J Mol Model (2011) 17:1473–1482
DOI 10.1007/s00894-010-0850-1

ORIGINAL PAPER

Integrating computational and mixture-based screening of combinatorial libraries

Austin B. Yongye · Clemencia Pinilla · Jose L. Medina-Franco · Marc A. Giulianotti · Colette T. Dooley · Jon R. Appel · Adel Nefzi · Thomas Scior · Richard A. Houghten · Karina Martínez-Mayorga

Received: 5 August 2010 / Accepted: 6 September 2010 / Published online: 21 September 2010

Abstract Mixture-based synthetic combinatorial library (MB-SCL) screening is a well-established experimental approach for rapidly retrieving structure–activity relationships (SAR) and identifying hits. Virtual screening is also a powerful approach that is increasingly being used in drug discovery programs and has a growing number of successful applications. However, limited efforts have been made to integrate both techniques. To this end, we combined experimental data from a MB-SCL of bicyclic guanidines screened against the κ-opioid receptor and molecular similarity methods. The activity data and similarity analyses were integrated in a biometric analysis–similarity map. Such a map allows the molecules to be categorized as actives, activity cliffs, low similarity to the reference compounds, or missed hits. A compound with IC_{50}=309 nM was found in the "missed hits" region, showing that active compounds can be retrieved from a MS-SCL via computational approaches. The strategy presented in this work is general and is envisioned as a general-purpose approach that can be applied to other MB-SCLs.

Keywords Molecular similarity · Mixture-based screening · Biometric analysis · Combinatorial chemistry · Virtual screening

Electronic supplementary material The online version of this article (doi:10.1007/s00894-010-0850-1) contains supplementary material, which is available to authorized users.

A. B. Yongye · J. L. Medina-Franco · M. A. Giulianotti · C. T. Dooley · A. Nefzi · R. A. Houghten · K. Martínez-Mayorga (✉)
Torrey Pines Institute for Molecular Studies,
11350 SW Village Parkway,
Port St Lucie, FL 34987, USA
e-mail: kmartinez@tpims.org

C. Pinilla · J. R. Appel · R. A. Houghten
Torrey Pines Institute for Molecular Studies,
3550 General Atomics Court,
San Diego, CA 92121, USA

T. Scior
Departamento de Farmacia, Facultad de Ciencias Químicas,
Benemérita Universidad Autónoma de Puebla,
Av. San Claudio y 14 Sur, Colonia San Manuel,
72570 Puebla, Mexico

Introduction

Improvements in high-throughput chemical synthesis have made possible the rapid and efficient generation of molecules, giving rise to thousands or millions of compounds in combinatorial libraries. These libraries have been successfully used to identify active molecules for a variety of biological targets [1–3]. Advances in molecular biology have also enabled the evaluation of millions of individual compounds against a number of different biological targets via high-throughput screening (HTS). However, some assays, such as in vivo studies, are not amenable to the high-throughput miniaturization required to screen millions of individual compounds. In such cases, screening libraries using a mixture-based format [4–7] (also known as positional scanning synthetic combinatorial libraries, or PS-SCL) enables the evaluation of thousands to millions of molecules in approximately a hundred to a few hundred samples. This technique has recently found new applications, for instance in the search for conotoxins [8] and in vivo screening [9].

The workflow for the mixture-based screening strategy is shown in Fig. 1. Three main steps are involved in the use of these libraries: synthesis, biological evaluation, and deconvolution (a detailed description of their design and

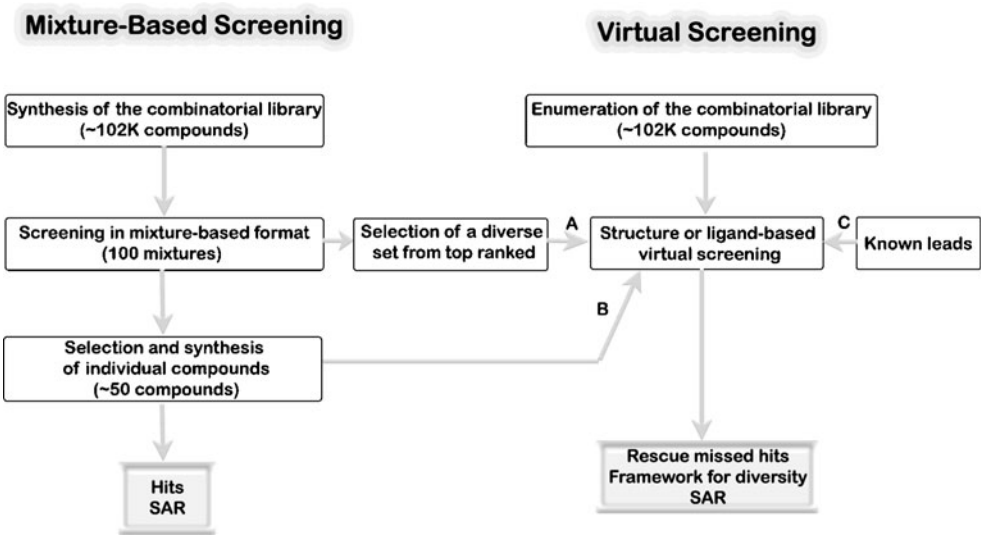

Fig. 1 General workflow involving experimental mixture-based screening and virtual screening

use is described elsewhere [5, 7, 10]). Briefly, the synthesis of small molecules according to the PS-SCL methodology consists of a core template that typically contains three or four substituents (R groups). For each mixture, the substituent at one R position is defined while all of the other R positions are enumerated using ~30–50 reagents for each position. Ultimately, the number of compounds studied in each library corresponds to the product of the number of substituents included, while the total number of mixtures is the sum of the number of substituents for each fixed R position. Each mixture is then screened in a biological assay that can vary from single point measurements to full dose responses. This first screen inherently provides initial structure–activity relationships that are utilized to select promising molecules for synthesis as individual compounds (deconvolution). Typically, the selection of molecules is based on the combination of R groups that showed the best biological response from the mixture-based screening, referred to as traditional positional scanning deconvolution [5]. Another deconvolution method, called biometric analysis (BA) [6, 11, 12], is a scoring matrix that systematically ranks compounds in a library employing information derived from a PS-SCL. BA involves the calculation of the predicted biological activity of each individual molecule from the activity measured at each R position. Using the positional scanning and BA methods, highly active peptides [13, 14] and peptidomimetics which target opioid receptors have been identified [14] .

In silico methods can be incorporated at different stages of the drug discovery process, from library design to lead optimization and metabolism [15, 16]. Computational methods are largely applied to corporative chemical collections [17] as well as combinatorial chemical libraries [18]. However, limited efforts to explicitly integrate information from mixture-based combinatorial libraries and computational techniques have been reported so far [6, 19]. The structural analogy contained in combinatorial libraries in general and in mixture-based libraries in particular deserves particular consideration. For instance, highly dense libraries can be conceptualized as focused libraries and therefore offer a good opportunity to find selective ligands. While heterogeneous libraries will certainly have a greater coverage of the chemical space for the same library size, they have a greater probability of missing hits due to the lower coverage of activity cliffs [20].

Virtual screening may aid the downsizing of large compound libraries and the selection of a smaller set of promising hits, whereas mixture-based screening may screen out some of the false positives of virtual screening by cross-checking their location in a BA ranking. An illustration of how mixture-based and virtual screening methodologies can be used synergistically is presented in Fig. 1. Depending on the information available, the integration of these two procedures can be performed at different stages, as represented in Fig. 1 by the letters A and B. In Case A, the binding affinities of the mixtures have been determined, but individual compounds have not yet been synthesized. At this level, in silico techniques can be utilized to enumerate virtual libraries, with the respective substituent positions occupied by substituents derived from the top-ranked mixtures for each specific position. In Case B, in addition to the screening data from the mixtures, some measure (IC_{50}, percent inhibition) of the activities of individual compounds is also available. These compounds can be screened directly against the structural model of a target or by assessing their similarities to other known drugs (Case C).

Some of the methods utilized in virtual screening include 2D methods (2D fingerprints, substructures), 3D methods

(volume/surface matching, 3D pharmacophore, pharmacophore fingerprints), and molecular docking [21]. The two elements used for similarity searching are a query or reference molecule(s) and a test database to be compared (screened) to the reference. The query molecules are typically known active compounds, leads or drugs. A similarity measure can be used to select molecules with high structural similarity to the reference compound(s) with the aim of finding active molecules [22]. In the case of diversity, compounds that are structurally dissimilar to the query are preferred. It is worth noting that these similarity applications depend on the assumption that a linear relationship exists between the similarity metric and biological activity [23].

In previous studies we used several structural representations to identify consensus activity cliffs [24]. We also explored the use of 3D similarity to analyze SARs of datasets obtained from combinatorial libraries [10, 25].

In this work, we present a step towards the integration of mixture-based combinatorial library screening data and virtual screening information. Here, the predicted activity obtained from the experimental mixture-based screening is combined with structural similarity methods. As a test case, we employed a combinatorial library of bicyclic guanidines screened against the κ-opioid receptor. The activity data and similarity analyses were integrated into a biometric analysis–similarity map. Such a map allows the molecules to be categorized as actives, activity cliffs, diverse, or missed hits.

Methods

Experimental

The bicyclic guanidine PS-SCL was screened in a radioreceptor binding assay specific for the κ-opioid receptor. The entire library was screened at 4 μg/mL, and each of the mixtures was incubated for 2.5 h at 25 °C with 3 nM [^{3}H] U69,593 in a total volume of 0.65 mL of guinea pig brain homogenate [26]. The selected individual bicyclic guanidines (vide supra) were synthesized and biologically tested in a similar manner to the combinatorial library. For an extensive review of the screening results of the bicyclic guanidine PS-SCL in the κ-opioid binding assay, see Houghten et al. [5].

Computational

Query molecules

A set of bicyclic guanidines with high binding affinity was previously identified [5]. A table with showing the substituents at R^1, R^2, and R^3 positions is provided in the "Electronic supplementary material" section (Table S2). From this set we selected four compounds with submicromolar IC_{50} values, but which were as structurally dissimilar to each other as possible (comparison not shown), as our query molecules. In order to identify active molecules that would not be selected when only internal queries were used, we added an external set of query molecules from the Wombat database [27]. This encompasses 70 diverse structures with IC_{50}<250 nM to the κ-opioid receptor, and did not cross the blood–brain barrier in mice assays.

2D and 3D structural similarities

A shape-based similarity comparison was performed with the program ROCS (Rapid Overlay of Chemical Structures) [28]. ROCS reports the volume overlap of two compounds as a single Tanimoto coefficient score, or a linear combination of the Tanimoto score with an additional score (color score) that takes into account heteroatom similarities. The combined score is termed the combo score, with a maximum possible value of 2; this value can only be obtained by comparing a molecule with itself in the same conformation. A multiconformer library was generated utilizing OMEGA [29]. Conformational similarities and rankings were performed between these ensembles of conformers and each query molecule using the ROCS software [28]. The maximum (max) and average (mean) scores for each compound with all the queries were then computed and halved. The ROCS combo scores were also halved. Henceforth, we refer to these halved combo scores as (*rocs*/2).

The virtual library containing ~102 K compounds was enumerated using the CombiGen module of the Molecular Operating Environment (MOE) software. Maximum and average 2D similarities between the queries and database were computed in MOE, employing the Tanimoto coefficients for the following maximally orthogonal fingerprints: MACCS keys, typed-graph distances (TGD), typed-graph triangles (TGT) and graph-based three-point pharmacophores (GpiDAPH3).

Results and discussion

Bicyclic guanidines screened against the κ-opioid receptor

The positional scan synthetic combinatorial library (PS-SCL) evaluated in this study consisted of 102,459 bicyclic guanidines. Details of the synthesis are described elsewhere [5, 30] and a brief description is provided in the "Methods" section. A schematic representation of the positional

scanning library screening results and the BA for the bicyclic guanidines library is shown in Fig. 2. The histograms at the top show the percent inhibition for each mixture containing a particular substituent at the R^1, R^2, or R^3 position. Trends in the preferred type of substituent at each position can be observed, as well as the substituents that on average (in the mixture) produce inactive compounds; see, for instance, the *S*-methylsulfonylethyl group at the R^1 position compared to *R*-benzyl or *R*-cyclohexylmethyl in that same plot. In a previous screening of these compounds [5], defined R^1, R^2 and R^3 substituents in mixtures with percent inhibitions of >80% were selected, and a sublibrary was generated and screened in a mixture-based format. The most active mixtures in the sublibrary were selected, and all possible combinations of the substituents were synthesized, giving rise to 48 individual compounds [5], resulting in molecules with activities ranging from 37 nM to 10,000 nM in the κ-opioid receptor radioligand binding assay. In general, when a clear SAR in the library data is observed, hits or individual compounds with affinities in the nanomolar range can be found in ~90% of the cases (this statistic is based on observations of the libraries analyzed in-house over 20+ years).

The distribution of the normalized biometric analysis (BA) score is shown at the bottom of Fig. 2. This experimental ranking can be compared with other estimations of activity, for instance scores obtained from virtual screening methods. In the following sections, the BA score is analyzed in combination with structural similarity to query molecules.

Molecules evaluated in a single-dose manner (percent inhibition)

A set of 149 molecules were synthesized and evaluated at a single concentration; see Table S1 in the "Electronic

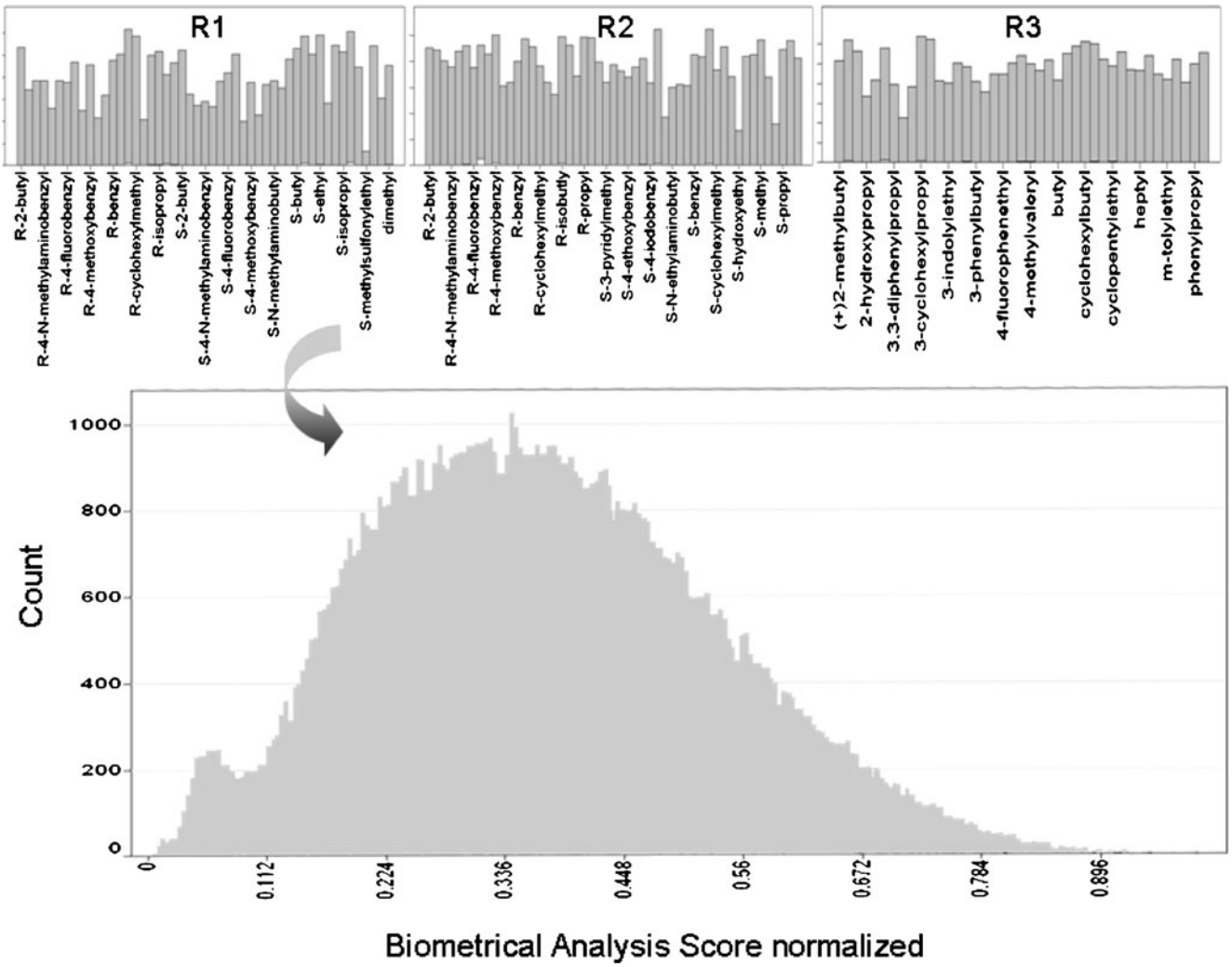

Fig. 2 Positional scanning and normalized biometric analysis (BA) for the bicyclic guanidine library. The BA is the product of the activities of each fixed R group in an enumerated compound. Typically, active molecules are found in the tail of the curve 90% of the time

supplementary material." These molecules were selected in order to explore different areas of the BA distribution. The criteria used to select these compounds for synthesis were based on 3D structural similarities to reference compounds (Fig. S1, "Electronic supplementary material"). Molecules with percent inhibitions above the first quartile (% inhibition>27.46) were considered active in the percent inhibition analysis.

Internal query molecules used

Four molecules were selected based on a pairwise 3D shape structural comparison among the 48 previously reported bicyclic guanidines [5]. Molecules that better differentiated the rest of the set in terms of structural dissimilarity were chosen as "internal query" (iq) molecules (data not shown).

Molecules with combo scores greater than 1.3 (rocs/2= 0.65) after comparison to internal queries

After comparing the entire library (~102 K compounds) with the internal queries, a set of 53 molecules with *rocs*/2 values that were greater than 0.65 (1.3/2) were chosen for synthesis. Previously, the threshold of 1.3 was identified as a reasonable cutoff for internal queries [25]. This selection corresponds to a wide range of BA scaled scores (from 0.05 to 0.9). In other words, these molecules would all be considered potential actives based on *rocs* similarity, but are also distinguishable on the BA axis. Although a few molecules with % inhibition <27.46 (inactives) were located at BA scaled scores of >0.5, no molecules with % inhibition>27.46 (actives) were found below that same BA value. Nonetheless, 47% (25/53) of the selected molecules showed % inhibition values above the cutoff (27.46).

Top-ranked molecules by combo score after comparison to external queries (Wombat database) (rocs/2> 0.67)

In addition to the internal queries, a set of 70 compounds was selected from the Wombat database to serve as external queries. These compounds showed affinity toward the κ-opioid receptor (see "Methods"). As expected, the similarity values (*rocs*/2) were lower than those obtained when internal queries were evaluated, and none of the molecules gave *rocs*/2 > 0.9. The top 73 ranked molecules (*rocs*/2> 0.67) were selected for synthesis (Fig. S2). It was suggested previously that a similarity cutoff of 0.99 (*rocs*/2=0.49) can be applied when using external queries [31]. However, in this work, employing a cutoff of *rocs*/2 = 0.49 resulted in a large number of compounds; the similarity value employed here (*rocs*/2>0.67) allowed the selection of a manageable number of compounds for synthesis and biological evaluation. Of the 73 compounds selected with *rocs*/2 > 0.67, only five (7%) were active according to the percent inhibition cutoff. Although expected, this low ratio of actives highlights the great challenge of identifying active compounds based on external reference compounds rather than internal queries.

Stereoisomers of the hydroxylated analog of the most active compound obtained from previous studies

During the course of a previous study, the most active bicyclic guanidine (R^1=*S*-methyl, R^2=*S*-4-methoxybenzyl, R^3=3-cyclohexylpropyl) had a 4-methoxybenzyl group at the R^2 position [5]. In addition, 3D ROCS overlays revealed that the 4-methoxy moiety of the (*S*)-4-methoxybenzyl substituent coincided with the hydroxyl group found in known opiates [25]. By analogy with opiates [25], it was hypothesized that the hydroxyl analog of the most active compound in the library would be a good or even better binder than the methoxylated analog. Therefore, four stereoisomers of the hydroxyl analog of this bicyclic guanidine were synthesized and evaluated for κ-activity. Structurally, these four compounds are closely related. However, they become distinguishable on the BA scale. Only the *S*-4-hydroxybenzyl stereoisomer (R^1=*S*-methyl, R^2=*S*-4-hydroxybenzyl, R^3=3-cyclohexylpropyl) had a higher % inhibition than the cutoff employed here (27.46, see above). Surprisingly, the hydroxyl analogs were not as active as their methoxylated counterparts. This suggests that, even though they could adopt a conformation similar to known opiates, their binding modes might not be the same. This finding shows how screening in a mixture-based format is able to provide counterintuitive hits.

Random selection

Fifteen molecules were selected at random, and three of these displayed favorable activity (>27.46%). This represents a proportion of 20% actives in the selected set. As expected, and regardless of the activity, the majority of the random points fell within the most frequent in the distribution of the normalized BA scores in Fig. 2.

Molecules with high combo score values (rocs/2> 0.9) and middling, low or high biometric analysis scores

Molecules with high similarity values (combo score > 1.8; *rocs*/2>0.9) but with different BA scores were also explored. From the standpoint of the similarity search, the three sets of molecules were all very similar to the reference compounds; however, they exhibited different BA scores (Fig. S3). Among these compounds, the molecule with a low BA score did not possess a percent inhibition value that

was greater than the percent inhibition cutoff. One of the two molecules with BA score that was slightly above 0.6 had a percent inhibition that was greater than the cutoff. Finally, two molecules had BA scores above 0.8, and both were active. These sets of molecules showed that mixture-based screening is able to distinguish molecules with high structural similarities.

The following aspects will be described below:

- Quantitative comparison of the performance of BA versus random selection
- Cross-enhancement gained from the integration of virtual screening and BA.

Quantitative comparison of the performance of BA versus random selection

The recovery of active compounds with respect to the percentage of screened molecules was assessed. In Fig. 3, the diagonal line represents random selection, whereas a hypothetical perfect recovery would ideally lie close to the vertical axis [32, 33]. Each false positive outcome is graphically registered as a data point, each of which deforms the vertical segment of the curve and brings it closer to the diagonal. Thus, curves above the diagonal are considered better for selecting active molecules than random selection. The curve obtained for BA is shown as open circles in Fig. 3. The categorical classification for activity based on percent inhibition was set to 27.46%, which corresponds to a percent inhibition above the first quartile. This curve shows that the retrieval of active individual molecules (in % inhibition) based on the BA score is better than random selection. One numerical way to evaluate this performance is by the area under the curve (AUC). As can be seen, the AUC for BA is 0.75, whereas for random selection it is 0.5. For the dataset presented here, BA clearly outperformed random selection when analyzing binding affinity data. Although the enhanced performance of BA compared to random selection has been observed before [7], this observation is shown in a quantitative manner here.

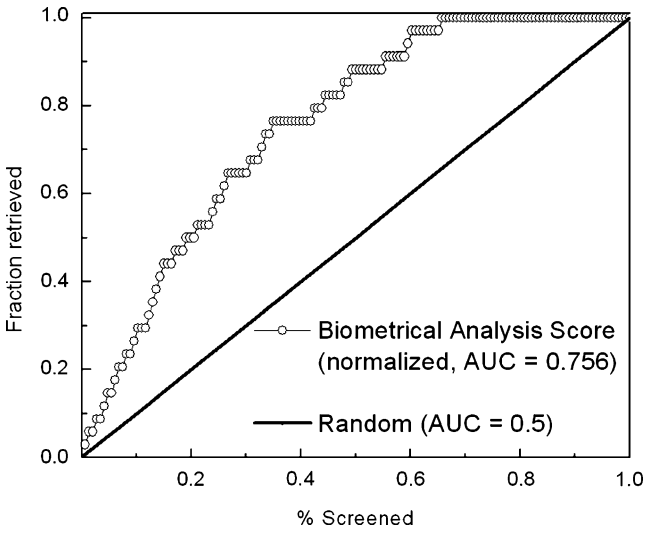

Fig. 3 Recovery of active compounds utilizing the biometric analysis score, as compared to random selection

Cross-enhancement obtained by the integration of virtual screening and BA

A schematic representation of similarity values to reference compounds versus BA scores is provided in Fig. 4a. Related graphs that elaborate on this type of analysis have been proposed [34, 35]. The *y*-axis may contain virtual screening scores, or even consensus among fingerprints [24] or methods [36, 37]. As depicted, data points in each quadrant provide insightful information about the structure–activity relationships of a group of molecules. Data points located on the diagonal signify that such molecules are equally scored by the BA and the virtual screening method. However, it should be noted that BA and virtual screening scores (similarity values in this case) are independent measures, and there are no fixed cutoffs defining each quadrant (the gray lines in Fig. 4a). Different similarity values have been used to identify a molecule that is very similar to a reference compound: 0.85 (2D similarity methods in general), 0.49 (*rocs*/2 when comparing to external queries) [31] and 0.65 (*rocs*/2 when comparing to internal queries) [25]. In a similar manner, there is no fixed cutoff to partition the BA scale. In fact, guidance on this partition is one of the issues of this study.

Quadrant I represents molecules with high BA scores but low similarity to the reference compounds. Selecting molecules in this quadrant facilitates the selection of compounds that are as structurally different as possible from the reference but that are still highly likely to be actives. Molecules located in quadrant II represent potential hits based on BA and similarity. Selecting molecules in this quadrant provides additional hits with a similar chemical nature to the reference. Molecules in quadrant III represent activity cliffs. Molecules with both low BA and low similarity scores are the least relevant.

An added value of the analysis shown in Fig. 4 is the distinction between activity cliffs and consensus hits. There is further differentiation along the BA axis for the same virtual screening score (*y*-axis); likewise, there is differentiation along the *y*-axis for the same BA score.

Since the BA axis provides a group of compounds that have a high probability of being actives, this selection appears suitable for applying computational filters; for instance, a diversity analysis can be performed. Although the intrinsic diversity within the combinatorial library is limited, further

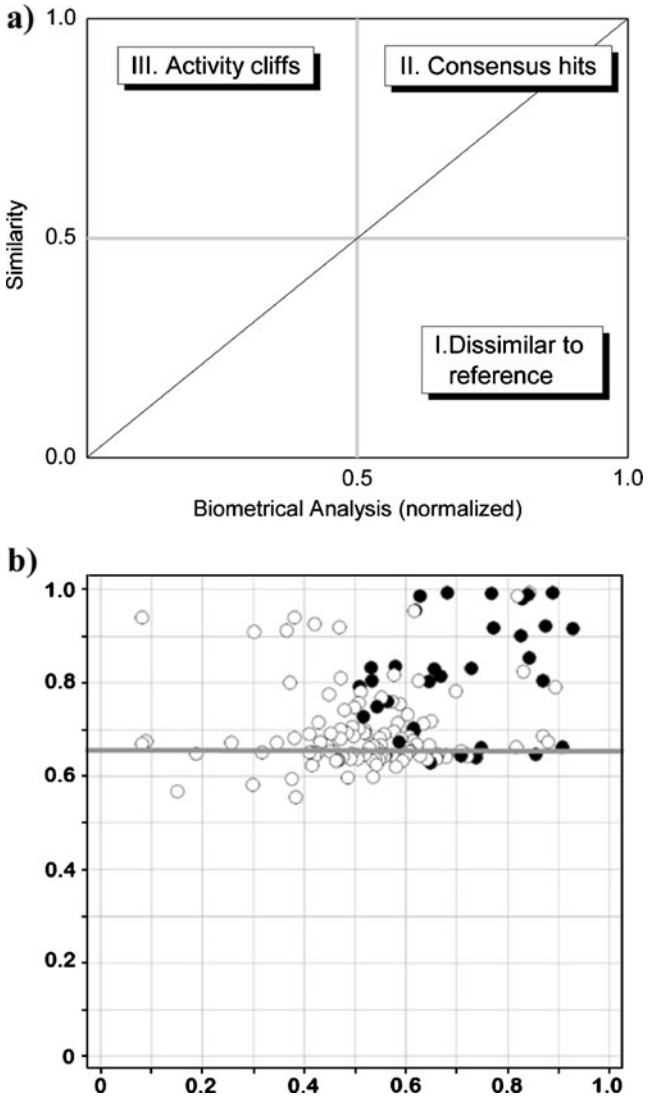

Fig. 4 **a** A schematic representation of a BA–molecular similarity map. Region I contains relatively dissimilar molecules to the reference compounds. Regions II and III contain data points that are equally likely to be actives that can be distinguished in the BA axis, consensus hits or activity cliffs. **b** The biometric analysis–molecular similarity map utilizing *rocs*/2 for 149 bicyclic guanidines. *Closed circles* represent molecules with % inhibition >27.46

selection based on diversity will help to retrieve underrepresented hits. A second filter that can be employed is based on physicochemical properties. Physicochemical properties, and in particular drug and lead likeness, are common criteria used when filtering and designing libraries. However, for the particular case of mixture-based combinatorial libraries, it is advantageous to perform this filtering once there is information at the mixture-based level.

BA–similarity map

The BA–similarity map was constructed with *rocs*/2 similarity values; see Fig. 4b. Closed circles show molecules with % inhibition>27.46. Note that the BA score is derived from the percent inhibition of the mixtures, and it allows the prioritization of individual compounds from a combinatorial library. The percent inhibition analyzed in this section is the actual biological evaluation of each individual compound. It is worth noting that all of the molecules with percent inhibitions above the cutoff (27.46) are also above the normalized BA score of 0.5, suggesting that using the percent inhibition in the mixture to infer the percent inhibition of an individual compound is sound.

For the set of compounds analyzed in this study, all of the active molecules were located towards the right-hand side of the BA–similarity map. The solid gray line shows a cutoff value of 0.65 on the similarity axis. Since the criteria for synthesizing compounds were based on high similarity (based on *rocs*/2 > 0.65) to the reference compounds, most of the molecules are located above *rocs*/2 = 0.65. Molecules below this threshold correspond to those selected at random. In terms of a BA cutoff, these data show that no active compounds were found below a BA scaled score of 0.5, suggesting that a BA scaled score of 0.5 may be used as a threshold when analyzing % inhibition. The general applicability of this threshold remains to be evaluated. This threshold depends on each library and biological assay, so it should be considered only a guideline.

For comparison, 2D similarity measures were calculated using four fingerprints, and a cutoff of 0.85 was used. The fingerprints utilized were GpiDAPh3, MACCS keys (166 bits), TGD, and TGT. The results are provided in Fig. S2 of the "Electronic supplementary material." The differentiation of the data points achieved with MACCS keys, TGD and TGT was much less than that obtained with ROCS (3D shape similarity method). This result is not surprising, and can be attributed to the low resolutions of these fingerprints. Interestingly, GpiDAPH3 provided similar data distribution results to ROCS.

Molecules evaluated in a dose–response manner (IC_{50})

Even though there is not always a direct correlation between the IC_{50} and binding affinity values, it has been observed over the years [5–7, 13] that it is possible to guide the selection of candidate compounds for further evaluation in a dose–response manner based on binding affinities. A subset of 32 new compounds with % inhibitions of >27.46 were chosen to measure IC_{50} values. For this analysis, molecules synthesized in previous studies [5] were incorporated into the dataset. The complete list consisted of 80 compounds and is documented in the "Electronic supplementary material." Three compounds belong to both sets; four were internal queries (see Table S2). Hence, 76 molecules were compared to the four internal queries.

Retrieving missed hits from BA

The activities obtained in a dose–response manner were examined. The BA–similarity map based on IC_{50} values is shown in Fig. 5b. It should be noted that this set contained 76 compounds, compared to 145 (excluding the four queries in each case) for the set evaluated based on % inhibition. Gray lines show the partition on the similarity axis as well as in the BA scaled score. When the synthesized molecules were considered using the traditional deconvolution method (see the "Electronic supplementary material," Fig. S3), the lowest BA scaled score among the actives was 0.75. A further partition can be suggested based on this BA score (Fig. 5a). This new partition further differentiates between "consensus hits" and "missed hits" in quadrant II. Thus, three active molecules were located in the vicinity of a normalized BA score of ~0.90. Three moderately active compounds were located between 0.80 and 0.87 in the normalized BA score, followed by an inactive compound that was located below a BA value of 0.80. This shows a good rank order based on BA score. Interestingly, one active compound (IC_{50}=309 nM) was found at a BA scaled score of 0.65. This compound, which was not chosen in the original screening, was identified as a well-behaved candidate, judging by its high similarity to the query molecules employed here. This compound was ranked on the BA scale at position 4,931 (out of 102,459). A compound with IC_{50}=993 nM was also found. This exemplifies the notion that ligands within nM activity range can be located deeper in the BA distribution, and that similarity-based analyses can assist in retrieving such compounds. Among the molecules selected based on similarity to external queries, none had an IC_{50} value of less than 1000 nM. It should be noted that only five compounds were evaluated in this category. While the likelihood of finding promising molecules through the use of external queries is undoubtedly lower, this comparison, when successful, could provide additional molecules to those selected with internal queries. Lastly, among the three molecules evaluated with BA scores of 0.8–0.9 and *rocs*/2 similarity values of >0.9 (Fig. S3), one was moderately active and the other two were inactive in terms of IC_{50}.

Summary and conclusions

In this work, we investigated the combination of experimental (mixture-based combinatorial libraries) and virtual (similarity-based) screening approaches in order to find out the extent to which both methods can be used complementarily to increase the likelihood of finding active compounds in combinatorial libraries. Test compounds were derived from a library of 102,459 bicyclic guanidines screened against the κ-opioid receptor. A subset of 149 bicycle guanidines was synthesized and tested for the κ-opioid receptor in a radioreceptor binding assay at a single concentration. A total of 32 new compounds were further selected and tested in a dose–response manner. Four criteria were established to systematically explore different relevant regions in the biometrical analysis distribution: (i) similarity to internal queries; (ii) hydroxylated analogs to the hit previously found at an earlier stage; (iii) similarity to

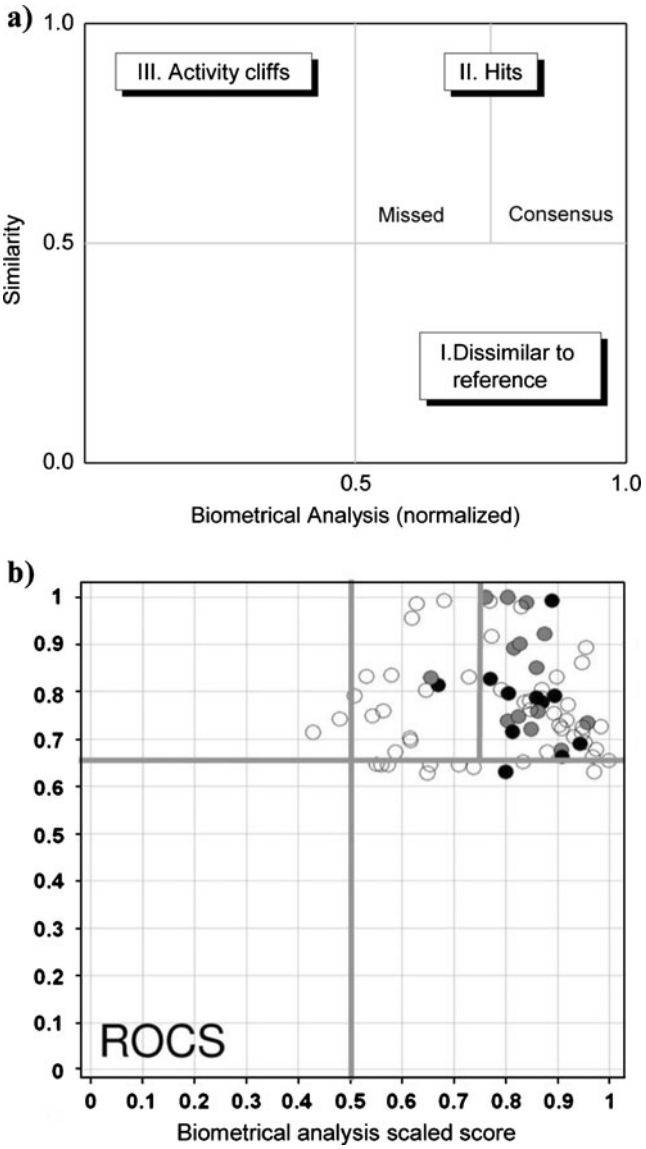

Fig. 5 **a** A schematic representation of a BA–molecular similarity map. Region I contains relatively dissimilar molecules to the reference compounds. Regions II and III contain data points with an equal likelihood of being actives but that are distinguished on the BA axis as hits or activity cliffs. Region II is further divided into consensus and missed hits. **b** The biometric analysis–molecular similarity map utilizing *rocs*/2 for 76 bicyclic guanidines. *Circles* are color coded by IC_{50} value: *black*, IC_{50}<500nM; *gray*, 1000nM<IC_{50}<500nM; *white*, IC_{50}>1000nM

external queries; (iv) random choice. Based on data on percent inhibition of activity and in silico methods, we suggest a threshold of 0.5 for the normalized biometric analysis scaled score in order to discriminate the subset of compounds with high probabilities of being active. Biometric analysis–similarity maps allowed the categorization of subsets of molecules into those with a high likelihood of being either active compounds or activity cliffs. In addition, a region containing molecules that are putatively active but structurally different from the reference compounds was also distinguished. In the "missed hits" region, we identified a new active compound (IC_{50}=309 nM) and one moderately active compound (IC_{50}=993 nM) that were not found in the first deconvolution cycle. The present approach can be extended to use scores obtained from other virtual screening approaches, such as docking or pharmacophore modeling. In turn, similarity analyses based on fragments and R groups can provide additional information for integrating virtual screening and mixture-based screening data. Analyses of other mixture-based combinatorial libraries will show the extent to which these two methodologies cross-validate and complement each other in the search for useful hits for drug optimization.

Acknowledgments This work was supported by the State of Florida, Executive Officer of the Governor's Office of Tourism, Trade and Economic Development. We thank OpenEye Scientific Software for providing the OMEGA, ROCS, and VIDA programs, Dr. Tudor Oprea for providing a subset of the Wombat database, and Kyle Kryak for assistance. The work was funded in part by the National Institutes of Health grants 5R21DA019620-02 (RAH), 5P41GM081261-03 (RAH), 3P41GM079590-03S1 (RAH), and 1F03DAO2S850-D1A1 (AN).

References

1. Fox S, Farr-Jones S, Sopchak L, Boggs A, Comley J (2004) High-throughput screening: searching for higher productivity. J Biomol Screen 9:354–358. doi:10.1177/1087057104265290
2. Goode DR, Totten RK, Heeres JT, Hergenrother PJ (2008) Identification of promiscuous small molecule activators in high-throughput enzyme activation screens. J Med Chem 51:2346–2349. doi:10.1021/jm701583b
3. Hertzberg RP, Pope AJ (2000) High-throughput screening: new technology for the 21st century. Curr Opin Chem Biol 4:445–451
4. Dolle RE (2001) Comprehensive survey of combinatorial library synthesis: 2000. J Comb Chem 3:477–517. doi:10.1021/cc010049g
5. Houghten RA, Pinilla C, Appel JR, Blondelle SE, Dooley CT, Eichler J, Nefzi A, Ostresh JM (1999) Mixture-based synthetic combinatorial libraries. J Med Chem 42:3743–3778. doi:10.1021/jm990174v
6. Houghten RA, Pinilla C, Giulianotti MA, Appel JR, Dooley CT, Nefzi A, Ostresh JM, Yu YP, Maggiora GM, Medina-Franco JL, Brunner D, Schneider J (2008) Strategies for the use of mixture-based synthetic combinatorial libraries: scaffold ranking, direct testing, in vivo, and enhanced deconvolution by computational methods. J Comb Chem 10:3–19. doi:10.1021/cc7001205
7. Pinilla C, Appel JR, Borras E, Houghten RA (2003) Advances in the use of synthetic combinatorial chemistry: mixture-based libraries. Nat Med 9:118–122. doi:10.1038/70946
8. Armishaw CJ, Singh N, Medina-Franco JL, Clark RJ, Scott KC, Houghten RA, Jensen AA (2010) A synthetic combinatorial strategy for developing alpha-conotoxin analogs as potent alpha7 nicotinic acetylcholine receptor antagonists. J Biol Chem 285:1809–1821. doi:10.1074/jbc.M109.071183
9. Reilley KJ, Giulianotti M, Dooley CT, Nefzi A, McLaughlin JP, Houghten RA (2010) Identification of two novel, potent, low-liability antinociceptive compounds from the direct in vivo screening of a large mixture-based combinatorial library. AAPS J 12:318–329. doi:10.1208/s12248-010-9191-3
10. Yongye AB, Appel JR, Giulianotti MA, Dooley CT, Medina-Franco JL, Nefzi A, Houghten RA, Martinez-Mayorga K (2009) Identification, structure–activity relationships and molecular modeling of potent triamine and piperazine opioid ligands. Biorg Med Chem 17:5583–5597. doi:10.1016/j.bmc.2009.06.026
11. Hemmer B, Gran B, Zhao YD, Marques A, Pascal J, Tzou A, Kondo T, Cortese I, Bielekova B, Straus SE, McFarland HF, Houghten R, Simon R, Pinilla C, Martin R (1999) Identification of candidate T-cell epitopes and molecular mimics in chronic Lyme disease. Nat Med 5:1375–1382. doi:10.1002/0471142735.im0905s45
12. Zhao Y, Gran B, Pinilla C, Markovic-Plese S, Hemmer B, Tzuo A, Whitney LW, Biddison WE, Martin R, Simon R (2001) Combinatorial peptide libraries and biometric score matrices permit the quantitative analysis of specific and degenerate interactions between clonotypic TCR and MHC peptide ligands. J Immunol 167:2130–2141
13. Dooley CT, Chung NN, Wilkes BC, Schiller PW, Bidlack JM, Pasternak GW, Houghten RA (1994) An all D-amino-acid opioid peptide with central analgesic activity from a combinatorial library. Science 266:2019–2022. doi:10.1126/science.7801131
14. Houghten RA, Dooley CT, Appel JR (2006) In vitro and direct in vivo testing of mixture-based combinatorial libraries for the identification of highly active and specific opiate ligands. AAPS J 8:E371–E382. doi:10.1208/aapsj080242
15. Venhorst J, ter Laak AM, Commandeur JN, Funae Y, Hiroi T, Vermeulen NP (2003) Homology modeling of rat and human cytochrome P450 2D (CYP2D) isoforms and computational rationalization of experimental ligand-binding specificities. J Med Chem 46:74–86. doi:10.1021/jm0209578
16. Brooijmans N, Kuntz ID (2003) Molecular recognition and docking algorithms. Annu Rev Biophys Biomol Struct 32:335–373. doi:10.1186/1471-2105-10-58
17. Martin YC (1992) 3D database searching in drug desing. J Med Chem 35:2145–2154. doi:10.1021/jm00090a001
18. Boehm M, Wu T-Y, Claussen H, Lemmen C (2008) Similarity searching and scaffold hopping in synthetically accessible combinatorial chemistry spaces. J Med Chem 51:2468–2480. doi:10.1021/jm0707727
19. Medina-Franco JL, Maggiora GM, Giulianotti MA, Pinilla C, Houghten RA (2007) A similarity-based data-fusion approach to the visual characterization and comparison of compound databases. Chem Biol Drug Desig 70:393–412. doi:10.1111/j.1747-0285.2007.00579.x
20. Maggiora GM (2006) On outliers and activity cliffs: why QSAR often disappoints. J Chem Inf Model 46:1535–1535. doi:10.1021/ci060117s
21. Bajorath J (2002) Integration of virtual and high-throughput screening. Nat Rev Drug Discov 1:882. doi:10.1038/nrd941
22. Johnson MA, Maggiora GM (1990) Concepts and applications of molecular similarity. Wiley, New York
23. Nikolova N, Jaworska J (2003) Approaches to measure chemical similarity—a review. QSAR Comb Sci 22:1006–1026. doi:10.1186/1471-2121-8-S1-S6

24. Medina-Franco JL, Martinez-Mayorga K, Bender A, Mari'n RM, Giulianotti MA, Pinilla C, Houghten RA (2009) Characterization of activity landscapes using 2D and 3D similarity methods: consensus activity cliffs. J Chem Inf Model 49:477–491. doi:10.1021/ci800379q
25. Martinez-Mayorga K, Medina-Franco JL, Giulianotti MA, Pinilla C, Dooley CT, Appel JR, Houghten RA (2008) Conformation–opioid activity relationships of bicyclic guanidines from 3D similarity analysis. Bioorg Med Chem 16:5932–5938. doi:10.1016/j.bmc.2008.04.061
26. Smith JAM, Hunter JC, Hill RG, Hughes J (1989) A kinetic analysis of κ-opioid agonist binding using the selective radioligand [^{3}H]U69593. J Neurochem 53:27–36. doi:10.1111/j.1471-4159.1989.tb07291.x
27. Olah M, Mracec M, Ostopovici L, Rad R, Bora A, Hadaruga N, Olah I, Banda M, Simon Z, Mracec M (2004) WOMBAT: world of molecular bioactivity. In: Oprea TI (ed) Chemoinformatics in drug discovery. Wiley-VCH, New York, pp 223–239
28. OpenEye Scientific Software (2007) ROCS v.2.3.1. OpenEye Scientific Software, Santa Fe (see http://www.eyesopen.com)
29. OpenEye Scientific Software (2007) OMEGA v.2.2.1. OpenEye Scientific Software, Santa Fe (http://www.eyesopen.com)
30. Ostresh JM, Schoner CC, Hamashin VT, Nefzi A, Meyer JP, Houghten RA (1998) Solid-phase synthesis of trisubstituted bicyclic guanidines via cyclization of reduced *N*-acylated dipeptides. J Org Chem 63:8622–8623. doi:10.1208/aapsj080242
31. Sykes MJ, McKinnon RA, Miners JO (2008) Prediction of metabolism by cytochrome P4502C9: alignment and docking studies of a validated database of substrates. J Med Chem 51:780–791. doi:10.1021/jm7009793
32. Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L (2005) The use of receiver operating characteristic curves in biomedical informatics. J Biomed Inform 38:404–415. doi:10.1186/1471-2105-8-331
33. Triballeau N, Acher F, Brabet I, Pin J-P, Bertrand H-O (2005) Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. J Med Chem 48:2534–2547. doi:10.1021/ci800101j
34. Shanmugasundaram V, Maggiora GM (2001) Characterizing property and activity landscapes using an information-theoretic approach. Abstr Pap Am Chem Soc 222:32-CINF
35. Patterson DE, Cramer RD, Ferguson AM, Clark RD, Weinberger LE (1996) Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. J Med Chem 39:3049–3059. doi:10.1021/ci025635r
36. Whittle M, Gillet VJ, Willett P, Loesel J (2006) Analysis of data fusion methods in virtual screening: similarity and group fusion. J Chem Inf Model 46:2206–2219. doi:10.1016/S1367-5931(00)00110-1
37. Whittle M, Gillet VJ, Willett P, Loesel J (2006) Analysis of data fusion methods in virtual screening: theoretical model. J Chem Inf Model 46:2193–2205. doi:10.1016/S1367-5931(00)00110-1