Journal of Computer-Aided Molecular Design, **14:** 199–213, 2000.
KLUWER/ESCOM

Morphological similarity: A 3D molecular similarity method correlated with protein-ligand recognition

Ajay N. Jain
UCSF Cancer Center, Box 0128, San Francisco, CA 94143-0128, U.S.A. E-mail: ajain@cc.ucsf.edu

Received 4 March 1999; Accepted 2 July 1999

Key words: molecular diversity, molecular similarity, protein-ligand recognition

Summary

Recognition of small molecules by proteins depends on three-dimensional molecular surface complementarity. However, the dominant techniques for analyzing the similarity of small molecules are based on two-dimensional chemical structure, with such techniques often outperforming three-dimensional techniques in side-by-side comparisons of correlation to biological activity. This paper introduces a new molecular similarity method, termed morphological similarity (MS), that addresses the apparent paradox. Two sets of molecule pairs are identified from a set of ligands whose protein-bound states are known crystallographically. Pairs that bind the same protein sites form the first set, and pairs that bind different sites form the second. MS is shown to separate the two sets significantly better than a benchmark 2D similarity technique. Further, MS agrees with crystallographic observation of bound ligand states, independent of information about bound states. MS is efficient to compute and can be practically applied to large libraries of compounds.

Introduction

Recognition of small molecules by proteins is largely mediated by molecular surface complementarity. Structure-based drug design approaches exploit this in cases where an X-ray structure of the biological target protein exists and docking approaches such as Hammerhead [1–3] and DOCK [4] are applicable. In practice though, it is often the case that a high-resolution structure of an interesting target protein does not exist. In these cases, computational techniques for *quantitative comparison* of molecules become important (e.g., for molecular diversity optimization or virtual screening of chemical libraries). To be of practical utility, these techniques must also be computationally efficient. Several studies have been published comparing the application of various 2D and 3D methods for molecular comparison [5–9]. The 2D methods capture topological similarity in chemical structure and fall into two categories: structural key methods (e.g. the MACCS key method) and 2D hashed fingerprint methods (e.g. work following Willett et al. [10]). Two-dimensional structural key methods count the frequency of occurrence of specific substructures in a molecule and represent the information as a vector. The hashed fingerprint methods enumerate all bond paths up to some length and represent the information as a compressed binary vector. The 3D methods include classic pharmacophoric methods, shape-based methods, and 3D fingerprint-based methods. The pharmacophoric methods seek to identify important points on molecules and represent the distances between them, sometimes including vectorial information. These methods largely ignore hydrophobic shape, which is taken into consideration in the latter methods, to varying degrees. The recent work of Mount et al. offers a 3D method with dense characterization of molecular shape but makes no quantitative comparison to other methods on the protein-ligand recognition problem [11]. Several other 3D shape-based similarity techniques have been described which have focused on superpositioning instead of quantitative similarity comparisons [e.g. 12–15]. The relationship of these methods to those described here will be discussed later.

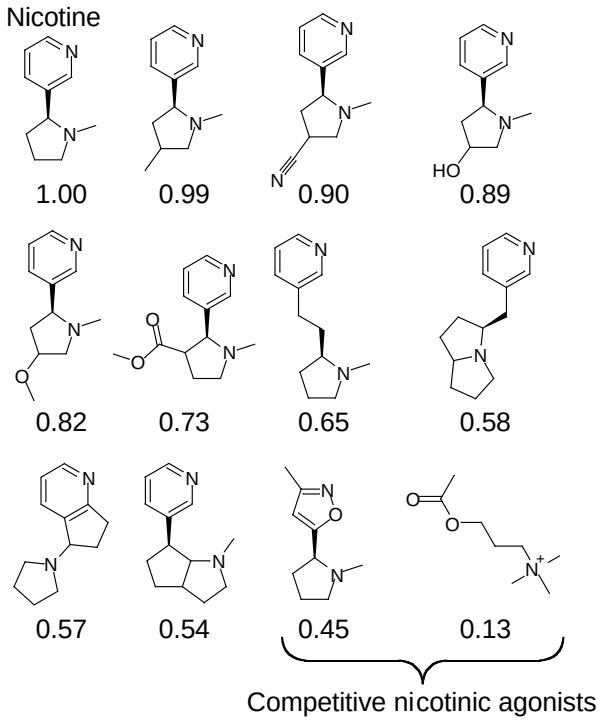

Figure 1. Nicotine (top left), related pyridine analogs and competitive nicotinic receptor ligands (lower right: oxazole derivative and acetylcholine), sorted by topological similarity score.

The comparative studies reach the same conclusion: 2D methods generally outperform 3D methods. Brown and Martin compared 2D structural key methods, 2D hashed methods, and several 3D pharmacophoric methods in their ability to correlate with molecular properties, including activity versus specific protein targets [5, 6]. They found that 2D structural key methods performed better than 2D hashing methods and that all of the 3D pharmacophoric methods tested performed the worst. Patterson et al. explored a similar concept by looking for 'neighborhood behavior' in the space of bioactivity versus molecular diversity [7]. They found that 2D hashed fingerprints performed better than their 3D shape-based topomeric approach (when the complication of common core substructures was subtracted). Briem and Kuntz developed a 3D fingerprint-based method using vectors of DOCK scores, which measured 3D ligand complementarity to a series of protein binding sites [9]. As with the other 3D approaches, a 2D hashed fingerprint-based method performed better in enriching for compounds with activity against specific protein targets.

The common benchmark in all of the above studies was the 2D hashed fingerprint method implemented in the Daylight software package [16]. It performed the best in nearly all cases, but there are important limitations. Figure 1 illustrates this with an example due to Y. Martin [17]. Nicotine and several analogs are shown along with a known oxazole-containing nicotinic agonist and acetylcholine, the natural ligand. The molecules are listed in order of decreasing similarity by the Daylight method. The simple nicotine analogs show high computed similarity to nicotine. However, the oxazole-containing compound, a known, potent, competitive ligand with obvious structural similarity has relatively low computed similarity, and acetylcholine, the natural ligand, is judged to be unrelated using this metric. While the success of the Daylight method is well documented, and its utility in terms of rapid similarity searching is uncontroverted, it is nevertheless counterintuitive that such a technique predicts protein-ligand recognition events better than 3D approaches.

In what follows, a new molecular similarity technique, termed morphological similarity (abbreviated MS), is introduced. It is dependent only on surface shape and charge characteristics of the ligands, and it is computationally tractable even with the complication of conformation and alignment optimization. A molecular similarity benchmark is defined using a data set comprised of ligands whose protein-bound states are known. Ligand pairs observed to bind the same sites form a set of positive examples, and ligand pairs not observed to bind the same sites form the negatives. The performance benchmark is quantitative separation of the set of positive pairs from the negative pairs based on molecular similarity. MS is compared with the topology-based Daylight Tanimoto similarity technique mentioned above (abbreviated TS). MS separates the ligand pair sets significantly better than TS. Further, the specific conformations and alignments of ligand pairs that maximize MS are correlated with crystallographically observed ligand binding geometries.

Methods and data

This section will define morphological similarity and give particulars of the optimization techniques, procedures, and data used in the experiments. Figure 2 shows nicotine and the oxazole derivative from Figure 1 superimposed. In this alignment, the protonated amines are coincident, and the pyridine nitrogen and oxazole nitrogen are capable of accepting a hydrogen-

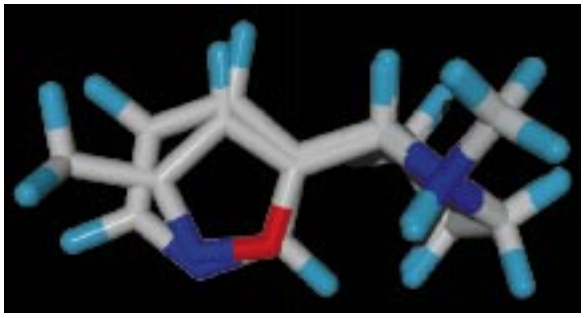

Figure 2. Optimal superposition of nicotine and a competitive oxazole derivative by morphological similarity. The differences in heterocycle substructure do not manifest in the placement of the hydrophobic and polar moieties of the molecules.

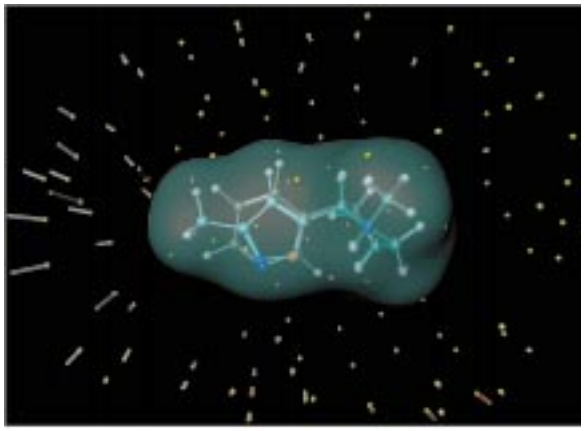

Figure 3. Differences in measurements made of nicotine (shown in gray) optimally superimposed with a competitive oxazole derivative (shown in cyan) from observation points in space. Yellow spheres indicate the position of observation points with significant weight. Gray rods indicate differences in the hydrophobic surface being observed, blue ones indicate differences in positive polar surfaces and red ones negative. The major difference arises from the protruding methyl on the oxazole, not the difference in 2D substructure of the oxazole and pyridine.

bond from the same part of space. Each molecule, viewed from the perspective of a protein, is presenting a similar surface. That is, the molecules shown in Figure 2 look roughly the same from the outside, at least with respect to their surface shapes and disposition of charge. This despite the substructural differences that can confound topological methods.

Morphological similarity compares measurements of molecular surfaces from points in space (as in Compass and the SPERM program [18, 19]). Figure 3 shows a graphical depiction of the measured differences from a set of observation points of the two molecules in their optimal superposition. Yellow spheres indicate the position of the observation points. Gray rods indicate differences in the hydrophobic surface being observed, blue ones indicate differences in positive polar surfaces, and red ones negative. The two primary areas of difference are due to the protruding methyl group and the oxygen of the oxazole derivative. These differences are not substantial when taken in the context of the areas of agreement between the molecules, and the similarity score is 0.94 in the conformations and alignment shown (on a scale of 0 to 1).

Morphological similarity is defined as a Gaussian function of the differences in molecular surface distances of two molecules at weighted observation points on a uniform grid (the yellow spheres in Figure 3 are those points with significant weight). A historical practical limitation of shape-based techniques has been the speed of optimization of conformation and alignment of molecule pairs. Here though, rapid optimization is possible because the molecular observations are *local* and are not dependent on the absolute coordinate frame. So, two unaligned molecules that have some degree of similarity will have some corresponding set of observers that are 'seeing' the same things. Optimization of the similarity of two unaligned molecules is performed by finding sets of observers of each molecule that form triangles of the same size, where each pair of corresponding points in the triangles are observing similar features. The transformation that yields a superposition of the triangles will tend to yield high-scoring superpositions of the molecules.

Morphological similarity definition

The similarity of one molecule to another, each in a particular conformation and alignment in a coordinate frame, is defined in terms of the differences in observations made at a set of observation points on a uniform three-dimensional grid with a spacing λ_0. The positions of the points are denoted o_i. At each point, a weight is defined to limit the points that contribute to the subsequent similarity computation to those that are on the outside of one or another of the two molecules.

At each point, for each molecule, three distances are computed: the minimum distance to the vdW surface, the minimum distance to a hydrogen-bond acceptor or negatively charged atom, and the minimum distance to a hydrogen-bond donor or positively charged atom. In addition, a directionality term is computed for the observations of polar moieties. This corresponds to the directional concordance of the vec-

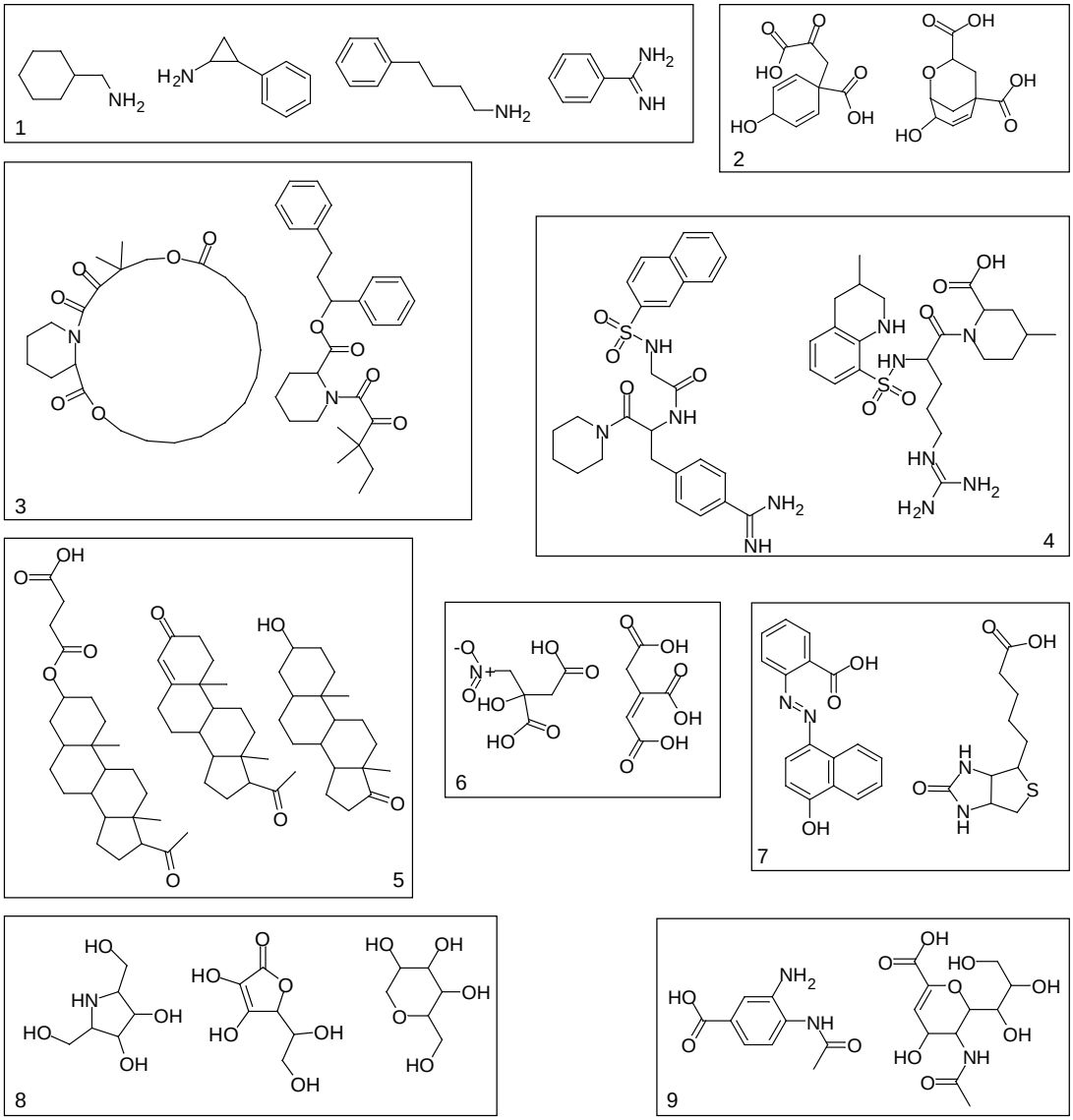

Figure 4. Representative examples of sets of competitive ligands of different proteins used to generate pairs of related molecules for the separability tests: (1) trypsin, (2) chorismate mutase, (3) FKBP, (4) thrombin, (5) DB3 steroid binding monoclonal antibody, (6) aconitrase, (7) streptavidin, (8) xylose isomerase, (9) neuraminidase.

tor from the point to the polar atom and the atom's favored interaction vector.

The similarity between two molecules is simply a normalized sum of weighted Gaussian-like functions of differences in distances from observation points to the molecules. The equations will be presented top-down:

$$f(a,b) = \frac{\sum_i (w_i^a + w_i^b) \begin{bmatrix} \sigma(s_i^a - s_i^b, \lambda_1) \\ + \\ \max(S^{a^+}, S^{b^+})\sigma(s_i^{a^+} - s_i^{b^+}, \lambda_1)\sigma(S_i^{a^+} - S_i^{b^+}, \lambda_2) \\ \max(S^{a^-}, S_i^{b^-})\sigma(s_i^{a^-} - s_i^{b^-}, \lambda_1)\sigma(S_i^{a^-} - S_i^{b^-}, \lambda_2) \end{bmatrix}}{\sum_i (w_i^a + w_i^b)[1 + \max(S_i^{a^+}, S_i^{b^-}) + \max(S_i^{a^-}, S_i^{b^-})]}, \quad (1)$$

$$s_i^a = \min_{j \in a}(d(o_i, a_j) - r_j), \quad (2)$$

$$w_i^a = \sigma(s_i^a - \lambda_3, \lambda_4), \quad (3)$$

$$S_i^a = (1 + c_j)\omega((u_{ij} \cdot v_j) - \lambda_5, \lambda_6), \quad (4)$$

$$\sigma(x, \lambda) = e^{(-x^2/\lambda)}, \quad (5)$$

$$\omega(x, \lambda) = \frac{1}{1 + e^{-x/\lambda}}. \quad (6)$$

Equation 1 defines the similarity *f* of molecules *a* and *b*. It is the sum over all observation points *i* of the weighted difference in observations (*s* and *S*) of the two molecules. Equation 2 defines the direct observations of molecular surface distances (*d* is Euclidean distance, o_i denotes position of observer *i*, a_j denotes position of atom *j*, and r_j denotes the vdW radius of atom *j*). The atom set *a* can be either the full atom set *a*, the positive atom set a^+, or the negative atom set a^-. Equation 3 defines the observation point weighting based on a Gaussian-like function σ (Equation 5) of the nearest distance to the molecular surface. Equation 4 defines the polar strength component, dependent on a sigmoidal function ω (Equation 6) of directionality weighted by a formal charge term (for the atoms sets a^+ and a^-). In Equation 6, *j* denotes the atom of *a* with the surface nearest to o_i, and c_j denotes its charge. The vectors u_{ij} and v_j denote the direction from o_i to a_j and from a_j to the centroid of its substituents, respectively.

Assignment to the sets *a*, a^+, and a^- is done heuristically and is designed to account for the common cases well. The scheme is similar to that used by Hammerhead and Compass [2, 18]. All atoms belong to the set *a*. All hydrogens connected to non-carbon atoms are identified as positive atoms (in the set a^+) as are formally positively charged atoms. All oxygen atoms are identified as negative atoms (in the set a^-). Sulfur atoms not involved in disulfides or bonded to oxygens are in a^-. Nitrogen atoms with three substituents not bonded to hydrogen that are non-planar with respect to their substituents are in a^-. All atoms with no formal charge are assigned a charge of 0. Atoms with positive formal charge distribute the total charge over any attached hydrogens. Atoms with negative formal charge distribute the total charge across any resonant atoms (e.g. the negative charge of a carboxylate is distributed between the two oxygens).

For the experiments reported here, λ_0 was 2.0 Å, λ_1 was 2.0, λ_2 was 0.5, λ_3 was 4.0, λ_4 was 0.2, λ_5 was 0.5, and λ_6 was 0.3. The parameters were not varied to optimize the performance of the technique on the separability problem or geometric prediction task. They were chosen primarily for computational speed. Other parameter sets were not systematically explored. For the similarity computation, only those observation points whose weight exceeded 0.1 were considered.

Similarity optimization

There are two problems in similarity optimization of one molecule to another. The first is rigid alignment of one molecule, or fragment, to another. The second is conformational optimization superimposed onto the alignment problem. The conformational optimization problem is solved as in Hammerhead [1], with a process of fragmentation, conformational search, alignment and scoring, and reconstruction from high-scoring aligned fragments:

1. The input molecule is fragmented at rotatable bonds, avoiding fragmentation into fragments smaller than 6 heavy atoms.
2. Each fragment is conformationally searched up to a user-settable number of conformations (the default is 10).
3. Each fragment conformation is aligned to the target molecule to maximize its morphological similarity to the target (see below). Only the atoms of the fragment contribute to the surface similarity computation.
4. Fragments with the top 10 similarity scores are retained.
5. The placed fragments guide the iterative addition of the remaining molecular fragments, one fragment at a time, taking bond geometries into account while maximizing similarity to the target.

The alignment optimization method differs substantially from Hammerhead and is described in detail below.

Recall from above, that all measurements made of the molecules in the similarity definition are *local*. So, a molecule that is arbitrarily translated and rotated from an initial position will still have a set of observation points that are measuring roughly the same values. Certainly this is true if the observation grid is arbitrarily dense. The problem of optimizing the alignment of one molecule to another is reduced to finding corresponding sets of observers, where each corresponding pair must yield a high local score according to the similarity function. Of course, the internal geometries of the corresponding point sets should also be consistent.

In this work, matching point-pairs from observations of different molecules were generated that had high local similarity (>0.7). From these, triples of point pairs were identified where each edge length difference was small (<1.5 Å) and each edge was greater than 4.5 Å in length. For each matched triangle pair, an alignment was computed to superimpose one to the other. Similar transforms were grouped and incremented in score, with the resulting high-scoring transforms being applied to the molecule to be aligned. Each resulting alignment was then evaluated according to the similarity function.

Since the similarity function is continuous and piecewise differentiable, gradient-based optimization was utilized on partially matched molecules as well as the final alignments. A restraining term preventing excessive steric clashes was employed in the conformational sampling process as well as the conformational optimization process as in Hammerhead [2, 3].

Implementation and speed

The program implementing morphological similarity is written in C and runs on Macintosh, SGI, and PC platforms. A typical full alignment optimization (no conformational component) takes less than 2 s on an SGI R10000 250 MHz machine. Using the conformation subsampling alluded to above, alignment and conformations optimization of a flexible molecule to a single conformation of a target molecule scales roughly linearly, with molecules having 1 rotatable bond being optimized in approximately 5 s, 3 rotatable bonds in 10 s, 5 in 15 s, and so on.

Data

The work of Jones et al. on the GOLD docking technique provides a convenient dataset on which to test the ability of a similarity method to discriminate related molecules from unrelated ones [20]. The set includes 134 crystal structures of small molecules bound to proteins, where there are many cases of multiple different ligands bound to the same (or highly similar) proteins. In the case of pairs of molecules that bind the same site, one expects that an effective similarity technique will yield high scores relative to pairs of molecules not observed to bind the same sites. That is, the technique should *separate* the populations of molecule pairs.

Figure 4 shows representative examples of grouped sets of ligands that were used to generate a population of positive pairs (the full set comprised 74 pairs). The ligands in each group all bind competitively at the same or highly similar sites. The identity mapping was excluded (i.e., the ligand of 3PTB was not compared to itself). A population of negative pairs was generated by randomly choosing 680 pairs of ligands, each of which did not belong to the related pairs set. The full list of ligand pairs used in the calculation, cross-referenced to the names from the GOLD set, is provided in the Appendix.

Pairwise similarity estimation

For each pair, the morphological similarity was computed as follows. For each comparison (molecule A to molecule B), the target molecule B was conformationally sampled, and the similarity reported was the average of the optimum match of A to each of the conformations of B. In this manner, the ability of A to mimic B was measured. Note that A to B is different than B to A, so both measurements are included in the resulting distribution. In cases where, for example, A is significantly more flexible than B, A may mimic B well, but not vice-versa. In such cases, the maximum similarity over all conformations may be a reasonable choice of overall similarity of B to A. However, the average of the optimized matches was used in order to avoid spuriously relating two compounds where a small subset of their mutual conformational space matched, but the remainder did not.

A maximum of ten conformations was used for sampling the target molecules for matching. The conformations were selected to be maximally different by rms deviation. Also, in the interests of computational speed, a maximum of ten conformations were used in sampling the fragments of the molecules to be aligned. Control experiments to determine the impact of the conformational subsampling showed that there was not a significant impact on the estimated similarity scores. Note that for ligands that were identical but from different complexes (a very small fraction of the related pairs used), the morphological method still was required to match a molecule to sampled conformations of itself beginning from unrelated conformations.

The topological similarity technique was applied using the Daylight Toolkit and the supplied contributed code for SMILES format generation (syb2smi) as well as for the Tanimoto similarity computation (simatrix). The default parameters were employed. Note that for TS, the identical molecule pairs automatically yielded similarities of 1.

Results and discussion

This section first presents the results of the separability analyses of the positive pairs and negative pairs. Second, examples of the geometric predictions made by the technique on three examples from the GOLD dataset are presented. Lastly, MS is applied to superpositioning diverse ATP competitive kinase inhibitors.

Separability of related from unrelated ligands

Both MS and TS were applied to the positive and negative pair sets described above. Figure 5 shows the smoothed probability density functions and corresponding cumulative distribution functions for the two sets of data for MS (panel A) and TS (panel B). For MS, the distributions are partially overlapping, but are separable to a large degree. The distance between the inflection points of the cumulative distribution functions is indicative of the separation. For TS, the distribution of related pairwise similarities is quite diffuse, with a nearly uniform distribution from 0.1 to 1.0. However, the distribution of unrelated pairwise similarities remains partially separable from the broad distribution of the positives.

The two methods have different distributional characteristics, but direct comparison of performance can be made using receiver operator characteristic (ROC) curves, which compare true and false positive rates for all possible thresholds of similarity. ROC curves for the two methods are plotted in Figure 6. The abscissa corresponds to the false positive rate, and the ordinate corresponds to the true positive rate. For the morphological technique, it is possible to reject 95% of the unrelated pairs while rejecting only 23% of the related pairs. For the topological technique, to achieve 95% rejection of unrelated pairs, one must also reject 42% of the related pairs (the vertical line highlights this difference). This corresponds to an 87% increase in false negative rate using TS to achieve statistical significance at $p = 0.05$ in a single pairwise comparison. Conversely, if one is willing to tolerate a particular false negative rate, say 30%, there is a 7-fold increase in the false positive rate (14% versus 2%) in TS versus MS (highlighted by the horizontal line). The morphological technique performs uniformly better than the topological technique at separating related from unrelated ligand pairs.

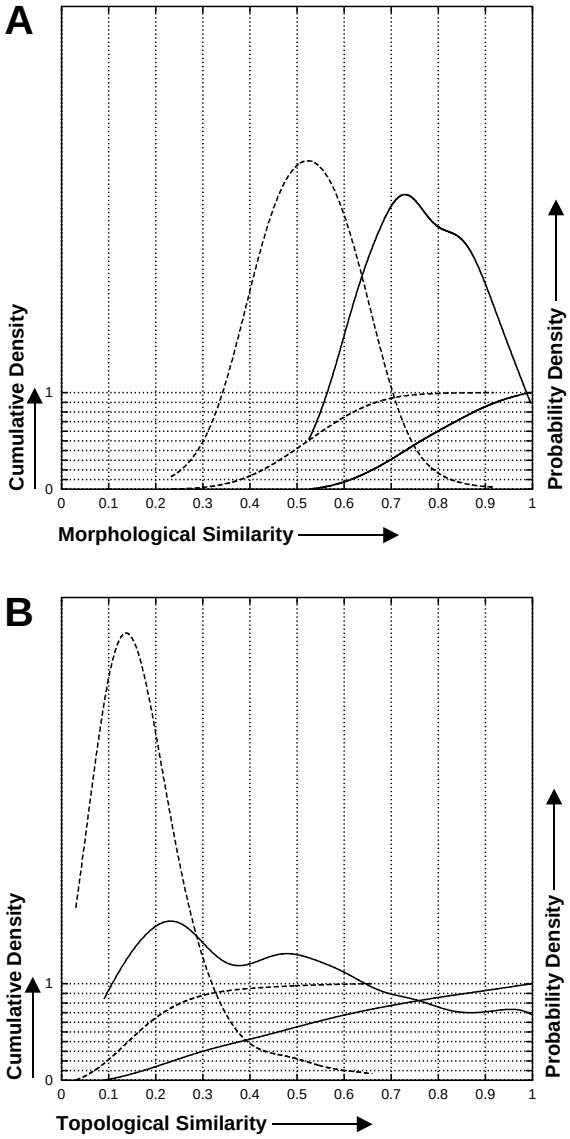

Figure 5. (A) Smoothed probability distributions of related and unrelated ligand pair similarities for the morphological similarity technique. Solid lines represent the probability density function for related pairs of molecules and the corresponding cumulative distribution function (in the range 0 to 1). Dashed lines are for the unrelated pairs. (B) Smoothed probability distribution functions of related and unrelated ligand pair similarities for the topological similarity technique.

Relationship to bound states of ligands

If MS is detecting aspects of ligand similarity that are directly correlated with protein recognition, then the maximally similar conformations and alignments of molecule pairs under MS should resemble their relative bound states. Figure 7 shows three examples

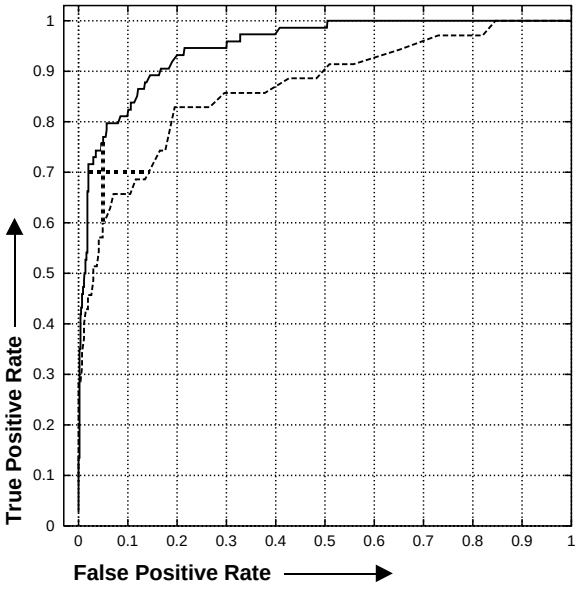

Figure 6. ROC curves for morphological versus topological similarity methods for separation of related from unrelated ligand pairs. Morphological similarity is plotted with a solid line and topological similarity with a dashed line.

of typical pairs missed by topological similarity that are retained by morphological similarity at the $p = 0.05$ cutoff of similarity. They are displayed in their maximally similar conformations and alignments. The molecules exhibit quantitatively similar surfaces in these alignments. Figure 8 shows the same molecule pairs in their experimentally determined configurations within the active sites of their cognate proteins compared with those predicted by MS. The proteins were aligned to minimize alpha-carbon rms deviation. The bound ligands are shown with gray carbons in each panel. The MS-optimized conformations are shown with cyan carbons. The MS-optimized pairs from Figure 7 were aligned to each respective active site by superposition of the ring of the left-hand molecule of each pair to the bound ligand. The right-hand molecule was carried in its relative alignment to the left-hand molecule.

For the ligands of chorismate mutase, there is almost no variation in the predicted versus crystallographically observed conformations. In the trypsin case, the amine of the two aminomethylcyclohexane conformations and the C4 carbons are well matched, but the cyclohexane itself is flipped. Since benzamidine is symmetric, there is no way for the method to prefer one superposition versus the other of the two ligands. In the xylose isomerase case, the best-matching similarity induced model generates an incorrect mapping of functional elements of the two molecules. Ascorbic acid is rotated and shifted relative to the correct orientation. However, the gross relationship of the two molecules is maintained.

Superpositioning of multiple ligands

For competitive ligands numbering more than two, it is also possible to generate geometric hypotheses, but it is slightly complicated by the combinatorics. There are order N^2 pairwise similarities to consider in seeking a set of conformers in a particular alignment that maximizes the joint similarity of all ligand pairs to each other. Any population fitness optimization procedure may be used to do this, and an approach similar to a genetic algorithm [21] was used in the following example where four diverse ligands were under simultaneous consideration.

Figure 9 shows five ligands of a cyclin-dependent kinase (CDK2): ATP, staurosporine, olomoucine, deschloro-flavopiridol (DCF), and isopentenyl adenine (IPA). The binding geometry of each ligand is known through crystallographic observation [22]. The drawings in the figure depict the relative orientations in the ATP-binding site. Note that the three purine derivatives all have different orientations of the purine ring, and the other two molecules have very different underlying chemical structures.

Beginning from conformations unrelated to those observed crystallographically, the first four molecules were optimized for mutual similarity (ATP was truncated to adenosine for computational speed). Figure 10 shows the crystallographically observed relationship of ATP (orange) and staurosporine (green), aligned based on the alpha carbons of CDK2 (PDB references: 1HCK and 1AQ1, respectively). Staurosporine extends its hydrophobic 'arms' into areas not probed by ATP, and ATP extends into a phosphate binding domain that is organized around a magnesium ion.

Staurosporine is a convenient background on which to display the relative orientations of the other four ligands in the similarity-optimized model. The first three overlays of Figure 11 show the mutually optimized conformations of adenosine, DCF, and olomoucine, with staurosporine's optimal conformation in orange. IPA was optimized for summed similarity to the overlap of the first four, resulting in the last overlay. Adenosine's predicted configuration corresponds very closely to that shown in Figure 10. The structures of DCF and olomoucine bound to CDK2 are not

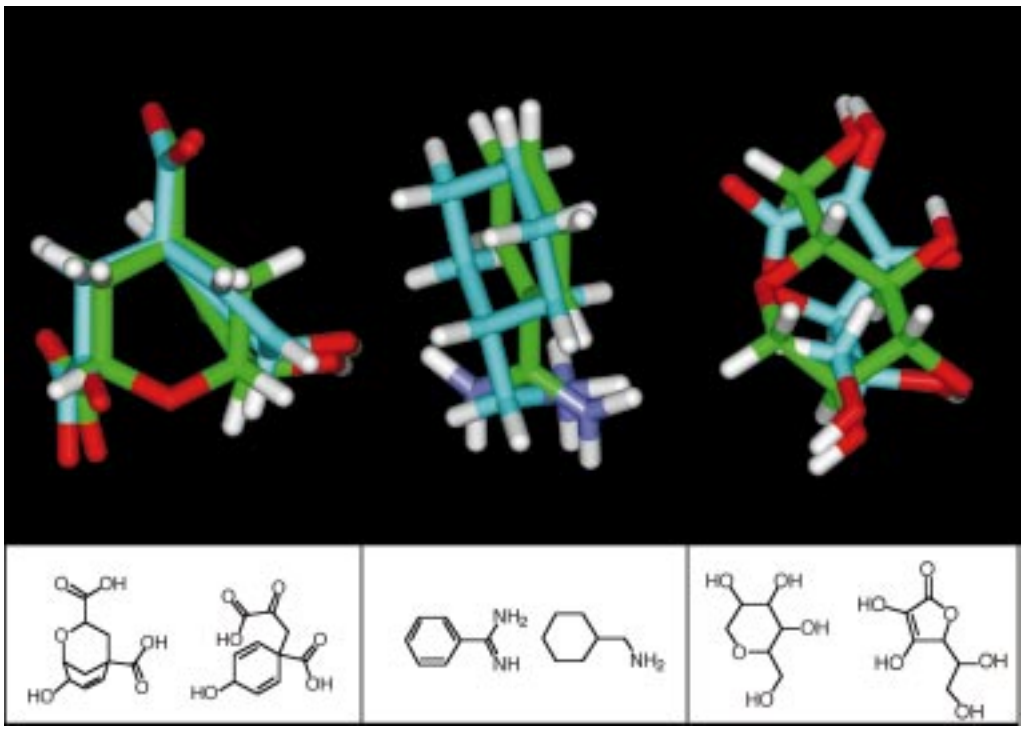

Figure 7. Examples of ligand pairs detected at $p = 0.05$ for morphological similarity that are missed by topological similarity. The ligands are shown in their optimal mutual superpositions by MS. Left are two ligands of chorismate mutase, middle are two ligands of trypsin and right are two ligands of xylose isomerase. Each pair has the left-hand ligand in green, with the right-hand ligand in cyan.

yet available in the Protein Data Bank, but the optimized relationship shown corresponds well to those described in Kim et al. [22] (this is cartooned with the 2D depictions). The minor exception has to do with the pendant phenyl of olomoucine, which is folded back in the similarity optimized model but is coincident with the phenyl of DCF in the experimentally observed conformation. Another overlay of olomoucine scores almost as well, with the molecule flipped left to right, superimposing its pendant phenyl onto the right-hand side of staurosporine and its N-H onto the N-H of staurosporine.

This predicted orientation of IPA corresponds quite closely to the experimentally observed relationship, despite the potentially complicating factor of a different purine orientation relative to adenosine and olomoucine. The mean similarity score of IPA against the other 4 was in the 81st percentile of all compounds from the separability experiment tested against the same model. Of the molecules that scored as well as IPA, nearly half were nucleoside analogs. IPA is a weak inhibitor, with an IC_{50} of 50 μM versus CDK2 [22]. In a library of compounds, one could eliminate roughly 80% of the compounds and still detect IPA and other weak inhibitors using this crude unweighted similarity-based model.

To achieve better enrichment, more refined, possibly weighted, similarity models would need to be employed. In weighted models, lack of dependence of binding affinity to structural variation in ligands that are consistent with a single part of space could be represented explicitly with low weight. Similarly, tight dependence on specific molecular features would also be explicit (e.g. the common hydrogen-bonding interactions among the CDK2 ligands). Weighted models should yield more accurate geometric alignments, as well as offering the potential of quantitative predictions, as in Compass [18].

Discussion

It should not be surprising that a three-dimensional technique that observes only the molecular surface is able to outperform a two-dimensional topological approach in recognizing ligands that share properties that allow them to bind the same protein sites. Consider that a protein binding site can be well-

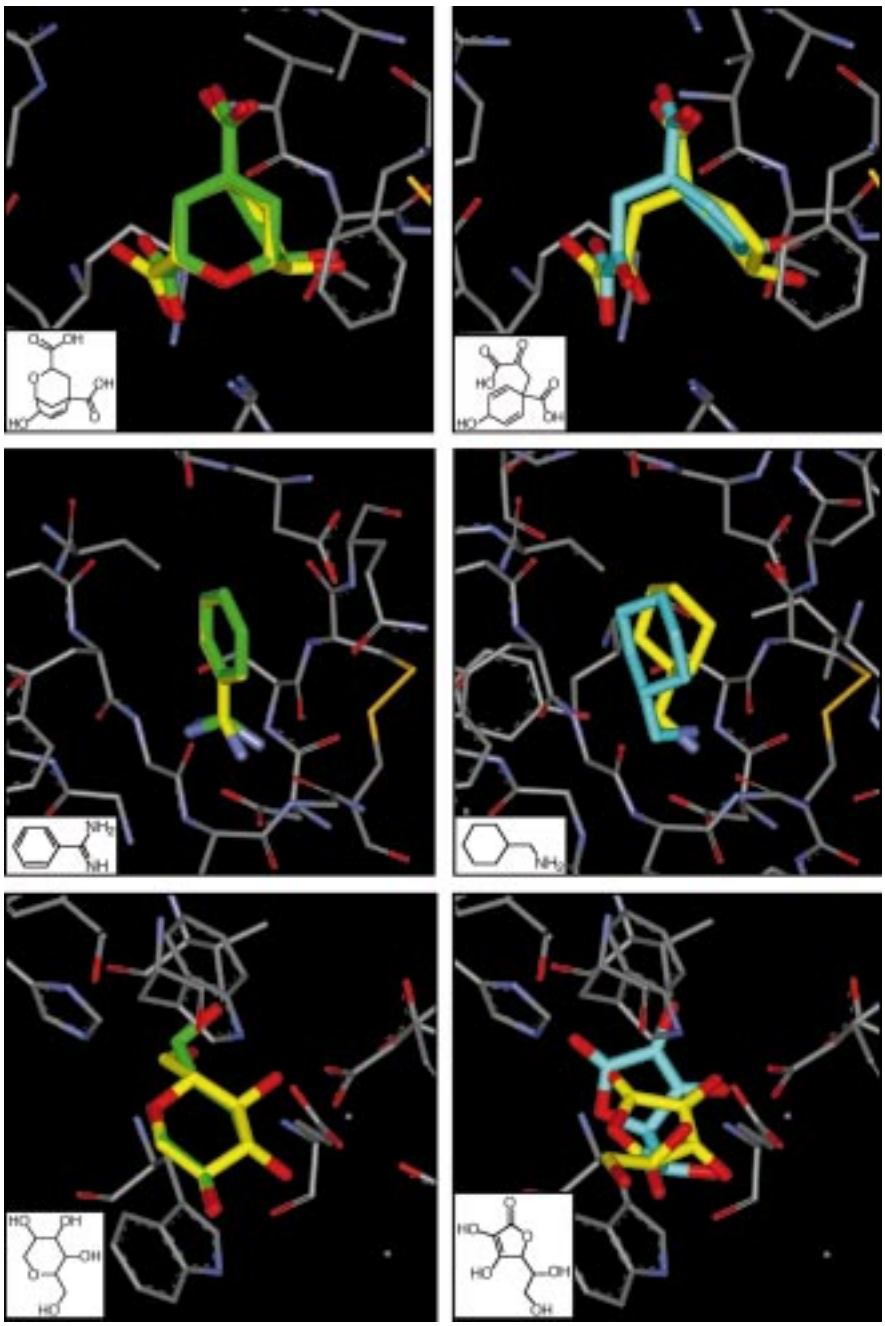

Figure 8. Comparison of expected versus experimentally determined protein-bound geometric relationships between ligand pairs. Each left-hand panel shows the relationship of the predicted conformation of the left-hand ligand (in green) to its observed bound state (in yellow). Each right-hand panel shows the relationship for the right-hand ligand (in cyan) compared to its observed bound state (in yellow). The molecules are aligned as in Figure 7.

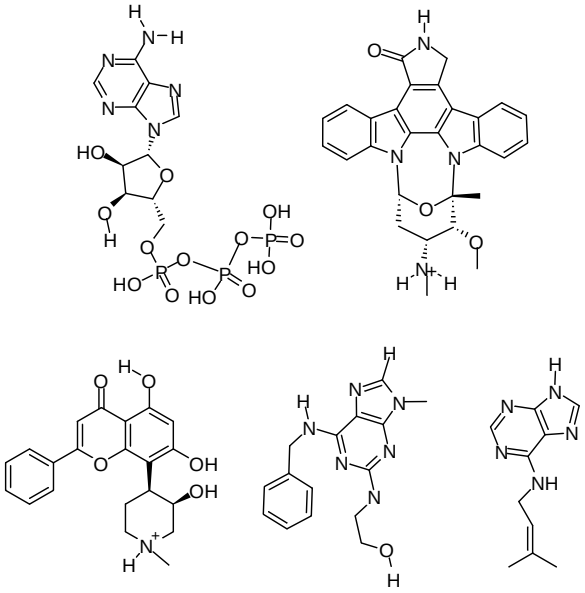

Figure 9. CDK2 ATP-binding site ligands: ATP, staurosporine, des-chloro-flavopiridol, olomoucine, and isopentenyladenine (shown in orientations corresponding to their relative relationship in the CDK2 active site). Hydrogens participating in important interactions are explicitly bonded.

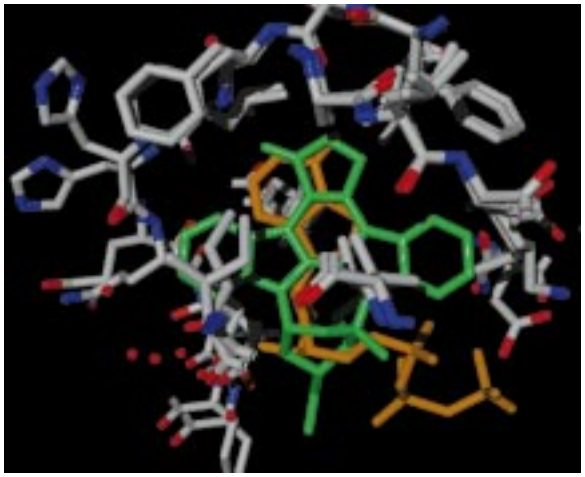

Figure 10. Experimentally determined overlap of ATP (orange) and staurosporine (green) bound to CDK2.

approximated by a collection of spheres of different sizes which respond differently to hydrophobic, positively charged, and negatively charged surfaces. This is essentially the definition of morphological similarity. The weighted observation points form a perfectly shaped binding pocket around a ligand. The distances from the observation points induce spheres whose shape is complementary to the ligand's. Fitting another ligand into that binding pocket and matching the distribution of surface properties is the process of similarity optimization.

Returning to the nicotine example presented in the Introduction (see Figure 1), MS yields similarity scores to nicotine of 0.94 for the oxazole derivative and 0.65 for acetylcholine, with the remaining ligands spread between these two, excepting the simple methyl analog, which is ranked between nicotine and the oxazole derivative. The MS value of 0.65 for acetylcholine is just at the level of significance at $p = 0.05$, based on the distributions computed for the unrelated pairs set. For TS, the value was 0.13, which is embedded in the unrelated pairs distribution for TS. Both the relative and absolute MS scores of these ligands are more in concordance with expectations than for TS.

Shape-based molecular similarity techniques that address some of these issues have also been described by other researchers. Masek's molecular skins approach [12,13] defines a shell around ligands with finite thickness and relates molecular similarity to the maximal intersection volume of the skins of two molecules. The HBMATCH approach of Mills et al. [14] defines favorable hydrogen-bonding interaction sites around ligands and bases superpositions on minimizing deviations between the preferred hydrogen-bonding points of one molecule to another. Both approaches are similar in spirit to what is described here and have demonstrated success on the superpositioning problem.

The work of Lemmen et al. on FlexS is also related [15], and reports results on enrichment for similar biological activity, following Briem and Kuntz [9]. The FlexS approach follows a fragment-based paradigm where the user defines a base-fragment from which a molecule is incrementally constructed to match a target. FlexS uses a similarity function based on soft volumes defined by Gaussian functions. The authors present extensive tests of superpositioning accuracy for ligands matched to the bound conformations of other ligands of the same cognate protein. On a single enrichment experiment involving RGD peptidomimetics, the technique showed enrichment for active molecules versus unrelated molecules. In the present work, the emphasis has been on recognition of similarity and dissimilarity across a broad variety of ligand types. More extensive tests of quantitative superpositioning are planned.

The fastest of the techniques mentioned above is FlexS, with flexible match times of a few minutes on flexible ligands for which an appropriate base frag-

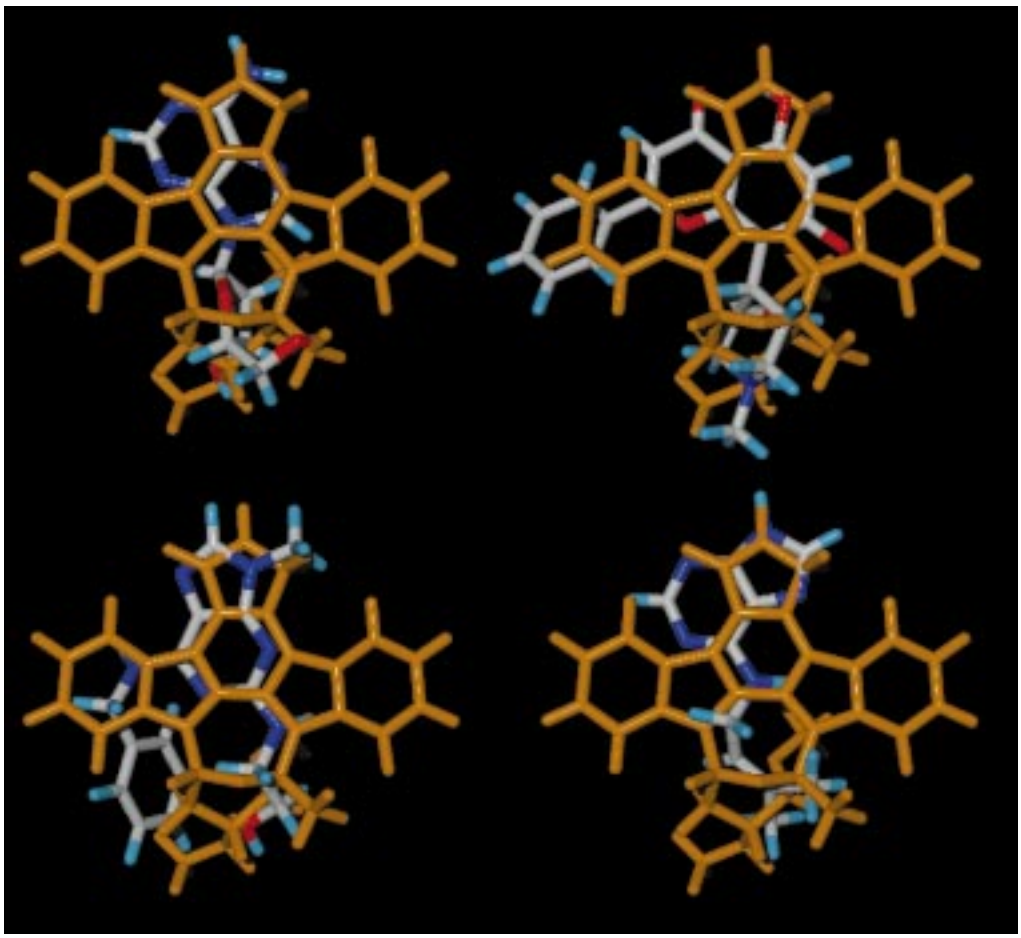

Figure 11. Predicted geometric relationship of five ligands of CDK2, generated independent of experimental observation. Staurosporine is used as a common reference (in orange).

ment has been placed with respect to the target by the user or another program. The morphological similarity function can be optimized (fully automatically) for the match of a flexible ligand to a rigid target in approximately 5 s per rotatable bond. A single rigid optimization takes less than 2 s. The speedup is primarily due to the similarity definition being *local* at different points in space around the molecules to be matched. This is exploited in the alignment procedure. The procedure only generates and tests putative alignments that are expected to be high-scoring based on a local similarity computation involving three corresponding points for each ligand.

Practical applications for such 3D methods exist in molecular diversity optimization and lead optimization. Over 200 000 molecules have been profiled by MS. Computations involving millions of MS-based molecular comparisons are performed routinely using estimation techniques similar to those reported by Mount et al. [11]. Note that even with the encouraging results on geometric prediction, methods such as MS do not obviate the need for crystallographic determinations of the bound states of ligands. Such data will continue to be important for the development of better empirical computational techniques as well as the design of potent and specific ligands. However, in the many cases where crystallographic determinations are not feasible, techniques such as those described here offer a way to reduce the number of competing hypotheses for the geometric relationship of a set of ligands. This can focus lead optimization strategies.

Despite many attempts to define and quantify 3D molecular comparison techniques in terms of correlation with protein-ligand recognition, the field has met with little success compared with widely used, computationally efficient 2D methods. This apparent paradox

has been resolved by applying a technique that is exclusively dependent on the 3D surface-shape and charge of ligands. By directly modeling the dominant factors in protein-ligand recognition, the technique mimicked the outcome of protein-ligand recognition. The practical consequence is a general molecular similarity method. Since the method lends itself to a fast optimization strategy, the technique has broad application in molecular diversity optimization, quantitative structure-activity prediction, de novo ligand design, and related processes important to drug discovery. The broader lesson is that reasoning about protein-ligand recognition events should lie at the molecular surface.

Acknowledgements

I would like to thank Sung-Ho Kim for pointing out the CDK2 example and Alan Roter, Dave Morgans, and Keith Bostian of Iconix Pharmaceuticals for corporate support and technical discussions.

References

1. Welch, W., Ruppert, J. and Jain, A.N., Chem. Biol., 3 (1996) 449.
2. Jain, A.N., J. Comput.-Aided Mol. Design, 10 (1996) 427.
3. Ruppert, J., Welch, W. and Jain, A.N., Protein Sci., 6 (1996) 524.
4. Kuntz, I.D., Blaney, J.M., Oatley, S.J., Landgridge, R. and Ferrin, T.E., J. Mol. Biol., 161 (1982) 269.
5. Brown, R.D. and Martin, Y.C., J. Chem. Inf. Comput. Sci., 37 (1997) 1.
6. Brown, R.D. and Martin, Y.C., J. Chem. Inf. Comput. Sci., 36 (1996) 572.
7. Patterson, D.E., Cramer, R.D., Ferguson, A.M., Clark, R.D. and Weinberger, L.E., J. Med. Chem., 39 (1996) 3049.
8. Cramer, R.D., Clark, R.D., Patterson, D.E. and Ferguson, A.M., J. Med. Chem., 39 (1996) 3060.
9. Briem, H. and Kuntz, I.D., J. Med. Chem., 39 (1996) 3401.
10. Willett, P., Similarity and Clustering in Chemical Information Systems. Research Studies Press, Letchworth, 1987.
11. Mount, J., Ruppert, J., Welch, W. and Jain, A.N., J. Med. Chem., 42 (1999) 60.
12. Masek, B.B., Merchant, A. and Matthew, J.B., Proteins, 17 (1993) 193.
13. Masek, B.B., Merchant, A. and Matthew, J.B., J. Med. Chem., 36 (1993) 1230.
14. Mills, J.E.J., Perkins, T.D.J. and Dean, P.M., J. Comput.-Aided Mol. Design, 11 (1997) 229.
15. Lemmen, C., Lengauer, T. and Klebe, G., J. Med. Chem., 41 (1998) 4502.
16. James, C., Weininger, D. and Delany, J., Daylight Theory Manual. http://www.daylight.theory.com/dayhtml/theory/theory.toc.html
17. Martin, Y., Brown, R.D., Bures, B.G. and DeLazzer, J., Experience with the Application of Computers to Library Design. Cambridge Healthtech Institute's Second Annual Conference on Chemoinformatics, Boston, MA, June 1998.
18. Jain, A.N., Harris, N.L. and Park, J.Y., J. Med. Chem., 38 (1995) 1295.
19. Van Geerestein, V.J., Perry, N.C., Grootenhuis, P.D.J. and Haasnoot, C.A.G., Tetrahedron Comput. Methodol., 3 (1990) 595.
20. Jones, G., Willett, P., Glen, R.C., Leach, A.R. and Taylor, R., J. Mol. Biol., 267 (1997) 727. http://www.ccdc.cam.ac.uk/prods/gold/index.html
21. Koza, J.R., Genetic Programming: On the Programming of Computers by Means of Natural Selection, MIT Press, Cambridge, MA, 1991.
22. Kim, S.-H., Schulze-Gahmen, U., Brandsen, J. and Filgueira de Azevedo, W., Prog. Cell Cycle Res., 2 (1996) 137.

Appendix

The following two lists reference the test case names in the GOLD dataset [16]. Ligands that formed covalent adducts with the protein were eliminated due to the difficulty in interpreting the atoms with partially filled valence. Enzymes with catalytically active Zn^{++} ions were also eliminated from consideration due to the overriding influence of Zn-based interactions with specific small substructural components of the ligands.

All related pairs used: (1tng 1tni) (1tng 1tnl) (1tng 3ptb) (1tni 1tng) (1tni 1tnl) (1tni 3ptb) (1tnl 1tng) (1tnl 1tni) (1tnl 3ptb) (3ptb 1tng) (3ptb 1tni) (3ptb 1tnl) (1dwd 1etr) (1etr 1dwd) (1dr1 4dfr) (4dfr 1dr1) (1ghb 8gch) (8gch 1ghb) (1acj 1ack) (1ack 1acj) (1nis 1aco) (1aco 1nis) (6abp 1abe) (1abe 6abp) (1com 2cht) (2cht 1com) (1fki 1fkg) (1fkg 1fki) (1dbb 1dbj) (1dbb 2dbl) (1dbj 1dbb) (1dbj 2dbl) (2dbl 1dbb) (2dbl 1dbj) (1glp 1glq) (1glq 1glp) (1aaq 1hef) (1aaq 4phv) (1hef 1aaq) (1hef 4phv) (4phv 1aaq) (4phv 1hef) (1lah 1lst) (1lst 1lah) (1mrg 1aha) (1aha 1mrg) (1ive 2sim) (2sim 1ive) (2phh 1pbd) (1pbd 2phh) (1epb 1cbs) (1epb 1fen) (1cbs 1epb) (1fen 1epb) (6rsa 1rob) (1rob 6rsa) (1srj 1stp) (1stp 1srj) (1tka 1trk) (1trk 1tka) (1tph 7tim) (7tim 1tph) (1did 1die) (1did 1xid) (1did 1xie) (1die 1did) (1die 1xid) (1die 1xie) (1xid 1did) (1xid 1die) (1xid 1xie) (1xie 1did) (1xie 1die) (1xie 1xid)

All unrelated pairs used: (1aaq 1eed) (1aaq 1srj) (1aaq 1ive) (1abe 2lgs) (1abe 1dr1) (1abe 1cdg) (1abe 1hdc) (1abe 1rds) (1acj 1srj) (1acj 1glp) (1acj 1hri) (1ack 3ptb) (1ack 2cgr) (1ack 1icn) (1ack 1icn) (1ack 5p2p) (1ack 1etr) (1ack 1byb) (1ack 1byb) (1ack 1hdc) (1acm 2cmd) (1acm 2lgs) (1acm 2phh) (1acm 1xie) (1acm 3cla) (1acm 3cla) (1acm 1did) (1acm 2gbp) (1aco 1com) (1aco 2gbp) (1aco 1tdb) (1aco 1ukz) (1aco 5p2p) (1aco 1poc) (1aco 1eed) (1aha 1ukz) (1aha 3hvt) (1aha 5p2p) (1aha 1glq) (1apt 1cbs) (1baf

1hsl) (1baf 1imb) (1baf 2dbl) (1baf 1com) (1bbp 6rsa) (1bma 1rds) (1bma 1hsl) (1bma 1com) (1byb 1slt) (1byb 1nco) (1byb 6rsa) (1byb 2r07) (1byb 1die) (1byb 1tka) (1byb 1wap) (1byb 1ack) (1byb 1ack) (1byb 1lst) (1byb 1mup) (1byb 3ptb) (1cbs 1icn) (1cbs 1eap) (1cbs 3tpi) (1cbs 1apt) (1cdg 1coy) (1cdg 1abe) (1cdg 1pbd) (1com 1aco) (1com 1ghb) (1com 2dbl) (1com 1baf) (1com 1fkg) (1com 1bma) (1coy 1mrk) (1coy 1cdg) (1coy 1xid) (1coy 1fkg) (1coy 4phv) (1coy 4cts) (1ctr 1icn) (1ctr 4fab) (1ctr 1dwd) (1ctr 2cgr) (1ctr 1did) (1ctr 2lgs) (1dbb 1mrk) (1dbb 8gch) (1dbb 4dfr) (1dbb 1mrg) (1dbb 1lst) (1dbj 1tka) (1dbj 2phh) (1dbj 2r07) (1dbj 1eed) (1did 2sim) (1did 1mrk) (1did 1acm) (1did 3tpi) (1did 1ldm) (1did 1ctr) (1die 5p2p) (1die 1byb) (1die 4phv) (1dr1 1ulb) (1dr1 1abe) (1dr1 8gch) (1dr1 1mrg) (1dr1 1glq) (1dr1 1wap) (1dr1 1tka) (1dr1 1lic) (1dr1 2lgs) (1dwd 1hef) (1dwd 1ctr) (1dwd 1ukz) (1dwd 1wap) (1dwd 2lgs) (1eap 1hdc) (1eap 1cbs) (1eap 8gch) (1eap 2sim) (1eap 1xie) (1eed 1aaq) (1eed 1tdb) (1eed 1dbj) (1eed 1rob) (1eed 3hvt) (1eed 2phh) (1eed 2cmd) (1eed 1aco) (1epb 7tim) (1epb 1lmo) (1epb 1poc) (1eta 1icn) (1eta 1ulb) (1eta 1fen) (1eta 1tph) (1etr 6rsa) (1etr 1tdb) (1etr 1ack) (1etr 1tph) (1fen 1eta) (1fkg 1coy) (1fkg 8gch) (1fkg 1com) (1fkg 1ulb) (1fkg 3ptb) (1fki 1tdb) (1fki 2mcp) (1fki 1hsl) (1ghb 1com) (1ghb 1mup) (1glp 1tdb) (1glp 2sim) (1glp 1hef) (1glp 1pbd) (1glp 1acj) (1glq 1lmo) (1glq 2cgr) (1glq 1slt) (1glq 1dr1) (1glq 1srj) (1glq 1aha) (1hdc 1eap) (1hdc 1lmo) (1hdc 2cgr) (1hdc 1ack) (1hdc 1abe) (1hdc 1mrg) (1hef 1dwd) (1hef 2r07) (1hef 1glp) (1hef 1wap) (1hri 1icn) (1hri 8gch) (1hri 1nco) (1hri 4dfr) (1hri 1ida) (1hri 1acj) (1hsl 1lah) (1hsl 2lgs) (1hsl 1baf) (1hsl 2dbl) (1hsl 1bma) (1hsl 1snc) (1hsl 1fki) (1icn 1cbs) (1icn 1ctr) (1icn 1hri) (1icn 2r07) (1icn 1eta) (1icn 1mcr) (1icn 1lst) (1icn 6rsa) (1icn 1ack) (1icn 1ack) (1icn 1tph) (1icn 1xid) (1ida 1slt) (1ida 1mcr) (1ida 1hri) (1ida 2pk4) (1igj 1tdb) (1igj 6rsa) (1igj 1wap) (1igj 3tpi) (1imb 1tdb) (1imb 1slt) (1imb 1trk) (1imb 1baf) (1imb 1tng) (1imb 1nco) (1imb 1rne) (1ive 7tim) (1ive 1mcr) (1ive 6rsa) (1ive 2ak3) (1ive 1tni) (1ive 1aaq) (1lah 1hsl) (1lah 1tng) (1ldm 2cht) (1ldm 1did) (1ldm 1rob) (1ldm 1tdb) (1ldm 6rsa) (1lic 3tpi) (1lic 1dr1) (1lmo 1glq) (1lmo 1hdc) (1lmo 1lst) (1lmo 1epb) (1lst 1icn) (1lst 1lmo) (1lst 1byb) (1lst 1tph) (1lst 1dbb) (1mcr 1ive) (1mcr 1ukz) (1mcr 1ukz) (1mcr 1ida) (1mcr 1icn) (1mcr 1tph) (1mdr 1pbd) (1mrg 6abp) (1mrg 1pbd) (1mrg 1dr1) (1mrg 1mrk) (1mrg 3cla) (1mrg 1srj) (1mrg 1snc) (1mrg 1dbb) (1mrg 1poc) (1mrg 1hdc) (1mrk 2gbp) (1mrk 1did) (1mrk 3cla) (1mrk 1coy) (1mrk 1mrg) (1mrk 1dbb) (1mrk 2pk4) (1mup 2cht) (1mup 1ghb) (1mup 3cla) (1mup 1byb) (1nco 1byb) (1nco 3cla) (1nco 2cgr) (1nco 1hri) (1nco 1stp) (1nco 1imb) (1nis 3aah) (1nis 3ptb) (1nis 1rds) (1pbd 1mdr) (1pbd 1mrg) (1pbd 3aah) (1pbd 3aah) (1pbd 1cdg) (1pbd 1glp) (1pbd 2plv) (1poc 1epb) (1poc 1mrg) (1poc 2gbp) (1poc 1ulb) (1poc 1aco) (1rds 5p2p) (1rds 1bma) (1rds 2gbp) (1rds 1abe) (1rds 1nis) (1rne 2cgr) (1rne 1stp) (1rne 1imb) (1rob 3hvt) (1rob 2lgs) (1rob 1tph) (1rob 2cgr) (1rob 1eed) (1rob 1ldm) (1slt 1byb) (1slt 2ak3) (1slt 1glq) (1slt 1imb) (1slt 1ida) (1slt 1tni) (1slt 1tng) (1snc 2dbl) (1snc 1mrg) (1snc 6abp) (1snc 1xie) (1snc 1hsl) (1srj 1aaq) (1srj 1ulb) (1srj 1ulb) (1srj 4phv) (1srj 1glq) (1srj 1mrg) (1srj 1acj) (1stp 1tdb) (1stp 1nco) (1stp 1rne) (1tdb 1imb) (1tdb 1stp) (1tdb 1glp) (1tdb 1igj) (1tdb 3hvt) (1tdb 3hvt) (1tdb 5p2p) (1tdb 1fki) (1tdb 3tpi) (1tdb 1ldm) (1tdb 1etr) (1tdb 1eed) (1tdb 1aco) (1tka 6rsa) (1tka 1dbj) (1tka 3cla) (1tka 1dr1) (1tka 1byb) (1tng 1lah) (1tng 3hvt) (1tng 1imb) (1tng 6rsa) (1tng 1slt) (1tni 1ive) (1tni 1ukz) (1tni 6rnt) (1tni 1slt) (1tnl 1wap) (1tnl 8gch) (1tph 1xie) (1tph 2mcp) (1tph 1rob) (1tph 3aah) (1tph 1icn) (1tph 1ukz) (1tph 1lst) (1tph 1mcr) (1tph 1etr) (1tph 1eta) (1tph 3ptb) (1tph 3ptb) (1trk 2sim) (1trk 3aah) (1trk 1imb) (1ukz 1mcr) (1ukz 1mcr) (1ukz 1aha) (1ukz 1dwd) (1ukz 1aco) (1ukz 1tph) (1ukz 1tni) (1ulb 1dr1) (1ulb 1srj) (1ulb 1srj) (1ulb 3aah) (1ulb 5p2p) (1ulb 1eta) (1ulb 2r07) (1ulb 1fkg) (1ulb 1poc) (1wap 1dwd) (1wap 1dr1) (1wap 2sim) (1wap 1hef) (1wap 1tnl) (1wap 1igj) (1wap 1byb) (1xid 4cts) (1xid 6rsa) (1xid 1coy) (1xid 1icn) (1xie 2cmd) (1xie 2cht) (1xie 2lgs) (1xie 1tph) (1xie 2pk4) (1xie 1acm) (1xie 2plv) (1xie 1eap) (1xie 1snc) (2ak3 1slt) (2ak3 1ive) (2cgr 1glq) (2cgr 2pk4) (2cgr 2yhx) (2cgr 3aah) (2cgr 1rne) (2cgr 4phv) (2cgr 1rob) (2cgr 1ack) (2cgr 1nco) (2cgr 1hdc) (2cgr 1ctr) (2cht 6abp) (2cht 1xie) (2cht 1mup) (2cht 1ldm) (2cht 5p2p) (2cmd 1acm) (2cmd 1xie) (2cmd 2pk4) (2cmd 2plv) (2cmd 1eed) (2dbl 1snc) (2dbl 4dfr) (2dbl 1com) (2dbl 1hsl) (2dbl 1baf) (2gbp 1mrk) (2gbp 1aco) (2gbp 1acm) (2gbp 1rds) (2gbp 1poc) (2lgs 1hsl) (2lgs 1abe) (2lgs 4cts) (2lgs 1acm) (2lgs 1xie) (2lgs 1rob) (2lgs 1dr1) (2lgs 1dwd) (2lgs 1ctr) (2mcp 1tph) (2mcp 1fki) (2mcp 2r07) (2phh 1acm) (2phh 1dbj) (2phh 2plv) (2phh 2plv) (2phh 1eed) (2pk4 2cgr) (2pk4 1xie) (2pk4 2cmd) (2pk4 1mrk) (2pk4 1ida) (2plv 6abp) (2plv 2r07) (2plv 1xie) (2plv 1pbd) (2plv 3aah) (2plv 2phh) (2plv 2phh) (2plv 2cmd) (2r07 1dbj) (2r07 1icn) (2r07 1byb) (2r07 1hef) (2r07 2plv) (2r07 2mcp) (2r07 1ulb) (2sim 1did) (2sim 1glp) (2sim 6abp) (2sim 1eap) (2sim 1trk) (2sim 1wap) (2yhx 2cgr) (3aah 1pbd) (3aah 1pbd) (3aah 1nis) (3aah 1trk) (3aah 2cgr) (3aah

1tph) (3aah 1ulb) (3aah 4dfr) (3aah 2plv) (3aah 3ptb) (3cla 1mrk) (3cla 1acm) (3cla 1acm) (3cla 1tka) (3cla 1mrg) (3cla 1nco) (3cla 1mup) (3hvt 1aha) (3hvt 1tdb) (3hvt 1tdb) (3hvt 1rob) (3hvt 1tng) (3hvt 1eed) (3ptb 1ack) (3ptb 1nis) (3ptb 3aah) (3ptb 1tph) (3ptb 1tph) (3ptb 1byb) (3ptb 1fkg) (3tpi 1did) (3tpi 1tdb) (3tpi 1cbs) (3tpi 1igj) (3tpi 1lic) (4cts 2lgs) (4cts 1xid) (4cts 4fab) (4cts 5p2p) (4cts 1coy) (4dfr 2dbl) (4dfr 5p2p) (4dfr 3aah) (4dfr 1hri) (4dfr 1dbb) (4fab 1ctr) (4fab 4cts) (4phv 1coy) (4phv 2cgr) (4phv 1srj) (4phv 1die) (5p2p 1tdb) (5p2p 1aha) (5p2p 2cht) (5p2p 1rds) (5p2p 4dfr) (5p2p 1die) (5p2p 1ulb) (5p2p 1ack) (5p2p 4cts) (5p2p 1aco) (6abp 1mrg) (6abp 2cht) (6abp 2sim) (6abp 2plv) (6abp 1snc) (6rnt 1tni) (6rsa 1ive) (6rsa 1xid) (6rsa 1igj) (6rsa 1tka) (6rsa 1byb) (6rsa 1bbp) (6rsa 1etr) (6rsa 1icn) (6rsa 1ldm) (6rsa 1tng) (7tim 1ive) (7tim 1epb) (8gch 1dr1) (8gch 1eap) (8gch 1fkg) (8gch 1hri) (8gch 1dbb) (8gch 1tnl)