Journal of Computer-Aided Molecular Design **18:** 529–536, 2004.
DOI 10.1007/s10822-004-4067-1

Measuring CAMD technique performance: A virtual screening case study in the design of validation experiments

Andrew C. Good[a,*], Mark A. Hermsmeier[b] & S.A. Hindle[c]
[a]*Bristol-Myers Squibb, 5 Research Parkway, Wallingford, CT 06492, USA;* [b]*Bristol-Myers Squibb, P.O. Box 4000, Princeton, NJ 08543, USA;* [c]*BioSolveIT GmbH, An der Ziegelei 75, D-53757 Sankt Augustin, Germany*

Received 30 April 2004; accepted in revised form 8 September 2004

Key words: atom pairs, chemotypes, Daylight, fingerprints, Ftrees, pharmacophores, validation, virtual screening

Summary

The dynamic nature and comparatively young age of computational chemistry is such that novel algorithms continue to be developed at a rapid pace. Such efforts are often wrought at the expense of extensive experimental validations of said techniques, preventing a deeper understanding of their potential utility and limitations. Here we address this issue for ligand-based virtual screening descriptors through design of validation experiments that better reflect the aims of real world application. Applying the newly defined chemotype enrichment approach, a variety of two- and three-dimensional (2D/3D) similarity descriptors have been compared extensively across data sets from four diverse target types. The inhibitors within said data sets contain molecules exhibiting a wide array of substructure functionality, size and flexibility, permitting descriptor comparison in myriad settings. Relative descriptor performance under these conditions is examined, including results obtained using more typical virtual screening validation experiments. Guidelines for optimal application of said descriptors are also discussed in the context of the results obtained, as is the potential utility of fingerprint filtering.

Introduction

Ready access to ever increasing CPU and biological data drives a continuous evolution in novel algorithm development within computational chemistry. While such advances are often interesting and potentially useful, analyses of the data and methods applied in their validation frequently highlight less than optimal experimental design. The reasons for this are understandable, since data collection and validation experiments are both time consuming and generally less intellectually attractive than algorithm development. Consequently there is a natural tendency to implement the validation techniques used in earlier papers, since they set the standard by which new articles are generally judged.

The issue of data quality and applicability is recognized as a problem within the comparatively mature discipline of QSAR, where a significant effort has been undertaken in the analysis of techniques and data used in validation experiments. A search of the literature quickly highlights the need for such studies [1–7], given the inherent limitations of long standing techniques such as cross-validation. The same cannot be said for most other areas of computational chemistry, despite the fact that many of the same problems permeate

*To whom correspondence should be addressed. E-mail: andrew.good@bms.com

the field. A pertinent example of this can be found with ligand-based virtual screening descriptor design. While still a dynamic area of technique development, it still represents a mature discipline within computational chemistry. Despite this, the same basic flavor of validation experiment has propagated the literature for many years [8–12]. Said calculations generally take a small (often one) random selection of template compounds. These are then used to search a database seeded with compounds known to be active at the same target. Generally only one or two other (often related) descriptors are compared, and limited effort is expended analyzing the nature (size, shape, flexibility, etc.) of the compounds and their associated target to gain insight on how said properties effect performance. Consequently, while the descriptors presented are often both novel and potentially useful, their associated validation experiments provide modest insight regarding optimal application (one of the authors has also been equally culpable of such experimental design [9]).

If technique validation relevance is to be maximized, it needs to reflect the challenge a given methodology is ultimately attempting to address. In the case of virtual screening this is relatively straightforward, since descriptors are most often utilized to discover novel structural lead chemotypes that can be leveraged internally to create in-house intellectual property. To measure chemotype enrichment in a comprehensive manner, there are four primary issues that must be considered. First, experiments that mitigate the active analogue bias that permeates many data sets are required. This is accomplished through the assignment of each compound in the active data set to it own relevant substructure class [13]. Next, enrichment needs to be measured across multiple targets of different classes to determine performance in wholly unrelated regions of biological space. In addition, a comprehensive exploration of descriptor comparison space is required to remove artifacts produced by random selection of template molecules. Finally, results comparisons with other widely used descriptors are necessary to permit insight regarding relative performance. In the experiments below we address each of these topics in turn, detailing the results and insights such modifications produce.

Methods

The following experimental design was applied in order to optimize virtual screening validation protocols:

(1) Active compound data sets were collected for four unrelated targets previously explored

Table 1. Target active data set composition and chemotype (as defined by project chemists) summary.

Target type + data set composition	Primary chemotype definition drivers	Target chemotype property comparisons
GPCR (Melatonin receptor) 106 actives 20 chemotypes	Indole core mimic	Smallest and most rigid molecules Highest Daylight inter-chemotype similarity Pharmcount[a] = 431
Fatty acid binding protein (AP2) 28 actives 8 chemotypes	Core linking acid moiety to remaining substituents	Relatively rigid compounds, many with a high degree of branching Some compounds exhibit collapsed binding modes Pharmcount[a] = 1057
Kinase (CDK2) 95 actives 15 chemotypes	Moiety mimicking adenine/ main core of molecules	Widest variety of chemotypes; lowest Daylight inter-chemotype Pharmcount[a] = 1932
Serine protease (FXa) 38 actives 19 chemotypes	P1 substituent/P1-P4 linker substituent	Pharmacophorically most complex data set Extended binding mode Pharmcount[a] = 2639

[a]Pharmcount equals the average number of 4 center pharmacophores present in the fingerprints of each target data set. This is provided to give the reader a feel for the molecular complexity of said data.

in-house: Melatonin receptor [14, 15] (GPCR), AP2 [16, 17] (fatty acid binding protein), CDK2 [18] (kinase) and Factor Xa [19] (FXa serine protease). Table 1 provides further details regarding data set size and properties.

(2) For each target, the data set was analyzed by the project chemists (medicinal and computational) and divided into chemotypes based on the variation of key substructure features. Said features were differentiated based upon the structural novelty of known key binding motifs and molecule cores (see Table 1).

(3) A variety of 2D descriptors were selected for comparison in order to provide significant insight into relative performance. Molecular connectivity descriptors employed included Daylight fingerprints [20], Ftrees [21, 22], plus in-house implementations of atom pairs [23] and binding (property) pairs [10]. The Tanimoto similarity coefficient [24] was employed to calculate similarity for each of these measures except for Ftrees. The Feature Tree descriptor [21] as implemented in the FTrees software [22] describes molecules using a tree representation, rather than a more familiar bitstring or vector. The features of the chemical building blocks (steric properties such as atomic volume and ring closures, plus a chemical property profile) of the molecule are stored at the nodes of the tree, while the edges between the nodes retain the topological relationship of these bulding blocks. The similarity between two Feature Trees is defined as the score for the best possible alignment or superposition of the two trees. Here, Feature Trees were compared using the dynamic match search algorithm with default parameters (made available to us in a pre-release version of FTrees version 2.0). This algorithm calculates alignments between descriptor pairs in a similar manner to algorithms designed for aligning protein sequences.

(4) In addition a number of 3D distance-based descriptors were also analyzed. These include in-house implementations of binary two, three and four center pharmacophore fingerprints (2/3/4 binary) [25]. Two and three center pharmacophore fingerprints were derived via simple deconstruction of the parent four center data. The similarity equation used varied depending on the number of centers in an attempt to reduce sensitivity due to pharmacophoric promiscuity as detailed in Equation 1 [25]:

$$\frac{O}{w(T-O)+w(P-O)+O} \tag{1}$$

where O = number of pharmacophores in common, T = total number of pharmacophores found in the template molecule, P = total number of pharmacophores found in probe molecule and w = weighting factor, which varies as follows: 2 centers = 1.0, 3 centers = 0.5, 4 centers = 0.11. Three and four center normalized and filtered 'trend vector' pharmacophore descriptors (3/4 count) detailed in the previous article [26] were also applied. Similarity for these descriptors was again calculated based simply on the sum of pharmacophore overlap (numerator term in Equation 1). To test the effect of the inclusion of conformational flexibility, these simple overlap calculations were repeated using fingerprints derived solely from the Concord structure.

(5) Enrichment calculations were designed to be as indicative as possible of a descriptor's ability to pick out novel actives from a hit list [26]. For each target the similarity of each molecule was calculated in turn versus all other molecules in the active data set together with a 9524 compound noise database selected to mimic inactive molecules. These compounds were selected from a 10,000 compound random selection of our in-house database with a conformational flexibility distribution that mirrored the parent data set. All molecules with more than 15 flexible bonds and Concord [27] build failures were removed. The enrichment rate was incremented each time a molecule from a chemotype not in the template chemotype was found. Each chemotype was only allowed to increment enrichment once, based on the assumption that all remaining molecules in the chemotype would subsequently be found using substructure searches. The resulting enrichment values for each molecule were averaged across all other molecules in the same chemotype. This provides an average enrichment value for that chemotype as a template. Finally the average of all the resulting chemotype enrichments was calculated to produce an overall performance measure.

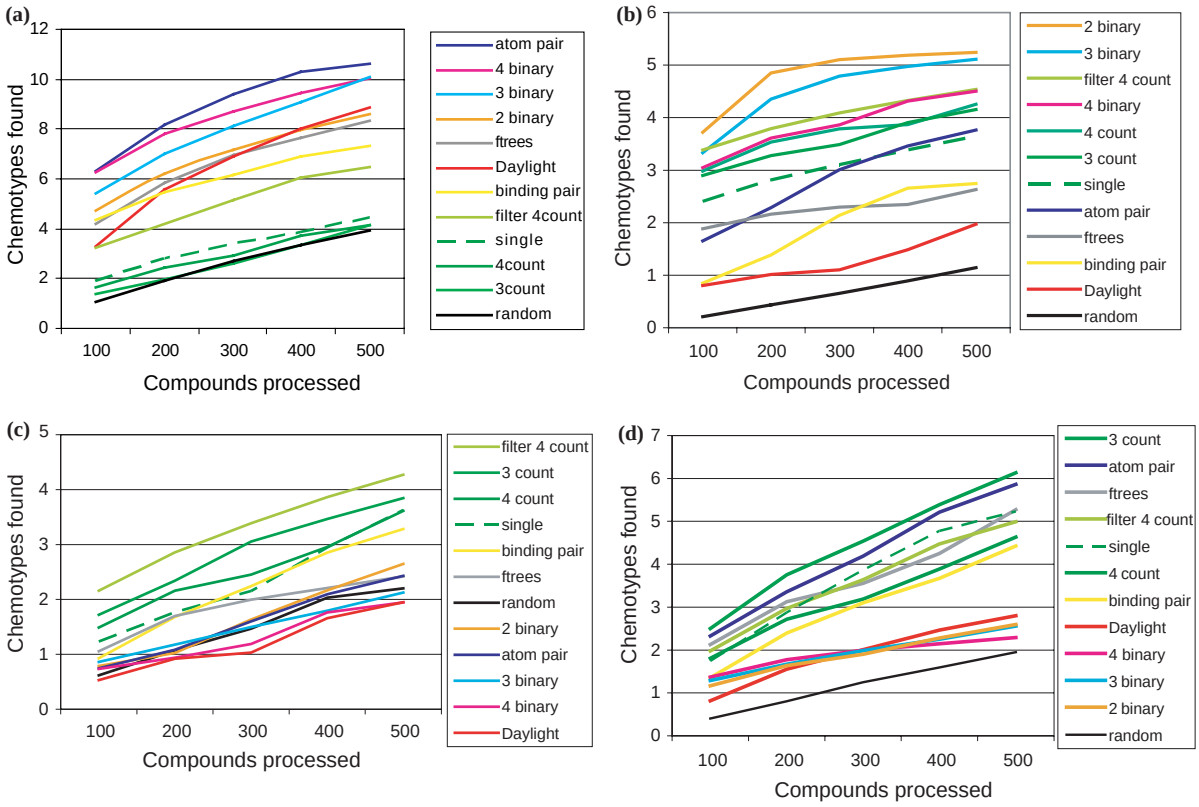

Figure 1. Primary chemotype enrichment data. (a) Melatonin (20 chemotypes), (b) AP2 (8 chemotypes), (c) CDK2 (15 chemotypes), (d) FXa (19 chemotypes). Descriptors compared are atom pairs (atom pair), binding property pairs (binding pair), Daylight fingerprints (Daylight), Ftrees (ftrees), 2/3/4 center binary pharmacophore fingerprints (2/3/4 binary), 3/4 center normalized pharmacophore histograms (3/4 count) and 4 center normalized and filtered pharmacophores (filter 4 count), pharmacophore overlap for 4 center pharmacophore fingerprint derived solely from Concord conformation (single). Random enrichment is also provided for comparison.

Results

The following experiments were undertaken:

(1) Chemotype enrichment calculations as defined in Methods section (5) were executed for all the targets data sets detailed in Table 1. The results of these calculations are shown in Figure 1a–d. Randomized enrichment was also calculated by randomizing the rankings 10 times for each template and averaging the resulting chemotype enrichment scores. These results were then processed as defined in Methods section (5) to determine randomized chemotype enrichment.

(2) Three additional control tests were undertaken to highlight the effects of validation experiment design decisions.

(i) In the first study, the effect of changing the noise data set was tested. For the second noise set, all compounds with a heavy atom count inside the range seen in the CDK2 actives were selected from our in-house database. About 10,097 compounds were then randomly selected from this set, and all compounds successfully converted by Concord were kept (a total of 9993 molecules, with zero overlap relative to the first noise data set). The CDK2 enrichment studies were then repeated using the new noise data. Results are shown in Figure 2a. This test was also repeated using the same additional noise set for the AP2 data. Figure 2b contains the results of this study.

(ii) For the second study, chemotype enrichment was replaced by the straight average hit rate enrichment for the CDK2 studies. This was done to highlight the potential

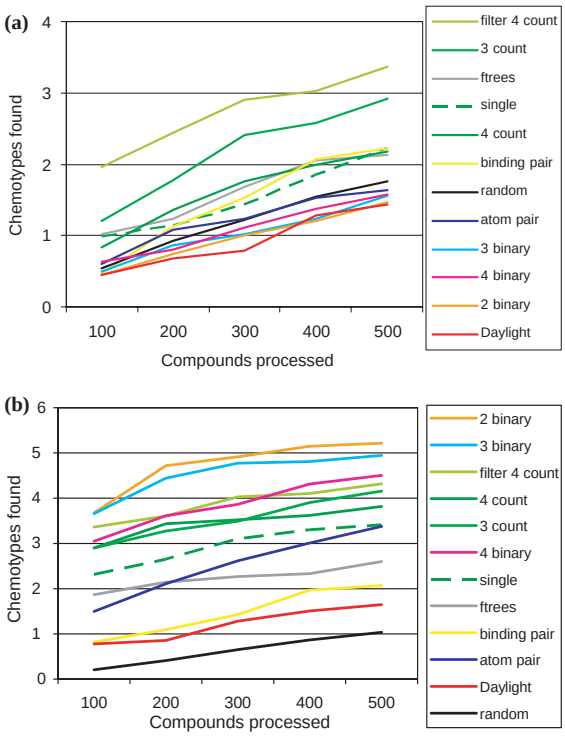

Figure 2. Chemotype enrichment data in conjunction with noise data set 2. (a) CDK2, (b) AP2.

differences that can be observed when active analogue bias is ignored. The CDK2 data set was chosen since >70% of the actives were contained in just three of the chemotype classes, providing a highly biased set. Average hit rate enrichment was determined by averaging the headline enrichment rates observed when each active data set molecule in turn was used as the template, with no account being taken of chemotype assignment. Results are shown in Figure 3.

(iii) In the final experiment, the chemotype enrichment data of each AP2 chemotype was analyzed in turn. This was undertaken to determine if selection of a particular class of template could produce enrichment behavior that varied significantly from the data set as a whole. This was done to highlight the potential problems that can occur on random template selection. The results for one such chemotype are shown in Figure 4.

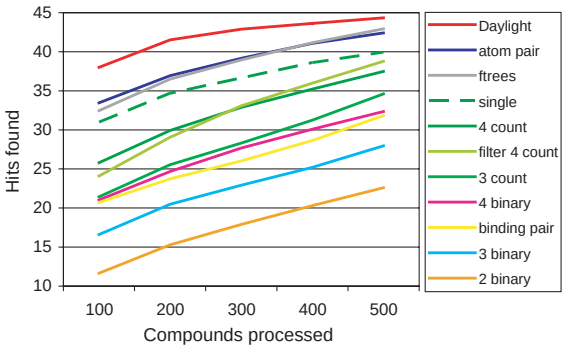

Figure 3. CDK2 headline hit rate enrichment data (chemotype differentiation ignored).

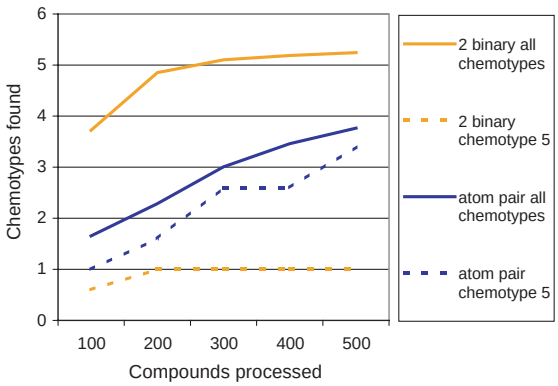

Figure 4. Overall chemotype enrichments versus enrichments for chemotype 5 templates alone, comparing 2 center binary pharmacophore and atom pair performance.

Discussion

Overall results shown in Figure 1 highlight that, as has been shown previously [28, 29], there is no single virtual screening solution that continually delivers superior enrichment values. Nevertheless, there are some interesting features within the data that may prove useful in methods selection. 2D descriptors, as one should intuitively expect, perform best when chemotypes exhibit a higher topological similarity and for systems with extended binding modes. Of the 2D descriptors applied, atom pairs and to a slightly lesser extent Ftrees perform well in certain cases (melatonin and FXa – Figures 1a and d). Daylight fingerprints are clearly the worst of the 2D descriptors at finding novel chemotypes. This is not surprising, given the high degree of substructure encoding present in the fingerprint. In general such substructure biased

descriptors should be avoided for screens where scaffold hopping is key.

For these studies 3D descriptors perform best with targets whose inhibitors exhibit a high degree of branching and collapsed binding modes (of hydrophobic moieties in close proximity) (AP2 – Figure 1b). As a consequence the resulting interatomic distances likely bear little correlation with bond count, a correlation that is intrinsic to most 2D descriptors. Binary pharmacophore descriptors provide good enrichment with active moieties containing less flexibility and functionality (as judged by molecule size and average pharmacophore count, see Table 1 – melatonin and AP2). As templates become more promiscuous, the inherent descriptor noise begins to mask signal (CDK2 and FXa – Figure 1c and d). Further, there is little evidence that the use of 4 center pharmacophores provides any real advantage relative to 2 or 3 center descriptors. This is likely due to the fact that the improved descriptive power of pharmacophore spatial definitions with higher complexity is tempered by increasing sensitivity to small changes in molecule makeup. Pharmacophore count per conformer varies as $n(n-1)/2!$, $n(n-1)(n-2)/3!$ and $n(n-1)(n-2)(n-3)/4!$ times the number of pharmacophore centers for 2, 3 and 4 center pharmacophore descriptors, respectively. As a result small molecular changes can result in large changes in similarity as center count is increased. In addition, as compounds become more complex, the inability of binary fingerprints to deal with saturation effects and weight relative pharmacophore importance begin to weigh on performance. This can be seen from the steadily improving relative enrichments of normalized pharmacophore descriptors as template complexity increases (CDK2 and FXa – Figures 1c and 1d). Indeed, for CDK2 none of the binary pharmacophore descriptors show a signal significantly in excess of that seen in random analyses. Normalization is not a universal panacea, however, as can be seen for melatonin, where the results are little better than random. There is some suggestion in the literature that the binding mode of melatonin antagonists is somewhat collapsed in nature [15], and it may be that normalization is removing the collapsed binding mode signal. Comparison of multi-conformer normalized fingerprints with those generated from a single conformer yields interesting results. Visual inspection of the enrichment data suggests only limited correlation between chemotype enrichments using the two techniques. Nevertheless, while using a single Concord conformation in descriptor generation degrades performance a little overall relative to the inclusion of conformational flexibility, the differences are small, and one could actually argue a marginal improvement for FXa. The FXa results make sense from the perspective that both bioactive and Concord conformations are typically in an extended binding mode. The relatively small changes seen in overall performance for the remainder eloquently highlight the signal to noise trade-off inherent in multiple conformation fingerprint generation, when only one of said conformers is in fact bioactive. This is emphasized by the most general trend seen in the data, the fact that filtering low signal pharmacophores [26] from normalized descriptors improves performance in every case tested. This result highlights the potential utility of descriptor noise removal, while providing further testimony regarding the noise buried within ensemble pharmacophore fingerprints. In essence, while the rapid analysis permitted by such generic fingerprint comparisons is attractive and useful, it is still no substitute for the pursuit of a focused set of potential bioactive conformers/pharmacophores when sufficient SAR data is available.

Overall it is clear that atom pairs, Ftrees, and pharmacophore-based descriptors are all capable of delivering significant novel lead enrichments given the right target type and data set. Descriptor selection is thus best made in the light of the scaffold hop size required, and the nature of the active molecule templates and the binding modes they are likely to adopt. Further study of how best to blend descriptors into a form of consensus score may be useful, though achieving a universally optimal consensus score is likely to prove as elusive as creating a universal descriptor. Based on the enrichment performance improvements seen in these studies, a more fruitful avenue to descriptor optimization would likely come from additional efforts in noise filtering.

Analyses of the validation test modifications highlight the potential effects of experimental design choices. The use of a different noise data set with CDK2, while having some effect on absolute enrichment, does little to change relative enrichment rankings (comparing Figures 2a and 1c).

Given that many of the results are little better than random for CDK2, however, it was decided to repeat the calculations using the AP2 data set. Again relative rankings are left virtually unaltered (compare Figures 1b and 2b). Further assessment of noise set selection would be required to determine whether this result can be generalized. Nevertheless, these results suggest that the technique used for noise selection is not the most crucial of design choices in the case of ligand-based virtual screening. Figure 3 clearly illustrates what can happen when headline hit rate is used to measure enrichment using a data set with a chemotype distribution skewed toward certain analogue sets, as is the case for CDK2 (where $>70\%$ of actives fall into a single chemotype). Daylight and atom pair descriptors, both of which perform poorly in terms of chemotype enrichment (Figure 1c), now move to the top of the enrichment rankings. The results shown in Figure 4 provide a nice example of what can happen when arbitrary template selections are made for enrichment calculations. For AP2 the overall average enrichment values clearly favor 2 center binary fingerprints over atoms pairs. However, if only molecules from chemotype 5 are considered, the relative enrichments are reversed. Only by selecting all active data set molecules in turn as templates can such bias be minimized.

Conclusions

The studies detailed in this article highlight a more rigorous enrichment validation strategy based on scaffold hopping for comparison of ligand-based virtual screening approaches. Contrasting this approach with the methods more typically applied for a number of different data sets clearly highlights the potential advantages for results interpretation and descriptor insight. The features used to define a key molecular scaffold are always open to interpretation. Nevertheless, the resultant clustering of active data sets is highly compatible with approaches applied in hit analysis post screening, virtual or otherwise. It is therefore envisaged that others engaged in the development of ligand-based virtual screening techniques will begin to take up variants of the strategies detailed here. Further, it is hoped that the work will stimulate the development of similar application oriented validation experiments in other areas of computational chemistry.

The time consuming nature of data set collection and analyses led the authors to select in-house data sets for study. This provided data that was both readily available and for which expertise existed to quickly and objectively divide the active compounds into relevant chemotypes. Efforts to undertake the arduous but necessary task of repeating such analyses on more publicly accessible data, for example using the WDI [30], would greatly expand the general applicability of the technique. The application of a program capable of dividing data sets automatically by common substructure would greatly simplify such a project. The resultant compound sets could then become the standard for subsequent validations of new virtual screening approaches.

References

1. Wold, S., Quant. Struct.-Act. Relat., 10 (1991) 191.
2. Eriksson, L., Johansson, E. and Wold, S., In Chen, F. and Schuurmann G. (Eds.), Quantitative Structure-Activity Relationships in Environmental Sciences-VII, Proceedings of QSAR 96, Elsinore, Denmark, June 24–28, 1996. SETAC, 1997, pp. 381–397.
3. Giuliani, A. and Benigni, R., In van de Waterbeemd, H., Testa, B. and Folkers, G. (Eds.), Computer-Assisted Lead Finding and Optimization: Current Tools for Medicinal Chemistry, Proceedings of the 11th European QSAR Symposium, VHCA & Wiley-VCH, Weinheim, Germany, 1997, pp. 51–63.
4. Gaudio, A.C. and Zandonade, E., Quim. Nova, 24 (2001) 658.
5. Kubinyi, H., Quant. Struct.-Act. Relat., 21 (2002) 348.
6. Golbraikh, A. and Tropsha, A., J. Comput.-Aided Mol. Des., 16 (2002) 357.
7. Baumann, K., Trends Anal. Chem., 22 (2003) 395.
8. Nilakantan, R., Bauman, N. and Venkataraghavan, R., J. Chem. Inf. Comput. Sci., 33 (1993) 79.
9. Good, A.C., Ewing, T.J.A., Gschwend, D.A. and Kuntz, I., J. Comput.-Aided Mol. Des., 9 (1995) 1.
10. Kearsley, S.K., Sallamack, S., Fluder, E.M., Andose, J.D., Mosley, R.T. and Sheridan, R.P., J. Chem. Inf. Comput. Sci., 36 (1996) 118.
11. Raymond, J.W. and Willett P., J. Comput.-Aided Mol. Des., 16 (2002) 59.
12. Putta, S., Lemmen, C., Beroza, P. and Greene, J., J. Chem. Inf. Comput. Sci., 42 (2002) 1230.
13. Good, A.C., Cheney, D.L., Sitkoff, D.F., Tokarski, J.S., Stouch, T.R., Bassolino, D.A., Krystek, S.R., Li, Y. and Mason, J.S., J. Mol. Graph. Mod., 22 (2003) 31.
14. Witt-Enderby, P.A. and Li, P.-K.,Vitam. Horm., 58 (2000) 321.

15. Harris, P.W.R., Hugel, H.M. and Nurlawis, F., Mol. Simul., 28 (2002) 889.
16. Chilmonczyk, Z., Siluk, D. and Kaliszan, R., Exp. Opin. Ther. Pat., 11 (2001) 1301.
17. Robl, J.A., Sulsky, R.B. and Magnin, D.R., Heterocyclylbiphenyl AP2 inhibitors. WO 2000059506 PCT Int. Appl. (2000).
18. Wadler, S., Drug Resist. Updates, 4 (2001) 347.
19. Walenga, J.M., Jeske, W.P., Hoppensteadt, D. and Fareed, J., Curr. Opin. Invest. Drugs, 4 (2003) 272.
20. Daylight fingerprints are produced using the Daylight Toolkit, part of the software suite from Daylight Chemical Information Systems: www.daylight.com.
21. Rarey, M. and Dixon, J.S., J. Comput.-Aided Mol. Des., 12 (1998) 471.
22. Ftrees is part of the software suite from BioSolveIT: www.biosolveit.de.
23. Cahart, R.E., Smith, D.H. and Ventkataraghavan, R., J. Chem. Inf. Comput. Sci., 25 (1985) 64.
24. Ellis, D., Furner-Hines, J. and Willett, P., Perspect. Inf. Manag., 3 (1993) 128.
25. Mason, J.S., Morize, I., Menard, P.R., Cheney, D.L., Hulme, C. and Labaudiniere, R.F., J. Med. Chem., 42 (1999) 3251.
26. Good, A.C., Cho, S.-J. and Mason, J.S., J. Comput.-Aided. Mol. Des., 18 (2004) 523 (this issue).
27. Concord 3D structure builder, distributed by Tripos and Optive Research Inc.: www.tripos.com, www.optive.com.
28. Pickett S.D., McLay, I.M. and Clark, D.E., J. Chem. Inf. Comput. Sci., 40 (2000) 263.
29. Sheridan, R.P. and Kearsley, S.K., Drug Discov. Today, 7 (2002) 903.
30. World Drug Index, distributed by Derwent Publications Ltd.: www.derwent.com