Similarity Searching in Databases of Flexible 3D Structures Using Autocorrelation Vectors Derived from Smoothed Bounded Distance Matrices

Nicholas Rhodes,[†,‡] David E. Clark,*[,§] and Peter Willett[†]

Department of Information Studies, The University of Sheffield, Regent Court, 211 Portobello Street,
Sheffield, South Yorkshire, S1 4DP, United Kingdom, and Argenta Discovery Ltd., 8/9 Spire Green Centre,
Flex Meadow, Harlow, Essex, CM19 5TR, United Kingdom

Received September 9, 2005

This paper presents an exploratory study of a novel method for flexible 3-D similarity searching based on autocorrelation vectors and smoothed bounded distance matrices. Although the new approach is unable to outperform an existing 2-D similarity searching in terms of enrichment factors, it is able to retrieve different compounds at a given percentage of the hitlist and so may be a useful adjunct to other similarity searching methods.

INTRODUCTION

The use of similarity searching to locate compounds in a database that resemble a query structure in some way is commonplace within drug discovery.[1−3] For instance, it may be applied to identify compounds similar to a hit from a high-throughput screen or to "lead-hop" from one chemical series to another.[4,5] Similarity searching methods differ from one another in the manner in which the chemical structures are represented (i.e., the descriptors) and in the method used to evaluate similarity between two such representations.

One type of molecular representation for similarity searching is based on autocorrelation vectors, originally proposed by Moreau and Broto[6] to encode topological information derived from a structure diagram. Given that each atom, I, has an associated property $f(I)$, an autocorrelation vector, $\mathbf{V}$, is defined such that

$$\mathbf{V} = (V_1, V_2, ... V_J, ... V_{\mathrm{Max}})$$

where V_J contains the sum of products $f[K]*f[L]$ for all pairs of atoms, K and L, separated by exactly J bonds, $0 \le J \le$ Max, and Max is the maximum possible interatomic separation within that structure.

Autocorrelation vectors have subsequently been applied to similarity analysis by Moreau and Turpin,[7] Moreau et al.,[8] and Bauknecht et al.[9] The method was also included in Tripos' extensive evaluation of the neighborhood behavior of molecular descriptors.[10] The use of autocorrelation vectors derived from 3-D structures has been reported by Moreau and Turpin.[7] Other workers have developed the idea for use with molecular surfaces (notably Wagener et al.[11]).

While implementing the autocorrelation method for 2-D structures, it became clear that, by analogy with flexible 3-D database searching, if the method were extensible to 3-D "rigid" molecules (as reported by Moreau and Turpin[7]), then perhaps a 3-D "flexible" approach would be viable. Few time-efficient 3-D flexible similarity searching methods exist at the present time (although progress has been reported recently[12]) and so it was felt this idea was worth pursuing.

IMPLEMENTATION

Autocorrelation vectors were created following the work of Moreau and Turpin.[7] The eight atomic properties used were as follows: number of heavy atom connections, van der Waals radii, indicator for saturated/unsaturated, Pauling electronegativity, indicator for carbon/heteroatom, indicator for hydrogen-bond donor (SMARTS definition from Gillet et al.[13]), indicator for hydrogen-bond acceptor (SMARTS definition from Gillet et al.[13]), and indicator for ring/nonring.

With 3-D flexible autocorrelation vectors, there is a need to encode the molecular flexibility somehow. In this work, smoothed bounded distance matrices (SBDMs) were investigated, inspired by earlier applications in flexible 3-D substructure searching[14] and, more recently, in 3-D similarity searching.[12] The SBDMs were created using the Sybyl DIST_GEOMETRY commands.[15] In the work of Moreau and Turpin,[7] the autocorrelation vector for each atomic property comprised 16 elements, representing bond paths of length 0−15 atoms. In this work, 16 elements were also used, this time representing the range from 0.0 to 20.8 Å divided into bins of 1.3 Å. By analogy with the creation of distance keys for flexible 3-D substructure searching, when creating the flexible autocorrelation vector, for a given atom pair IJ, all the distance bins in the range between lowerbound$_{IJ}$ and upperbound$_{IJ}$ have the product property$_I$ × property$_J$ added to them.

The resulting vectors have the same dimension as their 2-D counterparts (126 elements = 8 × 16). Two vectors are compared by computing the square of the Euclidean distance between them—although other metrics may also be worth investigating. These comparisons are very rapid, giving rise to short search times, even for large databases.

[*] Corresponding author phone: +44-(0)1279−645611; e-mail: david.clark@argentadiscovery.com.

[†] The University of Sheffield.

[‡] Current address: Bio-Layer Pty. Limited, Unit 4, 26 Brandl Street, Brisbane Technology Park, Eight Mile Plains, Queensland QLD 4113, Australia.

[§] Argenta Discovery Ltd.

10.1021/ci0503863 CCC: $33.50 © 2006 American Chemical Society
Published on Web 01/04/2006

Table 1. Summary of the 11 Activity Classes and the Database Used in This Study[a]

class no.	identifier	class description	actives	class similarity mean (sd)	rotatable bonds mean (sd)	molecular weight mean (sd)	H-bond acceptors mean (sd)	H-bond donors mean (sd)	surface area, Å^2 mean (sd)
06233	A	5 HT3 antagonist	493	0.143 (0.098)	2.72 (1.25)	314.89 (46.58)	3.10 (1.03)	0.97 (0.75)	298.64 (41.46)
06235	B	5HT1A agonist	530	0.137 (0.098)	6.05 (2.43)	367.00 (77.42)	4.03 (1.87)	0.81 (0.85)	356.45 (72.64)
06245	C	5HT reuptake inhibitor	155	0.155 (0.141)	4.68 (2.37)	356.82 (73.96)	3.33 (1.32)	0.94 (0.72)	339.81 (70.27)
07701	D	D2 antagonist	213	0.15 (0.124)	5.68 (2.02)	401.75 (64.21)	3.94 (1.12)	0.92 (0.71)	380.23 (55.76)
31420	E	renin inhibitor	692	0.32 (0.106)	17.09 (2.54)	663.52 (82.74)	7.17 (1.54)	4.66 (1.13)	662.04 (77.98)
31432	F	angiotensin II AT1 antagonist	889	0.237 (0.105)	8.90 (2.58)	505.92 (88.43)	5.84 (1.51)	1.43 (0.75)	477.21 (76.80)
37110	G	thrombin inhibitor	533	0.183 (0.102)	10.20 (3.22)	490.55 (80.74)	6.00 (1.59)	3.63 (1.47)	478.57 (75.46)
42731	H	substance P antagonist	1002	0.153 (0.084)	8.41 (3.09)	513.50 (104.61)	4.07 (1.52)	1.38 (1.15)	476.76 (96.92)
71523	I	HIV protease inhibitor	632	0.207 (0.104)	12.70 (4.05)	595.53 (118.69)	6.48 (1.96)	3.36 (1.45)	580.43 (118.52)
78331	J	cyclooxygenase inhibitor	589	0.113 (0.083)	4.62 (2.64)	347.70 (72.28)	3.89 (1.26)	1.35 (0.86)	330.65 (68.69)
78374	K	protein kinase C inhibitor	314	0.133 (0.123)	6.67 (4.26)	429.38 (104.02)	5.03 (2.51)	2.33 (1.42)	399.57 (96.35)
MDDR		*(whole data set, 82 153 molecules)*	*(6042)*	*n/a*	6.86 (4.13)	410.48 (121.44)	5.15 (2.48)	1.80 (1.62)	392.50 (114.41)

[a] Statistics were calculated using Pipeline Pilot.[19] Class similarity is expressed as the mean intraclass Tanimoto similarity calculated using ECFP-6 fingerprints.[19]

TEST DATA

All experiments reported here were carried out using version 2001.1 of the MDDR (MDL Drug Data Report) database.[16] A series of filters was applied to the database to remove the following classes of compound: (1) those with syntactically incorrect SMILES or SMILES of length > 200 characters, (2) those consisting of two or more fragments, and (3) those containing unusual elements (i.e., elements other than C, N, O, P, S, F, Cl, Br, and I).

3-D structures were generated using Concord[17] with default parameters. 8710 compounds were excluded by the program leaving 82166 structures. A further 13 structures failed the distance geometry routines leaving a final searchable 3-D autocorrelation vector database containing representations of 82153 structures. The rate-limiting step in the construction of the autocorrelation vector database was the calculation of the SBDMs (an O(N^3) process). The time required could be of course reduced by splitting the job over multiple CPUs.

Eleven classes of actives were selected,[18] and, using SQL, compound names were prefixed with an identifier to facilitate processing of the results by scripts. The database contains a total of 6042 compounds that are active in a single category, and each of these was used as a query to evaluate its effectiveness in retrieving the other members of the activity class from which it was drawn (Table 1). The approach is designed to provide a novel method of searching databases using available bioactive reference structures.

To assess the effectiveness of this approach it was compared to 2D ECFP-6 fingerprints[19] applied in an analogous manner to identical data. Two fingerprints were compared by calculating the similarity by Tanimoto coefficient. The similarity S between two molecules A and B is given by

$$S_{A,B} = \frac{[\sum_{j=1}^{j=n} x_{jA}x_{jB}]}{[\sum_{j=1}^{j=n}(x_{jA})^2 + \sum_{j=1}^{j=n}(x_{jB})^2 - \sum_{j=1}^{j=n} x_{jA}x_{jB}]}$$

where n is the size of the vectors, i and j are elements of the vectors, and x_{jA} is the value of the jth element in molecule A.

The autocorrelation vector matching program (unoptimized Fortran 77 code) took approximately 118 s for each search against the 82153 structures (i.e., ca. 700 structures/second) on a single processor of a dual 180 MHz R10000 processor SGI machine with 128MB of memory. Search time is independent of the size of the query or database molecule but requires prior generation of the autocorrelation vectors. For the purposes of comparison, the Pipeline Pilot protocol[19] took around 2.7 s on a medium-specification Windows PC to perform each equivalent 2D search with no prior generation of the fingerprint representations.

RESULTS AND DISCUSSION

To estimate the effectiveness of the method in retrieving other members of a class, the results of searching using each active were sorted in order of increasing (squared) Euclidean distance, the “identity” at the head of the list was removed, and rankings were calculated using shell scripts. A Python script was used to count the number of members of that class occurring in the top 1%, 2%, and 5% of the rankings. Enrichment factors were calculated as the number of times this result exceeded that which would be expected from a purely random selection. Thus, a method as effective as random would have an enrichment factor of unity. Results in terms of enrichment factors are summarized in Table 2. The percentage recoveries of the actives in the top 1%, 2%, and 5% of the rankings are summarized in Table 3.

As can be seen, the results from the 3-D autocorrelation vector (ACV3D) method are always better than random but never better than those from the ECFP-6 method (Figure 1). Interestingly, the ACV3D enrichment is almost constant rather than dropping off, which is the normal behavior with larger hitlist percentages. The reasons for this unusual behavior are currently unclear. Of all the classes, the ACV3D method performs best for the renin inhibitors, although the ECFP-6 method outperforms it significantly here too. This class has the highest intraclass similarity of all the activity classes (as measured using the Tanimoto similarity coefficient applied to ECFP-6 fingerprints), so this result is perhaps not surprising. Both methods struggle with the cyclooxygenase inhibitors, and this class has the lowest intraclass similarity.

Table 2. Summary of Results Obtained by Searching the MDDR Database with 6042 Active Molecules Drawn from 11 Activity Classes Using 3D Autocorrelation Vector Representations and ECFP-6 Fingerprints[a]

	ACV3D enrichment factor level, mean (s.d.)			ECFP-6 enrichment factor level, mean (s.d.)		
class description	1%	2%	5%	1%	2%	5%
5 HT3 antagonist	7.11 (4.49)	5.74 (3.28)	6.82 (2.91)	17.86 (9.19)	10.82 (5.03)	7.28 (2.26)
5HT1A agonist	5.41 (4.18)	4.12 (2.86)	4.39 (2.31)	14.92 (8.51)	9.32 (4.77)	6.28 (2.12)
5HT reuptake inhibitor	6.46 (4.27)	4.66 (2.63)	4.71 (1.77)	17.28 (12.68)	11.07 (7.68)	7.44 (3.68)
D2 antagonist	4.56 (3.87)	3.67 (2.90)	4.60 (2.60)	9.86 (4.35)	6.35 (2.43)	6.05 (1.75)
renin inhibitor	17.21 (8.00)	13.59 (6.06)	13.10 (5.39)	82.74 (15.51)	45.72 (6.84)	19.35 (1.79)
angiotensin II AT1 antagonist	5.68 (2.48)	4.56 (1.75)	5.12 (1.64)	34.22 (7.43)	29.52 (7.71)	17.32 (2.85)
thrombin inhibitor	4.86 (2.79)	4.07 (2.10)	5.17 (2.07)	25.83 (13.74)	16.41 (8.01)	11.11 (3.69)
substance P antagonist	3.65 (2.31)	3.17 (1.91)	4.68 (2.37)	16.05 (8.44)	10.35 (5.06)	7.66 (2.44)
HIV protease inhibitor	8.31 (5.13)	6.77 (4.12)	7.37 (4.10)	30.23 (16.45)	20.21 (9.56)	13.72 (3.62)
cyclooxygenase inhibitor	2.31 (1.22)	2.09 (0.93)	3.44 (1.20)	6.69 (3.69)	4.32 (2.17)	3.72 (1.21)
protein kinase C inhibitor	2.89 (2.62)	2.08 (1.70)	2.47 (1.18)	17.56 (14.14)	10.08 (7.75)	5.55 (3.45)

[a] Mean and standard deviation of enrichment factors calculated at the 1%, 2%, and 5% levels.

Table 3. Summary of Results Obtained by Searching the MDDR Database with 6042 Active Molecules Drawn from 11 Activity Classes Using 3D Autocorrelation Vector Representations and ECFP-6 Fingerprints[a]

	ACV3D percentage of actives recovered in top-*n%*, mean (s.d.)			ECFP-6 percentage of actives recovered in top-*n%*, mean (s.d.)		
class description	1%	2%	5%	1%	2%	5%
5 HT3 antagonist	7.11 (4.49)	11.47 (6.56)	21.19 (10.39)	17.85 (9.18)	21.65 (10.06)	28.50 (10.87)
5HT1A agonist	5.40 (4.18)	8.23 (5.71)	14.43 (9.00)	14.91 (8.51)	18.65 (9.54)	24.57 (10.32)
5HT reuptake inhibitor	6.46 (4.27)	9.33 (5.26)	15.81 (7.01)	17.26 (12.67)	22.14 (15.37)	29.53 (17.79)
D2 antagonist	4.56 (3.87)	7.34 (5.79)	13.93 (9.18)	9.85 (4.35)	12.71 (4.87)	19.90 (6.52)
renin inhibitor	17.20 (8.00)	27.18 (12.13)	48.69 (21.74)	82.69 (15.50)	91.44 (13.69)	95.35 (11.30)
angiotensin II AT1 antagonist	5.67 (2.47)	9.11 (3.49)	16.27 (5.52)	34.20 (7.42)	59.03 (15.42)	78.19 (17.03)
thrombin inhibitor	4.86 (2.79)	8.14 (4.19)	15.77 (6.90)	25.81 (13.73)	32.81 (16.02)	44.44 (18.10)
substance P antagonist	3.65 (2.31)	6.34 (3.82)	13.36 (7.41)	16.04 (8.44)	20.69 (10.12)	28.75 (11.66)
HIV protease inhibitor	8.30 (5.12)	13.54 (8.25)	25.08 (15.28)	30.22 (16.44)	40.42 (19.12)	56.32 (20.13)
cyclooxygenase inhibitor	2.31 (1.22)	4.17 (1.86)	9.22 (3.58)	6.68 (3.69)	8.64 (4.33)	12.86 (5.16)
protein kinase C inhibitor	2.89 (2.62)	4.16 (3.41)	7.59 (4.89)	17.55 (14.13)	20.16 (15.50)	23.95 (16.84)

[a] Mean and standard deviation of percentage recovery of actives calculated at the 1%, 2%, and 5% levels.

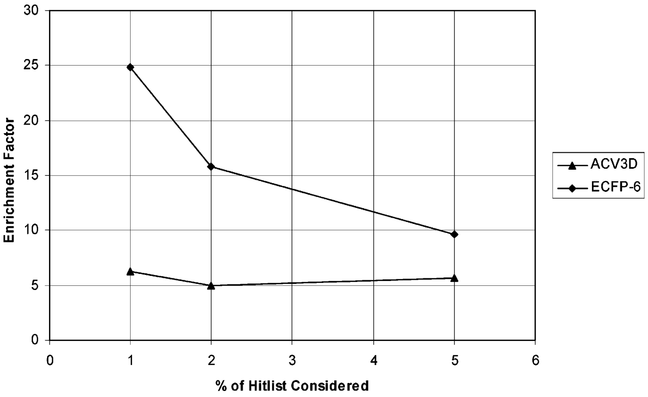


Figure 1. Mean enrichment factors over all 11 classes for ECFP-6 and ACV3D.

The angiotensin II AT1 antagonists are a class of active for which the ECFP-6 method performs very well, while the ACV3D method performs relatively poorly. In terms purely of enrichment factors, there is no advantage to be gained by using the ACV3D rather than ECFP-6 method.

Although the enrichment factors from the ACV3D method are not comparable to those from ECFP-6, it was hoped that the former method might retrieve compounds not identified by the latter within a given percentage of a hitlist. To compare the two sets of results we selected one active from each class for detailed examination, chosen on the basis that its performance was as close as possible to the average for that class. Though we accept that the distributions within each class may be highly skewed it seemed a reasonable basis on which to choose a class representative with a medium-sized hitlist. A comparison of the similarity of the top 100 actives from the two hitlists produced by compound 151019 reveals that the ECFP-6 hitlist has a mean intrahitlist similarity of 0.289 (s.d. = 0.147) compared to that given by the ACV3D method, 0.178 (s.d. = 0.165). This indicates that the actives obtained by the latter method are a more heterogeneous set. The top 10 actives retrieved by each method are presented in Tables 5 and 6. Table 5 shows that all but one of the hits obtained from the ECFP-6 fingerprints contain a tricyclic ring system comprising an indole fused to a six-membered lactam, whereas the same system is only present in three of the structures shown in Table 6. This observation is in keeping with the lower intrahitlist similarity obtained using ACV-3D.

In each case, the ECFP-6 fingerprints recalled more actives at the 1% retrieval level, but the ACV3D method retrieved actives not highly ranked by the ECFP-6 method (Table 4). The worst case was the large and flexible renin inhibitors (only 5 of the 119 actives retrieved by ACV3D were not included in the 606 member ECFP-6 hitlist); conversely, the best case was the D2 antagonist class where 9 of 10 compounds retrieved by ACV3D were not included in the 19-member ECFP-6 list. Thus, despite their poorer enrichment factors, ACV3D searches may still prove to be a useful adjunct to other similarity searching methods.[1]

Table 4. Comparison of the "Hitlists" Produced by Searching Using 3D Autocorrelation Vector Representations and ECFP-6 Fingerprints

class no	identifier	class description	actives	ACV3D 1% recovery	average performer retrieves ∼	average performer	no. retrieved by ACV3D @ 1%	no. retrieved by ECFP-6 @ 1%	no. unique to ACV3D
06233	A	5 HT3 antagonist	493	7.11	35	A_151019	38	148	17
06235	B	5HT1A agonist	530	5.40	29	B_151387	32	44	17
06245	C	5HT reuptake inhibitor	155	6.46	10	C_157413	10	50	4
07701	D	D2 antagonist	213	4.56	10	D_158269	10	19	9
31420	E	renin inhibitor	692	17.20	119	E_154703	119	606	5
31432	F	angiotensin II AT1 antagonist	889	5.67	50	F_182272	50	350	25
37110	G	thrombin inhibitor	533	4.86	26	G_206625	34	164	18
42731	H	substance P antagonist	1002	3.65	37	H_182504	37	127	18
71523	I	HIV protease inhibitor	632	8.30	52	I_203385	52	336	14
78331	J	cyclooxygenase inhibitor	589	2.31	14	J_149731	14	46	9
78374	K	protein kinase C inhibitor	314	2.89	9	K_164741	9	62	3

Table 5. Top Ten Actives Obtained with a Search Based on Compound 151019 Using ECFP-6 Fingerprints

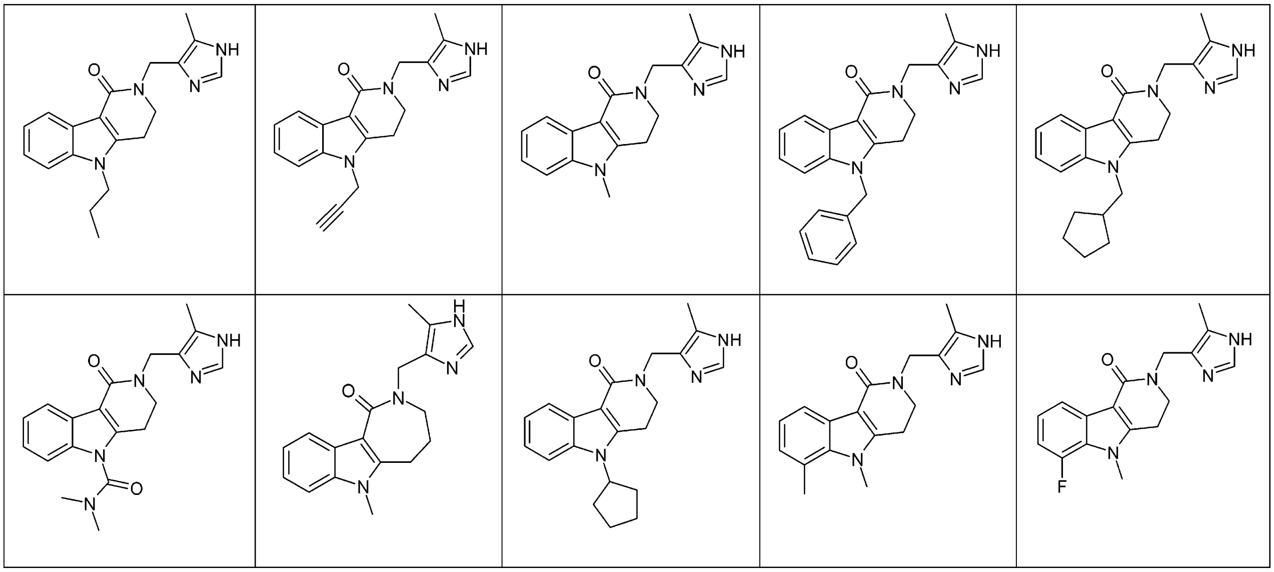
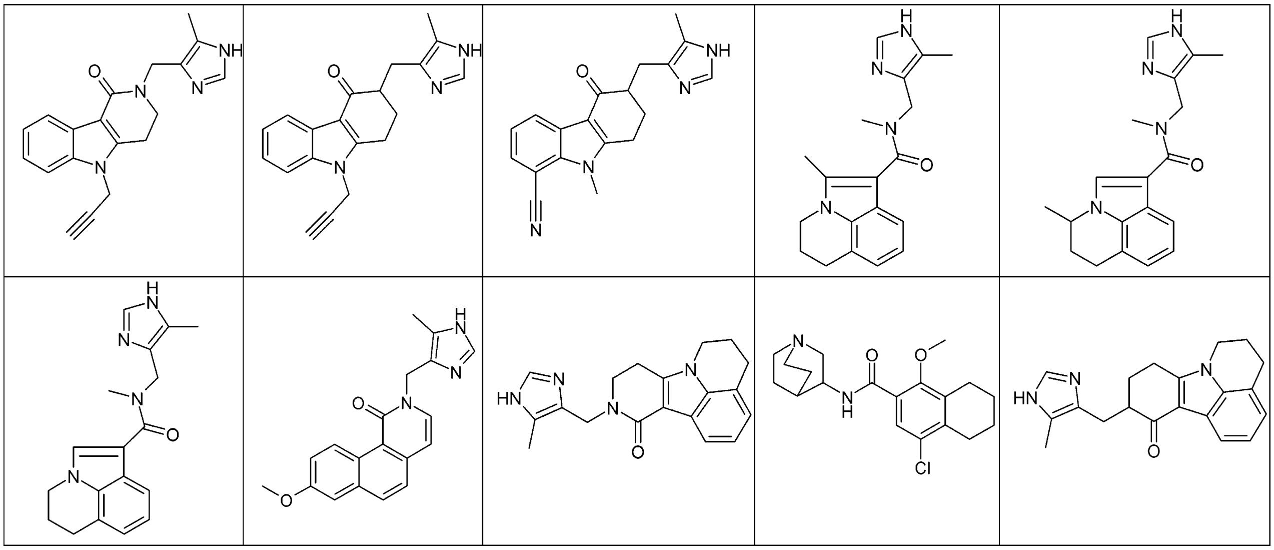


Table 6. Top Ten Actives Obtained with a Search Based on Compound 151019 Using 3D Autocorrelation Vectors



CONCLUSIONS

We have presented an exploratory study of a novel method for flexible 3-D similarity searching based on autocorrelation vectors and smoothed bounded distance matrices. Although the new approach was unable to outperform an existing 2-D similarity searching method in terms of enrichment factors, it did prove able to retrieve different compounds at a given percentage of the hitlist. Thus, in an ideal situation, one would use both methods to give enhanced recall. This is in keeping with received wisdom.[1]

REFERENCES AND NOTES

(1) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discovery Today* **2002**, *7*, 903−911.

(2) Lengauer, T.; Lemmen, C.; Rarey, M.; Zimmermann, M. Novel technologies for virtual screening. *Drug Discovery Today* **2004**, *9*, 27−34.

(3) Willett, P. Searching techniques for databases of two- and three-dimensional chemical structures. *J. Med. Chem*. **2005**, *48*, 4183−4199.

(4) Cramer, R. D.; Jilek, R. J.; Guessregen, S.; Clark, S. J.; Wendt, B.; Clark, R. D. "Lead hopping". Validation of topomer similarity as a superior predictor of similar biological activities. *J. Med. Chem*. **2004**, *47*, 6777−6791.

(5) Rush, T. S., III; Grant, J. A.; Mosyak, L.; Nicholls, A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein−protein interaction. *J. Med. Chem.* **2005**, *48*, 1489−1495.

(6) Moreau, G.; Broto, P. The autocorrelation of a topological structure: a new molecular descriptor. *Nouv. J. Chim*. **1980**, *4*, 359−360.

(7) Moreau, G.; Turpin, C. Use of similarity analysis to reduce large molecular libraries to smaller sets of representative compounds. *Analusis* **1996**, *24*, 17−21.

(8) Moreau, G.; Broto, P.; Fortin, M.; Turpin, C. Computer-conducted screening of molecular structures of potentially anxiolytic substances by means of an autocorrelation technique. *Eur. J. Med. Chem*. **1988**, *23*, 275−281.

(9) Bauknecht, H.; Zell, A.; Bayer, H.; Levi, P.; Wagener, M.; Sadowski, J.; Gasteiger, J. Locating biologically active compounds in medium-sized heterogeneous datasets by topological autocorrelation vectors: dopamine and benzodiazepine agonists. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 1205−1213.

(10) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem*. **1996**, *39*, 3049−3059.

(11) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of molecular surface properties for modeling corticosteroid binding globulin and cytosolic Ah receptor activity by neural networks. *J. Am. Chem. Soc*. **1995**, *117*, 7769−7775.

(12) Raymond, J. W.; Willett, P. Similarity searching in databases of flexible 3D structures using smoothed bounded distance matrices. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 908−916.

(13) Gillet, V. J.; Willett, P.; Bradshaw, J. Identification of biological activity profiles using substructural analysis and genetic algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165−179.

(14) Clark, D. E.; Willett, P.; Kenny, P. W. Pharmacophoric pattern matching in files of three-dimensional chemical structures: Use of bounded distance matrices for the representation and searching of conformationally flexible molecules. *J. Mol. Graphics* **1992**, *10*, 194−204.

(15) Sybyl v7.0. Developed and distributed by Tripos, Inc., 1699 South Hanley Road, St. Louis, MO 63144-2319, U.S.A. http://www.tripos.com

(16) MDL Drug Data Report database version 2001.1. Available from MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577, U.S.A. http://www.mdli.com

(17) Concord v5.1.2. Balducci, R.; McGarity, C. M.; Rusinko, A., III; Skell, J.; Smith, K.; Pearlman, R. S (University of Texas at Austin). Distributed by Tripos, Inc., 1699 S. Hanley Rd., Suite 303, St. Louis, MO 63144-2913, U.S.A. http://www.tripos.com

(18) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Schuffenhauer, A. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177−1185.

(19) Pipeline Pilot (Server version 4.1.0.200). Developed and distributed by SciTegic, 9665 Chesapeake Drive, Suite 401, San Diego, CA 92123-1365 U.S.A. http://www.scitegic.com

CI0503863