Calculation of Intersubstituent Similarity Using R-Group Descriptors

John D. Holliday, Stephen P. Jelfs, and Peter Willett*

Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield,
Western Bank, Sheffield S10 2TN, United Kingdom

Peter Gedeck

Novartis Horsham Research Centre, Novartis Pharmaceuticals UK Ltd., Wimblehurst Road, Horsham,
West Sussex, RH12 5AB, United Kingdom

Received August 22, 2002

This paper discusses the calculation of the similarities between pairs of substituents on ring systems. An R-group descriptor characterizes the distribution of some atom-based property, such as elemental type or partial atomic charge, at increasing numbers of bonds distant from the point of substitution on the parent ring. The similarity between a pair of descriptors is then calculated by a comparison of the corresponding property vectors. Experiments with the BIOSTER database demonstrate the ability of such similarity measures to discriminate between bioisosteric and nonbioisosteric functional groups.

INTRODUCTION

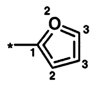
The measurement of similarity underlies many of the techniques that have been developed for the correlation of molecular structure with biological activity. Thus far, the principal focus of study has been the calculation of intermolecular similarity, i.e., a quantification of the degree of structural resemblance between pairs of complete molecules and many different types of similarity measure have been used for this purpose.[1,2] It is possible to group these measures into two broad classes. Global measures are more common and provide an overall measure of how similar one molecule is to another, e.g., a similarity measure based on 2D fragment bit-strings and the Tanimoto coefficient[3] might specify that two molecules had a similarity of 0.76. Local similarity measures additionally align the two molecules that are being compared,[4] thus facilitating the identification of those parts of the molecules that are most closely related.

In this paper we consider the calculation of intersubstituent similarity, i.e., the quantification of the degree of resemblance between pairs of substituents located at the same relative position on a ring system that is common to a pair of molecules. Our measure is a local one, in that while yielding an overall measure of how similar one substituent is to another, it is also possible to ascertain the extent and the location of the similarities and differences between two substituents. The measurement of intersubstituent similarity has been studied for very many years in the context of designing congeneric series for QSAR studies; early examples of this approach are reported by Hansch et al.[5] and by Wooton et al.[6] More recently, the introduction of combinatorial synthesis has led to the development of a whole range of substituent-based library design procedures, following the seminal paper of Martin et al.[7] The work reported here has arisen from an interest in bioisosterism, i.e., atoms, functional groups, or molecules that are associated with similar types of biological activity.[8,9] While the concept of bioisosterism has been known for many years, it is only recently that large-scale computational studies have started to appear. Watson et al.[10] discuss the use of crystallographic information in the ISOSTAR database to quantify the extent to which pairs of substituents are involved in sets of similar nonbonded interactions, and Sheridan[11] has recently reported an analysis of the MACCS Drug Data Report (MDDR) database to quantify the extent to which pairs of substituents occur in similar structural environments. These approaches are elegant in approach but both have limitations: the ISOSTAR work is constrained by the lack of detailed crystallographic information for many of the substituents of interest, while the MDDR work involves a time-consuming clique-detection procedure that meant that only a small sample of the database could be analyzed.

This brief communication describes an approach to the calculation of intersubstituent similarities that draws on much of the referenced previous work and has the following characteristics: it characterizes substituents using physicochemical information that can be readily calculated from a 2D structure diagram; it is based on a type of structural descriptor that has previously been used successfully for the design of combinatorial libraries; and it is sufficiently simple to be applied to very large files of substituents at minimal computational cost. The next section of the paper introduces the descriptor that we have used, which we refer to as an R-group descriptor, and we then describe how pairs of such descriptors can be compared to measure the similarity between the corresponding substituents, in a manner analogous to the Web system for substituent-based similarity searching reported by Ertl.[12] The fourth section reports an extensive series of experiments to validate the use of the R-group descriptor as an appropriate representation of a substituent, and the paper concludes with a summary of our findings and suggestions for future work.

* Corresponding author e-mail: p.willett@sheffield.ac.uk.

10.1021/ci025589v CCC: $25.00
Published on Web 02/01/2003



Distance:	1	2	3	4	5	6
Atomic Weight	12.01	29.02	26.04	0.00	0.00	0.00
Hydrophobicity	0.08	0.44	0.56	0.00	0.00	0.00
Molar Refractivity	3.24	5.49	8.81	0.00	0.00	0.00
Atomic Charge	112.00	-277.00	165.00	0.00	0.00	0.00
Polar Surface Area	0.00	13.14	0.00	0.00	0.00	0.00
Hydrogen Bond Acceptor	0.00	1.00	0.00	0.00	0.00	0.00
Hydrogen Bond Donor	0.00	0.00	0.00	0.00	0.00	0.00

Figure 1. R-group descriptors based on seven different atomic properties.

THE R-GROUP DESCRIPTOR

The descriptors considered here have their basis in work by Martin et al.[7] that involved several different ways of calculating the dissimilarity between pairs of substituents. In one of their approaches, a range of atomic properties is calculated for each of the atoms some number of bonds away from a fixed position in a molecule, such as the point of substitution for a functional group; the sum of the values for each such property at each number of bonds distant is then calculated and a substituent characterized by the resulting set of summed values. This provides a structural representation that is related to the autocorrelation functions described by Broto et al.[13] and by Schuur et al.[14] Martin et al. showed that these property-sums could assist in the design of combinatorial libraries when used in combination with a diversity-selection procedure based on D-optimal design but do not appear to have investigated the characteristics of this representation in detail; this we have done in the work reported here.

The basic approach of Martin et al. can be generalized to encompass a range of types of descriptor. Specifically, an R-group, *i*, is defined in terms of a specific atomic property, *p*, by a descriptor D_i^p that is a vector of length *n* containing a series of values $d^p{}_{ij}$, each of which is the sum (or some other combination) of the chosen atomic property values at a distance of *j* bonds from the point of attachment of the substituent to the central ring scaffold, i.e.,

$$D_i^p: = \{d_{i,j}^p\}_{j=1,n} = (d_{i,1}^p,...,d_{i,n}^p)$$

The index *j* encodes the through-bond distance (i.e., the number of bonds) from the point-of-attachment of the R-group, so that small values of *j* describe positions on the substituent close to the start, while larger values of *j* denote positions further away from the point of attachment. In this paper, we consider distance to be defined by the number of bonds, but one could equally use a binned version of the through-space (Euclidean) distance between an atom and the point of attachment, calculated via use of a structure generation program such as CONCORD or CORINA; the use of such distances is currently under investigation. We also consider only the sums of the property values at each specific distance *j*. However, even with these restrictions, it is possible to define a range of different R-group descriptors, depending on the particular atomic properties that are used. For example, Figure 1 shows substituent vector values for seven different descriptors: atomic weight; atomic charge;[15] hydrogen-bond donor (HBD) count; hydrogen-bond acceptor (HBA) count; atomic contribution to molar refractivity (MR);[16] atomic contribution to hydrophobicity (logP);[17] and atomic contribution to polar surface area (PSA).[18] In this case, the maximum through-bond distance is just three bonds; however, as discussed in the next section, the largest substituents considered in our experiments have through-bond distances of up to six bonds, and the vector values in Figure 1 have thus been right zero-filled.

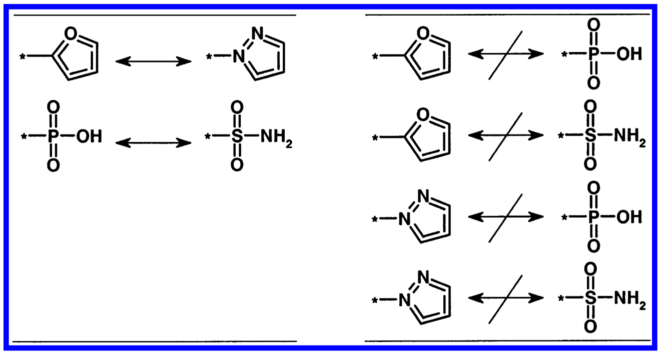


Figure 2. Example bioisosteric (left) and nonbioisosteric (right) pairs of functional groups derived from the BIOSTER database.

Having introduced the R-group descriptor, the next section reports a validation study that seeks to determine the extent to which this descriptor encodes meaningful structural information.

DISCRIMINATION OF BIOISOSTERIC AND NONBIOISOSTERIC FUNCTIONAL GROUPS

Validation Method. The validation study draws on work by Watson et al.[10] that used the BIOSTER database, a compilation of several thousand pairs of molecules that have been identified from the literature as being bioisosterically equivalent.[19] Watson et al. analyzed these pairs of bioisosteric molecules to identify pairs of bioisosteric functional groups, where two functional groups are regarded as bioisosteric if a pair of bioisosteric molecules differ only in the presence of these groups. There was a total of 418 such pairs of functional groups for which the R-group descriptors could be calculated: these 418 pairs will be referred to as the *bioisosteric data set*. This data set contained a total of 106 unique groups, representing a total of 5565 distinct pairs of groups. Taking away the 418 bioisosteric pairs gives 5147 pairs, constituting what will be referred to subsequently as the *nonbioisosteric data set*. Figure 2 shows two bioisosteric pairs and the four potential nonbioisosteric pairs that can be derived from them.

It must be emphasized that the separation into bioisosteric and nonbioisosteric is open to criticism, since both data sets contained numerous questionable pairs, i.e., bioisosteric pairs that may not necessarily be bioisosteric, and likewise nonbioisosteric pairs that may well be bioisosteric; moreover, such descriptions may well apply only in the context of a specific biological target. However, the BIOSTER data does provide the only public source of carefully evaluated literature data of which we are aware that can be used for an analysis of the sort described below. The basic idea that we have adopted[10] is that one would expect that a bioisosteric

Profile-dependent, $D_i^p := (d_{i,j}^p)_{j=1,6} = (d_{i,1}^p, \ldots, d_{i,6}^p)$ (a)

Position-dependent, $D_j^p := (d_{i,j}^p)_{i=1,106} = (d_{1,j}^p, \ldots, d_{106,j}^p)$ (b)

Property-dependent, $D^p := (d_{i,j}^p)_{i=1,106;\, j=1,6} = (d_{1,1}^p, \ldots, d_{1,6}^p, \ldots, d_{106,1}^p, \ldots, d_{106,6}^p)$ (c)

Figure 3. R-group descriptor vectors across which standardization could proceed.

pair of substituents would tend be more similar to each other (using some measure of intersubstituent similarity) than a nonbioisosteric pair of substituents, only if the similarity measure was capable of encoding information that was relevant to the contributions to biological activity provided by those substituents. If this is indeed the case, then we would expect that a distribution of the similarities between each of the pairs of substituents in the bioisosteric data set would tend to be shifted toward higher similarity values than would the corresponding distribution for the similarities between pairs of substituents in the nonbioisosteric data set. Conversely, if this is not the case, then we would expect no significant difference between the two distributions.

Calculation of Similarity Measures. A chemical similarity measure comprises three parts:[1] the structural representation; the standardization that is applied to such representations; and the similarity coefficient that is used to quantify the degree of resemblance between two standardized representations. This section describes the calculation of these three components for the R-group descriptors.

Seven R-group descriptors were generated for each of the unique 106 functional R-groups identified in the BIOSTER data, as described in the previous section. The maximum R-group length was found to be only six bonds from its attachment point, and a 6-element vector could hence be used to represent each R-group descriptor. Therefore, for each R-group i ($1 \leq i \leq 106$) and property p ($1 \leq p \leq 7$), a vector was calculated spanning the distance j ($1 \leq j \leq 6$) from the R-group's attachment point (as exemplified in Figure 1). The resulting representations were then used for the calculation of a range of different intersubstituent similarity measures, each based upon one of the individual atomic properties; similar experiments were carried out in which all of the properties were combined.

Standardization is sometimes omitted in the calculation of a similarity measure but is necessary here in view of the very different ranges of values associated with the various R-group descriptor elements. Our experiments employed Z standardization, so that each variable has a mean and a standard deviation of zero and unity, respectively. Given the data set of R-group descriptors, three alternative standardization approaches were carried out for each specific atomic-property, as shown in Figure 3: a profile-dependent method that involved the standardization of values across each R-group descriptor individually; a position-dependent method that involved the standardization of values across each particular position, or distance; and a property-dependent method that involved the standardization of values across each entire property-group. An example R-group and its associated standardized R-group descriptors based on molar refractivity are shown in Figure 4.

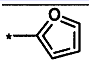
	None	3.24	5.49	8.81	0.00	0.00	0.00
	Profile	0.09	0.70	1.61	-0.80	-0.80	-0.80
	Position	-0.32	-0.36	0.86	-0.50	-0.41	0.00
	Property	0.06	0.67	1.56	-0.81	-0.81	-0.81

Figure 4. Example R-group descriptors based on molar refractivity and different standardization methods.

Euclidean Distance, $S^p{}_{A,B} = \left[\sum_{j=1}^{6}\left(d_{A,j}^p - d_{B,j}^p\right)^2\right]^{1/2}$ (a)

Euclidean Distance, $S_{A,B} = \left[\sum_{p=1}^{7}\sum_{j=1}^{6}\left(d_{A,j}^p - d_{B,j}^p\right)^2\right]^{1/2}$ (b)

Figure 5. Euclidean distance (a) for a single property and (b) for all properties.

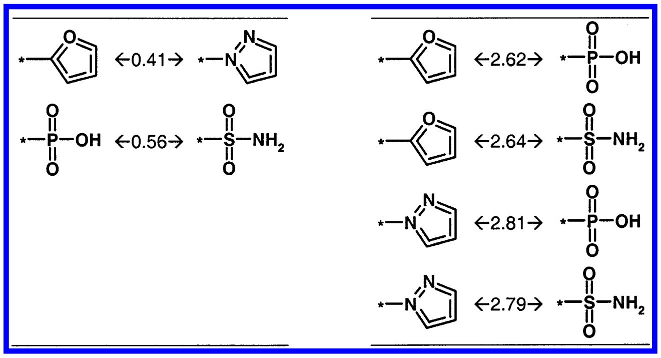


Figure 6. Euclidean distances between groups based on unstandardized MR R-group descriptors.

Three popular similarity coefficients were used to calculate the similarities between each pair of (un)standardized representations: the Cosine coefficient, the Tanimoto coefficient, and the Euclidean distance.[3] Two versions of each coefficient were implemented, one allowing the derivation of similarity scores based upon each individual atomic-property and the other allowing a single similarity score to be derived based on all the atomic-properties combined (as shown for the Euclidean distance in Figure 5).

Using the three similarity coefficients, seven atomic property descriptions and one combined property description, with unstandardized and three standardized forms, it was possible to calculate a total of 96 different similarity scores for each pair of R-groups: examples of these scores are shown in Figure 6. The distributions of the scores were subsequently studied to determine which combinations of the similarity coefficients, atomic properties, and standardization methods performed the best, i.e., which implementation resulted in the bioisosteric R-groups as having the greatest intersubstituent similarities when compared with the nonbioisosteric intersubstituent similarities. The significance of the differences, if any, between the bioisosteric distributions and the corresponding nonbioisosteric distributions was assessed using the Student t-test and the Mann−Whitney U test. The Mann−Whitney test was used in addition to the standard t-test as the former, nonparametric test does not make any assumption as to the normality of the distributions that are being compared. In both cases, a one-tailed test was used, so as to test whether the bioisosteric similarities were significantly greater than the nonbioisosteric similarities. A difference at the $p \leq 0.05$ was taken to be significant; in fact, many of the differences observed were significant at the $p \leq 0.01$ level.

RESULTS

Some of the distributions based on Euclidean distance and property-dependent standardization are shown in Figure 7.

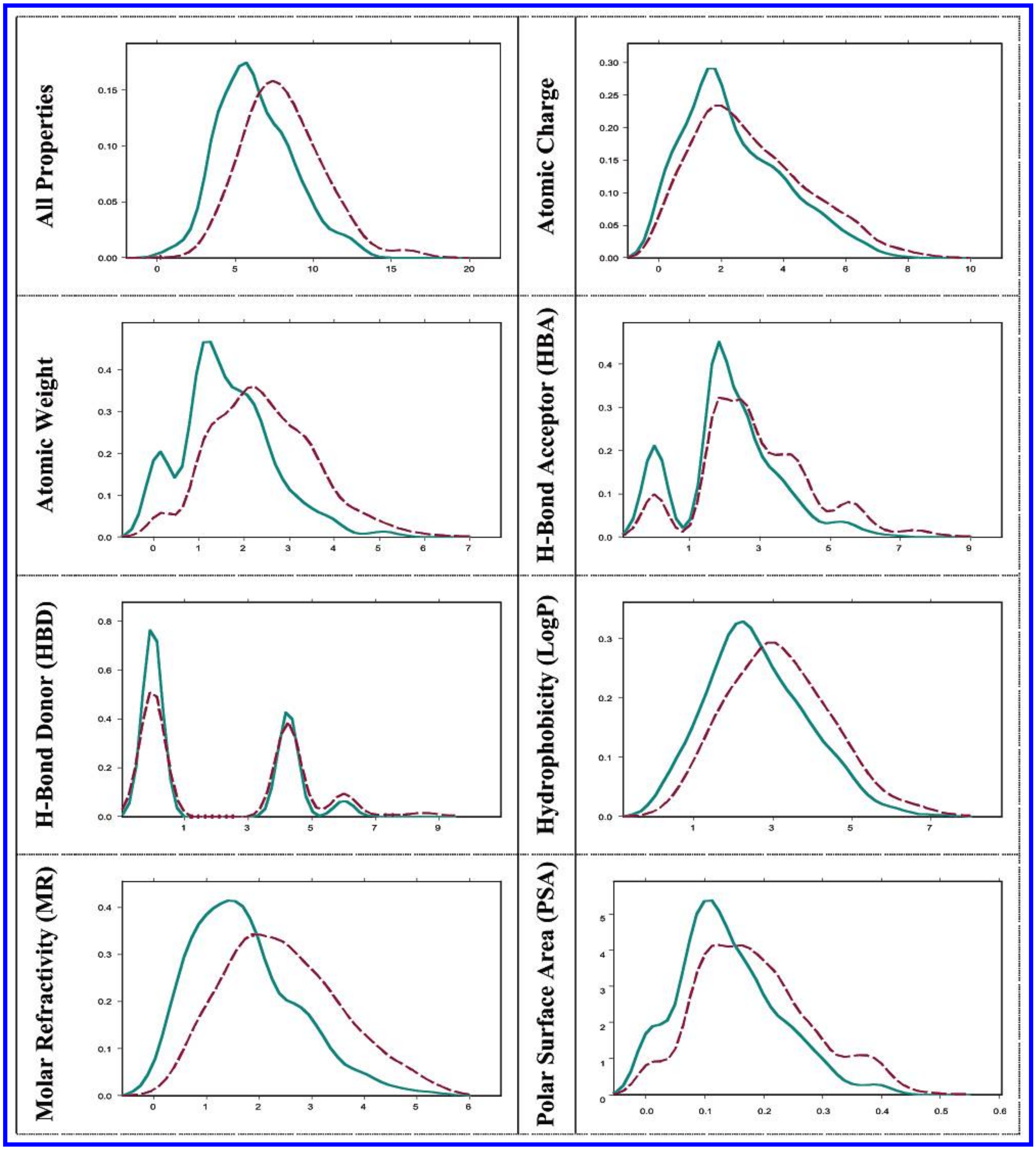


Figure 7. Similarity score distributions based on the Euclidean distance and property-independent standardization: the bioisosteric distribution is the solid-line and the nonbioisosteric distribution the dashed-line.

These plots highlight the shift of the bioisosteric data set distribution to the left, (i.e., smaller distance and hence more similar) of the nonbioisosteric data set distribution for each of the representations.

The results of the Student *t* and Mann−Whitney *U* tests are summarized in Tables 1 and 2. Since increasing *t*-test values and decreasing *U* values imply a greater separation between the bioisosteric and nonbioisosteric distributions, it can be seen that both of the tests rank the various atomic-properties, similarity coefficients, and standardization methods in a broadly similar manner, even for distributions that are far from normally distributed (e.g., those obtained from hydrogen-bond donor counts as shown in Figure 7). Specifically, the properties MR and atomic weight performed the best (as denoted by the highest average *t*-values and lowest average *U*-values), followed by PSA, HBA, and logP, with atomic charge and HBD performing the worst.

All three standardization methods gave results that were consistently superior to the unstandardized data; hardly surprisingly, this was especially the case when all of the properties were combined. Considering the average values in the right-hand columns of Tables 1 and 2, property-based standardization normally gives the best results, with Euclidean distance giving the best average performance of the three similarity coefficients. The best single result in Table 1 is property-standardized MR data with the Cosine coefficient,

Table 1. Student *t*-Test Values (Large Values Better) Using Different Types of Property for the Generation of the R-Group Descriptor[a]

coefficient	standardization	all	charge	weight	HBA	HBD	lopgP	MR	PSA	average
cosine:	none	4.40	1.83	12.59	6.70	1.31	4.84	11.73	8.40	6.8
	profile	10.92	1.74	13.17	6.59	1.37	4.66	13.23	8.26	7.49
	position	13.12	3.14	13.27	8.88	4.55	4.18	15.10	10.09	9.04
	property	14.08	3.55	14.95	7.46	4.25	6.29	17.11	8.92	9.58
tanimoto:	none	5.26	5.13	13.37	6.51	4.01	6.35	12.80	6.59	7.50
	profile	10.11	4.25	12.46	6.66	4.01	4.97	12.68	7.39	7.82
	position	11.76	4.37	11.78	8.11	4.15	5.04	13.62	9.20	8.50
	property	12.62	5.17	14.32	7.17	4.02	6.72	15.58	8.29	9.24
euclidean:	none	6.24	6.07	15.78	10.76	5.29	8.61	14.54	10.25	9.69
	profile	11.01	3.84	12.28	7.60	3.98	4.95	12.46	8.69	8.10
	position	9.41	3.91	10.02	6.70	4.23	3.85	13.60	8.38	7.51
	property	14.36	6.07	15.78	10.76	5.29	8.61	14.54	10.25	10.71
	average	10.27	4.09	13.31	7.83	3.87	5.76	13.92	8.73	

[a] The critical value ($p \leq 0.05$) is 1.96.

Table 2. Mann-Whitney *U* Test Values $\times 10^5$ (Small Values Better) Using Different Types of Property for the Generation of the R-Group Descriptor[a]

coefficient	standardization	all	charge	weight	HBA	HBD	LogP	MR	PSA	average
cosine:	none	9.26	10.12	6.86	9.01	10.59	9.07	7.02	8.56	8.81
	profile	7.40	10.14	7.12	9.00	10.59	9.20	7.10	8.57	8.64
	position	6.59	9.79	6.68	7.92	9.40	9.46	6.28	7.56	7.96
	property	6.58	9.52	6.68	8.32	9.48	8.81	6.39	7.93	7.96
tanimoto:	none	9.22	9.38	6.70	8.96	9.67	8.89	6.80	9.00	8.58
	profile	7.41	9.57	7.12	8.48	9.67	9.20	7.10	8.26	8.35
	position	6.52	9.77	6.65	7.75	9.47	9.43	6.29	7.43	7.91
	property	6.48	9.38	6.57	8.42	9.57	8.70	6.31	7.95	7.92
euclidean:	none	8.98	9.04	6.25	7.76	9.43	8.19	6.58	7.84	8.01
	profile	7.23	9.73	7.12	8.44	9.67	9.20	7.10	8.06	8.32
	position	7.43	9.22	7.20	8.18	9.60	8.94	6.85	8.07	8.19
	property	6.57	9.04	6.25	7.76	9.43	8.19	6.58	7.84	7.71
	average	7.47	9.56	6.77	8.33	9.71	8.94	6.70	8.09	

[a] The critical value ($p \leq 0.05$) is 11.38.

while unstandardized or property-standardized atomic weight data with the Euclidean distance yielded the best results in Table 2.

CONCLUSIONS

In this paper, we have discussed the use of R-group descriptors for identifying structurally similar pairs of substituents. These descriptors encode information about the distribution of a range of atomic properties at increasing distances from a substituent's point-of-attachment to a central ring scaffold. They are simple to compute from a 2D structure diagram, and validation experiments using the BIOSTER database demonstrate clearly that they do encode meaningful structural information. Pairs of descriptors can be compared rapidly to identify those substituents from a database of the same that are most similar to a user-defined query substituent, thus providing a simple way of suggesting new members of a congeneric series in a lead-optimization program. These results have been used as the basis for an intranet-based system for accessing the Novartis internal database of substituents, which now contains some 85 000 substituents. The system, which is similar in many respects to that described by Ertl,[12] has been written using the Java programming language and also making use of the Chemistry Development kit[20] and of OpenEye's OELib.[21]

There are many ways in which the work reported here could be extended. First, it would be possible to cluster the substituent database and thus to identify groups of substituents that might be regarded as being mutually interchangeable, e.g., for the design of focused combinatorial libraries. Second, the descriptors provide an obvious basis for either qualitative or quantitative SAR studies using, e.g., PLS correlation. Finally, we could calculate R-group descriptors based on a through-space, rather than through-bond, distances, thus providing an alternative to existing approaches to 3D QSAR. We are currently investigating these possibilities.

ACKNOWLEDGMENT

We thank the following: Novartis Pharmaceuticals UK Ltd. and the Biotechnology and Biological Sciences Research Council for funding; the Royal Society, Tripos Inc., and the Wolfson Foundation for software and hardware support; Synopsys Scientific Systems for the provision of the BIOSTER database; and Peter Ertl for compiling and providing the list of R-groups. The Krebs Institute for Biomolecular Research is a Biomolecular Sciences Centre of the Biotechnology and Biological Sciences Research Council.

REFERENCES AND NOTES

(1) *Concepts and Applications of Molecular Similarity*; Johnson, M. A.; Maggiora, G. M.; Eds.; Wiley: New York, 1990.

(2) *Molecular Similarity in Drug Design*; Dean, P. M.; Ed.; Chapman and Hall: Glasgow, 1994.

(3) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(4) Lemmen, C.; Lengauer, T. Computational Methods for the Structural Alignment of Molecules. *J. Comput.-Aid. Mol. Design* **2000**, *14*, 215−232.

(5) Hansch, C.; Unger, S. H.; Forsythe, A. B. Strategy in Drug Design. Cluster Analysis as an Aid in the Selection of Substituents. *J. Med. Chem.* **1973**, *16*, 1217−1222.

(6) Wooton, R.; Cranfield, R.; Sheppey, G. L.; Goodford, P. J. Physicochemical Activity Relationships in Practice. 2. Rational Selection of Benzenoid Substituents. *J. Med. Chem.* **1975**, *18*, 607−613.

(7) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring Diversity: Experimental Design of Combinatorial Libraries for Drug Discovery. *J. Med. Chem.* **1995**, *38*, 1431−1436.

(8) Burger, A. Isosterism and Bioisosterism in Drug Design. *Prog. Drug. Res.* **1991**, *37*, 287−371.

(9) Patani, G. A.; LaVoie, E. J. Bioisosterism: a Rational Approach in Drug Design. *Chem. Rev.* **1996**, *96*, 3147−3176.

(10) Watson, P.; Willett, P.; Gillet, V. J.; Verdonk, M. L. Calculating the Knowledge-Based Similarity of Functional Groups Using Crystallographic Data. *J. Comput.-Aid. Mol. Design* **2001**, *15*, 835−857.

(11) Sheridan, R. P. The Most Common Chemical Replacements in Drug-Like Compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 103−108.

(12) Ertl, P. World Wide Web-based System for the Calculation of Substituent Parameters and Substituent Similarity searches. *J. Mol. Graph. Model.* **1998**, *16*, 11−13.

(13) Broto, P.; Moreau, G.; Vandycke, C. Molecular Structures: Perception, Autocorrelation Descriptor and SAR Studies. *Eur. J. Med. Chem.* **1984**, *19*, 66−70.

(14) Schuur, J. H.; Selzer, P.; Gasteiger, J. The Coding of the Three-Dimensional Structure of Molecules by Molecular Transforms and Its Application to Structure-Spectra Correlations and Studies of Biological Activity. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 334−344.

(15) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity − a Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219−3228.

(16) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.

(17) Ghose, A. K.; Crippen, G. M. Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure−Activity Relationships. I. Partition Coefficients as a Measure of Hydrophobicity. *J. Comput. Chem.* **1986**, *4*, 565−577.

(18) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43*, 3714−3717.

(19) The BIOSTER database is available from Accelrys Inc. at http://www.accelrys.com/.

(20) The Chemistry Development kit is available at URL http://cdk.sourceforge.net.

(21) OELib is available from OpenEye Scientific Software at URL http://www.eyesopen.com/oelib.html.

CI025589V