

Selection of Molecules Based on Shape and Electrostatic Similarity: Proof of Concept of “Electroforms”

Andy Jennings*,† and Mike Tennant‡

Takeda San Diego, 10410 Science Center Drive, San Diego, California 92121, and Syrrx, Inc., c/o Takeda San Diego, 10410 Science Center Drive, San Diego, California 92121

Received December 11, 2006

Molecular shape and electrostatic distribution play a crucial role in enzyme and receptor recognition and contribute extensively to binding affinity. Molecular similarity and bioisosterism are much-discussed topics in medicinal chemistry. Many molecular representations and similarity metrics are available to help drug discovery, and activities such as compound hit explosion and library design can be undertaken using them. The quality of the resulting compound series is highly dependent upon the molecular representation and similarity metric used. We have used a range of software to investigate whether molecules can be represented and compared effectively using measures of three-dimensional shape and electrostatic distribution (“electroforms”). We find that these descriptors allow for the assessment of molecular similarities using standard molecular visualization tools and offer a method for comparing molecules that may be considered superior to other methods.

INTRODUCTION

Molecular shape and electrostatic distribution play a crucial role in enzyme and receptor recognition and contribute extensively to binding affinity. To a first approximation, molecules with similar shapes and electrostatic distributions should bind to receptors in a similar manner, and a consequence of this similarity is reflected in the development of structure–activity relationships.¹

Molecular similarity^{2,3} and the search for bioisosteres^{4–6} is a much-discussed topic in medicinal chemistry. Many molecular representations and similarity metrics are available to help drug discovery, and activities such as compound hit explosion and library design can be undertaken using them. The quality of the resulting compound series is highly dependent upon the molecular representation and similarity metric used, and this ‘quality’ is often somewhat subjective. For example, a molecular representation as simple and abstract as that described by pharmacophores should be able to capture the gross details of a molecule but should not be expected to capture the finer details that may be essential for lead optimization, for example, specific delocalization effects. Alternatively, abstract representations run the risk of classifying compounds as (dis)similar for no obvious reasons.

We have used the Omega,⁷ Shape,^{8,9} and EON^{10,11} software developed by OpenEye Scientific Software¹² to investigate whether molecules can be represented and compared effectively using measures of three-dimensional shape and electrostatic distribution. We find that these descriptors allow for the assessment of molecular similarities using standard molecular visualization tools. We have compared the method to the industry standard fingerprinting method provided with the Daylight¹⁸ package and find that the method here provides a finer-grain description of similarity than does the Daylight method.

METHODS

The shape similarity between molecules A and B can be determined, in part, by comparing the shapes of those molecules. This effectively reduces to calculating the overlap volume between two molecules and has been detailed elsewhere.⁸ Once the overlap has been optimized the similarity between molecules A and B can be calculated using the Tanimoto equation

$$\text{Tanimoto}_{A,B} = \frac{O_{A,B}}{I_A + I_B - O_{A,B}}$$

where $O_{A,B}$ is the overlap between molecules A and B, and I_A and I_B are the self-overlap for molecules A and B, respectively.

A normalized shape Tanimoto of unity indicates that the molecules are identical, and the Tanimoto tends to zero as molecules become less similar.

The ROCS (Rapid Overlay of Chemical Structures) package⁸ provides a fast and accurate¹³ method for superimposing molecules and has proved useful in a recent drug discovery program.¹⁴ However, ROCS does not contain an accurate notion of charge distribution and therefore is not a complete solution to the search for molecular similarity.

Electrostatic potentials of the two molecules can be effectively compared using a Tanimoto metric used in an analogous manner to that described for shape comparison but applied to the electrostatic fields of the two molecules.¹²

The electrostatic field of molecules A and B can be calculated and the similarity between fields expressed as the electrostatic Tanimoto

$$\text{Tanimoto}_{A,B} = \frac{\int A(\vec{r}) * B(\vec{r})}{\int A(\vec{r}) * A(\vec{r}) + \int B(\vec{r}) * B(\vec{r}) - \int A(\vec{r}) * B(\vec{r})}$$

where $\int A(\vec{r})$ is the electrostatic field of A.

* Corresponding author e-mail: andy.jennings@takedasds.com.









† Takeda San Diego.

‡ Formerly Syrrx, Inc., c/o Takeda San Diego.

Table 1. Common Linker Groups Examined in This Work^a

molecule name		
1-methyl-1H-imidazole	imidazole	oxazole
N-methylacetamide	acetic acid ethyl ester	N,N-dimethylacetamide
N-ethyl-N-methylacetamide	acetic acid methyl ester	

^a See Chart 1 for molecular structures.**Chart 1.** Structures and Names of Linkers Groups Used in This Work^a

			
Acetic acid methyl ester (1 conformation)	Acetic acid ethyl ester (1 conformation)	N-Methyl-acetamide (1 conformation)	N,N-Dimethyl-acetamide (1 conformation)
			
N-Ethyl-N-methyl-acetamide (4 conformations)	Oxazole (1 conformation)	Imidazole (1 conformation)	1-Methyl-1H-imidazole (1 conformation)

^a The figures in brackets indicate the number of conformers found using Omega⁷ with a 0.6 Å root-mean-square (rms) cutoff and an energy window of 5 kcal/mol. Information supporting these compounds as isosteres is contained in Accelrys' BIOSTER DB: http://www.accelrys.com/products/chem_databases/databases/bioster.html (accessed April 05, 2007).

The electrostatic Tanimoto ranges from −0.3 to +1 (see ref 12 for specific algorithm details), indicating dissimilar and similar electrostatic fields, respectively.

Electrostatic Tanimoto (ET) scores can be calculated using the EON program. Molecules are superimposed using ROCS and electrostatic potentials calculated using Openeye's ZAP Poisson–Boltzmann solver. Two ETs can be calculated using an external dielectric of 2 and of 80. The latter accounts for dampening of the electrostatic field by aqueous solvent and may better represent the field experienced upon binding to a protein.

A good charge model is essential to the accuracy of the electrostatic potential calculation, and we have employed the AM1-BCC method of Bayly¹⁵ as implemented in quacpac.¹² This method is extremely rapid and has been shown to be robust and correlate highly with HF/6-31G* fitted charges for a large number of organic compounds. We preferred this over the MMFF94 charge model¹⁶ as the AM1-BCC method

takes the charge-dependency of conformers into account when assigning charges, whereas MMFF94 uses conformer-independent templates to assign charges to molecules.

DETAILS OF CALCULATIONS

We explored two data sets: a set of eight commonly used linker moieties (Table 1) and a set of 60 single cycle/fused bicyclic rings taken from the Maybridge heterocyclic ring-numbering chart,¹⁷ shown in Table 2.

The 2H-tautomer of indole was added to the Maybridge set. Formal charges were assigned by hand where appropriate. Conformations of the molecules were generated in Omega. An energy window of 5 kcal/mol and a duplicate removal rmsd of 0.6 Å was used. Atomic charges were calculated for each conformation using the single-point AM1-BCC method implemented in quacpac, and these charged molecules were used in all subsequent stages of the analysis. The charged conformers were concatenated into a single database and superimposed in an all-vs-all manner using ROCS with chemistry optimization turned on, which biases the overlay by adding a positive weight to similar chemical groups, and all hits are reported. Each overlay was scored with EON using a terminal rotor spin value of 30 degrees, chosen to minimize the computational overhead. The best pairwise scores were written to a square matrix, and the similarity score between conformations *i* and *j* was set to the maximum value of the score between either *i,j* or *j,i*. We found that this was necessary as ROCS seems sensitive to input order and results in occasional slight asymmetric differences in the overlays. We chose to set the score to the maximum value as we are interested in the 'best' overlay possible from these methods, which is conventional in these types of studies. An alternative approach would be to select the median score for any *i,j* comparison. However, for our Maybridge data set the number of alternative overlays between pairs of structures was small due to the planarity and rigidity of the molecules, and consequently the spread of scores derived from these alternative overlays was limited. Hence, we lack the data to investigate this alternative and feel that the decision to choose the best possible overlay was the most sensible choice.

As stated in the methods section, the data were reranged from −0.3 to +1.0 to 0.0 to +1.0, and five similarity matrices

Table 2. Compounds Represented by the Maybridge Identifier^a

rings	ring names (numerical identifiers)			
1	furan (1)	1,2,3-oxadiazole (19)	thiazole (10)	morpholine (28)
	thiophene (2)	1,2,3-triazole (20)	imidazole (11)	1,4-dithiane (29)
	2H-pyrrole (3)	1,3,4-thiadiazole (21)	2-imidazoline (12)	thiomorpholine (30)
	pyrrole (4)	benzene (22)	imidazolidine (13)	pyridazine (31)
	2-pyrroline (5)	2H-pyran (23)	pyrazole (14)	pyrimidine (32)
	3-pyrroline (6)	4H-pyran (24)	2-pyrazoline (15)	pyrazine (33)
	pyrrolidine (7)	pyridine (25)	pyrazolidine (16)	piperazine (34)
	1,3-dioxolane (8)	piperidine (26)	isoxazole (17)	1,3,5-triazine (35)
	oxazole (9)	1,4-dioxane (27)	isothiazole (18)	1,3,5-trithiane (36)
	indolizine (37)	1H-indazole (44)	isoquinoline (51)	quinuclidine (58)
	indole (38)	2H-indazole (45)	cinnoline (52)	indene (59)
	isoindole (39)	benzimidazole (46)	phthalazine (53)	azulene (60)
	3h-indole (40)	benzthiazole (47)	quinazoline (54)	norbornane (61)
	indoline (41)	purine (48)	quinoxaline (55)	
2	benzofuran (42)	4H-quinazoline (49)	1,8-naphthyridine (56)	
	benzothiophene (43)	quinoline (50)	pteridine (57)	

^a The numbers in parentheses refer to Figure 7. See ref 17 for molecular structures.

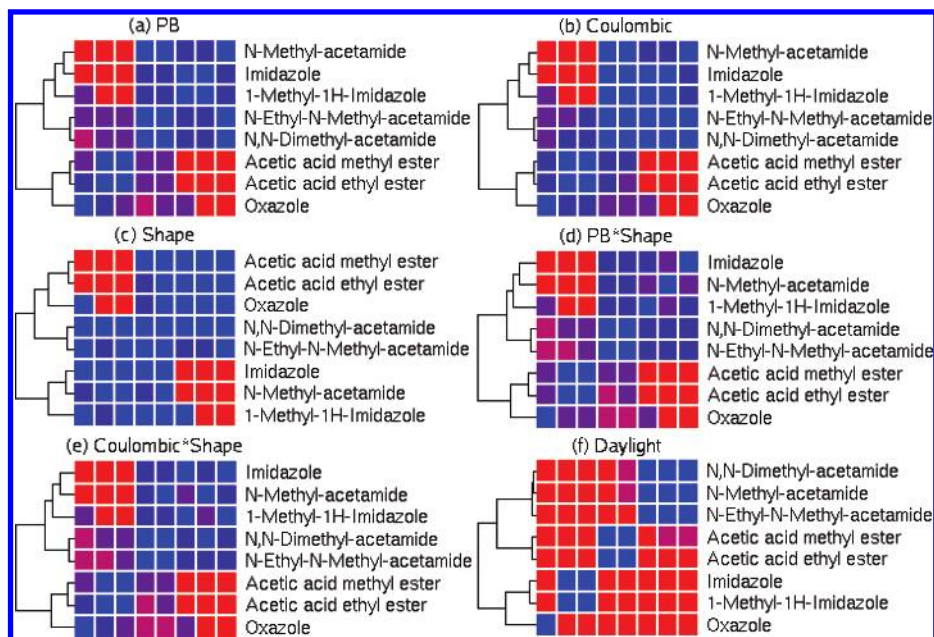


Figure 1. Color-coded Tanimoto similarity matrices for the linker compounds. Blue indicates a similarity of 1.0, red a similarity of 0.0. Parts (a)–(c) are Tanimoto similarity matrices for the individual representations, PB electrostatics, Coulombic electrostatics, and shape, respectively. Parts (d) and (e) are Tanimoto similarity matrices for the products of PB electrostatics and shape and Coulombic electrostatics and shape, respectively. Part (f) is the analogous Tanimoto similarity matrix using Daylight fingerprints as the molecular descriptor.

were ultimately output. These represented for the best overlay of a conformation from molecules A and B (1) the shape similarity; (2) the electrostatic similarity using a Poisson–Boltzmann (PB) electrostatic model; (3) the electrostatic similarity using a Coulombic electrostatic model, (4) the similarity using the product of shape and the PB model; and (5) the similarity using the product of shape and the Coulombic model.

To provide a comparison between our method and a much-used standard similarity method, analogous calculations were also performed using Daylight fingerprints^{18,19} and the corresponding Tanimoto metric. These fingerprints are a rapid and robust method of profiling molecules based upon bond paths and are used extensively in computational chemistry to identify (dis)similar molecules to a query. Given that it has no explicit concept of molecular shape one can appreciate its limitations when applied to molecules with distinct 3-D conformations.

Clustering of the dissimilarity matrices was performed using the statistical software R,²⁰ with the agnes method in the ‘cluster’ package.²¹ Agglomerative clusters were generated using the single-linkage and Ward’s methods. Visualization of the cluster trees was performed directly in R. Maximum similarity matrices were viewed using in-house code and the heatmap() function in R.

DISCUSSION

The ability to accurately predict the similarity or dissimilarity between molecules is crucial for SAR exploration activities in medicinal chemistry projects. A large number of commonly used molecular descriptors can result in a set of ‘similar’ compounds that have no physical significance, and hence it is difficult to explain why particular molecules are similar, beyond the trite explanation that ‘they look the same’. In this work we have used physically reasonable quantities, shape, and electrostatics, to compare molecules. These descriptors can usually be directly related to a compound’s relative ability to bind to a protein and illicit a

biological response¹ and can be viewed with simple visualization tools in order to explain (dis)similarity.

We have calculated various similarity scores for the best overlay of two sets of small molecules in an attempt to ascertain whether explicit shape and electrostatic-based descriptors could provide an accurate way to classify molecules into similar subsets. The first set of molecules was composed of a small number of commonly used linker groups. The second set was composed of either one or two fused ring systems and no rotatable bonds: the only conformational freedom was associated with ring flips and rotational and translational degrees of freedom. These molecules were chosen to validate the methodology as they provide a simple and intuitive test set.

ANALYSIS OF COMMONLY USED LINKER SET

A commonly used set of 8 linkers was constructed in order to explore whether the electrostatic/shape method could classify simple commonly used linker groups as least as well as Daylight fingerprints. The standard Daylight fingerprint comparison algorithm is presented such that different atom codes are not matched, and our linker set is contrived to emphasize that fact. For example, under the Daylight scheme the imidazole molecules should not match the acetamide molecules well due to the different coding of the aromatic and aliphatic atoms, respectively. However, we know that these molecules are generally bioisoteric and would hope that the EON method would capture some similarity due to the inherent shape and electrostatic similarity between the molecules. Although our method scales as N^2 , these molecules have very little conformational freedom, and, given the small data set, the calculations are relatively rapid, taking under 30 s to run the suite of calculations. We used a color-coded similarity matrix to rapidly identify similar compounds, (Figure 1). This method identifies self-matches consistently and accurately, with all self-scores being unity except *N,N*-dimethylmethanamide (0.986). This small dis-

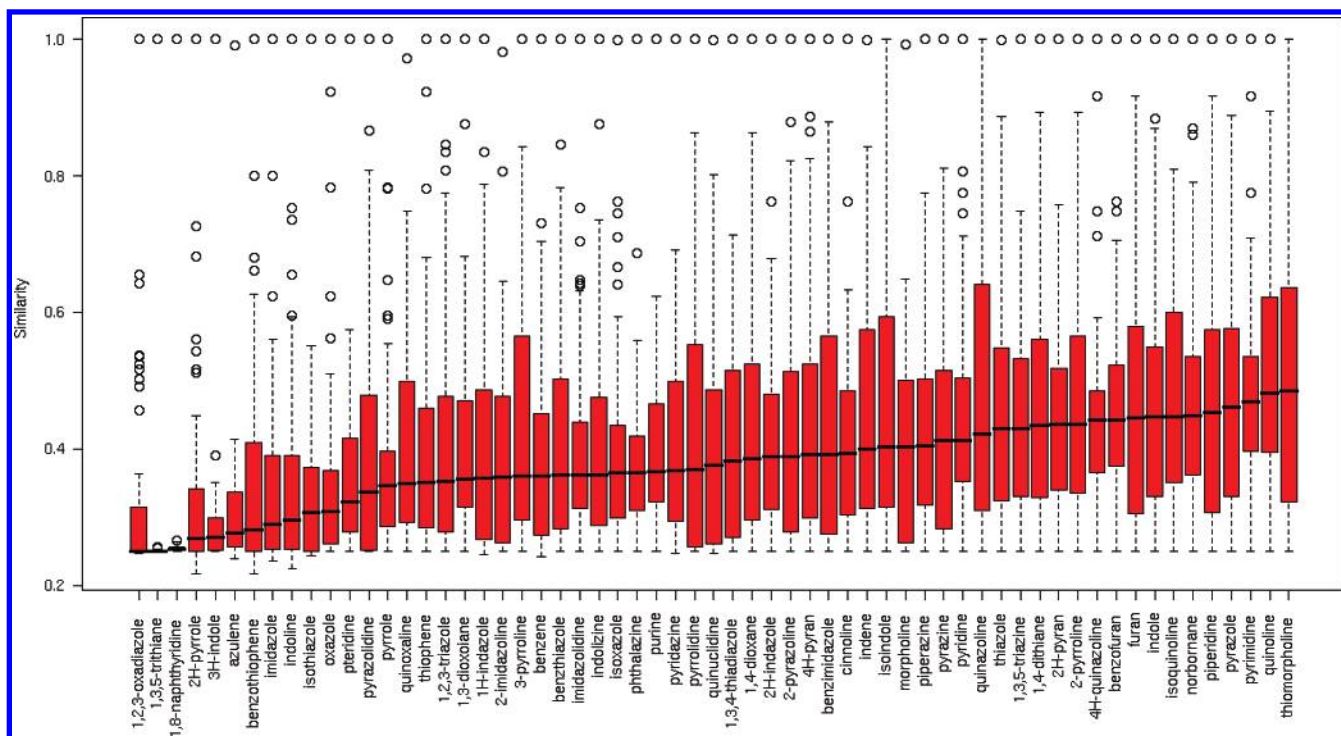


Figure 2. Tukey boxplot for the Maybridge compounds. The plot represents the PB*Shape similarities of each compound with all other compounds in the data set. The bold line in the middle of each box is the median similarity. The lower and upper extremes ("hinges") of the red boxes are the first and third quartiles of the data, respectively. Whiskers are drawn at 3/2 times the interquartile range from the median, and data points lying outside of these whiskers are considered outliers. In this case the outliers are, in fact, the few compounds with high similarity to the query compound.

crepancy is probably due to a slightly offset superimposition by ROCS, to which EON is very sensitive.

These similarity plots indicate that shape alone is not sufficient to distinguish between molecules (Figure 1(c)). The ET matrices (Figure 1(a), Poisson–Boltzmann and Figure 1(b), Coulombic) provide better resolution, but the product of ET*Shape gave the best resolution (Figure 1(d),(c)). There is minimal difference between the PB and Coulombic electrostatic Tanimoto matrices. We used these findings to focus all further studies on the PB*Shape matrices: we felt that the rationale for using the PB electrostatic measure was stronger as we were interested in aqueous-based phenomena.

We were interested in comparing the PB*Shape method to a classification using path length fingerprints as the descriptor. Path length fingerprints are standard methods used extensively in drug discovery, and we would expect that the EON method should perform as least as well in order to be of use. Figure 1(f) shows the maximum similarity matrix calculated using Daylight fingerprints. This method effectively classifies the data set into two subclusters: rings and nonrings, with no linkage between the subclusters. Contrast this with the PB*Shape matrix, where similarities between the sets are apparent, for example between oxazole and acetic acid ethyl ester; imidazole; and *N*-methylacetamide. These are similarities that are exploited in SAR development. The results from this small data set show that the methods used here can be applied to classifying similar small linkers. Using the EON method we are able to pick up useful similarities that would have been missed using a standard path-descriptor method. We subsequently applied the same methodology to the larger Maybridge set.

ANALYSIS OF MAYBRIDGE DATA SET

The Significance of the PB*Shape Score. We visually assessed the overall similarities between the molecules using the Tukey boxplot statistic (Figure 2). The general level of similarity is about 0.2, with few significant outliers. These outliers indicate molecules that are significantly similar, according to the Tukey statistic, to the query molecule. The Maybridge data matrix is sparse with respect to similar molecules and consequently contains a number of singletons, which may pose problems for accurate clustering.

To further understand the significance of the similarity scores, we ran the above protocol using a larger number of randomly chosen molecules containing either one or two ring systems. For the generated superimposition of each molecular pairwise comparison, similarities were transformed into 'score densities' within R. Score densities for the molecules aligned and scored against themselves ('self-score'), and molecules aligned and scored against different molecules ('nonself-score') are presented in parts (a) and (b), respectively, of Figure 3. These plots indicate that, for a molecule match against a different molecule, the PB*Shape density peaks at a score of 0.2 (Figure 3(b)). The densities observed when superimposing molecules against themselves show three major peaks at a PB*Shape score of 0.4, 0.6, and 0.8. The low end of the scores represents residual core overlap: the volume of two molecules that will overlap regardless of explicit atomic detail. The discretization of the scores is a reflection of the few degrees of freedom these molecules have during alignment and may not be observed as the number of rotatable bonds in a molecule increases. These two plots indicate that a score of less than 0.4 is probably a random match for the PB*Shape metric. Scores between 0.4

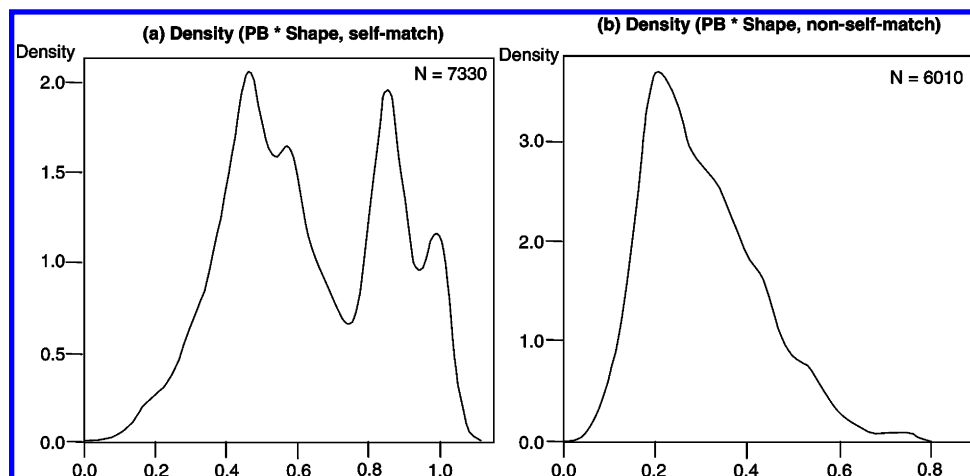


Figure 3. Density plots showing the PB*Shape similarities between randomly selected mono- and bicyclic ring systems. Part (a) shows that, for molecules aligned and scored against themselves (“self-score”) ($N = 7330$), there are peaks at similarity levels 0.4 and 0.6. For molecules aligned and scored against other molecules ($N = 6010$) there is a peak at a similarity of 0.2 (b), but there is a long tail extending up to 0.6. This indicates that the background similarity for small ring systems is centered around 0.2, but there is significant overlap between the “self-scored” true positive systems up to 0.5–0.6.

and 0.6 could indicate a true match, but there is significant density in that region for the nonself-molecule matches, and so false positives should be suspected. We consider that a PB*Shape score of 0.6 or greater indicates a positive match. These findings support the data seen in the boxplot for the Maybridge molecule set.

THE EFFECT OF DIFFERENT CLUSTERING METHODS

The normalized dissimilarity matrices were clustered using the single-linkage (nearest-neighbor) and Ward’s agglomerative methods implemented in the ‘agnes’ package in R. Single-linkage clustering assigns two points as being in the same cluster if the distance between them is smaller than the distance to any other point. Single-linkage clustering tends to result in long, or stringy, clusters. Ward’s method clusters points by minimizing the sum of square errors within clusters compared to the cluster mean. This method tends to produce small, spherical, clusters. We find here that single-linkage clustering provides a representation that fits with our intuitive view of the sparse chemistry space that is presented here. This method may not be useful if larger and less sparse data sets are used but should not be immediately discounted.²² Conversely clustering with Ward’s method sometimes leads to fragments clustering that are not intuitively chemically similar. Other methods implemented in the `hclust()` function in R were investigated but are not reported here.

These methods produce dramatically different clusters of the dissimilarity data, as shown in Figure 4(a),(b). Figure 4(a) shows the dissimilarity data clustered using the single-linkage method. There is some structure in the clusters and a large number of singletons. Intuitively this makes sense, as the similarities between molecules are generally low. The data clustered using Ward’s method (Figure 4(b)) show compact clusters, as is characteristic of this method.

In an attempt to evaluate and rationalize these methods we plotted a color-coded similarity matrix using the best score for molecule A against molecule B, as shown in Figure 5. A PB*Shape score of 1.0 is blue, and the color tends to red as PB*Shape tends toward zero. It is not entirely correct to compare the maximum similarity method to the tree method, as they are effectively one- and two-dimensional

representations, respectively. However, we can gain some clarity by comparing them, especially with regards to some of the outliers. Figure 5(b) shows the same color matrix with the scores binned for clarity: 0.0–0.4, red; 0.4–0.6, green; and 0.6–1.0, blue.

We can see from Figure 5 that norbornane, 1,3,5-triazine, and 1,2,4-trithiane have no high scores with other molecules and can be classed as outliers in this data set. These three molecules are represented as extreme outliers in the single-linkage tree but are forced into nonintuitive clusters with Ward’s method.

The next three outliers in the single-linkage tree, pteridine, quinazoline, and quinoxaline, do have a small number of similar neighbors, as indicated in Figure 5, and they are clustered with similar structures in the Ward’s tree (Figure 4(b)).

The single-linkage method contains some structure where the molecules are intuitively similar, for example pyrrolidine/piperidine, but these clusters are sparse. The Ward’s method, however, can cluster molecules that are not similar, for example the outlier norbornane with 1,3,5-triazine and pyrimidine.

We believe the threshold value for deciding cluster inclusion for the single-linkage method does not allow for large enough clusters, and too many singletons are present. This could be an artifact of the sparse data set we are exploring and could possibly be remedied using a larger, more similar, data set. Alternatively a different clustering method may be more appropriate, although we did not see any more intuitively appropriate clusters generated with any of the algorithms available in the `hclust()` package of R. One possible use of the single-linkage method would be to undertake ‘hole-filling’ SAR exercises, where compounds that fill gaps between singletons and clusters are exploited.

The Ward’s method, on the other hand, results in clusters that are sometimes too large, and where molecules that are true singletons are not treated as such. This could result in SAR clusters containing a number of ‘false positive’ structures. With this caveat, we feel that a good use of this method would be to explore and expand on close SAR trends within well-defined structural clusters.

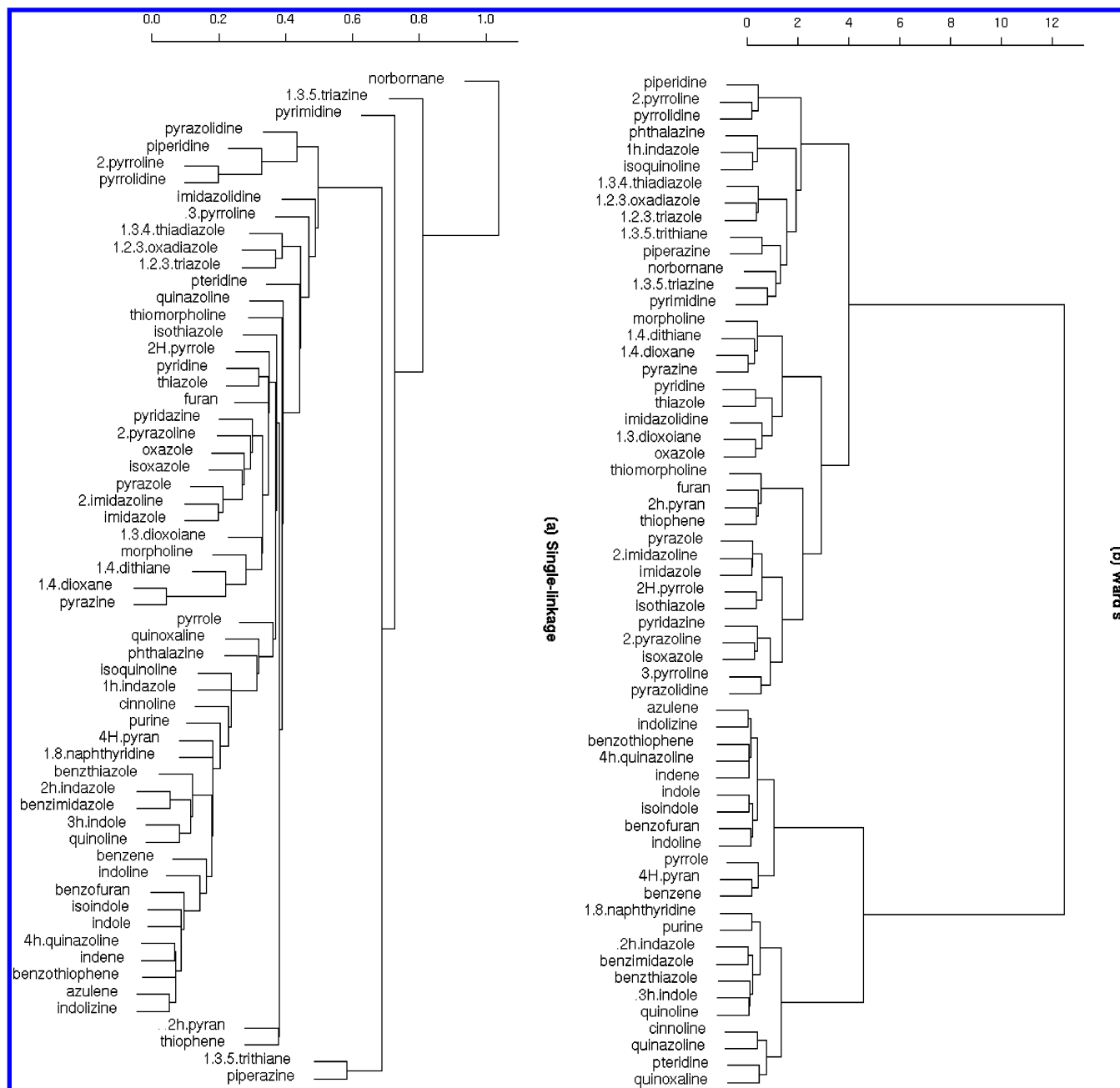


Figure 4. The Maybridge PB*shape similarity scores clustered using single-linkage (a) and Ward's (b) methods. The single-linkage method results in long thin clusters. The Ward's method describes the data as compact spherical clusters.

A COMPARISON OF EON AND DAYLIGHT FINGERPRINT SIMILARITIES

To quantify the limitations of 2-D fingerprints, such as those implemented in Daylight, we calculated the Tanimoto similarity tables for the Maybridge set using the Daylight fingerprints as descriptors. The maximum similarity matrix is shown in Figure 6. This is analogous to that created using the EON similarities (Figure 5) and was used to visually compare the results of using the two methods. It is immediately obvious that the EON method used here captures similarities that the 2-D fingerprint method misses. There are numerous examples of the failure of 2-D fingerprints to capture useful similarities present in even this small data set. It is worth noting that this has nothing to do with 3-D conformation as the rings we have chosen are, for the most part, planar and inflexible. This is a key observation that

even in the most favorable case 2-D fingerprints do not capture true chemical similarity. For instance, isothiazole is shown to be similar in shape and charge to 1,3,4-thiadiazole using our method but dissimilar using 2-D fingerprints. Another interesting comparison is how each method considers thiophene and benzothiophene. It is obvious that these two molecules are very different in size and shape, properties for which the protein target will be sensitive to, and this is reflected in the matrix generated using our method. The 2-D fingerprint method, utilizing common bond paths, finds these molecules to be relatively similar. These observations are repeated over much of the two matrices, with the shape method being superior to the Daylight, for these data, for identifying true positive molecules, and not identifying false positives.

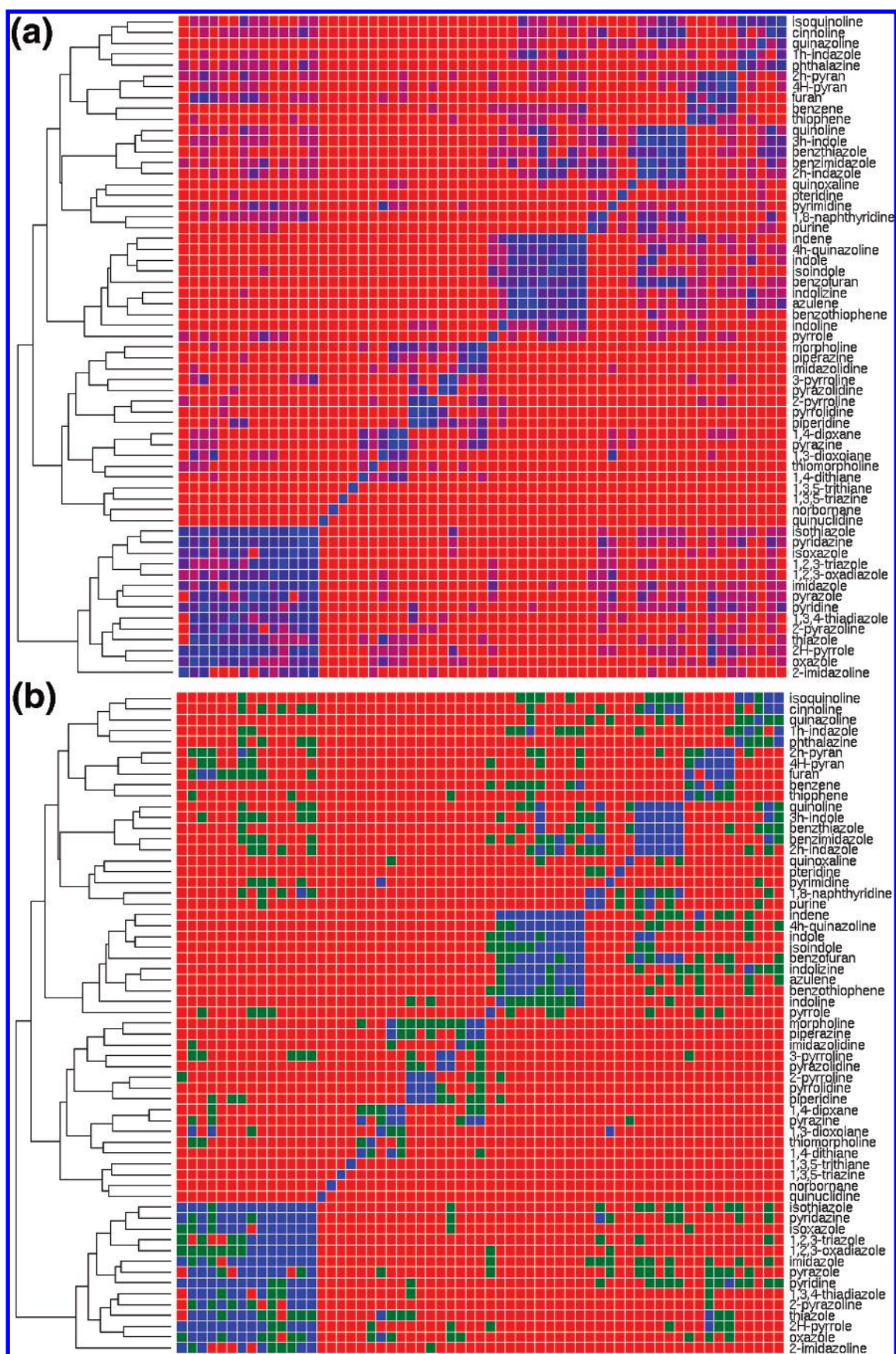


Figure 5. (a) The PB*Shape similarity for the Maybridge compounds represented as a color-coded matrix, where blue indicates a similarity of 1.0 and red a similarity of 0.0. (b) The PB*Shape similarity for the Maybridge compounds represented as a three-way color-coded matrix (0.0–0.4 red, 0.4–0.6 green, and 0.6–1.0 blue).

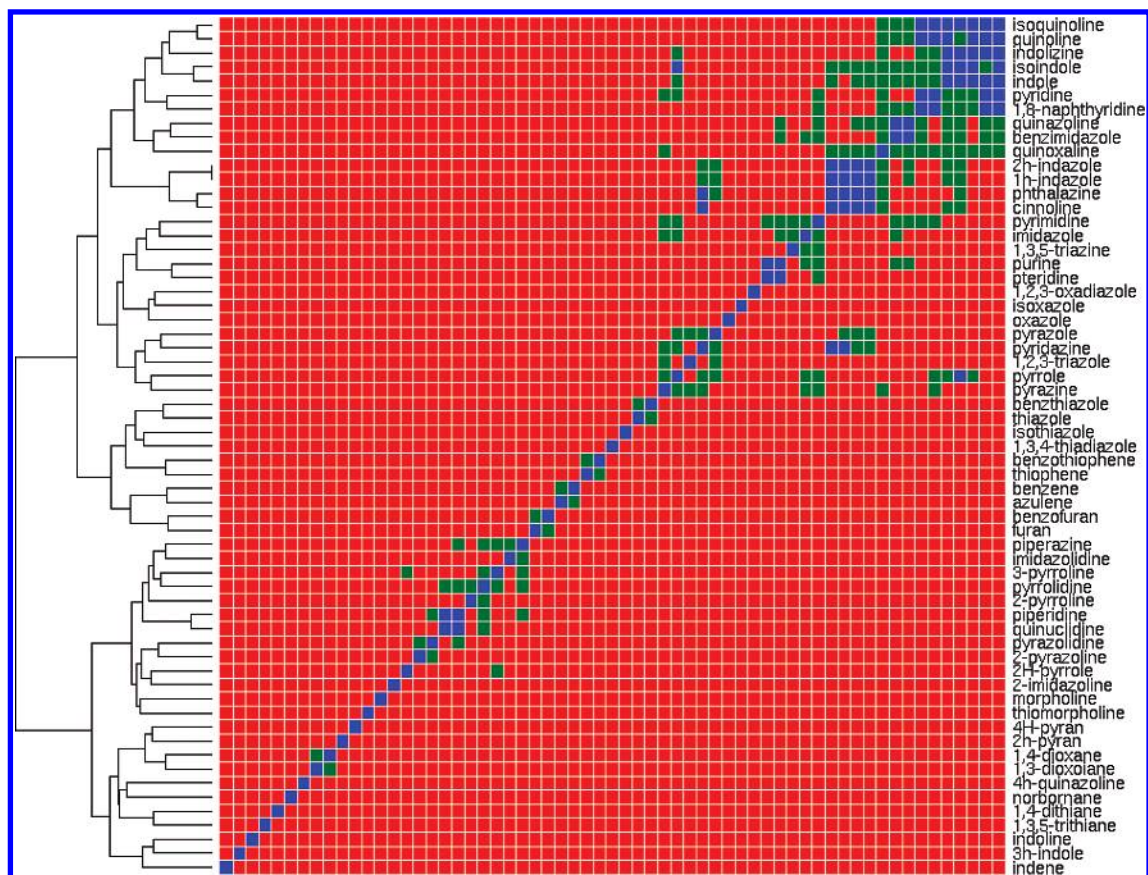


Figure 6. The similarity matrix for the Maybridge compounds using Daylight fingerprints as the molecular representation (0.0–0.4, red; 0.4–0.6, green; and 0.6–1.0, blue).

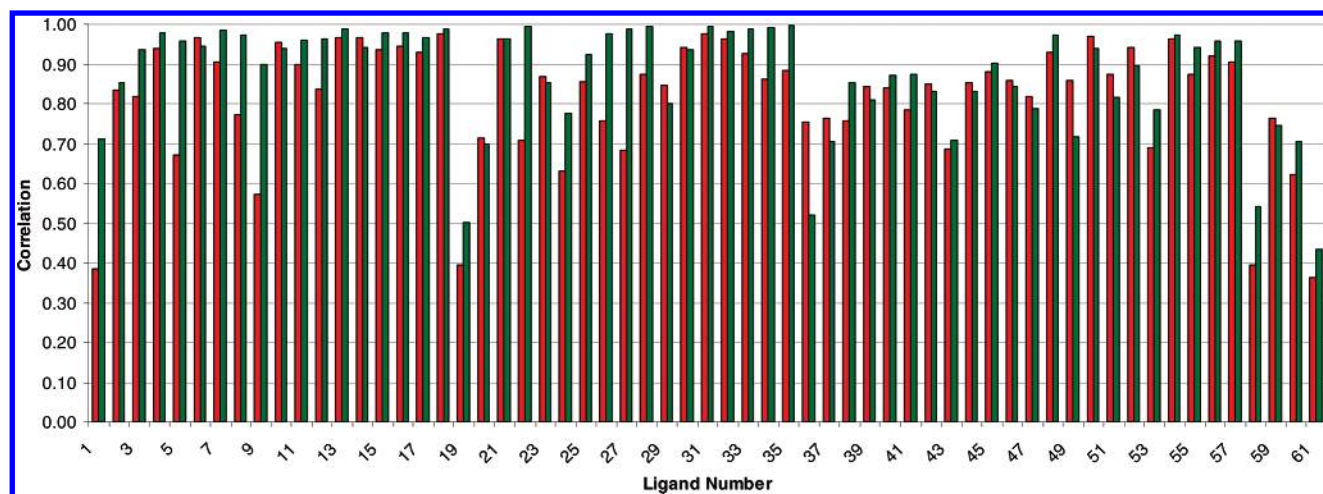


Figure 7. The Spearman (red) and Pearson (green) product-moment rank correlation coefficients for the Maybridge data set. The numerical labels correspond to those in Table 1.

THE EFFECT OF CHARGE MODELS ON CLUSTERS

We investigated the effects of the AM1-BCC and MMFF94 charge models on the structure of the clusters. Both models are used extensively in current computational chemistry applications as they calculate charges rapidly and are accurate and robust.¹⁵ AM1-BCC uses a bond-corrected AM1 semiempirical wave function to assign atom-based charges, whereas MMFF94 uses a template method based upon 6-31G* calculated charges. Ultimately we decided to use AM1-BCC for our explorations as this method is able to take conformational flexibility into account when assigning charges and can charge a more extensive set of molecules than

MMFF94: norbornane, for example. We also calculated the charges of the molecules at the 6-31G* level of theory using GAMESS-US.^{23–25} This resulted in minor rearrangements of the trees compared to the AM1-BCC charge model (data not shown), but the charges were similar enough to give us confidence in the AM1-BCC method. In Figure 7 we compare the AM1-BCC with the 6-31G* charges for each molecule in the Maybridge data set using Spearman and Pearson product-moment rank correlation coefficients. This shows that the charge schemes correlate well for most molecules. However, a number of molecules correlate poorly: furan, 1,2,3-oxadiazole, 1,3,5-trithiane, quinuclidine,

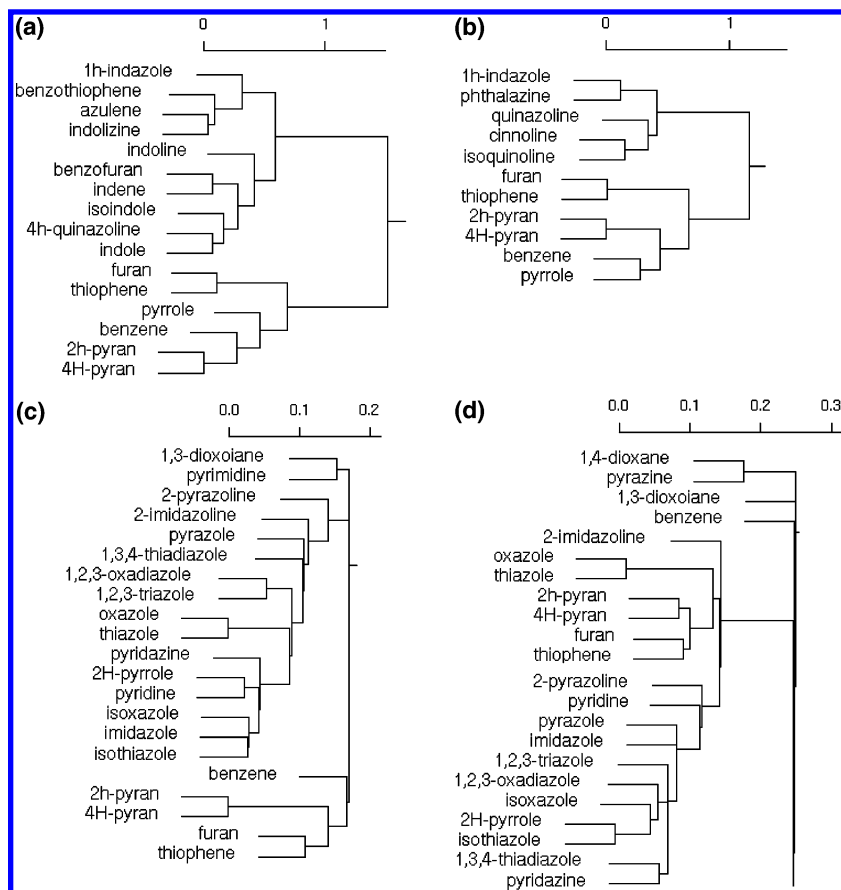


Figure 8. The effects of charge on the tree structure environment around benzene. Parts (a) and (b) show the differences between AM1-BCC and MMFF, respectively, using Ward's clustering, and parts (c) and (d) show the differences between AM1-BCC and MMFF using single-linkage clustering: (a) AM1-BCC charge model with Ward's clustering, (b) MMFF charge model with Ward's clustering, (c) AM1-BCC charge model with single-linkage clustering, and (d) MMFF charge model with single-linkage clustering.

indene, azulene, and norbornane. Upon inspection of these molecules, it was found that AM1-BCC partial charges were deviating significantly for certain carbons atoms and their associated hydrogens. In many cases, the sign of the partial charge on the carbon was incorrect and leads to the relatively poor correlations observed.

Figure 8 shows the effects of charge scheme on the structure of subclusters containing benzene. This leads to subtle, but significant, differences in the clusters.

Figure 8(a),(b) shows the effect of charge models using the Ward's clustering method. The benzene-containing clusters are similar, but the effect of the charge model on the location of pyrrole is interesting. With the MMFF94 charge scheme, pyrrole is clustered with benzene, whereas with the AM1-BCC method pyrrole moved into a node of its own. Also, 1H-indazole clusters with phthalazine with the MMFF94 charge scheme, whereas with the AM1-BCC method phthalazine is not present on any of the nearby nodes.

The analogous single-linkage trees are shown in Figure 8(c),(d). The AM1-BCC method leads to a similar structure as seen with Ward's clustering, but the benzene is considered a singleton, and the neighbor nodes here are H-bond acceptor heterocycles. The MMFF94 charge scheme against results in a similar cluster structure, but benzene is placed as an extreme outlier.

Both clustering method and charge scheme have effects on the structure of the clusters, some subtle, some more profound. Doubtless, this is also a function of the sparse data

matrix. Ideally, a bootstrapping method should be used to assign some confidence to the accuracy of the tree node placement.²⁶ As yet, we have not undertaken this exercise, and consequently the fine structure of the clusters should be evaluated with care.

CONCLUSIONS

We have compared and clustered sets of molecules using physically realistic shape and electrostatic field similarity metrics. The methods described provide a way to assign similarities between molecules that would not necessarily be found using classical path-length fingerprint descriptors, although the latter have many useful applications not possible with shape descriptors alone.

We show that both charge and cluster methods affect the results, sometimes subtly, but in some cases more drastically. Various clustering schemes can be used to explore different aspects of SAR design, from series explosion to hole-filling exercises. However, it is not clear that the clustering methods used above are ideal, and caution should be used when looking at potential false positives and negatives. The maximum similarity matrices we present can aid in this exercise.

Overall, it is gratifying to see that the method employed here finds molecules which one would deem to be similar based upon both intuition and empirical data. Both ring and nonring systems are captured with the method, and it is therefore expected to be broadly applicable to the problem

of bioisostere identification and design. The failure of 2-D fingerprint methods to meet this level of performance was to be expected, and the comparative performance is instructive. Such 2-D methods will continue to be of great value when very large sets of data are to be analyzed as these 3-D methods are computationally demanding.

ACKNOWLEDGMENT

The authors wish to thank the journal reviewers for their useful insight and comments. The authors also wish to thank the Takeda reviewers for their comments.

REFERENCES AND NOTES

- (1) Boström, J.; Hogner, A.; Schmitt, S. Do Structurally Similar Ligands Bind in a Similar Fashion? *J. Med. Chem.* **2006**, 49(23), 6716–6725.
- (2) Stahura, F. L.; Bajorath, J. New methodologies for ligand-based virtual screening. *Curr. Pharm. Des.* **2005**, 11(9), 1189–1202.
- (3) Jain, A. N. Virtual screening in lead discovery and optimization. *Curr. Opin. Drug Discovery Dev.* **2004**, 7(4), 396–403.
- (4) Olesen, P. H. The use of bioisosteric groups in lead optimization. *Curr. Opin. Drug Discovery Dev.* **2001**, 4(4), 471–8.
- (5) Wagener M.; Lommerse J. P. The quest for bioisosteric replacements. *J. Chem. Inf. Model.* **2006**, 46(2), 677–85.
- (6) Schuffenhauer, A.; Gillet, V. J.; Willett, P. Similarity searching in files of three-dimensional chemical structures: analysis of the BIO-STER database using two-dimensional fingerprints and molecular field descriptors. *J. Chem. Inf. Comput. Sci.* **2000**, 40(2), 295–307.
- (7) Boström, J. Reproducing the conformations of protein-bound ligands: A critical evaluation of several popular conformational searching tools. *J. Comput.-Aided Mol. Des.* **2001**, 15, 1137.
- (8) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A fast method of molecular shape comparison. A simple application of a Gaussian description of molecular shape. *J. Comput. Chem.* **1996**, 17, 1653–1666.
- (9) Haigh, J. A.; Pickup, B. T.; Grant, J. A.; Nicholls, A. Small Molecule Shape-Fingerprints. *J. Chem. Inf. Model.* **2005**, 45(3), 673–684.
- (10) Nicholls, A.; MacCuish, N. E.; MacCuish, J. D. Variable Selection and Model Validation of 2-D and 3-D Molecular Descriptors. *J. Comput.-Aided Mol. Des.* **2004**, 18, 451–474.
- (11) Nicholls, A.; Grant, A. J. Molecular shape and electrostatics in the encoding of relevant chemical information. *J. Comput.-Aided Mol. Des.* **2005**, 19(9–10), 661–686.
- (12) <http://www.eyesopen.com> (accessed May 20, 2006).
- (13) Boström, J.; Greenwood, J. R.; Gottfries, J. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J. Mol. Graphics Modell.* **2003**, 21, 449–462.
- (14) Rush, T. S.; Grant, J. A.; Mosyak, L.; Nicholls, A. A Shape-Based 3-D Scaffold Hopping Method and its Application to a Bacterial Protein-Protein Interaction. *J. Med. Chem.* **2005**, 48, 1489–95.
- (15) Jakalian, A.; Jack, D. B.; Bayly C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.* **2002**, 23(16), 1623–41.
- (16) Merck Molecular Force Field: I-V: Halgren, T.A. *J. Comput. Chem.* **1996**, 17, 490–641.
- (17) http://www.maybridge.com/Images/pdfs/ring_numbering.pdf (accessed May 26, 2007).
- (18) Daylight. <http://www.daylight.com> (accessed May 20, 2006).
- (19) Godden, J. W.; Stahura, F. L.; Bajorath, J. Anatomy of fingerprint search calculations on structurally diverse sets of active compounds. *J. Chem. Inf. Model.* **2005**, 45(6), 1812–9.
- (20) The R Project for Statistical Computing. <http://www.r-project.org/> (accessed May 20, 2006).
- (21) The 'cluster' package for R can be installed from the CRAN repository. cran.r-project.org/src/contrib/Descriptions/cluster.html (access May 20, 2006).
- (22) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, 38(2), 983–996.
- (23) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S. J.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. General Atomic and Molecular Electronic Structure System. *J. Comput. Chem.* **1993**, 14, 1347–1363.
- (24) Gordon, M. S.; Schmidt, M. W. Advances in electronic structure theory: GAMESS a decade later. In *Theory and Applications of Computational Chemistry, the first forty years*; Dykstra, C. E., Frenking, G., Kim, K. S., Scuseria, G. E., Eds.; Elsevier: Amsterdam, 2005; pp 1167–1189.
- (25) Gordon Group/GAMESS Homepage. <http://www.msg.ameslab.gov/GAMESS/GAMESS.html> (accessed May 20, 2006).
- (26) Highton, R. Molecular phylogeny of plethodonine salamanders and hyliid frogs: statistical analysis of protein comparisons. *Mol. Biol. Evol.* **1991**, 8(6), 796–818.

CI600549Q