The Most Common Chemical Replacements in Drug-Like Compounds

Robert P. Sheridan†

Department of Molecular Systems, RY50SW-100 Merck Research Laboratories, Rahway, New Jersey 07065

Received August 10, 2001

We have written a method that extracts one-to-one replacements of chemical groups in pairs of drug-like molecules with the same biological activity and counts the frequency of the replacements in a large collection of such molecules. There are two variations on the method that differ in their treatment of replacements in rings. This method is one possible approach to systematically identify candidate bioisosteres. Here we look at the MDDR database because it has a large diversity of drug-like compounds in a large number of therapeutic areas. The most frequent replacements in MDDR seem generally consistent with medicinal chemistry intuition about what chemical groups are equivalent or with groups that are easily converted by synthetic or metabolic pathways. This method can be applied to any set of molecules wherein the molecules can be paired by similar biological activity.

INTRODUCTION

Bioisosterism is the concept that a chemical group in a biologically active compound can be replaced with another such that the new molecule retains the biological activity. The presumption is that the groups to be substituted are similar in some important physical property. For example, a phenyl ring and a thiophene ring are about the same size and both are hydrophobic, carboxylate and tetrazole are both anions at physiological pH, etc. The reader is referred to reviews in this area, e.g. ref 1. Many papers have been written wherein group X2 is substituted for group X1 in molecule M1 to make molecule M2. If M1 and M2 have similar biological activities, the claim is often made that X1 and X2 are bioisosteres. (References 2−4 are recent examples.) This may not be generally true for the following reasons:

1. The substituent X1/X2 may be in an unimportant part of the molecule, i.e., a part that does not make a critical interaction with the receptor.

2. If X1 and X2 are relatively small, M1 and M2 are very similar molecules, and it is not surprising that similar molecules will have similar biological activities.

3. All that can be inferred is that X1 might be equivalent to X2 at that one position and only for that bioactivity.

A strict definition of bioisosterism might require that X1 and X2 be equivalent in a number of properties (hydrophobicity, size, charge, etc.). However, a more liberal definition may be groups that can be substituted for each other in a variety of chemical classes for a variety of bioactivities. Our aim here is to gather statistics on how often groups are substituted for others in drug-like molecules. Replacements that occur often may be worth considering in lead development projects.

METHODS

The overall scheme for counting replacements is shown in Chart 1. Details are provided below.

Chart 1

```
For each type of biological activity {
    Cluster molecules with that biological activity by overall topological similarity.
    For each cluster{
        For each pair of molecules M1 and M2 in that cluster{
            Extract the parts of M1 and M2 that are different and
            that correspond to a one-to-one replacement. Put them
            in a database of fragment-pairs specific for that activity.
        }
    }
}
Count the occurrence of identical fragment-pairs over all biological activities.
```

Clustering. Examining all pairs of molecules with a given biological activity is impractical because the extraction of replaced groups is computationally expensive. We are interested in pairs of molecules that differ only in one place. Therefore we clustered compounds using a method described previously,[5] which uses topological descriptors to calculate overall simlarity of two molecules. Only pairs of compounds within a cluster are examined. For this work we used the regular atom pair descriptor and a cosine similarity cutoff of 0.9.

Extraction of Replaced Parts. This has several steps, which are illustrated with an example in Figure 1.

1. Identify the corresponding atoms of molecules M1 and M2. These are the "match atoms".

2. Label the corresponding atoms with corresponding names.

3. Remove the bonds between match atoms.

4. Delete atoms with no bonds. What is left is called a "fragment-pair".

5. Filter the fragment-pairs.

By this algorithm, the fragments in each fragment-pair have at least one match atom that is the attachment point to the common parts of the molecules.

† Corresponding author phone: (732)594-3859; fax: (732)594-4224; e-mail: sheridan@merck.com.

10.1021/ci0100806 CCC: $22.00
Published on Web 01/03/2002

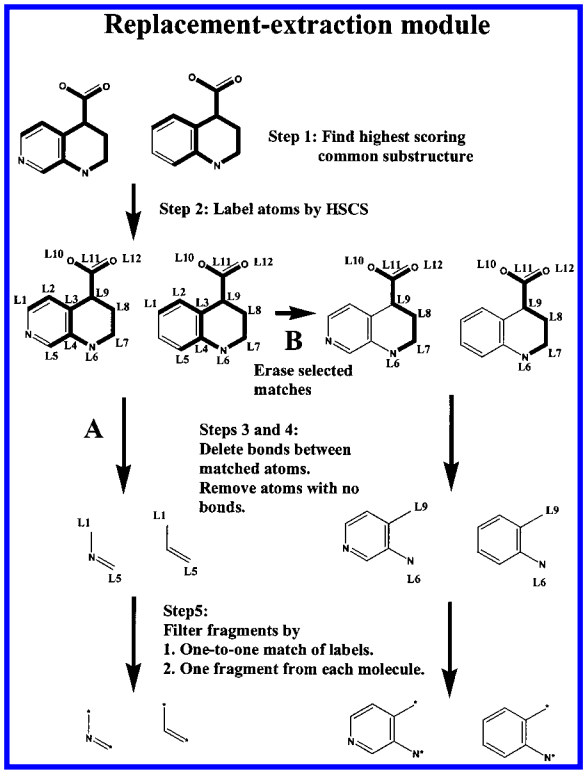

Figure 1. A schematic diagram of how replaced groups are parsed as a fragment-pair from a pair of molecules. There are two alternative paths (marked "A" and "B") depending on whether match atoms in rings or other groups are declared unmatched after Step 2 (see text). Steps 3−5 are the same thereafter. Bonds connecting match atoms are shown as bold. At the bottom of the figure, the match atoms remaining after processing are indicated by "*".

For step 1 we use the maximum common substructure detection method in Sheridan and Miller.[6] This method, based on clique detection, can generate substructures that are discontinuous. A clique-defined substructure is a set of pairs of atoms, one from M1 and one from M2, such that the paired atoms are of the same "atom type" and the topological distances (in bonds) between the atoms in M1 are the same as the corresponding distances between the atoms in M2. The score of a common substructure equals the number of atoms in the common substructure minus a "discontinuity penalty" ($p = 1$) that penalizes having discontinuous fragments in the substructure. We keep only the highest scoring common substructure (HSCS) for molecules M1 and M2. In the original work we typed atoms based on element, hybridization, and "physiochemical type". Here we use only element and hybridization. In the original work, we filtered out HSCSs that were not significantly larger than expected for two randomly selected molecules of the same size. However, for the current application, where size is irrelevant, we keep all HSCSs.

For step 5, we make the following requirements to ensure that there is a one-to-one replacement of one group with another:

a. The labels of the matched atoms have to be in one-to-one correspondence.

b. There is exactly one fragment extracted from M1 and one from M2.

Figure 2 illustrates what kinds of pairs of molecules would produce fragment-pairs that would pass the filter.

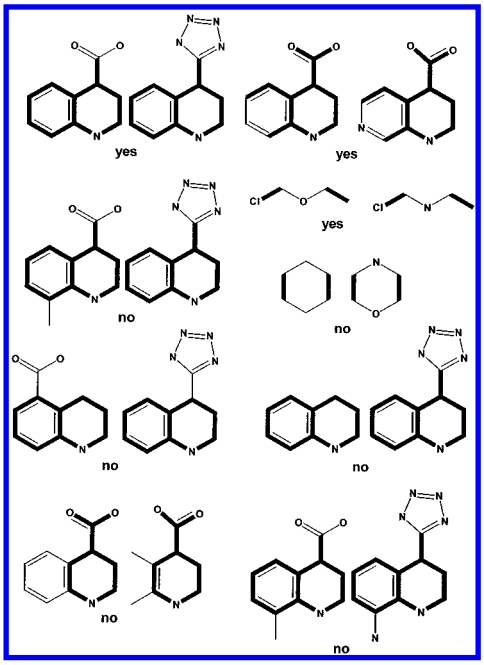

Figure 2. Examples of pairs of molecules where the extracted fragment-pair would pass (yes) or fail (no) the filters for algorithm A. The filters are meant to detect one-to-one replacements of a single chemical group.

We will call the algorithm described above "algorithm A". One optional modification is for certain atom matches to be erased after step 2. If there is at least one match atom in a 3,4,5,6,7, or 8-membered ring, all atoms in the ring are declared unmatched. Also, if there is a unmatched atom adjacent to a NO, CO, SO_2, or PO_2 in which the atoms are matched, the atoms in the NO, CO, SO_2, or PO_2 are declared unmatched as well. Steps 3−5 are followed as before. We will call this alternative "algorithm B". The difference in results is also illustrated in Figure 1. Algorithm B will be justified later on the basis of keeping more of the context of the replaced groups.

Counting the Occurrence of Fragment-Pairs. A "hash string" is calculated from connection table of each fragment-pair using a method similar to that of Burden.[7] A modified adjacency matrix **Q** is constructed such that the diagonal elements for atom i are made of the sum

$$\mathbf{Q}_{ii} = \text{atomic number} + 0.1^{*}\text{ number of non-hydrogen neighbors} + 0.01^{*}\text{ number of } \pi \text{ electrons} + 0.001^{*}\text{match state}$$

where the match state is 1 if the atom is a match atom and 0 otherwise.

The off diagonal elements are

$$\mathbf{Q}_{ij} = 0.4/\text{ topological distance between i and j}$$

If i and j are from different fragments, the distance is set to an arbitrary high number.

The hash string is a concatenation of the highest and lowest eigenvalues of **Q** expressed to six decimal places. (It is usually necessary to calculate the eigenvalues in double precision.) The hash string depends only on the atoms and the bonds between them; the order of the atoms and the order of the fragments in the pair is irrelevant. The inclusion of match information in the hash string helps us distinguish

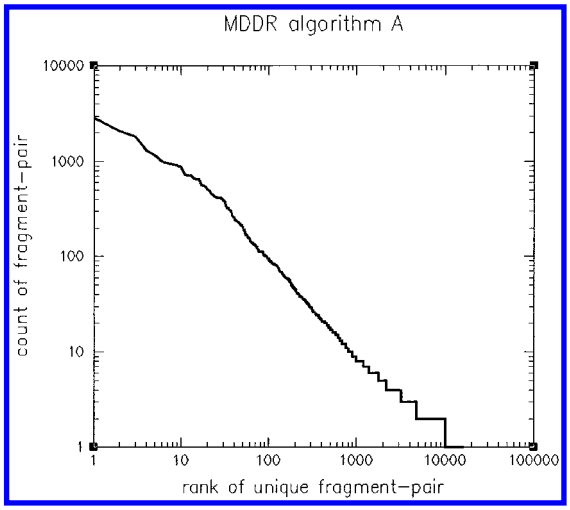

Figure 3. The count of a unique fragment-pair vs the rank in decreasing count.

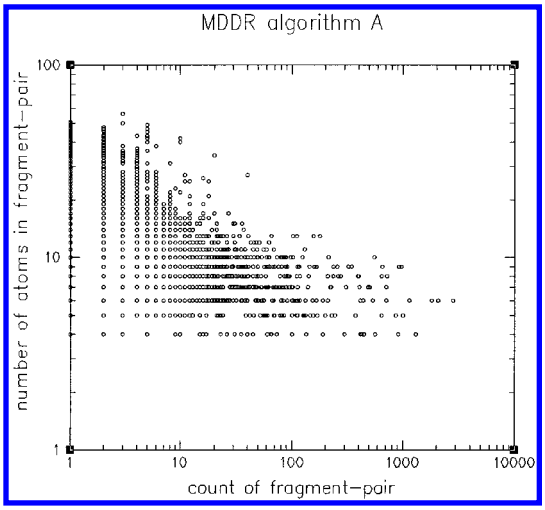

Figure 4. The number of non-hydrogen atoms in a unique fragment-pair vs its count.

fragments that have a match atom (*) at one vs two sides (e.g. C*−O−C* in ether vs C*−O−C in methoxy).

Counting unique fragment-pairs becomes a matter of counting the frequency of a unique hash strings over all biological activities. A counting method where the fragment-pairs are weighted as the inverse of the number of fragment-pairs in that activity did not produce significantly different results, at least for the most frequent fragment-pairs.

Source of Molecules. One source of drug-like compounds is the MDL Drug Data Report (MDDR),[8] a licensed database compiled from the patent literature. A small percentage of molecules in the MDDR are very large (e.g. peptides) and some are very small. Because we want to consider drug-sized molecules, we kept only those molecules within the range of 7−50 non-hydrogen atoms. (This is expedient as well because molecules with > 50 atoms tend to slow calculations of maximum common substructure.) Molecules in the MDDR are assigned a "therapeutic category" by the vendor. For the purposes of executing Chart 1, we will assume that molecules in the same therapeutic category have the same biological activity. Some therapeutic categories (e.g. "antihypertensive") contain molecules that work by different mechanisms, but this is not a problem here because we are looking only at pairs of very similar molecules, and these almost certainly work by the same mechanism.

There are 647 therapeutic categories. A molecule may be in more than one therapeutic category, and some therapeutic categories are nearly synonymous, but we did not make any special compensations for this.

RESULTS

In the MDDR there were 98 445 unique molecules in the size range and 556 therapeutic categories that had at least one pair of similar molecules in that range. A total of 527 985 pairs of molecules were compared. Algorithm A kept 90 095 fragment-pairs, of which 16 536 were unique. Figure 3 shows the distribution of the count of each unique fragment-pair as a function of its rank. This is a log−log relationship, i.e., the counts fall very quickly with rank. Figure 4 shows the size of the replaced groups, measured in the number of non-hydrogen atoms in each unique fragment pair, as a function of the count. The fragment-pairs with the highest counts are all small, the smallest possible being 4 (one replaced atom + one attachment point times two molecules). On the other hand, the rare fragment-pairs can be small or large.

The top 10 fragment-pairs, plus some other interesting ones are shown in Figure 5. By "drilling down" to molecule pairs that contribute to the fragment-pair count, one can get an interpretation of the context of the replacement. Many of these seem to be "classical" replacements in medicinal chemistry. The most common replacement (labeled A1 in Figure 5) is the replacement of C with N in an aromatic ring. This can occur in phenyl ⇔ pyridine, pyridine ⇔ pyrazine, pyridine ⇔ pyrimidine, etc. The next most common (A2) is −O− ⇔ −S−. This can occur in aliphatic chains, aliphatic rings, and aromatic rings. Similarly, replacement of −N− for −O− can happen not only in chains, aliphatic rings, and aromatic rings, but also in amides ⇔ esters. A4 and A5 occur only in aromatic rings. A change from a six- to a five-membered aliphatic ring is A7. The replacement phenyl ⇔ thiophene is A8, carbonyl ⇔ thiocarbonyl (in ureas and amides) is A26, amide ⇔ sulfonamide is A33, phenyl ⇔ furan is A40, and phenyl ⇔ pyrrole is A76. Some replacements (e.g. A30, A90, A99) appear to be moving heteroatoms around a ring. A95 is the reversal of an amide bond. We were surprised to see that replacements such as A11, A15, and A18 occur with a high frequency, but in retrospect they may not be so surprising. For instance, the reduction of a ketone to an alcohol (A15) is a common metabolic transformation.

We need to go fairly far down the list to see replacement of charged groups; for instance, A115 and A132 represent anionic replacements. We expected to see carboxylate ⇔ tetrazole as an anionic substitution, but a one-to-one replacement is surprisingly rare; it occurs only 10 times in the MDDR, giving it a rank of A881. Replacements of cations also occur far down the list. For instance, −CH2-guanidine ⇔ −CH2−NH3 (as in the amino acids Arg and Lys) occurs 21 times, making it A429. Guanidine ⇔ amidine is A873.

The fragment pairs in Figure 5 are generally small and do not contain much information about the context of the replacement, especially with regard to rings. That is why we used an optional step that involved "unmatching" ring atoms and ester, thioesters, phosphonates, etc. Using the option in algorithm B, there were 116 060 total fragment-

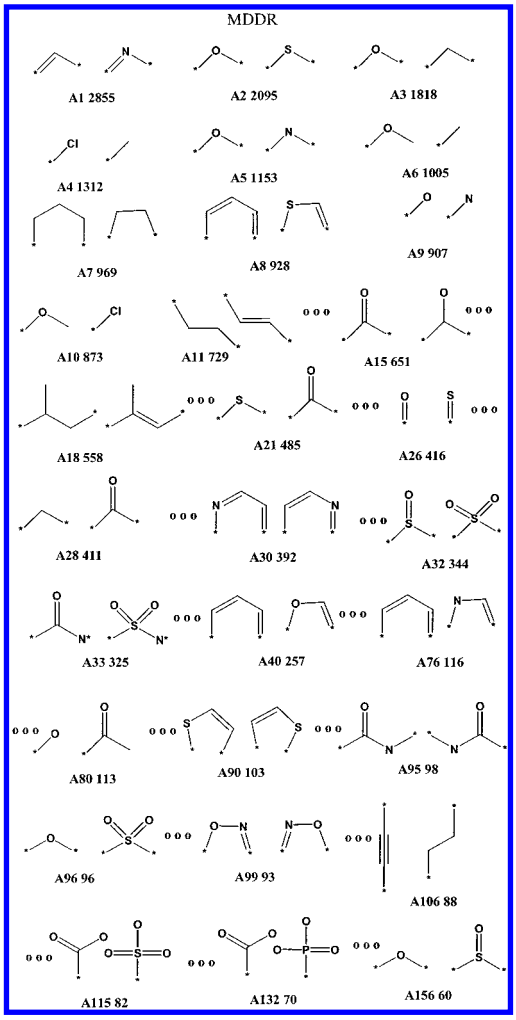

Figure 5. Selected unique fragment-pairs from the MDDR algorithm A. For each, the rank is given and its count, e.g. A1, is the most frequent fragment-pair with a count of 2855. A "*" indicates a match atom (the connection point to the conserved parts of the molecules). The order of the two fragments in each pair is arbitrary since the replacements are treated as symmetrical. The symbol "ooo" indicates one or more pairs were skipped.

pairs of which 18 275 were unique. Since more of the context is preserved, including substituents on the rings, we expected that there would be more unique fragment-pairs, but there are more total fragment-pairs as well. This is due to the fact that in algorithm A changes in two places in a single ring, e.g. cyclohexane ⇔ morpholine, are rejected by the filter, whereas algorithm B would treat cyclohexane ⇔ morpholine as a single change. The statistics for algorithm B look qualitatively very much like that of algorithm A shown in Figures 3 and 4, except that the counts are somewhat smaller, and the number of atoms in an average fragment-pair are somewhat larger, as expected.

The most frequent fragment-pairs plus others are shown in Figure 6. Many replacements are the same as with the algorithm A, but the ranks have changed. We can now distinguish between acyclic and ring replacements. The most common are now small acyclic replacements. Phenyl ⇔ benzyl (B9) is the most common replacement involving a ring. The next most common ring replacement is phenyl ⇔ thiophene (B11). This time, since the context is retained, we see that the more common replacement is 2-thiophene; 3-thiophene does not show up until B75. Similarly, we see

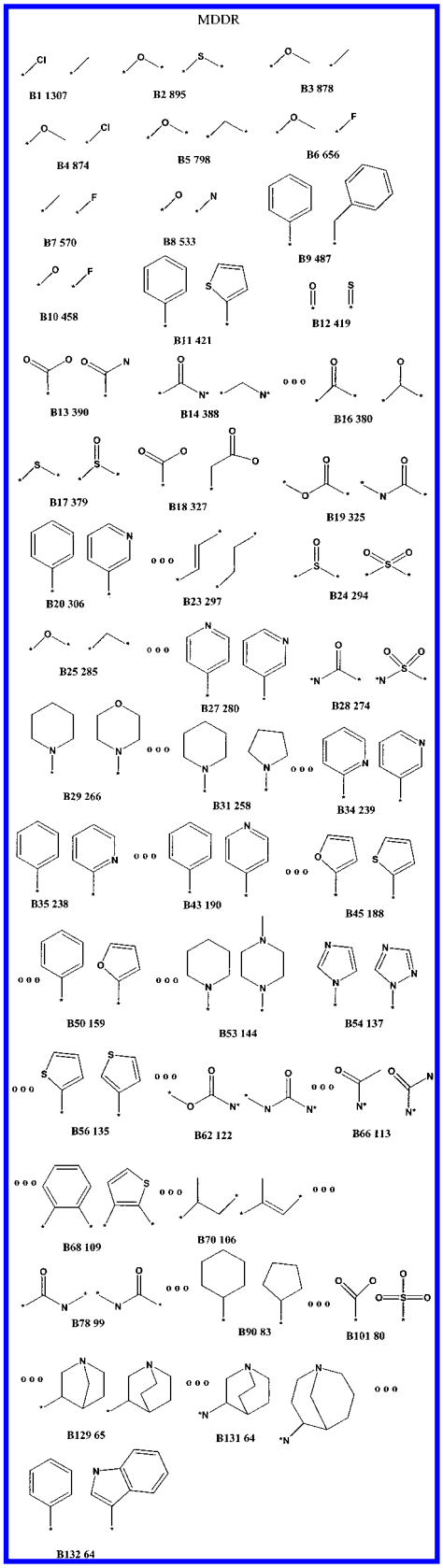

Figure 6. Selected unique fragment-pairs from the MDDR algorithm B.

the most common phenyl ⇔ pyridine replacement is 3-pyridine (B20). We can now see the explicit replacement of amide ⇔ ester (B19) and urea ⇔ carbamate (B62) instead of the more generic −N− ⇔ −O−.

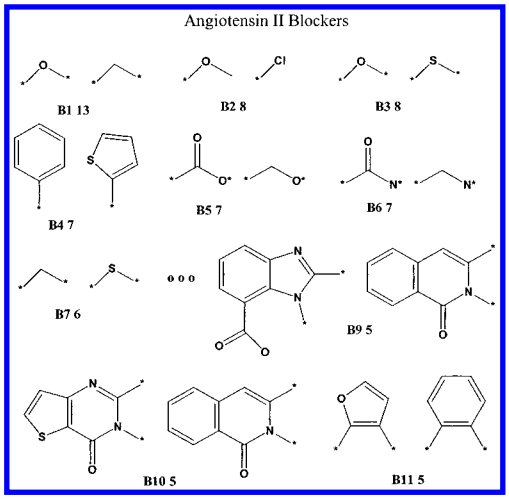

Figure 7. Selected unique fragment-pairs from only the molecules with Angiotensin II Blocker activity, algorithm B.

Besides looking at MDDR as a whole, one can look at a subset of activities, or a single activity, with the caveat that the counts are much smaller and some of the replacements are more likely to be associated with a particular position around a "core". For instance, Figure 7 shows the algorithm B results for the activity "Angiotensin II Blockers". There are 1304 compounds with 7−50 non-hydrogen atoms, 3468 replacements, of which 3132 are unique. Many of the more frequent replacements (B1−B5) are generic in the sense that they are also frequent for the MDDR as a whole. Further down the list (e.g. B9 and B10) we see activity-specific replacements. This is true for many of the individual activities we examined.

DISCUSSION

We have presented a method of identifying frequently replaced groups in drug-like molecules. Of the two variations, algorithm B appears to be more appealing because it retains more of the contextual information of the replacement. This method can be applied to any set of molecules wherein the molecules can be paired by similar biological activities.

There are certain limits to our MCS algorithm in determining equivalent groups, and many further refinements are possible. First, clique-based methods such as ours depend on having matched topological distances. Thus, molecules of the form R1-X1-R2 and R1-X2-Y2-R2 could not be matched at R1 and R2 because they are different distances apart in the two molecules, so the -X1- ⇔ -X2-Y2- replacement would not be detected. Second, the hash method of identifying unique fragment-pairs is very efficient, but it is rather stringent. A single atom change is enough to distinguish fragment-pairs (e.g. phenyl ⇔ 2-pyridine is different from phenyl ⇔ 2-pyrimidine), but it is not clear whether that level of discrimination is always useful. For example, many chemists would not consider phenyl ⇔ 2-pyridine conceptually different from phenyl ⇔ 3-methyl-2-pyrimidine. A clustering based on similarity rather than strict identity could consolidate many low-frequency fragment-pairs. Finally, more fragment pairs could be generated if we removed the constraint that a single group be replaced by another. Allowing double substitutions (e.g. the pair in the lower right of Figure 2) might provide additional insight.

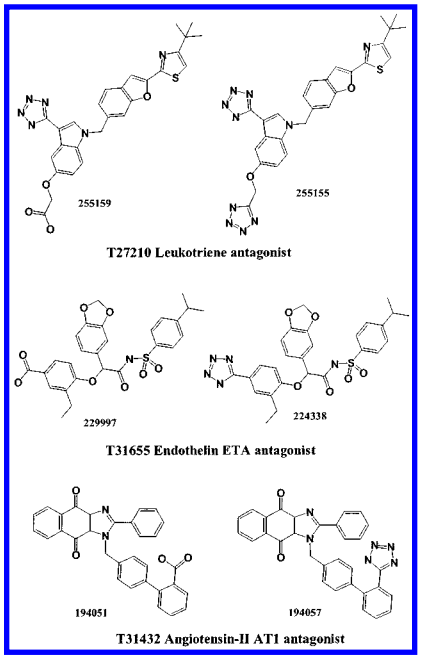

Figure 8. Examples from the MDDR where a carboxylate ⇔ tetrazole replacement is made in different therapeutic categories.

It should be emphasized that having a list of replacements and their frequencies is not the same as having a list of definitive bioisosteres. Because groups are often substituted does not necessarily mean the groups are physiologically equivalent. For instance, the ketone ⇔ alcohol replacement may be so common because of the ease of transformation. Also, since our approach is retrospective, it can highlight only those replacements that have already been made many times in some collection of molecules. That is, chemical groups that might be equivalent to a receptor, but do not appear in a one-to-one substitution in the database, cannot be detected.

That said, we note that we have probably captured at least some flavor of bioisosterism from the MDDR. The most frequently made replacements in MDDR correspond to widely known replacements in medicinal chemistry,[1] and generally the replacements make physical sense in terms of size and bond angle, although the replacements are less often equivalent in hydrogen bonding character. Conversely, at least some groups that are known to be equivalent in some property, e.g. carboxylate and tetrazole, are in the list, although they are perhaps not as frequently substituted in the MDDR as might be expected. It could be argued that, since the MDDR is derived from patent literature, we may be seeing the reflection of already established medicinal chemistry intuition about what groups are equivalent. Thus it is not at all surprising that our results would be consistent with such intuition. Even if this is true, however, it is useful to be able to systematize such intuition by an automatic method.

One goal here is to systematically detect and organize replacements as a resource to be mined for synthetic ideas. The most frequent replacements in MDDR are not necessarily the most interesting, because they are mostly small and already well-established. More insight might be derived from replacements that are relatively infrequent but show up enough times (say 10 or more) and in enough different

therapeutic areas that one can have confidence that they are "real". For instance, the carboxylate ⇔ tetrazole replacement is infrequent, but it does occur in unrelated molecules in a least three different therapeutic areas. Some examples are shown in Figure 8.

Although the MDDR is a very valuable database in that it contains many diverse molecules in many different therapeutic areas, making it nearly ideal for the work presented here, it has the limit that the activity data may not always be reliable. For instance, compounds may be claimed in a patent to have a specific activity, but the activity may not be quantitative, and thus not always comparable to compounds with a similar claimed activity from another laboratory. Also, whether the activity is in vivo or in vitro is not always consistent. A very useful approach might be to use our method extract frequent replacements from smaller databases with more consistent measures of activity. For instance, given sets of IC_{50} data, one could pair only those molecules from the same binding assay that have IC_{50}'s within a factor of 10. The frequent replacements from those pairs would more closely reflect bioisosterism for those specific activities.

ACKNOWLEDGMENT

The authors thank Dr. Scott Berk, Dr. James Doherty, and Dr. Arthur Patchett for useful comments. The tools for this work were written in MIX, Merck's in-house modeling system, and the author thanks the other members of the MIX team.

REFERENCES AND NOTES

(1) Wermuth, C. G. Molecular variations based on isosteric replacements. In *The Practice of Medicinal Chemistry*; Wermuth, C. G., Eds.; 1996; pp 202−237.
(2) van Vliet, L. A.; Rodenhuis, N.; Dijkstra, D.; Wikstrom, H.; Pugsley, T. A.; Serpa, K. A.; Meltzer, L. T.; Heffner, T. G.; Wise, L. D.; Lajiness, M. E.; Huff, R. M.; Svensson, K.; Sundell, S.; Lundmark, M. Synthesis and pharmacological evaluation of thiopyran analogues of the dopamine D-3 receptor selective agonist. *J. Med. Chem.* **2000**, *43*, 2871−2882.
(3) Balsamo, A.; Macchia, M.; Martinelli, A.; Rossello, A. The [(methyloxy)imino]methyl moiety (MOIMM) in the design of a new type of beta-adrenergic blocking agent. *Eur. J. Med. Chem.* **1999**, *34*, 283−291.
(4) Mederski, W. W. K. R.; Osswald, M.; Dorsh, D.; Anzali, S.; Christadler, M.; Schmiteges, C.-J.; Wilm, C. Endothelin antagonists: evaluation of 2,1,3-benzothiadiazole as a methylendioxyphenyl bioisostere. *Bioorg. Med. Chem. Lett.* **1998**, *8*, 17−22.
(5) Sheridan, R. P. The centroid approximation for mixtures: calculating similarity and deriving structure−activity relationships. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1456−1469.
(6) Sheridan, R. P.; Miller, M. D. A method for visualizing recurrent topological substructures in sets of active molecules. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 915−924.
(7) Burden, F. R. A chemically intuitive molecular index based on the eigenvalues of a modified adjacency matrix. *QSAR* **1997**, *16*, 309−314.
(8) Molecular Design Drug Data Report, version 99.1 distributed by Molecular Design Ltd.: San Leandro, CA.

CI0100806