Fast Small Molecule Similarity Searching with Multiple Alignment Profiles of Molecules Represented in One-Dimension

Norman Wang, Robert K. DeLisle, and David J. Diller*

Pharmacopeia, Inc., CN5350, Princeton, New Jersey 08543-5350

Received June 15, 2005

Multiple sequence alignment has proven to be a powerful method for creating protein and DNA sequence alignment profiles. These profiles of protein families are useful tools for identifying conserved motifs, such as the catalytic triad of the serine protease family or the seven transmembrane helices of the G-protein coupled receptor family. Ultimately, the understanding of the critical motifs within a family is useful for identifying new members of the family. Due to the complexity of protein−ligand recognition, no universally accepted method exists for clustering small molecules into families with the same or similar biological activity. A combination of the concept of multiple sequence alignment and the 1-dimensional molecular representation described earlier offers a new method for profiling sets of small molecules with the same biological activity. These small molecule profiles can isolate key commonalties within the set of bioactive compounds much like a multiple sequence alignment can isolate critical motifs within a protein family. The small molecule profiles then make useful tools for searching small molecule databases for new compounds with the same biological activity. The technique is demonstrated here using the human ether-a-go-go potassium channel and the kinase SRC.

Introduction

In bioinformatics, DNA and protein sequence alignment is the foundation for predicting structural similarity and thus the likelihood of similarity in function. In cheminformatics, a host of chemical similarity measures are used to assess the similarity between pairs of small molecules and thus the likelihood of them having similar function, i.e., biological activity. The list of small molecule similarity identification techniques includes substructure fingerprints,[1,2] pharmacophore fingerprints,[3−7] and overall 3-D alignments.[8,9] For the purpose of finding molecules with similar function, DNA and protein sequence alignment have several advantages over methods for determining chemical similarity. A protein sequence typically has 100−1000 amino acids. A small molecule on the other hand typically has 20−40 non-hydrogen atoms. Thus, the problem of determining chemical similarity does not have the same statistical power of large numbers. In addition, the biological activity of a small molecule often hinges on a single atom. Indeed, there are numerous cases in which the change, deletion, or addition of a single atom completely abolishes the biological activity of a small molecule. To further complicate matters, a single small molecule will often have multiple seemingly unrelated biological activities. In this sense the small molecule similarity problem is context specific. Figure 1a shows a molecule that both inhibits the p38 kinase and acts as an antagonist of the glucagon receptor, a class B GPCR. Figure 1b shows a related molecule that is an inhibitor of the p38 kinase but not an antagonist of the glucagon receptor. Figure 1c shows a second related molecule that retains the glucagon antagonist activity but is devoid of the p38 activity. From the viewpoint of p38 molecule 1a is similar to 1b but not 1c whereas from the perspective of the glucagon receptor 1a is similar to 1c but not 1b.

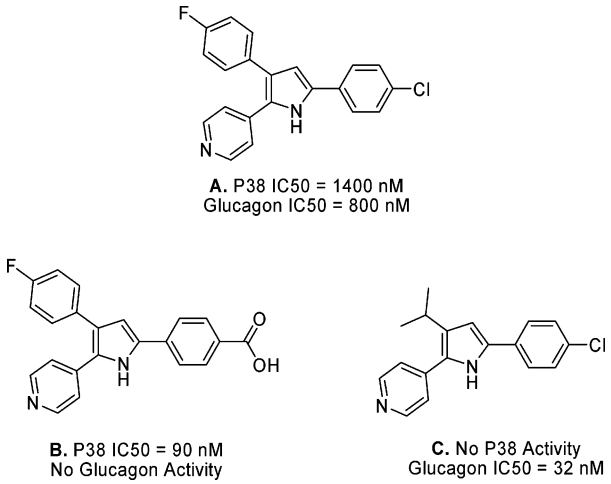

Figure 1. An example of a small molecule showing multiple biological activities and subtle changes that distinguish between these activities.[50] This example shows that assigning function to a small molecule, i.e., its biological activity, based on similarity is a context specific problem. (A) Compound **A** both inhibits the p38 kinase and acts as an antagonist of the glucagon receptor which is a GPCR from the class B family. Compound **B** is an example of a small molecule that is closely related to compound **A** and retains the p38 inhibitory activity but is devoid of the glucagon activity. Compound **C** is an example of a small molecule that is closely related to **A** and retains the glucagon activity but is devoid of the p38 activity. Thus, from the perspective of p38, molecule **A** is similar to **B** but not **C** whereas from the perspective of the glucagon receptor, **A** is similar to **C** but not **B**.

Despite the aforementioned differences, the problems of assigning function to a protein and assigning function to a small molecule share a common challenge. Often

* To whom correspondence should be addressed: ddiller@pharmacop.com, (609) 452-3783 (phone), (609) 655-4187 (fax).

10.1021/jm050563r CCC: $30.25
Published on Web 09/30/2005

two proteins with little or no overall sequence identity share a common function. An example is the serine protease family in which only the catalytic triad (Ser, Asp, and His) is required for function while the remainder of the amino acids largely serve to precisely position the catalytic triad.[10−13] Over a few hundred amino acids, three residues have no effect on overall sequence identity. Similarly, two small molecules with little or no obvious similarity often possess the same biological activity. Much like the example of the serine protease family, the binding of a small molecule to a specific protein often relies on a few key interaction points, such as hydrogen bond acceptors, donors, aromatic rings, etc., with the remainder of the molecule serving to precisely position the key interaction points. The set of interaction points and their precise arrangement is referred to as the pharmacophore.[14]

Bioinformatics addresses the issue of identifying relationships between proteins in large part through multiple sequence alignment. The advantage of multiple sequence alignment over pairwise alignment is that multiple sequence alignment can find residues that are conserved over an entire protein family thereby distinguishing the critical amino acids or motifs such as the catalytic triad of the serine protease family. The most comparable technique in computational chemistry is pharmacophore modeling. In this case, multiple small molecules are aligned in 3 dimensions. The features that the majority of the small molecules share are analogous to the conserved amino acids of a multiple sequence alignment. Pharmacophore models are often used to search through a small molecule database for molecules with the same biological activity much like a multiple sequence alignment can be used to search for related family members.

A key difference between multiple sequence alignment and pharmacophore modeling is that multiple sequence alignment is inherently a 1-dimensional search problem and is therefore amenable to a host of specialized algorithms whereas pharmacophore modeling is inherently a high dimensional problem involving for each molecule 3 rotational, 3 translational, and the sometimes many conformational degrees of freedom. In computational chemistry, 3-dimensional methods in general and pharmacophore modeling in particular suffer from the problem of conformational explosion as the number of rotatable bonds increases. This conformational explosion makes the initial elucidation of the pharmacophore extremely challenging and affects both the quality and speed of the search of conformational databases.

Recently, Dixon and Merz[15] developed a method for assessing molecular similarity that is directly analogous to pairwise sequence alignment. In this method the atoms of the small molecule, either from 3-dimensional coordinates or 2-dimensional topological distances, are projected onto 1 dimension using multidimensional scaling.[16] This results in the molecule being represented as a 1-dimensional string of atoms where the atom pairwise distances retain as much information as possible concerning their true 2- or 3-dimensional distances. Once two molecules have been projected to 1 dimension they can be rapidly aligned using the techniques of pairwise alignment. The chief differences between pairwise protein sequence alignment and pairwise 1-dimensional molecular alignment are; first, for 1-dimensional molecular alignment, both relative orientations must be considered, second, insertions and deletions do not make sense in the context of a small molecule, and third for small molecule alignment, the 1-dimensional representations can be aligned continuously relative to one another rather than in discrete steps. In a variety of tests Dixon and Merz have shown this molecular similarity method to be superior to a variety of other similarity methods.

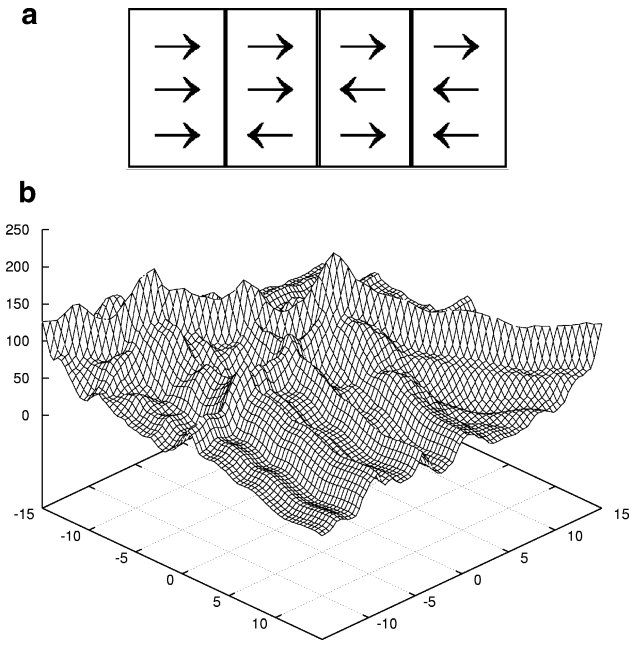

Figure 2. The search space in a multiple 1-dimensional alignment of three molecules. (a) The 4 possible relative orientations of the 3 1-dimensional objects. Note that the four remaining orientations in which the top arrow points to the left are by flipping all three arrows mathematically equivalent to one of the four relative orientations shown in part a. (b) The 2D search space for a single relative orientation of the three molecules Cisapride, Thioridazine, and Ziprasidone.

In this work, we extend the method of Dixon and Merz to multiple ligand alignment. The goal is to create 1-dimensional profiles from many molecules with the same biological activity that identify the features common to all or most of the molecules. Such a profile could potentially isolate the key features of the molecules, much like a pharmacophore model isolates key interaction sites and a multiple sequence alignment isolates conserved amino acids. Furthermore these profiles would avoid the difficulties associated with the high dimensionality encountered in the conformational search problem.

To align 1-dimensional representation of K molecules, there are a total of 2^{k-1} possible relative orientations. As an example, Figure 2a shows the 4 relative orientations of 3 1-dimensional objects. With 10 molecules there are 512 possible relative orientations, and with 20 molecules there are over 50 000 possible relative orientations. To find the global maximal alignment, one must solve a continuous global optimization problem in $K - 1$ dimensions for each of the possible 2^{K-1} relative orientations. As an example the 2-dimensional similarity landscape for 3 molecules, Cisapride,[17] Thio-

ridazine,[18] and Ziprasidone[18] are shown in Figure 2b. The number of local maxima is certain to increase exponentially with the number of molecules. Because of the complexity of the high dimensional similarity landscape and the large number of discrete relative orientations, a brute force solution to the overall global optimization problem is not practical. In addition, the large number of local maxima and the presence of discrete variables preclude the exclusive use of gradient-based methods.

To produce a consensus 1-dimensional alignment of multiple molecules, a heuristic approach is used: evolutionary programming[19] to address the continuous variables and genetic algorithms[20] to handle the discrete variables. The combination of genetic algorithms and evolutionary programming has been shown through numerous examples to robustly handle optimization problems with continuous and discrete variables[21,22] and thus is particularly well suited to the problem of multiple 1-dimensional ligand alignment. Due to the complexity of the optimization problem involved, we adopt the heuristic combination of evolutionary programming and genetic algorithms for building a near optimal multiple ligand alignment profile. We then show these profiles to be useful for rapidly searching large compound databases for novel molecules with the same biological activity. Finally, we compare the results with the 1-dimensional profiles to results using comparable 3-dimensional methods. We demonstrate the utility of this approach with the human ether-a-go-go potassium channel and the kinase SRC.

Methods

To obtain multiple 1-dimensional ligand alignments, a heuristic approach utilizing evolutionary programming and genetic algorithms was used to 'evolve' a solution, resulting in a consensus multiple alignment profile of the selected set of compounds. Each small molecule to be aligned is first converted, using only 2-dimensional topological distances, to its 1-dimensional representation through the multidimensional scaling and BFGS optimization procedure as described for the pairwise case.[15] Each small molecule's 1-dimensional encoding consists of information about the type of each atom and its 1-dimensional coordinate.

For the purposes of this description, it is assumed that each of the 1-dimensional representations has been created such that its center of mass is 0, i.e., the sum over the 1-dimensional coordinates of its atoms is 0. To put this molecule's 1-dimensional representation into a new frame of reference, its orientation and translation must be specified. An orientation takes a value of 1 or -1 where a value of 1 means to make no change in orientation and a value of -1 means to flip the 1-dimensional representation of the molecule. The translation is a continuous variable, takes any real value, and essentially shifts the representation along the 1-dimensional axis. Mathematically, the 1-dimensional coordinates are transformed by $x_i^{\text{new}} = \rho x_i^{\text{initial}} + \Delta x$ where ρ is the orientation and Δx is the translation.

Organism/Gene Encoding. The first step in a genetic algorithm or evolutionary program is the generation of an initial population of potential solutions referred to as organisms. Each organism consists of a set of genes, with each gene representing the translation and orientation of a single molecule. The initial set of generated genes contains translation values distributed via a Gaussian distribution with mean 0 and standard deviation equal to one-quarter of the molecule's 1-dimensional width. The initial orientation of each gene was chosen purely at random.

Alignment Scoring/Organism Fitness. To assess the fitness of an organism, the quality of any given multiple alignment must be scored. The score is given by

$$\text{alignment score} = \sum_{i<j}^{N_i}\sum_{k=1}^{N_j}\sum_{m=1} S(a_{ik},a_{jm}) f(x_{ik} - x_{jm}) \quad (1)$$

where the outer sum is over all pairs of molecules in the alignment, the second sum is over the atoms of the ith molecule, the inner sum is over the atoms of the jth molecule, $S(a_{ik},a_{jm})$ is the similarity of the kth atom of the ith molecule to the mth atom of the jth molecule, x_{ik} and x_{jm} are the 1-dimensional coordinates of the two atoms, and f is a distance dependent measure of the overlap of two atoms. The full description of the scoring scheme requires two components: the atom pairwise similarity matrix, S, and the distance dependent overlap function, f. These are described below.

The overlap function, f, in eq 1 could reasonably be any number of functions. For this work the original overlap function proposed by Dixon and Merz[15] was used:

$$f(\Delta x) = \begin{bmatrix} 0 \text{ if } |\Delta x| \geq 1 \\ 1 - |\Delta x| \text{ if } |\Delta x| < 1 \end{bmatrix} \quad (2)$$

This overlap score is best thought of as each atom having a width of 1 bond unit and the overlap being the area under the intersection of the two atoms.

Atom Pairwise Scoring Matrixes. There are no well-established atom similarity matrixes such as the Dayhoff,[23] PAM,[23] MDM,[23] BLOSUM,[24] GCB,[25] and PET[26] similarity matrixes for amino acids. There are a large number of factors that could be taken into account in determining atom similarity. These factors include, actual atom type, atom hybridization, partial charge, polarizability, aromaticity, hydrophobicity, etc. As many of these factors depend on both the particular atom and its neighbors in the molecule, there could be in essence an infinite number of atom types. A reasonable atom similarity matrix must have a sufficiently detailed atom description without having an unmanageable number of atom types.

Ultimately, the atom types used were the same as those used by Dixon and Merz.[15] In this scheme an atom's type is determined by its atomic number, its hybridization, number of bonded hydrogens, and whether it is a member of a ring. The most obvious atom pairwise similarity matrix is the identity matrix, i.e., atoms have a similarity of 1 if they have identical type; otherwise they have a similarity of 0. This matrix proved to be too strict. The analogy for protein sequence alignment would be to classify amino acids such as leucine and isoleucine as completely different. Experience has shown that similarity matrixes such as BLOSUM, PAM, and the hydrophobicity matrixes are significantly more sensitive than the identity matrix. With this in mind similarity matrix 1 (denoted by M below) was developed with the following rules:

Except for the halogens, atoms of different atomic number have no similarity. The similarity between atoms of the same atomic number decreases from 1 to a minimum of 0 via the rules: decrease by 0.4 for each change in hybridization; decrease by 0.2 if one of the atoms is in a ring while the other is not; decrease by 0.2 for the difference in number of hydrogens. The similarity for halogens is $1-0.2 \times n$ where n is the difference in the rows of the periodic table of the two halogens.

Ultimately, similarity matrix 1 was not sufficiently discriminating because the alignments were always dominated by the carbon atoms of the molecules. Often, the carbon atoms act more as the framework of the molecule and less as prominent interactions. To overcome this issue, the MDDR[27] database was profiled for the frequency of occurrence of each possible atom type. From the vector of occurrences, P, the weighted occurrence vector, W, is defined by $W = MP$ where M is similarity matrix 1. The element W_j, of the weighted occurrence vector, W, is the expected similarity between an

atom of type j and a randomly chosen atom from the MDDR database. Similarity matrix 2 is then derived by scaling similarity matrix 1 by the weighted occurrences via the formula

$$S_{ij} = \frac{M_{ij}}{\sqrt{W_i W_j}} \tag{3}$$

Similarity matrix 2 was used for all atom pairwise similarity calculations.

Selection Process. There are many ways in which a new generation of potential solutions can be created from the previous generation. For this work the selection process was done through roulette wheel selection[21,22] with (mu,lambda) population replacement from one generation to the next. Given a population of organisms, the probability Q_i for organism i to have offspring in the next generation is dependent on its fitness score, E, and the temperature constant, T, via the relationship, $Q_i \propto e^{(Si/T)}$, where the constant of proportionality is chosen so that the sum of the Q_i over the population is 1 and T is a constant used to control the global versus local nature of the search.

Recombination and Mutation Processes. A genetic algorithm was used to perform crossover with two parents to yield a single offspring. Each gene of the offspring was inherited at random with uniform crossover with 50% probability. At this point the genome of the offspring is a combination of the genomes of its parents.

After the entire offspring generation is created, two types of random mutations were used to further modify each organism: flip mutation and shift mutation. The flip mutation takes the form of bit flipping which is typical of a genetic algorithm but usually not seen in evolutionary programming. The flip mutation negates the orientation of the 1-dimensional atom coordinates resulting in a 1-dimensional representation that is flipped. The shift mutation uses a Gaussian distribution to change the translation of a 1-dimensional representation essentially shifting the molecule along the 1-dimensional axis. This is typical of evolutionary programming but usually not seen in genetic algorithms. Both the shift and flip mutations were applied to each molecule with a probability of 0.05%. A Gaussian distribution with a standard deviation of 0.5 bond units was used to randomly perturb those molecules selected for a shift mutation. Also, for this work a population size of 128 was used with 500 steps of the evolutionary process.

The Compound Database Search. After the multiple alignment of the small molecules with the same biological activity of interest, we wish to search large databases of available or virtual compounds with the intent of finding novel small molecules with the same biological activity. To assess the likelihood of a small molecule having the desired biological activity, its 1-dimensional representation must be aligned to the multiple ligand profile. To find the optimal alignment, two 1-dimensional optimizations (both relative orientations) must be performed. With two straightforward techniques the database search can be extremely fast: ~900 molecules per second on a single SGI R10000 processor.

The first technique is to create a lookup table for each atom type at the beginning of the database search. To do so, a fixed step size (usually 0.1 bond units) is chosen. Beginning at the lower limit of the profile the score for an atom of type 1 is calculated and stored in the first position of the array. Next the atom is moved one step size to the right, the score for the atom is calculated and stored in the second position of the array. The process is continued until the atom reaches the upper limit of the profile. The process is then repeated for each atom type. When the score for a molecule from the database needs to be calculated, the score for each atom can be found using its 1-dimensional coordinate from the appropriate array. The memory needed to store the arrays is less than 1 Mb, and the time necessary to calculate the lookup table is less than 10 s on an SGI R10000.

The second technique is a standard bracketing technique used for solving 1-dimensional optimization problems.[28] Again, a fixed step size (usually 0.5 bond units) and upper and lower limits (determined by the bounds of the profile and the bounds of the molecule) are chosen. Beginning at the lower limit the score for the molecule is calculated. The molecule is moved one step to the right and the score calculated again. The process is continued until the molecule moves past the upper limit. Three consecutive offsets of the molecule (x1 < x2 < x3) are said to bracket a maximum if S(x1) < S(x2) and S(x3) < S(x2): because the scoring function S is continuous, one can mathematically guarantee that under these conditions S will have a local maximum somewhere between x1 and x3.[28] Standard line optimization techniques can then be used to find any bracketed local maximum to any desired level of accuracy.

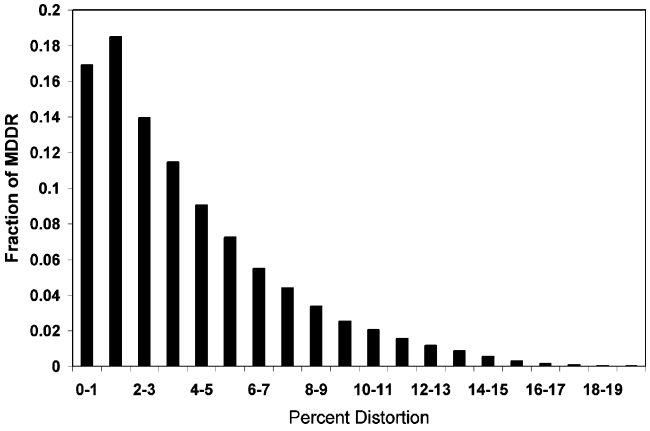

Figure 3. The distribution of the extent of distortion of each molecule caused by projecting from two-dimensional topological distances to its 1-dimensional representation. 90% of the molecules exhibit a distortion below 9%, 95% exhibit a distortion below 11%, 99% exhibit a distortion below 14%, and essentially all the molecules of the MDDR database exhibit a distortion below 20%. Thus, the 1-dimensional representation retains the majority of the information present in the 2-dimesional structure.

Results

All molecules used in these studies were projected into their 1-dimensional representation using only their interatomic topological distances via the procedure outlined by Dixon and Merz.[15] Thus, there is no need to produce intermediate 3-dimensional coordinates at any stage in the process. To quantify the extent of loss of information we calculated the average distortion with the 1-dimensional representation of each molecule in the MDDR database with respect to its 2-dimensional structure. The average distortion of a molecule is calculated via

$$\text{distortion} = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{i-1} \frac{|d_{ij} - d_{ij}^{1D}|}{d_{ij}}$$

where N is the number of atoms in the molecule, d_{ij} is the topological distance between atoms i and j, and d_{ij}^{1D} is the distance between atoms i and j in the 1-dimensional representation. In essence the formula captures the average distortion between all pairs of atoms in the molecule. Figure 3 shows the distribution of the distortion of the molecules in MDDR in their 1-dimensional representation. 90% of the molecules exhibit a distortion below 9%, 95% exhibit a distortion below 11%, 99% exhibit a distortion below 14%, and essentially all the molecules of the MDDR database exhibit a distortion below 20%. Thus, the 1-dimensional representation

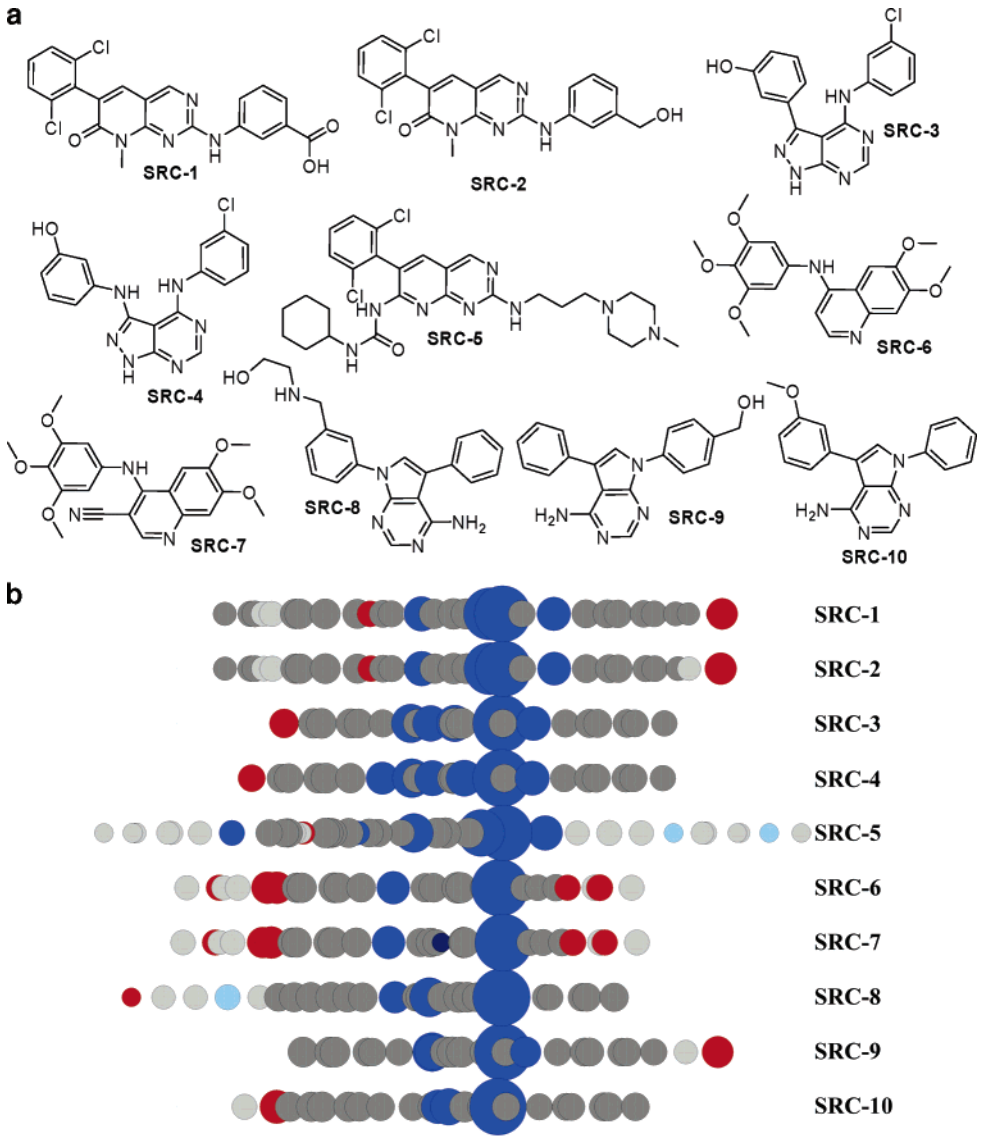

Figure 4. The SRC 1-dimensional multiple ligand alignment. (a) The compounds used in the alignment. (b) The resulting alignment. Each compound is shown in part a roughly oriented in the same manner as in the alignment in part b. The color indicates the type of atom, i.e., gray is carbon, red is oxygen, blue is nitrogen, etc. The darkness of the color indicates the hybridization: light colors indicate Sp3 atoms and dark colors indicate Sp2 or Sp1 atoms. The size of the atom is proportional to the amount the atom contributes to the score of the overall alignment.

retains the majority of the information present in the 2-dimesional structure.

To demonstrate the utility of the multiple ligand profiles for identifying novel small molecules with similar biological activity, we perform seeding experiments: a relatively large set of small molecules with the same biological activity are collected and split into a training and test set. For the purpose of this work, the training sets consist of 10 molecules and the test sets consist of at least 50 small molecules. The small molecules of the training set are used to build the 1-dimensional profiles. The profiles are then used to rank the molecules of the test set along with those from a database of random compounds. The quality of the search method is judged by the extent to which it differentiates the molecules of the test set from those of the random database.

The random molecules (~100 000) were drawn from the MDDR database[27] with a molecular weight cutoff of 600 Da. The MDDR database consists of drug-like molecules and is representative of the types of small molecules synthesized in medicinal chemistry programs. Thus, selecting random small molecules from the MDDR database makes these realistic examples.

The examples described below are: the human ether-a-go-go potassium channel (hERG) and the kinase SRC. The kinase SRC has been implicated in a variety of cancers.[29] In addition, SRC is used as a selectivity assay in many kinase programs. Thus, a SRC virtual screen would be useful as a means to generate new leads and as a virtual filter to identify potential selectivity issues in kinase programs. As a test case SRC is representative of a medicinal chemistry program in that there are a small handful of chemotypes all of which are represented in the training set. Thus, the test set contains molecules whose chemotypes are all represented in the training set.

The second example is the hERG potassium channel. Inhibition of the hERG potassium channel has recently been linked to cardiotoxicity such as prolonged QT

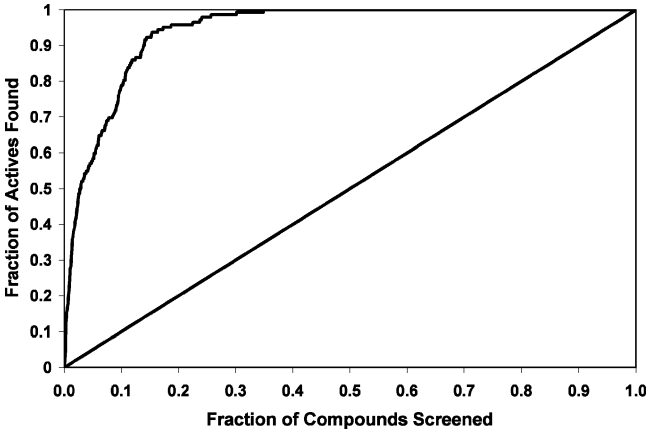

Figure 5. The performance of the SRC profile. To make these figures, the molecules are ranked and sorted by their score, in this case their similarity to the 1-dimensional profile. The top curve is the fraction of the known SRC actives found (*y*-axis) versus the fraction of the overall compounds (*x*-axis). The diagonal line is the expected performance of a method that selected compounds at random.

interval and torsades de pointes.[30] Drugs such as cisapride[17] and terfenadine[31] have been withdrawn from the market because of their propensity to cause fatal arrhythmias caused by blocking the hERG channel. Thus, the hERG channel is a counter screen in many medicinal chemistry programs. The hERG test set consists largely of chemotypes not represented in the training set. Thus, as an example the hERG data set is more representative of the challenges faced in discovering new chemotypes either during discovery efforts or lead hopping.

The SRC Validation Study. The final SRC profile and the 10 compounds used to build the profile are shown in Figure 4a. In this case the test set consists of 142 known SRC inhibitors.[32−42] The strongest feature in the SRC profile is a central aromatic nitrogen (see Figure 4b). One could hypothesize that these aligned nitrogens make the hydrogen bond with the backbone NH of the hinge region[43] often observed as critical in kinase−ligand interactions. In 9 of the 10 molecules used to make the SRC profile the structure−activity relationships strongly support the aromatic ring aligned to the left of the profile as being the ring that binds in the main hydrophobic pocket. The group that likely binds in the main hydrophobic pocket for molecule SRC-8 (see Figure 4) is the aromatic ring on the same side as the NH2. Thus, this molecule should likely be oriented the same as SRC-9 and SRC-10. The psuedo-2-fold symmetry makes the problem of orienting these three molecules particularly difficult.

The overall results of the database search with the SRC 1-dimensional profile are shown in Figure 5. Approximately 26% of the known SRC inhibitors are ranked within the top 1% of the MDDR compounds (enrichment = 26), 58% of the known SRC inhibitors are ranked within the top 5% of the MDDR compounds (enrichment of 11.6), and nearly 95% of the known SRC inhibitors are found within the top 17% of the MDDR compounds (enrichment of 5.6).

The hERG Validation Study. The final 1-dimensional profile and the 10 molecules used to build the hERG profile are shown in Figure 6. In this case, the hERG test set consists of 92 known hERG inhibitors. The central feature of the hERG profile is a basic amine; only Loratadine lacks this feature. The basic center is immediately surrounded by aliphatic carbons. Both ends of the profile are dominated by aromatic rings.

The performance of the hERG profile, shown in Figure 7, is not nearly as crisp as those with the SRC data set. In this case 6% of the known hERG inhibitors are ranked in the top 1% of the MDDR compounds, 25% of the known hERG inhibitors are ranked in the top 3% of the MDDR compounds, and 30% of the known hERG inhibitors are ranked in the top 5% of the MDDR compounds. Thus, the enrichment factors with the hERG data set range from 6 to 8 when considering the top 1−5% of the overall compounds compared to enrichment factors of 11−26 with the SRC data set over the same range. The difference in performance likely stems from the fact that the hERG potassium channel is much less specific in its binding requirements when compared to typical proteins such as SRC. This affects results such as these in numerous ways. First, there are in all likelihood far more hERG inhibitors than SRC inhibitors within the MDDR database. Second, the hERG data set is much more diverse than the SRC data set. Both of these considerations would be expected to increase the difficulty in tests such as these.

A Comparison to 3-Dimensional Methods. Here we compare the results with the 1-dimensional profiles to those with two 3-dimensional methods: explicit pharmacophore models and 3-point pharmacophore fingerprints. These methods are commonly used and well validated and therefore create good benchmarks for comparison.

In all cases, the conformations were generated for the molecules using Catalyst[44] with the fast option and a 15 kcal/mol strain cutoff. The pharmacophore fingerprints were generated from within Cerius2[45] using a 10 Å grid with a uniform 2.0 Å spacing and using all available features: negative charge, positive charge, negative ionizable, positive ionizable, hydrogen bond acceptor, hydrogen bond acceptor projection, hydrogen bond donor, hydrogen bond donor projection, aromatic ring, aromatic ring projection, and hydrophobic features. From the pharmacophore fingerprints of two molecules the Tanimoto coefficient was used as the similarity measurement. Finally, to rank the compounds of the MDDR database and the test sets, both the mean and maximum similarity to the compounds in the training set were used.

To create a pharmacophore model for hERG, we recreated the model described by Ekins and co-workers.[46] This model consists of a positive ionizable feature and four hydrophobic features. To create a pharmacophore model for SRC, the crystal structures 1qcf (HCK),[47] 1m17 (EGFr),[48] and 1m52(ABL)[49] are used, each of which is cocrystalized with a ligand similar to at least one of the compounds in the training set (see Figure 8). Two pharmacophore models, a four feature and a six feature model, were then created by manually mapping the observed conserved interactions. The four feature model consists of a hydrogen bond acceptor representing the aromatic nitrogen that interacts with the backbone NH of the hinge region (Met341 of SRC), an aromatic ring feature representing the heteroaromatic core, an aromatic ring representing the group in

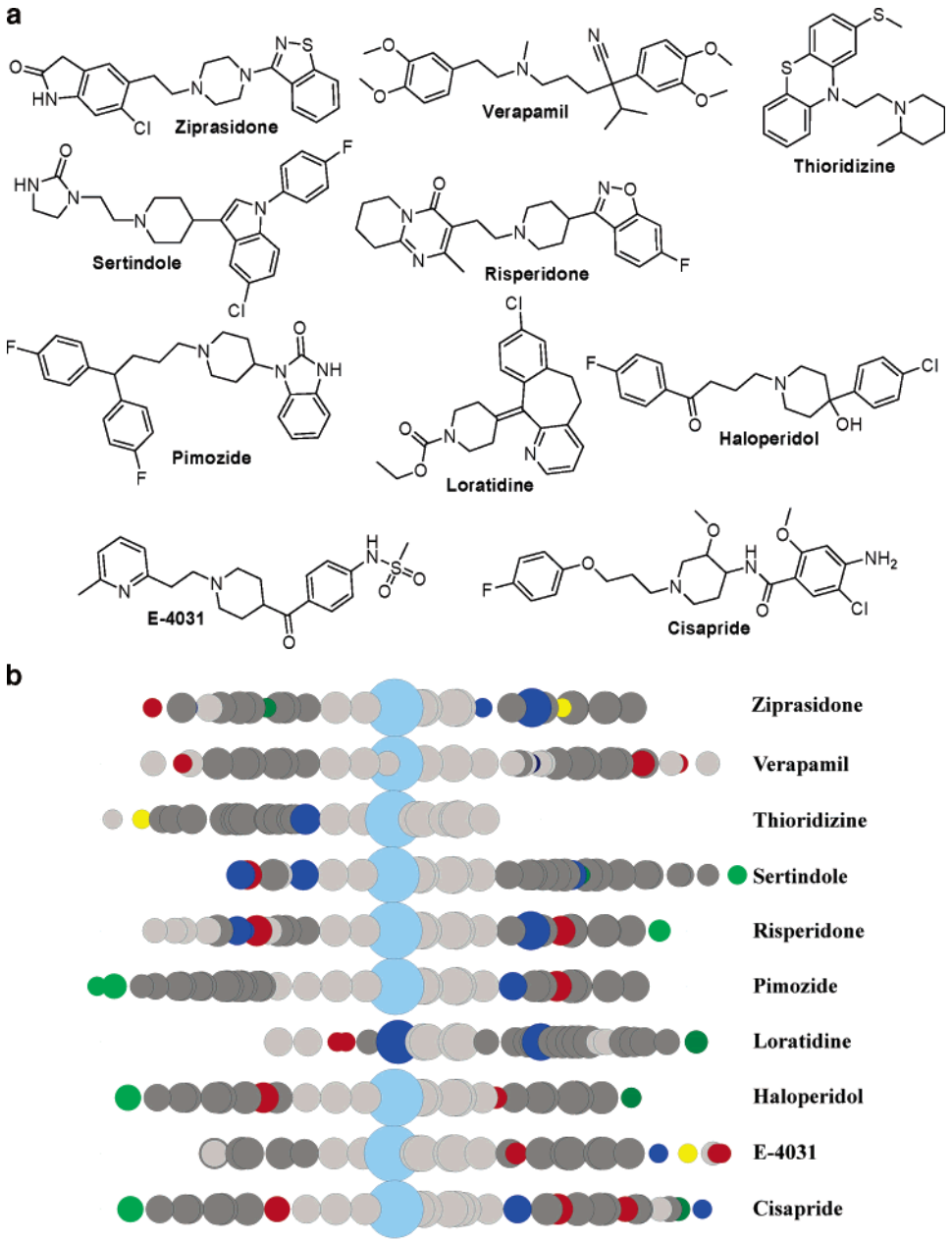

Figure 6. The hERG 1-dimensional Multiple Ligand Alignment. (a) The compounds used in the alignment. (b) The hERG 1-dimensional alignment. Each compound is shown in **6a** roughly oriented in the same manner as in the alignment in part b. The atoms are colored and sized as described in Figure 4b.

the main hydrophobic pocket, and a hydrophobic feature representing the groups in the secondary hydrophobic patch. The six feature model consists of these same four features with two additional hydrogen bond donors that represent the interactions with the backbone carbonyl oxygens of the SRC hinge region (Glu339 and Met341). The donors are observed in some cocrystallized kinase inhibitors of the same chemotype used in the SRC training set but not in others. With both hERG and SRC pharmacophore models the "Fit Value" produced by the Catalyst utility citest was used to rank the compounds in the MDDR database and in the test sets.

The results from the comparisons of the 1-dimensional profiles to the 3-dimensional methods are shown in Figure 9 (SRC) and Figure 10 (hERG). For both data sets the pharmacophore fingerprints and the explicit pharmacophore models produced reasonable enrichments. In both cases, however, the 1-dimensional profiles outperform the 3-dimensional methods. With the SRC data set, the 1-dimensional profile significantly outperforms either the pharmacophore model or the pharmacophore fingerprints. In particular, the 3-dimensional methods show significant enrichment initially but quickly resume random behavior. Compounds for which the bioactive conformation was generated were probably ranked very high whereas those for which the conformational search failed to produce a conformation reasonably close to the bioactive conformation were ranked within the noise produced by the MDDR compounds. Since it avoids any explicit conformation 1-dimensional multiple alignment does not face this problem.

Perhaps the largest difference between the methods is in the amount of time required to search a large database. For the discussion below all quoted times are from an SGI R10000. Searches with the 1-dimensional profiles can rank approximately 900 molecules per

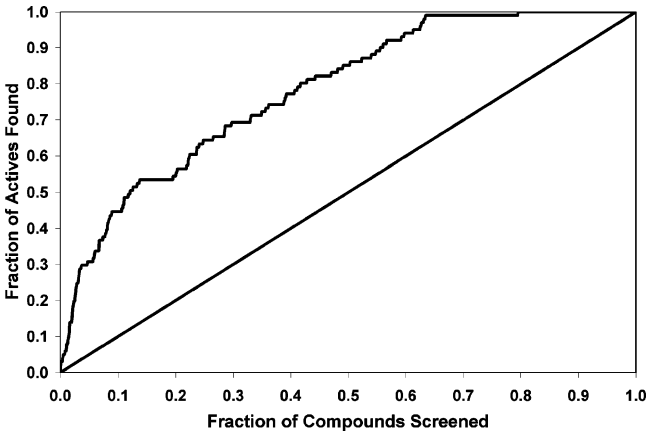

Figure 7. The performance of the hERG profile. This figure was created using the hERG 1-dimensional profile to rank the compounds of the MDDR and the hERG training set in the same manner as Figure 5.

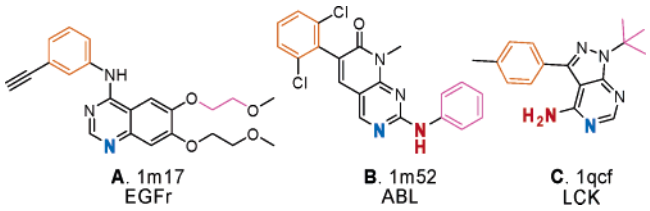

Figure 8. The compounds and the features used to build the SRC pharmacophore model. The three compounds were selected from crystal structures of related kinases because they are representative of the chemotypes found in the training set. **A** was taken from the crystal structure of 1m17[48] (EGFr). **B** was take from the pdb structure 1m52[49] (ABL). **C** was taken from pdb structure 1qcf[47] (HCK). The orange aromatic ring of the three ligands bound in the main hybrophobic pocket of the appropriate kinases. This interaction was represented by an aromatic ring feature. The blue nitrogen of the three ligands was observed to interact with the hinge NH (corresponds to Met341 of SRC) in each case. This interaction was represented by a hydrogen bond acceptor. The substructures highlighted in purple were bound in the secondary hydrophobic site and were represented as a hydrophobic feature. The central heteroaromatic core was represented by an aromatic ring feature. Finally, the NH of **B** was observed to interact with a backbone carbonyl oxygen of the kinase (corresponds to Met341 of SRC) while the NH2 of **C** was observed to interact with a second carbonyl oxygen of the kinase (corresponds to Glu339 of SRC). Each of these was represented by a hydrogen bond donor feature in the 6 feature model.

second meaning the search of the approximately 100 000 MDDR compounds takes less than 2 min. The conformation generation for the MDDR database took approximately 15 days of CPU time. Once the conformations had been generated the fingerprint generation took approximately 7 h. Once the conformations and corresponding pharmacophoric fingerprints were produced the pharmacophore similarities to the MDDR compounds were calculated in approximately 2 h. The generation of the pharmacophore fit value for the entire MDDR database took several weeks of CPU time, and as such this is not a practical means to search through large corporate databases. While speed is a secondary consideration it does dictate where the methods can be reasonably applied. Particularly, the speed and quality of searches with the 1-dimensional profiles makes them amenable to searching large databases of available or virtual compounds where 3-dimensional methods are not fast enough.

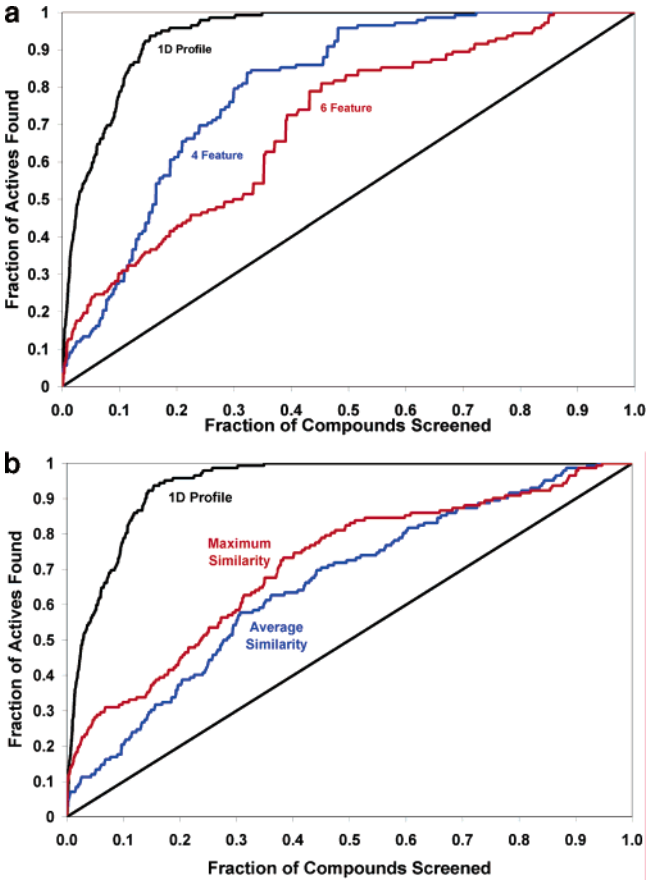

Figure 9. A comparison between the SRC 1-dimensional profile and 3-dimensional methods. (a) The performance of the 1-dimensional profile is in black. The performance of the 4 feature pharmacophore model is in red. The performance of the 6 feature pharmacophore model is in blue. The pharmacophore models are described in Figure 8. (b) The performance of the 3-dimensional profile is in black. The results with the 1-dimensional pharmacophore fingerprint are shown in blue and red. The maximum similarity between the training set compounds was used to create the red curve whereas the average similarity to the training set compounds was used to create the blue compounds.

Comparison with Pairwise 1-Dimensional Similarity. To demonstrate the value of the multiple alignment, we compare the performance of the multiple alignment with the 1-dimensional pairwise similarity. To do this, we compute the 1-dimesional similarity[15] between each of the molecules of the training set to each of the molecules of the test set and MDDR database. As with the 3-dimensional methods to rank the molecules of the test set and MDDR database, their maximum similarity and average similarity to the members of the training set were used.

For the SRC case in which the chemotypes of the test set are well represented by the training set the maximum similarity outperforms both the average similarity and the 1-dimensional profile (see Figure 11a). This result is not all that surprising as molecules of the same chemotype would be expected to show very high similarity to other molecules of the same chemotype. Thus, the maximum similarity metric should be well suited to finding highly related additional inhibitors. The average similarity and the 1-dimensional profile perform very similarly.

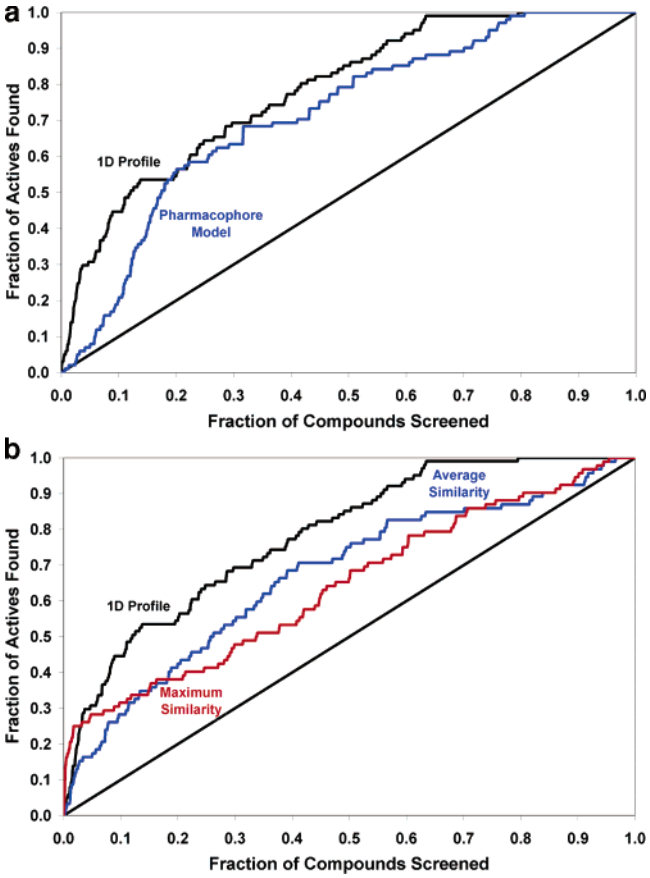

Figure 10. A comparison between the hERG 1-dimensional profile and 3-dimensional methods. (a) The performance of the 1-dimensional profile is in black. The performance of the hERG pharmacophore model is in blue. The hERG pharmacophore model used is a reproduction of the model by Ekins and coworkers.[46] (b) The performance of the 1-dimensional profile is in black. The results with the 1-dimensional pharmacophore fingerprint are shown in blue and red. The maximum similarity between the training set compounds was used to create the red curve whereas the average similarity to the training set compounds was used to create the blue compounds.

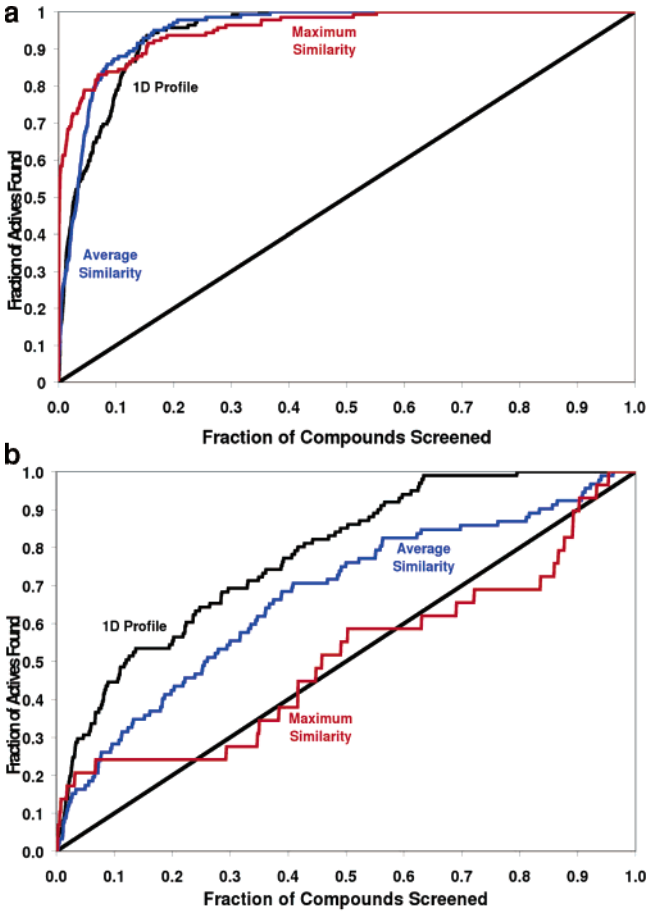

Figure 11. A comparison of the 1-dimensional profiles to pairwise 1-dimensional similarity. In both cases the results with the 1-dimensional profile are shown in black, the results with the maximum similarity metric are shown in red, and the results with the average similarity metric are shown in blue. (a) The SRC example. (b) The hERG case.

The hERG test case better demonstrates the value of the multiple alignment profiles over the pairwise alignment. In this case the majority of the chemotypes in the test set are not represented in the training set. In this case the multiple alignment profile outperforms both the maximum and average similarity metrics. As expected the maximum similarity metric performs quite poorly. After finding approximately 20% of the known inhibitors, the maximum similarity metric trails off to random performance. This is expected, as most of the test set was not represented in the training set. Interestingly, the average similarity performs significantly better than random even though the maximum similarity does not. The 1-dimensional profile significantly outperforms both the average and maximum similarity. For example the average similarity metric finds 28% of the known inhibitors within the top 10% of the overall compounds compared to 45% for the 1-dimensional profile.

It is also noteworthy that for the pairwise similarity methods the time required to search a large database will increase linearly with the number of molecules in the training set. The methods used to perform the database searches with a 1-dimensional profile are independent of the number of molecules in the training set. Thus, for these cases where we had 10 molecules in the training set the database searches with the 1-dimensional profile were over 10 times faster than the pairwise searches.

Discussion

Because it involves discrete and continuous variables with a complex landscape, multiple ligand alignment offers a challenging numerical optimization problem. Despite its heuristic nature the combination of a genetic algorithm and evolutionary programming proved to be a robust way to generate near optimal multiple ligand alignments. It is likely that for similar problems with mixed continuous and discrete variables this combination will prove fruitful.

Despite the unconventional nature of the approach, the 1-dimensional representation of small molecules retains much of the information present in a small molecule structure. Much like a 3-dimensional pharmacophore model, a 1-dimensional profile can effectively isolate the common features of a set of molecules with the same biological activity. This in turn makes the profiles useful for searching databases of available small molecules for compounds with the same biological activity. In addition, the improvements in running times are accompanied by improvements in enrichment. In

both cases examined here, the 1D methods outperformed the 3D methods. This is significant considering that virtual screening methods such as these are best positioned prior to synthesis and applied to large virtual chemical libraries. The observed improvements in accuracy and the considerable decrease in time required for virtual screening make searching large virtual collections tractable and allows the profiles to be used multiple times while speeding up the turn-around time in the cyclic design of combinatorial libraries.

Future work in this area will include more rigorous atom similarity calculations. While the atom similarity matrices used for this work showed improvement over the identity matrices there is little reason to believe they are optimal. Improvements could include focusing more on properties of the atoms such as hydrogen bonding capacity, aromaticity, etc., rather than explicit atom types. Furthermore, the alignment procedures could be tailored to include both molecules with the biological activity of interest and molecules known to be devoid of the biological activity of interest. Incorporating the negative information would further focus the profiles to the key features of the molecules and might help to eliminate many of the false positives found in the virtual screening experiments.

References

(1) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.

(2) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 1273−1280.

(3) Makara, G. M. Measuring molecular similarity and diversity: total pharmacophore diversity. *J. Med. Chem.* **2001**, *44*, 3563−3571.

(4) Mason, J. S.; Good, A. C.; Martin, E. J. 3-D pharmacophores in drug discovery. *Curr. Pharm. Des.* **2001**, 7, 567−597.

(5) McGregor, M. J.; Muskal, S. M. Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 569−574.

(6) McGregor, M. J.; Muskal, S. M. Pharmacophore fingerprinting. 2. Application to primary library design. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 117−125.

(7) Rarey, M.; Stahl, M. Similarity searching in large combinatorial chemistry spaces. *J. Comput. Aided Mol. Des.* **2001**, *15*, 497−520.

(8) Lemmen, C.; Lengauer, T.; Klebe, G. FLEXS: a method for fast flexible ligand superposition. *J. Med. Chem.* **1998**, *41*, 4502−4520.

(9) Lemmen, C.; Lengauer, T. Computational methods for the structural alignment of molecules. *J. Comput. Aided Mol. Des.* **2000**, *14*, 215−232.

(10) Barrett, A. J.; Rawlings, N. D. Families and clans of serine peptidases. *Arch. Biochem. Biophys.* **1995**, *318*, 247−250.

(11) Fischer, D.; Wolfson, H.; Lin, S. L.; Nussinov, R. Three-dimensional, sequence order-independent structural comparison of a serine protease against the crystallographic database reveals active site similarities: potential implications to evolution and to protein folding. *Protein Sci.* **1994**, *3*, 769−778.

(12) Krem, M. M.; Di Cera, E. Molecular markers of serine protease evolution. *EMBO J.* **2001**, *20*, 3036−3045.

(13) Rawlings, N. D.; Barrett, A. J. Families of serine peptidases. *Methods Enzymol.* **1994**, *244*, 19−61.

(14) Guner, O. F. History and evolution of the pharmacophore concept in computer-aided drug design. *Curr. Top. Med. Chem.* **2002**, *2*, 1321−1332.

(15) Dixon, S. L.; Merz, K. M., Jr. One-dimensional molecular representations and similarity calculations: methodology and validation. *J. Med. Chem.* **2001**, *44*, 3795−3809.

(16) Borg, I.; Groenen, P. *Modern Multidimensional Scaling: Theory and Applications*; Springer: New York, 1997.

(17) Mohammad, S.; Zhou, Z.; Gong, Q.; January, C. T. Blockage of the HERG human cardiac K+ channel by the gastrointestinal prokinetic agent cisapride. *Am. J. Physiol.* **1997**, *273*, 2534−2538.

(18) Kongsamut, S.; Kang, J.; Chen, X.; Roehr, J.; Rampe, D. A comparison of the receptor binding and HERG channel affinities for a series of antipsychotic drugs. *Eur. J. Pharmacol.* **2002**, *450*, 37−41.

(19) Fogel, L. J.; Owens, A. J.; Walsh, M. J. *Artificial intelligence through simulated evolution*; John Wiley and Sons: New York, 1966.

(20) Holland, J. H. *Adaptation is natural and artificial systems*; Univesity of Michigan Press: Ann Arbor, 1975.

(21) Baeck, T.; Fogel, D. B.; Michalewicz, Z.; Back, T.; Fogel, D. R. R. *Evolutionary Computation 1: Basic Algorithms and* Operators; Institute of Physics: Bristol, UK, 2000.

(22) Baeck, T.; Fogel, D. B.; Michalewicz, Z. *Evolutionary Computation 2: Advanced Algorithms and Operators*; Institute of Physics: Bristol, UK, 2000.

(23) Schwartz, R. M.; Dayhoff, M. O. Matrices for detecting distant relationships. *Atlas of Protein Sequence and Structure*; National Biomedical Research Foundation. Washington, D.C., 1978.

(24) Henikoff, S.; Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 10915−10919.

(25) Gonnet, G. H.; Cohen, M. A.; Benner, S. A. Exhaustive matching of the entire protein sequence database. *Science* **1992**, 256, 1443−1445.

(26) Jones, D. T.; Taylor, W. R.; Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **1992**, *8*, 275−282.

(27) MDDR *MDL Drug Data Report*; MDL: San Leandro, CA.

(28) Press: W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical recipes in C: The art of scientific computing*, 2nd ed.; Cambridge University Press: New York, 1993; 994.

(29) Frame, M. C. Src in cancer: deregulation and consequences for cell behaviour. *Biochim. Biophys. Acta* **2002**, *1602*, 114−130.

(30) De Ponti, F.; Poluzzi, E.; Cavalli, A.; Recanatini, M.; Montanaro, N. Safety of nonantiarrhythmic drugs that prolong the QT interval or induce torsade de pointes: an overview. *Drug Safety* **2002**, *25*, 263−286.

(31) Roy, M.; Dumaine, R.; Brown, A. M. HERG, a primary human ventricular target of the nonsedating antihistamine terfenadine. *Circulation* **1996**, *94*, 817−823.

(32) Wang, Y. D.; Miller, K.; Boschelli, D. H.; Ye, F.; Wu, B. et al. Inhibitors of src tyrosine kinase: the preparation and structure−activity relationship of 4-anilino-3-cyanoquinolines and 4-anilinoquinazolines. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 2477−2480.

(33) Trumpp-Kallmeyer, S.; Rubin, J. R.; Humblet, C.; Hamby, J. M.; Showalter, H. D. H. Development of a binding model to protein tyrosine kinases for substituted pyrido[2,3-d]pyrimidine inhibitors. *J. Med. Chem.* **1998**, *41*, 1752−1763.

(34) Traxler, P.; Green, J.; Mett, H.; Sequin, U.; P., F. Use of a pharmacophore model for the design of EGFR tyrosine kinase inhibitors: isoflavones and 3-phenyl-4(1H)-quinolones. *J. Med. Chem.* **1999**, *42*, 1018−1026.

(35) Traxler, P.; Bold, G.; Frei, J.; Lang, M.; Lydon, N. et al. Use of a Pharmacophore Model for the Design of EGF−R Tyrosine Kinase Inhibitors: 4-(Phenylamino)pyrazolo[3,4-d]pyrimidines. *J. Med. Chem.* **1997**, *40*, 3601−3616.

(36) Klutchko, S. R.; Hamby, J. M.; Boschelli, D. H.; Wu, Z.; Kraker, A. J. et al. 2-Substituted Aminopyrido[2,3-d]pyrimidin-7(8H)-ones. Structure−Activity Relationships Against Selected Tyrosine Kinases and in Vitro and in Vivo Anticancer Activity. *J. Med. Chem.* **1998**, *41*, 3276−3292.

(37) Hamby, J. M.; Connolly, C. J.; Schroeder, M. C.; Winters, R. T.; Showalter, H. D. et al. Structure−activity relationships for a novel series of pyrido[2,3-d]pyrimidine tyrosine kinase inhibitors. *J. Med. Chem.* **1997**, *40*, 2296−2303.

(38) Boschelli, D. H.; Wang, D. Y.; Ye, F.; Yamashita, A.; Zhang, N. et al. Inhibition of Src kinase activity by 4-anilino-7-thienyl-3-quinolinecarbonitriles. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 2011−2014.

(39) Boschelli, D. H.; Wu, Z.; Klutchko, S. R.; Showalter, H. D.; Hamby, J. M. et al. Synthesis and tyrosine kinase inhibitory activity of a series of 2-amino-8H-pyrido[2,3-d]pyrimidines: identification of potent, selective platelet-derived growth factor receptor tyrosine kinase inhibitors. *J. Med. Chem.* **1998**, *41*, 4365−4377.

(40) Thompson, A. M.; Connolly, C. J.; Hamby, J. M.; Boushelle, S.; Hartl, B. G. et al. 3-(3,5-Dimethoxyphenyl)-1,6-naphthyridine-2,7-diamines and Related 2-Urea Derivatives Are Potent and Selective Inhibitors of the FGF Receptor-1 Tyrosine Kinase. *J. Med. Chem.* **2000**, *43*, 4200−4211.

(41) Missbach, M.; Altmann, E.; Widler, L.; Susa, M.; Buchdunger, E. et al. Substituted 5,7-diphenylpyrrolo[2,3-d]pyrimidines: po-

tent inhibitors of the tyrosine kinase c-Src. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 945−949.
(42) Connolly, C. J. C.; Hamby, J. M.; Schroeder, M. C.; Barvian, M.; Lu, G. H. et al. Discovery and structure−activity studies of a novel series of pyrido[2,3-d]pyrimidine tyrosine kinase inhibitors. *Bioorg. Med. Chem. Lett.* **1997**, *7*, 2415−2420.
(43) Toledo, L. M.; Lydon, N. B.; Elbaum, D. The structure-based design of ATP-site directed protein kinase inhibitors. *Curr. Med. Chem.* **1999**, *6*, 775−805.
(44) Catalyst; 4.8 ed.; Accelrys, Inc.: San Diego, CA.
(45) Cerius2; 4.9 ed.; Accelrys Inc.: San Diego, CA.
(46) Ekins, S.; Crumb, W. J.; Sarazan, R. D.; Wikel, J. H.; Wrighton, S. A. Three-dimensional quantitative structure−activity relationship for inhibition of human ether-a-go-go-related gene potassium channel. *J. Pharmacol. Exp. Ther.* **2002**, *301*, 427−434.
(47) Schindler, T.; Sicheri, F.; Pico, A.; Gazit, A.; Levitzki, A. et al. Crystal structure of Hck in complex with a Src family selective tyrosine kinase inhibitor. *Mol. Cell* **1999**, *3*, 639−648.
(48) Stamos, J.; Sliwkowski, M. X.; Eigenbrot, C. Structure of the epidermal growth factor receptor kinase domain alone and in complex with a 4-anilinoquinazoline inhibitor. *J. Biol. Chem.* **2002**, *277*, 46265−46272.
(49) Nagar, B.; Bornmann, W. G.; Pellicena, P.; Schindler, T.; Veach, D. R. et al. Crystal structures of the kinase domain of c-Abl in complex with the small molecule inhibitors PD173955 and imatinib (STI-571). *Cancer Res.* **2002**, *62*, 4236−4243.
(50) de Laszlo, S. E.; Hacker, C.; Li, B.; Kim, D.; MacCoss, M. et al. Potent, orally absorbed glucagon receptor antagonists. *Bioorg. Med. Chem. Lett.* **1999**, *9*, 641−646.

JM050563R