Journal of Computer-Aided Molecular Design (2005) 19: 661–686
DOI 10.1007/s10822-005-9019-x

Molecular shape and electrostatics in the encoding of relevant chemical information

Anthony Nicholls[a,*] & J. Andrew Grant[b]
[a]*OpenEye Scientific Software, Inc., 3600 Cerrillos Rd. Suite 1107, Santa Fe, NM, 87507, USA;* [b]*AstraZeneca Pharmaceuticals, SK 10 4TF, Mereside, Macclesfield, Cheshire, England*

Received 17 May 2005; accepted 29 September 2005

Key words: analytic representation, database storage, electrostatics, molecular encoding, molecular shape, QSAR, shape fingerprints, similarity

Summary

We propose a molecule's chemistry can be hidden by representations of its shape and electrostatic field while retaining crucial, pharmaceutically relevant, information. Necessary, but not sufficient, to this proposition are the importance of shape and electrostatics to activity, the facility to easily represent, store and compare field properties, and knowledge of the density of possible drug-like molecules within a given radius of physical similarity. We provide methods and evidence to support the conclusion that a useful encoding is practical and propose tests for falsification.

Introduction

In the Sherlock Holmes story "The Adventure of the Dancing Men" the tale turns on a secret code where each letter of the alphabet is represented by a different stick figure. Holmes was able to decipher the code by frequency analysis, i.e. using the fact the letter 'E' occurs most commonly in the English language, followed by 'T' etc. By contrast, modern codes, based on technology derived from, for instance, large prime numbers or elliptical curves, strive to produce messages so random that the only practical line of attack is impractical, i.e. try every key or message. If we step back and consider the possibilities for molecular anonymity, it is not difficult to see that we are closer to stick figures than we are large prime numbers. The purpose of encryption in our context is to hide the chemical identity but not properties. We need some properties to make comparison possible and useful. But properties inevitably contain some information and so we are not presenting a random cipher. In fact, encryption is probably a misleading characterization for what is really required: obfuscation. Patents are awarded for precise definitions of molecular composition not vague descriptions, and so we do not have to prevent decryption to a similar message only to that precise message.

What properties should be used for molecular representation? Here again, traditional encryption may be misleading. It might seem obvious to use alphanumeric variables that arise from chemical descriptors as a substitute for cryptographic plain-text. However, we are not encrypting, we are obfuscating. If the purpose is to hide chemistry, using properties closely derived from chemistry is questionable. Our proposal is to instead use 3-D field properties, such as molecular shape and electrostatics. These are valid descriptions and can be thought of as an alternate branch in efforts to apply Schrödinger's Equation to chemistry (Figure 1). In fact, the first Hohenberg–Kohn

*To whom correspondence should be addressed. E-mail: anthony@www.eyesopen.com

theorem [1] establishes a formal basis for property calculation purely from electron density. Furthermore, although chemical descriptors been applied extensively, and with some success, to physical property estimation, the pursuit of pharmaceutically active compounds requires insight into intermolecular interaction, i.e. protein-ligand affinity. It is far from clear that chemical concepts are efficient in this domain. For instance, so-called "scoring functions", i.e. methods of estimating protein-ligand affinity from chemical typing are notoriously poor [2, G.L. Warren et al., submitted]. Shape and electrostatics represent the dominant properties of molecular interaction and would therefore are likely to carry the information appropriate for pharmaceutical discovery. There is now increasing evidence to this effect [3–6, Muchmore et al., submitted].

If our goal is obfuscation, an immediate concern is whether the shape and electrostatic fields surrounding a molecule exactly describe the molecule. From the practical experience of applying these concepts to molecular description, we conjecture that they do. If so, an exact representation of both is equivalent to the chemical structure. Fortunately, it is not necessary to describe shape and electrostatics exactly. Usefulness comes from similarity, not equivalence, and so the introduction of ambiguity, either deliberately or from coarseness of representation, does not preclude utility. The advantages and disadvantages of ambiguity will be discussed in some detail. However, it is instructive to consider exact representations and here a useful comparison to everyday cryptography can be made. Consider the common login interface of modern computing. Original implementations were found to be unsafe against dictionary look-up, i.e. trying all common words as possible passwords. This was greatly improved by delaying the results of a password attempt by a few seconds, i.e. slowing exhaustive attacks, and by requiring users to adopt passwords that are sufficiently different from common words, e.g. by adding numbers or special characters so as to make a comprehensive dictionary unrealistically large. Seen in this light, 3-D descriptors have the advantage of being slow to calculate, relative to chemical descriptors, and each compound can have many shape and electrostatic variants, e.g. from conformation and enantiomer expansion, and charge and tautomer state evaluation.

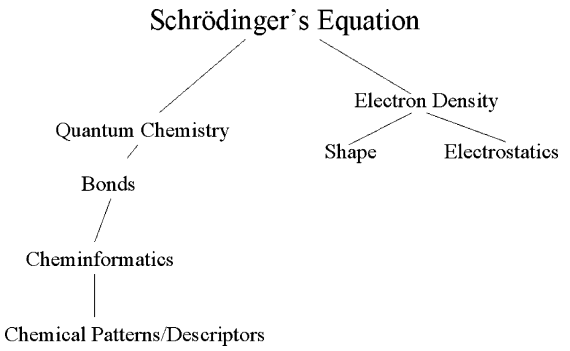

Figure 1. An argument that shape and electrostatics are equally valid descriptions of molecules.

Figure 2 illustrates different search methods assuming an exact description of shape and electrostatics. Method A is the "brute force" method, i.e. construct shapes and electrostatic profiles until a match is found. If the search is against N possible shapes, this will succeed, on average, in $N/2$ attempts. In this scenario, a description is every bit as safe as traditional encryption. The number of potential molecules of drug-like nature is astronomical. Estimates have ranged from 10^{25} to 10^{200} and up [7, 8]. A brute force search would be every bit as slow as a search for prime factors in an RSA key. However, as we have already noted, our descriptions contain information. This would enable a search of type B, namely a restricted search. For example, comparison against shapes of widely

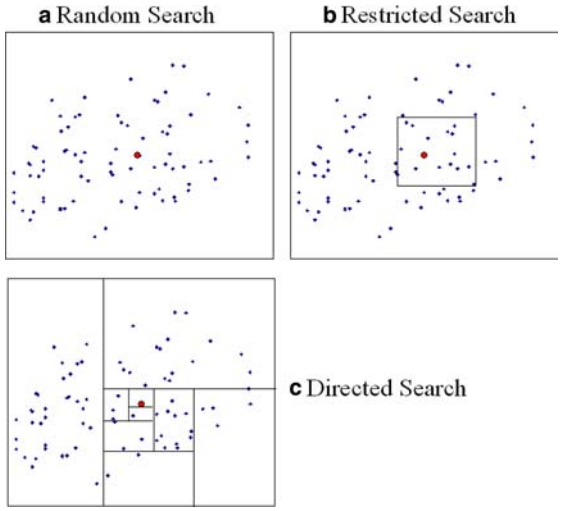

Figure 2. Search possibilities amongst N conformers in electro-shape space. (a) requires, on average, $N/2$ comparisons, (b) requires $kN/2$ and (c) $\log_2(N)$.

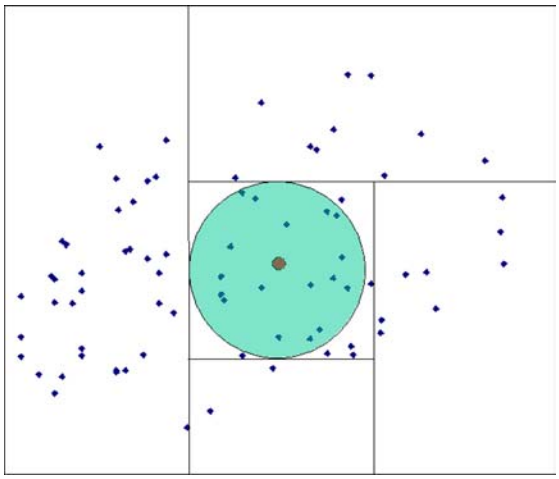

Figure 3. Directed search when the shape and electrostatics have ambiguity.

different volumes could be ruled out. Searches against fractions of possible molecules are clearly faster, but still likely to be widely impractical unless restrictions can be made stringent. However, beyond a few simple scalar properties such as volume or net charge, we know of no such properties that usefully restrict a search against shape and electrostatics. This does not mean that such characteristics do not exist, and if they did a progressive restriction, Method C, might exploit them. Here the space of possible solutions is gradually restricted until only one solution remains. Typically such methods take a sub-linear time, often proportional to $\log(N)$. Were such a solution to exist, shape and electrostatics would not, in general, be safe obfuscators.

Consider now the effect of ambiguity on a representation. Figure 3 illustrates the effect, represented by a circle around the precise definition, of deliberate or systematic error. Even a class C method can only pin down a region of descriptor space, albeit quickly. Of course, if the number of structures within the zone of ambiguity is small, one can then apply Method A to the remainder, i.e. enumerate all possible solutions confident that at least one was the actually structure. If, however, the remaining solutions are essentially still innumerable, the obfuscation is effective. In this work, and others [3, 9–12, J.A. Grant, in preparation], we have adopted a formal description of shape that admits a very precise consideration of molecular obfuscation. This description of shape in terms of volume difference at optimal alignment (see Figure 4) is both

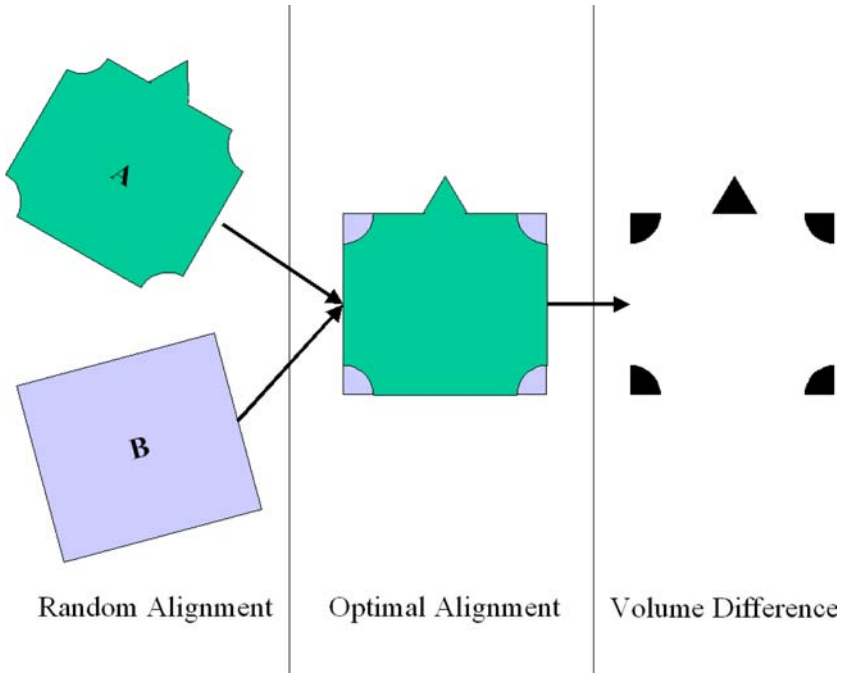

Figure 4. Illustration of the definition of shape used in this paper.

intuitive and mathematically compelling. This difference of fields is a metric quantity [3, 13] and, as such, forms a topological space [14]. Given a molecular size limit, such a space is finite and can be covered by representative forms. One practical application is a shape ''fingerprint'' we shall consider for anonymization. The dimensionality of this space is crucial for understanding the difficulties of deducing chemistry from a shape as search techniques that work well in one or two dimensions can rapidly fail in higher dimensions.

We shall combine results reported elsewhere, in particular the work of Fink et al. [15] on enumerating molecular graphs and previous work on our definition of shape [12], with new results on large-scale similarity searching to provide reasonable bounds on the dimensionality of electro-shape space and so provide concrete estimates of molecular obfuscation. We shall describe three potential methods of shape representation and the practical issues for each: shape fingerprints, volume-fitting ellipsoids, and analytic forms. These approaches arose in our investigations of effective methods of virtual screening but each also satisfies the requirements for molecular obfuscation. We will also consider potential issues concerning anonymization, such as the multi-conformer problem, the chance of shape or electrostatic outliers and the potential that chemical space as envisioned by chemists is actually far more finite than the mathematical descriptions proposed. We then suggest a possible alternate, namely surrogate structure representation, to avoid these potential pitfalls.

Methods

Shape and shape fingerprints

If we define the difference in shape between two objects as:

$$D_{A,B} = \min\left(\sqrt{\int_{\text{all-space}}\int\int (S_A(x,y,z) - S_B(x,y,z,\alpha,\beta,\gamma,t_1,t_2,t_3))^2 dxdydz}\right)$$

where (α,β,γ) are rotational variables, for example Euler angles, (t_1,t_2,t_3) translational variables and S_x is a shape, or *characteristic*, function of object X, then it can be shown that D is a metric distance. That is to say:

$$D_{A,A} = 0$$
$$D_{A,B} = D_{B,A} \geq 0$$
$$\text{if } D_{A,B} = 0, B = A.$$
$$D_{A,B} + D_{B,C} \geq D_{A,C}$$

As such, D forms a *topology* for shape. This space is continuous, i.e. any shape has a shape arbitrarily close to it (although this may not be a molecular shape) and is compact. In addition, if

$$\int_{\text{all-space}}\int\int S(x,y,z)^2 dxdydz < \infty$$
$$\frac{\partial^n S}{\partial r^n} < \infty; \frac{\partial^m S}{\partial r^m} = 0 \text{ if } m > N \text{ where } N < \infty$$

then the extent and dimensionality of the space are both finite. Unfortunately, the condition that derivatives disappear beyond a certain degree is not true for typical molecular representations i.e. the dimensionality is technically infinite. However, in practice, the error in truncating dimensions beyond a certain limit is small. It is possible to cover the space with a finite set of spheres or *balls* such that any point in the space is within one of these balls. Therefore, a finite number of molecular shapes can represent or *cover* all molecular shapes. This is the concept behind shape fingerprints. A predetermined set of structures are found such that any shape must be 'close' to at least one structure, and probably many. A bit is turned on for each such proximal structure and this set of bits is a representation of the locality in shape space (see Figure 5). The coverage set can be chosen in many ways, however an optimal set might consist of structures that overlap each other to a minimal extent. This has been explored via a d-optimal

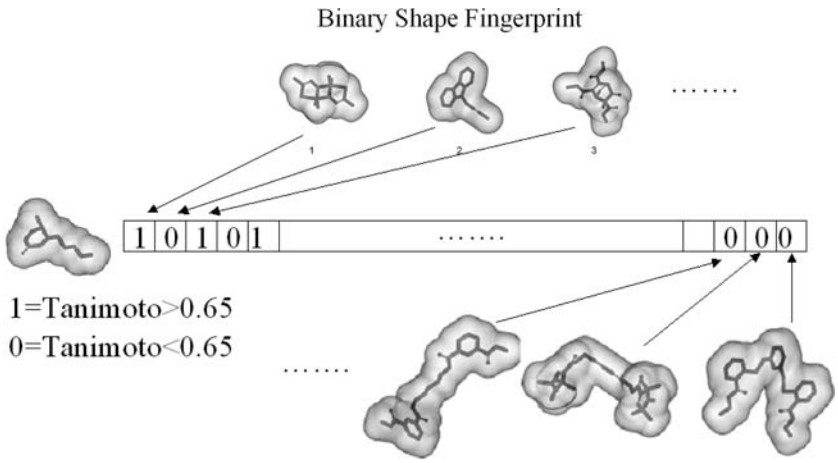

Figure 5. Generation of shape fingerprints. If the ST between the structure being fingerprinted and a reference structure is greater than a certain threshold, typically 0.65, a bit corresponding to that reference structure is set to from 0 to 1.

design methodology [12] that indicates a few thousand structures can adequately represent the shape space of molecules of heavy atom count between 12 and 32. Example structures from this process are shown in Figure 6. Other fingerprint methods for shape have also been proposed, e.g. [16], although not utilizing this nature of shape space.

A useful measure of shape similarity is the Shape Tanimoto (ST):

$$\mathrm{ST}_{A,B} = O_{A,B}/(O_{A,A} + O_{B,B} - O_{A,B})$$

where overlap is defined by

$$O_{A,B} = \int\limits_{\text{all}-\text{space}} \int\int S_A(x,y,z) S_B(x,y,z) dxdydz$$

STs are bounded by zero and one. As all molecules can overlap to some extent, the zero bound is only approached. Complete shape similarity, i.e. a ST of unity, is usually only found for a compound with itself, or with extremely conservative variation. As an intuitive measure, ST compares well with chemically based approaches. For instance, an ST above 0.7 generally means a visually similar shape. Above 0.75 and there is good shape similarity. Above 0.8 the similarity is striking. Above 0.85 the shapes are almost identical. Figure 7 illustrates this trend. Functional similarity seems to correlate well with shape similarity of 0.7 and greater [3]. The chemical structures corresponding to the shapes in Figure 7 are given in Figure 8, to illustrate the variation in chemical construction.

Shape ellipsoids

Ellipsoids are a natural geometric construct with which to represent molecules. When viewed purely as volumes, molecular fragments such as ring systems take on the appearance of oblate ellipsoids, whereas spacer regions or alkyl chains are prolate spheres. Finding a closely matching ellipsoid for such simple components is not difficult; the moments of inertia alone define a set of lengths and directions that are more than adequate. The more complex problem is to form representations that do not reflect chemistry, an essential step in anonymization. The approach taken here relies heavily on the mathematics of Gaussian functions. It has been shown that the volume of a molecule can be accurately reproduced by a set of atom-centered Gaussians [10]. Here, if there are N atoms in the molecule and each atom has position (x_i, y_i, z_i), we have:

$$V = \int\limits_{\text{all}-\text{space}} \int\int S(x,y,z).dxdydz$$

$$= \int\limits_{\text{all}-\text{space}} \int\int \left(1 - \prod_{i=1}^{N} (1 - G_i(x,y,z))\right) dxdydz$$

$$G_i = p_i e^{-\alpha_i\left((x-x_i)^2+(y-y_i)^2+(z-z_i)^2\right)}$$

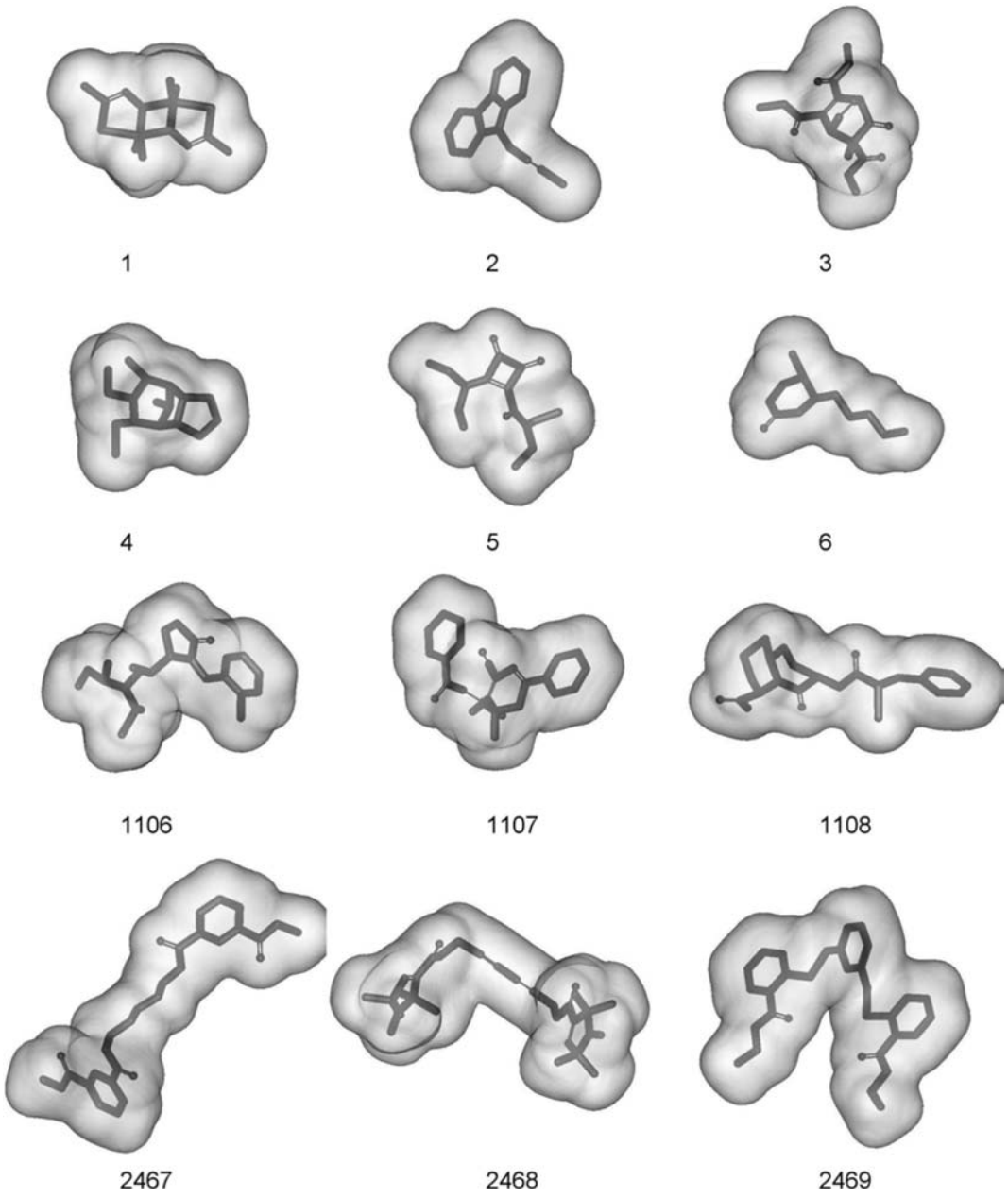

Figure 6. Examples of reference shapes that represent the shape space of molecules of heavy atom count between 12 and 39.

The Gaussian width α and prefactor p are constrained such that the volume of a single sphere is reproduced and that the overlap of two spheres is accurate. There have been many other attempts to use Gaussian functions for the representation of molecular forms [17–19]. The shape function S provides an analytic alternative of considerable utility compared to the binary inside/outside function of fused spheres. For example, the overlap of two molecules can be evaluated as a sum of overlaps of Gaussian functions that are simple to evaluate and differentiate. Here, we fit this function to one or more ellipsoidal Gaussians of the form:

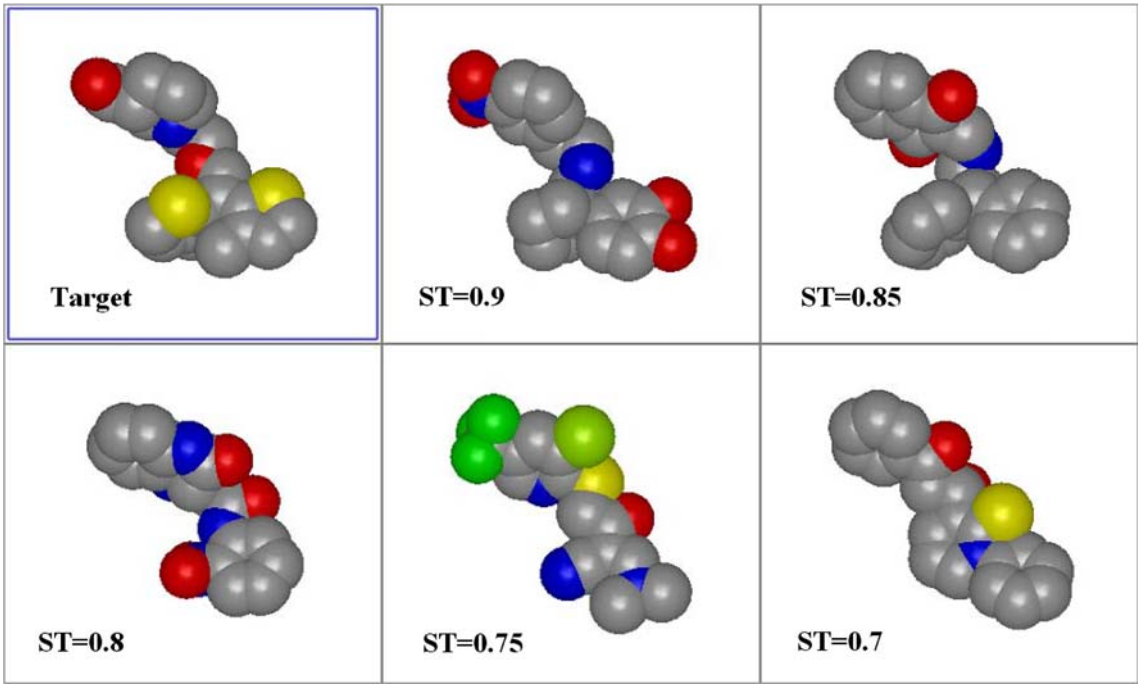

Figure 7. Visual similarity and the ST.

$$E_j = p_j e^{-\alpha_j(X-X_j)^2-\beta_j(Y-Y_j)^2-\chi_j(Z-Z_j)^2}$$

where

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} {}_jr_{1,1} & {}_jr_{1,2} & {}_jr_{1,3} \\ {}_jr_{2,1} & {}_jr_{2,2} & {}_jr_{2,3} \\ {}_jr_{3,1} & {}_jr_{3,2} & {}_jr_{3,3} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}; \begin{pmatrix} X_j \\ Y_j \\ Z_j \end{pmatrix}$$

$$= \begin{pmatrix} {}_jr_{1,1} & {}_jr_{1,2} & {}_jr_{1,3} \\ {}_jr_{2,1} & {}_jr_{2,2} & {}_jr_{2,3} \\ {}_jr_{3,1} & {}_jr_{3,2} & {}_jr_{3,3} \end{pmatrix} \begin{pmatrix} x_j \\ y_j \\ z_j \end{pmatrix}$$

i.e. where the center of ellipsoid j is at (x_j,y_j,z_j), has (Gaussian) widths $(\alpha_j,\beta_j,\chi_j)$ along axes described by the orthogonal (rotation) matrix ${}_jr$. The procedure is to minimize the function:

$$F_n = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(S(x,y,z) - \sum_{i=1}^{n} E_i(x,y,z)\right)^2 dxdydz$$

for a set of n Gaussian ellipsoids. Typical initial conditions are spheres placed randomly on different atoms. Each ellipsoid has 10 parameters: prefactor, widths, center and rotation and hence the target function is an optimization of $10n$ variables. Surprisingly, perhaps, this procedure works relatively efficiently. Experience has shown that multiple representations typically arise when $n>1$, but that each is a visually reasonable representation. If a single representation needs to be chosen, that with the lowest F_n can be assigned. However, the choice of an appropriate n is more challenging as $F_n^{\min} < F_{n-1}^{\min}$, simply because there are more variables available to fit against S. However, as more ellipsoids are used, they begin to overlap, i.e. over-sample. If we use the n ellipsoids to partition atoms into n sets, S_i, we can define an ancillary function,

$$P(n)=\sum_{i=1}^{n}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(S_i(x,y,z)-E_i(x,y,z))^2 dxdydz$$

where as the fitness function, $F(n)$, decreases with n, $P(n)$ increases. A minimum of $(F(n)+P(n))$ typically correlates with a well-formed, 'minimal' ellipsoidal representation of S. Figure 9 illustrates the progression of representations for Omeprazole and Figure 10 the optimal representation at the two ellipsoid level.

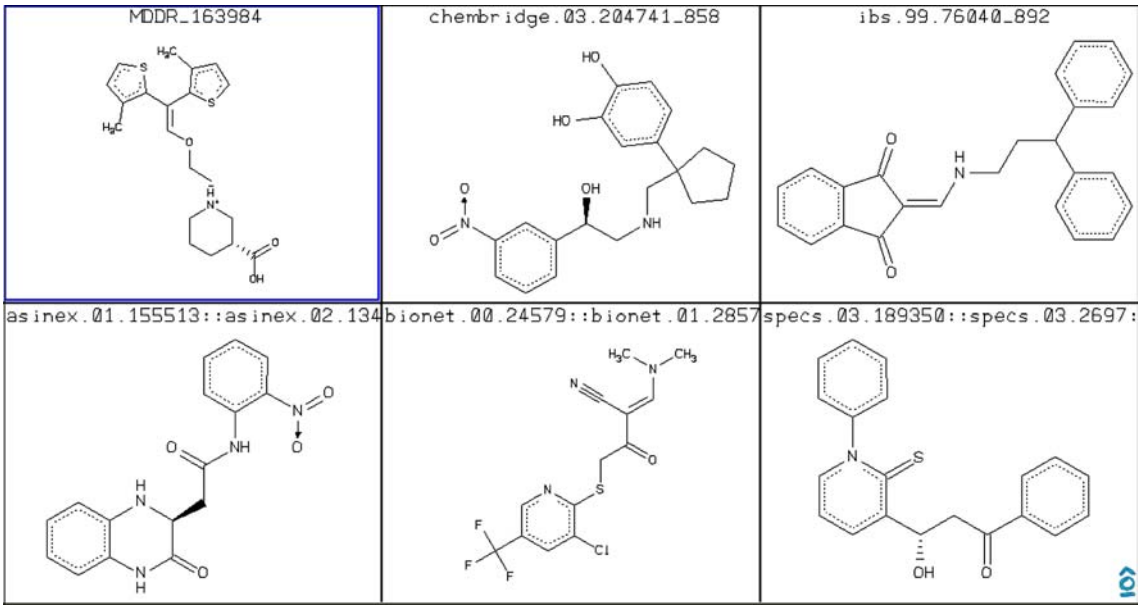

Figure 8. Illustration of the chemical dissimilarity for the same set as in Figure 7.

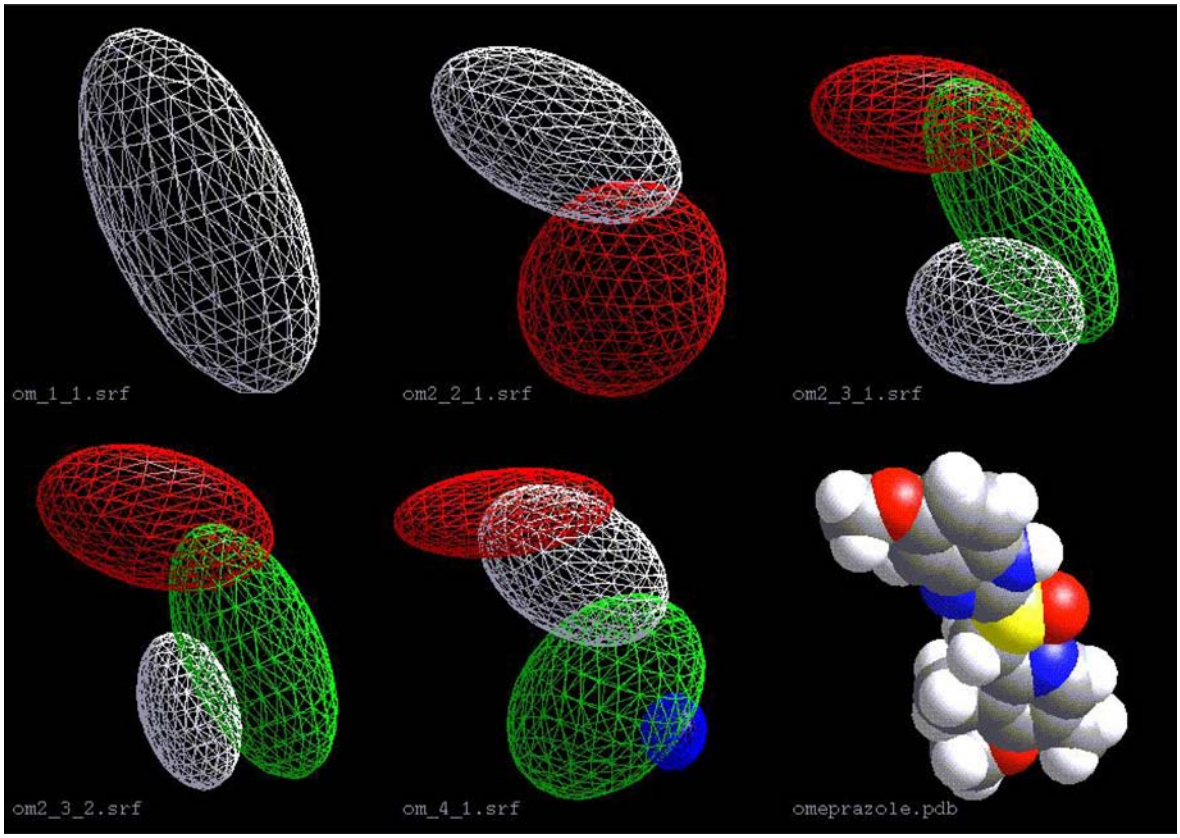

Figure 9. Illustration of ellipsoid matching. The structure, bottom left, of Omeprazole generates five representations, including two alternate representations at the three ellipsoid level.

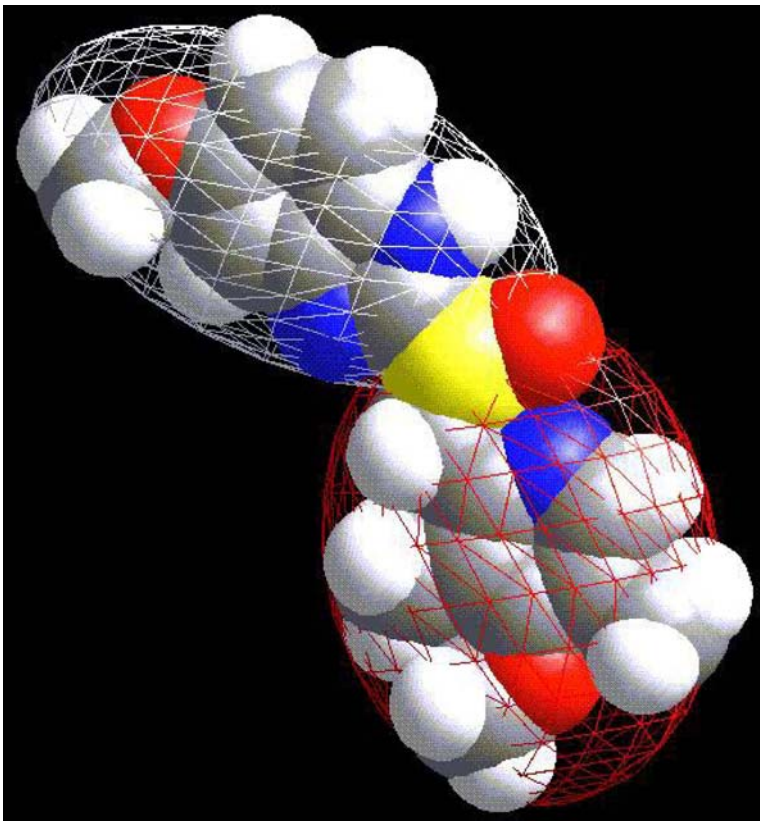

Figure 10. The best ellipsoid match against Omeprazole.

Polynomial representations

An alternative to representing fields as grids is to transform them into a set of analytic functions. For instance, the Discrete Fourier Transform (DFT) takes a grid of values and produces a set of cosine and sine functions to reproduce the original values. Curtailing the expansion produces a smoother, less accurate function. When the number of terms is equal to the number of data points, the function can be reproduced exactly. There are many other choices for expansion functions. The DFT works well for crystallographic density because the periodicity of the transform corresponds to the periodicity of the lattice. In quantum mechanics the basic orbital functions of electrons moving under a central potential are spherical harmonics for angular momentum and Laguerre polynomials for the radial component because they satisfy the separated form of the Schrödinger equation in spherical coordinates. Because of their natural application in representing rotational states, spherical harmonics have been used in the representation of molecular fields [20] and generic objects [21]. There have also been many other efforts in the direction of analytic representation of molecules, either for volumes or surfaces [22–27].

There are elegant aspects in the application of this approach. First, analytic forms are far more compact than grid representations, making storage and retrieval easier. Second, grid representations are difficult to transform, especially by rotation. In fact, such operations tend to be lossy, i.e. one cannot necessarily invert the operations and retrieve the same values. Function representations can be transformed backwards and forwards without loss except that due to numerical precision. Third, analytic functions give analytic gradients, useful for optimizing alignment whereas gradients from grids have to be obtained numerically. Fourth, functional representations can be smoothly truncated to give more approximate, but still valuable, representations. This provides a

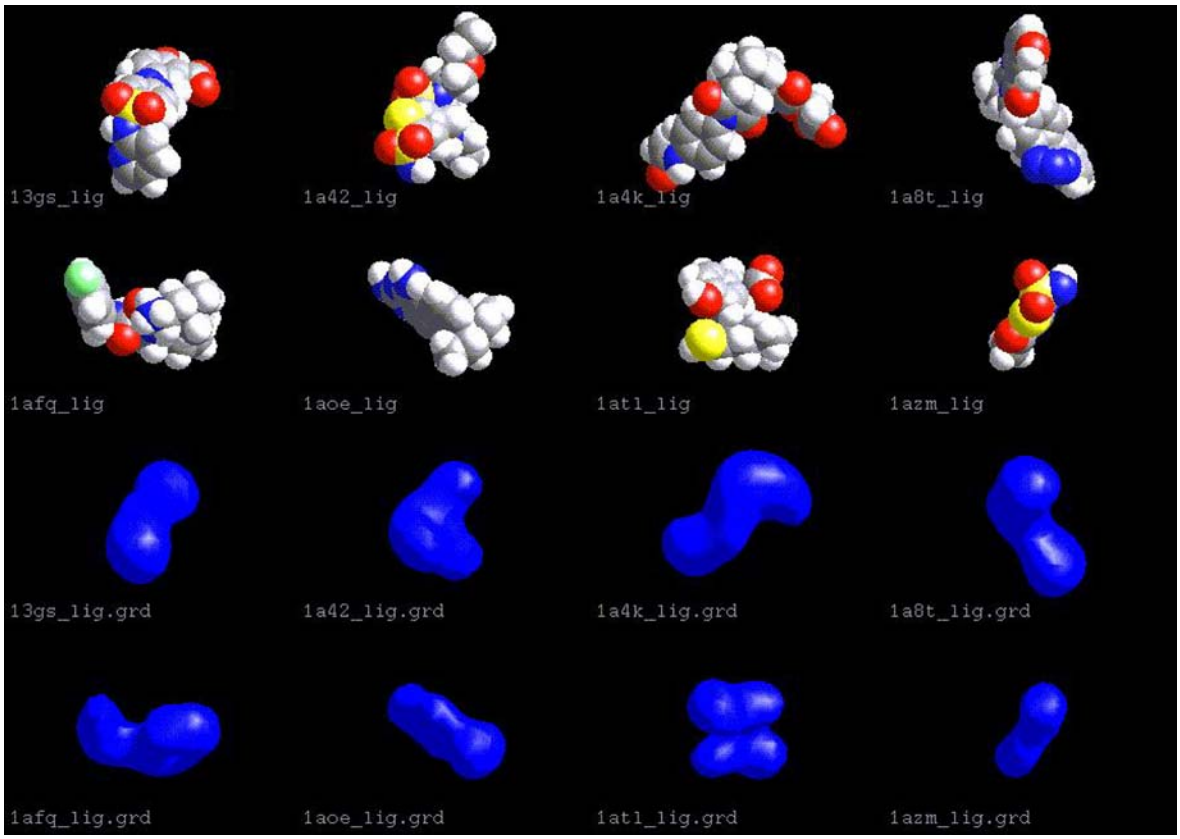

Figure 11. Illustration of analytic shape representations of eight small molecules.

simple way to anonymize the atomic positions in a shape representation, or the electrostatic field from charged atoms. This is an example of a more general feature of functional representations, i.e. the ease with which noise can deliberately be imposed on the shape and electrostatic signal such that useful comparison is still possible, but exact reconstitution (where were the atoms, and what charges did they carry?) becomes harder.

Our approach has been to use the functions:

$$f_{l,m,n}(x, y, z, \lambda) = x^l y^m z^n e^{-\lambda(x^2+y^2+z^2)}$$

in the representation of molecular shape and electrostatics. Such terms can be grouped by the sum of $(l+m+n)=p$, the *order* of the polynomial. For $\{p=0\}$ there is one term, $\{p=1\}$ has three terms, $\{p=2\}$ six terms and $\{p=N\}$ has $(N+1)*(N+2)/2$ terms. Including all terms up to fourth order requires only 35 terms and provides a reasonable reproduction of the basic shape and electrostatic profiles of small molecules (Figures 2 and 3). The choice of λ is crucial to the efficiency of the polynomial representation. Experience has shown that $\lambda = 1/a^2$, where a is the largest eigenvalue of the mass matrix, i.e. the major axis of a single ellipsoid representation, works well for electrostatics and $\lambda = 0.5/a^2$ is appropriate for molecular shape. The result of optimizing the coefficients for these functions so as to represent the underlying fields with least error gives rise to the examples shown in Figures 11 and 12.

Conformer construction

An important part of the generation of three-dimensional representations of molecules is conformational expansion. In many cases the bioactive form will not be known and an ensemble of shapes will be required. Equally important, though, is that these are appropriate conformations, i.e. high-energy conformations are not likely to be relevant to biological efficacy and greatly exceed, in number, low-energy conformations. Fortunately, recent research has provided guide-

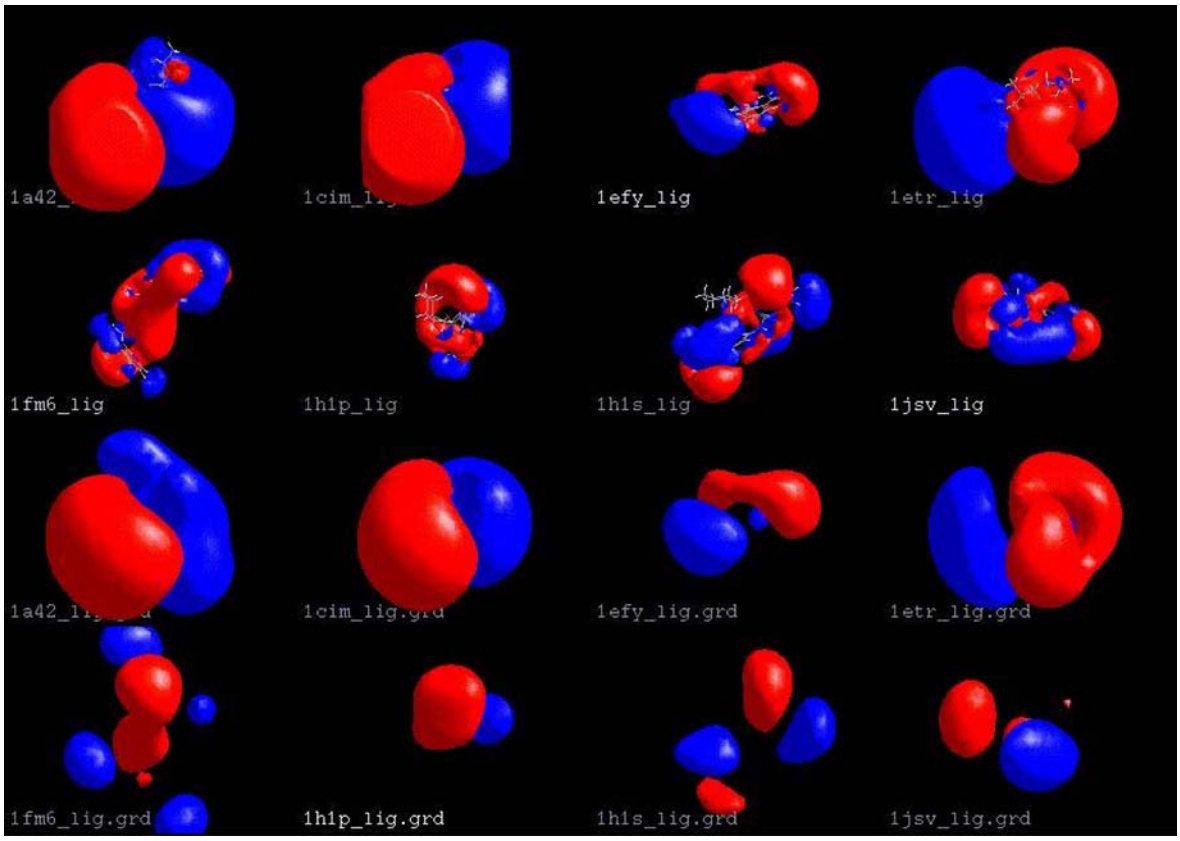

Figure 12. Illustration of analytic representations of the electrostatic fields of eight small molecules.

lines for the classification of high- and low-energy conformations. Perola and Charifson [28] examined both the strain energy of co-crystallized ligands from public and private sources and found that conformers are minimally strained (90% less than 5 kcal mol^{-1}) and that the conformational minima are, in general, within 5 kcal mol^{-1} of the solvated global minima. Boström et al. [29] found solvated minima were highly correlated with bioactive conformations, and also assessed the facility of OMEGA [30] to find such states [31]. MacCuish et al. [3] found a 5 kcal mol^{-1} window in conformational generation was sufficient to capture most relevant shapes for shape and electrostatic field QSAR.

Using a consensus value of 5 kcal mol^{-1} in a program such as OMEGA typically produces 80–100 structures for molecules of 5–6 rotatable bonds, which is a typical average for pharmaceutical collections. Such expansions usually employ root-mean-squared (distance) (RMS) reduction, i.e. conformations that are within an energy window of the lowest energy conformer but also within a certain RMS of a lower energy structure are not included. This potentially reduces the set of available shapes, but a low RMS typically implies a similar shape (although the converse is not true). As such, it is possible that the typical RMS list could be culled further by removing structures if a similarly shaped conformation of lower energy exists. Studies suggest this may contract the number of shapes by a factor of 10 or more [J. MacCuish, unpublished]. Electrostatic comparison may have the opposite problem. Small motions of highly polar groups can dramatically affect the electrostatic profile while leaving the shape essentially unchanged. This suggests an approach wherein shape is the primary search key, with multiple electrostatic profiles stored for each shape. This approach also applies to tautomeric, charge transfer and ionization state multiplicities.

Electrostatic field calculation and comparison

Electrostatic fields are compared in a manner analogous to molecular shape, i.e. by field integration:

$$D_{A,B} = \sqrt{\int\int\int_{\text{exterior}} (E_A(x,y,z) - E_B(x,y,z))^2 dxdydz}$$

Note that the degrees of freedom of the second field pertaining to rotation and translation have been removed. For present purposes these are determined by the best shape orientation between A and B. The other difference is the restriction of the integral to the exterior of both molecules. It seems reasonable to avoid comparison of the potential inside molecules. First, it is not germane to molecular interaction. Secondly, it avoids singularities from any point charge descriptions used to generate potentials. Typically, this occlusion of electrostatic potentials is achieved by multiplying each map by a molecular shape-based masking function. In this work, the form of the masking function is a switching function as defined below:

$$\mu(\bar{r}) = \{1.0 \text{ if } e^{-S(\bar{r})} < 0.06\}, \{0.0 \text{ otherwise}\}$$

$$\text{where } S = A\sum_{i=1}^{N} \chi_i(\bar{r})$$

$$\text{where } \chi_i(\bar{r}) = e^{-\alpha(\bar{r}-\bar{r}_i)^2} \text{ for atom } i \text{ at } \bar{r}_i \text{ from } N$$

Typical values of A and α are 10.0 and 0.7. The distance integral then becomes:

$$D_{A,B} = \sqrt{\int\int\int_{\text{all-space}} (E_A(x,y,z)\mu_A(x,y,z) - E_B(x,y,z)\mu_B(x,y,z))^2 dxdydz}$$

Likewise, an Electrostatic Overlap (EO) is defined as:

$$\text{EO}_{A,B} = \int\int\int_{\text{all-space}} E_A(x,y,z)\mu_A(x,y,z) * E_B(x,y,z)\mu_B(x,y,z)dxdydz$$

and an Electrostatic Tanimoto as:

$$\text{ET}_{A,B} = \text{EO}_{A,B}/(\text{EO}_{A,A} + \text{EO}_{B,B} - \text{EO}_{A,B})$$

Note that because an EO can be negative, for instance if $E_A = -E_B$, so is ET. For this condition of complete field reversal, ET is a minimal $-1/3$. It can be shown that an additional consequence of negative electrostatic field values is that the expected ET between any two fields is approximately zero. Consequently, the threshold for visual and functional similarity is much lower. Typically an ET of 0.2 corresponds to some similarity, with values above 0.4 representing strong homology.

An important issue in the use of electrostatic fields surrounding a molecule is the choice of the surrounding medium. The electrostatic fields around a molecule in vacuum are quite different from those in water or octanol. Vacuum fields may be calculated by applying Coulomb's Law to a partial charge assignment to each atom. Alternatively, aqueous potentials can be derived from similar charge sets and the Poisson–Boltzmann or Poisson Equation. It remains an open question as to which field is most correlated with activity. The argument for the Coulomb field is that it is a fundamental field, i.e. independent of medium. It might also be claimed to be most appropriate for bound states with solvent excluded. However, binding sites typically try to replace the effects of water. For instance, groups of opposite polarity from the protein compensate polar or charged groups on ligands that bind well. As such, the preorganization of the protein provides a reaction field similar to water and this might suggest solvated electrostatics as being more relevant.

A further complication is the choice of charge sets. It would seem here that one should use a set of charges least dependent on chemical identity. Those derived from correlation with quantum potentials would seem to fit this description. Quantum codes require only positions, element number, and multiplicity (charge-state) and not chemical connectivity. Conversely, charge sets

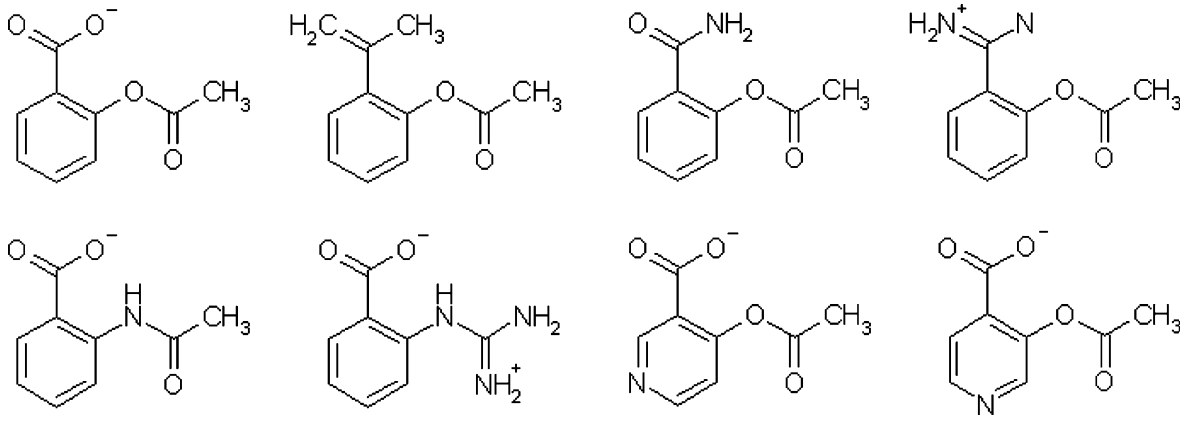

Figure 13. Examples of atomic level substitutions on the chemical structure of Aspirin (top left) generated using WABE.

such as Gasteiger and Marsili [32] are chemical-based and as such perhaps more likely to be reverse engineered.

De similis design

WABE [30] is a program that looks for atomic substitutions that maintain local hybridization. The dominant change in molecular energy is electrostatic while the three-dimensional shape of the molecule remains similar. The type of substitution permitted can be tailored from local graph environments present in any molecular dataset presented, and are applied in a combinatorial fashion. Examples of WABE output are shown in Figure 13. WABE is an example of a *de similis* design, as opposed to *de novo* design, as an initial graph pattern of a known molecule is the starting point. Further details on WABE will be published elsewhere. In this work, WABE is used to explore the diversity of possible electrostatic environments around molecules of similar shape.

Millions of available chemical entities (MACE)

OMEGA v1.1 was used to create a collection of 620 million conformers from a set of collected set of 2.3 million vendor-available compounds collected in-house. The parameters for OMEGA were an energy cut-off of five kilocalories, an R.M.S. threshold between conformers of 0.6 Å and maximum 1000 conformers per structure. MACCS keys fingerprints [33] were generated for each molecule via the *Fingerprint Module* of the Mesa Analytics [34] suite of tools. The total, compressed, file size is about 30 Gigabytes, split into 250 separate files.

Results

Shape representations

Figure 14 shows comparisons of the behavior of the three anonymizing representations; shape fingerprints, analytic functions, and ellipsoids were used to calculate the ST of one thousand query shapes to a target molecule Fesitin, an inhibitor of Human Cyclin-Dependant Kinase 6 (PDB code 1XO2) [35]. The query molecules were chosen at random from the MDDR [36]. Each representation uses the orientation generated and the ST calculated by the program ROCS [30].

Each form clearly has its own characteristics. Shape fingerprints tend to produce much lower STs than the numerical solution from ROCS and fail to discriminate when the shape similarity falls below a certain threshold, i.e. if a target and query are sufficiently different they may have no reference structures in common, and hence any bit-string overlap. This was anticipated because shape fingerprints were designed to discriminate between shape similar and not shape dissimilar molecules. The ellipsoid and functional representations are more similar. However, in the region of $ST > 0.7$ the correlation between the ellipsoid volumes and the ROCS results has a slope close to unity, whereas the functional forms overestimate the ST.

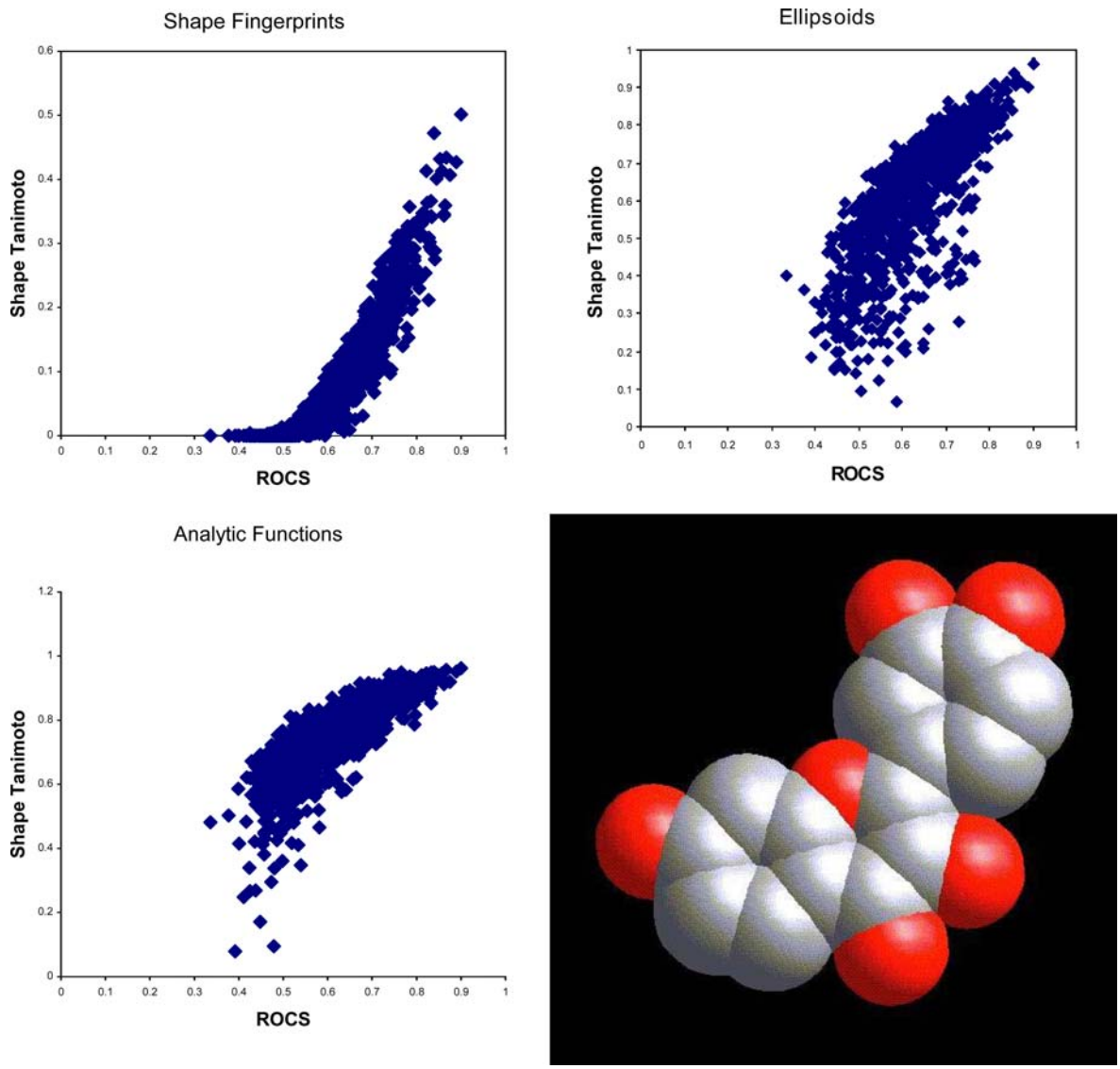

Figure 14. Ability of different shape representation to agree with ROCS shape similarity. Lower right is a space-filling image of the target, Fesitin.

With only 35 terms, the functional expansion tends to simplify the shape and, as a consequence, molecules tend to look overly similar. On the other hand, ellipsoids are uniformly matched to the volume of the molecule and produce a better correlative, if not more accurate, result.

Overall each correspondence is useful, i.e. would significantly enrich if used to predict which structures are close in shape to a given target. Substantial work has been done in the shape fingerprint domain to prove this is so regardless of the target [12]. Less has been done with either the ellipsoid or functional forms. The latter, in particular, seem promising as the accuracy can be continually improved as more terms are added. In addition, the possibility of combining the two approaches, i.e. to calculate a functional representation of the atoms within each ellipsoid, is an intriguing possibility. It should be possible to gain a significant improvement in the accuracy of the functional form and also open up the possibility of partial shape comparison with these anonymous representations.

The fingerprint approach can also be applied to electrostatic fields although it remains to be seen if this is a practical approach, given the imposed correlation between shape and electrostatics. This, however, would be important if fingerprints were to be the only anonymization. Alternatively, shape fingerprints could be used as a primary filter, followed by an alignment of electrostatic fields by polynomial functions or ellipsoids. It should be noted that ellipsoids also form an interesting basis for electrostatic comparison. Potentials can be mapped on to the surface of each ellipsoid and then stored as a discretized array or functional form. Comparison of electrostatics between two representations is then simply a matter of pairing up ellipsoids and assessing the difference between ordered arrays.

Some practical considerations of using the suggested anonymizing representations are ease of storage, retrieval and reconstitution. Shape fingerprints can be compressed to about 80 bytes, while the functional form and ellipsoid representations are of the order of 30–60 floating point numbers. It is straightforward to store this amount of data along with standard representations of molecular form. For instance, the polynomial forms described above for representing either shape or electrostatic fields can be stored to fourth order (35 coefficients) by a 184-character tag in an sdf file. The cost of calculation is another important characteristic. Shape fingerprints, at present, take around 3–4 s per conformer, i.e. much slower than the computation of that conformer from a connectivity record. However, it is likely this can be improved to fractions of a second per fingerprint. Ellipsoidal representations are also slow. An exhaustive evaluation, i.e. multiple constructions from different starting points and with different numbers of ellipsoids may take 10–12 s. However, the polynomial form representations of molecular shapes can be extremely rapid, i.e. of the order of several thousand per second. The calculation of electrostatic fields to be encoded is also rapid, i.e. of the order of a hundredth of a second. As such, the analytic forms are probably the best candidates for production on a mass scale, suitable for large virtual libraries.

Large-scale shape comparisons

A crucial aspect of this proposal is the number of similarly shaped molecules within a given threshold. If shapes were so distinct as to provide a direct index to chemistry the proposal would fail. To illustrate this is not the case we present results from a study on large-scale shape searching, the complete details of which will be presented elsewhere. A hundred shapes sampled from a set of representative shapes of MDDR (generated by CORINA [37]) were each compared against the MACE conformer library using ROCS. Figure 15

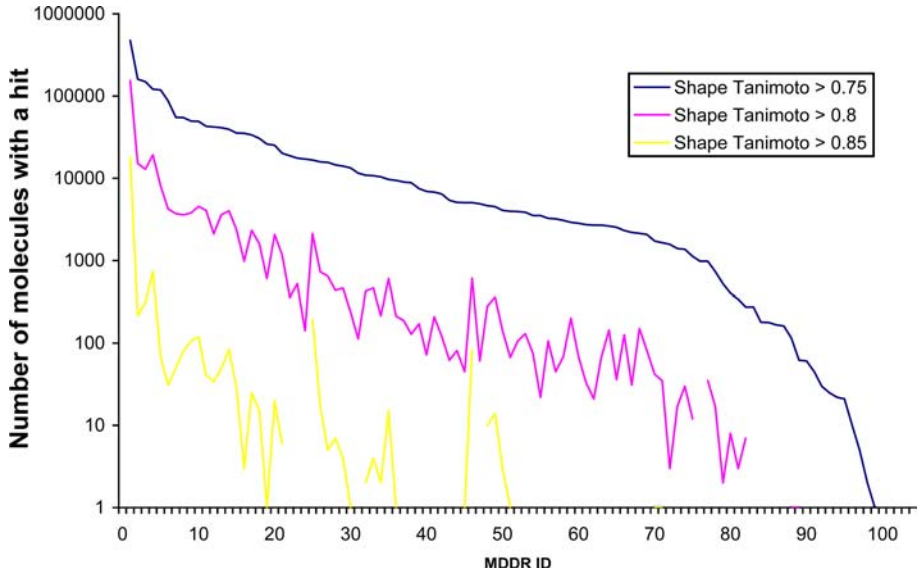

Figure 15. Number of molecules having at least one conformer with a ST greater than 0.75, 0.8 or 0.85 ST against one hundred diverse MDDR structures. All molecules chemically similar to a target (MACCS keys fingerprint Tanimoto >0.65) were pre-filtered.

shows the results at three different levels of shape similarity, ST 0.75, 0.8 and 0.85. The MDDR target structures have been ordered by number of hits greater than 0.75 ST. Hits from molecules that are chemically similar to their target, i.e. with a chemical fingerprint Tanimoto (CT) greater than 0.65, have been pre-filtered and do not appear in the count. Clearly, there is a wide variation in the number of similar structures at any ST and Figures 16 and 17 illustrates that this is primarily a decreasing log-linear function of size. The origin of this behavior is twofold. First, the distribution of molecular sizes in the test set and database as shown in Figure 18. The latter has a median of around 20 heavy atoms, whereas the MDDR targets average closer to 35. The difference reflects the source of the two collections. The vendor set is of typical drug-like size for purchase for lead development, whereas reference shapes predominantly sample larger sizes. The second reason is that the number of available shapes grows exponentially with size [J.A. Grant, in preparation]. As such, the probability of finding a match at larger molecular weight decreases rapidly (and log-linearly).

The results here might be seen as an indication of the limitation of shape encoding, i.e. larger molecules may have too few morphologs to be secure. In fact, the opposite is the case. The number of *possible* structures rises faster than the number of shapes. We postpone illustration of this until the discussion, but the correct conclusion is that chemical identity is only predictably insecure for very small molecules.

If many similarly shaped compounds can be found within a reasonable threshold, the remaining question is how similar are such molecules in their electrostatics? Figure 19 illustrates that a preliminary search on shape can produce a reasonable yield of electrostatically analogous compounds. The enrichment over the total set, i.e. including shape dissimilar compounds, is primarily due to the superposition of the mask function applied to the total electrostatic field, i.e. masked fields cannot overlap well if the masks do not. Also shown here is the effect of considering small, terminal rotors. Rotation about the bonds connecting these groups to the main body of the molecule does not greatly affect the shape of the molecule. However, if this group has a significant dipole perpendicular to the axis of rotation the angular dependence of the electrostatic field can be significant. By checking a

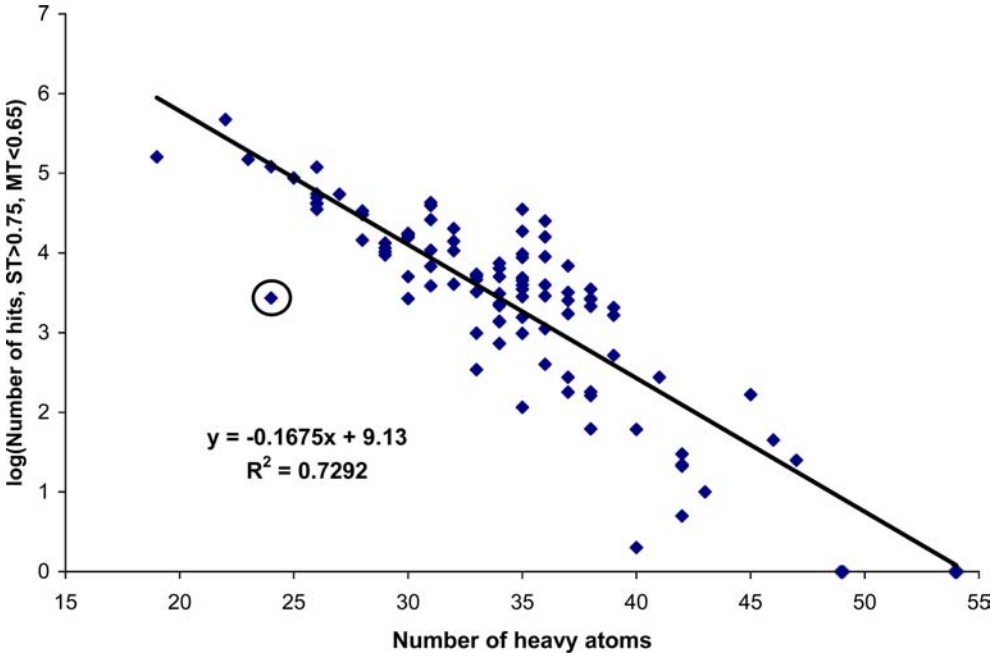

Figure 16. Number of compounds at least 0.65 MACCS Keys Fingerprint Tanimoto distances from a target with at least one conformer with a shape similarity of greater than 0.75 ST, versus the number of heavy atoms of each target. The outlier MDDR_287621 is circled and shown in Figure 17.

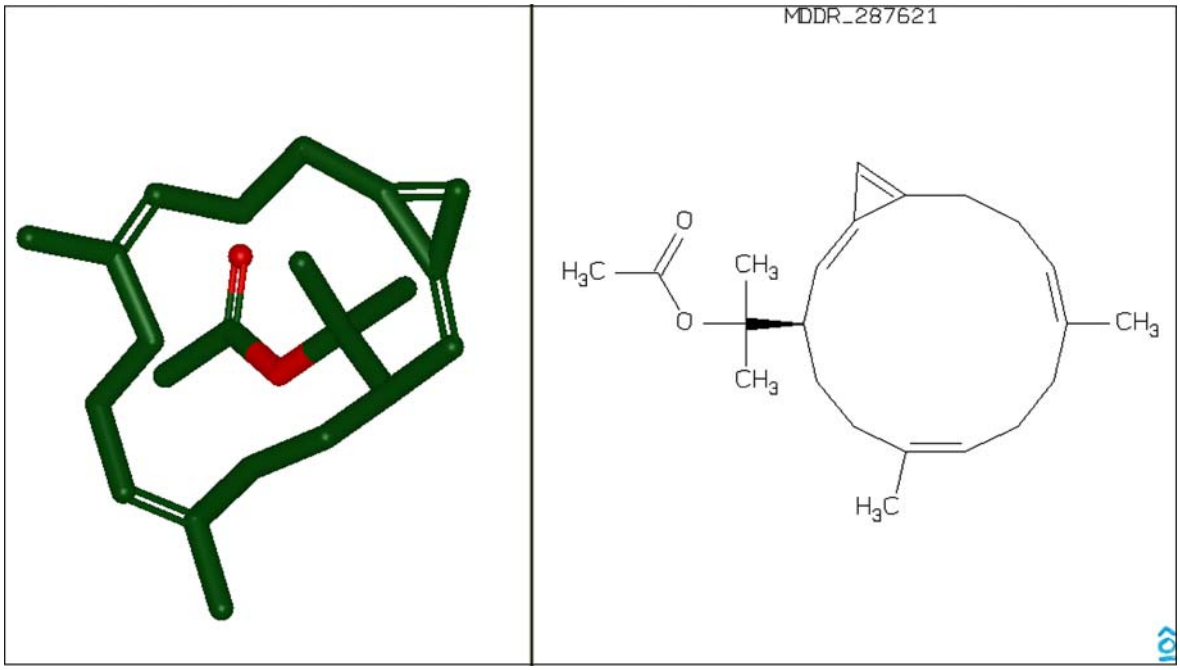

Figure 17. MDDR_287621. A molecule exhibiting an unusual shape.

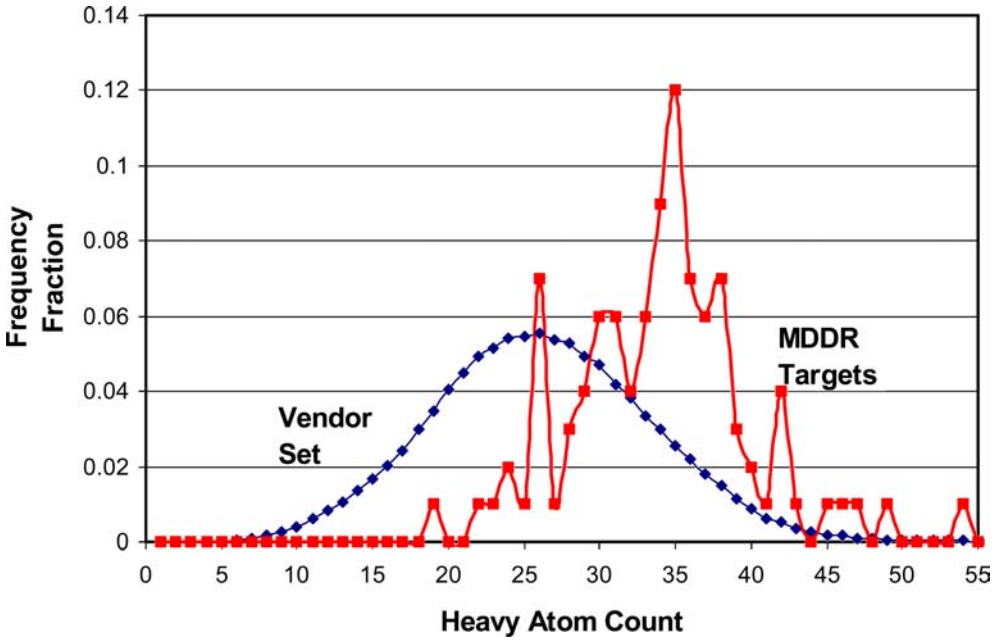

Figure 18. Heavy atom distribution in the 2.3 millions compounds from the vendor collection and of the set of 100 MDDR representative shapes.

set of rotations of such rotors for better EO higher-quality matches are often found. Significant enrichment in target activity has been observed for hits with shape similarity greater than 0.7 ST and electrostatic similarity greater than 0.2ET. Searches against the MACE database show that such compounds can always be found.

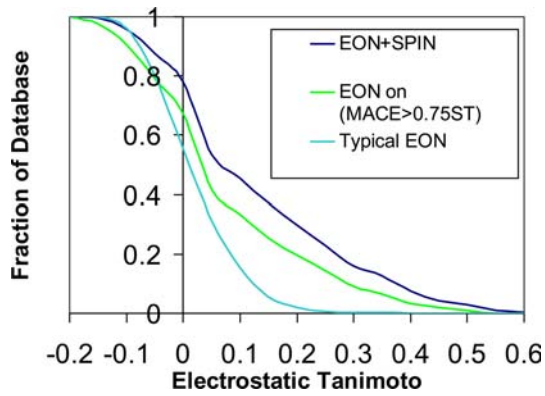

Figure 19. Typical distributions of Electrostatic Tanimoto (ET) over a large database using the program EON. Enriching first by shape greatly increases the probability of electrostatic matching. Furthermore, sampling polar terminal rotor positions can significantly improve this already enriched set.

Electrostatic optimization

The previous section illustrated shape and electrostatic similarity to existing compounds, but much of the argument for anonymity relies on the vastness of potential chemistry from *de novo* design, combinatorial chemistry, inspired synthetic chemistry etc. We want the anonymous compound to be lost in an almost infinite number with similar physical properties. Large-scale searches can find close shape matches in even modest datasets. However, it is rare to discover an ET greater than 0.5. As discussed below, it is likely that this is due to two effects: a greater dimensionality of electrostatic space, or the dual electro-shape space, and the necessity for a good, initial shape match. The former increases the need for sampling and the latter reduces the available sample size. The *de similis* design program WABE is a natural approach with which to investigate this problem. It varies atom types but locks hybridization, i.e. maintains shape but explores electrostatics . As an experiment, one thousand compounds were chosen from the set of shape hits to MDDR_163984. Each compound had at least one conformer with a ST > 0.75. Each was chemically dissimilar from the target (CT < 0.65) and was chemically distinct from each other (intra-CT < 0.65). WABE was then applied to every structure and the maximum electrostatic similarity for each to MDDR_163984 recorded. The results are shown in Figure 20.

As expected, there is substantial improvement in average electrostatic similarity. Interestingly, some compounds did not improve. Investigation proved this due to lack of sampling. WABE is restricted to substitutions that are observed in an ancillary database and, as such, the number of compounds generated varies from single digits to several million, depending on the constitution of

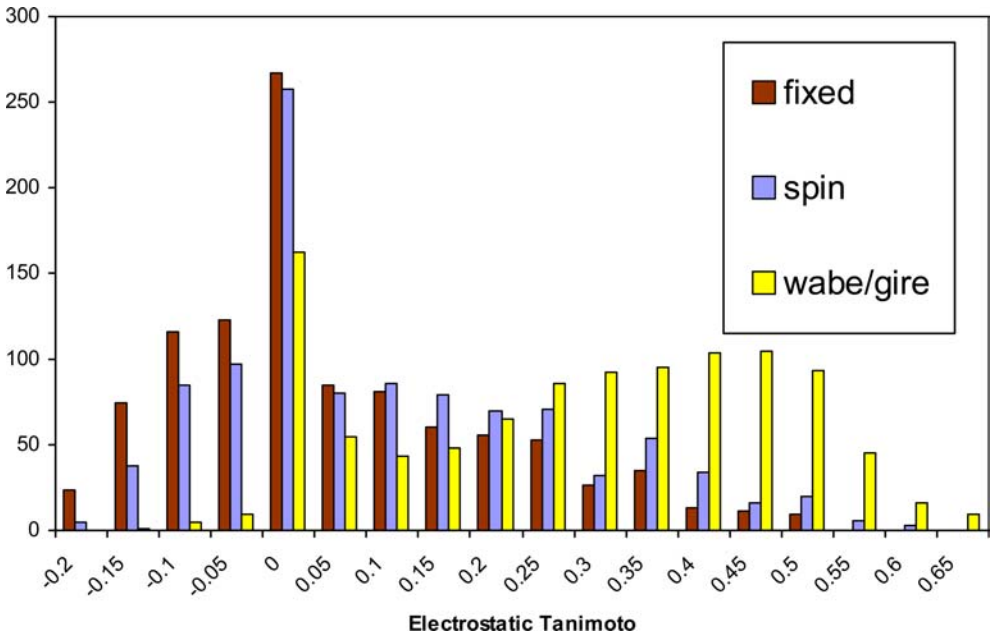

Figure 20. Graph of the distribution of Electrostatic Tanimotos of a thousand chemically distinct structures from the MACE conformer set that had a ST greater than 0.75 to MDDR_163984, and an intra-set Chemical Tanimoto of < 0.65.

the substrate molecule. There was no correlation between initial and optimized Electrostatic Tanimoto, but that between number of compounds sampled and improvement was substantial. The median *de similis* count of the 10 molecules that optimized to an ET greater than 0.6 was ~75,000, compared to ~5000 for the total set. In addition, the median ST was 0.84 compared to 0.78 for the total set, reinforcing the point that shape similarity is a prerequisite for good electrostatic matches.

Discussion

Of shape space and graphs

We will base this discussion mostly on the MACE large-scale searching results, previous work on shape [12, J.A. Grant, in preparation], and a recent publication by Fink et al. [15]. In the latter work, the authors considered molecules derived from all graphs of 11 atoms or less subject to chemical constraints. Graph nodes had a maximum connectivity of four. All three- and four-member rings were removed, as well as non-planar graphs, unfeasible ring bridges and tricyclic bridgeheads. The remaining set of unique, uncolored graphs (1830 in total) was transformed into molecular graphs by the combinatorial introduction of bond-saturations and elemental types of carbon, oxygen, nitrogen or fluorine, removing unfeasible bond-types. This produced a total of about 14 million molecules of less than 160 Da. Our interest is in asymptotic limits as a function of heavy atom count. Log plots of the number of unfiltered graphs, chemically filtered graphs and "reasonable" compounds all show quadratic behavior (parameters in Table 1). A log plot with a quadratic dependence, even a mild one, is surprising. A heavy atom count of 25 would, by extrapolation, produce 10^{26} molecules. And this is with only C, N, O and F, and without conformers, stereochemistry or proton placement. Earlier in this paper we refer to an upcoming work [J.A. Grant, in preparation] that shows the number of unique shapes, for a given ST threshold, is a *linear* exponential function of heavy atom count. As such, the number of graphs must eventually surpass the number of shapes. From our analysis [J.A. Grant, in preparation], the number of graphs exceeds the number of shapes after six atoms. At 25 the ratio is greater than 10^{12}, i.e. if these graphs are uniformly distributed through shape space there are a trillion different graphs close in 3D form to each representative shape assuming only a single conformer/enantiomer/tautomer per graph.

Some general statements can be made on the shape space these molecules occupy. The difference from unity of a Tanimoto measure, also known as the Soergel measure, is a metric for bit patterns. It is a reasonable conjecture that the Soergel of a continuous measure is also a metric, for instance the Shape Soergel (SS). In extensive tests of the SS, no violations to the triangle equality were found except those arising from incomplete optimization [38]. Assuming SS is a metric, shape space can be cast as a sphere of unit diameter in some high-dimensional space. Assuming shapes are uniformly distributed in this space and ignoring edge effects, the ratio of shapes within a radius of R_1 and within a radius of R_2 is $(R_1/R_2)^N$, i.e. the ratio of the volumes of two hyperspheres of dimension N. The curves in Figure 15 of the number of shape hits at 0.75, 0.8 and 0.85 ST, represent query radii of 0.25, 0.2, and 0.15 SS. The ratio of conformer hits, $h(S)$, at these different distances should reflect the dimensionality of the space, i.e.

$$h(S_1)/h(S_2) = (S_1/S_2)^N$$
$$N = \log(h(S_1)/h(S_2))/\log(S_1/S_2)$$

Table 2 gives the average number of conformer hits at each level for the 30 lightest MDDR target

Table 1. The quadratic forms of (log) numbers of graphs and compounds derived from heavy atoms counts one through 11 from the work of Fink et al (ref).

$AQ^2 + BQ + C$ Q = Heavy atom count	A	B	C	R^2	Values at $Q = 11$
Unfiltered graphs	0.0411	0.1142	−0.2976	0.9988	739,335
Chemically filtered graphs	0.025	0.0245	−0.1815	0.9997	1272
Molecules realized	0.0297	0.3125	0.1309	0.9988	11,864,872

Constants and correlation coefficients (R^2) calculated in Excel from values in their paper.

Table 2. The ratios of total shape hits for the 30 lightest MDDR target structures against the 600 million shapes in the MACE database.

ST/Soergel	0.75/0.25	0.8/0.2	0.85/0.15
Log_{10} (total conformer hits)	7.02	5.93	4.59
Log_{10} (Soergel measure)	−0.602	−0.699	−0.824
Predicted dimension (column N and $N+1$, modulo3)	11.2	10.7	10.9

molecules. We chose these 30 rather than the whole set for three reasons. First, a range of 20–30 heavy atoms is a reasonable range for drug-like molecules. Second, the statistics are less reliable for the rest of the set. Finally, we know from other work [J.A. Grant, in preparation] that the dimensionality is not constant with size, but grows. Therefore we choose a restricted range to simplify the analysis, typically for a hypothetical molecule with 25 heavy atoms. The three ratios available from this data and the associated prediction of N are given in the last row. The consensus value of around eleven can be independently checked from the total number of shape hits (ST > 0.75) for this subset. There are $6*10^8$ shapes in MACE and a ST of 0.75 corresponds to a Soergel radius of 0.25 and a diameter of 0.5. Comparing this to the maximum diameter of all of shape space of 1.0, this implies a volume reduced by 2^{11}, i.e. roughly two thousand. This would suggest a total of $(30*6*10^8/2^{11}) = 9$ million hits for the lightest thirty MDDR targets. In fact there are 10.5 million. It should be stressed that this estimation of a dimension to ''shape-space'' is only approximate and only valid for a small range of molecular size. For the purposes of the analysis of molecular obfuscation by shape and electrostatics, however, it is a useful estimate. Interestingly, it also coincides with estimates of dimensionality made by others from considerations of antibody-antigen diversity [39, 40].

Electro-shape

Given that the quadratic coefficients to the logarithmic growth of both graphs and molecules are quite similar (Table 1), the number of colorings for each graph, i.e. the ratio, is approximately log-linear. The number of colorings of any particular graph is roughly $= 10^{\{\text{Number of Heavy Atoms } * \ 0.288\}}$. At 25 heavy atoms there would be about 10^{10} such colorings, although Fink points out that the number derived from each graph is extremely variable, in accord with our experiments with WABE. What does this say to the dimensionality of electrostatic space? We can show the rate of growth in shapes is roughly half the rate of (log) growth of colorings from Fink [J.A. Grant, in preparation]. This suggests that electrostatic space may have up to twice the dimensionality of shape, i.e. 22 dimensions.

As the minimum possible ET is −1/3, the Electrostatic Soergel (ES) measure for electrostatics ranges from zero and 4/3. If our ET threshold is 0.45, our estimate of electrostatic dimensionality would predict about 1.5% of structures having > 0.45 ET from the ratio of hyperspheres. Clearly this is incorrect for the MACE database as a whole, but if we consider the one thousand structures shape-similar (ST > 0.75) to target structure MDDR_163984, this is reasonably accurate (1.9%). Increasing the threshold to ET > 0.6 gives a fraction of $\sim 10^{-5}$. However, compared to the total number of structures generated by WABE for the MDDR_163984 set (50 million) this appears to predict too many hits. Even after adjusting for uneven distribution of sampling we should expect about 60 or 70 structures greater with an ET > 0.6. We note that this is about the number of structures with an ET > 0.55. A likely explanation is that the

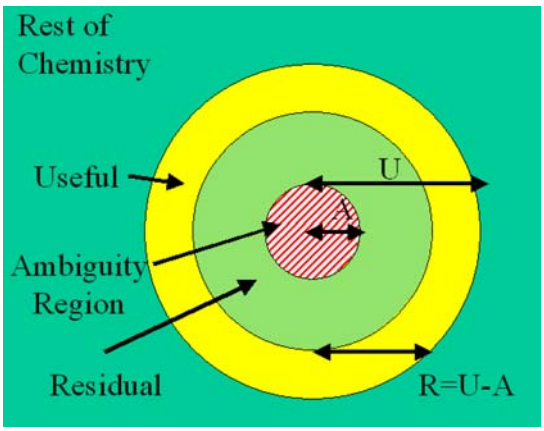

Figure 21. Ambiguity in molecular representation. Given a two-dimensional descriptor space and an inaccurate representation at the center of the concentric circles, while the precise representation is a distance A away (or less). If molecules are considered similar within a radius U, then the residual radius R, i.e. the radius that will not include false positives, is equal to $U-A$.

requirement for shape similarity becomes more stringent as the required electrostatic similarity increases. If we consider an even more stringent measure of ET > 0.75, we arrive at a factor of $4*10^{-10}$. This might explain our lack of observation from a database of only $6*10^{8}$ conformers. If we compare to the number of colorings for a single graph (10^{10}), this would suggest four matches per graph. If there are an astronomical number of graphs to a particular shape, anonymity is still safe, but it does illustrate the facility of electrostatic specificity to make a mockery of large numbers and this is worth considering in the context of ambiguity.

Ambiguity

It is likely a representation of a molecule's shape and electrostatics will not be completely accurate. The results from shape fingerprints, ellipsoids or analytic functions illustrate a degree of imprecision intrinsic to each. One way of considering the effect of error in is illustrated in Figure 21. The true representation lies a distance A from its actual representation in descriptor space. The radius of usefulness, i.e. such that similarity in descriptors is meaningful in terms of function, is U. However, because of the ambiguity introduced due to representational error, we can only assume usefulness with a radius of (U–A) or R. The volume of chemical space enclosed by R is less than that of U, i.e. we will miss some true positives. Figure 21 presupposes a dimensionality of two, but if we use the value of 11, as in the above analysis, the ratio of found positives to all positives is:

$$\text{True Positive Ratio} = \left(\frac{R}{U}\right)^{11} = \left(\frac{U-A}{U}\right)^{11} = \left(1-\frac{A}{U}\right)^{11}$$

Increasing R will recover some true positives, but at the cost of false positives. Of course, in practice things are never this black and white; there will be a probability of equivalent function given a similarity distance, and this function will be convoluted with the error function. However, in this *gedanken* the point at which all true positives

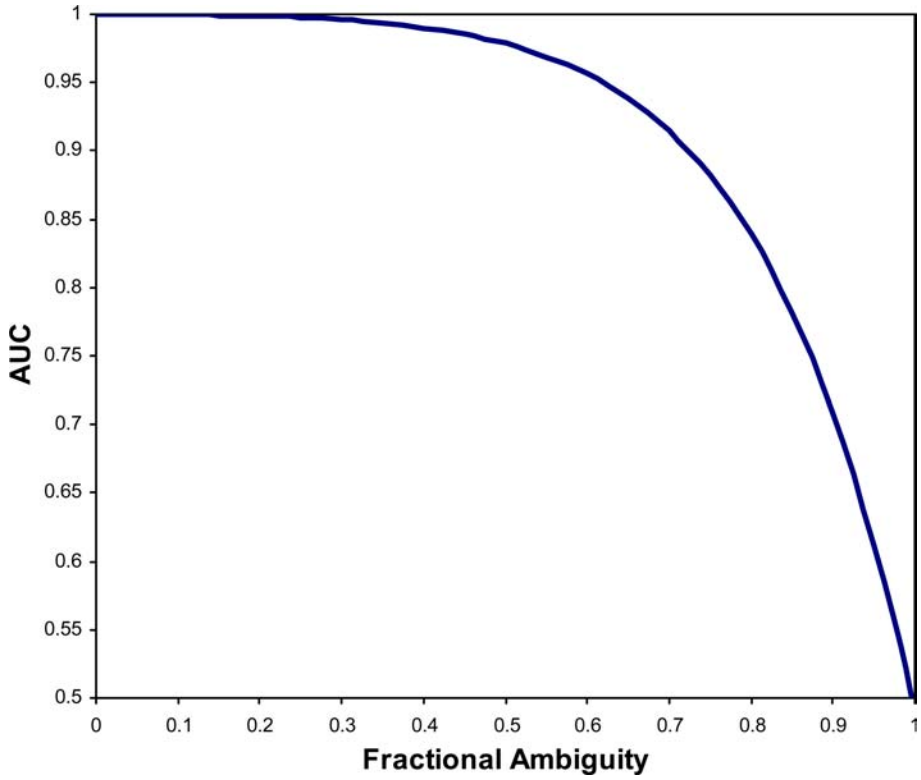

Figure 22. An unambiguous analysis of ambiguity: The AUC of a ROC curve of shape searching, assuming a ST > 0.75 implies functional similarity, as a function of fractional ambiguity.

are returned yields the following ratio of false to true:

$$\text{False Postive Ratio} = \left(\frac{U+A}{U}\right)^{11} - 1 = \left(1+\frac{A}{U}\right)^{11} - 1$$

As an example, at a fractional ambiguity of 10% the useful volume is reduced to ~31% while false positives become 185% more numerous than true positives; at 20% the numbers are 8.5 and 643% respectively.

A formal method of evaluating ambiguous shape searching is to consider the two cutoff radii (U−A) and ($U+A$) as forming a utility range. The radius assumed for shape similarity is then the parametric variable of a receiver-operator curve (ROC). As such, we can calculate an Area Under the Curve (AUC), the traditional measure of a value of a test, as a function of ambiguity. To first order, it can be shown:

$$\text{AUC} = 1 - 0.5 * (1 - ((U-A)/U)^{11}) * ((U+A)^{11} - U^{11})/0.5^{11}$$

The 0.5 in the denominator is the radius of the Tanimoto sphere assumed to contain all structures. Figure 22 shows the behavior of the AUC as a function of fractional ambiguity (A/U). As AUC values greater than 0.95 are usually considered excellent, it is surprisingly how much ambiguity can be introduced and yet still considered valuable.

Now consider the advantage of ambiguity for anonymity. The number of structures within the central sphere of ambiguity consists of all possible structures the anonymization might actually represent. Assume a fractional ambiguity of 10% and a ST cut-off of 0.75. This corresponds to a shape-space diameter of 2*(1−T), i.e. 0.5. Therefore, the fraction of all compounds within the sphere of ambiguity is: $(0.5)^{11}$*$(0.1)^{11} \sim 5$*10^{-15}. Fink et al. give the number of graphs with 25 heavy atoms, at ~10^{16} but this assumes no conformer, tautomers or enantiomers. Stereoisomers at a heavy atom count of 11 increased the count by a factor of three and so this is a lower limit for 25 atoms. We assume perhaps a hundred conformers. Tautomers produce new electrostatic patterns but not shapes. This would suggest around fifteen thousand graphs even in this tiny volume. At 20% ambiguity, close to what is seen in the ellipsoid and analytic representations, this increases to thirty million shapes.

Now let us consider electrostatics. Suppose we chose a shape ambiguity of 20% with, on average, 10^{10} colorings per graph, ignoring tautomeric rearrangements and terminal rotor adjustments for the moment, this would give ~3*10^{17} molecules in the ambiguity zone. This appears to be a very safe number. However, if we now consider the dimensionality of electrostatics, the situation initially looks bleak. There would be no structures predicted with an ET > 0.9 to the true compound, even from a selection of 3*10^{17} molecules, i.e. the actual molecule would be instantly discernable by its electrostatics. Of course, this is not really the case. The ambiguity in shape introduces some ambiguity in electrostatics. And, while shape-to-function is a relatively tight correlation, i.e. ST > 0.75, electrostatics is much more forgiving, e.g. ET > 0.2 representing significant similarity. As such, even a very considerable ambiguity in electrostatics, completely blurring any possibility of a one-to-one comparison would still easily allow retention of activity, albeit increasing the number of false positives. For example, the ROC analysis outlined above applied to electrostatics gives than an AUC of > 0.95 if (A/U) = 1 and U = 0.3, i.e. if we search for ET > 0.4 in an attempt to find actual ET values > 0.7.

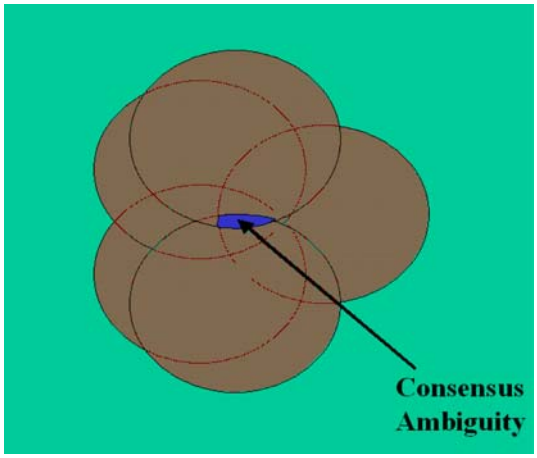

Figure 23. Illustration of the effect of multiple ambiguity regions for a single molecule. If multiple conformations of the same molecule are labeled as such, there is the potential for refinement to a region of shape space containing a practical number of potential compounds.

It is important the preceding discussion be viewed in the context of the possible search strategies outlined in Figure 2. If, and only if, an efficient strategy could be found to avoid fruitless evaluations (Method C), is ambiguity even required to aid obfuscation. Otherwise, it is clear from the number of shapes and electrostatic patterns, e.g. millions of possible close graphs and billions of possible colorings of these graphs, would make a brute force or even heavily restricted method quite infeasible.

Issues

The above analysis suggests it is practical to hide a single conformer's chemical identity with shape and electrostatics, especially given a judicious choice of ambiguity. However, there remains the question of whether a multi-conformer representation is also secure. After all, Holmes was able to decode the Dancing Man cipher because he obtained enough messages to validate his hypothesis. The most secure code is a "one-time pad", i.e. a translation of each word/letter in a message from a list that is never reused. However, the encoding of multiple conformers constitutes the opposite; it is multiple messages about the same molecule. As an example, even with the anonymity provided by lossy representations, if there are multiple representations of a single chemical composition the overlaps of the shape space spheres may be much smaller (Figure 23). False positives also increase, although this is inevitable when using multiple conformers as a target, i.e. if we do not know the bioactive conformation, a multi-conformer target will always be less precise.

In practice the potential drawbacks of multi-conformer presentation may not arise. Consider the following real world example of applying molecular anonymity. Two companies wish to compare chemical libraries, each consisting of many thousands of compounds, and agree to afterwards exchange at most a hundred molecules. Each company anonymizes its collection through an expansion into low-energy conformations, assigning a random label to every *conformer*. Both companies select one hundred shapes from the other's collection, informing each of their choices. At this point the total number of unique molecules, i.e. one hundred minus conformer duplicates, is disclosed. If A has chosen fewer unique molecules than B, then B discards choices until they have equal numbers or vice versa. Alternatively, A and B have a second selection round to fill out their quotas, followed by re-disclosure of unique counts and so forth, terminating when both have one hundred distinct molecules. The key is that at no time does either A or B have to tell each other *which* shapes are not unique. Given only the net number of unique molecules, the combinatorics as to which shapes on either's list are duplicates are such that the information released at any stage is insignificant in all but astronomically unlikely circumstances (for instance, all the shapes chosen belonging to the same compound). The same logic would apply if the transaction were anonymous supply rather than exchange, i.e. if B is a vendor wishing to supply company A with novel compounds and not wishing to disclose actual chemistry. Given that chemical suppliers have been known to sell tautomers of the same compound as different entities (and at different prices!) [38], one could even imagine this extended to molecular shapes!

Perhaps a more cogent concern is outlier behavior. Our experiments and analysis suggest that the density of states is sufficient as to ensure anonymity in general, but this does not mean for all cases. For instance, in Figure 16, illustrating the number of hits from MACE versus heavy atom counts of the MDDR targets, there is an outlier (circled) with 24 heavy atoms that is two orders of magnitude lower than expected. This structure is shown in Figure 17 and exhibits an unusually close contact between a terminal carbon and the 14-member ring. Fortunately, it is not difficult to test if a given shape is uncommon [J.A. Grant, in preparation] and as such to exclude compounds likely to be identified by their shape alone. For instance, it would be hard to disguise the shape of Buckminsterfullerene.

A more substantial problem might be "chemical tramlines". Fink et al. estimated around 14 millions molecules of 11 or less heavy atoms and yet could only find about 36 thousand unique compounds of this size in a set of commercial databases. They suggest this indicates current chemistry is much more restrictive than is generally thought. This may be a case of theoretical hubris, i.e. many of their molecules might turn out not to be practical, no matter how carefully the authors filter their collection. However, given almost three orders of magnitude between avail-

able and theoretical, it is also possible they have a point. If so, then rather than having to search vast hypothetical libraries for matches to shape and electrostatic similarity, it might be only necessary to search a far more conservative set of structures using heuristics as to typical chemistries. Therefore, the large and reassuring numbers presented above may be unsuitable: the 'list' of potential messages may be much smaller than we had hoped. With this in mind we suggest an additional approach: surrogate structures.

Surrogate structures

One method of potentially avoiding legal issues surrounding the disclosure of shape and electrostatic properties, and which avoids the need to blur molecular properties, is to instead disclose those of a different molecule. This approach was suggested by Abagyan et al. [41] and fits well with the work presented. The concept is that the molecule disclosed should have properties similar enough to the undisclosed structure that searches or comparisons made against it would also be useful. We term this approach the "King's Champion" device, or *Campio Regis*, from the Anglo-Norman concept of having a surrogate replace the king in matters of trial-by-combat. The particular attraction of *Campio Regis* is that the legal issues are pushed one level deeper. Not only is it difficult to discover which molecule is being represented, even if this veil is pierced there is the challenge of then discovering *Regis*, the real molecule this surrogate was derived from. Abagyan suggested a series of chemical mutations to arrive at a scrambled proxy but it is not proven that such a method will (a) retain properties of interest and (b) be irreversible.

Our proposal is to use a combination of shape search, or construction, followed by electrostatic optimization via a shape-preserving program such as WABE. An advantage to this scenario is that the *Campio* does not have to be a practical structure. Often in *de novo* or *de similis* design the proposed compound bares no resemblance to anything synthetically accessible. For this application, however, this is acceptable. The chemical properties are merely there to support a morphology and electrostatic profile. As such, the number of potential structures is vast. From Fink et al. the number of unfiltered graphs, i.e. connectivities thought too extreme to form real molecules, was not only significantly greater than the number of graphs judged reasonable (739,335 versus 1272 at 11 heavy atoms) but the rate of increase with respect to heavy atom count was much greater (Table 1). Extrapolating to 25 heavy atoms predicts 10^{28} unfiltered graphs compared to 10^{16} filtered, meaning there are a trillion more graphs in any given shape region. Finally, this approach also mitigates the possibility that molecular obfuscation is doomed because of "chemical tramlines"; the King might be prosaic, but the Champion can be as colorful as required.

Future work and conclusions

Ultimately, our proposal may fail for several reasons, perhaps the most important being paranoia. This is the guiding principle of the pharmaceutical industry and the primary reason for any discussion of anonymization. Of course, just because you are paranoid does not mean they are not out to get you, and much of the fear surrounding the release of information may be justified. When a single compound can generate 50 billion dollars over the lifetime of patent protection there is certainly reasonable cause. Some might see the release of even shape and electrostatic profiles as too risky. For instance, it might give clues as to the targets under investigation, suggest types of similar compounds or illustrate holes in a collection others might exploit. However, the industry also finds itself in a typical "Prisoner's Dilemma". In this scenario, two entities can choose to help each other or not. If one helps and the other does not the result is bad for the helper, good for the helped. If neither helps, the result is neither good nor bad for both. If both help the result is good for both. In 1944, von Neumann and Morgenstern gave weights to each outcome and invented game theory [42]. It is unclear what the weights are for the pharmaceutical industry; but it is entirely possible the best outcome is actually more sharing and less paranoia.

The second possibility is ignorance. Whenever large numbers are bandied about they are typically guesswork. The work of Fink et al. is an exception. Our understanding of the topological spaces that describe the steric and electrostatics fields of small molecules is also improving. The combination, an understanding of graph and field diversity, should make it possible to put reliable numbers to the

potential anonymity of any 3D approach. Even so, not all shapes are equally anonymous. One of the findings from our development of shape fingerprints is the great variability (several orders of magnitude) in the *popularity* of a shape. Losing ones self in a crowd is only possible if there is a crowd. Much more needs to be done to map any possible lacunae in shape space.

The third possibility is craft. Perhaps we underestimate the ability of those skilled in molecular modeling. Maybe the proper choice of algorithm or approach could render any anonymization worthless. Of the three possibilities for failure, this is the hardest to gauge. Stories of new 'unbreakable' algorithms being broken at the same meeting they are announced are part of the standard lore of cryptography [43]. The only practical approach, as with cryptography, would seem to be challenge and competition. Although failure to decrypt in of itself does not guarantee security, it has a pragmatic value. There is no proof that the product of two large prime numbers cannot be factored rapidly, but that has not prevented the RSA algorithm becoming the gold standard of encryption. One roadblock to a competition or challenge is the availability of tools for the production and comparison of the representations suggested here but this does not seem insurmountable, particularly for academic institutions. Molecular anonymization that satisfies the concerns of its potential audience may yet prove impossible, but to borrow once more from Sherlock Holmes, ''The game is afoot!''

Acknowledgements

The authors wish to thank Mike Tennant and Andrew Jennings of Syrrx, Inc. (Takeda Pharmaceuticals) for help in constructing the MACE database and running large-scale ROCS calculations, Geoff Skillman for the curation of the MACE dataset, Robert Tolbert for help with WABE, James Haigh for construction of shape fingerprints on the Fesitin dataset and Mesa Analytics for the use of their fingerprinting code.

References

1. Hohenberg, P. and Kohn, W., Phys. Rev. B, 136 (1964) 864.
2. Bissantz, C., Folkers, G. and Rognan, D., J. Med. Chem., 43 (2000) 4759.
3. Nicholls, A., MacCuish, N.E. and MacCuish, J.D., JCAMD, 18 (2004) 451.
4. Rush, T.S., Grant, J.A., Mosyak, L. and Nicholls, A., J. Med. Chem., 48(5) (2005) 1489.
5. Katz, A.H., Tawa, G.J., Mason, K., Gove, S. and Alvarez, J.C., Comput. Chem., 92 (2004), ACS, Anaheim.
6. Tawa, G.J., Katz, A.H. and Alvarez, J.C., Comput. Chem., 244 (2004) ACS, Philadelphia.
7. Petit-Zeman, S., 4th Horizon Symposium, Oct. 23–25, 2003.
8. Bohacek, R.S., McMartin, C. and Guida, W.C., Med. Res. Rev., 16 (1996) 3.
9. Masek, B.B, Merchant, A. and Matthews, J.B., Proteins, 17 (1993) 193.
10. Grant, J.A. and Pickup, B.T., J. Phys. Chem., 99 (1995) 3503.
11. Grant, J.A. and Pickup, B.T., J. Comput. Chem, 17 (1996) 1653.
12. Grant, J.A., Haigh, J.A., Nicholls, A. and Pickup, B.T., JCICS, Vol. 20, (2005) (In press).
13. Nicholls, A., Daylight Users Group Meeting, Feb. 24–27th, 1998.
14. Hall, D.W. and Spencer, G.L., Elementary Topology, Wiley, 1955, p. 59ff.
15. Fink, T., Bruggesser, H. and Reymond, J.-L, Angew. Chem. Int. Ed., 44 (2005) 1504.
16. Putta, S., Christian, L., Beroza, P. and Greene, J., JCICS, 42 (2002) 1230.
17. Kearsley, S.K. and Smith, G.M., Tet. Comp. Met, 3 (1990) 615.
18. Good, A.C., Hodgkin, E.E. and Richards, W.G., JCICS., 32 (1992) 188.
19. Good, A.C. and Richards, W.G., J. Chem. Inf. Comput. Sci., 33 (1993) 112.
20. Lin, J.-H. and Clark, T., J. Chem. Inf. Model (2005) 45 (in press).
21. Kazhdan, M.M., Thesis Dissertation, Comp. Sci. Dept., Princeton, June 2004.
22. Duncan, B.S. and Olson, A.J, Biopolymers, 33 (1993) 219.
23. Leicester, S.E., Finney, J.L. and Bywater, R.P, J. Mol. Graphics, 6 (1988) 104.
24. Max, N.L. and Getzoff, E.D., IEEE Comput. Graphics Appl., 8 (1988) 42.
25. Ritchie, D.W. and Kemp, G.J., Proteins: Structure, Function & Genetics, John Wiley & Sons, 1999.
26. Ritchie, D.W. and Kemp, G.J.L., J. Comput. Chem., 20(4) (1999) 383.
27. Grant, J.A. and Pickup, B.T., In W. Gunsteren and P.K. Weiner (Eds.), Computer Simulation of Biomolecular Systems, Vol. 3, Kluwer, 1997.
28. Perola, E. and Charifson, P., J. Med. Chem., 47 (2004) 2499.
29. Boström, J., Norby, P.-O. and Liljefors, T., JCAMD, 12 (1998) 383.
30. ROCS 2.0, Omega 1.1, WABE 0.9, EON 1.0, OpenEye Scientific Software, Santa Fe, NM (www.eyesopen.com).
31. Bostrom, J., Greenwood, J.R. and Gottfries, J., J. Mol. Graph Model, 5 (2003) 449.
32. Gasteiger, J. and Marsili, M., Tetrahedron Lett. (1978) 3181.
33. Durant, J.L., Leland, B.A., Henry, D.R. and Nourse, J.G., J. Chem. Inf. Comput. Sci., 42 (2002) 1273.

34. Fingerprint, M., Mesa Analytics & Computing, LLC, Santa Fe, NM, www.mesaac.com.
35. Lu, H., Chang, D.J., Baratte, B., Meijer, L. and Schulze-Gahmen, U., J. Med. Chem., 48 (2005) 737.
36. MDDR (MDL Drug Data Report), MDL, Inc and Prous Science (1995–2005).
37. Corina, Molecular Networks, GmbH Computerchemie Langemarckplatz 1, Erlangen, Germany.
38. Jones, H. Thesis Dissertation, Chemistry Dept., University of Sheffield, 2004.
39. Smith, D.J., Forrest, S., Hightower, R.R. and Perelson, A.S., J. Theor. Biol., 189 (1997) 141.
40. Lapedes, A. and Farber R., Santa Fe Institute working paper 00-01-006.
41. Abagyan, R., Raush, E. and Budagyan, L. Safe Exchange of Chemical Information Symposium, March 14th 2005, San Diego, ACS.
42. vonNeumann, J. and Morgenstern, O., Theory of Games and Economic Behavior, Princeton Univ. Press, 1944.
43. Schneier, B., Applied Cryptography, Wiley, 1995.