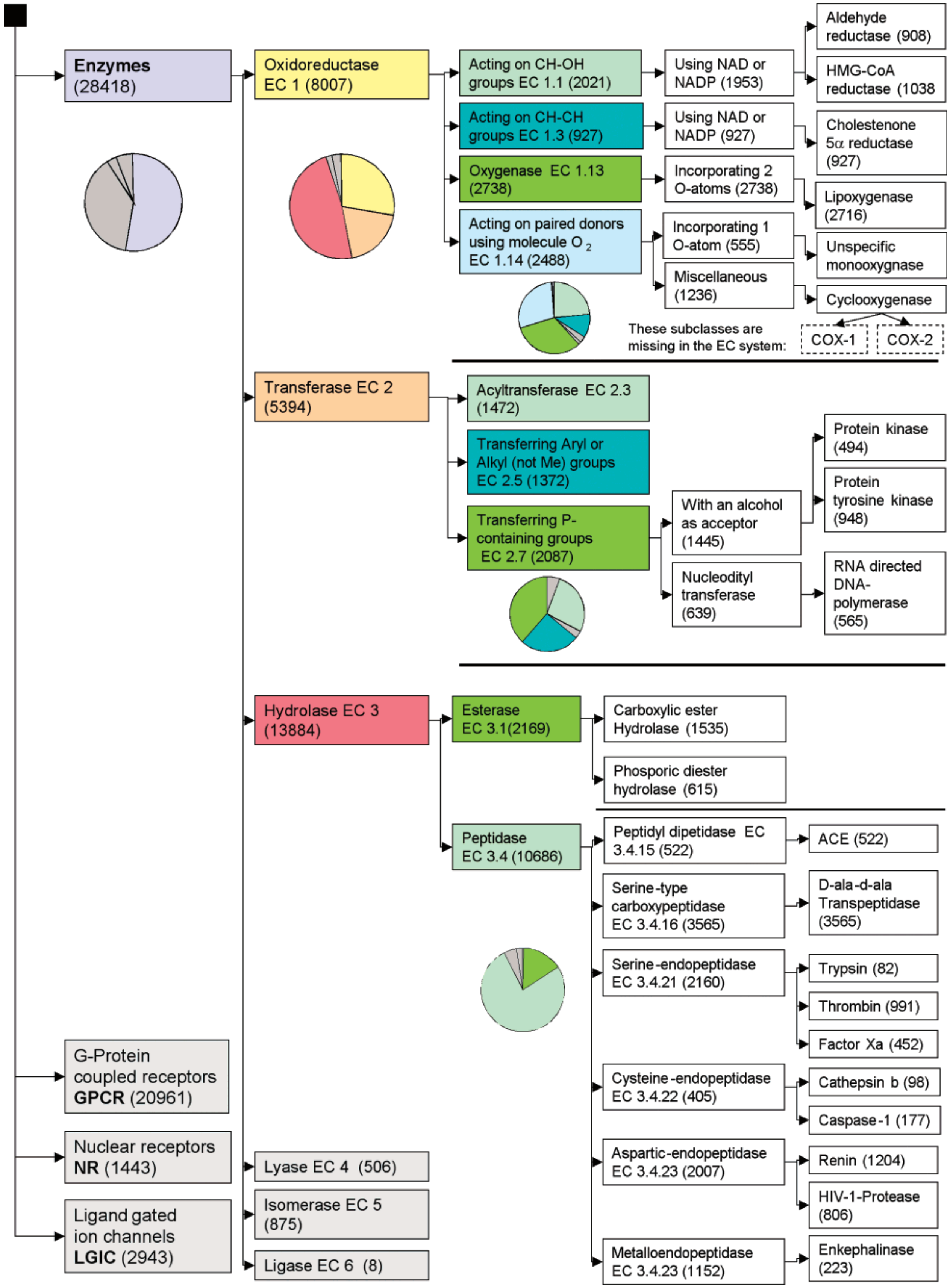

Enzymes (28418)
Oxidoreductase EC 1 (8007)
Acting on CH-OH groups EC 1.1 (2021)
Using NAD or NADP (1953)
Aldehyde reductase (908)
HMG-CoA reductase (1038
Acting on CH-CH groups EC 1.3 (927)
Using NAD or NADP (927)
Cholestenone 5α reductase (927)
Oxygenase EC 1.13 (2738)
Incorporating 2 O-atoms (2738)
Lipoxygenase (2716)
Acting on paired donors using molecule O 2 EC 1.14 (2488)
Incorporating 1 O-atom (555)
Unspecific monooxygnase
Miscellaneous (1236)
Cyclooxygenase
These subclasses are missing in the EC system:
COX-1
COX-2
Transferase EC 2 (5394)
Acyltransferase EC 2.3 (1472)
Transferring Aryl or Alkyl (not Me) groups EC 2.5 (1372)
Transferring P-containing groups EC 2.7 (2087)
With an alcohol as acceptor (1445)
Protein kinase (494)
Protein tyrosine kinase (948)
Nucleodityl transferase (639)
RNA directed DNA-polymerase (565)
Hydrolase EC 3 (13884)
Esterase EC 3.1(2169)
Carboxylic ester Hydrolase (1535)
Phosporic diester hydrolase (615)
Peptidase EC 3.4 (10686)
Peptidyl dipetidase EC 3.4.15 (522)
ACE (522)
Serine-type carboxypeptidase EC 3.4.16 (3565)
D-ala-d-ala Transpeptidase (3565)
Serine-endopeptidase EC 3.4.21 (2160)
Trypsin (82)
Thrombin (991)
Factor Xa (452)
Cysteine-endopeptidase EC 3.4.22 (405)
Cathepsin b (98)
Caspase-1 (177)
Aspartic-endopeptidase EC 3.4.23 (2007)
Renin (1204)
HIV-1-Protease (806)
Metalloendopeptidase EC 3.4.23 (1152)
Enkephalinase (223)
G-Protein coupled receptors GPCR (20961)
Nuclear receptors NR (1443)
Ligand gated ion channels LGIC (2943)
Lyase EC 4 (506)
Isomerase EC 5 (875)
Ligase EC 6 (8)

Figure 2. Overview of ligand-target classification applied to MDDR compounds. Part 1: Enzymes. As some compounds show activities on more than one target, the number of compounds in a class may be smaller than the sum over the members of the subclasses. Only the most investigated structures are shown as representatives. Numbers indicate the number of ligands annotated.

similar structures. To illustrate what kind of structures are retrieved by the reference set, we depict dopamine D3 receptor (Table 2, Figure 8) and serotonin 5HT1A receptor (Table 3, Figure 9) ligands, respectively, as examples for

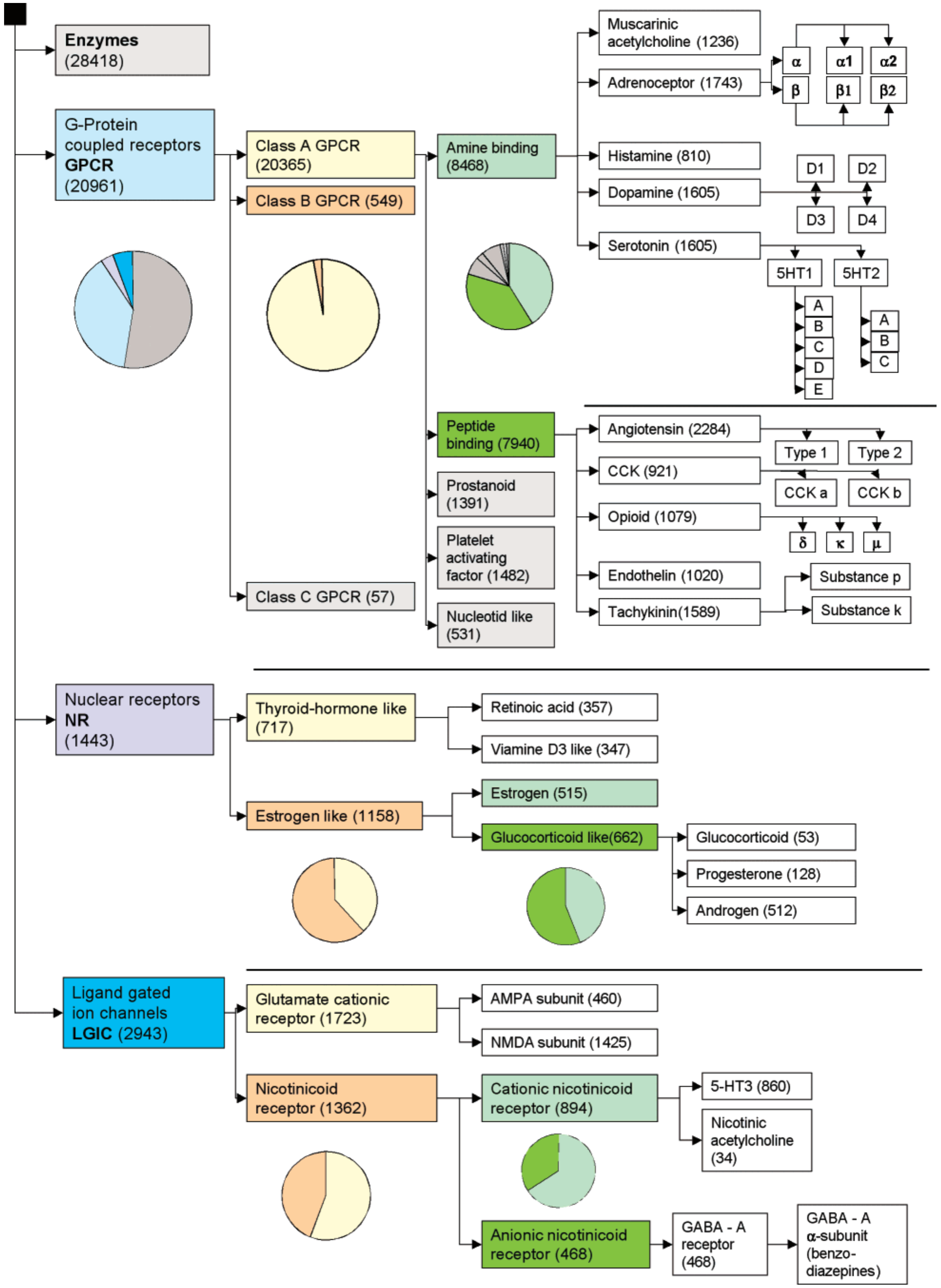


Figure 3. Overview of ligand-target classification applied to MDDR compounds. Part 2: GPCRs, NRs and LGICs. Details as in legend of Figure 2.

ligands of an homologous dopamine receptor and ligands of a more distant amine binding class A GPCR. Shown are selected compounds found at distinct Tanimoto similarity indices from 1 to 0.6 together with the reference set

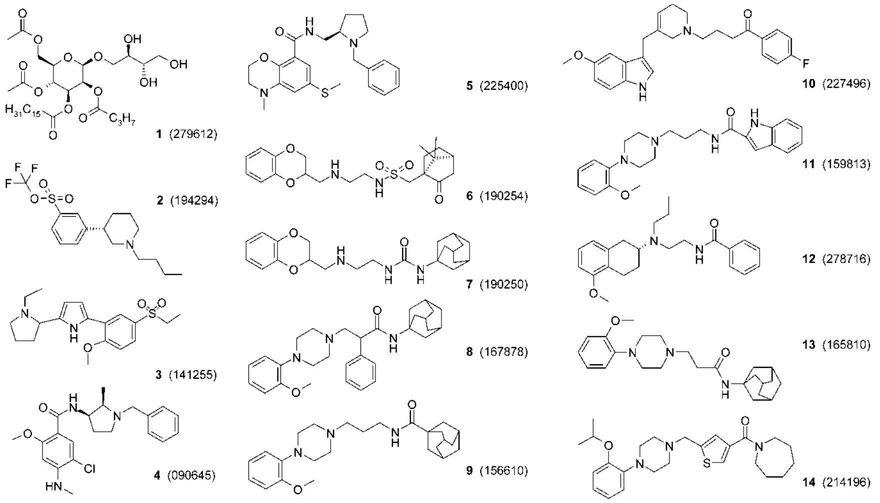


Figure 4. Examples of compounds in the dopamine D2 reference set (MDDR registry numbers in brackets).

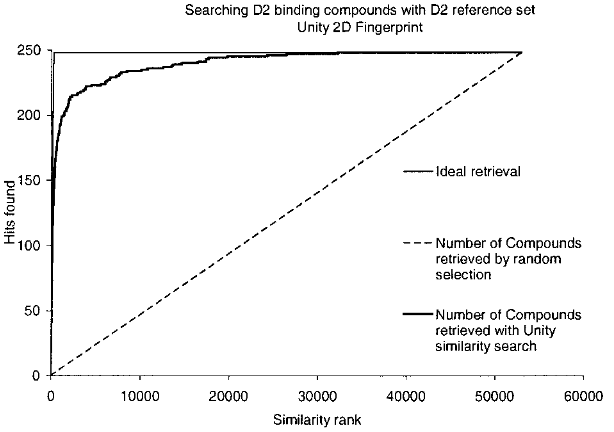


Figure 5. In silico screening – retrieval experiment – searching for dopamine D2 binding compounds, accumulated number of hits vs similarity rank.

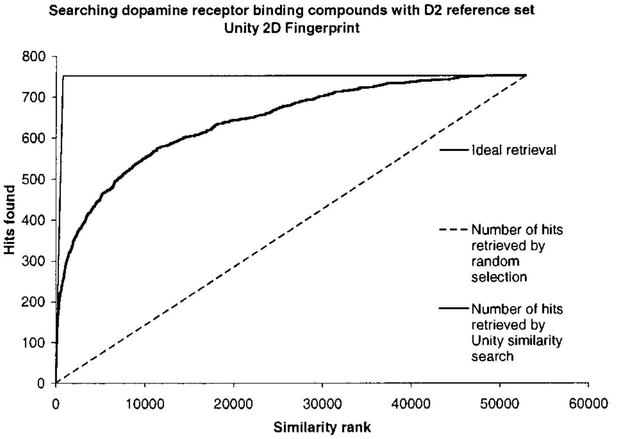


Figure 6. In silico screening – retrieval experiment – searching for combined dopamine receptor binding compounds, accumulated number of hits vs similarity rank.

compound which was the nearest neighbor responsible for their identification. At a similarity of 0.6, 10% of the total molecules screened were retrieved, and structures with a lower similarity were not considered. In the case of the D3 binding compounds, five of the shown structures were identified by one member of the reference set, whereas all shown 5HT1A actives were identified by a different member of the reference set underlining the usefulness of the multistructure reference approach.

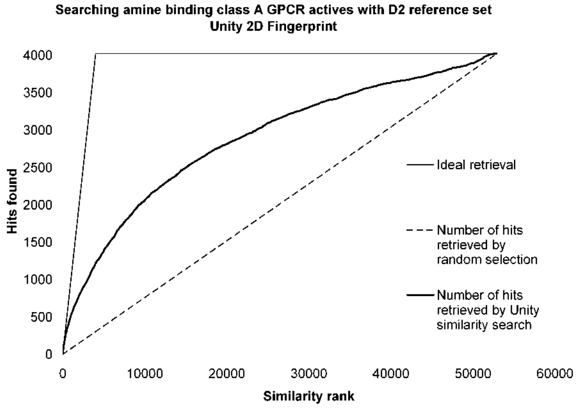


Figure 7. In silico screening – retrieval experiment – for combined amine binding class A GPCR compounds, accumulated number of hits vs similarity rank.

Table 1: In Silico Screening Results Using D2 Dopamine Receptor Ligands as Reference Set and Searching within the Candidate Set for Ligands of the Listed Classes[a]

		% of ligands retrieved			
total data set	D2	all dopamine receptor	all dopamine receptors except D2	amine binding GPCRs	amine binding GPCRs not dopamine
1	69	29	9	8	3
5	87	50	32	23	17
10	90	62	48	36	30

[a] Indicated are for each target class the retrieved fraction of actives ligands within the 1%, 5%, or 10% compounds of the total candidate data set most similar (ranked by similarity) to the D2 dopamine receptor reference set.

Library Design. Quantifying the focus of a combinatorial library constitutes a second application example, which relies on the possibility of accessing comprehensive sets of known reference compounds. Assessment of focus is in principle possible by comparing each compound of the candidate library with reference sets. Such a focus analysis is shown in Figure 10 for a tertiary amine combinatorial library compared to known amine binding class A GPCR, peptide binding class A GPCR, and LGIC ligands. As expected for tertiary amines, the library is predicted to have a particularly strong focus for monoamine binding GPCRs. Simultaneously a substantial number of compounds have similarities greater

Table 2: Selected Pairs of Retrieved D3 Binding Compounds and Their Nearest Neighbors in the D2 Reference Set at Distinct Similarity Values

similarity	D3 compd	nearest neighbor in reference set
1	**15**	**1**
0.95	**16**	**1**
0.94	**17**	**2**
0.90	**18**	**3**
0.85	**19**	**3**
0.80	**20**	**4**
0.75	**21**	**3**
0.70	**22**	**3**
0.65	**23**	**3**
0.6	**24**	**5**

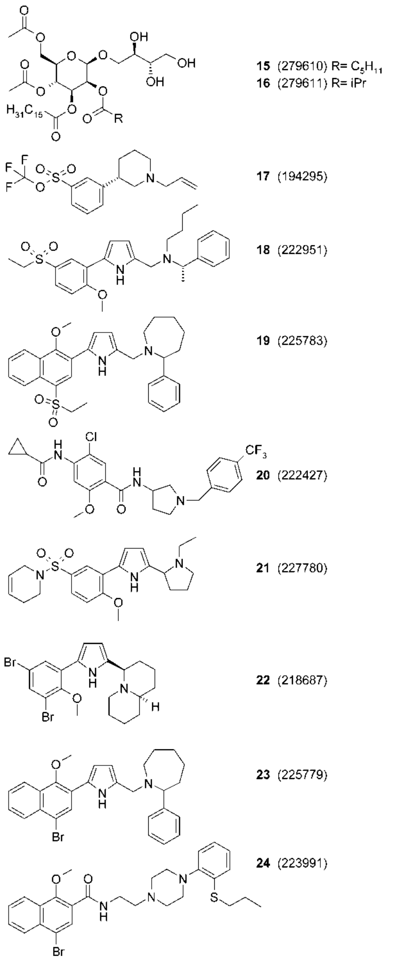


Figure 8. Examples of dopamine D3 receptor binding compounds identified with the D2 reference set (MDDR registry numbers in brackets).

than 0.6 compared to peptide binding GPCR and LGIC ligands.

DISCUSSION

Ligand Ontologies Bridging the Bio- and Cheminformatics Worlds. By linking MDDR activity keys to pharmaceutical targets within a classification scheme, we were able to group the MDDR structures by their macromolecular target classes. The four families, enzymes, GPCRs, NRs, and LGICs, cover most of the biological targets annotated in the MDDR activity keys. But there are targets still missing or incomplete in our ontology, the most important of these are the protein kinases, which need to be classified more in detail as well as some oxidoreductase families like the cyclooxygenases or the monoamine oxygenases. Missing completely in the ontology are the signal transducing membrane receptors such as the cytokine receptors. This shows how important it is to design a data structure for the classification scheme which can be expanded and amended as our knowledge about the pharmaceutical targets keeps growing.

Table 3: Selected Pairs of Retrieved 5HT1A Binding Compounds and Their Nearest Neighbors in the D2 Reference Set at Different Similarity Values

similarity	5HT1A compd	nearest neighbor in reference set
1	**25**	**6**
0.93	**26**	**7**
0.90	**27**	**8**
0.85	**28**	**9**
0.80	**29**	**10**
0.75	**30**	**11**
0.70	**31**	**12**
0.65	**32**	**13**
0.6	**33**	**14**

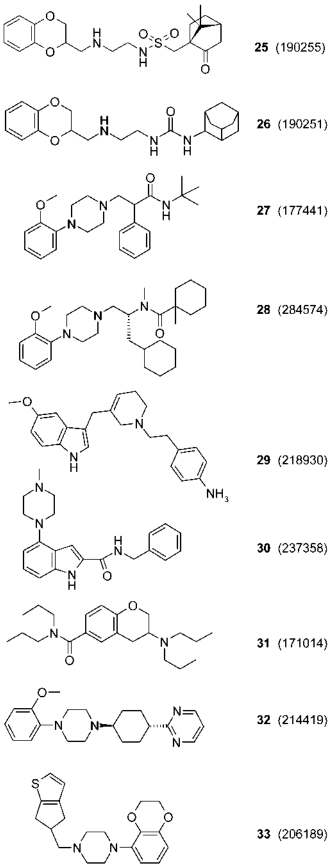


Figure 9. Examples of serotonin 5HT1A receptor binding compounds identified with the D2 reference set (MDDR registry numbers in brackets).

What also has to be discussed is how to deal with different binding sites on the same protein. It makes sense to treat each of them as a target of their own, as each of them has its own class of ligands which need not necessarily have chemical properties in common with the others. However, if one starts to screen for ligands of a new target, one often does not know how the ligands interact with the target. This will be known only later on in the drug discovery process

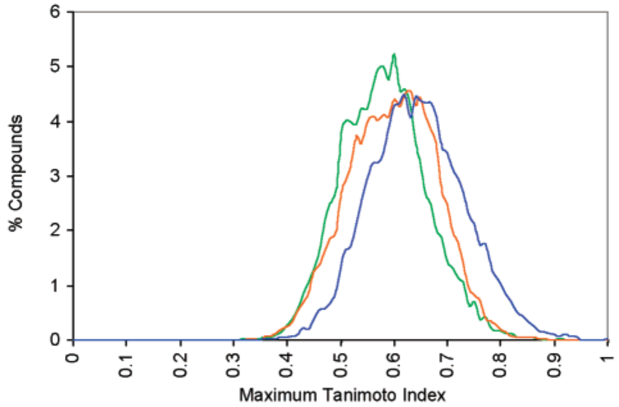


Figure 10. Focus of a tertiary amine combinatorial library. Histogram of the maximum Tanimoto similarity between each molecule of the library and reference sets obtained from the MDDR01.1 database, including 7176 amine binding GPCR ligands (blue), 2166 peptide binding class I GPCR ligands (orange), and 1165 LGIC ligands (green).

making an early classification of ligands and its use for in silico screening difficult.

As the classifications for the GPCRs and NRs are sequence analysis driven, one might consider working directly with phylogenetic trees because this provides an ordering without human decisions involved in building the classification system. In the phylogenetic tree, the nodes where the branches are separating remain anonymous. In a classification or taxonomy names are given to those groups which share commonalities. The names or identifiers for classes are very important because they can be referenced by other databases. Furthermore, not all existing drugs are yet classified to the last subtype of their receptor. For example in MDDR, there is an activity class for opioid agonists without further specifying the receptor subtypes. In our classification scheme this can still be linked to the common class of opioid receptors. A phylogenetic tree is no constant frame of knowledge as long as new proteins are still included. In contrast, the classification implemented here is more conservative. A node in the classification tree representing a target class will probably be kept, unless the whole classification scheme would have to be abandoned. If necessary, in an edge between two nodes an additional node may be inserted, but this will keep the identifier of the existing nodes intact, so all references pointing to them will still be valid. An additional table linking the leaf nodes of the ligand-target classification tree to the sequences identifiers (e.g., SWISS-PROT ID) of the precise molecular targets allows BLAST-type sequence similarity-based identification of ligands of next homologous receptors, relating in this sense ligands to the sequences.

In Silico Screening and Library Design − Knowledge-Based Approaches Founded on Ligand-target Classification. The main purpose of our ligand ontology is that ligands of specified levels can easily be collated to serve as comprehensive reference sets for cheminformatics-based similarity searches and for library design of target class focused compound collections. In our retrospective screening experiment for amine binding GPCR ligands, we have shown that such reference sets could not only be used to identify ligands binding to the same target as the reference compounds but also to find ligands for more distant targets. This illustrates the paradigm that ligands of closely homologous receptors can generally be accepted as a starting point in lead finding programs for receptors for which no specific ligands are yet known.[12] This clearly enlarges the scope of similarity searching which is classically applied for a specific target and not for target classes. In this case, the potential for ligand identification drops with increasing phylogenetic distance. The fact that different structures of the reference set were responsible for the identification of hits shows the superiority of a multistructure reference approach compared to screens with a single query compound. Even if most hits are identified by one reference compound, it is not known before the screening which compound this will be. While the nature of a simple Unity 2D type fingerprint tends to retrieve compounds which are closely related to the chemical scaffold of the query compound, a multistructure reference set, composed by the use of the ligand ontology, still ensures some diversity in the retrieved hits. This may lead to outliers in some cases when the reference set contains an outlier itself (e.g., structure **1** − Figure 4 which led to the identification of the D3 hits **15** and **16** − Figure 8 which do not have the tertiary amine group commonly found in dopamine receptor binding structures). Screens with a larger reference set as the herein used D2 set will probably be dominated less by outliers. Nevertheless a preprocessing of the reference set could be useful. One can cluster the reference set by similarity and can then identify singletons which may be outliers. These can be eliminated, and of those clusters which contain several compounds, a representative can be taken to form a preprocessed reference set leading to a reduced computational effort.

Although a multistructure reference set can ensure a certain diversity of the results, efforts to improve the quality of hits by using more sophisticated similarity measures depending less directly on the chemical scaffold is certainly worth investigating. Descriptors based on physicochemical properties,[13] graph-theoretical descriptors,[14] or molecular fields[15] may be considered. Thus, we are evaluating currently for the major target families how different similarity measures for ligands reflect the similarity of their targets.

The ligand-target classification can also be used for the analysis of corporate high throughput screening data. If one links each assay to a target node in the classification scheme, it is possible to select all assays related to a target family, and in a second step all compounds which showed activity in at least one of them can be used to collect all compounds active for a target family. These compounds can then be submitted to assays on related receptors or serve as reference structures for further in silico screening or target class focused library design. Both disciplines rely on the possibility of retrieving comprehensive sets of ligands which are likely to share a conserved molecular recognition mode.

CONCLUSION

Because biology works by applying prior knowledge ("what is known") to an unknown entity,[3] in post-genomic drug discovery research, targets can no longer be viewed as singular objects having no inter-relationship. The structure−activity relationship homology concept[16] and chemogenomics[17,18] approaches which attempt to identify all possible ligands of a given gene family are obviously well placed to further systematize the drug discovery process. To proceed within such family or system approaches we have here outlined the potential for in silico screening and library design of adapting annotation schemes which were originally only thought of for targets toward ligands. Our ligand ontology addresses the four major target classes and will have to be extended to include the novel target families of the genome. Also, for enzyme ligands, the classification system should be extended to differentiate subfamilies such as kinases. The models of classification for GPCRs and NRs, which are based on sequence analyses, are possibly best suited for the purpose of identifying ligands which share commonalities in molecular recognition. These developments will clearly benefit from GO type annotation projects but will also have to be adapted within the target view for their purpose in lead-finding. Future developments could also include search capabilities based on SCOP-3D folds or PROSITE motifs, as they have already been implemented within the LIGAND database of Japan's GenomeNet for enzyme reactions.[19]

ACKNOWLEDGMENT

Drs. P. Ertl, B. Rohde, and P. Selzer (Novartis) are acknowledged for insightful discussion. The Swiss KTI is acknowledged for the research grant "Information-based Approaches in Drug Design".

Supporting Information Available: Three tables of (1) the target classification tree with target keys; (2) the link between the recommended target names and the target keys; and (3) the link between the MDDR activity keys and the target keys. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Stadel, J. M.; Wilson, S.; Bergsma, D. J. Orphan G Protein-coupled Receptors: A Neglected Opportunity for Pioneer Drug Discovery. *Trends Pharmacol. Sci.* **1997**, *18*, 430−437.

(2) The Gene Ontology Consortium. Gene Ontology: Tool for the Unification of Biology. *Nature Genetics* **2000**, *25*, 25−29.

(3) Stevens, R.; Goble, C. A.; Bechhofer, S. Ontology-based Knowledge Representation for Bioinformatics. *Brief. Bioinform.* **2000**, *4*, 398−414.

(4) The Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB) enzyme database available on the web: http://www.chem.qmw.ac.uk/iubmb/enzyme/. Published as hardcopy: *Enzyme Nomenclature;* Academic Press: San Diego, California, 1992; ISBN 0-12-227164-5; Supplement 1: *Eur. J. Biochem.* **1994**, *223*, 1−5; Supplement 2: *Eur. J. Biochem.* **1995**, *232*, 1−6; Supplement 3: *Eur. J. Biochem.* **1996**, *237*, 1−5; Supplement 4: *Eur. J. Biochem.* **1997**, *250*, 1−6; Supplement 5: *Eur. J. Biochem.* **1999**, *264*, 610−650.

(5) Available on the web: http://www.gpcr.org/7tm. Published as hardcopy: Horn, F.; Weare, J.; Beukers, M. W.; Hörsch, S.; Bairoch, A.; Chen, W.; Edvardsen, Ø.; Campagne, F.; Vriend, G. GPCRDB: An Information System for G Protein-Coupled Receptors. *Nucleic Acids Res*. **1998**, *26*, 277−281.

(6) The IUPHAR committee on receptor nomenclature and drug classification. *The IUPHAR compendium of receptor characterization and classification*, 2nd ed.; IUPHAR Media: 2000.

(7) Available on the web: http://receptors.ucsf.edu/NR/. Published as hardcopy: Horn, F.; Vriend, G.; Cohen, F. E. Collecting and Harvesting Biological Data: The GPCRDB & NucleaRDB Databases. *Nucleic Acids Res.* **2001**, *29*, 346−349.

(8) Available on the web: http://www.pasteur.fr/recherche/banques/LGIC/LGIC.html. Published as hardcopy: Le Novère, N.; Changeux, J.-P. LGICdb: The Ligand-gated Ion Channel Database. *Nucleic Acids Res.* **2001**, *29*, 294−295.

(9) MDL Drug Data Report; MDL ISIS/HOST software, MDL Information Systems, Inc.

(10) Derwent World Drug Index; MDL ISIS/HOST software, Derwent Information Ltd.

(11) Comprehensive Medicinal Chemistry (CMC-3D) database; MDL ISIS/HOST software. CMC is an updated electronic version of *Comprehensive Medicinal Chemistry*; Pergamon Press: 1990; Vol. 6.

(12) Nishioka, T.; Sumi, K.; Oda, J. Finding Lead Structures From Amino Acid Sequence Similarities of Target Proteins. In *Probing Bioactive Mechanism*; Magee, P. S., Henry, D. R., Block, J. H., Eds.; American Chemical Society: New York, 1989; pp 105−122.

(13) Xue, L.; Bajorath, J. Molecular Descriptors for Effective Classification of Biologically Active Compounds Based on Principal Component Analysis Identified by a Genetic Algorithm. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 801−809.

(14) Gillet, V. J.; Willett, P.; Bradshaw, J. Identification of Biological Activity Profiles Using Substructural Analysis and Genetic Algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165−179.

(15) Schuffenhauer, A.; Gillet, V. J.; Willett, P. Similarity Searching in Files of Three-dimensional Chemical Structures: Analysis of the BIOSTER Database Using Two-dimensional Fingerprints and Molecular Field Descriptors. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 295−307.

(16) Frye, S. V. Structure−activity Relationship Homology (SARAH): A Conceptual Framework for Drug Discovery in the Genomic Era. *Chem. Biol.* **2001**, *6*, R3−R7.

(17) Jacoby, E. A Novel Chemogenomics Knowledge-based Ligand Design Strategy − Application to G protein-coupled Receptors. *Quant. Struct.-Act. Relat*. **2001**, *20*, 115−123.

(18) Caron, P. R.; Mullican, M. D.; Mashal, R. D.; Wilson, K. P.; Su, M. S.; Murcko, M. A. Chemogenomic Approaches to Drug Discovery. *Curr. Opin. Chem. Biol.* **2001**, *5*, 464−470.

(19) Goto, S.; Nishioka, T.; Kanehia, M. LIGAND: Chemical Database of Enzyme Reactions. *Nucleic Acids Res.* **2000**, *28*, 380−382.

CI010385K