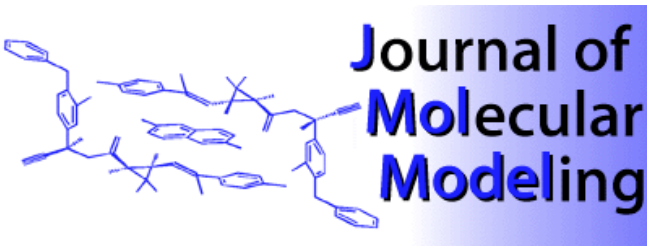
J. Mol. Model. **2000**, *6*, 550 – 562



FULL PAPER

Design of Array-Type Compound Libraries that Combine Information from Natural Products and Synthetic Molecules

Florence L. Stahura[1], Ling Xue[1], Jeffrey W. Godden[1], and Jürgen Bajorath[1,2]

[1]New Chemical Entities, Inc. (NCE), 18804 North Creek Pkwy. S., Bothell, WA 98011-8805, USA. Tel: (425) 424-7297; Fax: (425) 424-7299. E-mail: jbajorath@nce-mail.com

[2]Department of Biological Structure, University of Washington, Seattle, WA 98194, USA.

Received: 28 February 2000/ Accepted: 30 June 2000/ Published: 30 August 2000

Abstract A new approach to the design of compound libraries, named MetaFocus (Metabolite-Focused library), is presented that exploits information encoded in natural molecules and combines naturally occurring and synthetic compounds. An important goal of the MF approach is the identification of synthetic compounds that mimic properties of natural molecules that are difficult to obtain in sufficient quantities or to synthesize. Compounds in MetaFocus (MF) arrays are focused on natural molecules with attractive therapeutic effects. Similarity search and diversity design techniques are employed to generate compound arrays that start from a selected natural molecule, add similar molecules, either from natural or synthetic sources, and diversify scaffolds derived from these molecules. Since the identification of similar molecules from natural and synthetic sources plays a significant role in our library design efforts, the performance of fingerprint-type search tools was systematically assessed in a newly assembled test database consisting of 16 biological activity classes. MF arrays are organized as an easily expandable and searchable data structure and serve as a knowledge base for drug discovery applications. Here we introduce the design principles and organization of MF arrays and present example applications.

Keywords Library design, Molecular scaffolds, Molecular similarity, Naturally occurring molecules, Similarity searching, Synthetic compounds

Introduction

Due to the availability of large-scale combinatorial synthesis and high-throughput screening technologies, combinatorial libraries play a significant role in pharmaceutical research [1-5]. Thus, for many computational chemists, design of combinatorial libraries and analysis of large compound collections have become important topics [5-12]. A variety of design concepts are currently applied [13-16], and large and diverse compound libraries play a significant role in screening [17]. However, at the same time, efforts have increasingly expanded towards the design of smaller and more specialized libraries that are focused on specific therapeutic targets and/or include bioavailability criteria as design parameters [18-25].

Correspondence to: J. Bajorath, NCE, 18804 North Creek Pkwy. S., Bothell, Washington 98011, USA

The assessment of molecular similarity, definition of diversity space(s) for combinatorial exploration, and generation of chemical diversity have been, and continue to be, cornerstones of library design strategies [13-16]. This is also true for target-focused or analog libraries that usually rely on chemical diversification of preferred core structures or selected lead compounds [24-26]. A variety of concepts have been developed to facilitate diversity analysis and design [5-16]. Many of these approaches depend on the definition of chemical diversity space, typically by selection of various molecular descriptors [13-16,27-29]. In this context, it is important to consider that the choice of diversity criteria remains somewhat subjective and that it is still difficult to decide which concepts or strategies are most efficient in generating chemically meaningful diversity.

We have attempted to explore a conceptually different way to design compound libraries for drug discovery applications. In contrast to many contemporary approaches, we initially focus on the selection of naturally occurring molecules with known therapeutic effects, rather than on chemical reactions or preferred synthetic building blocks [30-34]. Although it is widely accepted that natural molecules are valuable drug sources, many of these molecules are chemically too complex to subject them to synthetic programs [35,36]. This may explain why little, if any, effort has so far been spent to generate libraries consisting of both natural and synthetic compounds. Given these limitations, why is it still attractive to "focus" compound libraries on selected natural molecules? A major reason is that many available natural products have associated therapeutic effects [37], for example, anti-bacterial or anti-fungal activity. Thus, careful exploration of structures and properties of such molecules may help to identify novel synthetically accessible molecules with related effects and specificity. In this regard, natural molecules provide a knowledge base for focusing compound libraries on certain therapeutic target areas.

We have selected a number of natural molecules of moderate chemical complexity and interesting biological effects and subjected them to computational design efforts. In these studies, we compute analogs of natural molecules and attempt to identify synthetically accessible mimics for diversity design. These compounds are indexed and organized in chemical arrays that consist of both natural and synthetic molecules. Here we report on our initial efforts to design MetaFocus arrays. We introduce the components of the approach, discuss the design principles and computational techniques, and present applications.

Materials and methods

Source databases

As a source of naturally occurring molecules, the Chapman & Hall compound collection was used [37]. Approximately 116,000 natural molecules were searched in our calculations. As sources for synthetic compounds, ACD [38] and Maybridge [39] were used. A total of ~199,000 synthetic compounds were screened.

Computations and data structures

MOE [40] served as a computational platform to visualize structures, calculate descriptors and properties, launch similarity search calculations (see below), create intermediate databases, and sample diverse compounds (see below). Compounds are indexed by incorporating a code into their data files, as illustrated in the Results section. MF arrays were implemented in MOE, ISIS [41] and also kept as UNIX files.

Similarity searching

The ability to search for compounds with biological activity similar to query molecules, regardless of their chemical origin, is an important aspect of the MF approach. For similarity searching, a binary mini-fingerprint (MFP1, also called SSKey-3DS), consisting of only 54 bit positions, was used [42]. MFP1 was originally designed to detect compounds with similar activity in a test database [42,43]. It consists of numerically encoded descriptors that account for molecular flexibility, aromatic character, and hydrogen bonding acceptors (a total of 22 bit positions) and, in addition, 32 bit positions accounting for the presence or absence of defined structural key-type [28,44] molecular fragments. Details of the MFP1 design were previously reported [42,43]. The fingerprint was generated using SVL code [45] and implemented in MOE for similarity searching.

Analysis of fingerprint performance

We have carried out systematic calculations to test the predictive value of MFP1 prior to its application in array design. To do so, we have assembled a new benchmark database consisting of 264 compounds belonging to 16 different biological activity classes. The detailed composition of the database is reported in the Results section. Seven of 16 activity classes consisted of synthetic compounds and were selected from the literature as described [46]. The other nine activity classes consisted of natural molecules and were assembled by searching for compounds with similar specific activity documented in the Chapman & Hall database [37]. The resulting database was imported into MOE for search calculations. The performance of MFP1 was assessed by systematic "one against all" calculations where each compound was separately searched against the remainder of the database and the percentage of correctly identified compounds (i.e., belonging to the same activity class) and false positives (i.e., belonging to different activity classes) was calculated. As a similarity criterion, fingerprint overlap between query molecules and database compounds was calculated using the Tanimoto coefficient (Tc)

[47]. Two series of calculations were carried out. First, compounds were considered similar at a Tc cut-off value of at least 0.85, a value often used to indicate chemical and/or functional similarity of compounds [6]. Second, Tc values were systematically varied between zero and one in increments of 0.01 to determine the Tc cut-off value that yielded overall best similarity search performance. Thus, for each of the 264 compounds, 101 similarity search calculations were carried out. On the basis of these calculations, overall performance values were determined. Furthermore, systematic calculations were carried out to assess class-specific influences on overall prediction performance. In these calculations, each of the 16 activity classes was omitted once from the database and the exhaustive "one against all" search calculations were repeated, yielding 16 performance values for different combinations of 15 activity classes. All programs required for these calculations were written in SVL [45] and implemented in MOE.

Scaffolds and R-groups

Our approach to generate diverse compounds, as described below, relies in part on a hierarchical description of molecules [32], which means that molecules are divided into scaffolds (or core structures) and R-groups. Isolation of scaffolds from natural molecules and their mimics provided the basis for computational design of derivatives. There are different ways to define molecular scaffolds. They can be isolated from compounds automatically by use of algorithms that break defined bonds in molecules [32,48], are knowledge-based [23,24], or encode reaction information [34]. In knowledge-based scaffold design, core structures of known inhibitors or lead compounds are selected and used as molecular building blocks for combinatorial exploration [23]. This requires prior knowledge about core structures that are active against given targets, for example, compounds that are ATP-analogs and known to bind to the cofactor binding site in tyrosine kinases [24]. Chemical reaction information can also be used to isolate molecular building blocks from compounds, which is well illustrated by the retro-synthetic RECAP approach [34]. Although we can generate molecular scaffolds automatically [48], they were, for the examples presented here, defined by selecting points of chemical diversity in a molecule as sites for R-group attachment. In addition, we implicitly incorporated reaction information by specifying intermediate products for a given reaction as separate scaffolds. This was done to ensure synthetic feasibility of designed compounds. Our R-group database consisted of ~1,500 different groups. Non-ring R-groups were isolated from Maybridge compounds [39] using a previously reported algorithm [48] and complemented by ring moieties identified by retro-synthetic compound analysis [34]. This R-group database consisted of moieties with carbon, oxygen, or nitrogen atoms as substitution points. Subsets of the database were created with groups only having either carbon, nitrogen, oxygen atoms as attachment points.

Diversity sampling

Following selection of scaffolds, diverse compounds were sampled by exploring many scaffold/R-group combinations. This design is more similar to product-based [24,31] than reaction-based design [7,10,31], as it relies on molecular scaffolds derived from whole molecules, as opposed to reagent lists. Sampling of scaffold/R-group combinations was performed using the QuaSAR-CombiDesign function of MOE [49] as follows. For addition of R-groups to scaffolds, between one and four substitution points per scaffold were defined. Initially, combinations of scaffolds and R-groups were randomly sampled and selected molecular descriptors calculated for each compound. As descriptors, we used a previously reported set of 57 structural-key type fragments [46], the number of aromatic bonds in a molecule, the fraction of rotatable bonds, and the number of hydrogen bond acceptors. These descriptors correspond to those encoded in our mini-fingerprint [42] (see above). For 100 randomly selected compounds, a principal component analysis [50] of molecular descriptor space was performed and the top three principal components, linear combinations of original descriptors, were selected for subsequent calculations. Monte Carlo (MC) simulated annealing calculations were then carried out to sample diverse compounds in descriptor space defined by three principal components. These calculations started at a normalized temperature (T) value of 1 and proceeded through at least 7,500 MC steps, while T was scaled using a factor of 0.95 from one iteration to the next, until T was smaller than 10^{-6}. During these calculations, compounds were randomly exchanged between the initially enumerated source database and the diverse sample. An entropy-based metric was applied as a diversity criterion: A compound was accepted in the diverse sample if its addition led to an increase in entropy of the descriptor distribution in principal component space. Diversity sampling was terminated when 2,000 compounds were produced from an initially specified ensemble of scaffolds.

Results and discussion

The MetaFocus concept

The MF approach attempts to expand and diversify structures and properties of natural molecules with therapeutically relevant activities. Practically, it relies on the ability to identify synthetic and/or natural molecules that have properties similar to a particular metabolite and to generate diverse derivatives. Figure 1 illustrates the different components of MF arrays using anisomycin as an example. Anisomycin is chemically a relatively simple metabolite from *Streptomyces griseolus* with known, yet little explored, protein synthesis inhibitory activity [51]. Thus, it represents a possible starting point to focus chemical libraries on such effects. The underlying idea is that chemical expansion and diversifica-

tion of the structure of, for example, anisomycin would increase the probability of generating compounds with modulated and perhaps more specific activities. This general principle often applies to the design of focused compound libraries [24,25].

Scaffold design and application

For the product-oriented compound design approach [30,31] (see Methods), the generation of molecular scaffolds plays an important role. Scaffolds are best defined "hierarchically"

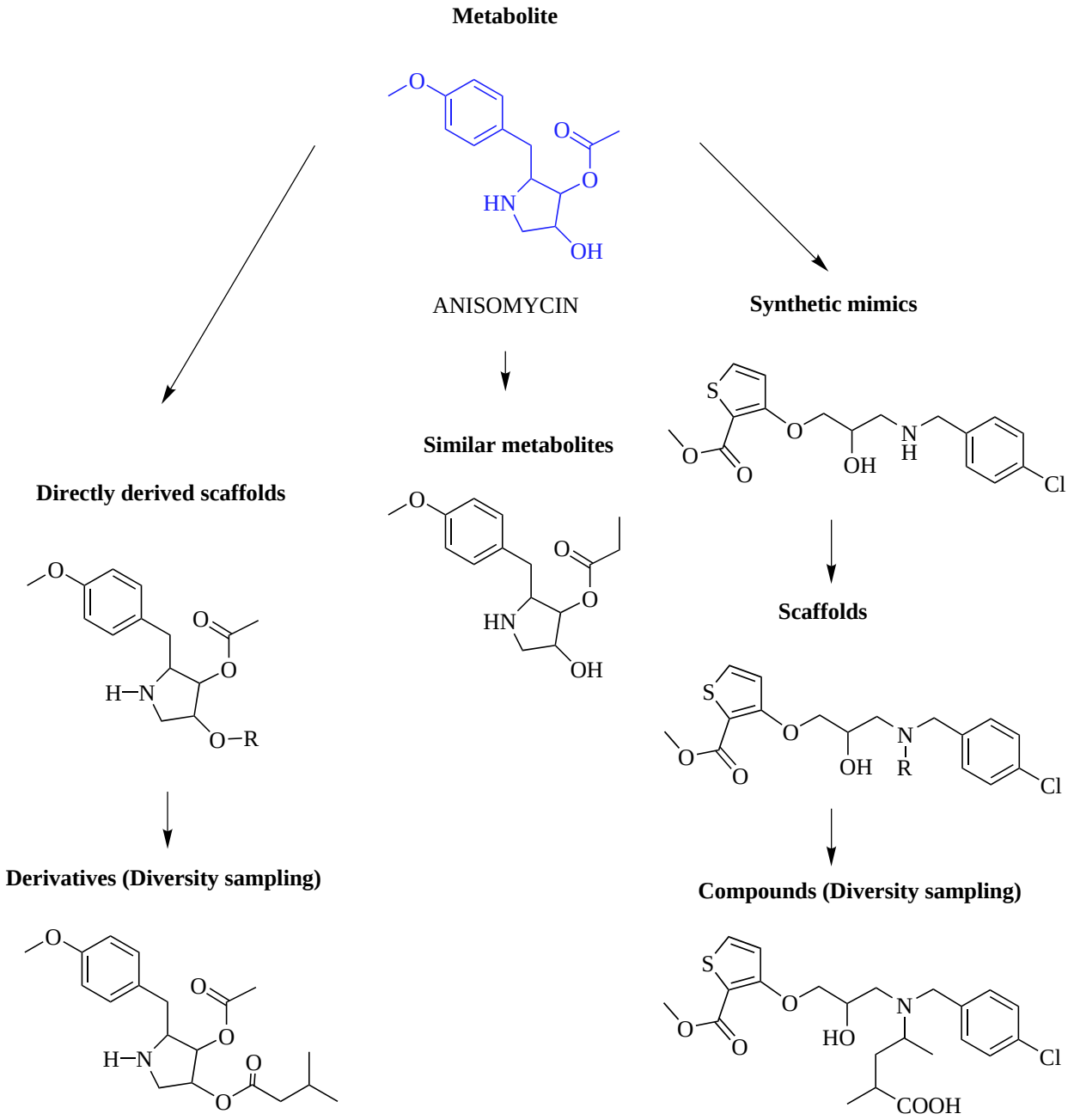


Figure 1 ***Components of MetaFocus arrays.*** *Anisomycin is shown as an example of a natural molecule. Several major components form an array: Directly derived scaffolds, similar natural molecules, scaffolds derived from similar natural molecules (omitted for clarity), synthetic mimics, scaffolds derived from synthetic mimics, and diverse derivatives and compounds generated from scaffolds. Representative examples are shown. Directly derived scaffolds are obtained, for example, by specifying points of diversity that can be targeted in synthetic chemistry efforts to produce derivatives. Similar metabolites and synthetic mimics are obtained by similarity searching. Scaffolds are submitted to diversity sampling of scaffolds/ R-group combinations*

Table 1 *Biological activity classes of compounds in the test database*

	Biological activity	Number of compounds
S_BEN	Benzodiazepine receptor ligands	22
S_CAE	Carbonic anhydrase II inhibitors	22
S_H3E	H3 antagonists	21
S_TKE	Tyrosine kinase inhibitors	21
S_5HT	Serotonin receptor ligands	21
S_HIV	HIV protease inhibitors	18
S_COX	Cyclooxygenase-2 inhibitors	17
N_5LP	5-Lipoxygenase inhibitors	17
N_ACE	Angiotensin converting enzyme inhibitors	9
N_CAT	Acyl-CoA: cholesterol acyltransferase inhibitors	20
N_BLC	β-lactamase inhibitors	14
N_PPD	Phosphodiesterase inhibitors	14
N_PA2	Phospholipase 2 inhibitors	12
N_PKC	Protein kinase C inhibitors	15
N_RVT	Reverse transcriptase inhibitors	14
N_TMB	Thrombin inhibitors	7

The database consists of compounds belonging to 16 biological activity classes. The first column shows abbreviations for each activity class. "S_" indicates synthetic compound classes and "N_" natural molecules

Table 2 *Performance of mini-fingerprint MFP1 in similarity searches including natural molecules*

	cut-off	correct (%)	incorrect (%)	cut-off	correct (%)	incorrect (%)
MFP1	0.85	19.7	0.1	0.75	34.1	1.0
PH2D	0.85	11.6	0	0.64	26.7	0.6
S_BEN	0.85	20.8	0.11	0.75	35.8	1.13
S_CAE	0.85	19.9	0.09	0.75	32.1	1.07
S_H3E	0.85	20.6	0.10	0.75	35.5	1.19
S_TKE	0.85	19.7	0.10	0.75	33.5	1.09
S_5HT	0.85	20.5	0.10	0.75	35.1	1.17
S_HIV	0.85	20.3	0.10	0.74	36.4	1.24
S_COX	0.85	18.9	0.09	0.75	32.1	1.07
N_5LP	0.85	20.1	0.10	0.75	35.1	0.87
N_ACE	0.85	19.8	0.07	0.74	32.8	1.03
N_BLC	0.85	17.7	0.10	0.74	32.8	1.18
N_CAT	0.85	18.9	0.07	0.74	34.2	1.02
N_PPD	0.85	19.9	0.10	0.75	34.6	1.08
N_PA2	0.85	19.4	0.06	0.74	35.3	0.94
N_PKC	0.85	19.0	0.10	0.74	34.6	1.10
N_RVT	0.85	19.8	0.03	0.75	34.8	0.79
N_TMB	0.85	19.7	0.09	0.74	34.4	1.16

Performance is reported for two similarity cut-off values of the Tanimoto coefficient (Tc). The second Tc value represents the similarity cut-off value at which best performance was achieved, as determined in our calculations. A Tc value of 0.85 is often used as a measure of chemical similarity of two molecules [6]. "Correct" reports the percentage of correctly identified compounds and "incorrect" the percentage of false positive matches. The first two rows report the overall performance of MFP1, consisting of only 54 bit positions, in exhaustive "one against all" searches in the test database, and compare its performance to a pharmacophore atom-type fingerprint, PH2D [52], implemented in MOE, that consists of 1,024 bits. Rows three to 18 report results of reference calculations. In each calculation, one activity class was omitted from the databases and exhaustive similarity searches were carried out for compounds belonging to the remaining 15 classes. Biological activity classes are abbreviated according to Table 1

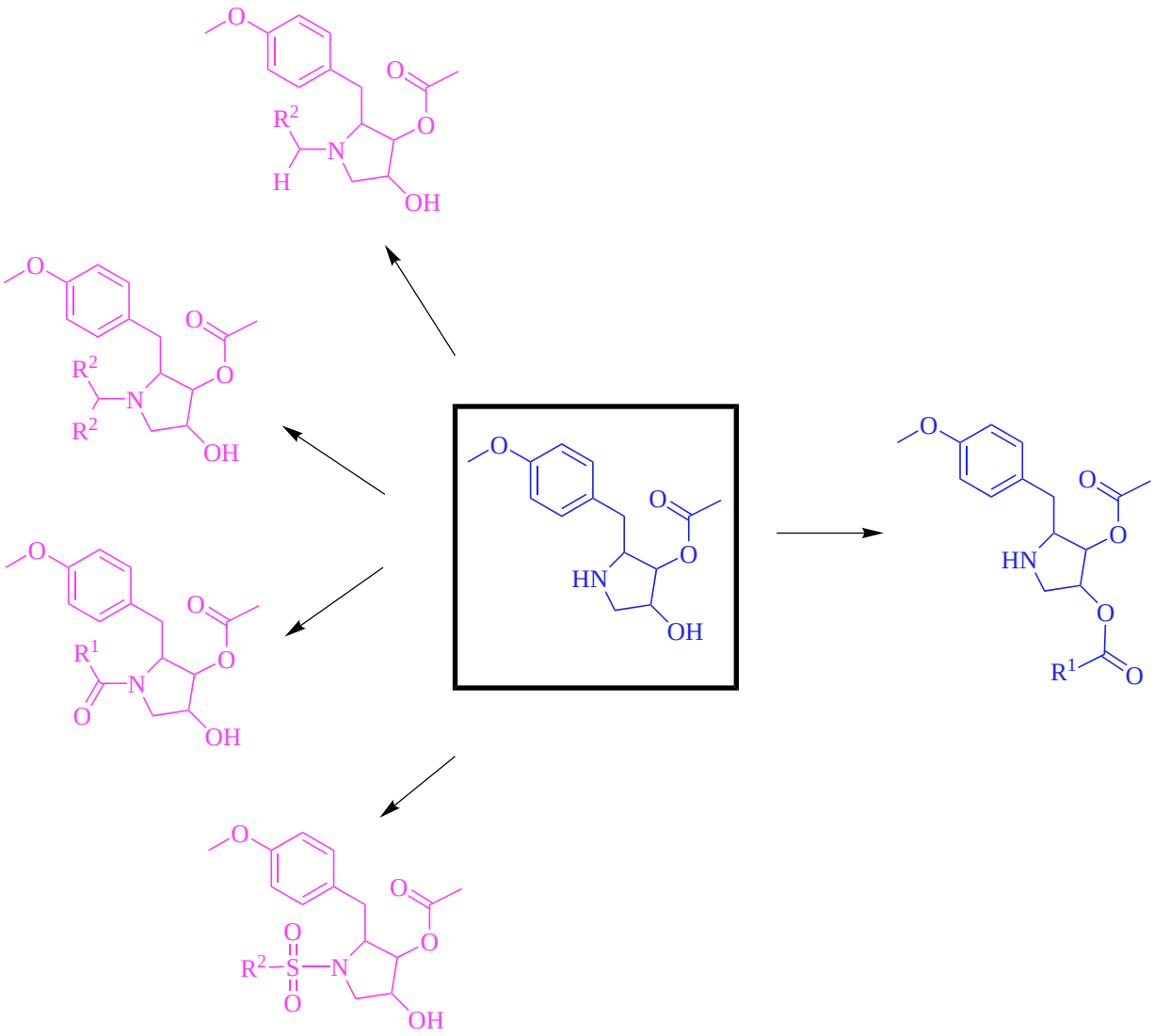


Figure 2 ***Design of directly derived scaffolds using reaction information.*** *Initially, two points of diversity were specified in anisomycin, a secondary amine and the hydroxyl group attached to the aliphatic ring. Then scaffolds were defined that are intermediates of different chemical reactions (four scaffolds for substitutions of the amine and one, an ester, for the hydroxyl group). Combination of these scaffolds yields a total of nine scaffolds for a two-step combinatorial reaction sequence. In this example, R1 substitutions were computed using R-groups with only carbon atoms as substitution points, while R2 substitutions were generated using R-groups with carbon, nitrogen, or oxygen atoms as substitution points*

as core structures obtained from molecules after subtraction of R-groups [32]. Figure 1 shows two different types of scaffolds used to generate MF arrays. In our approach, scaffolds are either directly derived from natural molecules by specifying points of diversity (i.e., sites that can be targeted by diverse chemical reactions) or derived from synthetic mimics. To ensure synthetic feasibility of MF compounds, we attempt to design scaffolds that implicitly incorporate reaction information. This is illustrated in Figure 2, which shows scaffolds directly derived from anisomycin. Two points of diversity were targeted, a secondary amine and a hydroxyl group. The use of intermediates of specific chemical reactions as separate scaffolds incorporates reaction information. For anisomycin, this resulted in four scaffolds for the first point of diversity (secondary amine) and one scaffold for the second (hydroxyl group). Products from two-step combinatorial reactions were designed by diversity sampling of scaffold/R-group combinations at the first point and subsequently submitting the products generated to diversity sampling at the second site. Synthetic feasibility of compounds is further considered by pre-selection of R-groups. For example, in the case of anisomycin, R1 substitutions, as shown in Figure 2, were computed using R-groups with only carbon atoms as attachment points to avoid the design of unstable compounds.

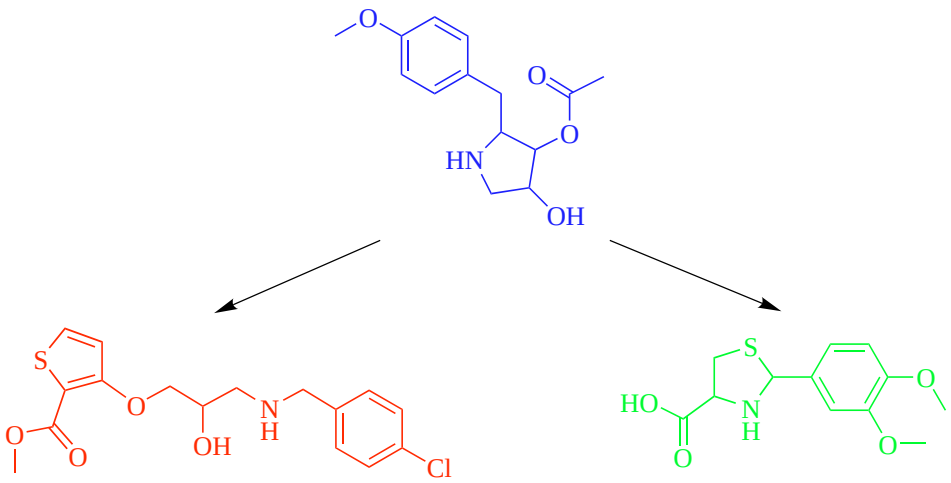


Figure 3 ***Synthetic mimics of anisomycin.*** *Two mimics of anisomycin are shown that were identified by similarity searching in ACD using mini-fingerprint MFP1 (see Methods). The Tanimoto coefficient (Tc) for MFP1 overlap between anisomycin and the two synthetic compounds was 0.8*

Similarity search tools

The identification of synthetically accessible mimics of metabolites is an important component of the MF approach. We have previously reported the generation of small binary fingerprints (termed mini-fingerprints or MFPs) that were specifically designed to recognize molecules with similar biological activity, rather than chemical similarity only [42]. In test calculations using a database consisting of seven compound classes, a total of 455 compounds, MFP1 correctly recognized approximately 50% of compounds belonging to the same activity class and only 2% false positives [42,43]. An important aspect of MFP design has been to balance the level of structural resolution at which compounds are evaluated and the ability to detect features responsible for a specific biological activity [43]. In other words, search tools designed to identify structure-activity relationships correctly must be capable of distinguishing critical features in compounds having different biological activities but should not be too sensitive to minor structural variations that are tolerated within the same activity class. MFPs generated so far were not specifically trained on natural molecules.

In the context of MF library design, we have tested MFP1 for its ability to recognize similarities in both synthetic and natural molecules. To do so, we have assembled a test database consisting of synthetic and natural compounds belonging to 16 biological activity classes (between seven and 22 compounds per class). The exact composition of the test database is reported in Table 1. In this more challenging test case, systematic similarity searches revealed an approximately 34% probability to identify molecules with similar biological activity correctly and only 1% false positives. This over-

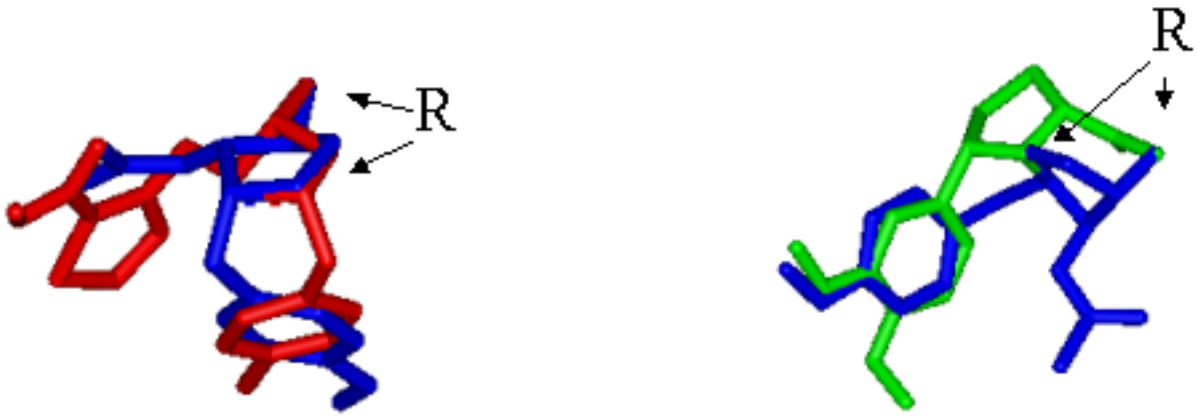


Figure 4 ***Comparison of anisomycin and its mimics.*** *The figure shows optimal three-dimensional alignments (calculated with the flexible alignment function of MOE [56]) of anisomycin and the two synthetic mimics shown in Figure 3 (using the same color scheme). The superpo-sition reveals the spatial correspondence of points of chemical diversity (or substitution points, labeled R) in these molecules*

all performance level was achieved at a Tc value of 0.75. If the Tc cut-off value was greater (and thus more stringent) than 0.75, fewer similarities were identified. The results are reported in Table 2. Consistent with previous findings [42], MFP1 performed better than a more complicated reference fingerprint [52]. Additional calculations shown in Table 2 demonstrate that omission of individual biological activity classes did not notably influence overall performance, indicating the absence of significant class-specific effects. Thus, on the basis of these calculations, similarity searching using MFP1 provided a reasonable chance to identify compounds with similar properties from different sources.

Anisomycin mimics

As the next step, following scaffold design, synthetic compounds were searched for potential mimics. Using a Tc threshold value of 0.8, eleven "similar" compounds were identified and Figure 3 shows two examples. Some structural similarities are evident when comparing these molecules. However, it would have been difficult, if not impossible, to identify these similarities by substructure matching. Figure 4 shows the results of flexible three-dimensional alignments of anisomycin and its mimics, which further illustrates similarities between these molecules. As can be seen, points of chemical diversity in anisomycin (as specified in Figure 2) spatially correspond to those in the mimics. The comparison supports the idea that meaningful similarities can be detected using relatively simple 2D metrics. In addition, search calculations identified 21 natural molecules similar to anisomycin. Since the molecular basis of the protein synthesis inhibitor activity of anisomycin is little explored, functional analogs may act in a variety of ways. This supports the strategy of generating compound libraries focused on anisomycin and its mimics to search for novel inhibitors of protein synthesis.

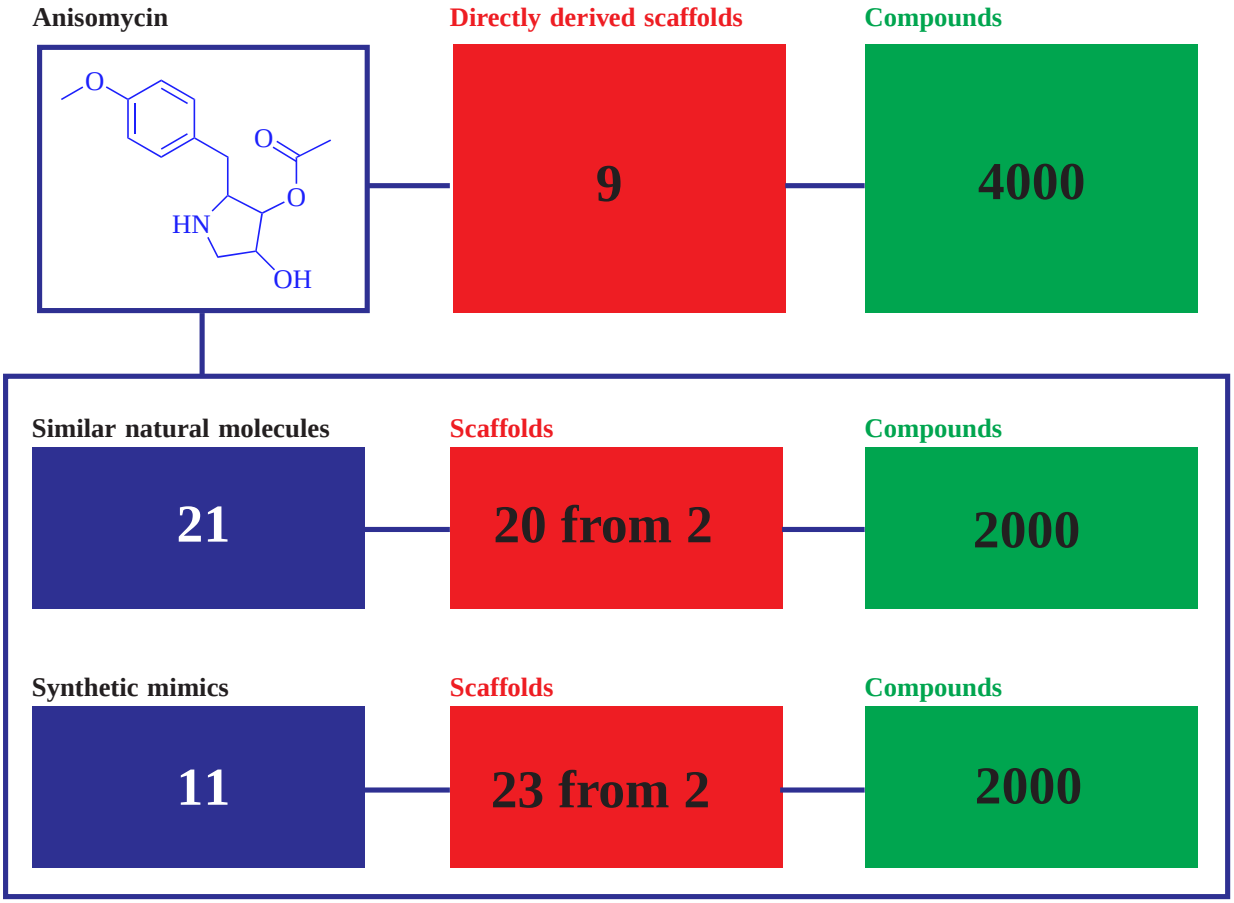


Figure 5 ***Structure of a MetaFocus array.*** *The anisomycin array represents a virtual library of approximately 8,000 entries resulting from directly derived scaffolds, from scaffolds derived from similar natural molecules, or synthetic mimics. "20 from 2" means that a total of 20 molecular scaffolds were derived from two similar metabolites, and "23 from 2" means that 23 scaffolds were derived from two synthetic mimics*

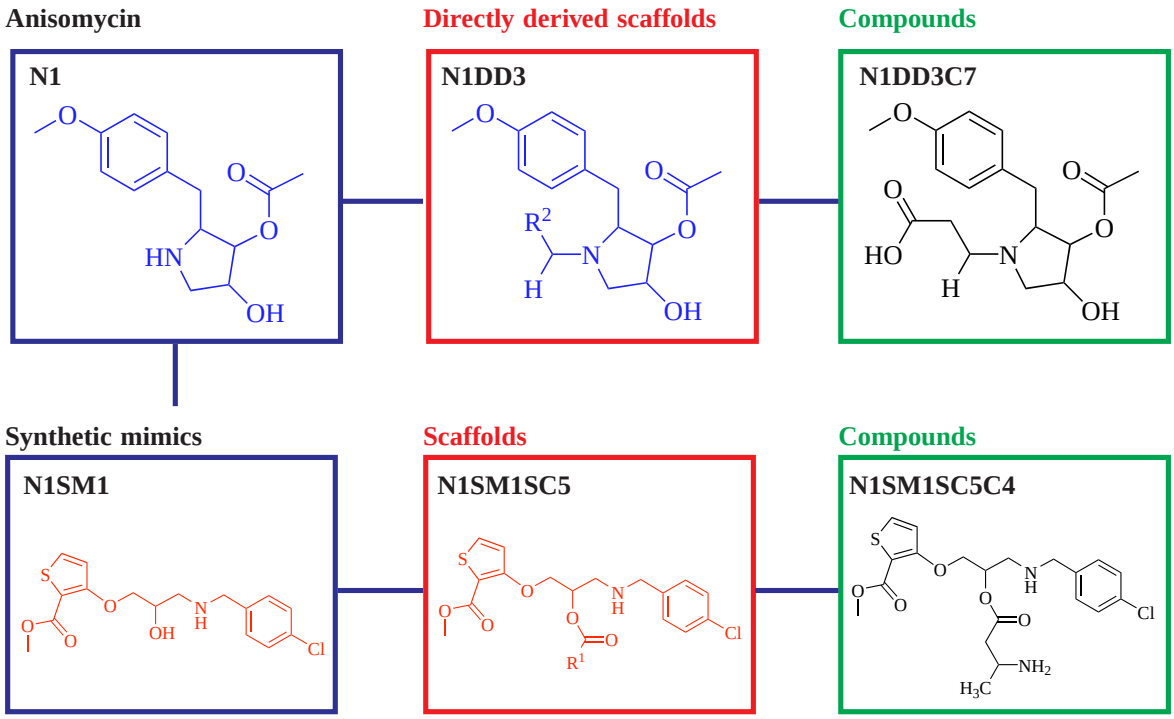


Figure 6 ***Compound indices.*** *The figure illustrates how molecules in MF arrays are indexed to capture their relationships to others. "N" indicates a natural molecule, "DD" means directly derived scaffold, "SC" scaffold, and "SM" synthetic mimic. Not shown here is "SN", the index for natural molecules similar to N. For example, "N1SM1SC5C4" (lower right) defines this entry as the fourth compound obtained by diversity sampling from scaffold number five derived from synthetic mimic number one of natural molecule one (anisomycin)*

Anisomycin array

The MF array was obtained by combining scaffolds, mimics, and diverse compounds generated using anisomycin as template. The organization of the array is shown in Figure 5. The key to the organization and expansion of MF arrays is the use of an indexing scheme. Compound indices are shown in Figure 6. This ensures that all entries and their relation to other compounds are clearly defined and makes it possible to search the array for analogs of natural products. Alternatively, arrays can be searched using synthetic compounds as input to find out whether these compounds are related to any of the natural molecules in the array.

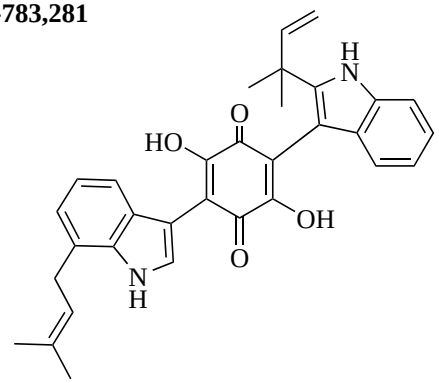
Focusing on protein kinases

Another array was designed based on an "insulin mimetic", a fungal metabolite capable of activating the insulin receptor tyrosine kinase (and thus inducing insulin-dependent signaling pathways) [53]. This molecule, a natural quinone derivative, named L-783,281 [53], is shown in Figure 7. It is chemically more complex than anisomycin. In contrast to anisomycin, its activity has been identified as target-specific. The molecular mechanism of action is yet to be determined but data available so far suggest that binding of L-783,281 to the insulin receptor kinase induces an (activating) conformational change in the region of the ATP (cofactor) binding site adjacent to the catalytic site [53]. Thus, L-783,281 presents an attractive starting point to generate molecules that potentially modulate the specificity and activity of protein kinases, similar to a previously reported approach [23].

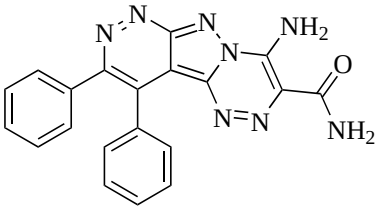
Similarity searching for L-783,281

Using a Tc cut-off value of 0.8 for MFP1 overlap, six synthetic mimics were identified and, using a slightly higher Tc value of 0.85, 21 related natural molecules, the majority being other quinone derivatives. Representative examples are shown in Figure 7. Although the compounds have not yet been tested, literature searches revealed that one of these compounds, bisindolylmaleimide III (BM), has known inhibitory activity against protein kinase C [54]. Comparison of its structure with other kinase inhibitors [23,24] suggested the possibility that it may bind to the cofactor binding site.

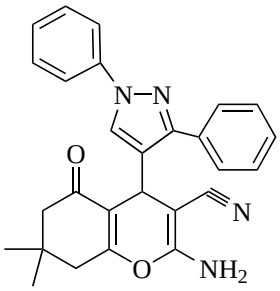
L-783,281



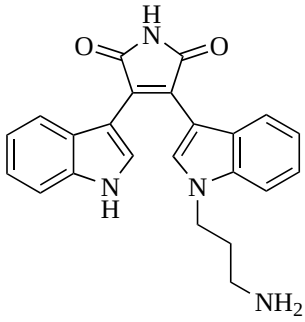
ACD



180424
0.80 (Tc)

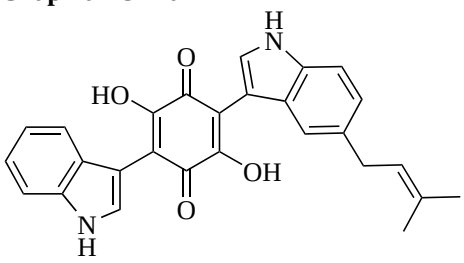


184243
0.80

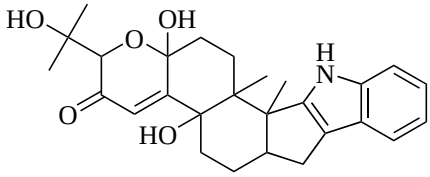


Bisindolylmaleimide III
0.80

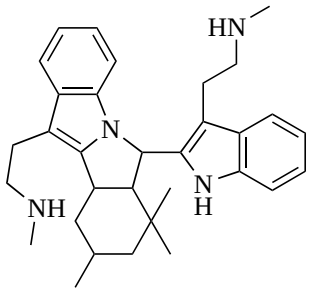
Chapman & Hall



Semicochliodinol A
0.90 (Tc)



14aα-Hydroxypaxilline
0.85



15'-Hyroxy-14',15'-dihydroisoborreverine
0.85

Figure 7 ***Similarity search for fungal metabolite L-783,281.*** *Representative results of similarity searches for "insulin mimetic" L-783,281 are shown. Tc provides the value of the Tanimoto coefficient for L-783,281 and selected mimics*

This region is largely conserved in protein kinases, yet sufficiently different to permit the generation of ligand with distinct specificity [23,55]. Thus, in this case, these compounds are thought to bind to similar sites in related enzymes, yet cause opposite effects. It follows that exploitation of these molecules is likely to yield additional compounds with further modulated effects, consistent with the idea behind MF array design.

Compound and array design

We initially focused compound design on BM. Nine scaffolds were derived and from these, 2,000 diverse compounds were sampled (Figure 8). The current structure of the MF array is shown in Figure 9. In contrast to the anisomycin example (shown in Figure 5), the L-783,281 array is only partially filled, since semi-synthetic derivatives of L-783,281 or

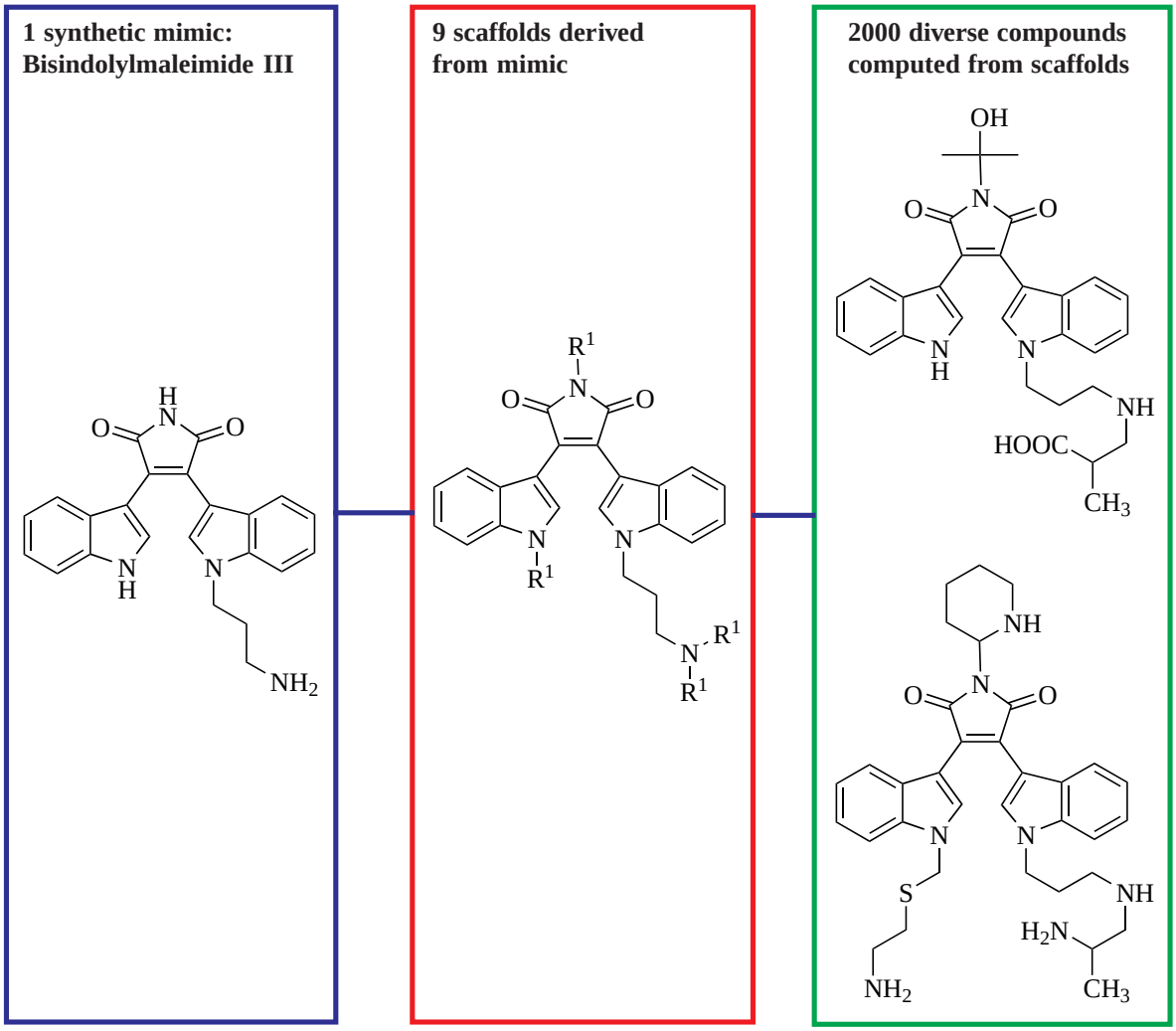


Figure 8 ***Compound design.*** *Bisindolylmaleimide III was identified by similarity searching as a synthetic mimic of L-783,281 (see Figure 7). Nine scaffolds were derived from this molecule and 2,000 diverse compounds were computed from these scaffolds. Representative structures are shown*

derivatives designed from similar natural molecules are presently not included. However, the array can be readily expanded to include, for example, compounds designed from other mimics.

Conclusions

We have attempted to develop a design strategy that relates structures and properties of naturally occurring molecules and synthetic compounds and provides a basis for the generation of natural/synthetic hybrid libraries. A major reason for doing so is that many natural molecules and their activities (even if not well characterized) provide a relatively unexplored knowledge base for focusing compound libraries and generating chemical diversity. The MetaFocus concept captures information encoded in natural molecules and translates this information into synthetically accessible molecules. Each array is focused on a specific natural molecule and presents a defined, yet flexible and expandable data structure. However, there are inherent limitations to the approach and a number of possibilities for improvement. Arrays can certainly not be generated for many natural molecules that are too complex for our current approach. Thus, the selection of suitable natural products will continue to depend, to some extent, on subjective criteria. Furthermore, since a critical component of this concept is the identification of synthetically accessible mimics of natural molecules, we aim to further improve the performance of our search tools to identify molecules with similar properties, regardless of their chemical source. However, irrespective of current computational details, design of MF

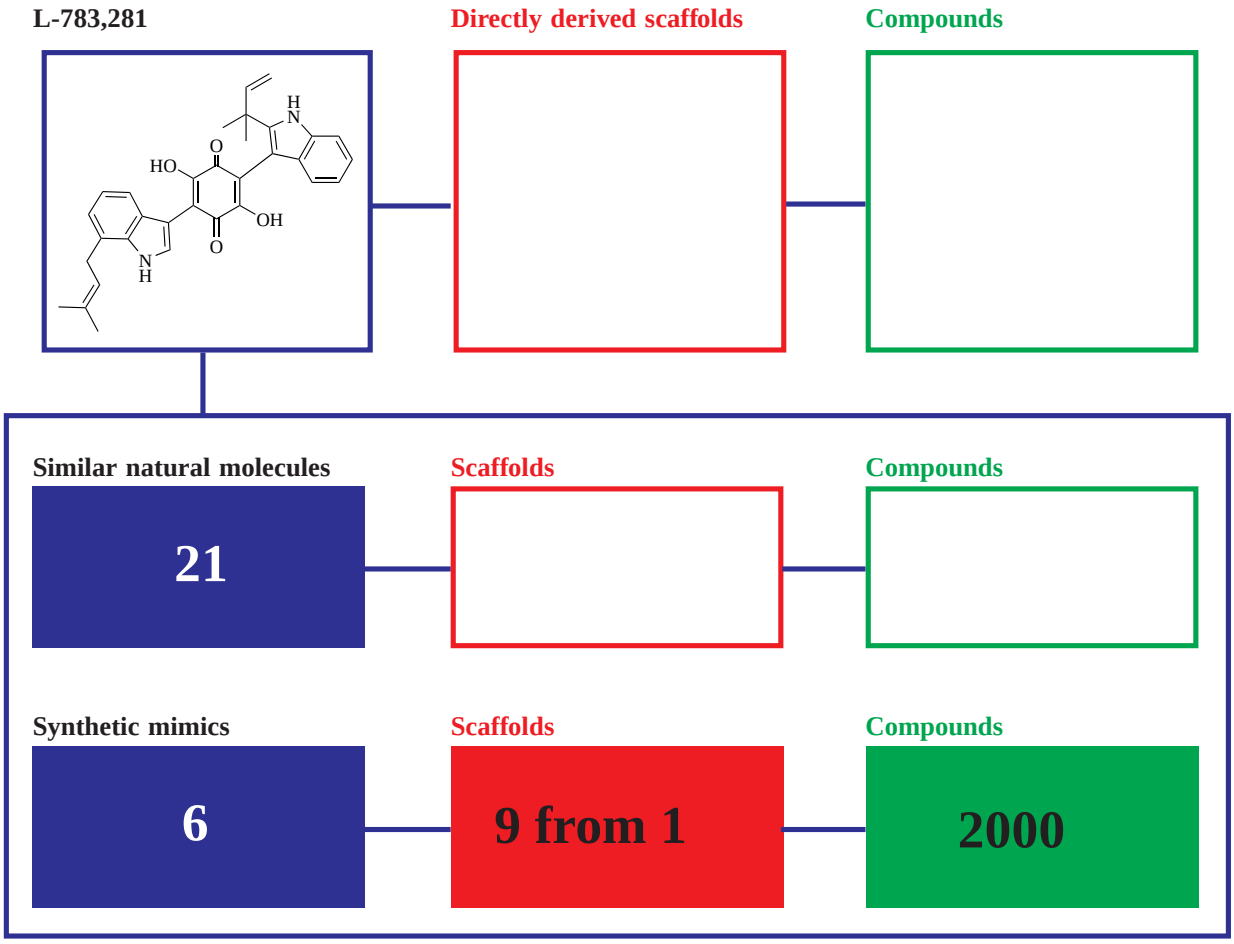


Figure 9 ***Array for L-783,281.*** *In contrast to anisomycin (see Figure 5), this array is only partially filled. The array does not contain directly derived scaffolds. Twenty-one related natural molecules were identified and six synthetic mimics. "9 from 1" scaffolds and "2,000 compounds" represent the design example for bisindolylmaleimide III (shown in Figure 8)*

arrays is beginning to set directions for chemical applications, as illustrated by our examples.

Acknowledgment The authors wish to thank Drs. Z. Chen, C. Garr, and U. Mocek for helpful discussions and suggestions.

References

1. Kubinyi, H. *Curr. Opin. Drug Discov. Develop.* **1998**, *1*, 16.
2. Leach, A. R.; Bradshaw, J.; Green, D. V. S.; Hann, M. M. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1161.
3. Kauvar, L. M.; Laborde, E. *Curr. Opin. Drug Discov. Develop.* **1998**, *1*, 66.
4. Houghten, R. A.; Pinilla, C.; Appel, J. R.; Blondelle, S. E.; Dooley, C. T.; Eichler, J.; Nefzi, A.; Ostresh, J. M. *J. Med. Chem.* **1999**, *42*, 3743.
5. Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. *J. Med. Chem.* **1995**, *38*, 1431.
6. Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. *J. Med. Chem.* **1996**, *39*, 3049.
7. Ferguson, A. M.; Patterson, D. E.; Garr, C. D., Underiner, T. L. *J. Biomol. Screen.* **1996**, *1*, 65.
8. Zheng, W.; Hung, S. T.; Saunders, J. T.; Seibel, G. L. *Pac. Symp. Biocomput.* **2000**, *8*, 588.
9. Schnur, D. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 36.
10. Cramer, R. D.; Patterson, D. E.; Clark, R. D.; Soltanshahi, F.; Lawless, M. S. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1010.
11. Zheng, W.; Cho, S.; Tropsha, A. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 251.

12. Dunbar, J. B., Jr. *Pac. Symp. Biocomput.* **2000**, *8*, 555.
13. Bures, M. G.; Martin, Y. C. *Curr. Opin. Chem. Biol.* **1998**, *2*, 376.
14. Pearlman, R. S.; Smith, K. M. *Perspect. Drug Discov. Des.* **1998**, *9*, 339.
15. Willett, P. *Perspect. Drug Discov. Des.* **1997**, *7/8*, 1.
16. Mason, J. S.; Hermsmeier, M. A. *Curr. Opin. Chem. Biol.* **1999**, *3*, 342.
17. Bajorath J. *Investig. Drugs* **2000**, *WH 5*, 50.
18. Murray, C. M.; Cato, S. J. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 46.
19. Ajay; Bemis, G. W.; Murcko, M. A. *J. Med. Chem.* **1999**, *42*, 4942.
20. Li, J.; Murray, C. W.; Waszkowycz, B.; Young, S. C. *Drug Discov. Today* **1998**, *3*, 105.
22. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Deliv. Rev.* **1997**, *12*, 3.
23. Gray, N. S.; 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.
24. Stahura F. L.; Xue, L.; Godden, J. W.; Bajorath, J. *J. Mol. Graph. Model.* **1999**, *17*, 1.
25. 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.
26. Szardenings, A. K.; Harris, D.; Lam, S.; Tien, D.; Wang, Y.; Patel, D. V.; Navre, M.; Campbell, D. A. *J. Med. Chem.* **1998**, *41*, 2194.
27. Brown, R. D.; Martin, Y. C. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572.
28. Brown, R. D.; Martin, Y. C. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1.
29. Brown, R. D. *Perspect. Drug Discov. Des.* **1997**, *7/8*, 31.
30. Gillet, V. J.; Willet, P.; Bradshaw, J. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 731.
31. Jamois, E. A.; Hassan, M.; Waldmann, M. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 63.
32. Bemis, G. W.; Murcko, M. A. *J. Med. Chem.* **1996**, *39*, 2887.
33. Calvet, A. *J. Mol. Graph. Model.* **1998**, *16*, 49.
34. Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. W. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511.
35. Cragg, G. M.; Newman, D. J.; Snader, K. M. *J. Nat. Prod.* **1997**, *60*, 52.
36. Wrigley, S. K.; Chicarelli-Robinson, M. I. *Ann. Rep. Med. Chem.* **1997**, *32*, 285.
37. Chapman & Hall, Dictionary of Natural Products, CD-ROM 1999 version, CRC Press LLC, 2000 NW Corporate Blvd, Boca Raton, FL 33431, USA.
38. ACD (Available Chemicals Database), MDL Information Systems Inc., 14600 Catalina Street, San Leandro, CA.
39. Maybridge Chemical Company LTD, Trevillet, Tintagel, Cornwall PL34 OHW, UK.
40. MOE, Molecular Operating Environment, Version 1999.05, Chemical Computing Group, Inc., 1255 University Street, Suite 1600, Montreal, Quebec, Canada, H3B, 3X3 (URL: www.chemcomp.com).
41. ISIS, MDL Information Systems Inc., 14600 Catalina Street, San Leandro, CA.
42. Xue, L.; Godden, J. W.; Bajorath, J. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 881.
43. Godden, J. W.; Xue, L.; Stahura, F. L.; Bajorath, J. *Pac. Symp. Biocomput.* **2000**, *8*, 566.
44. McGregor, M. J.; Pallai, P. V. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443.
45. SVL (Scientific Vector Language), Santavy, M.; Labute, P.; Chemical Computing Group, Inc., 1255 University Street, Suite 1600, Montreal, Quebec, Canada, (URL: www.chemcomp.com/feature/svl.htm).
46. Xue, L.; Godden, J.; Gao, H.; Bajorath, J. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 699.
47. Willett, P. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983.
48. Xue, L.; Bajorath, J. *J. Mol. Model.* **1999**, *5*, 97.
49. QuaSAR-CombiDesign, MOE, Chemical Computing Group, Inc., 1255 University Street, Suite 1600, Montreal, Quebec, Canada, H3B, 3X3 (URL: www.chemcomp.com).
50. Glen, W. G.; Dunn, W. J.; Scott, D. R. *Tetrahedron Comput. Methodol.* **1989**, *2*, 349.
51. Martinez, J. L., Jr.; Jensen, R. A.; McGaugh, J. L. *Prog. Neurobiol.* **1981**, *16*, 155.
52. Sheridan, R. P.; Bush, B. L. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 756.
53. Zhang, B.; Salituro, G.; Szalkowski, D.; Li, Z.; Zhang, Y.; Royo, I.; Vilella, D.; Diez, M. T.; Pelaez, F.; Ruby, C.; Kendall, R. L.; Mao, X.; Griffin, P.; Calaycay, J.; Zierath, J. R.; Heck, J. V.; Smith, R. G.; Moller, D. E. *Science* **1999**, *284*, 974.
54. Marano, C. W.; Laughlin, K. V.; Russo, L. M.; Mullin, J. M. *Biochem. Biophys. Res. Commun.* **1995**, *209*, 669.
55. Cohen, P.; Goedert, M. *Chem. Biol.* **1998**, *5*, R161.
56. Flexible alignment function of MOE, Chemical Computing Group, Inc., 1255 University Street, Suite 1600, Montreal, Quebec, Canada, H3B, 3X3 (URL: www.chemcomp.com/feature/malign.htm).