PZIM: A Method for Similarity Searching Using Atom Environments and 2D Alignment

Anders E. Berglund*,† and Richard D. Head†

Exploratory Immunobiology Inflammation & Immunology Research Unit, Pfizer Global Research and Development, Chesterfield, Missouri 63017, United States

Received May 24, 2010

The advent of extensive small molecule databases has brought with it the problem of searching these repositories for entities with desired properties. A multitude of similarity-searching algorithms, based on different underlying methods, currently exist for this purpose. However, due to the somewhat nebulous definition of "similar", all such approaches maintain different strengths and weaknesses. Presented here is PZIM, a new approach fundamentally based on a description of the atom environment that includes multiple adjustable features allowing for searches to be tailored on the basis of the user definition of similarity. In addition to flexible atom environment size, PZIM allows for the use of an atom-substitution matrix to identify similar pharmacophoric recognition elements. Finally, PZIM produces 2-dimensional alignments of all compared molecules that pass a user-defined threshold for similarity. To determine the usefulness of the approach, PZIM was compared to seven other similarity-searching methods on nine data sets recently published. PZIM achieved a rank of first or second in the majority of cases tested and obtained the highest average rank score across all methods tested. These results demonstrate the effectiveness of the PZIM approach across diverse search conditions.

INTRODUCTION

Over the past 20 years, chemical databases utilized by industry and academia have grown tremendously in size. With this expansion has come the challenge of searching these databases to recognize molecules with desired properties. One such problem, known as similarity searching, entails the identification of compounds with structural features analogous to a known starting molecule. Similarity searching is commonly performed as the basis for numerous computational analyses such as ligand-based virtual screening, HTS triaging/expansion, and nearly any type of molecular clustering. Several diverse techniques are available to perform such investigations. Many of these current solutions are based on fingerprints, atom environment, structural keys, or combinations of these. Fingerprints can be predefined sets of structural keys, calculated descriptors, or both.[1] Due to the nature of the precalculated keys and descriptors, fingerprinting programs tend to be very fast and memory-efficient, especially when searching databases of millions of compounds. The caveat is that the calculation of similarity is constrained to these predefined entities. An alternative approach makes use of the atom environment, which does not require prior definition of structural keys.[2−5] Instead, these techniques attempt to identify atoms with similar neighbors between the compared molecules. Most of these methods are 2-dimensional in nature, meaning that the molecule is viewed as flat, or planar. There are, however, both 1-dimensional and 3-dimensional approaches which attempt to optimize for speed or accuracy, respectively.[6−8] Three-dimensional methods, also called 3D-QSAR, are particularly useful for virtual screening where there are known inhibitors cocrystallized with a protein target.[9] All of these techniques have their own strengths and weaknesses with some best suited for specific types of searching task.

The primary drawback with the traditional, fast, fingerprint-based methods is that they are not always optimal in pharmacophore space. With, newer methods that use topological information,[2,4] this drawback has been minimized. If the right components are present within the given molecule, it may receive a high score independently of how they are interconnected. A common use of similarity searching is to scan through a large set of molecules using one or several known active compounds, query molecules, to find new molecules with similar activity. In such a case, it is essential that the functional groups responsible for the binding of the ligand to the receptor are oriented similarly in the query molecule and the hits from the database. One way to achieve a pharmacophore-based similarity measure is to align the two molecules and adjust the score on the basis of how well the components overlap. The concept of alignment based on connecting atoms and graphs was proposed some time ago, such as Morgan described it for determining unique structures as early as 1965.[10]

In this paper, we will present a new method entitled PZIM. PZIM is a 2D searching method that incorporates a number of concepts, including atom environment, atom similarity, and molecular alignment, within a single adjustable approach. PZIM is then compared to a number of other methods with a previously published benchmark dataset.

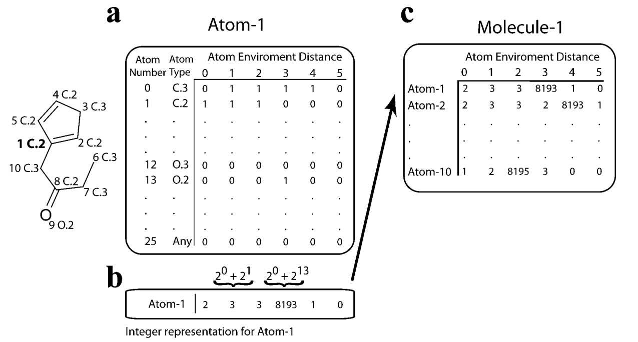
METHODOLOGY

Molecule Representation and Atom Types. The Tripos MOL2[11] file format was used for the representation of molecules in the work described here. In addition, the MOL2 atom-type definitions were also used for consistency. How-

* Corresponding author. E-mail: anders.e.berglund@pfizer.com.

† Current address: Indications Discovery Unit, Pfizer, Inc, 4320 Forest Park Avenue, Suite 302, St. Louis, MO 63108.

10.1021/ci1002075
Published on Web 10/01/2010



Atom Number	Atom Type	Atom Enviroment Distance 0	1	2	3	4	5
0	C.3	0	1	1	1	1	0
1	C.2	1	1	1	0	0	0
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
12	O.3	0	0	0	0	0	0
13	O.2	0	0	0	1	0	0
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
25	Any	0	0	0	0	0	0

Atom-1	2	3	3	8193	1	0

	Atom Enviroment Distance 0	1	2	3	4	5
Atom-1	2	3	3	8193	1	0
Atom-2	2	3	3	2	8193	1
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
Atom-10	1	2	8195	3	0	0

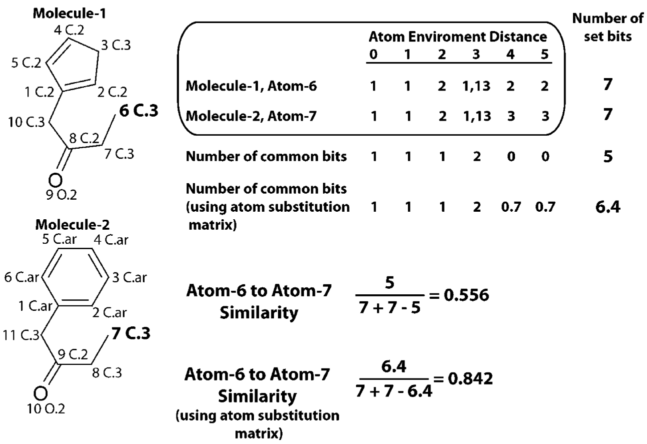
Figure 1. (a) Example of how atom-1 can be represented by a matrix of set bits. (b) The matrix can be converted to a vector of integers. (c) Example of how the whole molecule can be represented as a matrix of integers.

ever, it is important to note that the method is not restricted to this format or the given atom-type definitions. In all, 25 different atom types are utilized that represent most common atom types in typical small molecule libraries If an atom is encountered that is not contained within the given definition set, it is currently defined as "any". A list of the atom types is given in the Supporting Information.

Atom Environment. For this method, atom environment describes the surroundings of a given atom in terms of its bonded neighbors. The environment can be calculated on the basis of immediate neighbors, those directly bonded to the current atom, or can be expanded to additional levels that included neighbors of the immediate neighbors and so on.

In PZIM, the atom environment (AE) distance is defined by the number of bonding levels between two atoms. For example, an AE distance of 0 refers to the current atom. An AE distance of 1 refers to all atoms which are directly bonded to the current atom. A distance of 2 includes atoms which are bonded to level 1 atoms and so forth. The atom environment is then represented by a set of binary vectors. Each vector encodes data that captures all atom types present at a given distance from the current atom. Thus, to represent all atoms within an AE distance of 3 from a given atom, there will be four binary vectors generated, including the AE = 0 vector for the current atom. The vectors are fixed at a length of 25 bits, as there are currently 25 atom types defined within PZIM, and atom types present are represented by setting their corresponding bit. Figure 1 provides an illustration of how atom environment vectors are derived for an example molecule.

Figure 1a depicts how the AE vectors are derived for atom 1 (molecule 1) using an AE distance of 5. Each row in the matrix corresponds to an atom type, while each column represents a specific AE distance. For the first column, AE distance = 0, the second bit is set since atom 1 is a C.2. In the second column, the first and second bits are set (C.3 and C.2) for the atom types that are directly bonded to atom 1. This is repeated for all AE distances to be computed. This matrix of set bits can be converted to a vector of integers, as seen in Figure 1b. This is done by converting the binary vector to an integer, where the set bits make up the value of the integer. Two cases are given to exemplify how the integers are derived. This is repeated for every atom in the molecule, resulting in a matrix of integers, as is illustrated in Figure 1c. This final matrix represents the entire molecule to be evaluated and only has to be calculated a single time.



	Atom Enviroment Distance 0	1	2	3	4	5	Number of set bits
Molecule-1, Atom-6	1	1	2	1,13	2	2	7
Molecule-2, Atom-7	1	1	2	1,13	3	3	7
Number of common bits	1	1	1	2	0	0	5
Number of common bits (using atom substitution matrix)	1	1	1	2	0.7	0.7	6.4

Atom-6 to Atom-7 Similarity $\frac{5}{7+7-5} = 0.556$

Atom-6 to Atom-7 Similarity (using atom substitution matrix) $\frac{6.4}{7+7-6.4} = 0.842$

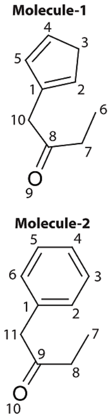
Figure 2. Calculation of the atom-to-atom similarities between atom 6 (molecule 1) and atom 7 (molecule 2). The figure also demonstrates how the similarity can be adjusted using an atom substitution matrix.

This allows for a very compact representation of the molecule that can then be searched against other entities. Together with the connectivity table, this is the entirety of the information that needs to be stored for calculating the similarity between two molecules.

Comparing the Similarity between Two Molecules. The first step in calculating the similarity between two molecules is to compute the similarity between every combination of atom pairs within them. This will result in a similarity matrix for the molecule pair. Contained within the matrix are the $N \times K$ atom-to-atom similarity scores, where N is the number of atoms in molecule 1 and K the number in molecule 2. The atom-to-atom similarity scores will be described next.

Atom Similarity Score. The similarity between two atoms is calculated by comparing the number of common atom types at each AE distance (Figure 2). Within this approach, the Tanimoto similarity index is utilized to compute the final similarity. The example given in Figure 2 compares atom 6 (molecule 1) to atom 7 (molecule 2). Atom 6 (molecule 1) has a total of seven set bits, one set bit for each AE distance, with the exception of two set bits (value 8193) at an AE distance of 3 due to the presence two different atom types at that level (C.3, O.2). Atom 7 (molecule 2) has the same number of set bits. For the first four AE distances, the sets bits are identical, resulting in five common bits between the two atoms. This results in a Tanimoto value of 0.556 (Figure 2), describing the similarity of atom 6 (molecule 1) to atom 7 (molecule 2).

Atom Similarity Correction and Atom Substitution Matrices. As is illustrated by the above description and Figure 2, the similarity between the two compared atoms is largely penalized by the fact that molecule 1 contains C.2 atoms in an equivalent location to C.ar atoms in molecule 2. The result of a discrete present/absent score is that comparing C.2 to C.ar is equally different to comparing C.2 to N.2. To account for physicochemical similarity in atom types, an atom-substitution matrix has been introduced. This substitution matrix borrows from the concept of PAM[12] and BLOSUM[13] matrices utilized to describe similarity between amino acids. Wang et al. also used an atom substitution in their similarity method based on multiple alignment profiles.[8] As there are few published matrices of this nature based on atom properties, a similarity matrix was defined on the basis



		Molecule-2											Selected Row Values
		1	2	3	4	5	6	7	8	9	10	11	
Molecule-1	1	**0.70**	0.38	0.30	0.36	0.30	0.38	0.53	0.43	0.44	0.32	0.48	0.70
	2	0.46	0.65	0.35	0.42	0.35	**0.65**	0.38	0.48	0.49	0.32	0.41	0.65
	3	0.32	0.36	0.67	0.41	**0.67**	0.36	0.54	0.43	0.34	0.39	0.38	0.67
	4	0.42	0.34	**0.62**	0.39	0.62	0.34	0.49	0.47	0.40	0.35	0.37	0.62
	5	0.38	**0.72**	0.38	0.46	0.38	0.72	0.36	0.39	0.48	0.35	0.37	0.72
	6	0.48	0.30	0.43	0.46	0.43	0.30	**0.84**	0.44	0.40	0.35	0.33	0.85
	7	0.30	0.37	0.37	0.39	0.37	0.37	0.40	**0.72**	0.38	0.34	0.50	0.72
	8	0.42	0.42	0.30	0.36	0.30	0.42	0.35	0.33	**0.62**	0.25	0.34	0.62
	9	0.27	0.41	0.45	0.35	0.45	0.41	0.32	0.34	0.30	**0.69**	0.35	0.69
	10	0.40	0.40	0.33	0.26	0.33	0.40	0.35	0.46	0.32	0.36	**0.67**	0.67
Selected Column Values		0.70	0.72	0.62	0	0.67	0.65	0.84	0.72	0.62	0.69	0.67	

$$\text{PZIM Score: } \frac{\text{SCV} + \text{SRV}}{\text{N} + \text{K}} = \mathbf{0.655}$$

Figure 3. Atom similarity matrix between molecule 1 and molecule 2. The bold red values indicate the optimal alignment of the two molecules.

of atom type, number of bonds, number of H-bond donors, and number of H-bond acceptors. The atom substitution matrix is given in Supporting Information Table S2.

The following rules have been applied for calculating the substitution matrix. First, two atoms of the same basic type (i.e., carbon) receive a value of 0.5. If the atoms share the same number of bonds, an additional increase of 0.2 is given; if the number of bonds differs by one, that increase is only 0.1. For atoms of any type, a value of 0.1 is added if they both act as H-bond donors, with an additional increase of 0.1 if they have the same number of donors. Similar criteria are used for scoring H-bond acceptors. Halogens are exempt from these rules, and their similarity is fixed to 0.7. Optimum scoring for atom similarity can certainly be debated; however, for this work, these rules produced reasonable results upon manual inspection.

With the atom substitution matrix, the number of common bits is adjusted from 5 to 6.4 for the example given in Figure 2. This increase is due to the substitution score of 0.7 replacing the original value of 0 when comparing C.2 to C.ar. This overall impact is to increase the final atom similarity from 0.556 to 0.842.

Molecular Similarity Matrix. The atom-to-atom similarity is calculated between all possible atom pairs in molecule 1 and molecule 2 (Figure 3). The resulting molecular similarity matrix is the foundation for the PZIM similarity score and is also used for the alignment of the two molecules, as will be described here.

Alignment of the Two Molecules and Calculation of the PZIM Score. To calculate a final molecular similarity score utilizing atom environments, as is done in PZIM, the two molecules require alignment. The following procedure has been developed for finding an optimal, or near optimal, alignment. To guarantee that the optimum is achieved in every instance, an exhaustive search would be required. As this would make the procedure too expensive computationally to compare large numbers of molecules to common databases of millions, a rapid assessment method was constructed. The top *N* values, where *N* is an adjustable parameter, in the similarity matrix are selected as alignment starting positions. The top value corresponds to the two most similar atoms in the two molecules, 0.84 in the example given in Figure 3, with respect to their local environment. After a given starting point is selected, all other values in that row and column are set to zero in the similarity matrix, ensuring that they will not be considered again. The atoms connected to the starting atom, in both molecules, are then selected, in the given example atom 7 (molecule 1) and atom 8 (molecule 2). Note that if there is more than one atom connected, each atom pair will be examined starting with the pair with the highest similarity value. Of these atom pairs, the first one equal to the max for their respective row and column will be added. In the current example, 0.72 is the highest value both in row and column for atom 7 (molecule 1) and atom 8 (molecule 2) and is thus selected. The search repeats in this fashion until a matching pairing is identified for all atoms, in effect, a greedy depth first search. If all atom pairs fail to meet these criteria and the list of pairs to be tested is empty, a second pass will be done where the criteria are relaxed. In this case, a pair need only be at the maximum in either a row or column, not both. Finally, if this fails, the highest scoring atom pair will be added, regardless of whether it represents the maximum for its row or column. For both of these cases, only one atom pair will be added, as this will make the list of atom pairs to be tested nonempty.

The PZIM score is calculated as the sum of all selected values divided by the average number of atoms in molecule 1 and molecule 2. The whole process is then repeated *N* times with new starting atom pairs in an attempt to identify the best global alignment.

Calculation of SimMax. The primary object of most procedures of this nature is to find molecules, typically from a large database, that are similar to a starting search molecule or molecules. Molecules which are very dissimilar are, normally, not of much interest. With this knowledge in mind, an additional optimization step has been added. On the basis

Table 1. Recall Rates for the Nine Different Data Sets[a]

Method	RR100 (%)	RR1442 (%)	Method	RR100 (%)	RR1442 (%)	Method	RR100 (%)	RR1442 (%)
low similarity								
aldose reductase			leukotriene synthesis			cyclooxygenase-2		
PZIM	**6.0**	**15.2**	**PZIM**	**4.9**	**7.2**	**PZIM**	**7.3**	**16.3**
Molprint 2 D	2.4	6.4	Molprint 2 D	2.7	6.1	BCI	1.4	4
BCI	2.1	6.2	PDR-FP	0.7	5.7	Molprint 2 D	1.2	4.6
TGT	1.5	3.2	BCI	0.7	2.5	TGT	0.9	2.8
TGD	1.2	6.2	MPMFP	0.6	2.3	TGD	0.7	1.6
MPMFP	1.1	6.5	TGT	0.1	1	MPMFP	0.2	1.3
PDR-FP	0.7	2.4	TGD	0	0.6	PDR-FP	0.1	1.4
medium similarity								
thrombin			IL-1b-converting Enzyme			endothelin		
PDR-FP	31.2	47	PDR-FP	55.2	65.6	PDR-FP	22.6	46.2
PZIM	**14.8**	**24**	Molprint 2 D	39.3	59.2	**PZIM**	**14.6**	**24.2**
TGT	13.3	25.8	**PZIM**	**38.9**	**62**	TGT	7.6	18.4
Molprint 2 D	11.3	15.7	TGT	23.7	41.4	TGD	7.4	28
MPMFP	7.6	18.9	TGD	21.8	32.6	Molprint 2 D	7.3	16.2
TGD	6.3	11.2	MPMFP	5.2	17.2	MPMFP	4.7	12.8
BCI	5.1	14.8	BCI	5.1	13.9	BCI	3.8	15.4
high similarity								
angiotensin-II			renin			HIV protease		
PZIM	**64.4**	**79.7**	PDR-FP	88.2	95.7	PDR-FP	68.3	87.6
Molprint 2 D	56.7	67.7	Molprint 2 D	77.3	88.1	**PZIM**	**57**	**70.9**
PDR-FP	33.3	49	**PZIM**	**75.4**	**84.7**	Molprint 2 D	50.5	69.8
TGD	30.5	49.9	BCI	65.7	85.6	TGT	39.2	58.9
MPMFP	29.9	49.5	TGT	59.2	71.3	MPMFP	29	49
BCI	25.8	44.3	MPMFP	45.3	63.7	BCI	23.4	46.7
TGT	17.9	28.8	TGD	33.8	47.9	TGD	14.6	24.5

[a] Included are the results for the other methods in the article by Tovar et al.[14] RR_100 and RR_1442 are the percent recovery rates for the known inhibitors not used as query molecules in the top 100 and top 1442 molecules, respectively.

of the way in which the PZIM score is calculated from the molecular similarity matrix, it is possible to circumvent the alignment of dissimilar molecules all together. The maximum similarity value that any molecule pair can have is the larger of the column max value or row max value means. This value is defined as the SimMax. If a minimum similarity value is given at the beginning of the procedure, it will first be compared to the SimMax for each molecule pair. If the SimMax value is less than this minimum threshold, the alignment will not be done, and the procedure will move on to the next comparison.

Data for Validation and Calculation of Performance. In this paper, the data set published by Tovar et al.[14] was utilized to compare PZIM to existing methods. Within this, there are nine different data sets divided into three categories of low, medium, and high diversity. For each data set, there are 22 to 27 known inhibitors taken from MDDR. About 1.44 million compounds from the ZINC[15] database are used as "inactives". For each data set, five molecules are selected as a training set, and the rest of the known actives are put among the ~1.44 million inactives. The recovery rate, or number of known actives retrieved, among the top 100, RR100, and top 1442 (0.1%), RR1442, scoring compounds were calculated. This was repeated 1000 times with five randomly selected training set molecules.

An atom environment (AE) distance of 7 was used for the PZIM method. In all the examples, the nearest-neighbor search technique was used for PZIM.

RESULTS

Table 1 provides the results from the nine different data sets, and 18 total criteria, used in this study. The results for the other methods are taken directly from the publication of Tovar et al.[14] In the case of the three high diversity data sets, PZIM obtained recovery rates in the 6−7% range (top 100) and 7−16% (top 1442). In all six cases, PZIM obtained the highest recovery rate. Within the second ranked approaches for these test sets, the equivalent recovery rate ranges were 1.4−2.7% and 4−6.4%, respectively. For the three medium diversity cases, PZIM ranked second in four cases and third for two. Here, the recovery rates were 15−39% (top 100) and 31−55% (top 1442). The top ranked method achieved rates of 24−59% (top 100) and 46−65% (top 1442). Finally, within the low diversity sets, PZIM achieved ranks from first to third. Recovery rates ranged 57−75% (top 100) and 71−85% (top 1442) for PZIM and 57−88% (top 100) and 68−96% (top 1442) for the other top ranking approach.

DISCUSSION

The results demonstrate that PZIM is effective in all of the test cases given. As can be seen in Table 2, PZIM obtained the highest average rank with respect to the nine data sets and 18 total test conditions. In addition, PZIM only ranked lower than third (of seven) in one instance. The performance was most notable in the very difficult high diversity training set data. While PZIM ranked first in each of the three test conditions, perhaps the more striking observation was that the overall recovery rate was often 2 to 3 times that of the other approaches. The consistent rankings of first, second, or third in the remaining medium and low diversity training set conditions indicate that the approach is well suited for general use and not just a specific situation.

Table 2. Ranking for the Different Methods[a]

	RR_100			RR_1442		
method	mean rank	min rank	max rank	mean rank	min rank	max rank
PZIM	1.8	1(4)	3	2.0	1(4)	4
PDR-FP	2.8	1(5)	7	2.9	1(5)	7
Molprint 2 D	2.8	2	5	3.0	2	5
TGT	4.4	3	7	4.7	2	7
BCI	5.1	2	7	5.0	3	7
MPMFP	5.6	5	7	5.1	2	7
TGD	5.6	4	7	5.3	2	7

[a] In parentheses is the number of times the method was ranked first.

Table 3. Recall Rate for PZIM with or without Atom Similarity Correction[a]

	RR_100%		RR_1442%	
data set	ASC	No ASC	ASC	No ASC
aldose reductase inhibitors	6.0	5.6	15.2	12.6
leukotriene synthesis inhibitors	4.9	4.3	7.2	9.7
cyclooxygenase-2 inhibitors	7.3	7.9	16.3	17.7
thrombin inhibitors	14.8	8.9	20.0	21.2
IL-1b-convering enzyme inhibitors	38.9	40.6	62.0	63.6
endothelin antagonists	14.6	12.8	24.2	17.2
angiotensin-II antagonists	64.4	56.4	79.7	73.5
renin inhibitors	75.4	73.8	84.7	86.1
HIV protease inhibitors	57.0	55.6	70.9	73.7

[a] ASC is with atom similarity correction, and No ACS is without.

PZIM has multiple adjustable parameters that were tested for sensitivity. One of these parameters is the AE distance used to calculate the environment for each atom. Values of 3, 5, and 7 were examined, though all data reported here were for 7. While small differences occurred for some conditions, the results were quite similar, see Supporting Information. Distances of greater than 7 have never demonstrated any advantage in the scenarios tested thus far. Furthermore, there is a significant performance trade-off that exists with greater atom-environment distances, as more calculations are required for each additional level. It is conceivable that for large molecules, particularly those which are cyclic, that larger AE distances may provide some benefit. However, for typical small molecules, such as those in the given examples, it is likely just adding redundant information at the cost of CPU cycles.

The other major adjustable parameter is the use of the atom-substitution matrix. As was discussed in the Methodology section, this functionality was added to overcome the different limitations of strict or broad atom type assignments. Specifically, the atom-substitution matrix provides a compromise that allows for molecules to be identified with similar pharmacophoric recognition elements. This may be of particular use when scaffold hopping. Given the nature of the test cases published here, it was not expected that the use of the substitution matrix would provide any significant benefit. However, in the interest of thoroughness, the method was run with and without substitution on the entire test sets with the results given in Table 3. In most cases, the results are very similar with differences of less than 3%. Nevertheless, in a few instances, use of atom substitution did increase the recovery rate by more than 5%. The accuracy of the substitution matrix presented here can certainly be argued. That said, in certain instances, it appears to provide substantial benefit, and with no current examples of significant down side, it is difficult to argue against its inclusion. In internal studies, it has also proven quite useful in expanding from starting chemistry (data not shown here).

Two issues that have not been discussed in significant detail, to this point, are the processing time of PZIM and the resulting 2-dimensional molecular alignments. In part, this is due to the fact that little work has been done to optimize or rigorously evaluate these components, respectively. Generally speaking, the implementation of PZIM reported here is considerably slower than most other similarity searching methods. Run times are very dependent on molecule size and the adjustable parameters used in a particular search. That said, an average comparison, utilizing the parameters reported here, is about 0.0225 s on a 2.8 GHz Pentium 4. This translates to 44 molecular comparisons per second. As mentioned, however, little has been done to optimize the code for speed, so it is likely that this could be substantially improved. The molecular alignment process is a significant component of this run time due to the optimization process. The only adjustable parameter in the procedure is the maximum number of starting points to be used in an attempt to identify the optimum. For the work presented here, this value was set to 15. This could potentially have been adjusted on the basis of both molecule size and similarity, which, in turn, could impact run times. Highly similar molecules will often find an optimal alignment fairly quickly, typically within a single attempt. Highly dissimilar molecules will not be aligned if SimMax is below the specified cutoff. Thus, molecules of moderate similarity are those most impacted by this value. While a larger value increases the probability of finding the optimal alignment, it also increases the amount of CPU time spent on the molecule, so a considerable trade-off exists. Other values were tested (data not shown); however, a thorough investigation of this parameter remains. In addition, a rigorous evaluation of the alignments themselves needs to be completed through a thorough visual inspection and a direct comparison with other structural alignment methods.

CONCLUSIONS

In this paper, we have described a new similarity searching algorithm, PZIM, and compared the performance against several previously published methods utilizing the benchmark data set from Tovar et al.[14] The method performed very well with the highest average ranking for recovery rate. The uniqueness of PZIM lies not only within the novel and adjustable description of the atom environment but also with the inclusion of an atom-similarity matrix and molecular alignment. While the use of these components can add significant computer overhead, the program can be run in a highly parallelized fashion, and as many of these components are user adjustable, the speed/recovery benefits can be optimized to fit the current need. We believe these factors and the data presented here demonstrate the potential usefulness of the PZIM algorithm and some of the advantages the method has as compared to existing methods.

ACKNOWLEDGMENT

We thank Gennadiy Poda for extensive testing of PZIM.

Supporting Information Available: Tables describing the following: MOL2 atom types, the atom substitution matrix, and recall rates for different atom environment distances. This information is available free of charge via the Internet at http://pubs.acs.org/.

REFERENCES AND NOTES

(1) Eckert, H.; Bajorath, J. Design and Evaluation of a Novel Class-Directed 2D Fingerprint to Search for Structurally Diverse Active Compounds. *J. Chem. Inf. Model.* **2006**, *46*, 2515–2526.

(2) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. *J. Chem. Inf. Model.* **2004**, *44*, 1708–1718.

(3) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naive Bayesian Classifier. *J. Chem. Inf. Model.* **2004**, *44*, 170–178.

(4) Rogers, D.; Brown, R. D.; Hahn, M. Using Extended-Connectivity Fingerprints with Laplacian-Modified Bayesian Analysis in High-Throughput Screening Follow-Up. *J. Biomol. Screen.* **2005**, *10*, 682–686.

(5) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(6) Dixon, S. L.; Merz, K. M. One-Dimensional Molecular Representations and Similarity Calculations: Methodology and Validation. *J. Med. Chem.* **2001**, *44*, 3795–3809.

(7) Melville, J. L.; Riley, J. F.; Hirst, J. D. Similarity by Compression. *J. Chem. Inf. Model.* **2007**, *47*, 25–33.

(8) Wang, N.; DeLisle, R. K.; Diller, D. J. Fast Small Molecule Similarity Searching with Multiple Alignment Profiles of Molecules Represented in One-Dimension. *J. Med. Chem.* **2005**, *48*, 6980–6990.

(9) Clark, R. D. Prospective Ligand- and Target-Based 3D QSAR: State of the Art 2008. *Curr. Top. Med. Chem.* **2009**, *9*, 791–810.

(10) Morgan, H. L. Generation of a unique machine description for chemical structures--a technique developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–13.

(11) Tripos Mol2 File Format. http://www.tripos.com/mol2/atom_types.html. Accessed August 20, 2010.

(12) Dayhoff, M. O.; Schwartz, R. M.; Orcutt, B. C. A model of evolutionary change in proteins. *Atlas Protein Sequence Struct.* **1978**, *5*, 345–352.

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

(14) Tovar, A.; Eckert, H.; Bajorath, J. Comparison of 2D Fingerprint Methods for Multiple-Template Similarity Searching on Compound Activity Classes of Increasing Structural Diversity. *ChemMedChem* **2007**, *2*, 208–217.

(15) Irwin, J. J.; Shoichet, B. K. ZINC - A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.

CI1002075