# Substructure Mining of GPCR Ligands Reveals Activity-Class Specific Functional Groups in an Unbiased Manner

Eelke van der Horst,[†] Yasushi Okuno,[‡] Andreas Bender,[†] and Adriaan P. IJzerman*[,†]

Division of Medicinal Chemistry, Leiden/Amsterdam Center for Drug Research, Leiden University, Einsteinweg 55, 2333CC Leiden, The Netherlands, and Department of PharmacoInformatics, Center for Integrative Education of Pharmacy Frontier, Graduate School of Pharmaceutical Sciences, Kyoto University, Japan

Received October 21, 2008

In this study, we conducted frequent substructure mining to identify structural features that discriminate between ligands that do bind to G protein-coupled receptors (GPCRs) and those that do not. In most cases, particular chemical representations resulted in the most significant substructures. Substructures found to be characteristic for the background control set reflected reactions that may have been used to construct this library, e.g. for the ChemBridge DIVERSet library employed these are ester and carboxamide moieties. Alkane amine substructures were identified as most important for GPCR ligands, e.g. the butylamine substructure, often linked to an aromatic system. Hierarchical analysis of targeted GPCRs revealed well-known motives and new substructural features. One example is the imidazole-like substructure common for the histamine binding receptor ligands. Another example is the planar ring system consisting of a fused five- and six-membered ring (indole-like substucture) common for the serotonin receptor ligands.

## INTRODUCTION

Chemical structure mining has a long tradition in the prediction of molecular properties. Methods for analyzing the structural features of molecules can be broadly divided into two categories: methods that focus on predefined structural parts (fragments) and methods that consider the complete set of possible substructures of a molecule.[53] Methods of the first category apply a set of fragmentation rules to partition the molecular structure into discrete fragments, which are then analyzed. Examples of such fragments are ring systems, linkers, and side chains,[1–5] synthetic building blocks,[6,7] or algorithmically defined molecular fingerprints.[8,9] Analysis of fragment frequencies has proven useful for the description and comparison of molecular databases and for the identification of 'chemical clichés'.[10,11] For instance, unexplored parts of chemical space, with only a few fragments, become apparent as well as the preferences of chemists for certain reaction types or starting materials, yielding a more densely populated chemical space. Analyzing the co-occurrence of fragments may further yield valuable information, i.e. on pairs that seem to avoid each other or pairs that constitute a common template. Analyzing fragment occurrences may also aid the design of new ligands in (chemical) fragment-based drug discovery[12] and is a prerequisite for similarity searching,[13] an approach in which predefined structural parts are utilized to construct molecular fingerprints. A fingerprint is a reduced representation of the molecule that holds information on the presence or absence of certain features.[13] Features included in the fingerprint may be aforementioned (discrete) fragments, such as rings and functional groups (so-called structural keys, e.g. MDL keys[14]), but also algorithmically defined, such as arbitrary structural elements of fixed size, of which the circular fingerprint has gained some popularity.[15] Since predefined fragmentation rules are dependent on the choices of the chemist, analyses and predictive models are inherently biased.

Complementary to analyses using predefined structural fragments as well as fingerprints are methods that consider all possible substructures that are found in the 2D structure of a molecule. These substructure-based methods thus avoid the bias that is intrinsic to the use of predefined fragments. However, they come at the price of computational expense. In a simple structure as for the amino acid alanine without explicit hydrogens, the number of substructures amounts to 20 already. Adding two methyl groups, i.e. the amino acid valine without explicit hydrogens, will yield 39 substructures. Because of the exponential growth of substructure count with increasing molecule size, most substructure methods seek ways to limit the number of substructures to be evaluated. Batista et al. described a method that uses repeated random fragmentation of molecules to generate profiles that served as a measure of molecular similarity[16] and later applied this to database screening.[17] Although this work represents substructure analysis in an unbiased manner, the success of the method depends on the choice of parameters (iterations, number of bonds cut). Besides that, the set of evaluated substructures is not complete, i.e. the evaluation of a substructure depends on chance and not on occurrence in the molecules. Two other methods that are substructure-based are maximal common substructure analysis and frequent substructure mining. Maximal common substructure analysis finds the largest connected substructure that a certain number of molecules have in common.[18,19] It is used for similarity

* Corresponding author phone: +31 71 5274460; fax: +31 71 5274565; e-mail: ijzerman@lacdr.leidenuniv.nl.

[†] Leiden University.

[‡] Kyoto University.

10.1021/ci8003896 CCC: $40.75
Published on Web 02/03/2009

and SAR analysis, for instance as implemented in commercial tools such as Pipeline Pilot (Scitegic) and ClassPharmer (Simulations Plus).[20,21] Frequent substructure mining finds the most common substructures in one or more sets of molecules by considering all substructures that occur in the molecules. It uses a minimum-frequency constraint to control the amount of substructures that are evaluated. It is an application of frequent subgraph mining, which finds all frequently occurring connection patterns from a set of graphs. Recent advances in graph-mining algorithms, together with the steady growth of computing power, have made it possible to mine data sets as large as 200,000 molecules on a standard PC.[22] Frequent substructure 'miners' have been successfully applied for prediction of CNS activity, bioactivity, and toxicity.[23−25] The SUBSTRUCT program developed by Engkvist et al. distinguished CNS active compounds with approximately 80% accuracy.[23] However, their approach was bound to a maximum size of the generated substructures, which was between 1 and 4 atoms by default. In contrast, Borgelt et al.[24] identified substructures that model the activity classes for HIV-1 antiviral screening data. Similarly, Kazius et al. applied the frequent subgraph miner Gaston[26,27] to mutagenic compounds and extracted a decision list of six discriminative structural features associated with mutagenicity.[28] For this, the authors adopted several different types of chemical representation. So-called elaborate chemical representation adds extra information to a molecule, for instance by adding extra labels to atoms or by replacing certain atoms with wildcards (abstractions). The authors obtained the most significant results when elaborate chemical representation was used. Similarly, others also reported improved findings when using, for instance, abstractions for rings and chains (reduced graphs).[29,30]

In the context of GPCR (G Protein-Coupled Receptor) ligands, the most important source of current medicines, only methods that analyze discrete fragments have been described. For instance, a property-based scoring scheme was constructed for the classification of GPCR ligands, intended for the creation of focused libraries.[31] Similarly, Schnur et al. identified frameworks or substructure classes that are common for families of ligands. In the context of GPCR ligands, these were defined as privileged structures albeit that these frameworks were not selective for GPCRs.[32]

While previous studies were limited to analysis of predefined fragments, in this study, we will use a complete method (i.e., frequent substructure mining) to analyze the structural features of GPCR ligands. This method represents an unbiased as well as an exhaustive way to mine information contained in chemical data sets of GPCR ligands. We build upon the approach described by Kazius et al.[28] to find frequently occurring substructures that are discriminative for GPCR ligands. This is accomplished by comparing the ligands against a control group and analyzing the frequencies of all possible substructures that occur in the sets. To include additional chemical details, both normal and elaborate chemical representations (atom and bond type abstractions with atom labels) were used. However, abstractions for molecular parts, such as special types for rings or chains, were omitted since these depend on the choice of the chemist, thereby introducing a bias. In addition, with reduced graph representations, information such as bond distance or substituent positions is lost, which led us to believe that the choice of the current algorithm is appropriate for the work performed here. To derive the significant features common to specific groups of ligands, we conducted two additional experiments on subsets of the original sets. For the first experiment, subsets were based on the presence of the previously found most-significant substructure. For the second experiment, ligands were grouped into subsets according to the hierarchical classification of their target GPCRs. In addition to the work of Kazius et al.,[28] we also analyzed which substructures are rarely found in the GPCR ligands compared to the control group. This type of analysis would be less useful for prediction of mutagenicity but has added value for prediction of receptor binding. In the latter case, there will be substructures that contribute not only to binding but also to lowering this possibility (e.g., steric hindrance, unfavorable pharmacophoric features). From this analysis, we established a comprehensive list of favorable and unfavorable features of this important ligand class.

## MATERIALS AND METHODS

**Data Sets.** GPCR ligands were collected from two publicly available online resources: the GPCR-ligand database (GPCR-Ligand Database or GLIDA) from the University of Kyoto[33] and a database for human GPCRs and their ligands (hGPCR-lig).[34] The set from GLIDA consisted of 22,122 ligands for human, mouse, and rat receptors, collected from various public data sources, such as PubChem,[35] $K_i$ Database,[36] and scientific literature. The 17,908 GPCR ligands from the hGPCR-lig database are taken from the scientific literature and the MDL Drug Data Reports (MDDR). Although the sets partly overlap, the first one is more illustrative for published research from academia, while the second is more representative for patented drugs recently launched or under development, i.e. commercial drugs, preclinical candidates, etc. These two sets were compared against a control set, denoted as the background set. For this, 15,993 compounds from ChemBridge's DIVERSet screening collection were used.[37] This set was meant to provide a suitable contrast to the GPCR ligands since we were interested in how GPCR ligands differ from a diverse collection of small molecules. In GLIDA, ligands are grouped according to the classification of their targets. Targets follow the pharmacological classification of GPCRs as used by the International Union of Pharmacology (IUPHAR).[38,39] Targets are arranged into a hierarchy of subfamilies, families, and classes, as used by the GPCRDB information system.[40] In GLIDA, some receptor classes are missing due to the low number of known ligands, e.g. trace amines in amine-binding GPCRs. Each ligand-target pair is annotated with an activity type, namely full agonist, partial agonist, agonist, antagonist, or inverse agonist. Since the annotation of GLIDA is not entirely completed yet, we only used 'agonist' (for full, partial, and agonist) and 'antagonist' (for inverse agonist and antagonist). Classification of compounds as active depends on the origin of the compounds. A reported affinity in one of the source databases classified a compound as active, independent of the reported binding affinity. The sets were cleaned and standardized using Scitegic's Pipeline Pilot 6.1.5.0 Student Edition.[20] Salts, counter-ions, and other small fragments associated with the molecules were removed, and zwitterions were neutralized. Charge and stereochemical information was discarded, and bonded hydrogen atoms were omitted from

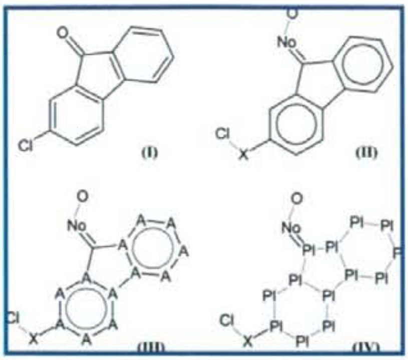


**Figure 1.** Example molecule in normal (**I**) and two elaborate chemical representations: one for aromatic bonds (**II**), one for aromatic atoms and bonds (**III**), and one for planar ring systems (**IV**). In the normal representation (**I**), aromatic bonds are represented as alternating single and double bonds, whereas in the first elaborate representation (**II**), a special type for aromatic bonds is used. In addition to this representation, the representation in (**III**) adds a special type for aromatic atoms (A). Since this example molecule has a planar ring system, all atoms that constitute the system are denoted as 'Pl' in the planar representation (**IV**). In both elaborate chemical representations, wildcards are used for heteroatoms ('No') and for halogens ('X') with a label attached specifying the actual atom-type.

the representation. After that, ChemAxon's standardizer[41] was used (for consistency with existing databases) to convert the structures into a uniform representation and to filter out duplicates. Only structures with molecular weight below 1000 Da were used. The final GPCR ligands sets consisted of 21,619 compounds for GLIDA, 16,509 compounds for hGPCR-lig, and 15,983 compounds for the background set. In some cases, analysis of large data sets using elaborate representation (see below) proved to be difficult since physical limits of system resources (maximum file size) were reached. In these cases, the experiment was continued using a sampled set of 5k ligands. These 'sampled sets' were constructed using Pipeline Pilot's Random Percent Filter.[20]

**Chemical Representation.** Molecular structures are represented as labeled graphs. Hydrogen atoms were excluded from representation. Four types of chemical representation were used: the initial chemical structure representation with the atom and bond types unchanged and three 'Elaborate' Chemical Representations (ECRs).[28] Figure 1 offers an example that accompanies the following description of the representations. Elaborate representation is a method to include extra information about the molecule by using abstractions, translations, and/or extra labels. The first elaborate representation includes a special bond type for aromatic bonds. In addition, the second one has a special type for aromatic atoms. The third representation offers a special type for planar ring systems, which has been successfully applied previously to predict the mutagenicity of compounds.[28] In elaborate chemical representation, aliphatic Nitrogen, Oxygen, and Sulfur atoms were represented as an aliphatic heteroatom by replacement with the symbol No. An extra label was attached to N and O to indicate the type and number of bound hydrogens, 'Ze' (zero) for no bonded hydrogens, 'On' for one bonded hydrogen, and 'Tw' for two bonded hydrogens. The halogen atoms, Cl, Br, I, and F, were replaced by X, and an extra label was attached to indicate their type. Note that the above-mentioned representation differs from Kazius et al.[28] in that only one heteroatom type was used and not two types for 'small heteroatom' (No) and 'large heteroatom' (Ps). In addition, fluor (F) was included as part of the halogen abstraction. Aromatic atoms and bonds were detected with basic aromaticity. Figure 1 has an example of a molecular structure in normal and chemical representation. The use of alternate representations may cause the same graph to appear multiple times. The aim of abstractions for atom and bond types is to raise the occurrence of similar substructures above the support threshold. Individually, these substructures might go undetected; however, the occurrence of their common representation sums the individual frequencies.

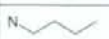
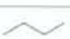
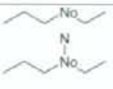
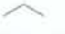
**Frequent Substructure Mining.** The frequent subgraph-miner Gaston was used to find all frequently occurring substructures in the data sets.[26,27] Frequent subgraph miners such as Gaston iterate over all molecules, extracting all possible substructures per molecule. Current subgraph miners utilize several approaches to keep the number of found substructures to a minimum. One reason is that a larger substructure can never occur more frequently than the smaller substructures it consists of. This allows numerous substructures to be pruned before being considered. Compared to other algorithms, Gaston is more efficient since computationally expensive operations take place in the last steps, when a large number of possible substructures has already been discarded. For a quantitative comparison of Gaston with other frequent subgraph miners, see ref 22. The importance of a substructure was determined by comparing its frequency against the frequency of occurrence in the control set. The most revealing substructures are those that occur frequently in one set and not in the other. As a measure of the importance of a substructure, the significance of association with one of the sets was determined by calculating the p-value of the finding. The p-value as used in this study is defined on page 3 of the Supporting Information of ref 42. It is the probability to find a statistical association with one of the two groups based on chance alone. On the assumption of a binomial distribution, it was calculated based on the number of ligands versus control group that were detected using that substructure. While this measure makes assumptions such as to the underlying distribution of features in each database, we still found it to be useful also in the ranking scheme described here. Using the p-value, the lists of frequently found substructures were ordered according to significance with the most-discriminating substructures at the top. The substructure with the lowest p-value was considered the most significant finding. The p-values of substructures are the same if they have the same absolute frequency. When two substructures had the same p-value and one substructure was a substructure of the other substructure, only the larger substructure was kept. On average, this was the case in 25% of all substructures. In case of substructures with equal p-values that were not substructures of each other, the larger substructures had preference over smaller ones in the list ordering. For example, if alanine and valine would be substructures with equal occurrences, and hence equal p-values, only the valine substructure would be kept in the list since alanine is a substructure of valine. In the case of leucine instead of alanine, both substructures would be kept since neither of the two is a substructure of the other. Another important parameter in frequent subgraph mining is the minimum support value, which is the relative number of molecules a substructure should occur in to be detected by the algorithm. Lowering the minimum support will result in finding an equal or higher number of substructures. A higher number of substructures increases the chance of finding a substructure that is more significant. However, there is a balance between minimum support and p-value of the most significant substructure. This will be illustrated with the following example of two sets of 100 compounds each. Presume that the most significant substructure found at a support threshold of 30 compounds occurred in 60 active and 20 control compounds. The p-value for this substructure is 5.32e-09, which is the chance of finding this substructure based on chance alone. Lowering the minimum support from 30 to 20 means that a new set of substructures is added to the already generated set. In theory, the most significant substructure that could be added with the new set has an occurrence of 29 active and 0 control compounds. The p-value of this theoretical substructure is 1.73e-10, which is more significant than the actual found substructure (with a p-value of 5.32e-09). Therefore, the experiment should be repeated at a lower minimum support of e.g., 20 to examine whether this theoretical substructure does actually exist. The experiment is completed if it results in the same substructure as found in the first run. This is because a more significant substructure cannot be found by lowering the support, i.e. the best theoretical substructure (19 active, 0 control) has a p-value of 7.42e-07. When another, more significant substructure is found at a lower support, the process is repeated until no theoretical substructure can be found that is more significant. Concluding, the minimum support value was chosen by iteratively lowering it per run until no better (more significant) substructures could be found, resulting in practice in support values between 10% and 30% for the data sets used here.

**Software.** For translating the molecules into elaborate representation, for partitioning the graph sets, and for the p-value calculations simple awk and bash scripts were used. These scripts and the substructure mining were run on Scientific Linux. (Sub)structures were edited and visualized using (MS Windows based applications) Pipeline Pilot 6.1.5.0 Student Edition and MDL ISIS/Draw 2.5.[20,43]

## RESULTS AND DISCUSSION

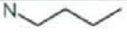
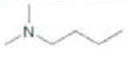
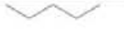
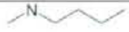
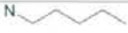
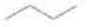
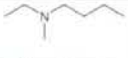
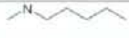
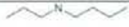
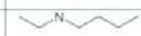
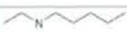
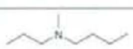
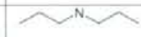
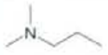
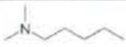
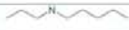
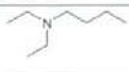
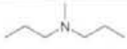
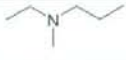
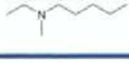
**Mining of Characteristic Substructures of GPCR Ligands.** Frequent subgraph mining was first applied to each individual set, both hGPCR-lig and GLIDA, for a broad analysis of structural features in GPCR ligands. Analysis of the substructure distributions revealed the best discriminating substructure for each of the four elaborate representations (see Materials and Methods). The substructure originating from the 'normal' representation performed best in discriminating GPCR ligands (both from hGPCR-lig and from GLIDA) from background compounds, i.e. the ChemBridge DIVERSet library. The statistics for all representations are summarized in Table 1, demonstrating that within each of the four representations highly significant substructures are occurring. The largest substructure was found in the 'aromatic atoms and bonds' representation and suggests a symmetrical organization of lipophilicity (through aliphatic carbon atoms) around a heteroatom, which was specified as nitrogen for GLIDA. This chemistry implies that at physiological pH the nitrogen heteroatom is likely to be protonated and charged, in line with the notion that many GPCR ligands interact with biogenic amine receptors. Strader et al.,[44] in one of the first mutagenesis studies on GPCRs, identified a negatively charged aspartic acid residue in transmembrane domain 3 of the $\beta$-adrenoceptor to form a salt bridge with the ligands' protonated amino group. A similar, though one atom smaller, substructure is found in the 'normal' representation. Consequently, this substructure is also at the top of the list of most significant substructures of the normal representation (hGPCR-lig: Table 2) and it is found in 74% of GPCR ligands in hGPCR-lig compared to only 26% of background compounds. Figure 2 shows some examples of GPCR ligands with this substructure overlaid, in sometimes unanticipated ways. These types of overlay also illustrate the completeness of coverage compared to the chemical fragment approach discussed in the Introduction. The other significant substructures in Table 2 are essentially variations of the first; the only differences are in number and length of carbon chains/atoms attached to the nitrogen atom. By scanning through the lists (e.g., Tables 1 and 2 and Tables 1 and 2 in the Supporting Information), a recurring theme becomes apparent. The topmost significant substructures are alkyl chains, some in combination with nitrogen, aromatic bonds, or combinations of these. Note that these lists represent substructures that occur often in GPCR ligands and not in the background molecules, thus being the distinguishing features that set these ligands apart from other organic compounds. In the 'normal' representation, a recurring theme is the alternating single/double bond feature, most likely being the substitute for aromatic bonds. Furthermore, the top significant substructures in this representation are alkylamines, chains of single-bonded carbon atoms, or combinations of both. The amine-containing substructures differ in number and length of bonded alkyl substituents; similarly, the length of the carbon tail differs as well as the position of the nitrogen within the tail. The substructure profiles for hGPCR-lig and GLIDA ligands were nearly identical, which implies that the method is stable for either data set. For instance, there was only one difference in the top 20 most significant substructures, and the ordering was virtually the same (compare Table 2 with Table 1 in the

**Table 1.** Best Discriminating Substructure for Each of the Four Elaborate Representations of the GPCR Ligands in hGPCR-lig ('Active' Column; Upper Line) and in the GLIDA Collection ('Active' Column; Lower Line)[a]

| Representation | Substructure | Active | Control | P-Value |
|---|---|---|---|---|
| Normal |  | 12,162 (74%)<br>16,434 (76%) | 4,081 (26%) | 7.35e-1707<br>3.46e-2138 |
| Ar-Bonds |  | 11,926 (72%)<br>16,142 (75%) | 4,065 (25%) | 7.25e-1611<br>3.04e-2026 |
| Ar-Atoms/Bonds |  | 8,627 (52%)<br>11,166 (52%) | 1,977 (12%)<br>1,517 (9%) | 1.06e-1354<br>1.50e-1760 |
| Planar |  | 13,050 (79%)<br>17,739 (82%) | 5,534 (35%) | 7.45e-1479<br>2.28e-1959 |

[a] For each representation, the same substructure is found in both databases, except for the 'aromatic atoms and bonds' representation. Here, the substructures differ in the presence of an extra atom label. The column labeled 'Control' refers to the background DIVERSet compound library. See Materials and Methods for an explanation of the p-value calculation.

**Table 2.** List of the 20 Most Frequent Substructures Found in GPCR Ligands (hGPCR-lig) Compared to the DIVERSet Background Compounds ('Normal' Representation)[a]

| | | Occurrence | | |
|---|---|---|---|---|
| Nr | Substructure | GPCR ligands | Background compounds | P-Value |
| 1 |  | 12,162 (74%) | 4,081 (26%) | 7.35e-1707 |
| 2 |  | 8,397 (51%) | 1,273 (8%) | 4.70e-1696 |
| 3 |  | 12,028 (73%) | 4,128 (26%) | 2.72e-1626 |
| 4 |  | 11,432 (69%) | 3,741 (23%) | 2.61e-1551 |
| 5 |  | 9,298 (56%) | 2,096 (13%) | 2.62e-1535 |
| 6 |  | 14,900 (90%) | 7,844 (49%) | 6.02e-1522 |
| 7 |  | 7,847 (48%) | 1,225 (8%) | 1.39e-1520 |
| 8 |  | 8,669 (53%) | 1,900 (12%) | 1.30e-1411 |
| 9 |  | 7,443 (45%) | 1,178 (7%) | 3.46e-1405 |
| 10 |  | 10,355 (63%) | 3,152 (20%) | 1.00e-1400 |
| 11 |  | 7,724 (47%) | 1,539 (10%) | 2.31e-1282 |
| 12 |  | 5,707 (35%) | 594 (4%) | 4.60e-1218 |
| 13 |  | 8,210 (50%) | 2,039 (13%) | 1.12e-1179 |
| 14 |  | 10,433 (63%) | 3,761 (24%) | 1.75e-1166 |
| 15 |  | 5,403 (33%) | 534 (3%) | 5.96e-1163 |
| 16 |  | 5,579 (34%) | 666 (4%) | 7.51e-1120 |
| 17 |  | 5,532 (34%) | 712 (4%) | 3.68e-1071 |
| 18 |  | 6,386 (39%) | 1,147 (7%) | 4.91e-1067 |
| 19 |  | 9,829 (60%) | 3,559 (22%) | 6.48e-1045 |
| 20 |  | 4,901 (30%) | 516 (3%) | 4.27e-1011 |

[a] Substructures were sorted according to significance, with the most significant substructure at the top. Thus substructure 1 (in bold) is found in 12,162 GPCR ligands (74%) compared to 4081 background compounds (26%). This finding is highly significant as judged from the corresponding p-value (7.35e-1707).

Supporting Information). This means that the most significant substructures cover the same area of chemical space in both the clinically promising candidates (hGPCR-lig) and compounds originating from scientific and patent literature (GLIDA). The topmost significant substructures are dominated by a few substructural themes that are common to the group as a whole. The hypothetical 'parent' fragment from which all frequent substructures derive would be an amine connected to an aromatic system through a carbon chain. This complies with the most common substructure, as defined by Sheridan et al.,[18] which was found in 21% of the ligands in the GPCR set. Again, the abundance of this substructure is probably due to the high number of aminergic receptor ligands present in the database (see also below). Even though the top significant substructures provide chemical insights found in the largest number of compounds, they also might reflect an obvious bias. A very simple one might be the commercial availability of reagents. Also, the hGPCR-lig database is largely filled with drug candidates that have reached the market or later-stage clinical trials. Due to the high attrition rates in drug discovery and development[45] these advanced compounds must have additional features for druglikeness that made them not fail beforehand, reducing the 'randomness' of substructure occurrence.

**Hierarchical Partitioning To Distinguish GPCR Ligands from Other Ligands.** For the first hierarchical analysis (the hierarchical partitioning), we used the hGPCR-lig and DIVERSet compounds. To find a new set of significant features, the experiment was repeated on subsets of the original sets: a subset in which all structures contained the substructure and another set in which they did not. The most significant substructure from the best performing representation was used to split the sets, and substructure mining was then repeated for the two subsets. The results are represented hierarchically as a tree (Figure 3). At the top of this hierarchy, we find the butyl amine substructure ('normal' representation) which occurs in 76% of GPCR ligands compared to 24% of background compounds. More than half of the molecules containing this aminergic tail also contain a substructure consisting of a heteroatom substituted with a methyl group and a propyl group. This substructure may overlay the butyl amine group or may be located elsewhere in the molecule. Molecules without the butyl amine substructure have a six-atom aromatic containing as most important structural feature. This aromatic chain is not closed to form a six-membered ring, meaning that this substructure is found not only in six-membered rings but also in aromatic systems containing fused five-membered rings etc. Since the 'normal' representation does not discriminate between aliphatic and aromatic bonds, it may lead to substructure contributions that are part of an aromatic ring system (Figure 2, e.g., compound IV). We therefore tested whether the use of a special aromatic bond type was the more suitable representation. Substructure occurrences for the aromatic bonds representation are listed in Table 3. When comparing the substructures of this representation with those of the 'normal' representation, a large overlap of the common motif was observed. For both representations, the most important moiety is the nitrogen substituted with one or more alkane chains. From these data, a second hierarchy (Figure 4) was constructed that did not consider the 'normal' representation. This approach resulted in a different substructure set than with the 'normal' representation. Compared to the amine from the 'normal' representation, the best-discriminating substructure now had a shorter tail attached to the nitrogen. This carbon tail is probably shorter due to the absence of contributing aromatic bonds that were represented as single bonds in the 'normal' representation. The first substructure with an aromatic bond was found at position 24 in the substructure list in Table 3, whereas in the 'aromatic atoms and bonds' representation the first aromatic substructure was

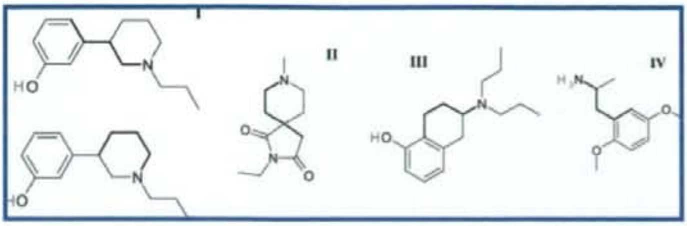


**Figure 2.** Example overlays on GPCR ligands of the most discriminative substructure, 'normal' representation. For the first ligand (I), two possible overlays are shown; multiple overlays are also possible for the ligands II and III. Note that the 'normal' representation uses Kekulé structures for aromatic systems and not separate types for delocalized bonds and aromatic atoms. This results in some interesting examples where the single bond of an aromatic ring is part of the aliphatic chain of the overlaid substructure, i.e. in structures I and IV.

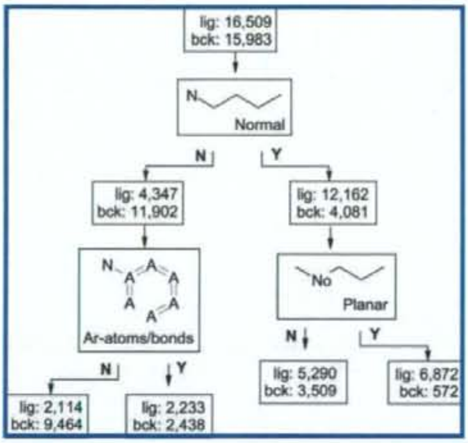
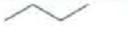
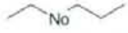
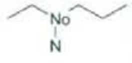
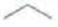
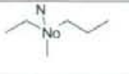
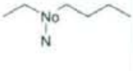
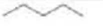
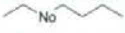
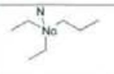
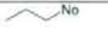
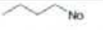
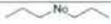
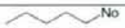
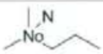
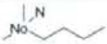
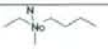
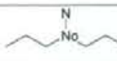
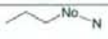
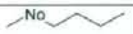
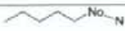
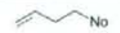


**Figure 3.** Hierarchical partitioning of GPCR ligands (hGPCR-lig-lig) versus background compounds (DIVERSet - bck). Ligands having the first substructure are taken out of the two sets (left path, N for 'No') or rather kept (right path, Y for 'Yes'), and the process of substructure mining is repeated. Numbers given are the absolute numbers of molecules used. Most discriminative substructures in any elaborate representation are shown. The name of the representation that yielded this substructure is given at the bottom of the box surrounding the substructures.

found much higher in the substructure list, i.e. at the third position. Iterating over the substructures with the 'aromatic atoms and bonds' representation leaves the impression that an aromatic ring is forming (Supporting Information: substructures 3 to 5 and 10 to 15 in Table 2). However, the fully closed ring is found much further down the list. The reason for this is that aromatic rings that differ in size contribute the same open substructures to the frequency tables. Since closed five-membered rings and closed six-membered rings cannot be overlaid, the support for these substructures is much lower. Note that even though the 'normal' representation may seem naïve, it still yielded the most significant substructure for GPCR ligands. Using this representation, a significant enrichment of the source sets can be accomplished, provided that bonds may be aliphatic or aromatic (Figure 1). Complementary to the analysis of typical substructures for the GPCR ligands was our analysis of substructures occurring more frequently in the DIVERSet background library. Thus, repeating the experiment for the background compounds yielded structural features that have low abundance in GPCR ligands. Table 4 gives an overview of the most-significant substructures for each representation. Similar as for the GPCR ligand analysis, a hierarchy was constructed for the background compounds (Figure 5). The tree is the same for analysis with and without the results of the 'normal' representation, since this representation did not produce any of the most significant substructures. The substructures that occurred often in the background set and not in the ligands are those that should be avoided when searching for GPCR ligands (since they seem to be related to inactivity). A carboxamide substructure was the most significant substructure found for the background (DIVERSet[37]) set. For this set, the aromatic bonds representation yielded the most significant finding (Table 4). Almost two-third of the compounds (depending on definition) in this set have a carboxamide or ester at the core of their scaffold, either linking two ring systems or linking a ring system with an aliphatic group. The high number of carboxamide and ester groups at the core of the molecules may reflect the simple organic reactions between alcohols and acids that have been used to construct the library. GPCR ligands differ not only in lower number of occurrence but also in position of these motifs. Where carboxamide and ester groups mainly form the linking groups between fragments in DIVERSet compounds, these motifs are also found as side-groups in GPCR ligands. This possibly reflects efforts to make drugs that are more soluble (for increased bioavailability) or to create prodrugs. The substructures in the GPCR hierarchy contained zero (Figure 4) or one heteroatom (Figure 3 and Figure 4), whereas those for the background set contained one, two, or three heteroatoms (Figure 5). This might reflect the lower number of hydrogen bond donors/acceptors in GPCR actives compared to inactives which was already noticed by Balakin et al.[31] Normally, as Feher and Schmidt pointed out,[46] the number of (small) heteroatoms is roughly equal in drugs as well as compounds in combinatorial libraries.

**GPCR Subfamilies.** We continued our analysis by focusing on individual classes of GPCR ligands. For the subfamily analysis, we continued with GLIDA and DIVERSet. For this second analysis, we derived smaller, sampled sets next to the two original full sets. The sampled sets, which were more convenient to work with, had the same substructure profiles as the full sets. For instance, the substructure lists of the full (Table 1 in the Supporting Information) and sampled set (Table 3 in the Supporting Information) of GLIDA are nearly identical. Substructures have the same order, e.g. only two shifts in positions occur in the 20 topmost structures for the 'normal' representation. Ligands were grouped hierarchically, based on the classification of the target GPCRs. The hierarchical grouping and levels where substructure analysis was performed are schematically presented in Figure 6. Similar to the analysis of GPCR ligands against

**Table 3.** Frequent Substructures for the GPCR Ligands (hGPCR-lig) Found with the 'Aromatic Bonds' Representation[a]

| Nr | Substructure | Occurrence | | P-Value |
|---|---|---|---|---|
| | | GPCR ligands | Background compounds | |
| 1 |  | 11,926 (75%) | 4,065 (25%) | 7.25e-1611 |
| 2 |  | 10,344 (65%) | 2,828 (17%) | 2.11e-1552 |
| 3 |  | 9,474 (59%) | 2,257 (14%) | 4.59e-1514 |
| 4 |  | 14,928 (93%) | 8,258 (50%) | 6.72e-1383 |
| 5 |  | 7,290 (46%) | 1,241 (8%) | 5.53e-1311 |
| 6 |  | 7,040 (44%) | 1,168 (7%) | 8.48e-1271 |
| 7 |  | 9,056 (57%) | 2,444 (15%) | 8.58e-1268 |
| 8 |  | 7,587 (47%) | 1,574 (10%) | 4.22e-1216 |
| 9 |  | 5,498 (34%) | 525 (3%) | 1.45e-1202 |
| 10 |  | 12,898 (81%) | 6,091 (37%) | 3.78e-1202 |
| 11 |  | 9,962 (62%) | 3,286 (20%) | 3.25e-1200 |
| 12 |  | 5,626 (35%) | 606 (4%) | 4.33e-1181 |
| 13 |  | 7,222 (45%) | 1,467 (9%) | 1.76e-1154 |
| 14 |  | 7,863 (49%) | 1,868 (11%) | 2.25e-1151 |
| 15 |  | 5,774 (36%) | 736 (4%) | 1.11e-1133 |
| 16 |  | 5,441 (34%) | 590 (4%) | 2.31e-1131 |
| 17 |  | 5,159 (32%) | 514 (3%) | 2.26e-1097 |
| 18 |  | 11,202 (70%) | 4,916 (30%) | 4.81e-997 |
| 19 |  | 8,939 (56%) | 2,999 (18%) | 2.10e-985 |
| 20 |  | 5,998 (38%) | 1,110 (7%) | 1.64e-966 |
| ... | ... | ... | ... | ... |
| 24 |  | 10,819 (66%) | 4,728 (30%) | 2.04e-937 |

[a] The first occurrence of an aromatic bond is found in the 24th substructure. See for further explanation the legends of Tables 1 and 2.

a background set we now compared the ligands of a subgroup against all other ligands in the entire group. A group within a group is called a *subgroup*; the group that contains the subgroup is denoted as the *supergroup*. Substructures that have a significant preference for either the subgroup or the supergroup are denoted as either *specific* or *avoiding*, respectively. Substructures that are specific for the supergroup are denoted as *generic*. Specific substructures are those that set ligands from one subgroup apart from ligands of the neighboring subgroups. Avoiding substructures are those that seem to avoid ligands of the subgroup but do occur in neighboring subgroups. Generic substructures are those that are common to a subgroup and the neighboring subgroups. Substructure lists not provided in this article are available as Supporting Information.

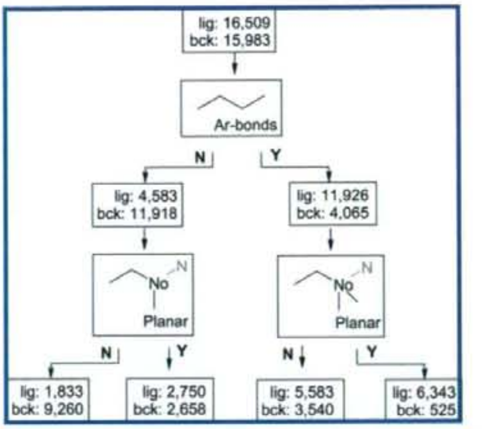
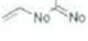
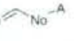
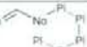
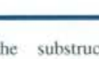


**Figure 4.** Hierarchical representation of the splits for GPCR ligands (hGPCR-lig) compared to the DIVERSet background collection, constructed without using the normal representation as in Figure 3. See the legend to Figure 3 for further explanation. As opposed to Figure 3, we now see a butyl chain without the amine group at the top of the hierarchy. Substructures containing the amine group are found one level down in the hierarchy.

**Table 4.** Best Discriminating Substructure Per Elaborate Representation for the DIVERSet Background Compounds Compared to the hGPCR-lig Collection[a]

| Representation | Substructure | Background compounds | GPCR ligands | P-Value |
|---|---|---|---|---|
| Normal |  | 6,247 (39%) | 1,511 (9%) | 9.03e-920 |
| Ar-bonds |  | 6,183 (39%) | 1,434 (9%) | 2.65e-938 |
| Ar-Atoms & Bonds |  | 6,998 (44%) | 2,140 (13%) | 1.92e-863 |
| Planar |  | 6,218 (39%) | 2,101 (13%) | 4.39e-658 |

[a] Note that the substructures do not have a geometric arrangement; the layout of double bonds and aromatic bonds is arbitrary.

First, the differences between aminergic receptor ligands and all other GPCR ligands in GLIDA were analyzed (Figure 7). The 'planar ring' representation yielded the most significant substructures, even though substructures with parts of a planar ring were found at lower positions in the list. The best-discriminating substructures were methyl- and ethyl-substituted amines; the amine group of the endogenous ligands (e.g., dopamine, (nor)epinephrine, acetylcholine, etc.)

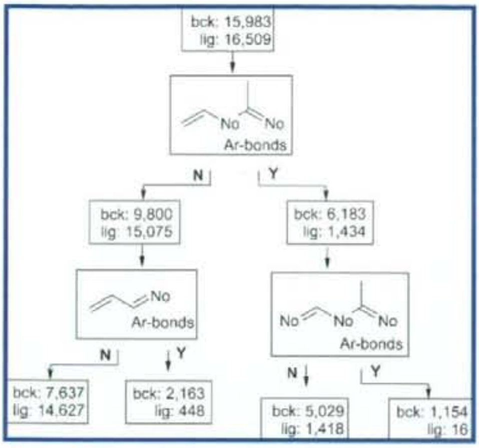


**Figure 5.** Hierarchical splits for the DIVERSet background compounds compared to GPCR ligands (hGPCR-lig). See legend to Figure 3 for further explanation. Here, it is particularly remarkable that the substructures all contain a double bonded heteroatom. The substructures at the top and right path match the carboxyl and ester groups, which are abundant in the DIVERSet.

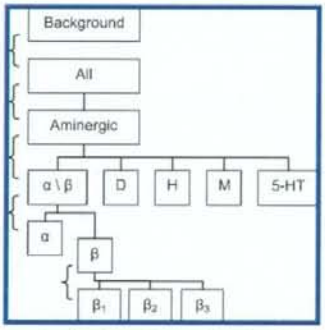


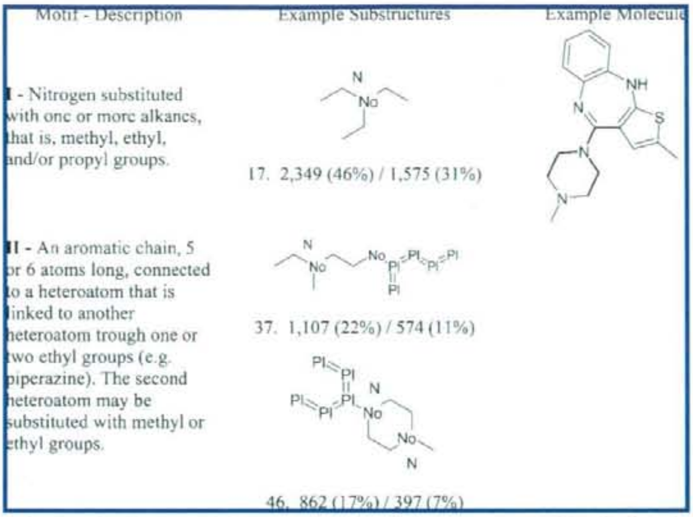
**Figure 6.** Schematic drawing of the subfamily hierarchy indicating the levels at which substructure analysis was performed (denoted by braces on the left side of the hierarchy). Boxes represent the sets and use the following labels: 'Background' - the ChemBridge DIVERSet, 'All' - total set of GPCR ligands found in GLIDA, 'Aminergic' - all aminergic receptor ligands, '$\alpha \setminus \beta$' - adrenoceptors, 'D' - dopamine receptors, 'H' - histamine, 'M' - muscarinic acetylcholine receptors, '5-HT' - serotonin receptors, '$\alpha$' - $\alpha$-adrenoceptors, '$\beta$' - $\beta$-adrenoceptors, and '$\beta_{1-3}$' - $\beta$-adrenoceptor subtypes 1 to 3.

is the common motif that accounts for naming of this group as biogenic amine receptors. The high occurrence of these substructures reflects efforts to mimic endogenous ligands by making analogs of these ligands (e.g., isoproterenol based on epinephrine). The opposite analysis was also conducted, yielding the structural features common to GPCR ligands excluding the aminergic ligands. The first, most significant, structural feature was a carbon atom connected to both a single-bonded heteroatom and to a double-bonded heteroatom. In the following positions, this heteroatom was specified as being a nitrogen atom, the second one as an oxygen atom. This reflects the carboxamide motif, found in peptide ligands (MW < 1000), which are part of other, nonaminergic, classes in GLIDA. The second important substructure consisted of two aromatic systems connected by a methylene group or by a single bond.

We continued by analyzing the five major aminergic targets individually against the other four. These five are the adrenoceptors (both alpha- and beta-), the dopamine receptors, the histamine receptors, the muscarinic acetylcholine receptors, and the serotonin receptors. Octopamine and trace amine receptors were not included due to scarce ligand information. For each analysis, the size of the aminergic control group was different due to the removal of duplicate entries, i.e. compounds that bind to more than one class. Although substructures found for the control group may be common to multiple GPCR targets, these are different from privileged substructures. Privileged substructures are discrete fragments, often scaffolds, found in one or more ligands for more than one target in the family.[47] Our analysis considers all possible substructures and yields only the most frequent substructures among the targets.

**Adrenoceptor Ligands.** An important feature of the adrenergic receptor ligands vs all other aminergic ligands is a substructure consisting of two heteroatoms connected by an ethyl group (Figure 8). The first heteroatom of this substructure is an oxygen atom specified as a hydroxy group, and the second is a nitrogen atom with a single hydrogen atom attached, meaning that this nitrogen is secondary. This chemical signature is representative for the motifs found in both $\beta$-adrenoceptor agonists and antagonists. An example containing this substructure is metoprolol, a $\beta_1$ antagonist (beta-blocker) used to treat hypertension. The second example substructure for motif I in Figure 8 has no atom specifiers for the heteroatoms, which means that this substructure also overlaps with the 1,2 diaminoethane substructure. A search for adrenoceptor (ant)agonists that have this substructure and not the hydroxyethylamine returned 58 hits, most of them specified as $\alpha$-adrenoceptor ligands in the database (second example structure of motif I). Note that both aforementioned substructures in the query had heteroatoms with one explicit hydrogen atom. At lower positions, the hydroxyethylamine motif reappears bonded to an aromatic system at the carbon atom that has the hydroxyl group attached. This is an exclusive element in $\beta$-adrenoceptor agonists. The substructures are essentially all part of the example substructure given for motif II which is found in 27% of the aminergic ligands. An example drug that has this motif is terbutaline, a $\beta_2$-adrenoceptor agonist used in the treatment of asthma. Substructures found less frequently in adrenergic ligands compared to aminergic ligands consisted of a nitrogen atom substituted at two or three positions, some as part of a largely saturated five- or six-membered ring, as found in e.g., apomorphine, a dopamine receptor ligand.

**Alpha- and Beta-Adrenoceptor Ligands.** We further examined the adrenergic receptor ligands, where we distinguished between $\alpha$- and $\beta$-adrenoceptors. The most significant features specific for the $\alpha$-adrenoceptor ligands (Figure 9) consist of a nitrogen atom substituted at three positions with methyl and ethyl groups (73% of ligands). One ethyl group can be connected to an aromatic system (33%) or to a heteroatom that is connected to an aromatic system (29%). An example drug containing this substructure is phenoxybenzamine, an $\alpha_1$-adrenoceptor antagonist used in the treatment of hypertension. The most significant substructures specific for $\beta$-adrenoceptor ligands (Figure 10) were all based on the 1-(ethylamino)propan-2-ol moiety (86% of ligands).



I - Nitrogen substituted with one or more alkanes, that is, methyl, ethyl, and/or propyl groups.

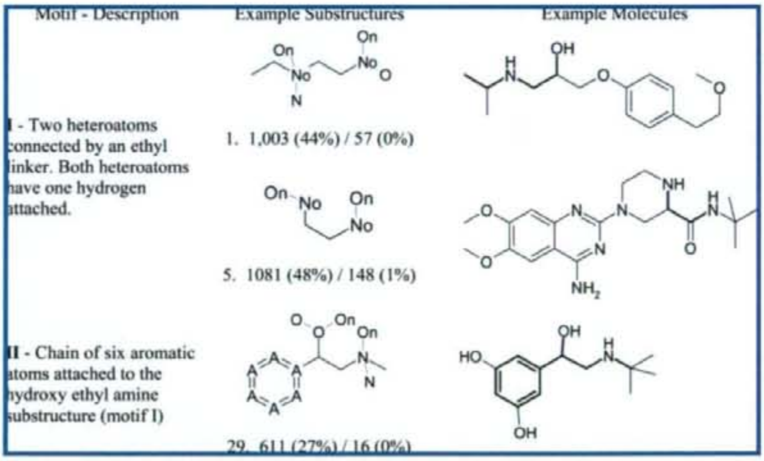
17. 2,349 (46%) / 1,575 (31%)

II - An aromatic chain, 5 or 6 atoms long, connected to a heteroatom that is linked to another heteroatom trough one or two ethyl groups (e.g. piperazine). The second heteroatom may be substituted with methyl or ethyl groups.

37. 1,107 (22%) / 574 (11%)

46. 862 (17%) / 397 (7%)

**Figure 7.** Common motif and example substructures for most significant substructures of aminergic ligands compared against all other GPCR ligands (GLIDA, 5k sampled), in planar ring systems representation. The 'Motif' number (Roman number, in bold) indicates the number of the found motif or structural theme. The motif number is followed by a short description, and one or more example substructures are provided. Below each example substructure, the position, occurrence in the active set (absolute and percentage), and occurrence in the control set (absolute and percentage) are listed. See the Materials and Methods section for further explanation about the representation of the substructures. For some motifs, an example molecule from the same class is provided, with the example substructure overlaid in bold. Here, an example drug containing motif I is olanzapine, which is used to treat schizophrenia, acting on dopamine $D_1$, $D_2$, and serotonin 5-$HT_2$ receptors (taken from ref 51).



I - Two heteroatoms connected by an ethyl linker. Both heteroatoms have one hydrogen attached.

1. 1,003 (44%) / 57 (0%)

5. 1081 (48%) / 148 (1%)

II - Chain of six aromatic atoms attached to the hydroxy ethyl amine substructure (motif I)

29. 611 (27%) / 16 (0%)

**Figure 8.** Common motif and example substructures for most significant substructures of the adrenoceptors ligands, in aromatic atoms and bonds representation. See the legend of Figure 7 for further explanation. The first example structure (for motif I) is metoprolol, a $\beta_1$-adrenoceptor antagonist (beta-blocker) used to treat hypertension (taken from ref 51). The second example structure for motif I is 4-(4-amino-6,7-dimethoxyquinazolin-2-yl)-N-tert-butylpiperazine-2-carboxamide, an α-adrenoceptor ligand and prazosin derivative found in GLIDA.[33] The example structure for motif II is terbutaline, a $\beta_2$-adrenoceptor agonist used in the treatment of asthma.

An example drug containing this substructure is propranolol, a nonselective beta-blocker, used in the treatment of hypertension. The most significant substructures specific for the $\beta_1$-adrenoceptor were all parts of a methylaminopropane substructure (81% of ligands). Here it should be noted that commercially available $\beta_2$-adrenoceptor ligands are *agonist*-shaving a structure such as terbutaline (Figure 8, first example motif II), whereas $\beta_1$-adrenoceptor ligands are mostly antagonists such as metoprolol (Figure 8, first example motif I). The most significant *avoiding* substructure for $\beta_1$-adrenoceptor ligands (50% of ligands), which at the same time occurs in $\beta_2$- and $\beta_3$-adrenoceptor ligands, consisted of an aromatic chain linked by an ethyl group to nitrogen that was linked by an ethyl group to an oxygen.

**Dopamine Receptor Ligands.** For the dopamine receptor ligands, two types of specific substructures were identified (Figure 11). The first substructure (in 30% of the ligands) consists of a chain of 4 to 5 aromatic atoms, connected to a nitrogen atom through a single carbon atom. This nitrogen is tertiary, as it is substituted with either two ethyl groups or one methyl and one ethyl group. The second substructure (12% of the ligands) consists of two aromatic chains of five or six atoms long that are linked through a heteroatom connected to N-methylethyleneamine, e.g. an N-methyleth-

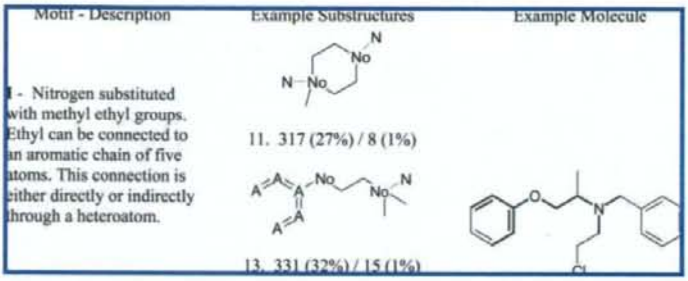

Motif - Description
Example Substructures
Example Molecule
I - Nitrogen substituted with methyl ethyl groups. Ethyl can be connected to an aromatic chain of five atoms. This connection is either directly or indirectly through a heteroatom.
11. 317 (27%) / 8 (1%)
13. 331 (32%) / 15 (1%)

**Figure 9.** Common motif and example substructures for most significant substructures of the α-adrenoceptors ligands versus β-adrenoceptor ligands, in aromatic atoms and bonds representation. An example is phenoxybenzamine, a $\alpha_1$-receptor antagonist used to treat hypertension. See the legend of Figure 7 for further explanation.

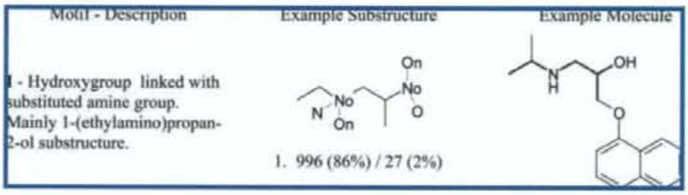

Motif - Description
Example Substructure
Example Molecule
I - Hydroxygroup linked with substituted amine group. Mainly 1-(ethylamino)propan-2-ol substructure.
1. 996 (86%) / 27 (2%)

**Figure 10.** Common motif and example substructures for most significant substructures of the β-adrenoceptor ligands versus α-adrenoceptors ligands, in aromatic bonds representation. An example drug containing this substructure is propranolol, a nonselective β-adrenoceptor antagonist (beta-blocker). See the legend of Figure 7 for further explanation.

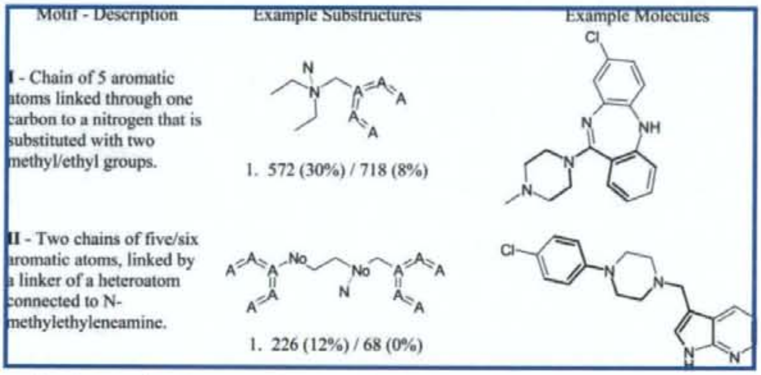

Motif - Description
Example Substructures
Example Molecules
I - Chain of 5 aromatic atoms linked through one carbon to a nitrogen that is substituted with two methyl/ethyl groups.
1. 572 (30%) / 718 (8%)
II - Two chains of five/six aromatic atoms, linked by a linker of a heteroatom connected to N-methylethyleneamine.
1. 226 (12%) / 68 (0%)

**Figure 11.** Common motif and example substructures for most significant substructures of the dopamine receptor ligands, in aromatic atoms and bonds representation. An example drug that has motif I is clozapine, an antipsychotic agent used in the treatment of schizophrenia.[52] Another example for motif I and also for motif II is compound L-745,870, a selective dopamine $D_4$ receptor antagonist.[53] See the legend of Figure 7 for further explanation.

ylenediamine linker. The terminal nitrogen of this linker may be substituted with an ethyl group. In both example molecules in Figure 11, the substructures overlap with the piperazine ring. The fact that the most significant substructures do not 'use' the entire piperazine moiety suggests that variations on the piperazine theme are possible when designing dopaminergic drugs. Similarly, the aromatic chains in motifs I and II overlap with several types of aromatic systems, e.g. five-membered or six-membered rings, containing either carbon or heteroatoms. This implies that aromaticity is the important feature and not so much the type of ring system that is used. Again, this finding offers further options for drug design.

**Histamine Receptor Ligands.** The most common motif (almost 50%) specific for histamine receptors is a chain of five aromatic atoms (Figure 12), where one or two aromatic atoms are specified as nitrogen atoms. These nitrogen atoms are separated by one aromatic atom; in some cases, one of the other neighboring aromatic atoms has an ethyl group attached. The majority of the significant substructures are chains, and actual ring closures, forming e.g. the imidazole ring as in histamine, are scarce. This seems counterintuitive at first sight, since the five-membered aromatic heterocycles are among the most obvious features when visually inspect-

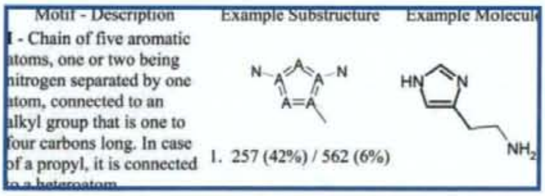

Motif - Description
Example Substructure
Example Molecule
I - Chain of five aromatic atoms, one or two being nitrogen separated by one atom, connected to an alkyl group that is one to four carbons long. In case of a propyl, it is connected to a heteroatom
1. 257 (42%) / 562 (6%)

**Figure 12.** Common motif and example substructures for most significant substructures of the histamine receptor ligands, in aromatic atoms and bonds representation. Histamine is provided as an example molecule containing this motif. See the legend of Figure 7 for further explanation.

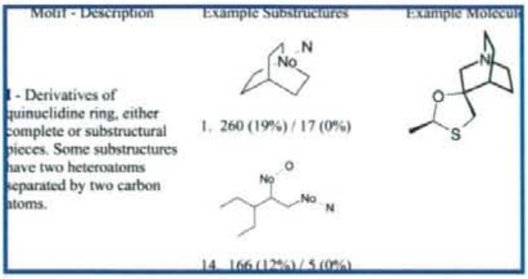

Motif - Description
Example Substructures
Example Molecule
I - Derivatives of quinuclidine ring, either complete or substructural pieces. Some substructures have two heteroatoms separated by two carbon atoms.
No
N
1. 260 (19%) / 17 (0%)
No
O
No
N
14. 166 (12%) / 5 (0%)
N
O
S

**Figure 13.** Common motif and example substructures for most significant substructures of the muscarinic acetylcholine receptor ligands, in aromatic bonds representation. An example molecule containing the quinuclidine ring is cevimeline, a muscarinic $M_3$ receptor agonist. See the legend of Figure 7 for further explanation.

ing the set. Although a common theme, the heterocycles in histamine receptor ligands all differ in size, ring-fusions, and heteroatoms. By considering substructures instead of complete ring fragments, it was thus possible to find structural similarities that have a much higher support among the ligands. This causes the high occurrence of the 'aromatic chains', since it is the most common feature among the diverse heterocycles.

**Muscarinic Acetylcholine Receptor Ligands.** Substructures of derivatives of the quinuclidine ring are the most common substructures for the muscarinic acetylcholine receptor ligands (Figure 13). It occurs in 19% of ligands for this class compared to 0% in other aminergic ligands. A second heteroatom may be attached, separated by two carbon atoms from the nitrogen. A typical example is civemeline, a muscarinic $M_3$ receptor agonist (Figure 13, first example). In this case, the second heteroatom would be the oxygen atom.

**Serotonin Receptor Ligands.** For the serotonin receptor ligands (Figure 14), the most specific substructure resembles the shape of the 2-ethylindole moiety (31% of ligands) that forms the core of serotonin, although a label specifying the nitrogen atom is missing. This is because this substructure covers the largest set of serotonin ligands (without becoming too general). In some cases, the ethyl group is attached to C3 rather than N1 of the core, which means that either the ethyl group or the atom specifier is not part of the substructure. The nitrogen atom can also be replaced by other heteroatoms or be absent in scaffolds that consist only of carbons, forming a planar ring system. This ring system consists of one aromatic ring instead of two. In fact, examples of all three cases were found among the ligands. At lower positions in the lists (position 22 in Table 16, Supporting Information) the same substructure (25% of ligands) is found with the nitrogen atom specifier. Examples containing this substructure are the endogenous ligand serotonin and the triptan antimigraine drugs such as sumatriptan (Figure 14).

**Aminergic Receptor Ligands.** When comparing the aminergic subgroups against all other aminergic ligands, the most significant features of the aminergic supergroup always have a low occurrence. This low occurrence is probably because these features are actually the features of one of the other subgroups of this class. For instance, the *avoiding* substructures for the serotonin receptor ligands are, among others, motif I from the adrenoceptor ligands (Figure 8). Therefore, these features can be considered substructures to avoid since they indicate possible side effects on other aminergic receptors. This is a more generalized way of a so-called antitarget analysis, to avoid GPCR-mediated side effects.[48] For dopamine, histamine, and serotonin receptor ligands, the most significant substructures of the other aminergic ligands (the *avoiding* substructures) are dominated by a motif of two heteroatoms connected by an ethyl linker. Since this resembles the substructures found for the ($\beta$) adrenoceptors (Figure 8), in both frequency and shape, this class probably dominates the avoiding substructure lists of the other four classes.

**General Observations.** In the following, we will discuss the representations and substructure selection criteria employed and their likely influence on the results obtained. First, the extraction of substructures discards any geometric information such as bond orientation. This loss of information may be appreciated, however, as it is beneficial for extracting more 'abstract' features in molecules. For instance, opposite *cis*−*trans* isomers may contribute to the same double bond in a substructure. Similarly, a chain of aromatic bonds may be part of one or multiple fused ring systems. Chirality is also lost in our approach, an issue that holds for all substructure search methods. Inclusion of 3D-conformational aspects in substructure searching is an open area for further research. Second, the p-value was used to sort the substructures according to significance. However, this value is very small for the top findings, and the differences in significance of the substructures are small. Therefore, choosing the most significant substructure to split the set is arbitrary; all substructures at the top of the occurrence lists would be a very good choice. Not only the significance of the finding is important but also what the finding predicts. It might therefore be better to focus less on the p-value and more on the ratio of retrieval of GPCR ligands and background compounds. In the end, a scientist might be more interested in the percentage of GPCR ligands that can be identified

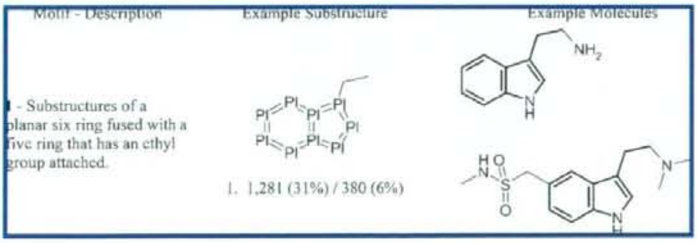


**Figure 14.** Common motif and example substructures for most significant substructures of the serotonin ligands, in planar ring systems representation. An example substructure is the endogenous ligand serotonin (above), or the antimigraine drug sumatriptan (below). See the legend of Figure 7 for further explanation.

from a set of molecules. Moreover, our approach focuses on selective substructures only; substructures that occur frequently in both sets are discarded. Although not selective, these substructures are still common structural features for GPCR ligands. Therefore, for the design of GPCR screening libraries, we also analyzed the features common in GPCR ligands in general, which were selected based on a minimum number of atoms/bonds (to remove trivial fragments) and frequency. A list of frequent, nonselective substructures is provided in Table 22, Supporting Information. This table lists all frequent (≥60% of molecules) substructures of more than six atoms that are found in GLIDA (5k sampled, normal representation). Some of the substructures (such as no. 23) can be associated easily with ligands of the aminergic receptor class, while other substructures (such as no. 1) seem to describe a more generic pattern of GPCR ligands. While this table will not be analyzed further by us at the current stage, fragments of the above type should be beneficial to be included in the design of GPCR libraries of any of the subtypes discussed in this work. As a final point, almost all significant findings were found when we used elaborate chemical representations, mainly with 'aromatic atom and bond types'. This suggests that this representation might be best for the application domain (i.e., GPCR activity prediction). Apparently, elaborate chemical representations add substantial value when searching for structural features typical for active compounds. Suggestions for further research would therefore be to extend the types of representations used, for instance by encoding the electronic properties of a molecule (for example, see ref 49).

## CONCLUSION

In this study, we analyzed frequently occurring substructures in GPCR ligands by comparing these with various control groups of compounds. Our analysis is complementary to employing privileged structures in ligand design,[50] in that it is not restricted to existing scaffold structures. It therefore offers further opportunities for introducing novelty in new chemical entities. We used different chemical representations for the molecules under consideration. As a result, we derived generalized substructural features for both ligands and control groups. The substructures found in the background set reflect the use of simple reactions that may have been employed to construct the library, for instance, the ester and carboxamide groups. In the GPCR ligand group, we found common as well as 'novel' substructures. First of all, our analysis identified well-known motifs (e.g., the side chain in $\beta$-adrenoceptor antagonists), which we considered a validation of our approach. In fact, the butylamine substructure (often linked to an aromatic moiety) occurred in 74% of the GPCR ligands compared to 26% of the background control group. Second, new structural patterns were also found, which may help medicinal chemists in their design efforts. As a typical example, we found fused 5:6 bicyclic ring systems in serotonergic ligands. These were identified in the so-called planar representation, indicating that aromaticity is not essential for both rings and that the precise location and nature of a heteroatom in the bicyclic core is not fixed. Indeed, the use of elaborate chemical representation gave the best, i.e. the most significant, description of the structural features that are important for a (sub)class of GPCR ligands.

## ACKNOWLEDGMENT

The authors thank Jeroen Kazius and Siegfried Nijssen for helpful discussions. This study was (partially) performed within the framework of the Dutch Top Institute Pharma, project number: D1-105.

**Supporting Information Available:** Lists of most significant substructures per compound set. This material is available free of charge via the Internet at http://pubs.acs.org.

## REFERENCES AND NOTES

(1) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39* (15), 2887–2893.

(2) Bemis, G. W.; Murcko, M. A. Properties of Known Drugs. 2. Side Chains. *J. Med. Chem.* **1999**, *42* (25), 5095–5099.

(3) Xue, L.; Bajorath, J. Distribution of Molecular Scaffolds and R-Groups Isolated from Large Compound Databases. *J. Mol. Model.* **1999**, *5* (5), 97–102.

(4) Xu, J. A New Approach to Finding Natural Chemical Structure Classes. *J. Med. Chem.* **2002**, *45* (24), 5311–5320.

(5) Nilakantan, R.; Nunn, D. S.; Greenblatt, L.; Walker, G.; Haraki, K.; Mobilio, D. A Family of Ring System-Based Structural Fragments for Use in Structure-Activity Studies: Database Mining and Recursive Partitioning. *J. Chem. Inf. Model.* **2006**, *46* (3), 1069–1077.

(6) Vinkers, H. M.; de Jonge, M. R.; Daeyaert, F. F. D.; Heeres, J.; Koymans, L. M. H.; van Lenthe, J. H.; Lewi, P. J.; Timmerman, H.; Van Aken, K.; Janssen, P. A. J. SYNOPSIS: SYNthesize and OPtimize System in Silico. *J. Med. Chem.* **2003**, *46* (13), 2765–2773.

(7) Todd, M. H. Computer-aided organic synthesis. *Chem. Soc. Rev.* **2005**, *34* (3), 247–66.

(8) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (5), 1708–1718.

(9) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naïve Bayesian Classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (1), 170–178.

(10) Lameijer, E. W.; Kok, J. N.; Back, T.; IJzerman, A. P. Mining a Chemical Database for Fragment Co-occurrence: Discovery of "Chemical Clichés". *J. Chem. Inf. Model.* **2006**, *46* (2), 553–562.

(11) Lipkus, A. H.; Yuan, Q.; Lucas, K. A.; Funk, S. A.; Bartelt, W. F.; Schenck, R. J.; Trippe, A. J. Structural Diversity of Organic Chemistry. A Scaffold Analysis of the CAS Registry. *J. Org. Chem.* **2008**, *73* (12), 4443–4451.

(12) Aronov, A. M.; Bemis, G. W. A minimalist approach to fragment-based ligand design using common rings and linkers: application to kinase inhibitors. *Proteins* **2004**, *57* (1), 36–50.

(13) Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2* (22), 3204–3218.

(14) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1273–1280.

(15) Glen, R. C.; Bender, A.; Arnby, C. H.; Carlsson, L.; Boyer, S.; Smith, J. Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs* **2006**, *9* (3), 199–204.

(16) Batista, J.; Godden, J. W.; Bajorath, J. r. Assessment of Molecular Similarity from the Analysis of Randomly Generated Structural Fragment Populations. *J. Chem. Inf. Model.* **2006**, *46* (5), 1937–1944.

(17) Batista, J.; Bajorath, J. Chemical Database Mining through Entropy-Based Molecular Similarity Assessment of Randomly Generated Structural Fragment Populations. *J. Chem. Inf. Model.* **2007**, *47* (1), 59–68.

(18) Sheridan, R. P.; Miller, M. D. A Method for Visualizing Recurrent Topological Substructures in Sets of Active Molecules. *J. Chem. Inf. Model.* **1998**, *38* (5), 915–924.

(19) Raymond, J. W.; Willett, P. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput.-Aided Mol. Des.* **2002**, *16* (7), 521–533.

(20) *Scitegic Pipeline Pilot, 6.1.5.0 Student Edition*; Accelrys, Inc.: San Diego, CA, 2007.

(21) *ClassPharmer, 4.5*; Simulations Plus, Inc.: Lancaster, CA, 2008.

(22) Wörlein, M.; Meinl, T.; Fischer, I.; Philippsen, M. A Quantitative Comparison of the Subgraph Miners MoFa, gSpan, FFSM, and Gaston*Knowledge Discovery in Databases: PKDD 2005*; 2005; pp 392−403.

(23) Engkvist, O.; Wrede, P.; Rester, U. Prediction of CNS Activity of Compound Libraries Using Substructure Analysis. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (1), 155–160.

(24) Borgelt, C.; Berthold, M. R. Mining Molecular Fragments: Finding Relevant Substructures of Molecules. In *Data Mining*, Proceedings of the 2002 IEEE International Conference on Data Mining, IEEE Computer Society: pp 51−58.

(25) Barratt, M. D.; Rodford, R. A. The computational prediction of toxicity. *Curr. Opin. Chem. Biol.* **2001**, *5* (4), 383–388.

(26) Nijssen, S.; Kok, J. N. A quickstart in frequent structure mining can make a difference. In *Conference on Knowledge Discovery in Data*, Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. Ronny, K., Gehrke, J., DuMouchel, W., Ghosh, J., Eds.; ACM Press: New York, U.S.A.: pp 647−652.

(27) Nijssen, S. *MULTI GASTON GrAph, Sequences and Tree ExtractiON algorithm, version 0.2*; Leiden Institute of Advanced Computer Science, Leiden University: Leiden, The Netherlands, 2004.

(28) Kazius, J.; Nijssen, S.; Kok, J.; Bäck, T.; IJzerman, A. P. Substructure Mining Using Elaborate Chemical Representation. *J. Chem. Inf. Model.* **2006**, *46* (2), 597–605.

(29) Heiko, H.; Christian, B.; Michael, R. B. *Large Scale Mining of Molecular Fragments with Wildcards*, Procedings of the 5th International Symposium on Intelligent Data Analysis, IOS Press: pp 495−504.

(30) Meinl, T.; Borgelt, C.; Berthold, M. R. *Mining Fragments with Fuzzy Chains in Molecular Databases*, Proceeding of the 2nd International Workshop on Mining Graphs, Trees and Sequences, Pisa, Italy, pp 49−60.

(31) Balakin, K. V.; Tkachenko, S. E.; Lang, S. A.; Okun, I.; Ivashchenko, A. A.; Savchuk, N. P. Property-Based Design of GPCR-Targeted Library. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1332–1342.

(32) Schnur, D. M.; Hermsmeier, M. A.; Tebben, A. J. Are Target-Family-Privileged Substructures Truly Privileged. *J. Med. Chem.* **2006**, *49* (6), 2000–2009.

(33) Okuno, Y.; Tamon, A.; Yabuuchi, H.; Niijima, S.; Minowa, Y.; Tonomura, K.; Kunimoto, R.; Feng, C. GLIDA: GPCR ligand database for chemical genomics drug discovery database and tools update. *Nucleic Acids Res.* **2008**, *36* (suppl_1), D907−912.

(34) hGPCR - lig. http://bioinfo-pharma.u-strasbg.fr:8080/hGPCRLig/index.jsp (accessed March 20, 2007).

(35) Wheeler, D. L.; Barrett, T.; Benson, D. A.; Bryant, S. H.; Canese, K.; Chetvernin, V.; Church, D. M.; DiCuccio, M.; Edgar, R.; Federhen, S.; Feolo, M.; Geer, L. Y.; Helmberg, W.; Kapustin, Y.; Khovayko, O.; Landsman, D.; Lipman, D. J.; Madden, T. L.; Maglott, D. R.; Miller, V.; Ostell, J.; Pruitt, K. D.; Schuler, G. D.; Shumway, M.; Sequeira, E.; Sherry, S. T.; Sirotkin, K.; Souvorov, A.; Starchenko, G.; Tatusov, R. L.; Tatusova, T. A.; Wagner, L.; Yaschenko, E. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2008**, *36*, D13−D21.

(36) Roth, B. L.; Lopez, E.; Beischel, S.; Westkaemper, R. B.; Evans, J. M. Screening the receptorome to discover the molecular targets for plant-derived psychoactive compounds: a novel approach for CNS drug discovery. *Pharmacol. Ther.* **2004**, *102* (2), 99–110.

(37) *DIVERSet*; ChemBridge Corp.: San Diego, CA, 2006.

(38) IUPHAR RECEPTOR DATABASE. www.iuphar-db.org (accessed May 9, 2007).

(39) Foord, S. M.; Bonner, T. I.; Neubig, R. R.; Rosser, E. M.; Pin, J.-P.; Davenport, A. P.; Spedding, M.; Harmar, A. J. International Union of Pharmacology. XLVI. G Protein-Coupled Receptor List. *Pharmacol. Rev.* **2005**, *57* (2), 279–288.

(40) Horn, F.; Bettler, E.; Oliveira, L.; Campagne, F.; Cohen, F. E.; Vriend, G. GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Res.* **2003**, *31* (1), 294–297.

(41) *JChem Standardizer, 3.2.11*; ChemAxon Kft.: Budapest, Hungary, 2007.

(42) Kazius, J.; McGuire, R.; Bursi, R. Derivation and Validation of Toxicophores for Mutagenicity Prediction. *J. Med. Chem.* **2005**, *48* (1), 312–320.

(43) *ISIS/Draw, 2.5*; MDL Information Systems, Inc.: San Leandro, CA, 2002.

(44) Strader, C. D.; Sigal, I. S.; Candelore, M. R.; Rands, E.; Hill, W. S.; Dixon, R. A. Conserved aspartic acid residues 79 and 113 of the $\beta$-adrenergic receptor have different roles in receptor function. *J. Biol. Chem.* **1988**, *263* (21), 10267–10271.

(45) Lindner, M. D. Clinical attrition due to biased preclinical assessments of potential efficacy. *Pharmacol. Ther.* **2007**, *115* (1), 148–175.

(46) Feher, M.; Schmidt, J. M. Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (1), 218–27.

(47) Bondensgaard, K.; Ankersen, M.; Thogersen, H.; Hansen, B. S.; Wulff, B. S.; Bywater, R. P. Recognition of Privileged Structures by G-Protein Coupled Receptors. *J. Med. Chem.* **2004**, *47* (4), 888–899.

(48) Klabunde, T.; Evers, A. GPCR Antitarget Modeling: Pharmacophore Models for Biogenic Amine Binding GPCRs to Avoid GPCR-Mediated Side Effects. *ChemBioChem* **2005**, *6* (5), 876–889.

(49) Marin, R. M.; Aguirre, N. F.; Daza, E. E. Graph Theoretical Similarity Approach To Compare Molecular Electrostatic Potentials. *J. Chem. Inf. Model.* **2008**, *48* (1), 109–118.

(50) Jacoby, E.; Fauchère, J.-L.; Raimbaud, E.; Ollivier, S.; Michel, A.; Spedding, M. A Three Binding Site Hypothesis for the Interaction of Ligands with Monoamine G Protein-coupled Receptors: Implications for Combinatorial Ligand Design. *Quant. Struct.-Act. Relat.* **1999**, *18* (6), 561–572.

(51) Klabunde, T.; Hessler, G. Drug Design Strategies for Targeting G-Protein-Coupled Receptors. *ChemBioChem* **2002**, *3* (10), 928–944.

(52) Meltzer, H. Y. Treatment-resistant schizophrenia--the role of clozapine. *Curr. Med. Res. Opin.* **1997**, *14* (1), 1–20.

(53) Bristow, L. J.; Kramer, M. S.; Kulagowski, J.; Patel, S.; Ragan, C. I.; Seabrook, G. R. Schizophrenia and L-745, 870, a novel dopamine D4 receptor antagonist. *Trends Pharmacol. Res.* **1997**, *18* (6), 186–188.

(54) van der Horst, E.; IJzerman, A. P. Computational Approaches to Fragment and Substructure Discovery and Evaluation. In *Fragment-Based Drug Discovery: a Practical Approach*, first edition; Zartler, E. R., Shapiro, M. J., Eds.; John Wiley and Sons, Ltd.: Chichester, West-Sussex, United Kingdom, 2008; pp 199–222.

CI8003896

BMC Genomics

BioMed Central

Research article

# Characterization of gene expression profiles for different types of mast cells pooled from mouse stomach subregions by an RNA amplification method

Soken Tsuchiya[1], Yuki Tachida[1], Eri Segi-Nishida[1,2], Yasushi Okuno[2,3], Shigero Tamba[1], Gozoh Tsujimoto[4], Satoshi Tanaka[1,5] and Yukihiko Sugimoto*[1]

Address: [1]Department of Physiological Chemistry, Graduate School of Pharmaceutical Sciences, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan, [2]Department of Systems Bioscience for Drug Discovery, Graduate School of Pharmaceutical Sciences, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan, [3]Department of PharmacoInformatics, Graduate School of Pharmaceutical Sciences, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan, [4]Department of Genomic Drug Discovery Science, Graduate School of Pharmaceutical Sciences, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan and [5]Department of Immunobiology, School of Pharmacy and Pharmaceutical Sciences, Mukogawa Women's University, Nishinomiya, Hyogo 663-8179, Japan

Email: Soken Tsuchiya - Soken.Tsuchiya@f13.mbox.media.kyoto-u.ac.jp; Yuki Tachida - yuki.tachida@090014.mbox.media.kyoto-u.ac.jp; Eri Segi-Nishida - eri.segi.nishida@pharm.kyoto-u.ac.jp; Yasushi Okuno - okuno@pharm.kyoto-u.ac.jp; Shigero Tamba - livegym-tmb@m6.gyao.ne.jp; Gozoh Tsujimoto - gtsuji@pharm.kyoto-u.ac.jp; Satoshi Tanaka - s_tanaka@mukogawa-u.ac.jp; Yukihiko Sugimoto* - ysugimot@pharm.kyoto-u.ac.jp

* Corresponding author

Published: 20 January 2009

BMC Genomics 2009, 10:35 doi:10.1186/1471-2164-10-35

Received: 9 July 2008
Accepted: 20 January 2009

This article is available from: http://www.biomedcentral.com/1471-2164/10/35

## Abstract

**Background:** Mast cells (MCs) play pivotal roles in allergy and innate immunity and consist of heterogenous subclasses. However, the molecular basis determining the different characteristics of these multiple MC subclasses remains unclear.

**Results:** To approach this, we developed a method of RNA extraction/amplification for intact *in vivo* MCs pooled from frozen tissue sections, which enabled us to obtain the global gene expression pattern of pooled MCs belonging to the same subclass. MCs were isolated from the submucosa (sMCs) and mucosa (mMCs) of mouse stomach sections, respectively, 15 cells were pooled, and their RNA was extracted, amplified and subjected to microarray analysis. Known marker genes specific for mMCs and sMCs showed expected expression trends, indicating accuracy of the analysis.

We identified 1,272 genes showing significantly different expression levels between sMCs and mMCs, and classified them into clusters on the basis of similarity of their expression profiles compared with bone marrow-derived MCs, which are the cultured MCs with so-called 'immature' properties. Among them, we found that several key genes such as *Notch4* had sMC-biased expression and *Ptgr1* had mMC-biased expression. Furthermore, there is a difference in the expression of several genes including extracellular matrix protein components, adhesion molecules, and cytoskeletal proteins between the two MC subclasses, which may reflect functional adaptation of each MC to the mucosal or submucosal environment in the stomach.

**Conclusion:** By using the method of RNA amplification from pooled intact MCs, we characterized the distinct gene expression profiles of sMCs and mMCs in the mouse stomach. Our findings offer insight into possible unidentified properties specific for each MC subclass.

## Background

Mast cells (MCs) are derived from hematopoietic stem cells and play important roles in allergic responses, innate immunity and defense against parasite infection. Unlike other blood cells, MCs migrate into peripheral tissues as immature progenitors and differentiate into mature mast cells. One of the unique features of MCs is that they show a variety of phenotypes depending on the different tissue microenvironment of their maturation [1]. In MCs, various MC-specific serine proteases are stored in the secretory granules, and their gene and protein expressions are dramatically altered when their cell environment is altered. For example, Reynolds *et al.* have shown that at least six distinct members of mouse MC-specific serine proteases are expressed in different combinations in different mast cell populations [2]. In addition, recent studies have shown that mature MCs vary in terms of what surface receptors and lipid mediators they express [3,4]. Because each mast cell population *in vivo* must play a specific role in the body, it is important to determine the character of each population of MCs.

Comprehensive gene expression analysis is a powerful approach to understand the characterization of various MC subpopulations. To date, several studies on microarray analysis of MCs have been conducted [5-7], but most of them dealt with MCs cultured *in vitro*. Alternatively, gene expression profiles of MCs isolated from skin and lung have been analyzed [3,8-10]. However, the numbers of MCs analyzed as one sample were relatively high and they were exposed to physical forces, enzymes and the anti-Kit antibody for purification, during which the original properties of the MCs may have been affected.

In the gastrointestinal tract, there are MCs that are mainly classified into two subclasses; mucosal MCs (mMCs) and submucosal MCs (sMCs) on the basis of their location, morphology (size and shape) and granule contents [11,12]. mMCs are mainly found in the mucosa of the gastrointestinal system, having chondroitin sulfate-containing granules, which are stained with toluidine blue but not safranin, and their activation occurs during parasite infection [13], while sMCs are localized in the submucosa of the gastrointestinal tract and their granules are rich in heparin and stained with both toluidine blue and safranin [1,11]. However, the molecular basis determining the differences in biochemical properties of these two MC subclasses remains uncertain, partially due to the difficulty of their isolation.

To overcome these problems, here we established a method of RNA amplification from intact MCs isolated from frozen tissue sections, which enables us to conveniently obtain the global gene expression pattern of MCs in various tissues. To validate this method, we first determined the minimum cell number required to achieve reproducible RNA amplification. We then compared the gene expression profiles obtained from small numbers of mMCs and sMCs in the mouse stomach, and found several key genes to be specifically expressed in one subclass of MCs, which may reflect some aspects of the distinct properties between the two MC subclasses in the gastrointestinal tract.

## Results and discussion

### *Development of an RNA amplification protocol to obtain gene expression profiles from a small amount of RNA*

To gain insight into the functional differences between the different subclasses of MCs, we employed three rounds of the T7-based RNA amplification method. Based on the preliminary experiments using peritoneal MCs and bone marrow-derived MCs (BMMCs), we estimated that a single MC yields 2 pg of RNA. Before we performed comparative analysis of MCs from different tissues, we first evaluated the accuracy and reproducibility of three rounds of the T7-based RNA amplification method, starting with the amount of RNA that can be obtained from a single MC. To assess this, we first compared the microarray results obtained from 5 μg of BMMC RNA prepared by the standard protocol with those obtained from the same RNA diluted $10^5$- or $10^6$-fold (30 pg, 10 pg and 2 pg) and subjected to three rounds of T7-based amplification (Figure 1a–c). Although three rounds of amplification yielded enough quantity of RNA for microarray analysis (>20 μg) even in the case of 2 pg RNA, scatter plot analysis revealed that the qualities of the obtained results were quite different between the samples from 5 μg and 2 pg RNA. The genes judged as 'Presence' in both 30 pg and 5 μg of RNA were 8,149 genes, which corresponded to 72% of genes judged as 'Presence' in the 5 μg of RNA (11,344 genes; Figure 1a), while only 4,116 genes were judged as 'Presence' in both 2 pg and 5 μg of RNA, which corresponded to only 36% of genes judged as 'Presence' in the 5 μg RNA (Figure 1c). The decrease in the number of genes judged as 'Presence' in the diluted samples (30 pg, 10 pg and 2 pg) may be due to the loss of low copy number RNA species during amplification.

We next examined the reproducibility of the microarray results obtained from two sets of 30 pg BMMC RNA samples (30 pg-1 and 30 pg-2) or two sets of 2 pg samples (2 pg-1 and 2 pg-2) (Figure 1d and 1e). In the 30 pg RNA samples, 7,537 (30 pg-1) and 8,777 (30 pg-2) genes were judged as 'Presence'. However, only 4,324 (2 pg-1) and 4,460 (2 pg-2) genes were judged as 'Presence' in each 2 pg RNA sample, again suggesting the loss of low copy number RNAs during amplification from a small amount of RNA. As to the reproducibility, 86% of the 'Presence' genes in the 30 pg-1 and 74% of 'Presence' genes in the 30 pg-2 sample were judged as 'Presence' in both 30 pg RNA

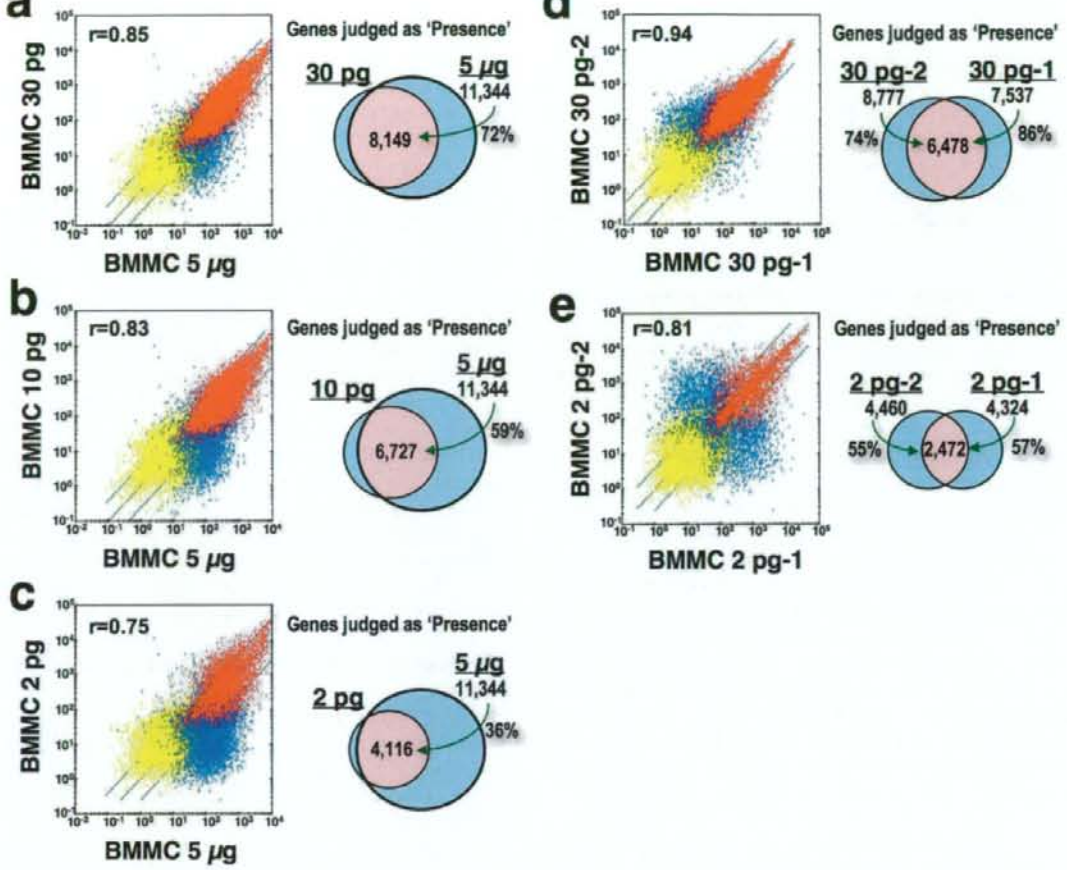


**Figure 1**

**Comparisons of three round-amplified products starting with very small quantities of RNA**. **(a-c)** Amplification biases in the products starting from a small quantity of RNA. Scatter plots of signal intensity obtained from 5 μg of BMMC RNA prepared by the standard protocol and from 30 pg (*a*), 10 pg (*b*) and 2 pg (*c*) of BMMC RNA prepared by three rounds of amplification are shown. **(d, e)** Reproducibility of the three-round amplification of a small quantity of RNA. Scatter plots of signal intensity between two independent products from 30 pg of BMMC RNA (BMMC 30 pg-1 and BMMC 30 pg-2) (*d*) or from 2 pg of BMMC RNA (BMMC 2 pg-1 and BMMC 2 pg-2) (*e*), are shown. Red dots show probe sets judged as "Presence", and yellow dots represent probe sets judged as "Absence" in both arrays. Blue dots show probe sets judged as "Presence" only in either array. The correlation coefficients (r) are presented. The same, four-fold induction and suppression thresholds are indicated as diagonal lines. Genes judged as "Presence" are placed in groups corresponding to pairwise overlaps shown in the accompanying Venn diagrams.

samples, while only 57% of 'Presence' genes in the 2 pg-1 and 55% of 'Presence' genes in the 2 pg-2 sample were judged as 'Presence' in both 2 pg RNA samples. These results suggested that the amplified products from the RNA from a single MC (about 2 pg) by the current method may include considerable amplification artifacts causing problems in accuracy and reproducibility. On the other hand, because of the higher reproducibility (>74%), we concluded that amplification from 30 pg RNA collected from 15 MCs would be suitable for the practical analysis of tissue MCs. Based on these results, we set our goal in this study to acquire gene expression profiles of MCs pooled from different regions. To minimize the influence of cell-to-cell variations within the same class and potential amplification artifacts, we prepared three sets of 15 MCs for each region and compared genes with significantly different expression between MCs from the different regions (Figure 2b). We chose stomach as the source organ, since we can isolate two kinds of MCs from the mucosa (mMC) and the submucosa (sMC) regions of the same sections, and mMCs and sMCs have been suspected to be different in several MC properties such as protease expression profile and sensitivity to safranin staining [1,11].

### *Gene expression profiles of submucosal and mucosal MCs from the stomach*

To visualize two kinds of MCs in the stomach without causing RNA degradation, the sections were fixed with carnoy's fixative and metachromatically stained with toluidine blue for a few seconds. sMCs and mMCs were microdissected using a patch pipette (Figure 2a and 2b). We prepared three sets of 15 MCs for each region, extracted their RNA and individually amplified them ($sMC_1$, $sMC_2$, $sMC_3$, and $mMC_1$, $mMC_2$, $mMC_3$). To improve the recovery of the extraction of as little as 30 pg of RNA, we used 'poly G' as a carrier, which does not interfere with the following RNA amplification or hybridization of the amplified product to the array (data not shown). To examine the effects of nonspecifically amplified artifact products, we performed the RNA extraction/amplification procedure without adding microdissected cells ("no cell") as a negative control (described in "*Materials and methods*"). The amplified RNAs of sMCs, mMCs and the "no cell" control were separately hybridized to a murine microarray. The signal values in the "no cell" sample were low in general and similar to the background levels (Figure 2c). The scatter plots of the samples independently prepared within the same group (e.g. $sMC_1$ vs $sMC_2$) showed a similar expression pattern; the average correlation coefficient for all probe-sets was 0.945 ± 0.004 and 0.893 ± 0.019 in sMCs and mMCs, respectively ($n = 3$). In contrast, the average correlation coefficient between sMCs and mMCs was 0.752 ± 0.034 ($n = 3$), which was much lower than those within the same group, suggesting that their gene expression patterns are different.

We further evaluated the accuracy and reproducibility of our method by other comprehensive analyses (hierarchical clustering analysis and principal component analysis [PCA]) using all probe sets. Microarray data obtained from sMCs, mMCs, skin-derived MCs, peritoneal MCs, BMMCs and non-MCs (macrophages and fibroblasts) were applied to these analyses. We first checked whether the amplification process in our method affects the global expression profile due to non-linear amplification. The results from the BMMC samples using RNA prepared by the standard protocol (BMMC-std) or the amplification method (BMMC-amp) were subjected to these analyses. Both hierarchical clustering analysis and PCA revealed that microarray data from BMMC-std and BMMC-amp were clustered in the same group (Figure 3a and 3b), suggesting that the global similarity in gene expression profiles is maintained during the amplification process. We next examined the similarity of expression patterns in three independent sMC or mMC samples. Upon clustering analysis and PCA, $sMC_{1-3}$ and $mMC_{1-3}$ were clustered in the same group, respectively. PCA also showed that the expression profiles of sMCs, mMCs and BMMCs are mutually different (Figure 3b).

We then compared the stomach-derived MCs (sMCs and mMCs) with skin-derived MCs, peritoneal MCs, BMMCs and non-MCs (macrophages and fibroblasts) by clustering analysis. The tissue-derived MCs (stomach MCs and skin MCs) were clustered separately from peritoneal MCs and BMMCs. These results may reflect different properties between tissue-derived MCs with firm adhesion to the neighboring cells and floating MCs without a tight contact. As to the similarity of MCs with fibroblasts and macrophages, it is reasonable that fibroblasts are most distant from MCs and macrophages are closer to MCs as a leukocyte family.

### *Validation of microarray results by real time RT-PCR analysis*

We next investigated whether the hybridization signals of known marker genes specific for sMCs and mMCs showed the expected expression trends [12,14]. The mMC-specific genes, mast cell protease 1 (*Mcpt1*) and 2 (*Mcpt2*) showed higher values in mMCs, while the sMC-specific marker genes, mast cell protease 4 (*Mcpt4*) and chymase 2 (*Cma2*), showed higher signal values in sMCs (Table 1 and Figure 4a) [15-29]. On the other hand, MC-common markers such as kit oncogene (*Kit*) and Fcε receptor (*Fcer1a*) showed significant signal values with no bias between mMCs and sMCs. To further evaluate the results, we measured the expression levels of these marker genes by real-time RT-PCR using RNA from the independently isolated MCs (Figure 4b). Moreover, we randomly selected three genes showing 'mMC-biased' expression and another three genes showing 'sMC-biased' expression; expression of these genes in MCs has not been reported previously (Figure 4a). There were no significant differences in the expression levels of *Kit* and *Fcer1a* between mMCs and sMCs. In contrast, the mMC-specific markers *Mcpt1* and *Mcpt2* and the 'mMC-biased' genes, *Anxa10*, *Ctse*, and *Fos* showed higher expression in mMCs, and the sMC-specific markers *Mcpt4* and *Cma2* and the 'sMC-biased' genes, *Cnn1*, *Ces3*, and *Cpe* showed higher expression in sMCs. These results indicate that the microarray

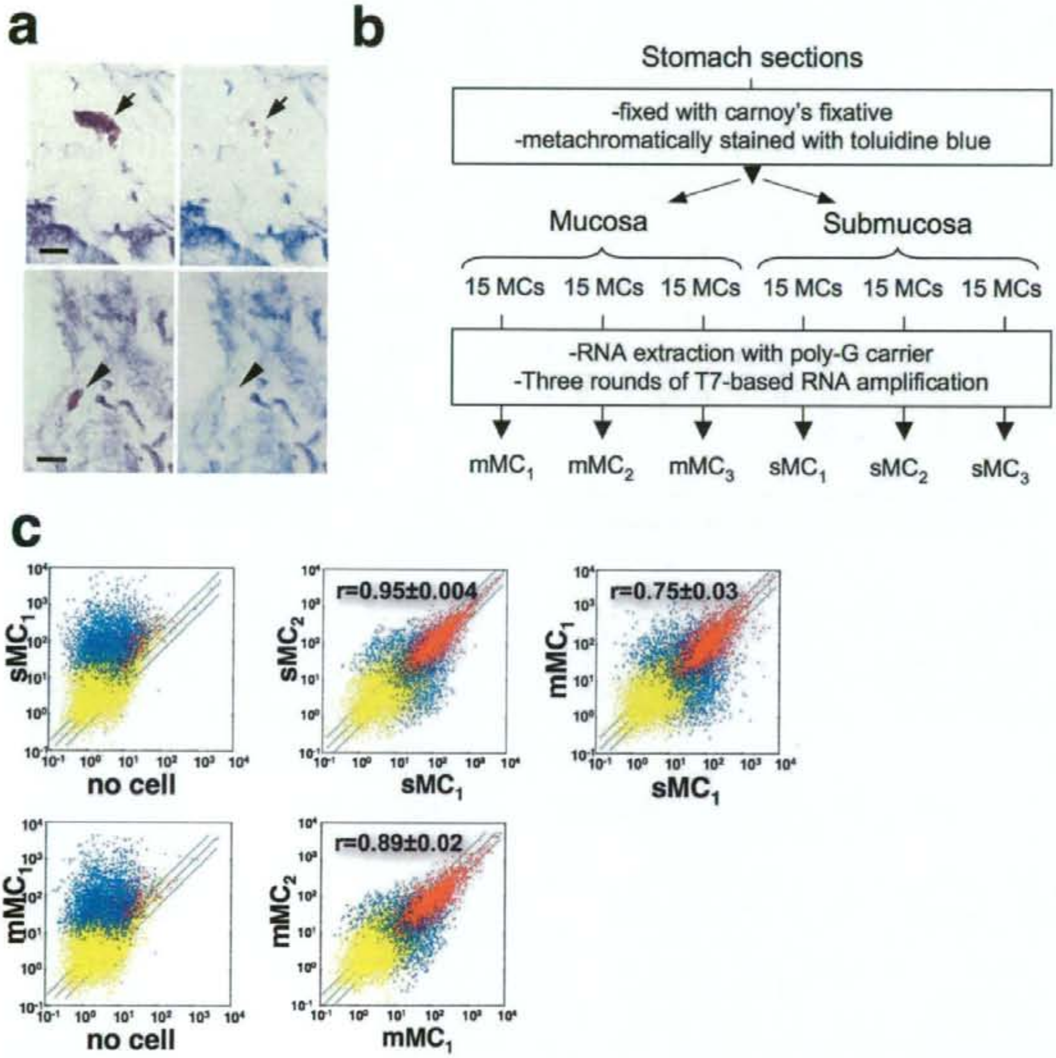


**Figure 2**

**Gene expression profiles of sMCs and mMCs from stomach tissue**. (**a**) Isolation of toluidine blue-stained MCs in the submucosa (sMC; *upper panels*) and the mucosa (mMC; *lower panels*) of stomach sections. A sMC (*arrow*) and mMC (*arrowhead*) that was metachromatically stained with toluidine blue before microdissection (*left panels*) disappeared after microdissection with a patch pipette (*right panels*). *Bars*, 10 μm. (**b**) Outline of the experimental strategy. (**c**) Labeled and fragmented antisense RNAs of three individual sMC samples, three individual mMC samples and the 'no cell' samples were hybridized to a Murine Array. Scatter plots for 'no cell' (x axis) and $sMC_1$ (y axis) (*upper left*), 'no cell' (x axis) and $mMC_1$ (y axis) (*lower left*), $sMC_1$ (x axis) and $sMC_2$ (y axis) (*upper center*), $mMC_1$ (x axis) and $mMC_2$ (y axis) (*lower center*), $sMC_1$ (x axis) and $mMC_1$ (y axis) (*upper right*) are shown. The correlation coefficients (r) for comparison within $sMC_{1-3}$, within $mMC_{1-3}$ and between sMCs and mMCs are presented as means ± S.D. Red dots show probe sets judged as "Presence", and yellow dots represent probe sets judged as "Absence" in both arrays. Blue dots show probe sets judged as "Presence" only in either array. The same, two-fold induction and suppression thresholds are indicated as diagonal lines.

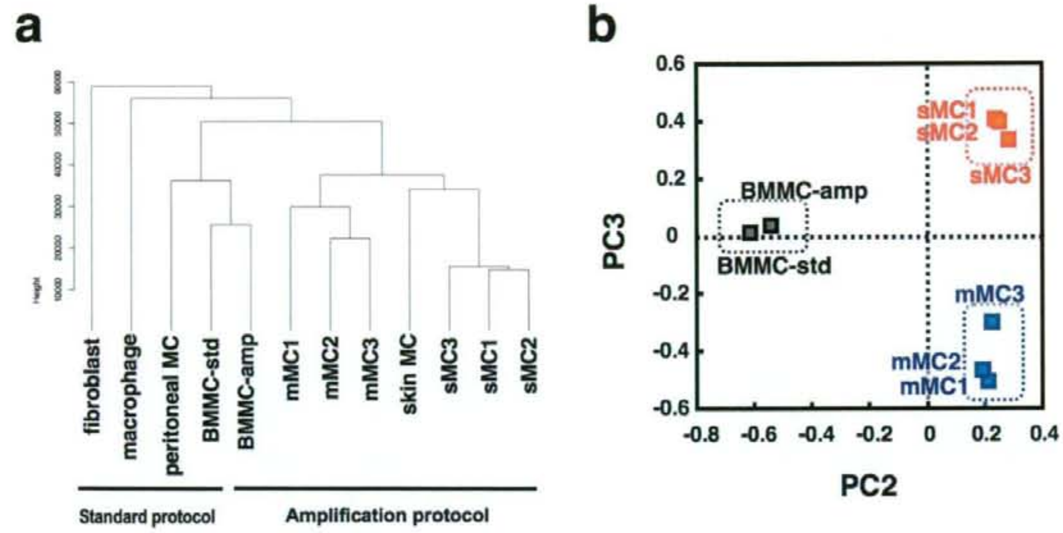


**Figure 3**

**Global gene expression analysis of $sMC_{1-3}$ and $mMC_{1-3}$.** (**a**) Hierarchical clustering of global gene expression of various preparations of MCs and non-MCs. Three-round amplified products of $sMC_{1-3}$, $mMC_{1-3}$, skin MCs and BMMCs, and the standard products of BMMCs, peritoneal MCs, macrophages and fibroblasts were analyzed. (**b**) The principal component analysis (PCA) reveals different gene expression profiles of $sMC_{1-3}$, $mMC_{1-3}$, and two preparations of BMMCs. The blue dotted square indicates mMCs, the red dotted square indicates sMCs, and the black dotted square indicates BMMCs.

results are reliable and reflect the gene expression profiles of intact sMCs and mMCs in the stomach.

### *Clustering analysis of the gene expression profiles and functional categorization between sMCs and mMCs*

Of the ~12,000 genes represented on the oligonucleotide array, we selected 1,272 genes whose expression levels between $sMC_{1-3}$ and $mMC_{1-3}$ were significantly different ($p < 0.05$, Limma's $t$ test). The expression level of each gene was normalized by its level in BMMCs, which are cultured MCs with so-called 'immature' properties, and the selected genes were classified into seven clusters using the $k$-means clustering algorithm (CL1-7; Figure 5a and Additional file 1). We also classified the genes into functional categories, and the representative genes are listed (Figure 5b). Among them, 666 genes (52.4%) showed sMC-biased expression (*CL1-3*); in 78% (519 genes) of sMC-rich genes, the expression levels were relatively low in BMMCs and augmented in sMC (*CL1&2*). For example, the expression level of *Mcpt4* was relatively low in BMMCs, and if the expression profile of BMMCs reflects the immature properties of MC progenitors, *Mcpt4* can be concluded to be induced during the final maturation into sMCs. Interestingly, the sMC marker genes *Mcpt5* and *Mcpt6* were classified into *CL2/3*, suggesting that these genes were expressed to some extent in 'immature' BMMCs, but their expression was suppressed during maturation into mMCs. On the other hand, 606 genes (47.6%) showed mMC-biased expression (*CL4-7*); in 51% (334 genes) of mMC-rich genes, their expression levels in BMMCs were low but were augmented in mMCs (*CL4&5*). For example, expression of *Mcpt1* was low in 'immature' BMMCs but was drastically induced during maturation into mMCs.

### *Protein expression of Notch4 in sMCs and Ptgr1 in mMCs in stomach tissue*

Among the genes showing differential expression (Figure 5b), we further focused on the expression of *Notch4* in sMCs and *Ptgr1* in mMCs, both of which have never been previously characterized in MCs. The *Notch4* gene product is a member of the Notch family, consisting of transmembrane receptors which are activated by cell surface ligands on adjacent cells. Recent studies have suggested that Notch signaling is involved in lymphocyte and mast cell differentiation [30,31]. We first confirmed that *Notch4* expression is significantly higher in the separately pooled sMCs than mMCs by real-time RT-PCR (data not shown).

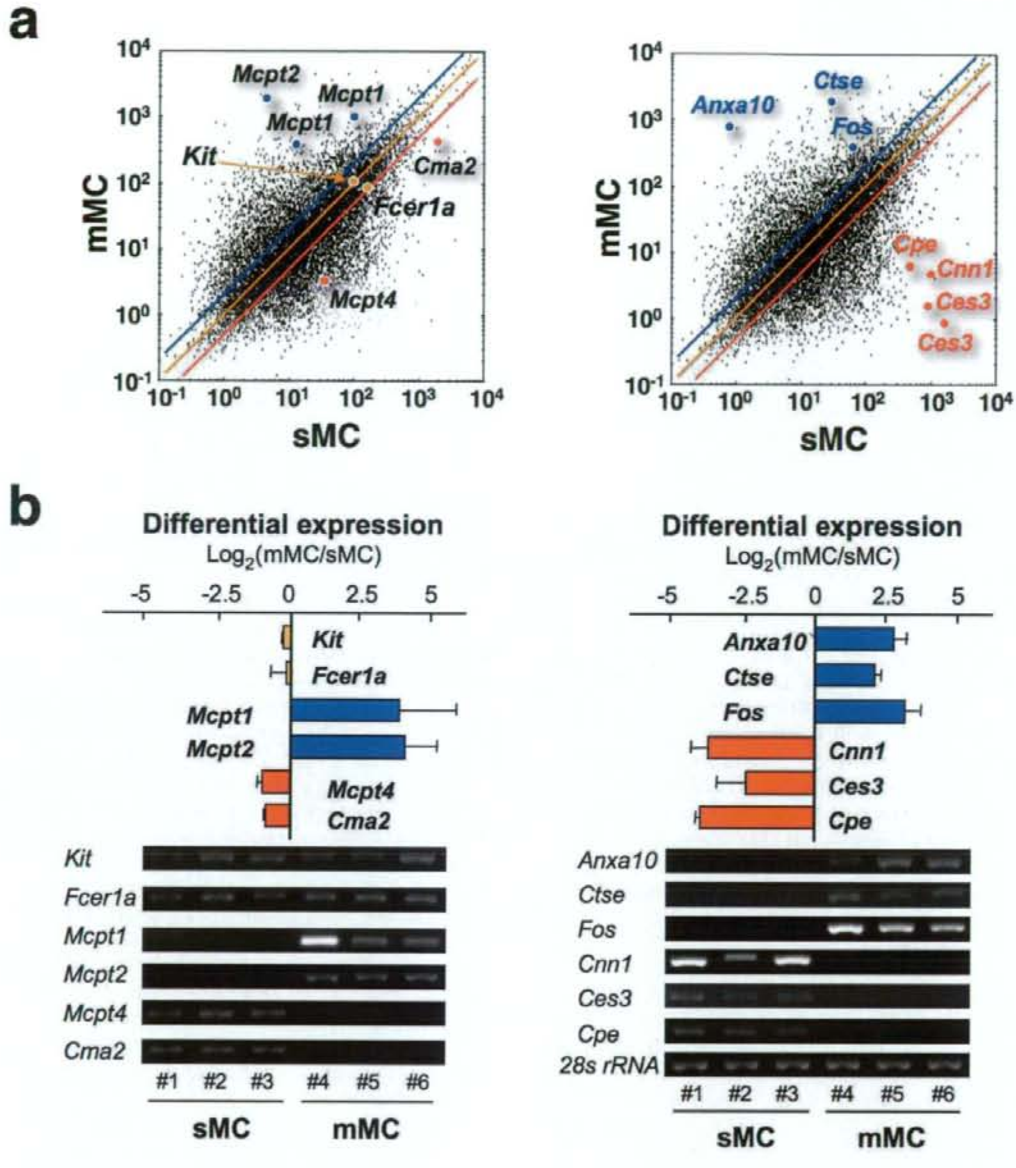


**Figure 4**

**Validation of the differentially expressed genes between sMCs and mMCs**. (**a**) sMC-specific (*Cma2*, *Mcpt4*), mMC-specific (*Mcpt1*, *Mcpt2*) and MC-common markers (*Fcer1a* and *Kit*) (*left panel*) and six randomly selected genes (*Ces3*, *Cnn1*, *Cpe*, *Anxa10*, *Ctse* and *Fos*) (*right panel*) are indicated in the representative scatter correlation graphs between $sMC_I$ and $mMC_I$. The same, two-fold induction and suppression thresholds are indicated as a yellow, blue and red line, respectively. (**b**) The expression levels of the genes in (a) were verified by real-time RT-PCR. The values represent the ratio of relative expression levels of mMCs to sMCs, and are shown as mean ± S.D. (n = 3). The specificity of the PCR product was confirmed by gel electrophoresis and analysis of the melting temperature. The expression level of each gene was normalized to 28S ribosomal RNA.