J Comput Aided Mol Des (2009) 23:725–736
DOI 10.1007/s10822-009-9297-9

Combinatorial library-based design with Basis Products

Joe Zhongxiang Zhou · Shenghua Shi · Jim Na · Zhengwei Peng · Tom Thacher

Received: 29 January 2009 / Accepted: 26 June 2009 / Published online: 11 July 2009

Abstract Uncovering useful lead compounds from a vast virtual library of synthesizable compounds continues to be of tremendous interest to pharmaceutical researchers. Here we present the concept of Basis Products (BPs), a new and broadly applicable method for achieving efficient selections from a combinatorial library. By definition, Basis Products are a strategically selected subset of compounds from a potentially very large combinatorial library, and any compound in a combinatorial library can represented by its BPs. In this article we will show how to use BP docking scores to find the top compounds of a combinatorial library. Compared with the brute-force docking of an entire virtual library, docking with BPs are much more efficient because of the substantial size reduction, saving both time and resources. We will also demonstrate how BPs can be used for property-based combinatorial library designs. Furthermore, BPs can also be considered as fragments carrying chemistry knowledge, hence they can potentially be used in combination with any fragment-based design method. Therefore, BPs can be used to integrate combinatorial design with structure-based design and/or fragment-based design. Other potential applications of BPs include lead hopping and consensus core building, which we will describe briefly as well in this report.

Keywords Basis Product · Fragment-based · Structural-based · Combinatorial design

Introduction

Lead optimization in drug discovery is a complex process. Many properties besides binding affinity have to be optimized for a lead series to produce compounds of an acceptable pharmacologic profile of a drug. Over time, this optimization process has evolved. Traditionally, individual properties of a lead series are sequentially optimized; this method has been gradually replaced by parallel and multi-property optimization strategies during potency optimization [1]. Today, optimization of ADME-Tox (absorption, distribution, metabolism, excretion, and toxicity) properties are routinely carried out in the early stages of the drug discovery process. This paradigm shift is driven by the cost analyses of historical drug discovery data and powered by new developments in various high-throughput technologies [1–3]. The discourse on "drug likeness" has helped this paradigm shift when researchers realized that starting with a "drug-like" lead can potentially be very cost-effective for a drug discovery program [4–6]. Also helpful is the recent discourse on ligand efficiency (LE), which calculates the binding affinity per heavy atom of a given compound [7, 8]. Ligand efficient leads allow more room for multi-property optimization as they tend to be smaller in molecular weight, and would therefore allow molecular tuning to attenuate various physico-chemical properties. By taking into account both drug-likeness and LE, the process of drug discovery for a given program may be greatly accelerated.

J. Z. Zhou (✉) · S. Shi · J. Na · Z. Peng · T. Thacher
Department of Computational and Structural Biology, Pfizer Global Research and Development, La Jolla Laboratories, 10777 Science Center Drive, San Diego, CA 92121, USA
e-mail: zjoe.zhou@gmail.com

Present Address:
S. Shi
Plexxikon Inc., 91 Bolivar Drive, Berkeley, CA 94710, USA

T. Thacher
Virtual Chemistry Inc., 12110 Salix Way, San Diego, CA 92129, USA

Fragment-based drug design (FBDD) methods that have been developed in the recent decade have the potential to provide very LE drug-like leads [9–15]. These methods typically involves biophysical techniques such as NMR to screen a set of fragments, producing hits with activity typically in micromolar to millimolar range. Alternatively, the screening can also be done computationally with various docking programs if the target structure is known. With low affinity hits, various analytical techniques are applied to ascertain the location of binding in the different sub-pockets of the active site of a given target. Depending on the outcome, either the fragment linking method or the fragment growing method is applied to increase potency. In the linking method, the fragments sitting in sub-pockets of an active site are linked via a chemical spacer that preserves the binding conformation of each fragment. In the growing method, the bound fragment "grows" into larger molecules to fill the binding site. The first successful applications of these fragment design methods are the designs of high affinity inhibitors for the FK506 binding protein (FKBP) and for the matrix metalloproteinase stromelysin (MMP-3) by a group in Abbott Laboratories, using a method called "SAR by NMR" [9, 10]. Since then, NMR and other FBDD methodologies have been successfully applied in various discovery projects by different pharmaceutical companies [13–15].

Before FBDD, combinatorial chemistry, catalyzed by the High Throughput Screening (HTS) technologies, has been developed into a powerful tool for drug discovery [16–19]. It has been widely used in lead discovery and lead generation, and has been integrated into medicinal chemistry as a powerful tool for lead optimization [19, 20]. The evolution of combinatorial chemistry has gone from the early days of an inexact science equivalent to throwing darts in dark to a much more target-focused discipline. This is evident in terms of chemical diversity being derived from smaller libraries representing more chemical reactions than from larger libraries made from fewer reactions, and also from the fact that target-oriented focused libraries now play a much more important role in drug discovery today than before. The molecular complexity is now more geared toward specific targets or target groups [11], and the role of chemical diversity has shifted to covering chemical space rather than searching for or achieving molecular complexity of a given set of compounds. And even though it is now possible to screen a large number of compounds against any given target for biological activities the number of compounds that can be physically synthesized is always only a miniscule portion of the entire combinatorial libraries. Therefore, to realize the full potential of a given combinatorial library, efficient in-silico screening methods to filter virtual libraries prior to syntheses has become a necessary drug discovery tool.

Depending on the chemistry involved, combinatorial libraries can be very large, with virtual libraries ranging from thousands to 10^8–10^{12} possible compounds. How to select the "right" compounds to synthesize then becomes very important part of drug discovery. Typically, general filters such as the Rule-of-5 or some other ADME-Tox related criteria derived from empirical data are applied [4–6, 21–24]. In addition, specific information related to the target can be used for further filtering when such information is gathered in the process of lead discovery. For instance, pharmacophore models can be built based on actives and inactives from a biochemistry screening assay such as an HTS run. Docking virtual libraries, when the target structure is known, can also be performed to filter large libraries. However, it is prohibitively expensive to use these high-level filtering tools even with the help of some initial filtering because of the enormous size of some combinatorial libraries.

While docking entire virtual libraries is not always feasible even with today's computational speed and resources, we believe we have a solution that can avoid costly computational expense while still able to adequately sample the molecular space of a given combinatorial library. By definition combichem derived compounds share molecular fragments with other compounds in the library, and a given molecular fragment should behave similarly in various molecular context (the SAR assumption). The high redundancy of fragments in combinatorial libraries is the basis for the efficiency of the method to be developed here. In the following sections, we will introduce a new and broadly applicable concept called the Basis Product (BP) method which can be applied for efficient filtering of large combinatorial libraries. We will show that this method can produce results comparable with the results produced from calculations with the full combinatorial library in a structure-based library design application. For property-based combinatorial library designs, we will derive (or re-derive) some key equations in order to gain additional insight about their applicability (see Ref. [25]). And lastly, we will also comment on other useful potential applications of the BP method to be published elsewhere.

Structure-based virtual screening of combinatorial libraries without docking the whole libraries has been explored by several groups before, majority of them by taking advantage of the array nature of the libraries [26–38]. Examples are CombiDOCK and DREAM++ by Kuntz and coworkers [27, 28], the VLSprout algorithm by Johnson and coworkers [29], Core-template-based method by Leach et al.[30], OptiDock by Sprout et al.[31], pharmacophore-combinatorial docking combination by Gastreich et al.[32], FLEXX^{C} by Rarey and Lengauer [33], and FLEXNOVO by Degen and Rarey [33]. Other efficient docking methods, such as LibDock by Diller and Merz [35, 36] and TrixX by Schellhammer and Rarey [37], can also be used for structure-based virtual screening of combinatorial libraries. Nevertheless, both LibDock and TrixX need full enumeration of combinatorial libraries, which may prove prohibitive. For property-based or similarity-based virtual screening, methods were developed to avoid full library enumeration [25, 39].

Before describing our method we would like to point out that the BP method is an integral part of a broader initiative called PGVL (Pfizer Global Virtual Library), an effort headed by Dr. Atsuo Kuki at Pfizer. Other components of PGVL will be discussed and presented in other publications [19, 40, 41].

Basis Products

Interactions important to the binding between a potential drug molecule and the target active site are derived from specific functional groups of the given compound. For a combinatorial library to produce a high potency compound the functional groups essential to the potency have to be in the R-groups of the reactants or formed in the product core. In a combichem library, all the functional groups present in the library can be completely represented by a selected product subset of the library. This strategically selected subset of compounds is called the Basis Products:

> *Basis Products* – Basis Products (BPs) are formed by combining the smallest reactants of all reaction components except one. This remaining reactant is then swept across all viable reactants for that reaction component while holding the other reaction components fixed, to provide a methodical sampling subset of the entire combinatorial library.

The pre-selected, fixed reactants for BP formation are also called capping reactants.

Put differently, for a two component reaction A + B → AB, the complete combinatorial array will be the entire set of {AB} with all possible combinations of reactants A and B. (Herein simplified notations will be used: "A" and "B" are used both as running indices and as reactants they represent while "AB" is used both as running index and as products it represents.) The BPs contain only $\{A_sB\}$ and $\{AB_s\}$ with the capping molecules A_s and B_s being pre-selected and preferably the smallest A and B respectively. Since A and B are both changing in the single set {AB} the number of compounds in it is much bigger than the sum of numbers of compounds in the double sets of $\{A_sB\}$ and $\{AB_s\}$, each with only one changing component. The number of virtual products in the BP library is $N_a + N_b - 1$ while the size of the virtual library is $N_a \times N_{b,}$ where N_a and N_b are the numbers of viable A and B available respectively. (The minus one is to correct the double counting of A_sB_s.)

Figure 1 and Table 1 illustrate how the BPs are selected from a combinatorial library. Figure 1 uses the reductive amination reaction as an example to illustrate the size reduction [42]. The number of amines pre-filtered for chemistry compatibility from ACD (Available Chemical Directory, 2007 edition) is 2,550, and the total of the compatible ketones and aldehydes equals 2,380. Complete cross reactions of these reactants will produce a virtual library of 6,069,000 potential products. For BPs, one pre-selected amine reacting with all aldehydes and ketones

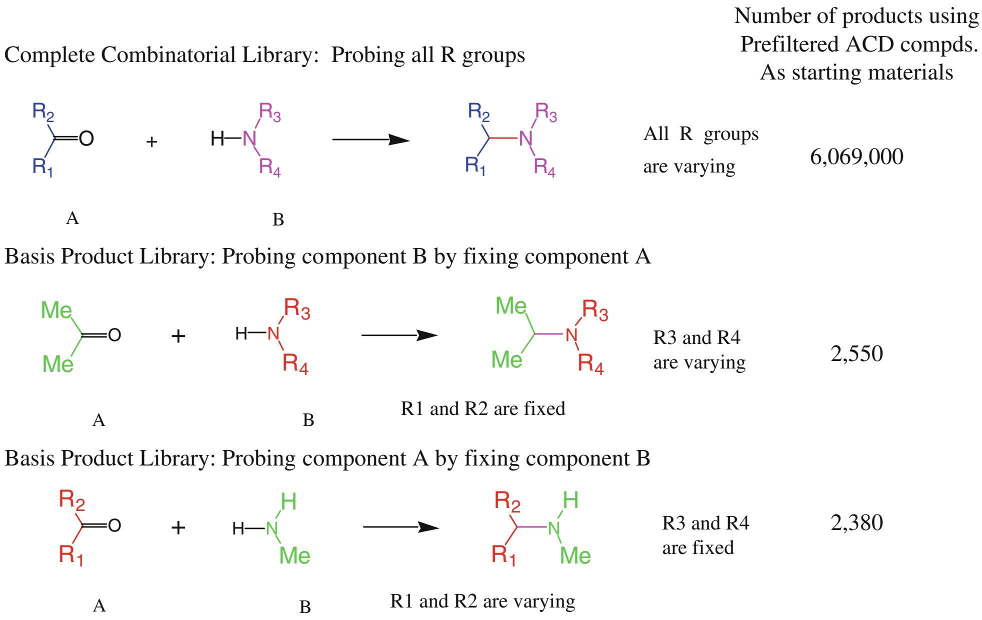
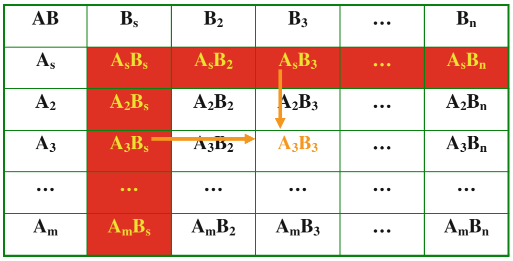


Fig. 1 Complete combinatorial library versus Basis Product library for a reductive amination reaction. The green portion of the molecules are pre-selected and fixed

Table 1 Basis Products (BPs)—strategically selected to cover building block space of a combinatorial library for: A + B → AB. red cells are BPs. Any product in this library either is a BP or can be represented by two BPs as illustrated by the golden arrows: A_3B_3 can be represented by $\{A_3B_s, A_sB_3\}$. A_s and B_s are the pre-selected capping reactants



AB	B_s	B_2	B_3	…	B_n
A_s	A_sB_s	A_sB_2	A_sB_3	…	A_sB_n
A_2	A_2B_s	A_2B_2	A_2B_3	…	A_2B_n
A_3	A_3B_s	A_3B_2	A_3B_3	…	A_3B_n
…	…	…	…	…	…
A_m	A_mB_s	A_mB_2	A_mB_3	…	A_mB_n

would give 2,550 BPs; and one pre-selected ketone reacting with all compatible amines would yield 2,380 BPs. Adding these two sub-libraries together forms the complete Basis Product library of 4,929 products, a much smaller value than the complete virtual library. (Note that the product from the capping reactants appear in both BP sub-libraries.) Furthermore, this smaller library represents the entire virtual library in terms of chemical diversity since it contains all the R-groups of both components.

Table 1 is a matrix representation of the selection process for BPs. It then follows that for a three-component reaction, three sub-libraries and its sum form the BP library, and the sum of four sub-libraries represent the BPs of a 4-component reaction, etc. In the reactions where the number of components is greater than two, in each sub-library one component is variable while the remaining components are pre-selected and fixed (the capping reactants). So for a three component reaction, the BPs of component A is the library of all variable component As with fixed components B and C, and BPs of component B is the library of all variable component Bs with fixed components A and C, etc. From the examples above, one can see that the number of BPs for any 2-, 3-, or 4-component reactions will be the sum of numbers of all the viable reactants of that reaction (less the multiple counts of the BP derived from all capping reactants). However, the size reduction from entire virtual libraries to BPs is more dramatic for 3- and 4-component reactions than for 2-component reactions (see Table 2).

The set of BPs is complete in the sense that any single virtual product in the combinatorial library can be represented by its BPs. For example, virtual product A_3B_3 can be represented by BPs A_sB_3 and A_3B_s as shown in Table 1. All chemical diversity represented by the R-groups in the virtual library are represented in the BP library since all the functional groups necessarily come from one of the reactants or from the newly formed product core.

Table 2 A snapshot of sizes of virtual libraries covering a broad set of over 500 Pfizer reactions: basis product library versus Pfizer global virtual library (as of March 2008)

Starting material	# Of Basis Products	# Of virtual products
In-house inventory		
2 component RXN	840,589	414,624,233
3 component RXN	2,243,608	7.666×10^{11}
4 component RXN	265,271	1.172×10^{13}
Total	3,349,468	1.249×10^{13}
Reagents from available chemical directory of MDL		
2 component RXN	4,542,321	2.539×10^{10}
3 component RXN	9,286,074	1.530×10^{14}
4 component RXN	1,102,339	3.263×10^{16}
Total	14,930,734	3.278×10^{16}

Pfizer has developed combinatorial synthetic protocols for over 500 different chemical reactions [40], ranging from two to four component reactions. Table 2 shows a snapshot of the library sizes for the Pfizer Global Virtual Libraries, and also illustrates the enormous size differences between the virtual libraries and the BP libraries. At Pfizer the BPs are used to tap into the vast product space of PGVL without actual enumeration of the entire virtual product space. Figure 2 illustrates how the BPs serve as indices to the PGVL products.

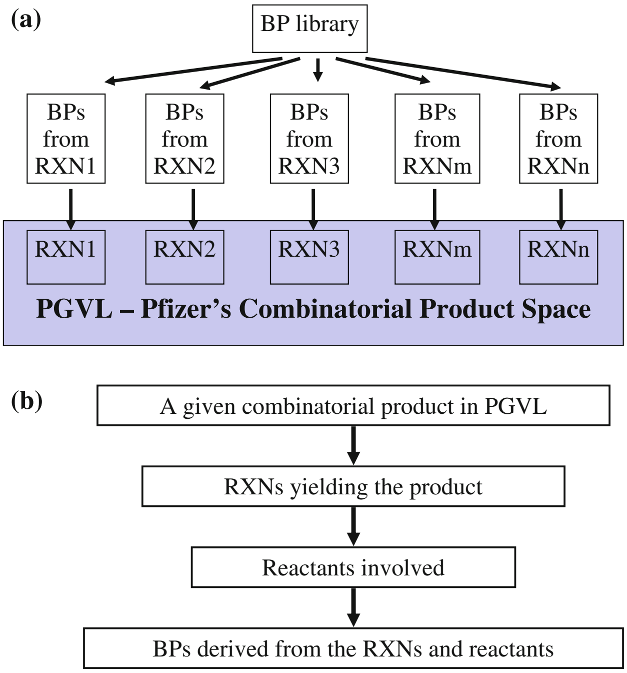

(a)
BP library
BPs from RXN1
BPs from RXN2
BPs from RXN3
BPs from RXNm
BPs from RXNn
RXN1
RXN2
RXN3
RXNm
RXNn
PGVL – Pfizer's Combinatorial Product Space
(b)
A given combinatorial product in PGVL
RXNs yielding the product
Reactants involved
BPs derived from the RXNs and reactants

Fig. 2 **a** Hierarchical indexing of combinatorial products using BPs as annotated basis set. **b** BP representations of combinatorial products

By using similar filtering criteria for fragment libraries, BPs can also be used for fragment-based design [9–15, 43–45]. The Basis Product library can be viewed as a fragment library, and it possesses one important feature that distinguishes it from other fragment libraries. Due to its combinatorial origin, BPs have chemistry knowledge encoded within, hence any BP hit can be easily followed up with a combinatorial library design.

Structure-based combinatorial library design with Basis Products

Method

Typically the docking process is the bottleneck for virtual screening of combinatorial libraries in an SBDD effort. However, exhaustive docking can be avoided based on the following arguments. For a given ligand bound to the active site of a receptor, the energetic contribution from the portions of a ligand contributing to the binding affinity are typically additive. When fully optimized, these portions will bind to the target more tightly than the corresponding portion in the original ligand. Therefore, for a given combinatorial product to have sufficient binding affinity to a certain target, the portions of the product, which can be thought of as fragments, also have to be sufficiently potent. This implicit premise of all fragment-based design holds true for enthalpy-driven systems [46, 47].

Figure 3 schematically illustrates a typical flowchart for structure-based filtering of virtual combinatorial libraries using BPs. Given a target protein structure, selected BPs from a BP library can be docked just like any other compound library. However, in some cases even the BP library may be too large for full docking, hence pre-dock filtering becomes a necessary procedure. These filters can be either property-based or chemistry-based, and typically this filtering process reduces the BP library to be docked to a more reasonable size of around 500 K compounds.

Once the BP library is reduced to a reasonable size, various docking programs can be used for the docking and scoring process [24, 26, 48–50]. The BPs with the highest scores are then selected for the hit follow-up process. This usually involves designing libraries based on the BPs, which are effectively enabled due to their combinatorial chemistry origin.

Once the docking is complete, BP hits need to be resolved into individual reactions that created them. Note that a BP can be formed from two distinct reactions. For example, reductive amination and amine alkylation can form the exact same products from a fixed amine and an aldehyde and an alkylating agent with the exact same R-group. However, this is uncommon and most BPs are derived from a single reaction.

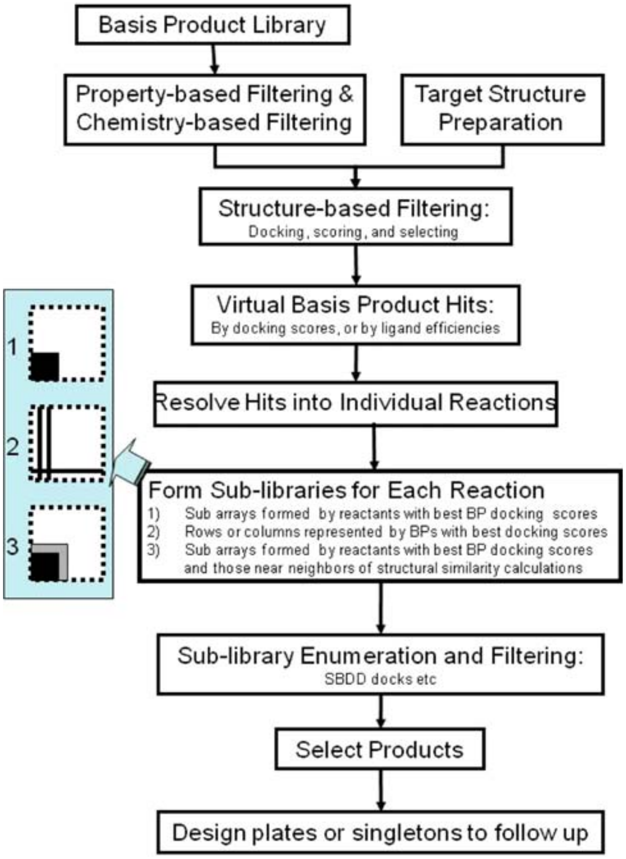

Basis Product Library
Property-based Filtering & Chemistry-based Filtering
Target Structure Preparation
Structure-based Filtering:
Docking, scoring, and selecting
Virtual Basis Product Hits:
By docking scores, or by ligand efficiencies
Resolve Hits into Individual Reactions
Form Sub-libraries for Each Reaction
1) Sub arrays formed by reactants with best BP docking scores
2) Rows or columns represented by BPs with best docking scores
3) Sub arrays formed by reactants with best BP docking scores and those near neighbors of structural similarity calculations
Sub-library Enumeration and Filtering:
SBDD docks etc
Select Products
Design plates or singletons to follow up
1
2
3

Fig. 3 A typical flowchart for structure-based virtual screening of combinatorial library using Basis Products. The graphics at the middle left illustrate graphically the 3 ways the sub-arrays are formed

For a given reaction, there are several strategies to form the sub-libraries for a given set of BP hits:

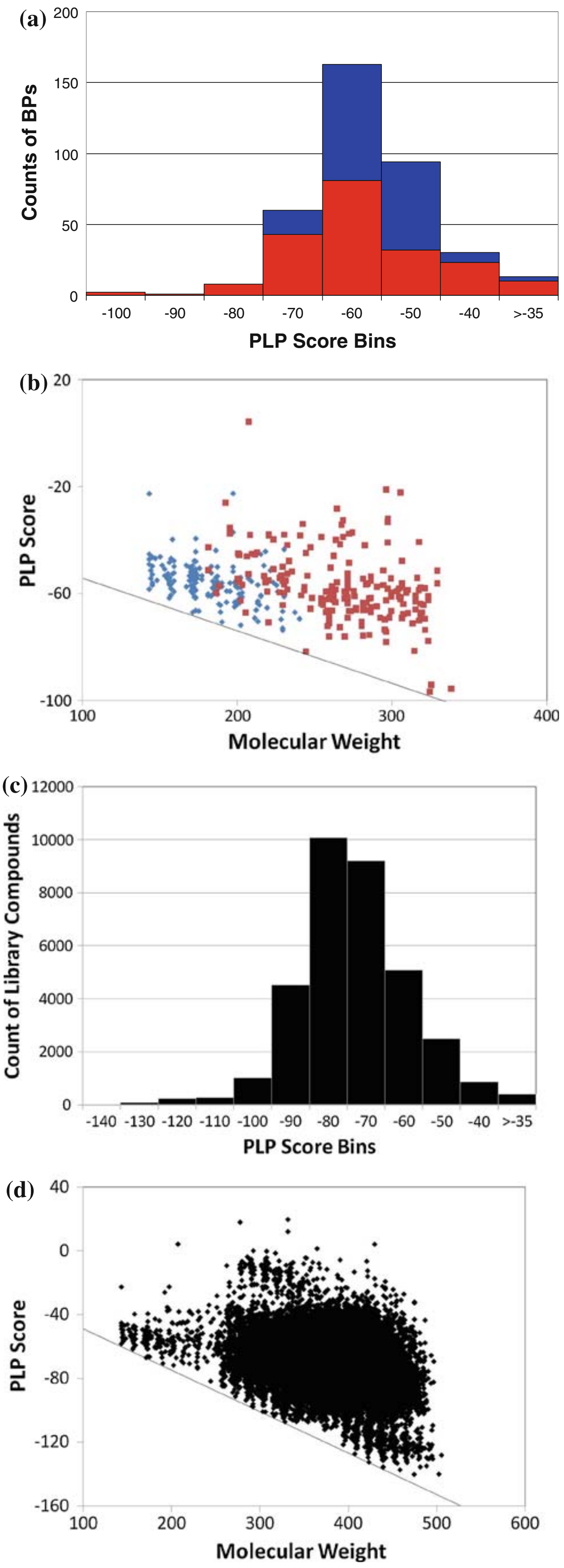
(1) Sub-arrays formed by the reactants corresponding to the variable components of the BP hits;
(2) Rows or columns represented by the BP hits;
(3) Sub-arrays formed by the reactants corresponding to the BP hits and their nearest neighbors in the similarity space.

The diagram in the middle left portion of Fig. 3 illustrates the ways in which the sub-arrays are formed. The virtual libraries generated by these sub-arrays can be sent directly to library production, or they can be re-docked into the protein binding site for further iterations of docking and filtering.

Test case

To validate the utility of BPs for structure-based virtual screening, we docked an entire virtual library into the

Fig. 4 **a** Histogram of PLP scores for the 371 BP compounds (the BP from both capping reactants is counted twice). The red portions are from contributions of the acid component while the *blues* are from amine component. **b** PLP score versus molecular weight for 371 BP compounds. *Red squares* are for acid component and *blue diamonds* are for amine component. **c** Histogram of PLP scores for 34,200 compounds in the complete library. **d** PLP score versus molecular weight for 34,200 compounds in the complete library



active site of dihydrofolate reductase (DHFR) [51]. The target structure used for this study is "3dfr" which we retrieved from the publically available PDB database. The ligand molecule, methotrexate (MTX), was extracted from the structure. Both NDP and water molecules were retained, and the pocket occupied by MTX defined the active pocket for docking. The combinatorial library came from the following amide-bond formation reaction:

$$R_1COOH + R_2R_3NH \rightarrow R_1(CO)NR_2R_3 + H_2O \qquad (1)$$

This reaction was chosen in order to include MTX, which has the same product core, into our library. A total of 200 acids and 171 amines were selected from the Available Chemical Directory of MDL for the library [52], and the capping reactants chosen for BP formation were acetic acid and methylamine. The entire virtual library of 34,200 compounds were enumerated and docked into the active pocket of the prepared DHFR structure. Note that the number of BPs is 370 instead of 371: the BP from acetic acid and methylamine appears in both BP sub-libraries.

Pfizer's in-house docking program, AGDOCK, and the corresponding PLP scoring function were used for the docking studies [53, 54]. Another in-house program was used to calculate the binding free energies and design energies for the various docked poses (P Rose, DK Gehlhaar et al., unpublished). The design energy is the binding free energy per heavy atom with a constant cratic entropic correction for translational and rotational degrees of freedom of the whole molecule, and is similar to the concept of ligand efficiency.

Ten flexible-ligand docks were performed for each molecule. The PLP score, calculated binding free energy, and design energy derived from the calculated binding free energy were computed for each pose. The pose with the best calculated binding free energy was chosen to represent the compound. All compounds were then ranked by the PLP score, the binding free energy, or the design energy, and the top ranked compounds were selected as the virtual hits. The docking results are shown in Figs. 4, 5, 6, and 7.

Test results and discussion

Figure 4a shows the histogram of the PLP docking scores for the 371 BP compounds (the BP from the capping reactants was counted twice). The *x*-axis indicates the docking score bins while the *y*-axis shows the counts of

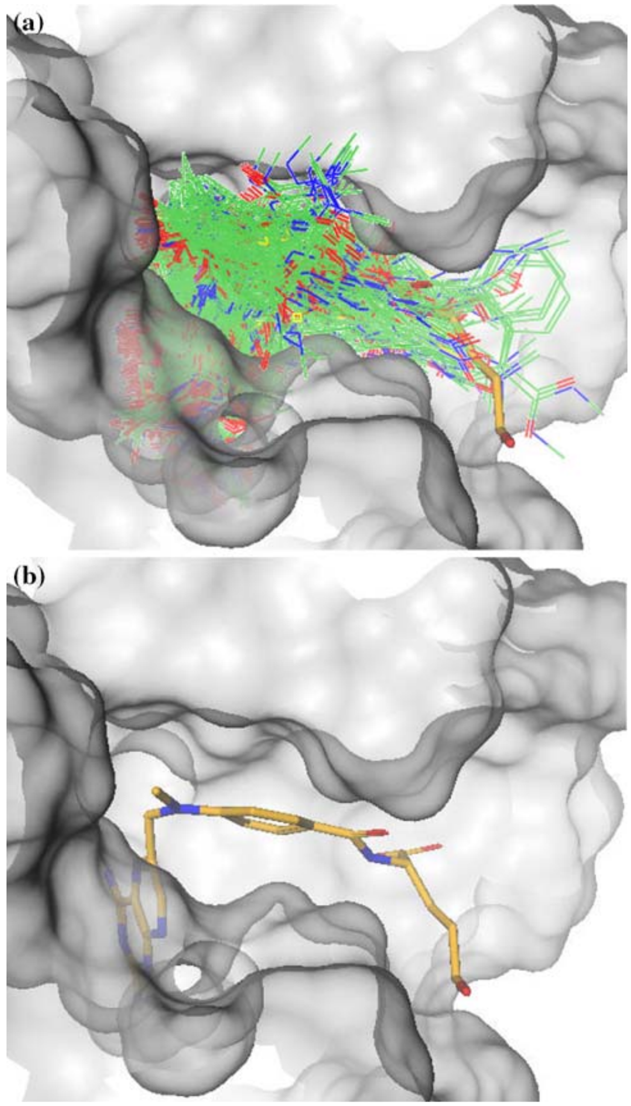


Fig. 5 **a** Docked poses of all 370 BP compounds at the active site of dihydrofolate reductase (DHFR). The *shade* is the connolly surface of DHFR near the active site as defined by the methotretate (MTX) in the crystal structure of MTX-DHFR complex [46]. The *thick yellow* molecule is the MTX in its crystal position. **b** The same MTX as in (**a**) without the BP compounds

BPs within each bin. Each count is further divided into two parts: the red color represents the number of BPs from the acid component, and the blue represents the amine component BPs. The PLP scores from the acid component ranged from −96.8 to 4.2 while those from the amine component ranged from −73.8 to −22.6. The difference in PLP score ranges for the two sets of BPs can be partially attributed to their differences in the molecular sizes. As shown in Fig. 4b, the amine BPs are clearly smaller overall than the acid BPs; the molecular weight range for acid BPs is 181–338 while the range for amine BPs is 143–297. In theory, larger ligands tend to bind more tightly to

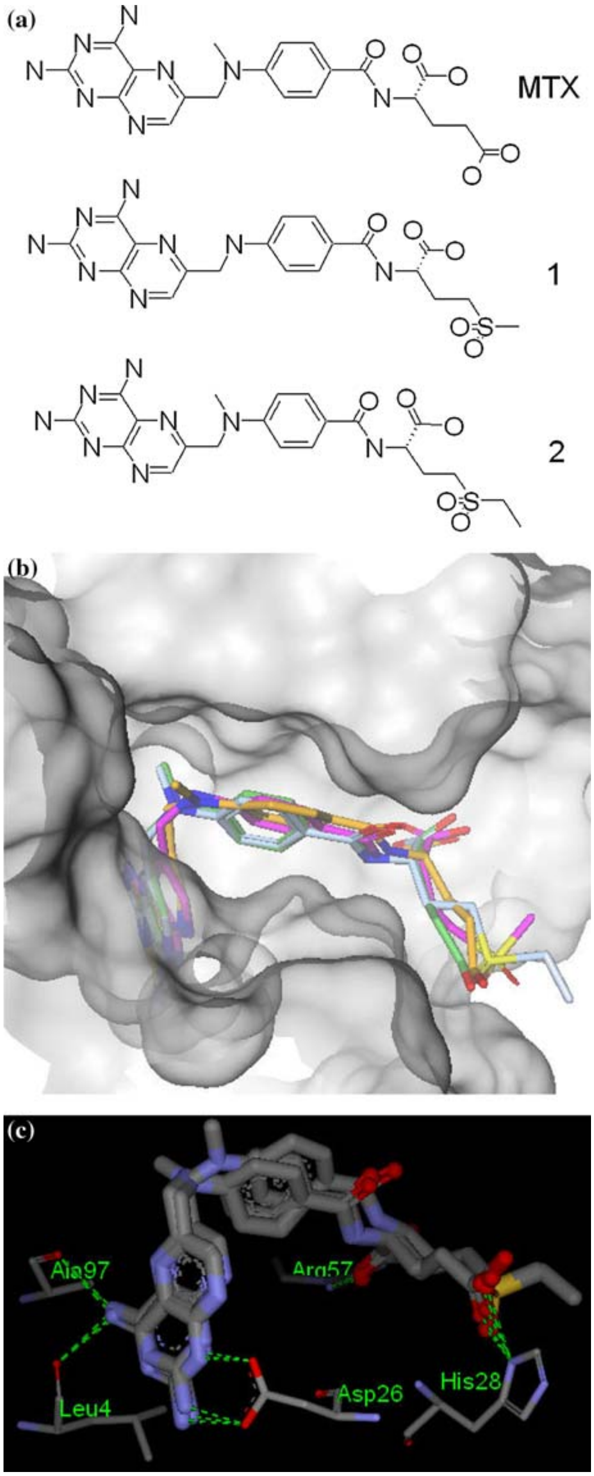


Fig. 6 **a** Structures of methotretate (MTX) and two other compounds selected based on BP method. **b** At the dihydrofolate reductase (DHFR) active site: crystal structure of MTX (*golden*), the best docked pose of MTX (*green*), and the best docked poses of the other two compounds **1** and **2** selected by BP method. **c** All docked poses preserve the essential interactions seen in the crystal structure. Shown here are hydrogen bonds between docked compounds (MTX 1 and 2) and 5 residues from DHFR: Leu 4, Asp 26, His 28, Arg 57, and Ala 97

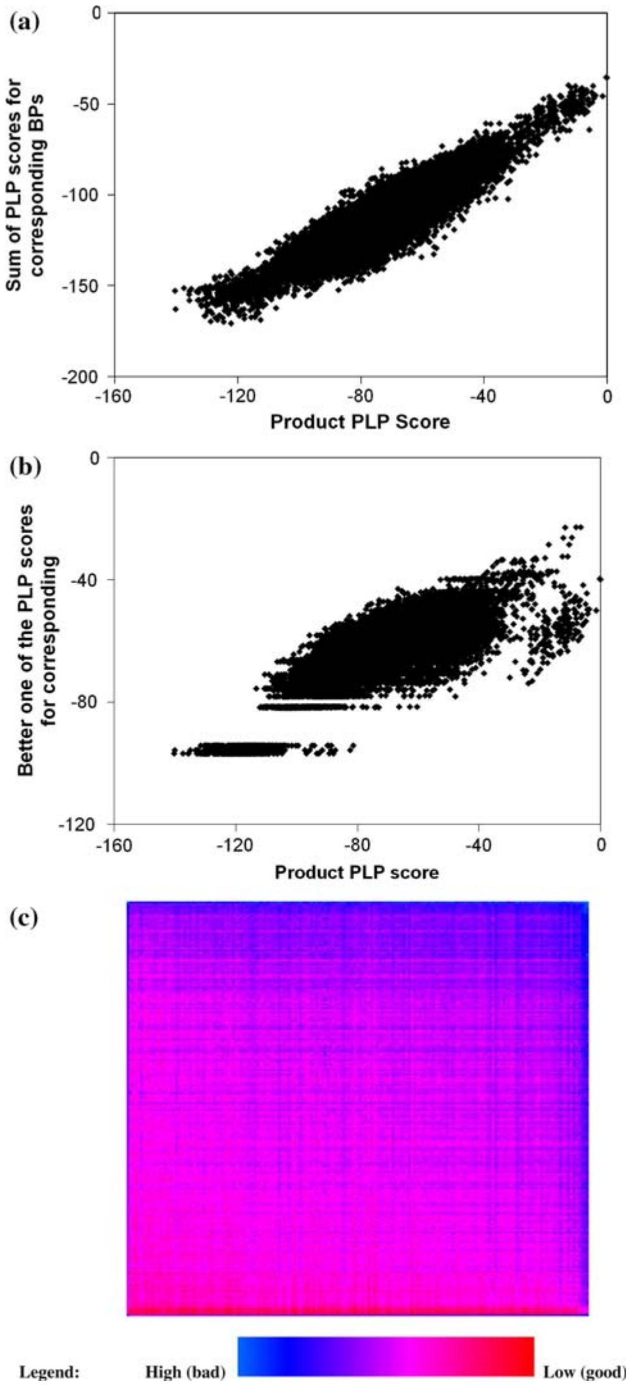


Fig. 7 PLP scores of docked results for a combinatorial library of 34,200 products from an amide bond formation reaction: **a** Product PLP score directly from docking versus PLP score estimated using sum of PLP scores of representing BPs. **b** Product PLP score directly from docking versus PLP score estimated using the best of PLP scores of the representing BPs. **c** Heatmap of product PLP scores. For the heatmap (**c**), the *x*-axis is for the amine component and the *y*-axis is for the acid component. The reactants are sorted by their BP PLP scores from low to high values (*good to bad*), except the first compound. The first *x* and the first *y* are capping compounds for BP formation. The color in each cell represents the value of the PLP score of the library product formed from the corresponding *x* reactant and *y* reactant. The cells in the first row and the first column are for BP compounds

the protein. The solid line in Fig. 4b illustrates that the ideal binders have PLP scores proportional to molecular sizes.

Figure 5a shows all of the BP docking poses inside the active site of DHFR. As a point of reference, MTX in its crystal structure position is shown as the golden-colored structure, and again by itself in Fig. 5b. As Fig. 5a shows, majority of the BPs were docked into or close to the deep pocket of the active site. Only a few BPs docked into the shallow pocket occupied by the carboxylic acid tail of the MTX (the right portion of the active site in Fig. 5a). The MTX structure is shown in Fig. 6a.

Figure 4c shows the histogram of the PLP docking scores for the 34,200 compounds in the entire virtual library. The PLP scores for this library range from −140.3 to 19.6. Note that MTX has a docking score of −132.7, making it the 10th highest ranked compound in the library. The highest ranked compound (compound **1**) has a structure very similar to MTX, and their structures are shown in Fig. 6a.

Figure 4d shows the plot of the PLP docking score against the molecular weight for the full library. Again, the solid line shows that the ideal PLP score for a given molecular weight is proportional to the molecular size. Here we also begin to see the usefulness of the BP method: the PLP scores of the two BPs representing compound **1** ranked #1 for the acid component and #89 for the amine component while the PLP scores for the BPs representing MTX ranked #2 for the acid component and #62 for the amine component.

One of our main goals for the test case study is to compare the docking results of the full combinatorial library against the docking scores of its BP library. To begin we will select the top scoring hits from the BP library. There are two methods for hit selection based on PLP scores: the first method is to select the top scoring BPs from the entire BP library, regardless of which component it comes from (acid or amine in this case); the second method is to select a set of top ranked BP compounds per reaction component. Since the hit follow-up strategies are different for each method (see "Method"), we will describe below how these strategies apply for each of the hit selection methods.

For the first method, we will form sub-libraries represented by the best BP compounds. From the 370 BPs we chose the top 5 BPs based on the PLP score, which in this case all came from the acid component. Next, we form the sub-library represented by these 5 BPs, which is done by removing their capping group and replacing it with the full reactant array. Since all 5 selected BPs came from capping the acid component, the sub-library is then formed by reacting the 5 corresponding acids with all 171 amines, yielding 855 fully combinatorial products.

Table 3 Potential speedups of the BP method over traditional brute force methods when applied to virtual screening of combinatorial libraries

	Size of sub-library needs to be screened	Size of parent library	Speedup[a]
Test case shown in sections "Test case" and "Test results and discussion"			
First follow-up scenario	1,225	34,200	27.92
Second follow-up scenario	770	34,200	44.42
Screening of complete PGVL[b]			
Library with in-house reagents	3,349,468	1.249×10^{13}	3.73×10^{6}
Library with commercial reagents	14,930,734	3.278×10^{16}	2.20×10^{9}

[a] Speedup = (Size of parent library)/(Size of sub-library needs to be screened)

[b] The numbers in second and third columns are from Table 2. The cost of follow-ups is neglected here. This neglect is partially compensated by the fact that BPs are usually much smaller molecules than compounds they represent, hence a much smaller cost is associated with screening BPs than the same number of compounds from the parent library

The next step is to repeat the docking study with the sub-library. From these results we found that the top ranked compound was again compound **1**. Its docked pose and its interactions with DHFR are shown in Fig. 6b, c. These results indicate that we were able to find the top ranked compound from the full array of 34,200 simply by docking 1,225, or the sum of its BP library (370) and the hit follow-up library (855). The sub-library docking also re-confirmed that MTX was the 10th ranked compound. In fact, the top 350 ranked compounds from the sub-library were identical to the top ranked compounds in the original full combinatorial library.

The second method of hit follow-up selects top ranked BP hits from the acid and the amine components, for a total of 40 BPs. Selecting the top 20 BPs gave a PLP cutoff score of −70.9 for the acid component and −64.3 for the amine component. We then enumerated the uncapped amines and acids as a fully combinatorial library to form a 400-compound sub-library (20 uncapped acids reacting with 20 uncapped amines), and submitted the library for docking. The results showed that the best docked compound from this follow-up sub-library was compound **2**, as shown in Fig. 6a. The PLP score for this compound was −140.1, and it was the second highest ranked compound in the original library.

In this hit follow-up method we docked a total of 770 compounds (370 BPs, 400 from the sub-library) and found the second highest ranked compound from the original library, a compound very similar to the highest ranked compound from the full library (see Fig. 6). The PLP scores of the BPs representing compound **2** ranked #2 for the acid component and #9 for the amine component. Note that MTX was not present in this sub-library because the amine monomer representing MTX was not among the top 20 amine BPs.

As shown in Fig. 6, in both follow-up strategies we were able to identify highly ranked compounds that had similar structure and docking poses to MTX. More importantly, we were able to achieve this by docking a small number of compounds as compared to the fully combinatorial library. This demonstrates the advantages of BP docking over brute force docking in saving time and resources.

The results described above are not surprising because the PLP scores for the library products correlates very well with the sums of the PLP scores of its BPs. This is shown in Fig. 7a, where the *x*-axis is the PLP score for a given library product, and the *y*-axis is the sum of the PLP scores for the two corresponding BPs. This correlation makes it tempting to conclude that the PLP scores are additive just like some other properties to be discussed in "Property-based combinatorial library designs with Basis Products" (see also, Ref. [25]). However, the linear correlation can also be interpreted as a library product docking in a pose that is energetically equivalent to the docked poses of its representative BPs. There have been some discussions regarding the factors, such as entropic or enthalpic contribution as being for or against this additivity assumption (see, for example, [46, 47]).

The correlation between the product PLP score and the best of its representing BPs was not as good (see Fig. 7b). For ranking combinatorial library compounds, both schemes were satisfactory, although "sum of BP" scheme performed much better. Note that in the "Best BP" scheme we only consider one fragment of a library product while in the "sum of BP" scheme we consider all fragments. When one of the fragments is dominant we would expect both schemes perform equally well. Otherwise, the "sum of BP" scheme is expected to perform better.

Figure 7c shows that it is practical to apply the hit follow-up strategies as described above and in "Method". This is a heat map for the PLP scores of the full library. The PLP scores for BP compounds are represented by cells in the first row (for the amine component) or the first column (for the acid component). The indices of both *x* and *y* for the rest of the cells are sorted by the BP PLP scores (from low to high). As can be seen from the figure, the compounds with the best PLP scores are represented by BPs with the best BP PLP scores. The highly ranked PLP scores are clustered near the lower left corner. Therefore, the results showed that the PLP scores for BPs are predictive of PLP scores for combinatorial library products.

The similar exercises were done using either calculated free energies or design energies to rank compounds. The correlations between a library compound and its representing BPs were not as good though. Therefore, even though we can still use BP compounds to identify the best library compounds the predictions were not expected to be as good when calculated free energies or design energies are used to rank compounds (results not shown here).

It is worth noting that the speedup of docking BPs over the entire library was better than 44 (= 34,200/770) (see Table 3). Yet, the quality of the compounds identified was comparable. Actually, the saving will become more dramatic when the size of the library increases. For the full Pfizer virtual combinatorial library, PGVL, the speedups would be in the magnitude of $1.25 \times 10^{13}/3.35 \times 10^{6} = 3.73 \times 10^{6}$ for in-house reagents and $3.28 \times 10^{16}/1.49 \times 10^{7} = 2.20 \times 10^{9}$ for commercial reagents (see Tables 2 and 3). These latter estimates excluded the costs of follow-ups for BP hits. The cost for follow-ups of BP hits is usually negligible comparing with the initial docking when the library size increases. Furthermore, docking BPs costs much less than docking the same number of compounds from the parent combinatorial library since BPs are usually much smaller than the compounds they represent.

Unlike other combinatorial docking programs [27–34], the BP structure-based virtual screening method described here does not perform in situ chemical reactions during docking. The enumeration and docking steps are modular and completely independent. Hence, chemistry knowledge is not needed for the docking step, and any docking programs can be used in combination with the BP method. LibDock is a modular method too for structure-based virtual screening of combinatorial library [35, 36]. Nevertheless LibDock requires full enumerations of the full combinatorial library that may prove prohibitive for large libraries.

Compounds in a combinatorial library share a lot of molecular fragments. The current method assumes that the dominant binding fragments will have similar binding modes independent of their parents. The more this assumption is true, the more predictive the BP docking results will be of the complete combinatorial library. The recently developed TrixX docking method has also used fragments to achieve up to more than 60-fold speedup for docking some libraries [37]. The speedup for TrixX would be much higher for combinatorial libraries. The drawbacks for using TrixX are that libraries need to be fully enumerated, and conformations and their analyses need to be catalogued in a relational database. Again, these may prove to be prohibitive for certain large virtual combinatorial libraries.

Property-based combinatorial library designs with Basis Products

It is well-documented that efficient pre-synthesis filtering of virtual library is very valuable and in some cases imperative for library design (see, for example, [55, 56]). Basis Products have been used for efficient filtering of virtual library without complete enumeration of the libraries [25]. This usage can be derived directly from the representation property of BPs for the full library (see Ref. [25] for an alternative derivation).

The relationship between a product of a combinatorial library and its BP representatives can be expressed in a tabulated form as illustrated in Table 1. Or more quantitatively, it can be expressed as the following isodesmic reaction equations for two-component reactions (Eq. 2) and for three-component reactions (Eq. 3) [57]:

$$AB + A_sB_s = AB_s + A_sB \tag{2}$$

$$ABC + 2A_sB_sC_s = AB_sC_s + A_sBC_s + A_sB_sC \tag{3}$$

where subscript "*s*" indicates pre-selected capping reactants for the corresponding reaction components. The isodesmic reaction equation can be generalized to:

$$\prod_{i=1}^{k} R^i + (k-1)\prod_{i=1}^{k} R_s^i = \sum_{i=1}^{k} R^i \prod_{j\neq i}^{k} R_s^j \tag{4}$$

for a k-component reaction:

$$\sum_{i=1}^{k} R^i \to \prod_{i=1}^{k} R^i. \tag{5}$$

Again simplified notations are used here. R^i, the running index for the i-th reaction component, is used to represent the reaction component itself. And the products are represented by a set of running indices. The subscript "*s*" indicates pre-selected capping reactants for BP formations. It is worth noting that the BPs involving only the capping reactants appeared in all reaction Eqs. 2–4. These terms are used to balance the atom types and bond types.

For additive properties, Eq. 5 can be transformed into the following formula:

$$Q\left(\prod_{i=1}^{k} R^i\right) = \sum_{i=1}^{k} Q\left(R^i \prod_{j\neq i}^{k} R_s^j\right) - (k-1)Q\left(\prod_{i=1}^{k} R_s^i\right) \tag{6}$$

This equation is derived by moving the second term of the left hand side of Eq. 4 to the right hand side and by using the additive assumption. Note that the terms on the right hand side of Eq. 6 involve only BPs. Therefore, Eq. 6 can be used to predict the property for a library compound using pre-calculated BP properties. This is exactly the equation used in Ref. [25].

Because counts of different atom types and bond types are conserved in isodesmic reactions, Eq. 6 is exact for atomic and bond properties, such as atom counts, bond counts, molecular weight, and other properties that can be decomposed into atomic and bond contributions. Hence, ring counts from Eq. 6 are also exact. Equation 6 holds approximately true for those properties depending only on small fragments. Examples of these properties are: calculated Log P values, various surface areas, and exclusion volumes. It is not appropriate to use Eq. 6 to predict molecular properties that are global, such as correlation energies and dipole moments.

It is interesting to point out that hypothetical isodesmic reactions have been routinely used to estimate heats of formation ab inito in Quantum Chemistry [57] and from experimentally measured heats of formation in Organic Chemistry [58]. Hence, it is natural to estimate heats of formation for compounds of a combinatorial library from the pre-calculated heats of formation for BPs.

We want to emphasize the fact that the calculations based on BPs are operations in the product space of a combinatorial library but with a scaling equal to that of the reactants. Therefore, the BP-based property filtering of combinatorial libraries has some advantages over either reactant-based filtering methods or product-based filtering methods. Unlike the reactant-based methods, the BP-based property filtering is based on properties calculated from compounds in the product space while the reactant-based methods use properties calculated from compounds in reactant space. Also, the BP-based filtering method does not need any enumeration of the combinatorial libraries beyond the BPs. Other product-based methods need to enumerate the compounds for which the properties are to be calculated. The BP-based filtering makes decisions based on properties of all compounds from the combinatorial library without any calculations beyond those for BPs (see Eq. 6). On the other hand, other product-based property-based methods make decision either based on partial knowledge from the selected few or with a brute-force computation of properties for all products.

Examples of key applications of Eq. 6 were given in Ref. [25].

Other applications of Basis Products

Docking of BPs can generate consensus core. Clustering analyses of docked poses can lead to most frequent cores at the most frequent poses. Then these “consensus” cores can be used for further design.

Another important application of BPs is analoging and/or lead hopping [41]. Given a lead compound, Basis Product library can be used to find similar compounds in the combinatorial space represented by the BP library. To achieve this, BPs similar to the fragments of the lead are identified. Then the BP hits are resolved into reactions. Sub-libraries are formed for individual reactions from the BP hits. A similarity computation for compounds in these sub-libraries to the original lead will offer top candidates for analogs or for lead hopping (for details, see [41]).

Basis Products can also be used to prioritize HTS campaigns for combinatorial-derived compounds. Virtual hits or real hits from HTS can be used for this purpose. The prioritization can be either by representation or by similarity. Similar strategies can also be applied to hit follow-up, if the hit is a combinatorial-derived compound. Basis Products can be identified from the hit. Then compounds in the sub-libraries around the hits either by similarity or by representation should rank high in prioritization. These strategies are quite similar to the conventional hit follow-up for combinatorial library hits.

Concluding remarks

Basis Products are strategically selected subset of compounds from a combinatorial library, and any compound in the library can be represented by its BPs. This representation relationship between BP library and its full parent library is the basis for all efficient BP applications. The structure-based virtual screening with BPs can successfully identify top compounds among a full combinatorial library with much less computations. The speedup of the BP method compared with the full docking is greater than 20 for our test case, and it can be much larger as the library size increases (see Table 3). Furthermore, the BP method may be the only practical method for structure-based virtual screening of some large virtual combinatorial libraries such as PGVL (see Table 2).

The property-based virtual screening with BPs is also very efficient as described here and shown in Ref. [25]. Other potential applications of the BP method include consensus core building, lead hopping, and library compound prioritization [19, 41]. With a simple molecular weight filtering, BPs can also be used as fragments in any experimental or virtual fragment-based design. There is a considerable advantage to use BPs as fragments. BPs have chemistry knowledge embedded in them. Any elaborations of the fragment hits are made much easier. Therefore, Basis Products can integrate combinatorial design with both structure-based design and fragment-based design.

In conclusion the BP method as we described here is an efficient method for both property-based and structure-based virtual screening of combinatorial libraries.

Acknowledgments The authors would like to thank Atsuo Kuki for his visionary guidance of PGVL Hub project at Pfizer, Jerry Hu for his help in generating data for Table 2, and a long list of other colleagues who helped in one way or another. We'd also like to thank Beth Lunney for reading the manuscript and offering insightful feedbacks.

References

1. Caldwell GW, Ritchie DM, Masucci JA, Hageman W, Yan Z (2001) Curr Top Med Chem 1:353
2. Kerns EH (2001) J Pharm Sci 90:1838
3. White RE (2000) Annu Rev Pharmacol Toxicol 40:133
4. Ajay (2002) Curr Top Med Chem 2:1273
5. Egan WJ, Walters WP, Murcko MA (2002) Curr Opin Drug Discov Devel 5:540
6. Muegge I (2003) Med Res Rev 29:302
7. Hopkins AL, Groom CR, Alex A (2004) Drug Discov Today 9:430
8. Abad-Zapatero C, Metz JT (2005) Drug Discov Today 10:464
9. Shuker SB, Hajduk PJ, Meadows RP, Fesik SW (1996) Science 274:1531
10. Hajduk PJ, Sheppard G, Nettesheim DG, Olejniczak ET, Shuker SB, Meadows RP, Steinman DH, Carrera GM Jr, Marcotte PA, Severin J, Walter K, Smith H, Gubbins E, Simmer R, Holzman TF, Morgan DW, Davidsen SK, Summers JB, Fesik SW (1997) J Am Chem Soc 119:5818
11. Hann MM, Leach AR, Harper G (2001) J Chem Inf Comput Sci 41:856
12. Jahnke W, Erlanson DA (eds) (2006) Fragment-based approaches in drug discovery. Wiley, Weinheim
13. Erlanson DA, McDowell RS, O'Brien T (2004) J Med Chem 47:3463
14. Hajduk PJ, Greer J (2007) Nat Rev Drug Discov 6:211
15. Hubbard RE, Chen I, Davis B (2007) Curr Opin Drug Discov Devel 10:289
16. Lebl M (1999) J Comb Chem 1:3
17. Leach AR, Bradshaw J, Green DVS, Hann MM, Delany JJIII (1999) J Chem Inf Comput Sci 39:1161
18. Yasri A, Berthelot D, Gijsen H, Thielemans T, Marichal P, Engels M, Hoflack J (2004) J Chem Inf Comput Sci 44:2199
19. Peng Z et al. Paper on PGVL Hub: the tool
20. Weller HN, Nirschl DS, Petrillo EW, Poss MA, Andres CJ, Cavallaro GL, Echols MM, Grant-Young KA, Houston JG, Miller AV, Swann RT (2006) J Comb Chem 8:664
21. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (1997) Adv Drug Deliv Rev 23:3
22. Rishton GM (2003) Drug Discov Today 8:86
23. Jalaie M, Shanmugasundaram V (2006) Mini Rev Med Chem 6:1159
24. Reddy AS, Pati SP, Kumar PP, Pradeep HN, Sastry GN (2007) Curr Protein Pept Sci 8:329
25. Shi S, Peng Z, Kostrowicki J, Paderes G, Kuki A (2000) J Mol Graph Model 18:478
26. Ghosh S, Nie A, An J, Huang Z (2006) Curr Opin Chem Biol 10:194
27. Sun Y, Ewing TJA, Skillman AG, Kuntz ID (1998) J Comput Aided Mol Design 12:597–604
28. Makino S, Ewing TJ, Kuntz ID (1999) J Comput Aided Mol Design 13:513–532
29. Johnson P, Talk at the Innovative Computational Applications Conference, San Francisco, CA, USA, Oct, 1999. VLSPROUT is an extension version of SPROUT: Gillet V, Johnson AP, Mata P, Sike S, Williams P (1993) J Comput Aided Mol Design 7: 127
30. Leach AR, Bryce RA, Robinson AJ (2000) J Mol Graph Model 18:358
31. Sprous DG, Lowis DR, Leonard JM, Heritage T, Burkett SN, Baker DS, Clark RD (2004) J Comb Chem 6:530
32. Gastreich M, Lilienthal M, Briem H, Claussen H (2006) J Comput-Aided Mol Design 20:717
33. Rarey M, Lengauer T (2000) Persp Drug Discov Design 20:63
34. Degen J, Rarey M (2006) Chem Med Chem 1:854
35. Diller DJ, Merz KM Jr (2001) Proteins Struct Func Genet 43:113
36. Diller DJ, Merz KM Jr (2006) US Patent No. US 7065453, June 20
37. Schellhammer I, Rarey M (2007) J Comput Aided Mol Des 21:223
38. Zhou JZ (2008) Curr Opin Chem Biol 12:379
39. Boehm M, Wu T-Y, Claussen H, Lemmen C (2008) J Med Chem 51:2468
40. Na J, Kostrowick J et al. Paper on PGVL Hub: the foundation
41. Hu J et al. Paper on Lead-Centric design inside PGVL Hub
42. Kim S, Oh CH, Ko JS, Ahn KH, Kim YJ (1985) J Org Chem 50:1927
43. Bartoli S, Fincham CI, Fattori D (2007) Curr Opin Drug Discov Devel 10:422
44. Hajduk PJ (2006) J Med Chem 49:6972
45. Leach AR, Hann MM, Burrows JN, Griffen EJ (2006) Mol Biosyst 2:430
46. Jencks WP (1981) Proc Natl Acad Sci USA 78:4046–4050
47. Murray CW, Verdonk ML (2002) J Comput Aided Mol Design 16:741–753
48. Böhm H-J, Stahl M (2002) Rev Comput Chem 18:41
49. Halperin I, Ma B, Wolfson H, Nussinov R (2002) Proteins Struct Funct Genet 47:409
50. Rajamani R, Good AC (2007) Curr Opin Drug Discov Devel 10:308
51. Bolin JT, Filman DJ, Matthews DA, Hamlin RC, Kraut J (1982) J Biol Chem 257:13650
52. The Available Chemical Database (ACD): http://www.symyx.com/
53. Gehlhaar DK, Verkhivker GM, Rejto PA, Sherman CJ, Fogel DB, Fogel L, Freer S (1995) Chem Biol 2:317
54. Verkhivker GM, Rejto PA, Gehlhaar DK, Freer S (1996) Proteins Struct Funct Genet 25:342
55. Walters WP, Stahl MT, Murcko MA (1998) Drug Discov Today 3:160
56. Clarifson PS, Walters PW (2002) Mol Div 5:185
57. Hehre WJ, Ditchfield R, Radom L, Pople JA (1970) J Am Chem Soc 92:4796
58. Benson SW (1976) Thermochemical kinetics. Wiley, New York