ELSEVIER

Journal of Molecular Graphics and Modelling 21 (2002) 185–194

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

Mining the Chemical Abstracts database with pharmacophore-based queries

John B. Bremner [a], Kao Castle [a], Renate Griffith [b,*], Paul A. Keller [a], Damon D. Ridley [c]

[a] *Department of Chemistry, University of Wollongong, Wollongong, NSW 2522, Australia*
[b] *School of Environmental & Life Sciences, University of Newcastle, Callaghan, NSW 2308, Australia*
[c] *School of Chemisty, University of Sydney, Sydney, NSW 2006, Australia*

Received 10 February 2002; accepted 28 June 2002

Abstract

A method is described to convert 3D patterns of pharmacophoric groups into 2D queries for molecular substructure searches of the Chemical Abstracts database with SciFinder Scholar®. The results of such searches and the options for refinement of the hit lists are presented using a rigid tetrahydroisoquinoline-carbazole (IQC) hybrid molecule fitted onto previously developed pharmacophores for subtype-selective α_1-adrenergic receptor antagonists as an example. The compounds retrieved were further analysed by limiting their physical properties to 'drug-like' ranges and by enumerating the ring skeletons they contain. Selected ring skeletons were evaluated by fitting them on to the original pharmacophores. Several structurally novel rigid ring skeletons were found with this new database mining technique which are potentially useful as leads in the design of α_{1B} selective adrenergic receptor antagonists.

Keywords: Computer-aided drug design; 3D pharmacophore; 2D query; SciFinder Scholar®; Database mining; High-throughput virtual screening; Selective antagonists; α_1-Adrenoceptor subtypes

1. Introduction

Pharmacophores have been used extensively in ligand-based computer-aided drug design [1]. A pharmacophore can be described as "a set of structural features in a molecule that is recognised at a receptor site and is responsible for that molecule's biological activity" [2].

Once a pharmacophore has been developed for a particular set of biologically active molecules, it can be used to search databases of 3D structures with the aim of finding new, structurally different lead molecules with the desired biological activity. As an example, the National Cancer Institute database of compounds screened against cancer cell lines and against human immunodeficiency virus (HIV), and the Maybridge® database, are available for searching with pharmacophores with both Tripos® and Accelrys® software packages. Although novel approaches to database searching with pharmacophores are constantly being developed [1,3–5], both the availability and the limited content of the databases are often limitations in these approaches.

SciFinder Scholar® [6] allows access to databases from the Chemical Abstracts Service (CAS) and from the US National Library of Medicine, thus opening the largest compound and literature collection publicly available. The structures of over 18 million substances and over 15 million sequences may be accessed using various 2D query types, including substructure searches. No 3D queries are currently possible, although 3D structures may be viewed, and 2D structures may be imported into alternative software allowing 3D analysis. In addition to its size, a feature of the CAS database is the availability of calculated physical properties. Further, through SciFinder Scholar® there are many possibilities for query refinement and query analysis, and for direct access to available literature on a particular compound.

Given the potential power of SciFinder Scholar® for database mining, we investigated the following questions.

- Can a 3D pattern of pharmacophoric groups be converted into a 2D query for a molecular substructure search of the CAS database with SciFinder Scholar®?
- What are the results of such a query and what options for refinement exist?
- Are these results useful for drug design, i.e. are previously unrecognised and structurally novel potential lead compounds located?

* Corresponding author.
E-mail address: renateg@mail.newcastle.edu.au (R. Griffith).

1093-3263/02/$ – see front matter
PII: S1093-3263(02)00141-9

We report here on our efforts to answer these questions using previously developed pharmacophores for subtype-selective α_1-adrenergic receptor antagonists as a test case.

2. Methodology

2.1. Generation of initial query

In order to generate a query for SciFinder Scholar®, a lead compound, fitted onto the pharmacophore(s) of interest is used. A rigid ligand is chosen in this example in order to preserve the correct spatial arrangement of the pharmacophoric groups (features) as much as possible. The same strategy could be followed, however, with any lead compounds. In mining for substances in SciFinder Scholar®, the initial substance query should be as general as possible. It should include only minimal structural elements, since numerous options may always be utilized at a later stage to filter answers in a predictable manner. A general initial query also avoids the introduction of a bias in the search from the structure(s) the initial query is based on.

2.2. Structure Refine and Analyse options in SciFinder Scholar®

There are many ways to refine and analyse structure queries (or perform what is also known as filtering of hit lists from database searches) using SciFinder Scholar®, and the main alternatives are as shown in Fig. 1. SciFinder Scholar® is a version of the software only available to academic institutions. The software available to other users is called SciFinder Scholar®, and this is identical to SciFinder Scholar® except for the inclusion of some additional features.

Once the initial query has been searched, answers from the CAS substance database (REGISTRY) are presented and references may be obtained for these substances in CAPLUS (the bibliographic database from CAS) and MEDLINE (the bibliographic database from the US National Library of Medicine). The exact number of answers retrieved with any query will increase every day, as new compounds are added to the databases.

The options for filtering the results from the initial queries are of two broad types, *Refine* and *Analyse*, and the main difference is that the former requires a specific input while the latter tabulates alternatives from which choices may then be made. The applications of some of these analysis tools are discussed in Section 3 later. The outcomes from *Refine* or *Analyse* operations are new substance answer sets which may be further analysed or refined, or for which references may be obtained.

Apart from refinement by structure, answers may also be refined by commercial availability. Thus, in SciFinder Scholar®, REGISTRY is linked to the database CHEMCATS® which contains over 900,000 commercially available substances.

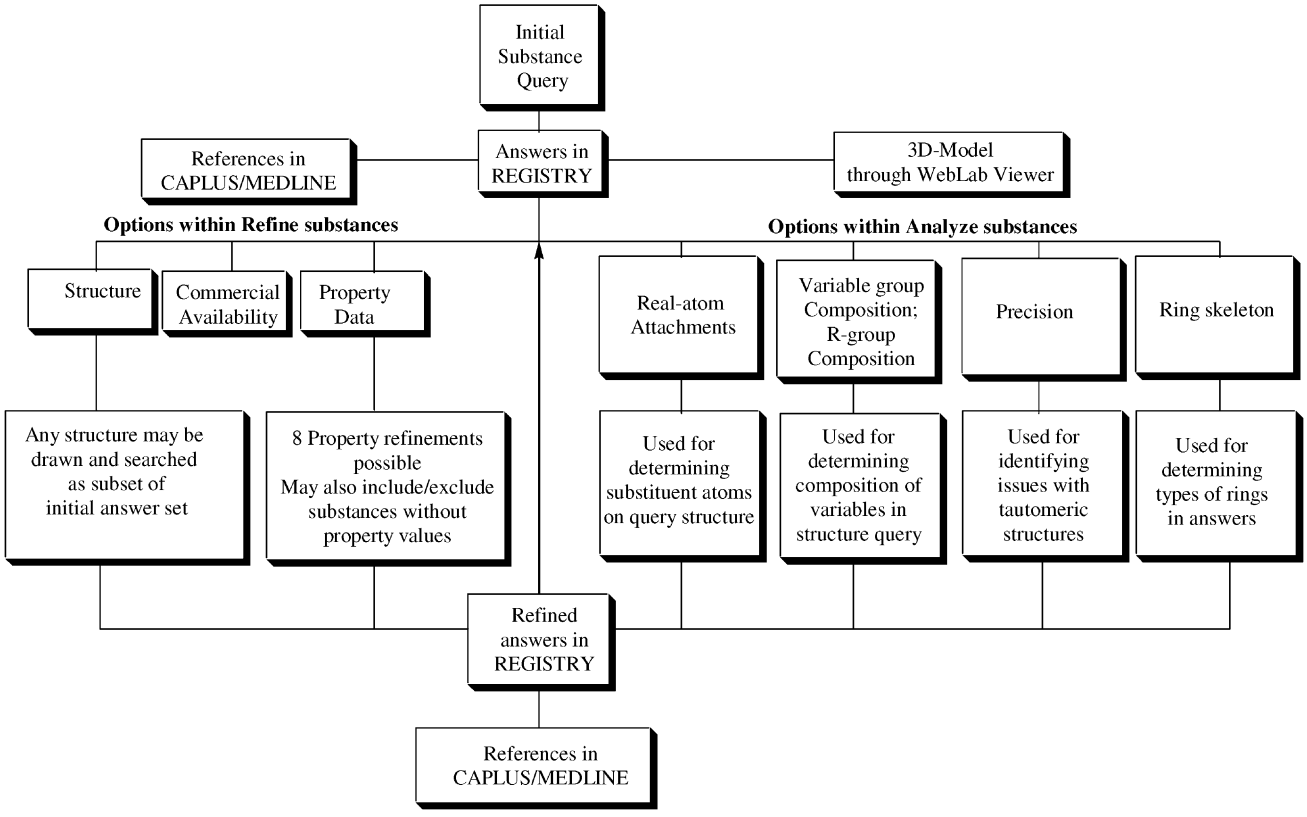


Fig. 1. Flow chart summarising the possibilities of refinement and analysis of the results from SciFinder Scholar® queries.

Finally, answers may be refined by property data. The physical property values are calculated using CAS connection tables and software developed by Advanced Chemistry Development, and information on the algorithms used for the calculations is available [7]. The record also contains a link to the 3D model which also is built from the structure connection tables [8].

The eight calculated properties are pK_a, number of hydrogen donors and hydrogen acceptors, freely rotatable bonds, molecular weight, molar solubility, log *P* and log *D*. This allows application of Lipinski's "Rule of Five" [9] as a filter to remove compounds that are unlikely to be successful drug candidates.

Currently, calculated physical properties are available for over 4.5 million substances, and those include substances which have been identified with a biological role by CAS and substances which are commercially available. Over 8 million additional substances will be added soon.

3. Results and discussion

To illustrate the methodology described above, a rigid antagonist against α_1-adrenergic receptors was used. The tetrahydroisoquinoline-carbazole hybrid (IQC), shown in Fig. 2, together with binding data for the three α_1-adrenoceptor subtypes, was suitable for this purpose because it displayed reasonably high affinity for all the three subtypes and also fits the three pharmacophores developed [10], as shown in Fig. 3. A previously identified α_{1C} subtype was later shown to be the same as the α_{1A} subtype, thus the three α_1 subtypes are now denoted as α_{1A}, α_{1B} and α_{1D} (capital letters in the subscripts denote the pharmacological characterisation, as opposed to the cloned receptors) [11].

3.1. Initial search

The first step was to design a query to retrieve molecules with the potential to bind selectively to the α_{1B} subtype over the α_{1A}- and α_{1D}-adrenoceptor subtypes. The α_{1B} subtype was chosen because there are currently no high-affinity,

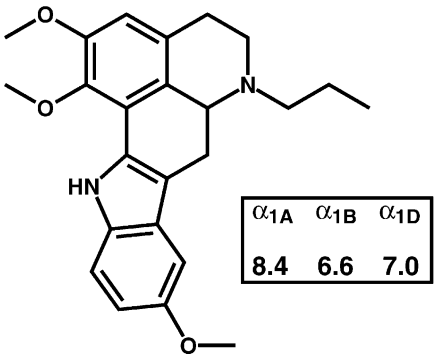


Fig. 2. Structure of the tetrahydroisoquinoline-carbazole hybrid (IQC) [10], together with binding data (pK_i) for the three α_1-adrenoceptor subtypes (determined for the racemic compound).

highly selective antagonists available for this receptor [12] and also because the pharmacophore differed markedly from the pharmacophores developed for the other two subtypes (see below).

When comparing IQC mapped onto the three pharmacophores for subtype selective α_1 antagonists (Fig. 3), it became obvious, that a potentially α_{1B} selective antagonist should, if possible, avoid all those features associated with high affinity towards the other two subtypes. Thus, any potential hydrogen bond acceptor groups should be avoided on both phenyl rings. Indeed, the phenyl ring (in this paper, 'phenyl ring' refers to a fused 'benzo' moiety) of the tetrahydroisoquinoline part of IQC should also be avoided. This left only the indole system as potentially important for a selective α_{1B} ligand, because it incorporated both features unique to the α_{1B} pharmacophore, i.e. a phenyl ring close to a hydrogen bond donor group. It is interesting to note in this context that bromotopsentin, a low affinity ligand which displays about 20-fold selectivity for α_{1B} over α_{1A}, is a very rigid molecule and contains two indole systems [13]. The basic tetrahydroisoquinoline nitrogen atom (positively charged at physiological pH) was also retained as a feature, despite the fact that it is common not only to all α_1-adrenergic ligands, but to all known high-affinity ligands for all bioamine receptors of the G-protein coupled receptor

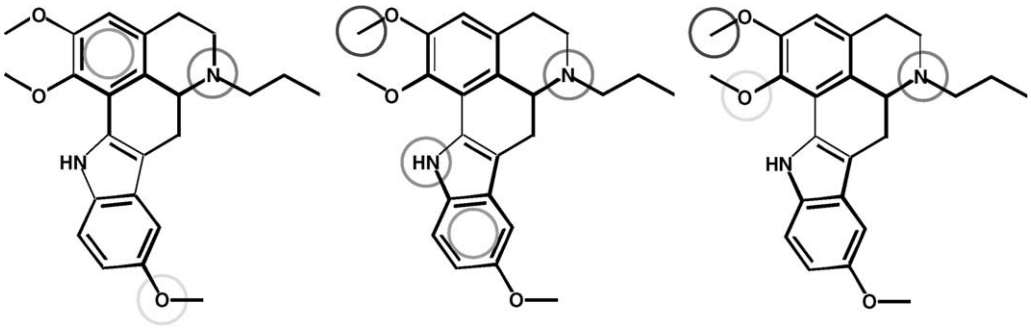


Fig. 3. The features of the three α_1 pharmacophores mapped on to IQC (from left to right, α_{1A}; α_{1B}; and α_{1D}). Features are represented by colour coded circles: red, positive charge at physiological pH; green, hydrogen bond acceptor; orange, aromatic ring system; blue, hydrophobic group.

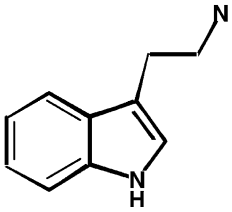


Fig. 4. The initial query (tryptamine structure).

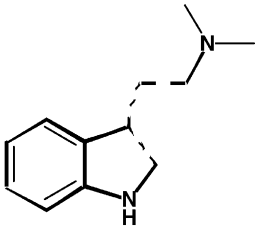


Fig. 5. The refined query with three unspecified bonds (with respect to bond order) indicated by dotted lines; see text for discussion.

superfamily. While the basic nitrogen atom is detrimental to selectivity, it has been shown to be crucial for high affinity of ligands for these receptors [14].

The initial query structure derived from IQC with the *minimum* structural features necessary for an α_{1B} selective ligand was, thus, a tryptamine unit, as shown in Fig. 4.

However, a substructure search on this structure produced over 190,000 answers of which over 113,000 were peptides containing tryptophane units. It was possible to eliminate these selectively through SciFinder Scholar®, but the process would require advanced options, so the better approach was to be a little more selective in the initial search.

Following further initial explorations, it was decided to commence the main search from the query as shown in Fig. 5. A substructure search on this query allowed retrieval of compounds with substituents at any of the available positions, as well as further rings attached to any of the bonds in the structures (including fused systems).

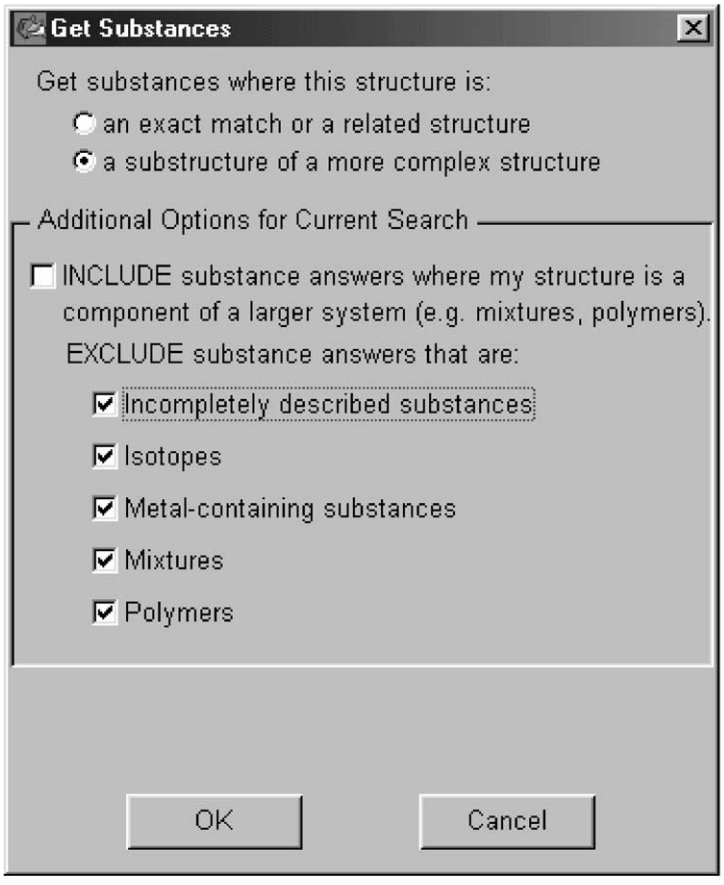


Fig. 6. Options chosen in SciFinder Scholar® for submission of the query as shown in Fig. 5 for a substructure search.

The ring bond indicated was unspecified (with respect to bond order) to allow for either indoles or dihydroindoles, and the "side chain" C–C bonds were unspecified in order to allow for the greatest number of options. On the other hand, it was not necessary to allow for double bonds to the side chain nitrogen since any such occurrence would necessarily give a quaternary nitrogen, not the tertiary basic nitrogen desired.

When a substructure search was requested, SciFinder Scholar® alerted the user to *Additional Options* and the choices made are indicated in Fig. 6. Thus, we were not interested in multi-component substances (mixtures) or polymers, nor in incompletely defined substances (those which were not fully described in the original article), isotopes or metal containing substances. This substructure search afforded almost 40,000 substances but this was an easily manageable answer set in SciFinder Scholar®.

3.2. Refining initial answers

It was immediately apparent from the initial answers that many included amides (e.g. Fig. 7). Since, α_{1B} selective

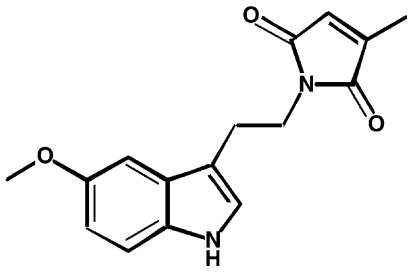


Fig. 7. An example of an amide retrieved with the search using the query described in Fig. 5 and the options described in Fig. 6.

ligands required a basic nitrogen, filtering to eliminate such amides was necessary. While there were several ways to achieve this, we chose first to refine the initial answers with the query shown in Fig. 8. We were conscious that such a refinement would eliminate quaternary carbons at the positions where hydrogens were now requested, but this was considered acceptable. The new answer set contained 35,986 substances including the structure, as shown in Fig. 7, and

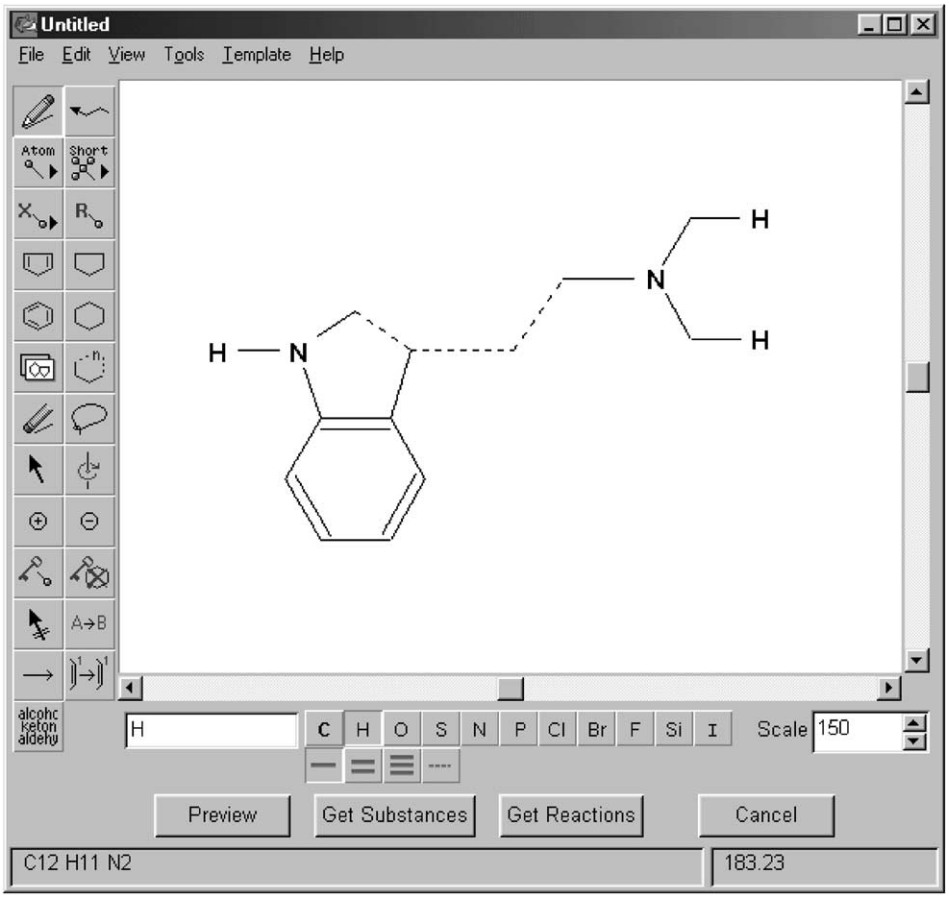


Fig. 8. Refinement of the answers from the query described in Fig. 5 with the options described in Fig. 6 to attempt to avoid retrieval of amide structures.

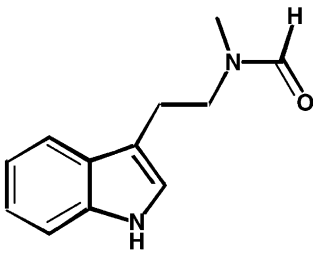


Fig. 9. An amide structure still "allowed" with the query described in Fig. 8.

numerous other amides. None of these had hydrogens at the positions requested.

It is helpful to understand here that, even with the query indicated in Fig. 8, SciFinder Scholar® needs to allow for substances of the type as shown in Fig. 9. Whenever amides (or indeed, whenever most carboxylic acids and derivatives, or carbonyl compounds) are possible answers, the tautomer issue needs to be addressed. For example, the compound in Fig. 10 can be interpreted as an amide (Fig. 10, left) or as hydroxypyridine (Fig. 10, right); 1,3,5-cyclohexanetrione can also be written as phloroglucinol. The full implications of the tautomer issue and elegant solutions have already been described elsewhere [15]. Accordingly, we next analysed

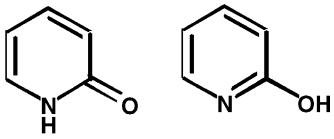


Fig. 10. Keto-enol tautomers: left, 2-oxo-1,2-dihydropyridine; right, 2-hydroxypyridine.

the substances by *Precision* and SciFinder Scholar® now indicated that there were 23,833 substances (Answer Set 1) which had the *Conventional Substructure*, another 11,860 substances were identified as *closely associated tautomers*, and a further 18 as *loosely associated tautomers*. All of the first set of compounds were tertiary amines and, thus were the required substances for more specific processing.

The next feature to address in order to probe for α_{1B} selective ligands was the requirement to avoid an attachment capable of hydrogen bonding in the indole/dihydroindole phenyl ring. This was achieved by analysing the *Real-atom Attachments* in Answer Set 1, as shown in Fig. 11. An alternative procedure would have been to specify a group unable to form hydrogen bond at this position in the initial structure query (Fig. 5), but it is always better to approach the problem from a more general position first and then to use

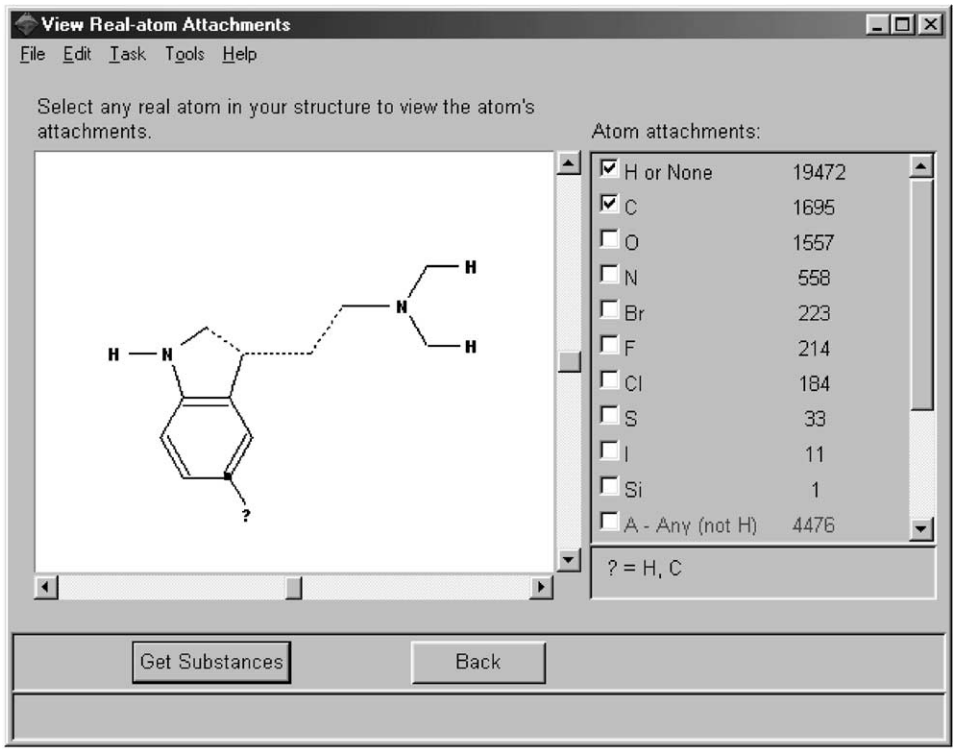


Fig. 11. Analysis of the *Real-atom Attachments* at the position indicated by a question mark. Allowing only carbon or hydrogen atoms at this position refines the original set of compounds retrieved with the query described in Fig. 8 to retrieve Answer Set 2.

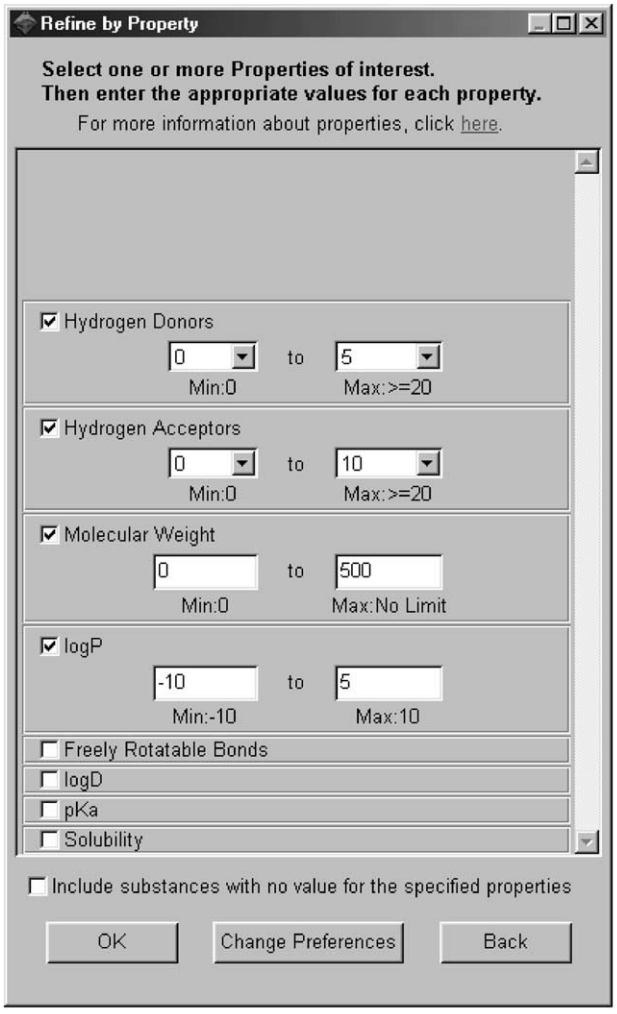
the *Analyse* and *Refine* options once an initial answer set has been obtained. In this case, it was apparent (Fig. 11) that the only non-hydrogen bonding atoms in actual structures at this position were carbon and hydrogen, so these were selected to afford Answer Set 2 (21,167 substances).

As anticipated, the original lead compound IQC was recovered in the first searches (Answer Set 1), but was filtered out after the *Real-atom Attachments* refinement.

The remaining feature required was to avoid an aromatic ring (phenyl ring of the tetrahydroisoquinoline system in IQC) and to avoid further hydrogen bonding attachments in this region, but the only way to effectively do this was to explore actual ring structures in Answer Set 2. When a *Ring skeleton analysis* for Answer Set 2 was conducted, a total of 567 different ring systems were identified and this was probably still too many to analyse manually by mapping on to the three α_1 pharmacophores.

3.3. Further structure refinements

A *Refine by Property* analysis was performed on Answer Set 2, choosing the options indicated in Fig. 12 (in accordance with Lipinski's "Rule of Five" [9]), which allowed the retrieval of 12,338 potentially 'drug-like' compounds (Answer Set 3). A *Ring skeleton analysis* on Answer Set 3 was then performed, and 391 different ring systems were found. The most common ones are shown in Fig. 13. The graphic display of ring structures in SciFinder Scholar® allows visual inspection of all structure types, and so it was a simple matter to work through the 391 ring skeletons with a view to selecting suitable candidates for fitting on to the α_1 pharmacophores. For example, it became clear that they fell into several categories with respect to pharmacophoric groups required for binding to the α_{1B}-adrenoceptor subtype. As discussed previously, an indole or dihydroindole



SciFinder Scholar® screens are reproduced with the permission of the

Chemical Abstracts Service, a Division of the American Chemical Society.

Fig. 12. Refinement of Answer Set 2 by physical properties, according to Lipinski's criteria for 'drug-like' molecules. This resulted in 12,338 answers (Answer Set 3).

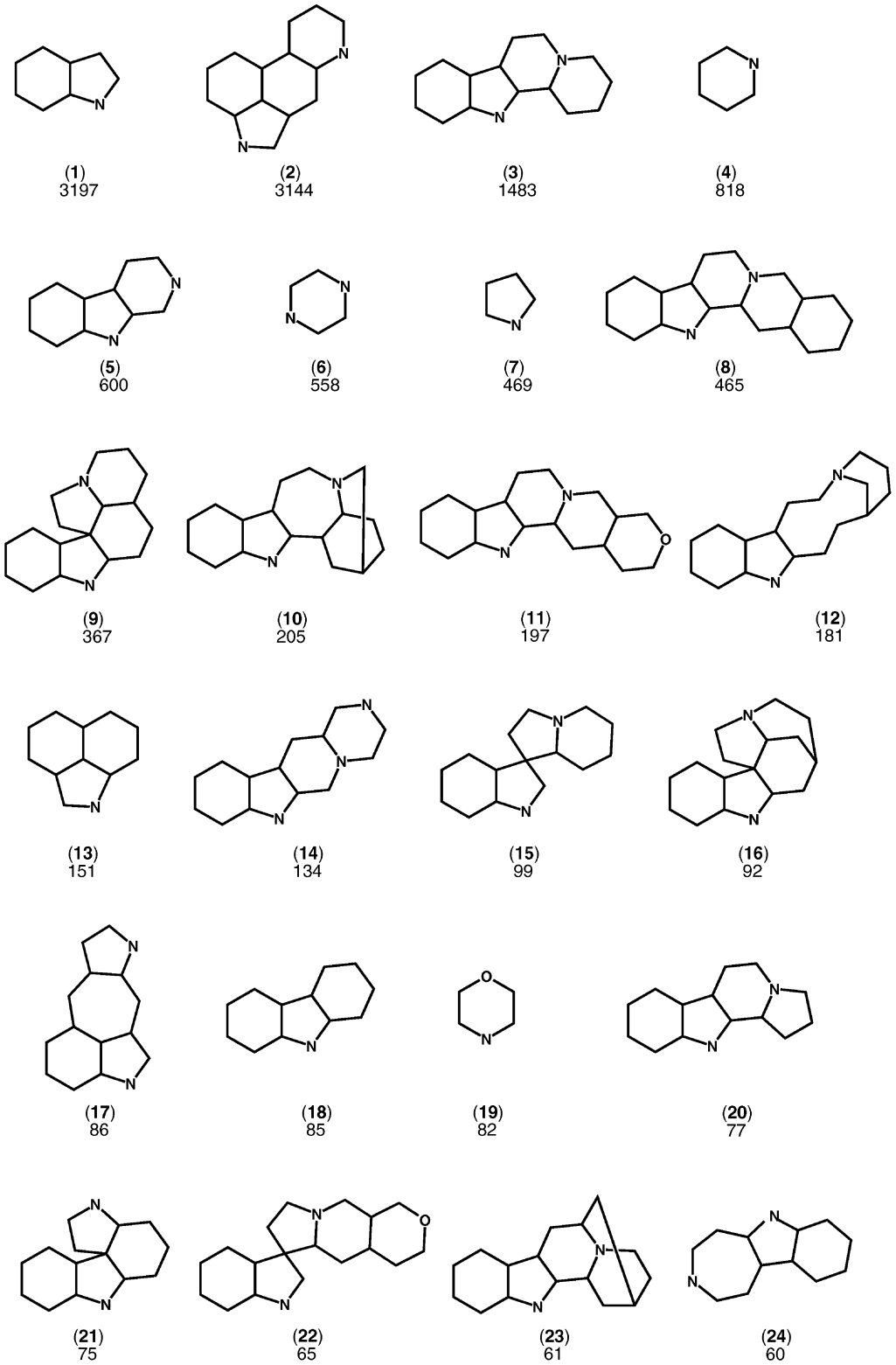


Fig. 13. The most common ring systems found by *Ring skeleton analysis* for Answer Set 3. Entry number in brackets for identification; the second number indicates how many compounds in Answer Set 3 contain the particular ring system.

system, as well as a basic nitrogen atom were required. Ring systems, such as (**1**) in Fig. 13 did not contain all these groups within the heterocyclic skeleton, whereas, (**2**) an ergoline derivative did. In the first instance it was decided to concentrate on ring systems containing all three important features. It was felt that such rigid molecules would be of greater value in elucidating the requirements for selective binding to α_{1B}-adrenoceptor sites. Systems, such as (**11**) in Fig. 13, were redundant, because with respect to the α_{1B} pharmacophore they were not different to other compounds in the list (in this case (**8**)). Some compound pairs, such as (**5** and **3**), were redundant with respect to the three groups mentioned previously, but (**3**) contained an extra aliphatic ring, which could potentially map the hydrophobic group of the α_{1B} pharmacophore (see Fig. 3). In such cases, the simpler parent ring system was investigated first, and only if it fitted the pharmacophore, was the expanded system mapped as well. Pairs, such as (**10** and **24**), were not considered redundant in this way because the methylene bridge in the former was likely to affect the spatial disposition of the pharmacophoric groups.

Selected ring systems from the 391 ring systems were, thus readily identified and, in the first instance, these were mapped on to the pharmacophores. Those fitting the α_{1B} pharmacophore in the same way as the original lead compound IQC are as shown in Fig. 14. These rigid ring systems

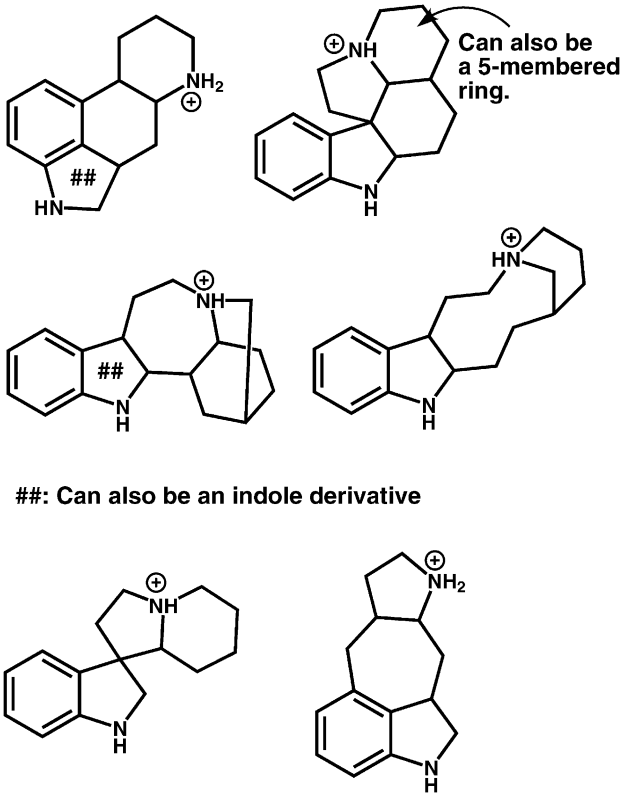


Fig. 14. Selected ring systems, illustrating the variety of potential lead compounds found which fit the features of the α_{1B} pharmacophore in the same way as IQC (see Fig. 3).

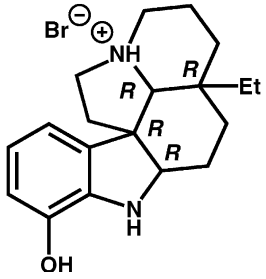


Fig. 15. Aspidospermidin-17-ol mono-hydrobromide, a commercially available lead compound retrieved with the search strategies detailed in the text.

contained three features important for selective antagonist binding to the α_{1B} receptor subtype in the correct 3D arrangement in space and were thus important lead compounds to be considered for further design.

It was also of interest to analyse which of the retrieved compounds are commercially available and can thus be obtained for binding studies very quickly, so that the overall strategy can be tested. Substance lists can easily be checked for commercial availability and the names of suppliers obtained within SciFinder Scholar®. Of the 21,167 compounds of Answer Set 2, a small subset of 192 compounds were commercially available, and these contained 30 different ring systems.

New substructure searches were also performed with the skeletons identified in Fig. 14 as input, and the structures returned were refined by commercial availability. A number of compounds of interest were retrieved, for example, the aspidospermidine derivative depicted in Fig. 15. This compound also fitted the α_{1B} pharmacophore, and it and other commercially available hits from this work will be investigated further. A literature search was also performed using 'aspidospermidin' as a term and this resulted in 231 references which could be further analysed by index term. This revealed that the majority of these references related to the synthesis or structure of this class of compounds. The titles of references relating to biological properties were scanned, but none related to adrenoceptors. Additionally, while it is acknowledged that specific receptor mediation of biological properties observed may not always have been identified, no references were found relating the search terms 'adrenergic receptor' or 'adrenoceptor' to the term 'aspidospermidin'. Therefore, this class of compounds promises to give rise to structurally novel lead compounds for the design of α_{1B} selective adrenoceptor antagonists.

Substructure searches with skeletons, in Fig. 13, and refinement by commercial availability also retrieved compounds with known activity as α_1-adrenergic antagonists, such as yohimbine derivatives, which are based on skeleton (**8**).

4. Conclusions

Using the design of subtype-selective α_1-adrenergic receptor antagonists as an illustrative example, we have

developed a methodology for converting previously determined 3D pharmacophores into 2D queries for molecular substructure searches of the CAS database with SciFinder Scholar®. We have presented the results of such a search and discussed the options for refinement of the list of compounds retrieved. Among the large number of analysis and refinement options available, some are unique to SciFinder Scholar® and are very valuable in this context, such as filtering out commercially available compounds from hit lists (this includes names of suppliers), linking directly into the available literature on any compound or class of compounds, as well as further restricting lists of compounds according to structural features (such as the kind of atom allowed to be attached to a particular position) or physical or biological properties. We have further shown that previously unrecognised and structurally novel potential lead compounds can be located with such a strategy, establishing its usefulness for drug design projects.

Acknowledgements

Support from the NHMRC is gratefully acknowledged, and we thank the Chemical Abstracts Service for permission to reproduce SciFinder Scholar® screens.

References

[1] O.F. Güner (Ed.), Pharmacophore Perception, Development, and Use in Drug Design, IUL Biotechnology Series, International University Line, La Jolla, CA, 2000.
[2] P. Gund, Three-dimensional pharmacophoric pattern searching, Prog. Mol. Subcell Biol. 5 (1977) 117–143.
[3] P.A. Keller, C. Birch, S.P. Leach, D. Tyssen, R. Griffith, Novel pharmacophore based methods reveal gossypol as a reverse transcriptase inhibitor, J. Mol. Graphics Model. (2001) in press.
[4] S. Grüneberg, B. Wendt, G. Klebe, Subnanomolar inhibitors from computer-screening: a model study using human carbonic anhydrase II, Angew. Chem. Int. Ed. 40 (2001) 389–393.
[5] P.A. Greenidge, B. Carlsson, L.-G. Bladh, M. Gillner, Pharmacophores incorporating numerous excluded volumes defined by X-ray crystallographic structure in three-dimensional database searching: application to the thyroid hormone receptor, J. Med. Chem. 41 (1998) 2503–2512.
[6] http://www.cas.org/SCIFINDER/SCHOLAR/.
[7] http://www.acdlabs.com.
[8] http://www.accelrys.com/viewer/index.html.
[9] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, Adv. Drug Delivery Rev. 23 (1997) 3–25.
[10] J.B. Bremner, B. Coban, R. Griffith, K.M. Groenewoud, B.F. Yates, Ligand design for α_1-adrenoceptor subtype selective antagonists, Bioorg. Med. Chem. 8 (2000) 201–214.
[11] R.M. Graham, D.M. Perez, M.T. Piascik, R.P. Riek, J. Hwa, Characterisation of α_1-adrenergic receptor subtypes, Pharmacol. Commun. 6 (1995) 15–22.
[12] S.P.H. Alexander, A. Mathie, J.A. Peters (Eds.), 2001 TiPS Nomenclature Supplement, 12th Edition, Current Trends, London, 2001, p. 15.
[13] D.W. Phife, R.A. Ramos, M. Feng, I. King, S.P. Gunasekera, A. Wright, M. Patel, J.A. Pachter, S.J. Coval, Marine sponge bis(indole) alkaloids that displace ligand binding to α_1-adrenergic receptors, Bioorg. Med. Chem. Lett. 6 (1996) 2103–2106.
[14] A. Beck-Sickinger, Structural characterisation and binding sites of GPCRs, Drug Discovery Today 1 (1996) 502–513.
[15] D.D. Ridley, Information Retrieval: SciFinder and SciFinder Scholar, Wiley, Chichester, 2002, ISSN 0479-84351-9.