J Mol Model (2001) 7:125–131
DOI 10.1007/s008940100019

ORIGINAL PAPER

Ling Xue · Jeffrey W. Godden · Florence L. Stahura
Jürgen Bajorath

A dual fingerprint-based metric for the design of focused compound libraries and analogs

Received: 13 September 2000 / Accepted: 27 February 2001 / Published online: 19 May 2001

Abstract A computational metric is introduced for the design of combinatorial libraries focused on small molecules with specific activity (e.g., enzyme inhibitors). The method follows a product-based design strategy and uses combinations of two binary molecular fingerprints to create chemical diversity around selected compounds and/or core structures. In the first step, compounds are sampled that are distinct from template molecules but likely to share similar biological activity. In the second step, designed compounds are accepted if they are not too similar to each other, as assessed by calculation of fingerprint overlap. Thus, it is possible to balance molecular "similarity" and "diversity" and control the degree of chemical diversity created in the vicinity of selected template molecules. In essence, the method aims to generate diverse arrays of compounds with a high probability of having activity similar to starting molecule(s) and is therefore well suited for the design of target-focused libraries or series of analogs. As an example, the method is applied to focus libraries on known protein kinase inhibitors.

Keywords Chemical diversity · Diversity metric · Focused libraries · Molecular fingerprints · Library design

Introduction

Computational design and analysis of compound libraries, [1, 2, 3, 4, 5, 6] have become integral parts of combinatorial chemistry programs. [7, 8, 9, 10, 11] A variety of algorithms and alternative strategies for diversity-oriented library design have been introduced. [12, 13, 14, 15, 16] In addition to chemically diverse libraries, which are typically used for screening against multiple targets, significant effort is being spent to generate more specialized libraries for drug discovery. These include chemical libraries enriched with compounds having drug-like properties, [17, 18, 19, 20, 21, 22] and libraries focused on specific biological targets, activities [23, 24, 25, 26], or therapeutic applications. [27] Approaches to the generation of target-focused libraries often involve a combination of structure-based design elements and combinatorial chemistry [28, 29, 30, 31], provided three-dimensional structures of therapeutically relevant targets are available. Alternatively, target focus can be achieved by concentrating on small molecules that display or mediate a specific biological activity, e.g., substrates, cofactors, or known inhibitors. [23, 24, 25, 26]

The design of combinatorial libraries is, in general, either reaction- or product-based. [32, 33, 34] In reaction-based design, encoded chemical transformations are applied to pools of reactants, the selection of which becomes the primary determinant of diversity in the resulting library. By contrast, in product-based design, selected molecular building blocks, often called frameworks, [35] or scaffolds [36] are combined with R-groups at prespecified attachment points. In this case, computational metrics are applied to sample diverse compounds from the vast chemical space defined by possible products. However, sampling of compounds is constrained by the choice of templates or scaffolds and the usually limited number of selected points of diversity. The relative performance of these alternative approaches depends, in addition to computational parameters, on the nature of their application. [32, 34] For example, reaction-based design is an effective approach for generating large and chemically diverse libraries required for screening, [2, 4] whereas a product-based strategy is particularly attractive when selected core structures or molecular scaffolds provide starting points for design. [25, 26] This is typically the case for focused libraries that are based on small molecules with specific activities or properties.

L. Xue · J.W. Godden · F.L. Stahura · J. Bajorath (✉)
Computer-Aided Drug Discovery, New Chemical Entities,
18804 North Creek Pkwy, Bothell, WA 98011, USA
e-mail: jbajorath@nce-mail.com
Tel.: +1 425 424-7297, Fax: +1 425 424-7299

J. Bajorath
Department of Biological Structure, University of Washington,
Seattle, WA 98195, USA

Such libraries are usually much smaller than diverse screening libraries and their performance depends, among other aspects, on the ability to extract information encoded in small molecules with desired properties (e.g., hits obtained from screening). Therefore, a critical step is the isolation of building blocks or scaffolds from these molecules. Suitable molecular scaffolds can be isolated by various means, e.g., knowledge of active core structures, [26] synthetic considerations, [35] or algorithms that follow hierarchical descriptions of molecules. [36, 37] Isolated scaffolds are then used to sample diverse scaffold/R-group combinations. This process aims to increase the probability of finding more potent and/or selective compounds by exploring the neighborhood of template molecules in chemical space. [26, 38] The design critically depends on the computational metric applied to generate compounds. Typically, this would be a distance- [34] or cell-based [14] measure of chemical diversity utilizing selected molecular descriptors. [39, 40]

In this report, we introduce a novel computational metric to focus combinatorial libraries and control their degree of diversity. In this approach, the use of different fingerprints and similarity criteria determines the balance between similarity to compounds of known activity and library diversity. The algorithm was implemented for product-based design of targeted libraries or analogs of biologically active molecules. Generation of virtual libraries for the identification of protein kinase inhibitors exemplifies this form of controlled (or directed) diversity design.

Materials and methods

The algorithm, as described in the Results section, was established using SVL code [41] and implemented in the Molecular Operating Environment (MOE). [42] The implementation makes use of a built-in function of MOE, the compound generator of the QuaSAR-CombiDesign module that randomly samples combinations of scaffolds and R-groups from different source databases. [43] As a scaffold database, 57 previously generated core structures were used that target the ATP binding site in protein kinases. [26] These scaffolds were mostly isolated from known kinase inhibitors. [26] As R-group database, we used an in-house generated set of ~1,500 different groups. Non-ring R-groups were automatically isolated from ACD compounds [44] using a previously reported algorithm [37] and ring moieties identified by the RECAP approach [35] were added.

In our calculations, three binary fingerprints of different length and complexity that could be calculated from two-dimensional molecular representations were used. “MFP” is a mini-fingerprint [45] consisting of only 62 bit positions [46] accounting for ranges of three numerical descriptors (the number of hydrogen bonding acceptors and rotatable and aromatic bonds in a molecule) and the presence or absence of 40 structural fragments or keys. [47, 48] Different mini-fingerprints were designed, on the basis of extensive descriptor analysis, to recognize molecules with similar biological activity specifically. [45, 46] In blind test calculations, MFP was found to have a greater than 50% chance of identifying molecules with similar activity but recognized only ~1% false positives. [46] The second fingerprint (“MACCS”) consists of 166 bits, each of which accounts for the presence or absence of one of 166 structural keys. [47, 48] The third fingerprint used in our calculations is a complex fingerprint consisting of 1,024 bit positions, implemented as “PH2D” in MOE. It accounts for all pairwise atomic distances in a molecule, on the basis of graph representations, and creates a signature pattern. [49] A complex fingerprint such as PH2D evaluates molecules at “high resolution” and thus has a higher intrinsic tendency to consider compounds “dissimilar” or generate “diversity” (when used for library generation). In terms of increasing complexity, the three fingerprints used here compare as follows: MFP<MACCS<PH2D.

Using combinations of these fingerprints, different libraries were calculated, each containing 500 to 1,000 compounds (generated from the set of 57 protein kinase inhibitor scaffolds and our R-group collection). During the design process, compounds were filtered according to Lipinski’s rules [17] to ensure basic drug-like properties. As a similarity criterion, fingerprint overlap was calculated for pairwise comparison of compounds using the Tanimoto coefficient (Tc). [50] It is defined as Tc=bc/(b1+b2–bc), with b1 being the number of bits set on in molecule 1, b2 the number of bits set on in molecule 2, and bc the number of bits set on common to both molecules. Scaffold and Tc value distributions of the generated libraries were compared using histograms. Compound distribution in three-dimensional “chemical space” was compared following principal component analysis (PCA) of those molecular descriptors used to generate the fingerprints. [51] For example, all unique descriptors encoded in MACCS and MFP were combined to create an initial multi-dimensional chemical space, the dimensionality of which was then reduced to three by PCA, carried out

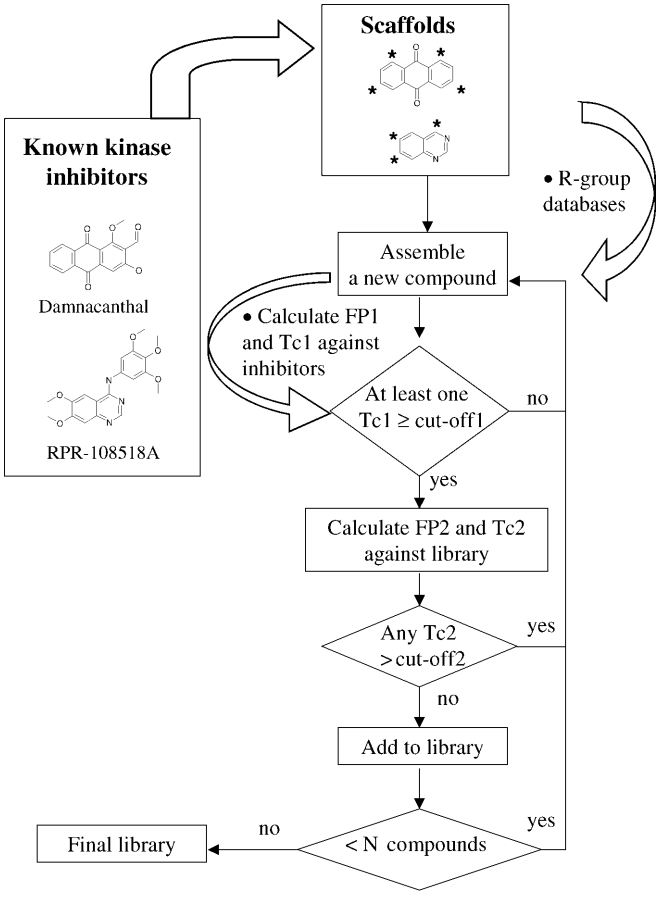


Fig. 1 Description of the method. Scaffolds are isolated from compounds of interest. Kinase inhibitors are used as an example. *Asterisks* indicate points for R-group attachment. Randomly sampled scaffold/R-group combinations are evaluated according to the flow chart. Calculation of FP1 represents the “similarity step”, as described in the text, and FP2 the “diversity step”. The calculations end once a specified number of compounds have been generated

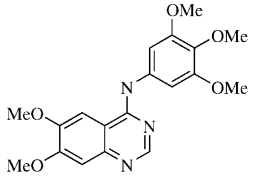
with MOE as described. [52] These first three principal components were then used as a coordinate system for graphical representation of libraries.

Results and discussions

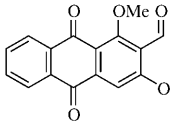
Introduction of the method

Our algorithm is described in Fig. 1. Initially, a collection of molecular scaffolds, isolated from compounds of interest (here kinase inhibitors), is used to sample one candidate compound at a time. In this case, scaffolds have between one and four points of diversity that are randomly decorated with R-groups. For each newly assembled compound, fingerprint 1 (FP1) is calculated and compared with FP1 of selected inhibitors from which scaffolds were isolated. If Tc1 is, at least once, greater than a specified cut-off value (cut-off 1), then the compound is accepted. This means that the newly assembled compound must be "similar" to (or in the chemical "neighborhood" of) at least one of the original inhibitors. In this case, each randomly generated compound was compared with four original kinase inhibitors. If a compound is accepted on the basis of Tc1 comparison, it is, in the second step, compared with all compounds previously accepted by Tc1 comparison (it follows that the first compound passing the Tc1 test will always be in the

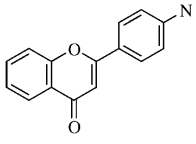
1. RPR-108518A



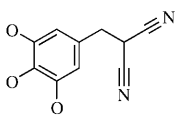
2. Damnacanthal



3. Aminogenistein



4. Tyrphostin



5. Methyl 7,8-dihydroxy-isoquinoline-3-carboxylate

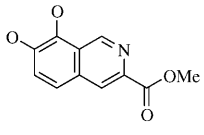


Fig. 2 Chart of selected protein kinase inhibitors. Structures of five protein kinase inhibitors (see [26] and references cited therein) are shown that were used to design a focused library (inhibitors **2**–**4**) or analog libraries (inhibitor **1**)

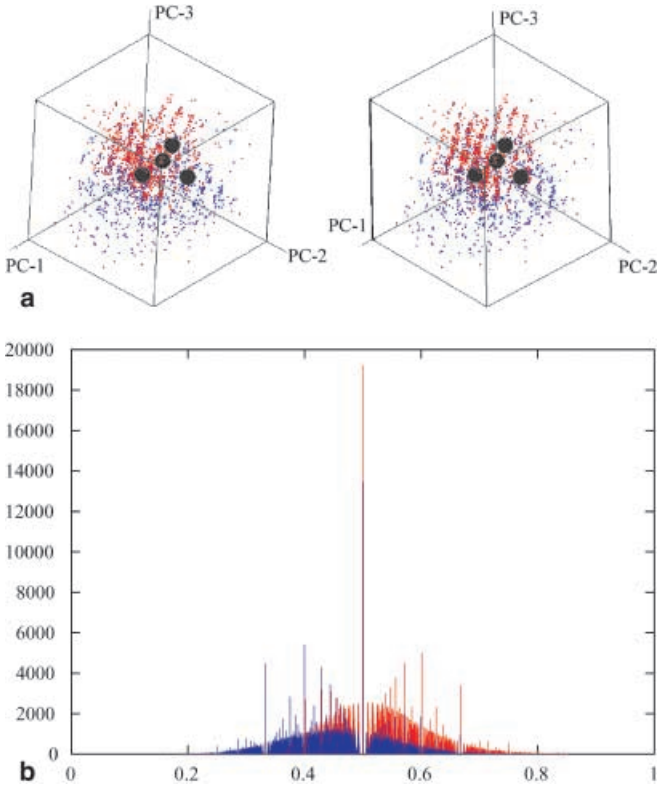


Fig. 3 Analysis of a focused library. In (**a**), a three-dimensional ("chemical space") stereo representation of the library is shown. PC -1, -2, and –3 are the principal component axes of molecular descriptor space. The positions of template inhibitors are shown in black. Compounds belonging to the focused library are colored *red* and compounds in the diverse reference library *blue*. (**b**) Distribution of the Tanimoto coefficient (Tc) in the focused (*red*) and diverse (*blue*) library. Each compound in the library was compared with all other compounds using MACCS as a fingerprint, and Tc values for complete pairwise comparisons are recorded. A shift of the distribution towards higher Tc values indicates increasing similarity of library compounds

library). For this second step, a different fingerprint (FP2) is used from which Tc2 values are calculated (it is important to note that comparisons are only made for assembled compounds and not at the level of molecular scaffolds). If Tc2 is, in at least one case, greater than a specified cut-off 2 value, the new compound is rejected. This means only those compounds are added to the array that are sufficiently "distant" from other library compounds, thereby generating diversity. Thus, in our implementation, FP1 comparisons represent the "similarity step" and FP2 comparisons the "diversity step".

Focusing versus diversity sampling

The similarity step is a requisite for focusing of the library in product space. FP1 comparisons ensure that each library compound is in the neighborhood of at least one biologically active compound selected as a starting point for library design. However, the approach, as implemented here, can also be used for more conventional diversity

design. If cut-off 1 is set to zero, the procedure defaults to diversity sampling in step 2. By contrast, if cut-off 2 is set to 1, only focusing is carried out and no diversity criteria are applied. Thus, dependent on the desired application, a balance between focusing and chemical diversification is achieved by adjusting the parameters.

Focused libraries versus analog design

In our approach, analog design can be rationalized as a special case of focusing where only one inhibitor is used to select compounds. In order to ensure that accepted compounds are closely related to the original inhibitor, a more stringent Tc1 criterion is used in the similarity step. Subsequently, a diversity criterion is applied to select analogs that are not too similar to each other. Focused and analog libraries are typically much smaller (~1,000–10,000 compounds) than libraries designed for screening on multiple targets (~100,000 compounds or more). Therefore, a vast number of possibilities exist in focused library and analog design for sampling chemical space close to template molecules, and our algorithm was designed to produce one possible solution for a library of limited size. Most importantly, it aims to balance the similarity of designed compounds to molecules with known activity with the diversity of the library.

Selection of fingerprints

Our compound design process is carried out using different fingerprint representations. For the similarity step, we apply mini-fingerprints (MFP; see Methods) specifically designed to recognize molecules with similar activity. [45] Thus, selection of compounds on the basis of MFP and using a cut-off 1 of 0.7 or greater is thought to increase the probability of generating compounds with activity similar to a template molecule. [45, 46] For the diversity step, fingerprints of medium (e.g., MACCS) to high complexity (e.g., PH2D) are used, which distinguish more chemical details than mini-fingerprints.

Design of focused libraries

As an application of the dual fingerprint approach, we have designed a library focused on known protein kinase inhibitors (compounds number **2** to **5** in the chart shown in Fig. 2). In this calculation, 57 scaffolds were used as a source (some of which were directly derived from these inhibitors). For the similarity step, MFP (FP1) was used and for the diversity step, MACCS (FP2) was selected. Tc1 and Tc2 cut-off values were set to 0.70 and 0.85, respectively, and a total of 1,000 compounds were sampled. These cut-off values were set as previously suggested for detection of molecules with similar activity [45] and diversity design, [38] respectively. In a reference calculation, 1,000 compounds were generated from 57 scaffolds by diversity sampling only with MACCS (cut-off 2=0.85). Under these conditions, both libraries were generated rapidly, each requiring less than 30 min of CPU time on an SGI Octane workstation. This is because compound sampling largely relies on simple pairwise fingerprint comparisons. Only the diversity step requires an increasing number of calculations as the library grows. Figure 3a shows a graphical representation of these two libraries in principal component space computed from the combined descriptor settings of the two fingerprints (MFP and MACCS). It illustrates the effect of the focusing step. Although the set of 57 scaffolds provided the basis for both calculations, compounds in the

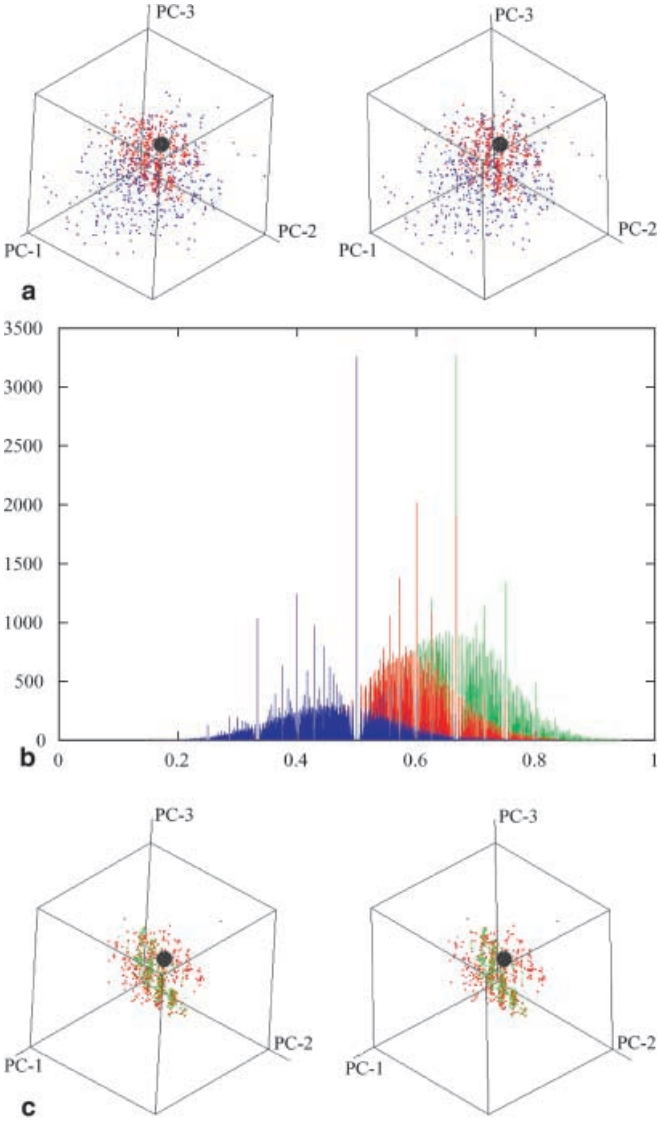


Fig. 4 Analysis of analog libraries. In (**a**), a chemical space representation of the first analog library (*red*; cut-off 1=0.80, cut-off 2=0.85) and the diverse reference library (*blue*) is shown relative to the position of the target inhibitor (*black*). The orientation of the stereo view is the same as in Fig. 3. (**b**) shows the comparison of Tc profiles of the reference library (*blue*) and the first (*red*) and second analog library (*green*; cut-off 1=0.88, cut-off 2=0.97). The Tc distributions were calculated using MACCS as described in the legend of Fig. 3. In (**c**), a chemical space comparison of the two analog libraries (*red* and *green*) is shown. The inhibitor is in *black*. The orientation of the stereo view is the same as before

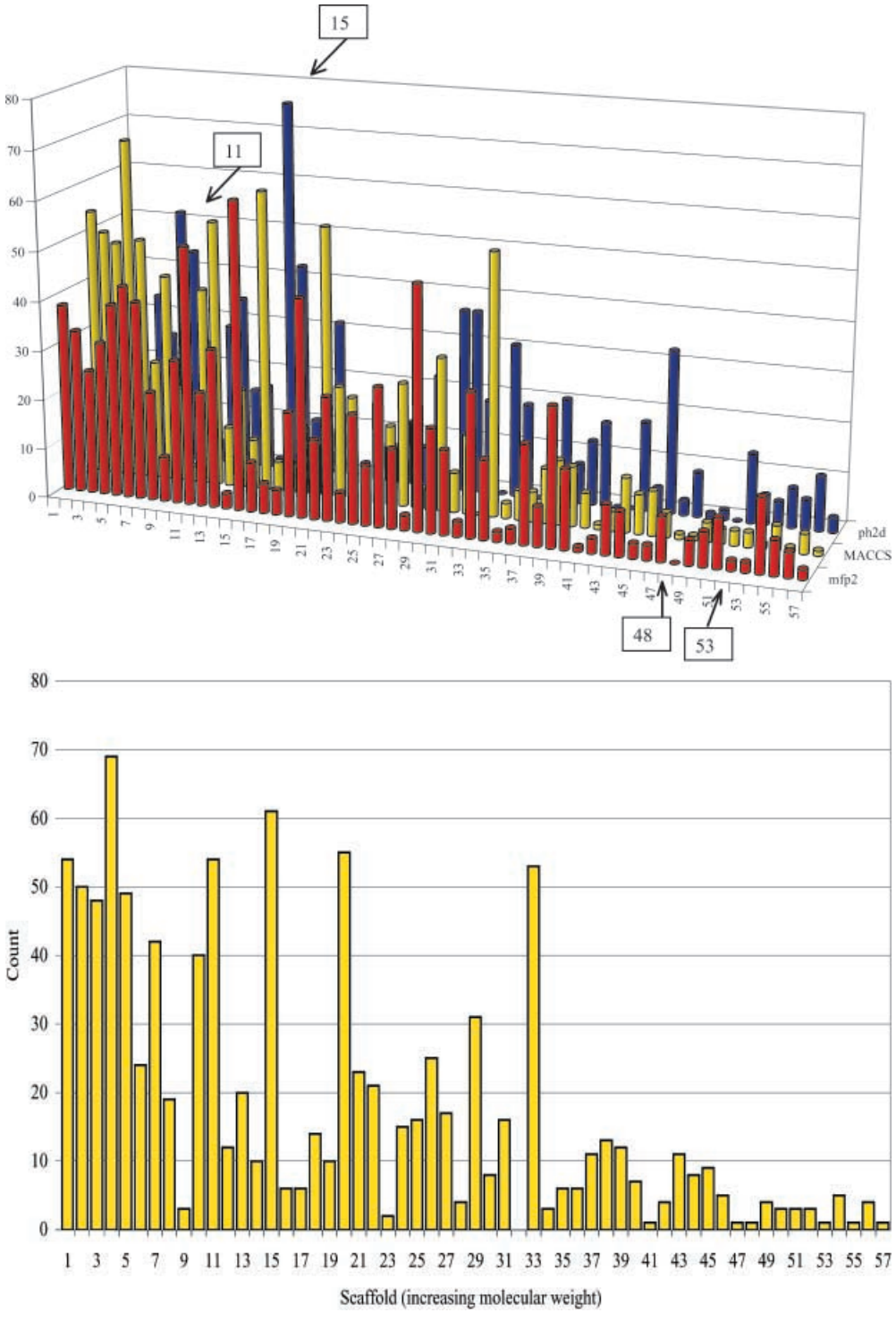


Fig. 5 Scaffold distribution in diverse libraries. The lower histogram shows the frequency of occurrence of 57 molecular scaffolds (ordered according to increasing molecular weight) in a library computed using MACCS. The upper representation compares results obtained for three different fingerprints, MFP (*red*), MACCS (*yellow*), and PH2D (*blue*). Positions of representative scaffolds are labeled

focused library (red) are much more concentrated in the vicinity of the original inhibitors (black) than compounds obtained by diversity sampling (blue). Figure 3b shows a histogram recording Tc values for all pairwise comparisons within each library. The Tc distribution also reflects the focusing effect. It is relatively broad for the diverse library and narrow for the focused library and shifted towards higher Tc values. Thus, compounds in the focused library are not only more similar to the selected inhibitors ("focal points") but also more similar to each other than in the diverse reference library.

Design of analog libraries

One of the inhibitors shown in the chart in Fig. 2 (RPR-108518A) was selected to build an analog library. The set of 57 scaffolds, MFP (FP1; similarity step), and MACCS (FP2; diversity step) were also used in this cal-

culation. However, cut-off 1 was now increased to 0.80. As before, cut-off 2 was set to 0.85. Since the similarity step focused compounds here more stringently on one (and only one) template molecule, it was more difficult to find molecules that passed the diversity test and, accordingly, several hours of calculation time were required to generate an analog library containing 500 compounds. We found that a total of 24 of 57 scaffolds were represented in this library. As a reference library, 500 compounds were generated by diversity sampling using MACCS (Tc2=0.85). Figure 4a illustrates the desired effect of focusing on one inhibitor. The majority of compounds generated surround the inhibitor in principal component space. Enhanced focusing by more stringent Tc1 criterion is revealed by comparison of Tc distributions (Fig. 4b). The distribution of the analog library (red) is now narrower, further separated from the reference library (blue), and shifted towards higher Tc values than was the case for the focused library shown in Fig. 3b. In an additional analog library calculation, cut-off 1 was set to 0.88 and cut-off 2 to 0.97. It follows that compounds more similar to the original inhibitor should be generated in the first step (corresponding to further enhanced focusing) and that, in the second step, only those compounds are rejected that are very similar or almost identical. As shown in Fig. 4b, the Tc distribution of this library (green) is even narrower and further shifted towards higher Tc values than the Tc distribution of the first analog library (red), thus confirming the anticipated effect. Figure 4c compares the two analog libraries in principal component space and illustrates that the distribution of compounds in the second analog library is more focused on the target inhibitor.

Diversity analysis

In an additional application, three different libraries, each containing 1,000 compounds, were computed by applying only the diversity step (cut-off 2=0.75) and using three fingerprints of different complexity (MFP, MACCS, and PH2D; see Methods). In each case, the distribution of our source scaffolds in library compounds was determined. The results reported in Fig. 5 show that scaffolds are not evenly distributed in designed compounds. In general, lower molecular weight scaffolds occur much more frequently in each library than larger scaffolds. Structures of representative scaffolds that are often or, alternatively, rarely utilized are shown in the chart in Fig. 6. The prevalence of lower molecular weight scaffolds is consistent with the idea that addition of R-groups to small core structures leads to greater relative diversity (i.e., more changes in fingerprint bit settings) than addition to large cores. Thus, at relatively low Tc cut-off values (accounting for fingerprint overlap), it is easier to sample compounds that pass the diversity test if core structures are small. However, the results in the chart in Fig. 6 also show that the scaffold distribution is influenced by the complexity of the applied fingerprint. In the case of PH2D (consisting of 1,024 bits), the distribution is shifted towards higher molecular weight scaffolds because this complex fingerprint is more sensitive to changes in minor chemical details. Thus, in this case, addition of R-groups creates “diverse” compounds more easily, even if larger scaffolds are utilized. These observations reinforce the application of conceptually different fingerprints in our calculations. For the similarity step, fingerprints designed to capture molecular features responsible for specific biological activities [45] are preferred. By contrast, for the diversity step, complex fingerprints more sensitive to minor variations in structure or chemical properties are more appropriate, since they are likely to provide a more even distribution of preselected scaffolds in library compounds.

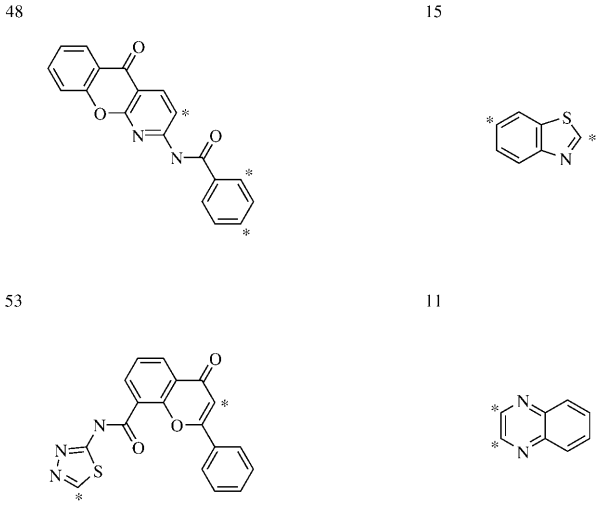


Fig. 6 Chart of structures of scaffolds in Fig. 5 either frequently or rarely utilized. *Asterisks* indicate points for R-group attachment

Although other diversity algorithms have been described in the literature, especially for clustering of compounds or selection of “representative” compounds from databases, e.g. [53, 54, 55], our method differs from previous work. It deliberately combines a similarity step with a diversity step and, in addition, it is entirely fingerprint-based. Since we start from randomly generated compounds, our approach is in essence a “filtering” method. Conceptually, it is perhaps most similar to, yet distinct from, the “OptiSim” approach, [53] which also uses fingerprints for “dissimilarity” selection.

Conclusions

We have introduced a fingerprint-based metric to control the level of chemical diversity achieved in product-based design of compound libraries. It was specifically implemented for focusing of libraries and can be applied to any collection of molecular building blocks or core structures of interest, however obtained. Dependent on the calculation parameters, the method can be used to design target-focused libraries, generate analogs, or, in its simplest form, sample diverse compounds. An attractive

feature of the dual fingerprint approach is that it provides an easily adjustable balance between the similarity of designed molecules to selected templates and the diversity of compounds in the library.

References

1. Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. *J. Med. Chem.* **1995**, *38*, 1431.
2. Ferguson, A. M.; Patterson, D. E.; Garr, C. D.; Underiner, T. L. *J. Biomol. Screen.* **1996,** *1*, 65.
3. Brown, R. D.; Martin, Y. C. *J. Med. Chem.* **1997**, *40*, 2304.
4. Cramer, R. D.; Patterson, D. E.; Clark, R. D.; Soltanshahi, F.; Lawless, M. S. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1010.
5. Pickett, S. D.; Luttman, C.; Guerin, V.; Laoui, A.; James, E. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 144.
6. Koehler, R. T.; Dixon, S. L.; Villar, H. O. *J. Med. Chem.* **1999**, *42*, 4695.
7. Terrett, N. K.; Gardner, M.; Gordon, D. W.; Kobylecki, R. J.; Steele, J. *Tetrahedron* **1995**, *51*, 8135.
8. Balkenhohl, F.; von dem Bussche-Hunnefeld, C.; Lansky, A.; Zechel, C. *Angew. Chem. Int. Ed. Engl.* **1996**, *35*, 2289.
9. Thompson, L. A.; Ellman, J. A. *Chem. Rev.* **1996**, *96*, 555.
10. Houghten, R. A.; Pinilla, C.; Blondelle, S. E.; Dooley, C. T.; Eichler, J.; Nefzi, A.; Ostresh, J. M. *J. Med. Chem.* **1999**, *42*, 3743.
11. Schreiber, S. L. *Science* **2000**, *287*, 1964.
12. Bures, M. G.; Martin, Y. C. *Curr. Opin. Chem. Biol.* **1998**, *2*, 376.
13. Kauvar, L. M.; Laborde, E. *Curr. Opin. Drug Discov. Develop.* **1998**, *1*, 66.
14. Pearlman, R. S.; Smith, K. M. *Perspect. Drug Discov. Design* **1998**, *9*, 339.
15. Mason, J. S.; Hermsmeier, M. A. *Curr. Opin. Chem. Biol.* **1999**, *3*, 342.
16. Polinsky, A. *Curr. Opin. Drug Discov. Develop.* **1999**, *2*, 197.
17. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Freeney, P. J. *Adv. Drug Deliv. Rev.* **1997**, *23*, 3.
18. Ajay; Walters, P.; Murcko, M. A. *J. Med. Chem.* **1998**, *41*, 3314.
19. Sadowski, J.; Kubinyi, H. *J. Med. Chem.* **1998**, *41*, 3325.
20. Walters, W. P.; Ajay; Murcko, M. A. *Curr. Opin. Chem. Biol.* **1999**, *3*, 384.
21. Martin, E. Y.; Critchlow, R. E. *J. Comb. Chem.* **1999**, *1*, 32.
22. Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. *J. Comb. Chem.* **1999**, *1*, 55.
23. Rice, R. L.; Rusnak, J. M.; Yokokawa, F.; Yokokawa, S.; Messner, D. J.; Boynton, A. L.; Wipf, P.; Lazo, J. S. *Biochemistry* **1997**, *36*, 15965.
24. Gray, S. N.; Wodicka, L.; Thunnissen, A.-M. W. H.; Norman, T. C.; Kwon, S.; Espinoza, F. H.; Morgan, D. O.; Barnes, G.; LeClerc, S.; Meijer, L.; Kim, S.-H.; Lockhart, D. J.; Schultz, P. G. *Science* **1998**, *281*, 533.
25. Szardenings, A. K.; Harris, D.; Lam, S.; Shi, L.; Tien, D.; Wang, Y.; Patel, D. V.; Navre, M.; Campbell, D. A. *J. Med. Chem.* **1998**, *41*, 2194.
26. Stahura, F. L.; Xue, L.; Godden, J. W.; Bajorath, J. *J. Mol. Graph. Model.* **1999**, *17*, 1.
27. Ajay; Bemis, G. W.; Murcko, M. A. *J. Med. Chem.* **1999**, *42*, 4942.
28. Salemme, F. R.; Spurlino, J.; Bone, R. *Structure* **1997**, *5*, 319.
29. Kubinyi, H. *Curr. Opin. Drug Discov. Develop.* **1998**, *1*, 16.
30. Antel, J. *Curr. Opin. Drug Discov. Develop.* **1999**, *2*, 224.
31. Kick, E. K.; Roe, D. C.; Skillman, A. G.; Liu, G.; Ewing, T. J. A.; Sun, Y.; Kuntz, I. D.; Ellman, J. A. *Chem. Biol.* **1997**, *4*, 297.
32. Gillet, V. J.; Willett, P.; Bradshaw, J. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 731.
33. Leach, R. A.; Bradshaw, J.; Green, D. V. S.; Hann, M. M. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1161.
34. Jamois, E. A.; Hassan, M.; Waldman, M. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 63.
35. Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511.
36. Bemis, G. W.; Murcko, M. A. *J. Med. Chem.* **1996**, *39*, 2887.
37. Xue, L.; Bajorath, J. *J. Mol. Model.* **1999**, *5*, 97.
38. Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. *J. Med. Chem.* **1996**, *49*, 3049.
39. Brown, R. D.; Martin, Y. C. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 731.
40. Bures, M. G.; Martin, Y. C. *Curr. Opin. Chem. Biol.* **1998**, *2*, 376.
41. Santavy, M.; Labute, P. Scientific Vector Language (SVL). Electronic publication: http://www.chemcomp.com/feature/svl.htm. Chemical Computing Group Inc., 1255 University Street, Montreal, Quebec, Canada, H3B 3X3.
42. Molecular Operating Environment (MOE), version 1999.05, Chemical Computing Group Inc., 1255 University Street, Montreal, Quebec, Canada, H3B 3X3 (see http://www.chemcomp.com/feature/v1999_05.htm).
43. Demers, J.; Murray, S. Molecular databases and MOE. Electronic publication: http://www.chemcomp.com/feature/dbview.htm. Chemical Computing Group Inc., 1255 University Street, Montreal, Quebec, Canada, H3B 3X3.
44. ACD (Available Chemicals Database), MDL Information Systems Inc., 14600 Catalina Street, San Leandro, CA 94577.
45. Xue, L.; Godden, J.; Bajorath, J. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 881.
46. Xue, L.; Godden, J.; Bajorath, J. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1227.
47. McGregor, M. J.; Pallai, P. V. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443.
48. MACCS structural keys, MDL Information Systems Inc., 14600 Catalina Street, San Leandro, CA 94577.
49. Sheridan, R. P.; Bush, B. L. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 756.
50. Willett, P.; Barnard, J. M.; Downs, G. M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983.
51. Glen, W. G.; Dunn, W. J. *Tetrahedron Comput. Methodol.* **1989**, *2*, 349.
52. Labute, P. QuaSAR-Cluster: A different view of molecular clustering. Electronic publication: http://www.chemcomp.com/article/cluster.htm. Chemical Computing Group Inc., 1255 University Street, Montreal, Quebec, Canada, H3B 3X3.
53. Clark, R. D. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1181.
54. Higgs, R. E.; Bemis, K. G.; Watson, I. A.; Wikel, J. H. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 861.
55. Reynolds, C. H.; Druker, R.; Pfahler, L. B. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 305.