Journal of Computer-Aided Molecular Design, **12:** 441–449, 1998.
KLUWER/ESCOM

The measurement of molecular diversity by receptor site interaction simulation

Camden A. Parks*, Gordon M. Crippen & John G. Topliss**
College of Pharmacy, University of Michigan, Ann Arbor, MI 48109–1065, U.S.A.

Received 14 November 1997; Accepted 3 March 1998

Key words: chemical leads, combinatorial chemistry, compound libraries, drug discovery, receptor targets, similarity measures

Summary

The assembly of large compound libraries for the purpose of screening against various receptor targets to identify chemical leads for drug discovery programs has created a need for methods to measure the molecular diversity of such libraries. The method described here, for which we propose the acronym RESIS (for *Re*ceptor *S*ite *I*nteraction *S*imulation), relates directly to this use. A database is built of three-dimensional representations of the compounds in the library and a set of three-point three-dimensional theoretical receptor sites is generated based on putative hydrophobic and polar interactions. A series of flexible, three-dimensional searches is then performed over the database, using each of the theoretical sites as the basis for one such search. The resulting pattern of hits across the grid of theoretical receptor sites provides a measure of the molecular diversity of the compound library. This can be conveniently displayed as a density map which provides a readily comprehensible visual impression of the library diversity characteristics. A library of 7500 drug compounds derived from the CIPSLINEPC databases was characterized with respect to molecular diversity using the RESIS method. Some specific uses for the information obtained from application of the method are discussed. A comparison was made of the results from the RESIS method with those from a recently published two-dimensional approach for assessing molecular diversity using sets of compounds from the Maybridge database (MAY).

Introduction

Automated high-throughput screening of large compound libraries is now widely employed in the pharmaceutical industry to identify chemical leads as starting points for new drug discovery programs. The structural or molecular diversity of a compound library used for this purpose is a matter of some consequence. Generally a maximally diverse collection of compounds within some broad parameters relating to their potential as practical structural leads is considered to be most desirable. The compound collections accumulated by pharmaceutical companies over many years reflect structural biases arising from concentrations of various compound types which were focussed on in the research programs. In adding compounds to these libraries, filling in areas of missing or underpopulated structural space should therefore be advantageous. Also, in constructing compound libraries using combinatorial chemistry it is important to have a measure of the resulting degree of molecular diversity so that the library characteristics are understood in terms of the degree of coverage of diversity space.

The measurement of molecular diversity is a relatively new endeavor. The term molecular diversity, while conveying a readily understood general meaning, has not been formally defined to our knowledge. The manner and degree in which one molecule differs from another can be expressed in many different ways. A number of studies bearing on this general subject have been published involving both two-dimensional and three-dimensional approaches [1–13]. It has been pointed out that molecular similarity-

*Present address: Department of Chemistry, Brandeis University, P.O. Box 9110, Waltham, MA 02254–9110, U.S.A.
**To whom correspondence should be addressed.

dissimilarity measures need to be considered in relation to a particular use, e.g., in drug receptor-ligand interactions [13].

The concept we chose to apply was to try to measure molecular diversity in a manner directly relevant to the intended use as described above, namely the discovery of new chemical leads for receptor targets. The method which we have termed *R*eceptor *S*ite *I*nteraction *S*imulation (RESIS) involves building a database of three-dimensional representations of compounds in the library and a set of three-point theoretical receptor sites based on putative hydrophobic and polar interactions. A series of flexible three-dimensional searches is then performed over the database, using each of the theoretical sites as the basis for one such search. The ability of the compounds to position functional groups at given distances is thereby determined which relates to the potential to fit various receptors and should thus provide a measure of molecular diversity in a manner relevant to the intended use.

Two similar three-dimensional methods for molecular diversity measurement involving 3-D pharmacophores have been described [9,10]. Our approach was conceived and executed independently. A comparison of the three methods is given in the section titled 'Comparison with other 3-D methods'.

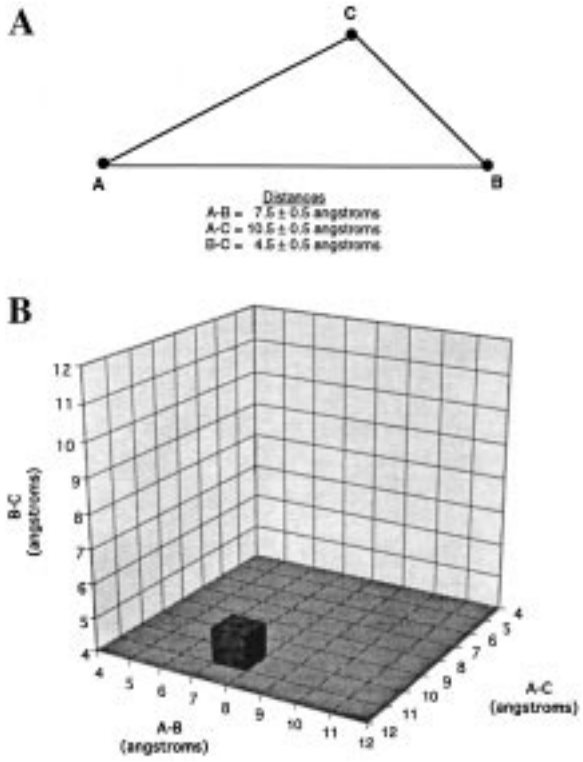

Figure 1. (A) A three-point site. (B) Site space representation of a three-point site.

Methods

Generating theoretical sites

The sites generated are based on a three-point receptor model. These are not meant to exactly model drug-receptor interactions, which often involve more than three sites, but are intended simply to be representative of the major possible binding interactions of a receptor. The central postulate of this model is that just three interactions often account for most of the binding energy [14]. These theoretical sites are used to measure the differences in the ability of the compounds to position functional groups at given distances which should roughly parallel differences in the ability to fit various receptors and should thus measure molecular diversity in a manner relevant to drug design.

A three-point receptor model takes the form of a triangle with one binding group at each vertex. To generate a set of theoretical receptor sites a range is selected for the length of each side of the site triangle as well as a tolerance for the lengths, for example plus or minus 0.50 Å. The lengths for each of the three sides are then varied systematically and independently through the given ranges.

One way of visualizing this site generation is to construct a site-space. Take a three-dimensional cartesian coordinate system and remove the negative portion of each axis. Assign each side of the site triangle to an axis. If a site is defined as being a set of three distances and an associated tolerance for these distances, then a site can be thought of as being a cube in this site space. The center of the cube is located at the coordinates corresponding to the distances of the site, and the length of the edges of the cube is twice the tolerance. For example, say a site had distances of 7.5, 10.5 and 4.5 Å and a tolerance of 0.5 Å. That site, shown in Figure 1A, would be represented in the site space by the cube shown in Figure 1B.

With this site space model in place, the generation of theoretical sites is relatively simple. Say the user specifies ranges of 4 to 12 Å for each of the three site distances. These ranges describe a cubical region in the site-space with sides extending from 4 to 12 Å on each of the three axes. The tolerance the user gives

determines the size of the site cubes into which that region is to be divided. The process of systematically and independently varying the three site distances now amounts to dividing the cubical region of interest into smaller cubes.

The sites used in the RESIS method are of three types. The first type of site is made up of three polar groups. The second is made up of two polar and one hydrophobic group, while the third is made up of one polar and two hydrophobic groups. Sites involving three hydrophobic groups were not included since it was thought to be very unlikely that the three primary binding interactions would all involve hydrophobic bonding.

Binding site interactions

Two types of interaction are used for the theoretical sites: polar and hydrophobic. Polar interactions are defined to include ion-ion, ion-dipole, dipole-dipole, and hydrogen bonding interactions. A polar group for the purposes of the RESIS method is defined as being any oxygen, nitrogen, or sulfur atom.

To account for hydrophobic bonding interactions it is first necessary to arrive at an appropriate definition of hydrophobic groups. At an early stage in the development of the method, emphasis was placed on identifying all such possible hydrophobic regions. The general definitions employed did indeed identify most of the valid hydrophobic groups. However, many groups which were at best questionable for hydrophobic interactions were also identified. Moreover, the generality of the hydrophobic group definitions meant that the average number of hydrophobic groups per compound was rather high. This slowed the searches to the point where it was not possible to use the method on compound libraries with more than one or two hundred compounds.

The hydrophobic group definitions now used in the RESIS method are more specific than those investigated initially. This has shortened the computer search time so that it is practical to process sizable compound libraries, of the order of tens or hundreds of thousands of compounds. It must be recognized, however, that there is a trade-off between an efficient generalized search procedure and the identification of all significant hydrophobic regions. Indeed, deciding which of these regions should be included is to some extent a judgement call. However, examination of a set of compounds (ca. 100) processed according to the current hydrophobe definitions showed that almost

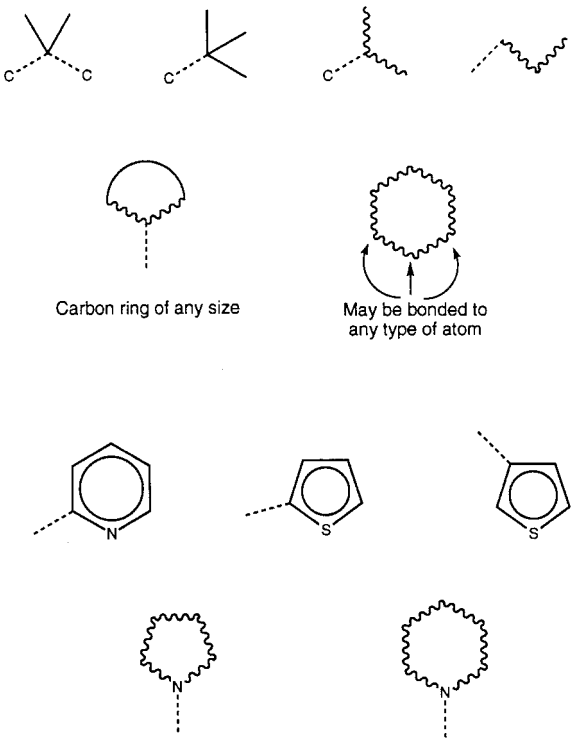

Figure 2. RESIS definitions of hydrophobic groups. Wavy lines indicate bonds of any valid bond order. Dashed bonds are bonds to the rest of the molecule.

all molecular regions that could be reasonably classified as significant hydrophobes with respect to putative drug-receptor interactions were identified. The current RESIS hydrophobes are shown in Figure 2. In that figure, the wavy lines indicate bonds of any valid bond order, while the dashed lines are the bonds to the rest of the molecule. In all the hydrophobes, halogens can be substituted for any of the hydrogens.

The RESIS hydrophobes can be broken down into three classes. The first class, the alkyls, include the gem-dimethyl and t-butyl groups and groups based on the isopropyl and n-propyl skeletons. The gem-dimethyl, t-butyl, and isopropyl groups must be attached to the rest of the molecule through a carbon, while the n-propyl may be attached to any atom. This ensures that the terminal carbons of the groups are at least two carbons away from a heteroatom.

The second class of RESIS hydrophobes is made up of carbon rings. A carbon ring of any size which is attached to the rest of the molecule through just one of the ring atoms is recognized as a hydrophobe. Ring atoms may carry methyl or halogen substituents. The other members of this class consist of six-membered

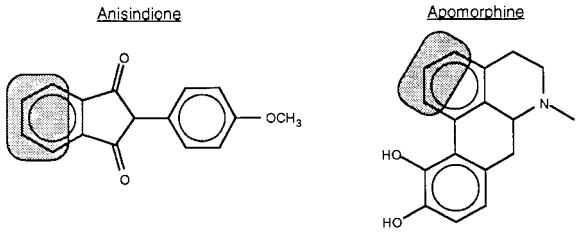

Figure 3. Examples of 'end-ring' hydrophobes.

carbon 'end-rings'. These rings have three consecutive ring atoms which may be bound only to hydrogen, halogen, methyl, and other atoms of the ring. The remaining ring atoms may be attached to any atom. Some examples of end rings appear in Figure 3.

The third class of RESIS hydrophobes consists of heterocycles. They include 2-pyridine, 2- and 3-thiophene, and saturated or partially saturated rings attached through a ring nitrogen such as pyrrolidine and piperidine. Ring atoms may carry methyl or halogen substituents as in the case of the carbon rings.

Software

The Unity 3/DB and Sybyl programs [15] were utilized together with a number of other programs which were written specifically to implement the RESIS method. To create a database to keep track of the compounds which hit each site, programs were written to parse the Unity output files and update the hit data for the queries. Since the Unity 3/DB search program was not set up to perform a series of searches, scripts and programs had to be written to run the series of searches required for the RESIS method. Also the Unity program had a limit to the number of characters for a search query, so it was usually not possible to put all of the hydrophobes in the library into the query. Accordingly, it was necessary to write software to use only the hydrophobes for the compounds in the library subset being searched, as well as software for breaking those hydrophobes into sets small enough for a Unity query.

Results

Characterization of a compound library

To illustrate the application of the RESIS method, studies were conducted using the CIPSLINEPC structural databases which are made up of drug compounds [16]. The Tripos programs SYBYL and UNITY [15] were used to convert CIPSLINEPC compounds into a database which could be searched by UNITY. A library of 7500 of these compounds was then processed by the RESIS system with searches conducted over the range 4 to 12 Å and tolerances of 0.50 Å for each of the three site distances. The search range was chosen after some initial experimentation and reflected a compromise between capturing as many three point interaction sites as possible in a large library of diverse molecules and the need to process such large libraries in a reasonable time frame.

The data for the compounds of particular pharmacological activity classes can be viewed, as shown in Figure 4. The classes of drugs shown are cancer, cardiovascular, central nervous system and endocrine. These four classes are the best represented in the set of 7500 CIPSLINEPC compounds. For each class of compounds three rows of boxes are shown. The top row corresponds to the data for the sites with three polar groups, the middle row for sites with two polar and one hydrophobic group, and the bottom row for sites with one polar and two hydrophobic groups. In each row there are four boxes, each of which is made up of sixteen squares. Each square in each box represents a $2 \times 2 \times 2$ Å cube in site space, which corresponds to one site with a tolerance of one angstrom or up to eight sites with a tolerance of 0.50 Å each. This coarser resolution was chosen to reduce the size of the figures and to give more of an overview of the pattern of hits. The distances increase from bottom to top and left to right, with x- and y-axes corresponding to the distance between the dissimilar groups and the z-axis to the distance between the two similar groups. (For both the second and third row of boxes the x- and y-axes are polar-hydrophobe distances, while the z-axis is the distance between the two similar groups. For the first row, which corresponds to the sites of three polar groups, the x-, y- and z-axes all correspond to polar-polar distances.) The darker a square, the greater the number of hits in its constituent sites. Each activity class has its own (linear) scale, the better to show differences in hit patterns between classes. As can be seen from the figure, there are some differences between the four classes shown. However, since each activity class includes compounds with diverse mechanisms of action this tends to reduce any difference in hit patterns between the classes and to preclude any useful conclusions relating these patterns to therapeutic class. The intent here is to illustrate that different drug compound libraries can be differentiated by the RESIS method and the differences characterized. Of course in addition to the visual representation,

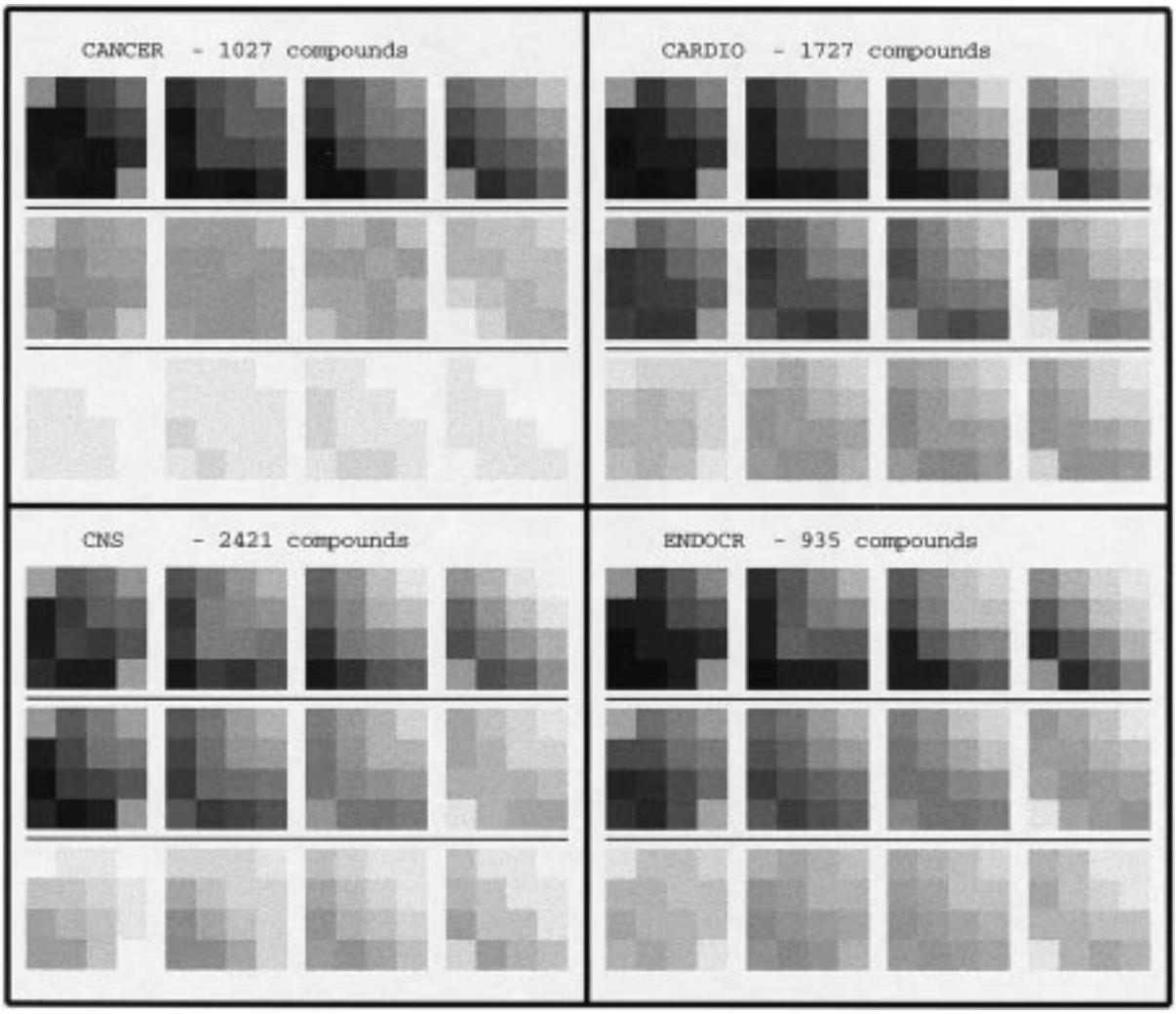

Figure 4. RESIS data for some activity classes of CIPSLINE compounds.

reference can be made to actual numbers of compounds hitting each site for a more precise evaluation of library characteristics as needed.

The data for the whole set of 7500 CIPSLINEPC compounds processed are shown in Figure 5. In this figure each square of each box represents one site with a tolerance of 0.5 Å. As in Figure 4, the rows are for sites of (top to bottom) three polar groups, two polar and one hydrophobic group, and one polar and two hydrophobic groups. The distances increase left to right and bottom to top, and the darker the square the more hits for the corresponding site. The crosses indicate sets of distances which cannot be used to form a three-point site. This figure suggests how the RESIS system can be used to fill in sparsely populated regions of site space to improve the diversity of a compound library. This search could be performed more quickly than a full search over the new library since only those sites which were sparsely populated in the first library will be used in the search over the new library. The new compounds which hit the sparsely populated sites can then be added to the first library.

The RESIS method may also be used to determine which compounds in a library are unlikely to provide active leads for drug discovery programs based on their potential for meaningful receptor site interaction. The user would first need to specify a site-space such that the range of distances would cover those found in most receptors. The minimum distance would be on the order of 2 to 3 Å and the maximum distance could be about 18 to 20 Å. The user would then run the system so that the only site (for each of the

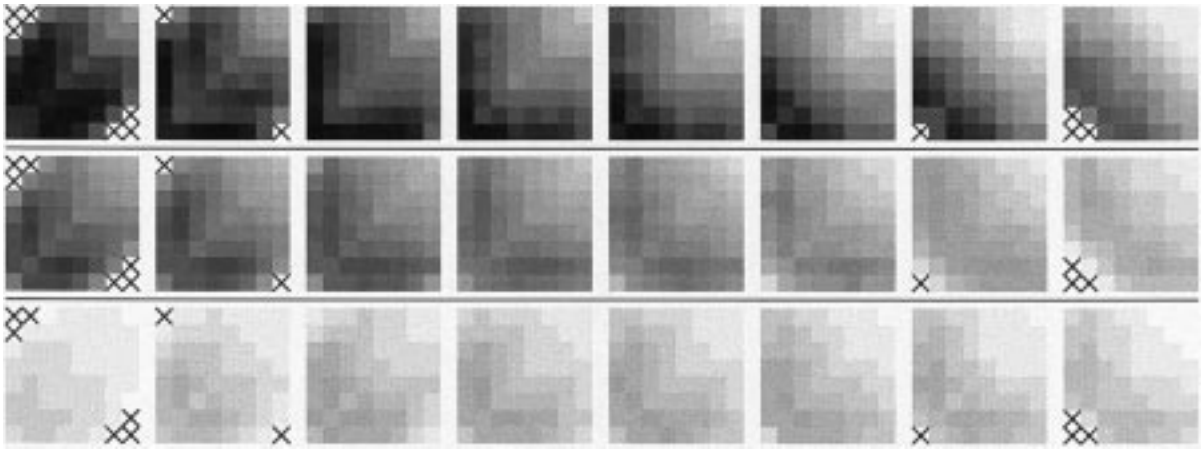

Figure 5. RESIS data for 7500 CIPSLINE compounds.

three site types) which is searched is that which corresponds to the entire site space specified in the first step. The compounds which did not hit the site on any of the three searches could be set aside as being of low priority.

Comparison with an existing two-dimensional method

We were interested in determining how the RESIS method performed in comparison with a recently published two-dimensional procedure for measuring molecular diversity [2], although the fundamentally different approaches of the two methods makes this an imprecise exercise. For this purpose two sets of compounds, consisting of 192 dissimilar compounds (**Dis** library) and 145 similar compounds (**Sim** library) were selected from the Maybridge database (MAY) [17] by a clustering analysis according to the procedure in reference [2].* The RESIS search series for each of those two libraries was performed using a range of 4 to 12 Å and a tolerance of 0.5 Å for each of the three distances in the theoretical sites, which resulted in 679 valid non-redundant sites.

One way of using the RESIS results to measure similarity is to find the number and size of sets of compounds which are identical in the RESIS system. Two compounds are identical in RESIS if they both hit all the same sites. The **Dis** library had three such sets of compounds, while the **Sim** library had four. The numbers of compounds and sites for each set are given in Table 1. A notable feature concerning the numbers presented in Table 1 is that the sets of compounds from the **Sim** library hit significantly more sites. One would expect that a library of dissimilar compounds would have fewer sites per set because those compounds would individually hit a greater variety of sites and thus those that hit more than a few sites would be unlikely to have hit the same sites as any other compound.

Table 1. Sets of RESIS-identical compounds

Dis		**Sim**	
# of compounds	# of sites	# of compounds	# of sites
3	1	4	8
3	1	3	4
2	1	3	3
–	–	2	9

On the other hand, compounds of a library of similar compounds would be expected not to have such a variety of sites hit by the individual compounds, making it easier to find sets of RESIS-identical compounds which hit larger numbers of sites. Figure 6 shows the compounds of the three sets of **Dis** compounds, while Figure 7 shows the compounds of one of the sets of **Sim** compounds. A factor here could be the number of rotatable bonds in the sets of compounds from the **Dis** and **Sim** libraries. The number of hits from a compound would be influenced by the number of such bonds. However, few are present in the sets of compounds in question and there appears to be no significant bias in this respect between the compound sets from the **Dis** and **Sim** libraries.

* We are indebted to Dr. J. Dunbar, Parke-Davis Pharmaceutical Research Division, Warner-Lambert Company, for providing these.

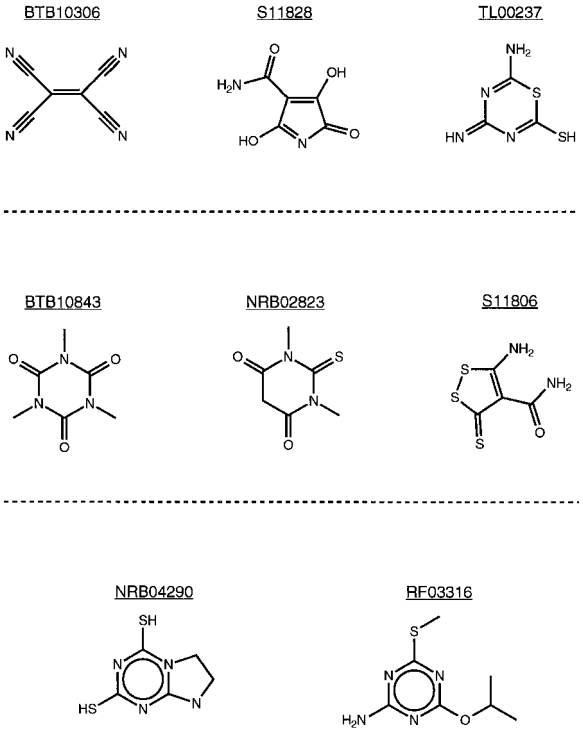

Figure 6. Three sets of RESIS-identical **Dis** compounds.

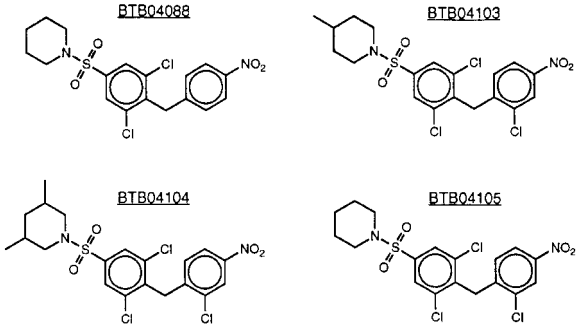

Figure 7. A set of RESIS-identical **Sim** compounds.

Another way of using data from a RESIS search series to assess similarity is to look at the number of sites which were hit by compounds. It would be expected that a set of similar compounds would hit fewer sites than would a set of dissimilar compounds, since the compounds of the similar set would be less diverse. This is seen in the data for the **Sim** and **Dis** search series. The 63 **Sim** compounds which hit at least one site, collectively hit 295, or 43.4%, of all the valid sites, while the 68 **Dis** compounds which hit sites, collectively hit 389, or 57.3% of the total sites.

One can also evaluate compound similarity by looking at sets of non-overlapping compounds. Such

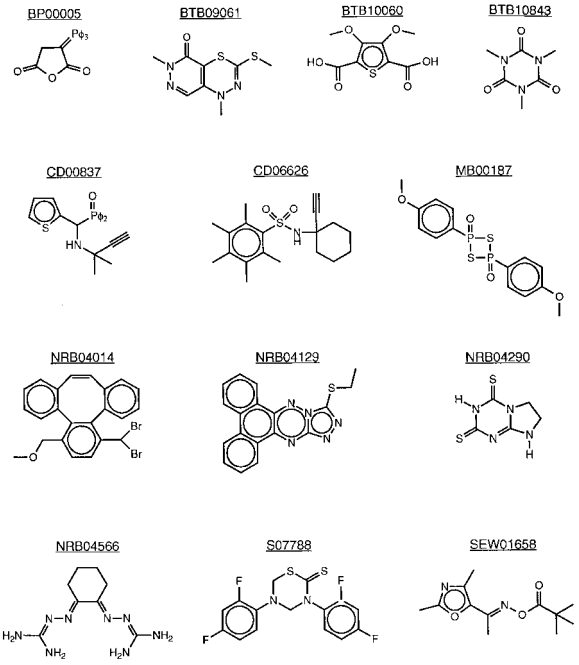

Figure 8. A set of non-overlapping **Dis** compounds.

sets consist of compounds which hit none of the sites hit by any of the other compounds in the set. The largest sets of non-overlapping **Dis** compounds had 13 compounds, while the largest such set of **Sim** compounds had only 8 compounds. Again, this is as one would expect since dissimilar compounds are less likely to overlap. A set of 13 non-overlapping **Dis** compounds is shown in Figure 8, while a set of 8 non-overlapping **Sim** compounds is shown in Figure 9.

The preceding comparative analysis illustrates the degree of consistency between the RESIS method and a recently published [2] two-dimensional approach with regard to the assessment of molecular similarity and dissimilarity. In some respects the results are compatible, however, there are significant differences and the three-dimensional receptor site oriented basis of the RESIS method provides additional characterization of, and insight into, molecular diversity. This can be seen, for example, in the sets of RESIS-identical **Dis** compounds (Figure 6) and the sets of non-overlapping **Sim** compouds (Figure 9). It has been observed [5] that the clustering protocol used in the two-dimensional method [2] is not very effective in grouping compounds in relation to their biological activity. We believe that the evaluation of molecular diversity afforded by the RESIS system may correspond well with the way receptor sites view diversity.

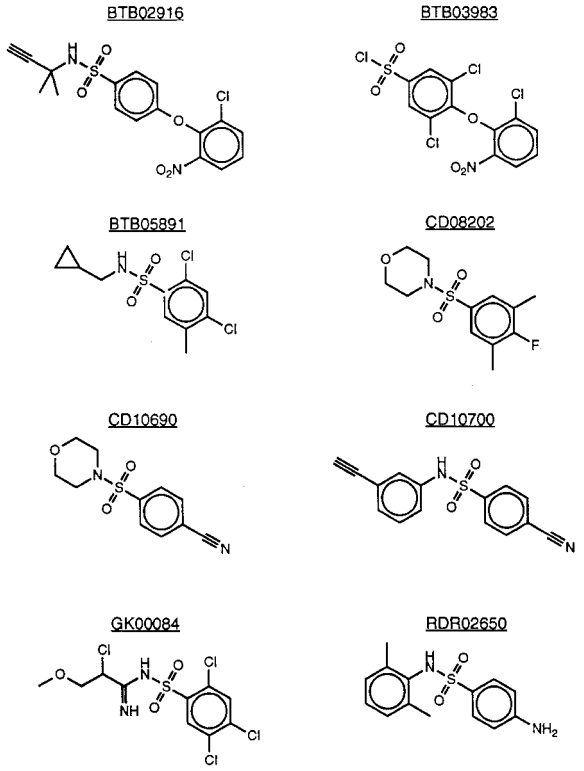

Figure 9. A set of non-overlapping **Sim** compounds.

Comparison with other 3-D methods

The ChemDiverse [9] and PDQ [10] methods are similar in concept to the RESIS method. All three methods are based on a three point interaction model. A detailed comparison of the ChemDiverse and PDQ approaches has been reported [10]. Both employ a wider range of center types than the RESIS method which utilizes two broad categories, hydrophobic and polar. The PDQ method subdivides these into more categories; hydrophobe and aromatic centroid, and hydrogen bond donor/acceptor, acid and basic centers, for a total of six in all. The ChemDiverse method uses these six plus, as of its October 1995 release, one more. Compared to the RESIS method this greater level of discrimination of bonding types in the PDQ and ChemDiverse methods provides a more detailed picture of potential drug-receptor site bonding interactions. However, significant redundancy occurs in considering separately sites termed hydrophobe and aromatic centroid where hydrophobe is an aliphatic or cycloalkyl species, since in many cases a hydrophobic drug-receptor interaction can occur where the hydrophobic region on the drug is either aliphatic/cycloalkyl or aromatic. Non-redundant combinations of center types totalled 3, 56 and 84 for the RESIS, PDQ and ChemDiverse methods respectively. In the RESIS method consideration was given to which combinations are likely to constitute the three most important sites for a particular drug-receptor interaction. On this basis the three hydrophobe combination was eliminated. This was not done for either the PDQ and ChemDiverese methods so that such improbable combinations as three hydrophobes, three aromatic centroids, three acid centers, three basic centers etc. were retained.

Different distance intervals between interaction centers were also employed by the respective methods. The RESIS method uses 0.5 Å intervals, the PDQ method employs coarser distance increments, ranging from 2.5 Å at shorter distances to 5.0 Å at the longer distances, and the ChemDiverse method uses the finest distance resolution at increments increasing progressively from 0.1 to 1.0 Å. The broader distance ranges employed in the PDQ method limits the degree of distance discrimination for the pharmacophores whereas it has been pointed out [10] that some of the additional information through the higher resolution of the ChemDiverse descriptor may not be significant. The RESIS approach is intermediate betwen these. The way RESIS performs searches – searching the whole space and then dividing that space up in successive steps - means that the distance intervals can be readily modified in this method. The distance range covered can be adjusted in all three methods.

The characterization of the molecular diversity of a compound library by each of the three methods will be dependent on the combination of center types and distance intervals employed. Compared to the PDQ method, RESIS has a more elementary breakdown of center types but uses a finer distance differentiation. ChemDiverse employs both a full range of center types and the most extensive set of distance intervals at the cost, however, of greatly increased complexity and computational demand and some redundancy in the information generated. The three methods, which all share a basic main principle, may be viewed as offering different trade-offs between the extent and detail of the information obtained and its accessibility and the resulting level of complexity and resources required. The RESIS method offers relative simplicity while still providing a useful level of differentiation between compound libraries with respect to molecular diversity.

Conclusions

The RESIS method gives a three-dimensional measure of the molecular diversity of a compound library which can be conveniently displayed as a density map providing a readily comprehensible visual impression of the library diversity characteristics. The method, which is based on simulated receptor site interactions according to a three-point site model, relates directly to the use of compound libraries for screening against various receptor targets. It has been illustrated by application to a compound library of 7500 compounds and could potentially be used on much larger libraries of up to 100 000 compounds.

The method can be applied to efficiently identify compounds which will enhance the molecular diversity of a compound library and can easily be used to identify compounds which may not be of any interest in drug design.

A comparative analysis showed that the RESIS method is by some measures consistent with a recently published two-dimensional approach. However, there are important differences in the results obtained arising from the three-dimensional format of the RESIS method which relates more directly to drug-receptor interactions. It differs from two other three-dimensional methods in utilizing a much simpler classification of putative drug-receptor bonding interactions. While providing less detailed information it has greater simplicity of use and provides an easily visualized measure of the molecular diversity of compound libraries.

Acknowledgements

The authors are indebted to the Chiron Corporation and the Parke-Davis Pharmaceutical Research Division of the Warner-Lambert Company for generous financial support.

References

1. Martin, E.J., Blaney, J.M., Siani, M.A., Spellmeyer, D.C., Wong, A.K. and Moos, W.H., J. Med. Chem., 38 (1995) 1431.
2. Shemetulskis, N.E., Dunbar, J.B., Dunbar, B.W., Moreland, D.W. and Humblet, C., J. Comput.-Aided Mol. Design, 9 (1995) 407.
3. Holliday, J.D., Ranade, S.S. and Willett, P., Quant. Struct.-Act. Relat., 14 (1995) 501.
4. Turner, D.B., Tyrell, S.M. and Willett, P., J. Chem. Inf. Comput. Sci., 37 (1997) 18.
5. Brown, R.D. and Martin, Y.C., J. Chem. Inf. Comput. Sci., 36 (1996) 572.
6. Good, A.C. and Kuntz, I.D., J. Comput.-Aided Mol. Design, 9 (1995) 373.
7. Boyd, S.M., Beverley, M., Norskov, L. and Hubbard, R.E., J. Comput.-Aided Mol. Design, 9 (1995) 417.
8. Chapman, D., J. Comput.-Aided Mol. Design, 10 (1996) 501.
9. Chem DBS-3D/Chem Diverse, developed and distributed as part of the Chem-X modeling package by Chemical Design Ltd., Roundway House, Cromwell Park, Chipping Norton, Oxfordshire, OX77SR, U.K.
10. Pickett, S.D., Mason, J.S. and Mclay, I.M., J. Chem. Inf. Comput. Sci., 36 (1996) 1214.
11. Lewis, R.A., Good, A. and Pickett, S.D., In van de Waterbeemd, H., Testa, B. and Folkers, G. (Eds) Computer-Assisted Lead Finding and Optimization, Wiley-VCH, Weinheim, 1997, pp. 137–156.
12. Lewis, R.A., Mason, J.S. and McLay, I.M., J. Chem. Inf. Comput. Sci., 37 (1997) 599.
13. Bradley, M., Richardson, W. and Crippen, G.M., J. Chem. Inf. Comput. Sci., 33 (1993) 750.
14. Farmer, P.S., In Ariens, E.J. (Ed.) Drug Design, Vol. X, Academic Press, New York, NY, 1980, p. 133.
15. Tripos Associates Inc., St. Louis, MO, 1994.
16. CIPSLINEPC Structural Databases comprising compounds taken from the Prous Science publications *Drugs of the Future* and *Drug Data report*. J.R. Prous Publishers, Barcelona, Spain.
17. Maybridge 94 Database, Daylight Chemical Information Systems.