

The Quest for Bioisosteric Replacements

Markus Wagener* and Jos P. M. Lommerse

NV Organon, Department of Molecular Design and Informatics, P.O. Box 20, 5340BH Oss, The Netherlands

Received September 12, 2005

To help advance drug discovery projects, a new and validated search method is presented by which potential bioisosteric replacements can be retrieved from a database of more than 700 000 structural fragments. The heart of the search method is an optimized topological pharmacophore fingerprint which describes each fragment as a combination of attachment points, hydrogen bond donors and acceptors, hydrophobic centers, conjugated atoms, and non-hydrogen atoms. In the fingerprint the influence of the attachment point is enhanced by giving it extra weight relative to the other descriptors. The Euclidean distance has proven to be the optimum distance measure to compare the fingerprints in a database search. The performance of the pharmacophore fingerprint based search method has been validated using more than 2200 bioisosteric fragment pairs extracted in an unbiased procedure from the BIOSTER database. The true bioisosteric pairs have been compared with pairs of random fragments originating from the WDI database. Normalized by the standard deviation of the random pairs distance distributions, an excellent separation of true pairs from random pairs was obtained for R-group fragments (2.2 standard deviation units) as well as for linkers (2.6 units) and cores (2.6 units). The bioisoster search method has been implemented as an intranet application called IBIS and is now routinely used by Organon researchers.

INTRODUCTION

In pharmaceutical research, it is a major challenge to convert a compound resulting from lead finding activities into a successful drug. Whereas the lead compound may already bind with high affinity to the biological target, it will usually have some undesirable characteristics regarding oral bioavailability, metabolic stability, selectivity, and/or toxicity. One strategy commonly used to address these issues is based on the concept of bioisosterism:¹ Structurally related compounds that both elicit the same biological activity are considered as bioisosters. A bioisosteric replacement transforms an active compound into another compound by exchanging a group of atoms with another, broadly similar group of atoms. The resulting new compound still has the original biological activity, and it is aimed for the improvement of undesirable characteristics present in the original compound.

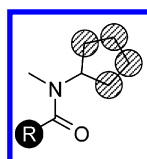
When planning a bioisosteric replacement, the two groups under consideration should be compared with respect to all their relevant properties such as size, shape, electronic properties, lipophilicity, pharmacophore characteristics, solubility, chemical reactivity, etc. Since it will be impossible for any two groups to match all these parameters, carefully balancing and weighing the various properties is required allowing for a replacement to be successful. Many reviews on bioisosterism have appeared in the literature,^{2–5} and a database of reported bioisosteric structures (BIOSTER) is available.^{6,7} These literature examples provide a starting point when searching for the replacement of a specific group. To speed up and increase the comprehensiveness of such a search, a computer-based search method is needed which is optimized for the identification of potential bioisosteric replacements.

Several computational methods have been described to automatically identify bioisosteric replacements. The methods differ in the way similarities between potentially bioisosteric groups are determined and how these methods have been validated. Similarity calculations can be based on propensity maps derived from small molecule crystal structures,⁸ on scores achieved when docking to a reference panel of proteins,⁹ on classical substituent properties (e.g. π - and σ -constants),¹⁰ or on R-group property vectors.¹¹ Weber et al. have validated their approach by deriving predictive QSAR models for a number of peptide data sets taken from the literature.⁹ Watson et al. and Holliday et al. both have used examples of bioisosteric groups manually extracted from the BIOSTER database to show the relevance of their approach.^{8,11}

We report a method for the retrieval of potential bioisosteric replacements using topological pharmacophore fingerprints. Based on this descriptor a similarity search in a database of more than 700 000 molecular fragments allows for the retrieval of the most promising potential bioisosteric replacements for a given fragment. To aim at an improved ADME/Tox profile, a number of search constraints (e.g. lipophilicity, flexibility, acidity/basicity) can be imposed during the search. The method has been implemented as IBIS (Intranet BioIsoster Search) at Organon.

For validation of our method we follow the strategy which has already been reported,^{8,11} viz. by the comparison of bioisosteric replacements collected from the BIOSTER database with random replacements. However, whereas previously only limited sets of bioisosteric fragments were manually extracted from the database, we have devised an unbiased method to fully automate the extraction process. This approach has greatly increased the number of bioisosteric fragments available for validation and thus allows for a more rigorous validation.

* Corresponding author e-mail: markus.wagener@organon.com.



atom pair	frequency
X4H	2
X5H	2
H1H	3
H2H	3

Figure 1. Building a topological pharmacophore fingerprint for a fragment with one attachment point X (filled circle with R) and four hydrophobe centers H (gray circles). All four atom pairs with the number of occurrences are listed.

METHODS

Topological Pharmacophore Fingerprints. A descriptor that can successfully be used to search for bioisosteric fragments should account for factors that are reported to influence the proper choice of suitable bioisosteric replacements:¹ size, shape, electronic properties, lipophilicity, pharmacophore content, solubility, chemical reactivity, etc. Although all these parameters need to be matching to some extent, we believe that it is most important to ascertain that a replacement can interact with the biological target in a fashion similar to the original fragment. This can be achieved efficiently with a kind of pharmacophore derived descriptor. Since the objective was to set up an interactive bioisoster search method, 3D methods were ruled out in order to focus on faster 2D methods. Therefore, it was decided to investigate various types of topological pharmacophore fingerprints.

The pharmacophore descriptors applied are closely related to the atom pair descriptors first introduced by Carhart et al.¹² Atom pairs are substructures of the form *<atom description> – <distance> – <atom description>* where *distance* is the length of the shortest path along bonds connecting the two atoms, and *atom description* is a suitable description of the atoms. For the atom descriptions we assigned any of the following pharmacophore properties: attachment point (X), hydrogen-bond acceptor (A), hydrogen-bond donor (D), hydrophobe (H), conjugated atom (C), aromatic atom (M), positively charged atom (P), and non-hydrogen atom (V). A topological fingerprint is subsequently generated by enumerating all pairs of atoms and their pharmacophore characteristics present in a molecular structure. For each combination found for two pharmacophore properties linked by a given distance, one or more bits are set at unique positions in the fingerprint. Although the descriptor is really based on pairs of pharmacophore entities separated by the topological distance, in the remainder of the text we will refer to the expression *atom pair* for the sake of simplicity.

The generation of a fingerprint is illustrated in Figure 1: If *n* defined pharmacophore properties are used and topological distances between 1 and *m* are taken into account, a fingerprint of length $mn(n + 1)/2$ will be generated. If the amide fragment shown is described just by the properties X (filled circle) and H (gray circles) and the topological distances allowed are between 1 and 16 (*n* = 2, *m* = 16), then the total fingerprint length is 48. Only four bits will be set, which correspond to the pairs X4H, X5H, H1H, and H2H.

Starting with this basic description of a topological pharmacophore fingerprint, a number of variations can be made: (a) the kind of pharmacophore properties used, (b) the relative weight of the pharmacophore properties used, (c) whether the frequency of occurrence of an atom pair is

structure	atom pair	simple fingerprint	fuzzy fingerprint
	X3A	...01000...	...011100000...
	X4A	...00100...	...000111000...
	X5A	...00010...	...000001110...

Figure 2. Three aldehydes each described by one atom pair only. In the simple fingerprints no common bits have been set. After fuzzification the first and third aldehyde still do not have a common bit set, but the second fragment does have one bit in common with both the first and third one.

taken into account, and (d) the way in which the topological distances are represented by one or more bits. All these variations were investigated using the same binary fingerprint representation so that no special versions of software were needed to process the different fingerprints. Details of the fingerprint calculation are explained in the following sections.

The presence of a pair of pharmacophore properties at a given distance is represented by setting a single bit in the bit string (fingerprint). Increasing the weight of one or more pharmacophore properties is established by increasing the number of bits set by the involved atom pair. If the pharmacophore properties X and H of the amide fragment in Figure 1 have equal weights, then four bits will be set. Now if the property X gets a relative weight of two, all pairs involving X will set two bits instead of one, and thus the total length of the bit string is increased to 80 with six bits set: two for X4H and X5H each and one for H1H and H2H each.

The frequency of occurrence of identical atom pairs is also represented in binary form, with extra bits set for multiple occurrence of an atom pair. This way of coding is required in order to maintain a sensible definition of similarity/distance between fingerprints, because most measures for similarity/distance between fingerprints use in some way the number of bits set in both fingerprints (*see below*). Different fingerprints can be generated by using an increasing number of bits set per atom pair to cover frequency information. In all cases, if *n* bits are used to code for frequency information, a set *n*th bit always means that the atom pair occurs at least *n* times. In Figure 1 there are four atom pairs present in the amide fragment shown: the pairs X4H and X5H occurring twice each and the pairs H1H and H2H occurring three times each. For example, when accounting for a maximum atom pair frequency of 2 using the pharmacophore properties X and H only, the length of the fingerprint increases from 48 to 96 with eight bits set, two for each of the atom pairs present.

Two structures with the same pharmacophore features, but with topological distances differing by just one bond, will end up with bit strings without a single common bit set. This would result in a high dissimilarity, even though the compounds might be considered rather similar. This can be illustrated with the three simple aldehydes shown in Figure 2. If only attachment points (X) and hydrogen-bond acceptors (A) are taken into account, each of the three structures contains just one atom pair with each of them involving the same pharmacophore properties but separated at different distances. The corresponding fingerprints as introduced above do not have common bits for any of the three pairs of structures, indicating that they are all equally dissimilar,

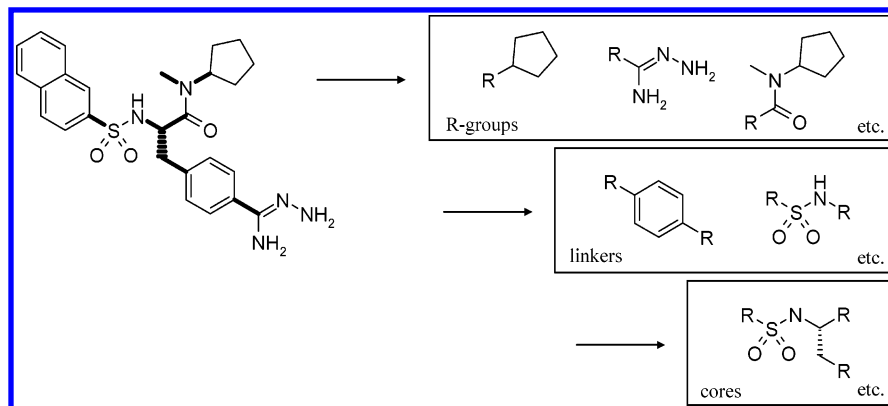


Figure 3. Fragmentation of a thrombin inhibitor. All cleavable bonds are highlighted in bold, and some examples of obtained R-groups, linkers, and cores are shown.

although the first and the second structure can be considered more similar to each other than the first and the third one.

This problem can be circumvented by fuzzification of the corresponding fingerprints: Each atom pair sets three bits instead of one, and the bits are assigned in a way so that atom pairs with distances d and $d+1$ share one common bit position. As a consequence, the corresponding fingerprints of the first and of the second structure in Figure 2 have one bit in common, whereas the first and third structure do not have any common bit set indicating their higher dissimilarity.

Distance Calculation. Several methods to compare fingerprints have been described in the literature. Holliday et al. have given an overview of 22 different similarity coefficients.¹³ In our study we focus on using the Euclidean and Soergel distances, which are probably the most widely used distance measures.¹⁴ In both cases the distance between two fingerprints A and B can be described using the same variables

$$D_{\text{Euclidean}} = (a + b - 2c)^{1/2}$$

$$D_{\text{Soergel}} = 1.0 - c/(a + b - c)$$

where a is the number of bits set in fingerprint A, b is the number of bits set in fingerprint B, and c is the number of bits set in both fingerprints A and B. Note, that the Soergel distance is the complement of the very popular Tanimoto coefficient (i.e. $D_{\text{Soergel}} = 1.0 - D_{\text{Tanimoto}}$). Since the results using either the Tanimoto coefficient or its complement will be equivalent, we use the Soergel distance allowing for a more convenient comparison with results achieved with the Euclidean distance.

Fragment Generation. The database of fragments that is needed to search for potential bioisosteric replacements was generated by a systematic fragmentation of structures from several source structure databases. Using existing chemical structures as a starting point is attractive since per definition only realistic chemical groups will be generated, and the statistics of the fragmentation process can be useful too as an additional property of each of the fragments stored. For instance, a high frequency of a fragment occurring in a database of known drugs may be an indication of its druglikeness, and a fragment that appears only once or a few times could be unattractive for various reasons or could even be due to a structure drawing error.

A few simple rules are controlling the structure fragmentation: Structures can only be split into parts at acyclic single bonds that either connect a heteroatom with a carbon atom or are connected to a branching point. This requirement ensures that structures are only split at “activated” bonds and thus prevents that featureless alkane chains are fragmented. Additionally, bonds that are part of a functional group (e.g. sulfonamide) cannot be cleaved.

The fragmentation process systematically breaks one, two, or three of the allowed bonds at a time and marks the attachment points with a special atom type R. This generates fragments with one, two, or three attachment points which we call R-groups, linkers, and cores, respectively. In Figure 3 a thrombin inhibitor is shown with the cleavable bonds highlighted in bold, and some examples of the resulting R-groups, linkers, and cores are given.

Validation Procedure. The validation of a method to suggest potential bioisosteric replacements depends on the availability of experimentally proven examples of such replacements (*true* pairs of bioisosteric fragments). Having these experimental data available, the validation of descriptors and search methods can be based on the assumption that a well performing approach will assign smaller distances (viz. higher similarity) to true pairs of bioisosteric fragments than to random pairs of druglike fragments. This validation strategy has first been suggested by Watson et al. who used 185 bioisosteric functional group pairs manually extracted from the BIOSTER database in their analysis.⁸

We also decided to make use of that valuable source of bioisosteric compounds for validating our methods. The version 2001.1 of the BIOSTER database contains 9816 pairs of bioanalogous structures abstracted from the chemical and biological literature including drugs, agrochemicals, enzyme inhibitors, prodrugs, etc.^{6,7} The structures in the database are organized as pseudoreactions in which the “starting material” and the “product” form a pair of bioisosters. Depending on choices and preferences of the BIOSTER database editor, the “reaction centers” of the pseudoreactions have been manually marked, and these substructures are considered to be bioisosteric. Additional information in the database includes the original literature references, reported biological activities, and a classification of the bioisosteric relationship.

For our validation purposes, we wanted to use only pairs of bioisosteric fragments from the BIOSTER database that fulfill a set of unbiased and objective criteria: Two substructures, one of them present in the first, the other one

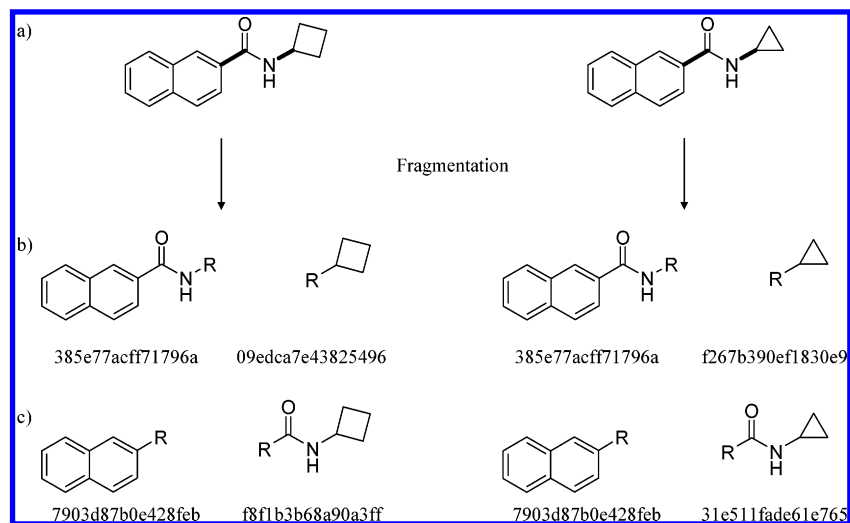


Figure 4. Extraction of bioisosteric R-groups from a pair of bioanalogous structures. Generated fragments are labeled by their canonical hash codes. The counterparts of fragments with identical hash codes form bioisosteric R-groups (b and c). The pairs with the smallest R-groups (b) are kept for the validation set.

present in the second structure of an entry from the BIOSTER database are considered as bioisosters if (a) the remaining parts of the two structures are identical, (b) the remaining parts of the two structures consist of at least two non-hydrogen atoms each, and (c) the number of non-hydrogen atoms in the remaining parts is greater than the number of non-hydrogen atoms in each of the two bioisosteric substructures. Furthermore, we only kept bioisosteric fragments with at least one and not more than 12 non-hydrogen atoms.

Based on these criteria an automatic procedure was devised to extract all the bioisosteric fragments from the BIOSTER database. Although the method is related to the one described by Sheridan,¹⁵ we followed a different approach avoiding the maximum common substructure detection. In the first step of our procedure, the two structures of a bioanalogous pair are exhaustively fragmented into all possible pairs of monovalent fragments (R-groups), two monovalent and one bivalent fragment (linkers), or three monovalent and one trivalent fragment (cores). The fragments generated at every fragmentation step are kept together, and for each fragment a canonical hash code is calculated.¹⁶ Then, the identification of identical substructures in a bioanalogous pair of structures can conveniently be done by comparing lists of hash codes generated for the fragments of a bioanalogous pair: In the case of R-groups it is sufficient that the hash code generated for one of the fragments of the first structure of a pair is identical to the hash generated for a fragment of the second structure of the same pair in order to conclude that the other two fragments (which have different hash codes) can be considered as bioisosters. By analogy, in the case of linkers and cores, if the generated monovalent fragments (4 for linkers, 6 for cores) are pairwise identical, the remaining two fragments form a bioisosteric pair. This method does not only identify the pair of smallest bioisosteric fragments (viz. containing fewest number of non-hydrogen atoms) but also all larger paired fragments up to a size restriction imposed. This can be seen in Figure 4 where the generation of bioisosteric R-groups is illustrated using two naphthoic acid amides. The smallest pair of bioisosteric fragments extracted is formed by the c-butyl and the c-propyl groups. This is the pair that can be identified by deleting the maximum common substructure from the two structures (Figure 4b).

In addition, two larger bioisosteric R-groups are also retrieved that correspond to the c-butyl and the c-propyl groups extended by an amide function (Figure 4c). For our validation purposes we only kept the pair of smallest bioisosteric fragments from each BIOSTER entry and discarded all other pairs from further analysis. This is necessary in order to avoid a bias in the validation set that would make the bioisosteric fragment pairs appear more similar than they in fact are. So from the two bioisosteric R-group pairs in Figure 4 only the c-butyl/c-propyl groups would be kept for the validation set. In the case of R-groups, keeping only the smallest fragments always leads to one unique fragment pair extracted from the original database entry. However, for linkers and cores there is sometimes not a unique pair of smallest fragments. In these cases, all pairs that share the same smallest fragment size have been kept for the validation sets.

Having the data set of bioisosteric groups extracted from the BIOSTER database available, it is now possible to validate descriptors and search methods by comparing the true bioisosteric pairs to random pairs of druglike fragments. The smaller the calculated distances of the true pairs in comparison with the distances of the random pairs, the better the descriptors and search method performs.

To quantitatively assess the quality of our methods we determined two measures as illustrated in Figure 5. As a first measure we used a normalized average distance between true pairs and random pairs (called Δ in Figure 5). Since the average distance cannot directly be used for comparison between different descriptors and settings, it was scaled by the standard deviation of the random pair distance distribution, i.e., Δ is expressed in units of this standard deviation. Not all of the distance measures have a symmetric and well-behaved Gaussian distribution that allows for coming to meaningful conclusions based on the mean values. Some of them are heavily skewed or even have multiple modes. Therefore, a second measure was introduced which is independent of the actual shape of a distribution: First, the 5-percentile of the random pair distribution is determined (upper limit of blue area in Figure 5). Then, the percentage of true bioisosteric pairs that have distances smaller than the 5-percentile identified in the first step can be used as an additional quality measure (red and blue areas added together

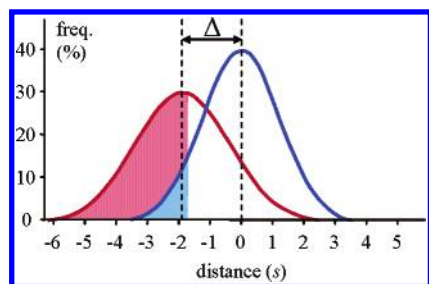


Figure 5. Two measures to quantify the separation between the true bioisosteric pair distance distribution (red line) and the random pair distance distribution (blue line) which have been normalized by the standard deviation (s) of the random pair distribution: (1) Δ , which is the difference between the averages of the true and random distributions and (2) *perc5*, which is the fraction of the true bioisosteric pairs (added red and blue areas) which have a distance lower than the 5-percentile of the random distribution (upper limit of blue area).

Table 1: Number of BIOSTER Database Entries Which Remain after the Indicated Filtering Steps

filtering step	remaining bioisosteric pairs
original database	9816
multiple reactants removed	9475
prodrugs etc. removed	8228
element filter applied	7382

in Figure 5). The greater this number, which we will call *perc5*, the better the separation between true bioisosteric and random pairs is.

RESULTS

Bioisosteric R-Groups. Before bioisosteric fragment pairs were extracted from the 2001.1 version of the BIOSTER database it was necessary to remove certain undesirable entries. Some of the pseudoreactions in the database contain more than one “starting material” and/or “product” in order to specify potential tautomerism. Since it is difficult to uniquely determine the bioisosteric fragment for these entries, we had to discard them from our data set. For a maximum scope, the BIOSTER database also contains entries with prodrugs, photolabels, chelating agents, protecting groups, etc. These molecular modifications potentially obscure a bioisosteric relationship, and we therefore removed the corresponding entries based on the annotations available in the database. Additionally, we removed all entries which pointed to other undesirable molecular properties such as toxicity. A complete list of annotations that led to removal of database entries from our data set is available as Supporting Information. Finally, we removed all pairs of structures with elements other than H, C, N, O, F, P, S, Cl, Br, and I. In this step all entries with questionable valences or pseudoelements such as R were discarded as well. Table 1 gives an overview of the filtering steps that were applied and the number of bioanalogous structures remaining.

Using the procedure introduced in the previous section, we extracted bioisosteric fragment pairs with one, two, and three attachment points from the remaining 7382 bioanalogous structures (Table 2). Applying our unbiased bioisosteric extraction method drastically reduced the number of entries containing true bioisosteric pairs of fragments: 86% of the original bioanalogous structure pairs do not

Table 2: Number of True Bioisosteric Pairs of Fragments Extracted from the BIOSTER Database and the Number of Unique Fragments They Contain

fragment	attachment points	bioisosteric pairs	unique fragments
R-group	1	1042	1208
linker	2	940	1140
core	3	299	434

contain a pair of R-groups fulfilling the criteria, and for only 4% of the original pairs a pair of bioisosteric cores could be identified. It is not surprising that the number of extracted fragments decreases in the order R-groups > linkers > cores. This merely reflects the fact that most of the examples from the BIOSTER database (and from medicinal chemistry publications) originate from compound series varying substituents attached to a common core. The extracted R-group, linker, and core fragment pairs are available as Supporting Information in the form of three structure files in Smiles format.

Descriptor and Distance Optimization and Validation.

To test the influence of the different options and parameters of the topological pharmacophore fingerprint, we used the 1042 pairs of R-groups extracted from the BIOSTER database as examples of true bioisosters. These were compared to a set of 10 000 random fragment pairs extracted with an analogous fragmentation scheme from structures taken from the World Drug Index (WDI).¹⁷ The following sets of parameters were used: normal and fuzzified fingerprints, Euclidean and Soergel distances, maximum atom pair frequencies of 1, 2, 3, 4, 7, and 23 different combinations and weights of pharmacophore descriptors. Additionally, we investigated whether it is sufficient to use the structures with formal charges removed as much as possible (“neutral fragments”), or whether it offers any advantages to charge the fragments corresponding to a physiological pH using simple substructure rules. For all 920 possible combinations of parameters the two quality measures (Δ versus *perc5*) introduced in the Methods section were determined. A scatter plot of Δ versus *perc5* for the 920 combinations is shown in Figure 6, where the yellow squares and circles represent the R-group results.

The scatter plot indicates a clear separation between the true bioisosters and the random pairs even for the worst parameter sets. The best separation for the R-groups in terms of the *perc5* criterion (*perc5* = 64.3%) was achieved with the following parameter set: normal fingerprints, neutral fragments, Euclidean distance similarity measure, maximum atom pair frequency limited to 3, and pharmacophore properties ADX⁴HCV, where the superscript indicates that attachment point X was given a weight of 4. This parameter set also scored extremely well based on the criterion Δ (Δ = 2.2): only 7 other parameter sets had slightly better values for Δ . The good correlation between the two quality measures is a general trend that can be seen for other parameter sets as well (cf. Figure 6). This indicates that, with only a few exceptions, most of the parameter sets have led to well-behaved distance distributions. In the following paragraphs, the individual influence of the different parameters will be discussed for the R-group bioisosters.

Both the type of the fingerprint (normal vs fuzzified) and the charge status of the fragments (neutral vs charged) have

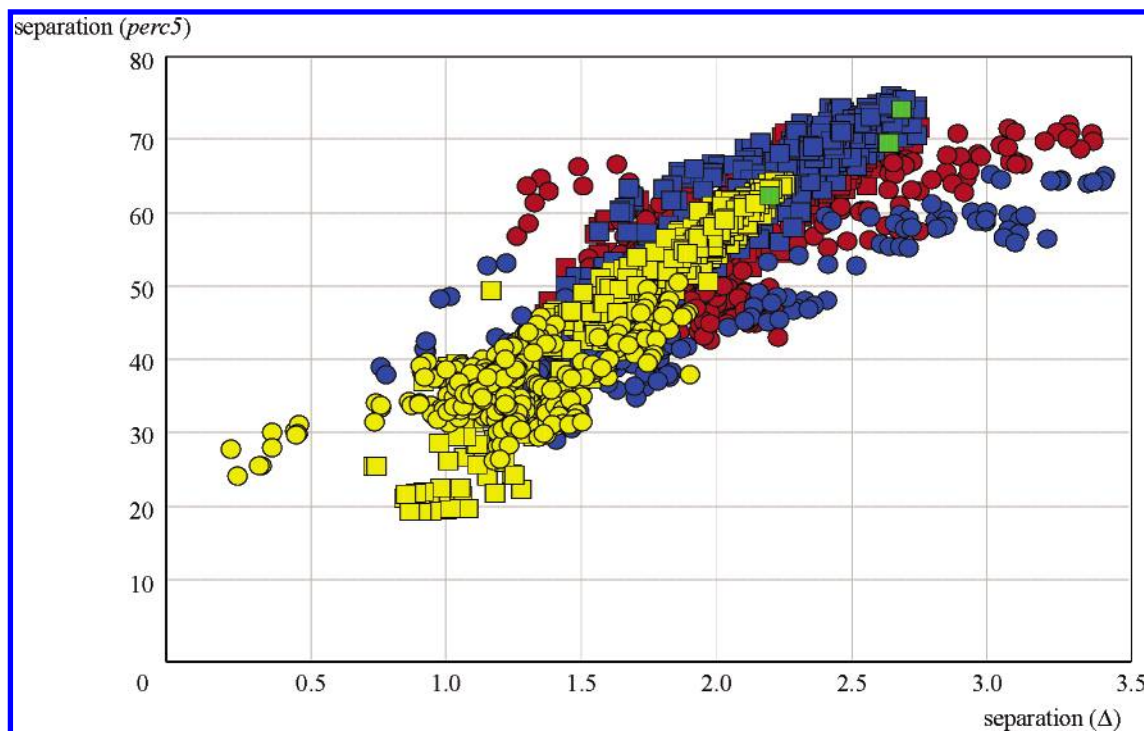


Figure 6. Scatter plot of the two quality measures Δ versus *perc5* for 2760 different topological pharmacophore descriptors/distance measures/fragment types. The larger Δ and *perc5* are, the better the true bioisosteric pairs have been separated from the random pairs. In yellow the results for 920 combinations tested on R-groups are shown, and the same combinations have been applied to linkers (in blue) and cores (in red). Combinations evaluated by the Euclidean distance are indicated with squares and by the Soergel distance with circles. The green squares represent a combination applied within the IBIS implementation: ADX³HCV, maximum atom pair frequency 3, charged fragments, fuzzified fingerprint, and the Euclidean distance measure.

no noticeable influence on the degree of separation between true bioisosters and random pairs. Charge status is most important when only the pharmacophore properties X, A, and D are present in the fingerprint, since in these cases the relative influence of the acceptor and donor properties are the most prominent. For instance, a tertiary sp^3 -nitrogen will act as an acceptor in its neutral form, but it changes into a donor atom upon charging. However, if additional pharmacophore properties are added the relative influence of acceptors and donors decreases, and the charge status of the fragments is no longer decisive.

In Figure 6 the experiments with Euclidean distances are represented by squares, the ones with Soergel distances by circles. As can be seen for the R-groups, there is a significant difference between the two sets of results with the Euclidean distance results outperforming the ones based on the Soergel distance. The best result achieved with the Soergel distance (normal fingerprints, maximum atom pair frequency 1, pharmacophore properties ADXP, charged fragments) is *perc5* = 50.65 and Δ = 1.9, which is considerably less than the best results achieved with the Euclidean distance (see above). This is a surprising finding, since the Tanimoto coefficient has for years been the measure of choice for fragment-based chemical similarity work.¹⁴ As the major difference between Euclidean and Soergel distances is the influence of a common absence of features, it can be argued that for the limited variety of feature types (i.e. pharmacophore properties) we are considering here, the common absence of a pharmacophore does indeed matter, which is accounted for in the Euclidean distance measure. For instance, two bioisosteric groups can indeed be considered

more similar if they both do not contain a large hydrophobic moiety.

The maximum atom pair frequency has a small but distinct influence on the degree to which true and random bioisosteric pairs can be separated, with an optimum value of two or three. This can be seen, e.g., when comparing the results achieved for different maximum atom pair frequencies and setting the rest of the parameters to the ones of the best parameter set (*perc5* = 64.3%, see above). For atom frequencies 1, 2, 3, 4, and 7 the following *perc5* values have been achieved, respectively: 63.4%, 64.2%, 64.3%, 62.6%, and 60.7%. Although it might be argued to what extent these differences are significant, this trend can be noticed throughout all the data generated.

Besides the distance function to be used, the proper choice of the pharmacophore properties has the greatest influence on the quality of the results achieved. Using only X and V and so just taking shape and size relative to the attachment point into account, the least favorable results were achieved with a value of *perc5* = 38.9% for the best combination of parameters keeping only XV fixed. In the remainder of the discussion only the pharmacophore properties used will be reported referring to the optimum parameter combination with these properties. Adding the hydrogen bonding properties A and D did considerably improve the results (ADXV, *perc5* = 52.1%). It is interesting to note that the combination ADX alone, i.e., without the shape and size parameter V, did not perform as well as the combination ADXV and that in this case the Soergel distance did outperform the Euclidean distance (ADX, *perc5* = 49.7%). An additional improvement could be achieved by adding hydrophobes to the pharma-

cophore description (ADXHV, *perc5* = 61.0%). It was not possible to further increase the *perc5* quality measure by adding various combinations of additional pharmacophore properties (P, M, C). Only by changing the relative weight of the attachment point X with respect to the other pharmacophore properties some further separation between true and random pairs of bioisosters could be realized. The optimum pharmacophore description is obtained when assigning a four times higher weight to attachment points than to other pharmacophore properties (ADX⁴HCV, $\Delta = 2.2$ *perc5* = 64.3%). This optimum was achieved in combination with neutral fragments and normal fingerprints, having a maximum atom pair frequency of 3.

Validation Using Bioisosteric Linker and Cores. The efficiency of the optimum pharmacophore description determined for R-groups was tested on linkers and cores using the same procedure as for R-groups. Now sets of 940 true bioisosteric pairs of linkers and 299 pairs of cores extracted from the BIOSTER database were compared to 10 000 random pairs of linkers and cores, respectively, originating from the WDI. The results of the same combinations of pharmacophore descriptors, fingerprint type, charge status, maximum atom pair frequencies, and similarity indices as used for the R-groups are also shown in Figure 6. Judging Δ and *perc5*, both the linker and the core distributions show an even better separation between true pairs of bioisosters and random pairs than the R-group distributions do.

Whereas the Euclidean distance measure behaves well, the Soergel coefficient shows a skewed profile especially at higher Δ for linkers and cores. All Δ values higher than 3.0 are results for either linkers or cores that have been described by the very simple pharmacophore description ADX or ADXP using the Soergel similarity measure. The difference between Δ and *perc5* here indicates the non-Gaussian nature of the underlying distance distributions, and in these cases Δ is clearly not a good performance indicator. Furthermore, another effect plays an important role when using the Soergel distance: With this distance measure the optimum pharmacophore descriptions for linkers (ADX, *perc5* = 65.2%) and for cores (ADX, *perc5* = 72.1%) do not seem to match our previous findings for R-groups. To understand this, we have to go back to the validation sets of bioisosteric linkers and cores. In these two sets another atom pair has come into play: the XnX atom pair (two attachment points separated by a topological distance of *n*) which is by default absent in R-groups. Analysis of the linker validation set shows that for 81% of the pairs, the XnX atom pair for both fragments are equal. In contrast, the random set of linkers contains only 14% of pairs with equal XnX atom pairs. The consequence is that, especially for the Soergel distance measure, the XnX atom pair is dictating the overall result for a pharmacophore description and that other descriptors such as V (shape/size) have hardly any influence. For the Euclidean measure the XnX atom pairs are less dominant, since the absence of features are equally important. The same effect is observed for cores. The core validation set contains 79% of pairs for which all XnX atom pairs are equal, whereas this figure is only 4% in the random set.

The optimum pharmacophore description for R-groups (ADX⁴HCV, neutral fragments, normal fingerprints, maximum atom pair frequency of 3), and using the Euclidean distance, also works perfectly well to separate the true

bioisosteric linkers and cores from the sets of random pairs (*perc5* = 76.0%, $\Delta = 2.6$ and *perc5* = 71.2%, $\Delta = 2.6$, respectively).

Implementation of IBIS. Based on this work we have implemented an interactive intranet application called IBIS (Intranet BioIsoster Search) to search for potential bioisosteric replacements for a given R-group, linker, or core. IBIS returns the most similar potential bioisosteric fragments according to the pharmacophore fingerprint description by searching large databases of R-groups, linker, and cores. These databases have been generated by exhaustive fragmentation of all structures available in the WDI,¹⁷ ACD,¹⁸ and Organon in-house structure databases. Only fragments with a maximum of 12 non-hydrogen atoms have been kept, resulting in 162 000 unique R-groups, 294 000 unique linkers, and 322 000 unique cores. The fingerprint is based on the pharmacophore description ADX³HCV (insignificantly different from ADX⁴HCV but slightly more efficient) in combination with a maximum atom pair frequency of 3. The user can control the charge status and the fingerprint type (fuzzified or not) to emphasize the kind of potential bioisosters to be retrieved.

Along with the distance distribution of true bioisosteric R-group pairs from the BIOSTER database and the random pairs, examples of normalized distances of true bioisosteric R-group pairs using the IBIS implementation are shown in Figure 7. An example of very similar R-groups is ethyl and chloromethyl. Actually, in our pharmacophore description ethyl and chloromethyl are identical since in this case the chlorine atom and the corresponding methyl group are both described by non-hydrogen (V). As a result, the Euclidean distance will be zero, which translates into 6 standard deviation units away from the random distribution ($d = -6.0$). According to our descriptions, a "ring closure" of a dimethyl aminium group into a pyrrolidinium results in very similar fragments ($d = -4.5$). Reversing the ester function, a common practice in drug design, also still leads to rather similar R-groups ($d = -3.6$). Going from 6-fluorbenz[d]-isoxazol-3-yl to *p*-fluorobenzoyl involves a ring opening. With $d = -2.8$ the resulting pharmacophore description is nevertheless still quite similar, consisting of a largely conjugated system (C) having a hydrophobic ring (H) and an important acceptor functionality (A) two bonds away from the attachment point (X). The next pair of true bioisosters, the acetyl and phenyl groups ($d = -1.0$), is more difficult to understand on the basis of our pharmacophore description. The true bioisosteric pairs and random pairs now occur almost equally frequent. The last example shown is the pair of a carbonate ester moiety and a carboxylate group. The distance ($d = +0.5$) is even worse than the average random replacement. Looking back in the original literature,¹⁹ we discovered that the carbonate ester moiety is actually a prodrug, which has not been annotated in the BIOSTER database (entry ACI105) and therefore was not removed from our validation set.

An important addition to the optimized pharmacophore fingerprint search methodology is the option to focus an IBIS search on fragments with certain properties. This is especially important if bioisosteric fragments are needed to modify structures such that compounds with better ADME/Tox profiles can be designed. The constraints currently implemented in IBIS are Molecular Weight, Polar Surface Area,

tion of $\Delta = 2.2$ (in standard deviation units of the random pair distance distribution). The optimal fingerprint has been validated against linkers and cores, which also led to an excellent separation between true bioisosteric pairs and random pairs ($\Delta = 2.6$ and $\Delta = 2.6$, respectively). Furthermore, it has been demonstrated that for our method the Euclidean distance is a more adequate similarity measure than the Soergel distance, likely due to the significance of mutual absence of important pharmacophore features.

Our bioisoster search method has been implemented as an intranet tool, called IBIS. Additional property constraints within IBIS allow to focus on fragments with improved properties. IBIS is now routinely used by Organon researchers and contributes to advance drug discovery projects.

Supporting Information Available: The complete list of annotations that led to removal of BIOSTER entries from our data set, the R-group, linker, and core fragment pairs extracted from the BIOSTER database as well as the three sets of 10 000 random fragment pairs extracted from the WDI. The six chemical structure files are in Smiles format. This material is available free of charge via the Internet at <http://pubs.acs.org>.

REFERENCES AND NOTES

- (1) Thornber, C. W. Isosterism and molecular modification in drug design. *Chem. Soc. Rev.* **1979**, 8, 563–580.
- (2) Patani, G. A.; LaVoie E. J. Bioisosterism: A Rational Approach. *Chem. Rev.* **1996**, 96, 3147–3176.
- (3) Wermuth, C. G. Molecular Variations Based on Isosteric Replacements. In *The Practice of Medicinal Chemistry*; Wermuth, C. G., Ed.; Academic Press: London, 1996; Chapter 13, pp 203–237.
- (4) Olesen, P. H. The use of bioisosteric groups in lead optimization. *Curr. Opin. Drug Discovery Dev.* **2001**, 4, 471–478.
- (5) Lima, L. M.; Barreiro, E. J. Bioisosterism: a useful strategy for molecular modification and drug design. *Curr. Med. Chem.* **2005**, 12, 23–49.
- (6) Ujvary, I. BIOSTER—A database of Structurally Analogous Compounds. *Pestic. Sci.* **1997**, 51, 92–95.
- (7) The BIOSTER database is available from Accelrys Inc. at <http://www.accelrys.com/>.
- (8) Watson, P.; Willett, P.; Gillet, V. J.; Verdonk, M. L. Calculating the knowledge-based similarity of functional groups using crystallographic data. *J. Comput.-Aided. Mol. Des.* **2001**, 15, 835–857.
- (9) Weber, A.; Teckentrup, A.; Briem, H. Flexsim-R: a virtual affinity fingerprint descriptor to calculate similarities of functional groups. *J. Comput.-Aided. Mol. Des.* **2002**, 16, 903–916.
- (10) Ertl, P. Cheminformatics analysis of organic substituents: identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups. *J. Chem. Inf. Comput. Sci.* **2003**, 43, 374–480.
- (11) Holliday, J. D.; Jelfs, S. P.; Willett, P.; Geddeck, P. Calculation of intersubstituent similarity using R-group descriptors. *J. Chem. Inf. Comput. Sci.* **2003**, 43, 406–411.
- (12) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, 25, 64–73.
- (13) Holliday, J.; Hu, C.; Willett, P. Grouping of coefficients for the calculation of intermolecular similarity and dissimilarity using 2D fragment bit-strings. *Comb. Chem. High Throughput Screening* **2002**, 5, 155–66.
- (14) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, 38, 983–996.
- (15) Sheridan, R. P. The most common chemical replacements in drug-like compounds. *J. Chem. Inf. Comput. Sci.* **2002**, 42, 103–108.
- (16) Ihlenfeldt, W. D.; Gasteiger, J. Hash codes for the identification and classification of molecular structure elements. *J. Comput. Chem.* **1994**, 15, 793–813.
- (17) *World Drug Index, version 2004/01*; Derwent Information: London, U.K., 2004.
- (18) *Available Chemicals Directory, version 1/04*; MDL Information Services: San Leandro, U.S.A., 2004.
- (19) Aungst, B. J.; Blake, J. A.; Rogers, N. J.; Saitoh, H.; Hussain, M. A.; Ensinger, C. L.; Pruitt, J. R. Prodrugs to Improve the Oral Bioavailability of a Diacidic Nonpeptide Angiotensin II Antagonist. *Pharm. Res.* **1994**, 12, 763–767.
- (20) *clogP v4.10*; BioByte Corp.: Claremont, U.S.A., Feb 2000.
- (21) *ACD/PhysChem batch v4.76*; ACD/Labs, Advanced Chemistry Development: Toronto, Canada, 1994–2001.
- (22) Rewinkel, J. B.; Lucas, H.; van Galen, P. J.; Noach, A. B.; van Dinther, T. G.; Rood, A. M.; Jenneboer, A. J.; van Boeckel, C. A. 1-Aminoisoquinoline as benzamidine isoster in the design and synthesis of orally active thrombin inhibitors. *Bioorg. Med. Chem. Lett.* **1999**, 9, 685–90.

CI0503964