Chimia 57 (2003) 224–228

ISSN 0009–4293

Library Design

Friederike Stoll*

Abstract: This review aims at giving a short introduction to the most important areas of library design. The description of compounds by descriptors and fingerprints, and similarity-based clustering techniques are illustrated in the context of untargeted library design. For lead finding and lead optimization libraries it touches on ligand-based combinatorial design, structure-based design, docking and scoring techniques, and fragment-based *de novo* design. It is shown that computational and combinatorial chemistry can be successfully combined in the design process.

Keywords: Combinatorial docking · Diversity · Drug-likeness · Library design · Similarity

1. Introduction

The number of potential compounds that could be synthesized using combinatorial methods is immense. It was estimated that about 10^{60}–100^{100} molecules are synthetically accessible and show drug-like properties [1]. The need for rational library design is obvious as it is impossible to synthesize a larger fraction of these molecules. With the help of rational methods the number of compounds that have to be synthesized can be minimized while using a maximum of the available knowledge, *e.g.* to ensure that the compounds have the desired physicochemical properties. A closely related task is the rational selection of subsets of compounds from proprietary or commercial databases for screening.

Libraries can be designed for different purposes [2], drug discovery being a major application area [3]:

– Discovery libraries are intended to increase the diversity of a compound collection (*e.g.* the corporate collection of a pharmaceutical company). In addition to diversity itself the possibility of efficient synthesis and the compounds' physicochemical and ADME (absorption, distribution, metabolism, excretion) properties should be considered to ensure that the hits are of true value for discovery projects.
– Targeted libraries for lead finding: the target can be a certain protein or a target family, *e.g.* GPCRs (G protein-coupled receptors), kinases, phosphatases *etc.* For the design of targeted libraries, well-known structural motifs ('privileged structures') are often used.
– Optimization libraries for lead optimization (focused libraries).

Combinatorial chemistry can be useful in all three cases. The methodology that can be used for targeted library design depends very much on the structural information that is available. If a set of ligands is known for the target, a pharmacophore hypothesis can be developed and used as a constraint in database searches. It is also possible to use ligands for homologous targets as reference for similarity searching [4][5].

If a three-dimensional protein structure has been solved one can use docking techniques (*i.e.* predict the binding modes between ligand and protein) or design a library *de novo*. This review aims at giving a short general introduction into the field of library design.

2. Discovery Libraries

As the number of imaginable drug-like molecules is so large it is important to choose well the compounds that are to be synthesized. An important goal for a discovery library is to cover the chemical space as fully as possible in order to increase the probability of finding hits or leads from different chemical classes.

A typical starting point for a discovery library is a large virtual library that contains all structures that could possibly be synthesized *via* a specific synthetic pathway. A subset of this large library can be chosen by either looking at the diversity of the actual synthetic products or at the diversity of the reagents involved in the synthesis [3]. However, there are studies that suggest that product-based selection methods may lead to more diverse sets than reagent-based methods, depending on the kind of descriptors used [6][7]. An additional advantage

*Correspondence: F. Stoll
Novartis Pharma AG
Postfach
CH–4002 Basel
Tel.: + 41 61 32 44852
Fax: + 41 61 32 46726
E-Mail: friederike.stoll@pharma.novartis.com

of product-based strategies is that whole molecule properties can be optimized such as the physicochemical property profile of a library [8]. A limitation to product-based selection methods is the combinatorial constraint. In library production the goal is to combine each building block with every other to obtain a fully combinatorial library. This is difficult to achieve when the desired compounds are picked directly from compound space because it is unlikely that the chosen compounds form a combinatorial ensemble (see Fig. 1a: Only nine compounds are picked but a combinatorial synthesis would give 7 A × 6 B = 42 compounds). Taking the combinatorial constraint into account makes library synthesis much more efficient (Fig. 1b). However, product-based design is computationally much more expensive than reagent-based strategies and may not be feasible for larger virtual libraries.

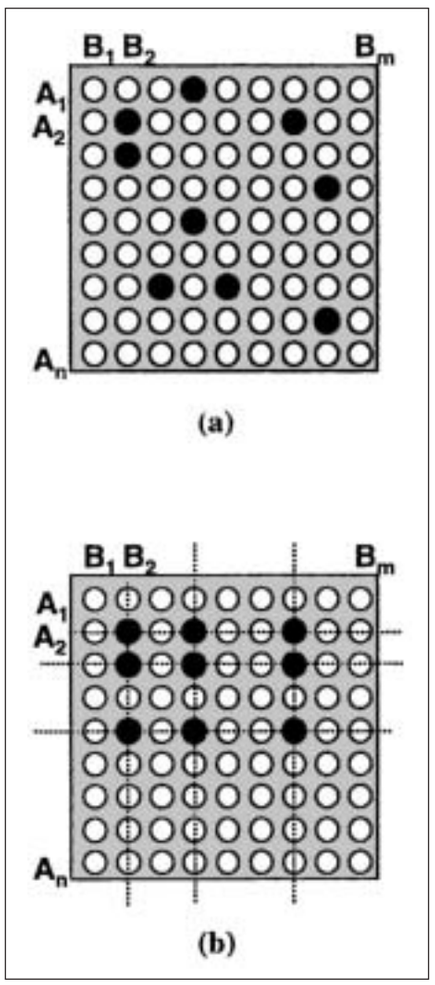

Fig. 1: A two-dimensional library is defined by two sets of reagents (rows A and columns B). (a) Compounds are picked without combinatorial constraint. (b) Taking the combinatorial constraint into account enhances the efficiency of library synthesis. (Fig. taken from [8].)

An ideal discovery library is representative of the chemical space with a minimum number of compounds. But how many compounds are needed in one series to make sure that hits are found with a high probability? A series of compounds is defined in this context as a number of molecules with a common core structure (scaffold). Even series that contain active compounds will mostly consist of inactive ones. If only a few compounds per scaffold are tested the probability of finding an active one amongst them is rather low. It was estimated that one needs 50–100 compounds per series to increase this probability to about 95% [9].

The basis of compound selection methods is the 'similar property principle' [10]. It assumes that structurally similar compounds should have similar physicochemical and biological properties so that even if the optimal substituent for a position is not a member of the subset, a similar one will be there [3]. Other important selection criteria are the properties of the compounds: Typically one tries to stay as much as possible within a property space that was defined as 'drug-like' or 'lead-like'. The most famous set of rules, the 'Rule of Five', was published by Lipinski [11]:

– molecular weight ≤ 500
– logP ≤ 5
– number of hydrogen bond acceptors ≤ 10
– number of hydrogen bond donors ≤ 5

Drug leads should not violate more than one of these limits, because experience has shown that compounds with multiple rule violations are unlikely to be orally absorbed. The 'Rule of Five' was derived from drugs and it may be necessary to adjust it to the purpose of lead finding [12]. One possibility of finding new rules would be to analyze a compound selection with desirable properties (*e.g.* a drug database or, in the case of targeted libraries, a set of hits from screening or drugs that bind to a target or target family).

In addition the pharmacokinetic (ADME) properties should be considered as early as possible [2][3]. The analysis of a compound database showed that bioavailability mainly depends on the flexibility of a compound (number of rotatable bonds) and its polar surface area [13]. This kind of rule can be used for filtering.

It should be kept in mind that the computation of molecular properties is not useful for the profiling of single compounds as the average rate of mis-classification is too high (about 20%) [1]. It is very useful however for the prediction of property distributions of compound collections (*e.g.* drug-likeness, GPCR modulator likeness, frequent-hitter liability) [1].

2.1. *Methods*

Choosing a diverse subset from a compound collection or a virtual library is a multi-step procedure [3]:

a) A number of descriptors or a fingerprint are computed for each compound.
b) The compounds are projected into a common descriptor space. Typically the number of descriptors is reduced by statistical methods such as principal component analysis (PCA) or factor analysis [14].
c) Subsets can be chosen using different strategies:

– the compounds with the lowest similarity coefficient
– the compounds with the largest distance in chemical space.

Different types of selection procedures have been described [15][16]:

– Compound clustering: the molecules are grouped into clusters that show a high degree of intra-cluster similarity and inter-cluster dissimilarity. One or several compounds per cluster are chosen [17].
– Grid-based sampling/partition-based selection: each dimension of the descriptor space (*i.e.* the range of values for the chosen molecular properties) is divided into bins (x-dimensional hypercubes). Afterwards compounds are sampled from each cube. This approach works only in property spaces of low dimensionality [6][17].
– The similarity of the fingerprints can also be assessed directly (direct sampling) [18][19].
– Optimization-based selections: As a prerequisite some quantitative measure of diversity has to be defined. The most diverse subset can be identified afterwards by combinatorial optimization methods, *e.g.* simulated annealing (SA) or genetic algorithms (GA) [15]. The latter imitate biological reproduction processes. In each generation the parent 'chromosomes' (*i.e.* the instruction set leading to a product) are changed by mutation, deletion, crossover *etc.* leading to a number of children. Of these children the best ranked are chosen as the next parent generation. This is done until an optimally diverse selection is found [20]. In the case of compound selection one chromosome would encode a whole sublibrary or a pool of reactants instead of individual compounds. GAs are able to optimize several properties simultaneously such as diversity and drug-likeness. An early example for this strategy was published by Gillet *et al.* [21], more recent studies are summarized by Egan, Walters, and Murcko [22].

2.1.1. *Descriptors*

Many different descriptors are used to capture molecular properties but not all of them are suited for the description of libraries. Physicochemical parameters such as molar refractivity or the partition coefficient are not available for virtual compounds, other parameters can be calculated but are computationally too expensive for large datasets, *e.g.* HOMO and LUMO energies [17].

In widespread use are descriptors such as the molecular weight, logP, number of hydrogen bond donors and acceptors, number of rotatable bonds, polar surface area, *etc*. [17][23] or topological descriptors that are based on molecular connectivity information or inter-atomic distances (for 3D conformers) [24][25]. One possibility to classify descriptors is their dimensionality: one-, two- and three-dimensional descriptors are deduced from 1D, 2D or 3D molecular representations (see Fig. 2).

Descriptors should have good neighborhood behavior: compounds which are close in property space to an active molecule should also be active and those that are close to an inactive compound should be inactive as well [17][23][26].

2.1.2. *Fingerprints*

Fingerprints are binary bit strings that encode diverse aspects of molecular structure and properties. Each bit (1 or 0) encodes the presence or absence of a feature or the value of a property descriptor and, from a different point of view, each bit represents one dimension in property space [27]. For similarity measures fingerprints are computed for all compounds and matched against each other. The size of fingerprints can vary a great deal. Simple 2D fingerprints have about 100 positions, more complex ones up to several thousand bits. 3D pharmacophore fingerprints are much more complex. They reflect all possible arrangements of three- or four-point pharmacophores in a molecule and consist of several million bits. Those pharmacophores have to be defined beforehand, *e.g.* by systematic conformational analysis. Fingerprints of higher dimensionality describe molecules in a much more complex way and are computationally expensive, but in many cases it will be sufficient to use simpler fingerprints. There is even evidence that 2D descriptors or fingerprints often perform better than 3D ones [23][26][28].

2.1.3. *Similarity Metrics*

A molecule that is represented in property space by a set of M descriptors can be represented by a point in M-dimensional space. The distance between these points can be measured by different metrics: the Euclidean distance, the Tanimoto coefficient or the cosine coefficient. The Euclidean distance between two compounds i and j ($d_{i,j}$) is calculated as

$$d_{i,j} = \sqrt{\sum_{k=1}^{M} (X_{ik} - X_{jk})^2} \qquad (1)$$

M is the set of numerical descriptors and X_ks are the values of the individual descriptors [17]. The Tanimoto and the cosine coefficients are calculated as shown in Eqn. 2 and 3.

$$S_{i,j} = \frac{\sum_{k=1}^{M} (X_{ik} - X_{jk})^2}{\sum_{k=1}^{M} X_{ik}^2 + \sum_{k=1}^{M} X_{jk}^2 + \sum_{k=1}^{M} X_{ik} \times X_{jk}} \qquad (2)$$

$$S_{i,j} = \frac{\sum_{k=1}^{M} X_{ik} \times X_{jk}}{\sqrt{\sum_{k=1}^{M} X_{ik}^2 \times \sum_{k=1}^{M} X_{jk}^2}} \qquad (3)$$

3. Targeted Libraries for Lead Finding

In principle the same methods can be applied in lead finding as were described above but in many cases the goal will not be the most diverse solution but one that shows good similarity with a specific hit. A very important difference between the design of diverse and targeted libraries is the amount of knowledge that is available: at least a number of hits is known, maybe even a 3D structure of the target protein. This structural information can be used to restrict the chemical space that has to be searched [3].

3.1. ***Ligand-based Combinatorial Design***

There are many targets for which no protein structure is known (*e.g.* G protein-coupled receptors). In the lead finding stage however some knowledge about structure–activity relationships is available and could be translated into a pharmacophore hypothesis. This pharmacophore can be used to do similarity searching in a virtual library. There is software available that not only searches for similar compounds but also optimizes simultaneously the properties of the set (*e.g.* HARPick [29], MoSELECT [30] or TOPAS [31]). With a suitable compound set and good quality biological data a quantitative structure–activity relationship (QSAR) analysis can be performed, *i.e.* structural parameters are correlated with binding affinity or activity. The QSAR model can then be used for the prediction of binding affinities of new compounds.

For hit optimization itself two strategies can be employed: either the topology (the 'skeleton' of the molecule) or the pharmacophoric patterns of the known hits can

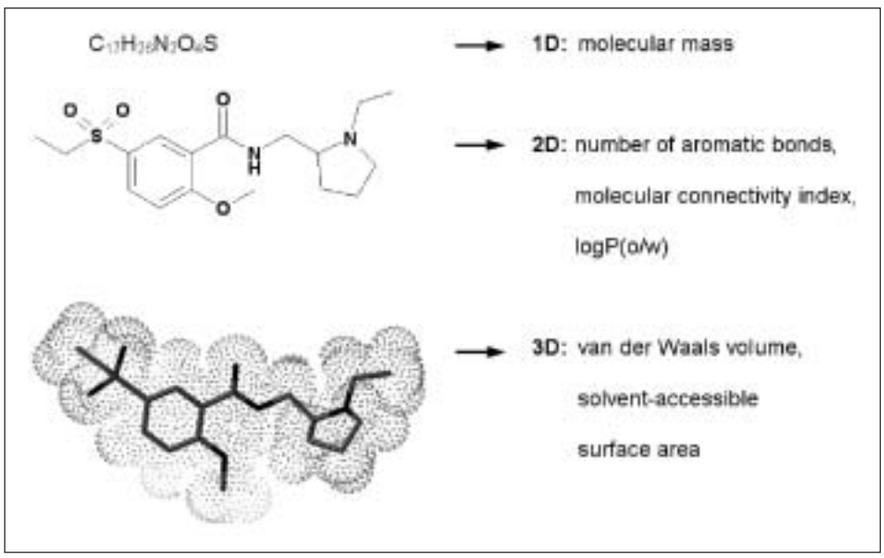

Fig. 2: 1D, 2D, and 3D molecular descriptors (adapted from [27]).

be preserved. It was shown that a variation of the topology leads to completely novel molecular skeletons while the search for compounds with a similar topology was better suited for the refinement of the pharmacophore hypothesis [32]. A fast automated search tool for compounds with similar topology is available with the program Ftrees and its extension Ftrees-FS [33][34], the latter being able to search not only in compound libraries but also in combinatorial fragment space. There are many other methods which cannot be explained here such as principal component analysis/partial least squares (PCA/PLS), encoder networks, self-organizing maps *etc.* [1].

Many examples are published that prove the success of design strategies. In two recent studies virtual compounds were constructed using libraries of diverse reagents. The resulting molecules were overlaid with the reference structure and scored. The most promising candidates were synthesized and tested. High-affinity compounds were obtained with a considerably increased hit rate [3][32].

3.2. *Structure-based Design*

If a 3D structure of the protein target is available it provides a well of additional information that can be used in the design of targeted libraries. It helps in scaffold selection as only those scaffolds have to be considered further that fit optimally into the binding pocket. Virtual libraries can be docked into the binding site or new compounds can be built up in the protein and ranked according to the quality of the fit. This review can only give a very short description of some docking and scoring tools, but many reviews can be found in the literature, *e.g.* [35–38]. Many examples of successful structure-based design of combinatorial libraries can be found in a review by Böhm and Stahl [38].

3.2.1. *Docking*

A number of different docking tools are available but not all of them are fast enough to be of use in the docking of large compound sets. Well-known fast docking programs are Dock [39], FlexX [40], Slide [41], and Fred [35]. The first two account for ligand flexibility by building up the compounds incrementally in the binding site. The latter two use pre-computed conformational ensembles that are stored in a database. Most docking tools do not take protein flexibility into account as this vastly increases the complexity of the task. There are some attempts, however, to allow for flexibility of the amino acid side chains (Slide [41], FlexE [42]).

3.2.2. *Scoring Functions*

A scoring function has to evaluate the different proposed binding modes of a compound in the binding site and to predict the correct one. In addition it has to rank the different compounds of the set by predicting their binding affinity [1][35]. For docking of large compound libraries fast scoring functions are necessary. Three different methodologies are the basis of these algorithms. Force field-based methods and knowledge-based methods are derived from protein–ligand complexes with high-affinity ligands. Empirical scoring functions are based on physicochemical properties such as hydrogen-bond counts [38][35]. Amongst the best known scoring functions are the FlexX [43] and the GOLD score [44]. Several studies have shown that the quality of a scoring function is dependent on the target and that in many cases a smart combination of multiple scoring functions seems to be superior to the use of individual functions [36][45–47].

3.2.3. *Fragment-based de novo Design*

De novo design of ligands can be used as an alternative to the docking of real or virtual compounds. In a first step the target protein has to be searched for putative protein–ligand interaction sites. These interaction sites are positions in the pocket where a ligand atom or functional group should be placed in order to interact favorably with the protein. In a next step new compounds are built by placing a small molecular fragment into the pocket and adding other fragments that fulfill the requirements ('seed and grow'). The quality of the new complexes can be rated using a scoring function (steric fit, chemical complementarity, pharmacophore similarity). The synthetic accessibility of the resulting compounds is usually better than with methods that build molecules atom by atom [48].

An alternative to this approach is the docking of several fragments and their subsequent connection *via* linkers ('dock and link'). These linkers can be chosen from a database or be designed *de novo* [48]. An experimental method that uses a similar approach is the SAR by NMR method [49]. An example of library design for NMR screening is described by Jacoby *et al.* [50].

Examples of *de novo* design programs are LUDI [51], BUILDER [52], SPROUT [53] or SMOG [54]. For the purposes of combinatorial design special combinatorial docking procedures have been developed. DREAM++ [55], CombiDOCK [56], and CombiSMOG [57] use algorithms that combine combinatorial ligand design and fast docking techniques [1].

3.3. *Lead Optimization Libraries*

For the design of lead optimization libraries the same tools can be used that were described above. The most important difference is that much more knowledge about the target can be used in the design process. The building blocks that are necessary to build a library that explores the chemical space around a lead can be chosen either by similarity analysis or with the help of a model. If a series of actives and their structure-activity relationship (SAR) is known, pharmacophoric constraints can be used as filtering tool. If it is possible to do a QSAR analysis the activity of compounds can be predicted and if there is a 3D structure of the target available the side chains can be chosen such that they are complementary to the protein structure. As described above docking of the compounds into the binding pocket can be used to predict their free energy of binding *via* a scoring function [3][17].

4. Summary

Finding compounds with good physicochemical and biological properties makes high demands on library design. This review gives some insight into the different tools that are available for the design of targeted or untargeted libraries and the complex requirements on reagents and products.

Received: March 24, 2003

[1] G. Schneider, *Curr. Med. Chem.* **2002**, *9*, 2095.
[2] D.W. Hobbs, T. Guo, *J. Rec. Signal Transduct. Res.* **2001**, *21*, 311.
[3] R.D. Brown, M. Hassan, M. Waldman, *J. Mol. Graph. Model.* **2000**, *18*, 427.
[4] A. Schuffenhauer, J. Zimmermann, R. Stoop, J.J. van der Vyver, S. Lecchini, E. Jacoby, *J. Chem. Inf. Comp. Sci.* **2002**, *42*, 947.
[5] A. Schuffenhauer, P. Floersheim, P. Acklin, E. Jacoby, *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 391.
[6] B.R. Beno, J.S. Mason, *Drug Discov. Today* **2001**, *6*, 251.
[7] E.A. Jamois, M. Hassan, M. Waldman, *J. Chem. Inf. Comp. Sci.* **2000**, *40*, 63.
[8] V.J. Gillet, *J. Comput-aided Mol. Des.* **2002**, *16*, 371.
[9] R. Nilakantan, F. Immermann, K. Haraki, *Comb. Chem. High Throughput Screen.* **2002**, *5*, 105.
[10] 'Concepts and applications of molecular similarity', Ed. M.A. Johnson, G.M. Maggiora, Wiley, New York, **1990**.
[11] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, *Adv. Drug Deliv. Rev.* **2001**, *46*, 3.

[12] S.J. Teague, A.M. Davis, P.D. Leeson, T. Oprea, *Angew. Chem. Int. Ed.* **1999**, *38*, 3743.
[13] D.F. Veber, S.R. Johnson, H.Y. Cheng, B.R. Smith, K.W. Ward, K.D. Kopple, *J. Med. Chem.* **2002**, *45*, 2615.
[14] D.K. Agrafiotis, V.S. Lobanov, D.N. Rassokhin, I. Izrailev , in 'Virtual screening for bioactive molecules', Ed. H.J. Böhm, G. Schneider, Wiley-VCH, Weinheim, New York, **2000**, 265.
[15] P. Willett, *Curr. Opin. Biotechnol.* **2000**, *11*, 85.
[16] P. Willett, *J. Comp. Biol.* **1999**, *6*, 447.
[17] A. Tropsha, W. Zheng, *Comb. Chem. High Throughput Screen.* **2002**, *5*, 111.
[18] D.K. Agrafiotis, *J. Chem. Inf. Comp. Sci.* **1997**, *37*, 841.
[19] M. Snarey, N.K. Terrett, P. Willett, D.J. Wilton, *J. Mol. Graph. Model.* **1997**, *15*, 372.
[20] L. Weber, in 'Evolutionary algorithms in molecular design', Ed. D.E. Clark, Wiley-VCH, Weinheim, New York, **2000**, p. 137.
[21] V.J. Gillet, P. Willett, J. Bradshaw, D.V.S. Green, *J. Chem. Inf. Comp. Sci.* **1999**, *39*, 169.
[22] W.J. Egan, P.W. Walters, M.A. Murcko, *Curr. Opin. Drug Discov. Dev.* **2002**, *5*, 540.
[23] R.D. Brown, Y.C. Martin, *J. Chem. Inf. Comp. Sci.* **1997**, *37*, 1.
[24] L.H. Hall, L.B. Kier, in 'Reviews in Computational Chemistry', Ed. K.B. Lipkowitz, D.B. Boyd, VCH, Weinheim **1991**, p. 367.
[25] R.E. Carhart, D.H. Smith, R. Venkataraghavan, *J. Chem. Inf. Comp. Sci.* **1985**, *25*, 64.
[26] H. Matter, *J. Med. Chem.* **1997**, *40*, 1219.
[27] J. Bajorath, *Nat. Rev. Drug Discov.* **2002**, *1*, 882.
[28] H. Matter, T. Potter, *J. Chem. Inf. Comp. Sci.* **1999**, *39*, 1211.
[29] A.C. Good, R.A. Lewis, *J. Med. Chem.* **1997**, *40*, 3926.
[30] V.J. Gillet, W. Khatib, P. Willett, P.J. Fleming, D.V. Green, *J. Chem. Inf. Comp. Sci.* **2002**, *42*, 375.
[31] G. Schneider, M.L. Lee, M. Stahl, P. Schneider, *J. Comput-aided Mol. Des.* **2000**, *14*, 487.
[32] R. Poulain, D. Horvath, B. Bonnet, C. Eckhoff, B. Chapelain, M.C. Bodinier, B. Deprez, *J. Med. Chem.* **2001**, *44*, 3378.
[33] M. Rarey, J.S. Dixon, *J. Comput-aided Mol. Des.* **1998**, *12*, 471.
[34] M. Rarey, M. Stahl, *J. Comput-aided Mol. Des.* **2001**, *15*, 497.
[35] P.D. Lyne, *Drug Discov. Today* **2002**, *7*, 1047.
[36] M. Stahl, M. Rarey, *J. Med. Chem.* **2001**, *44*, 1035.
[37] R.D. Taylor, P.J. Jewsbury, J.W. Essex, *J. Comput-aided Mol. Des.* **2002**, *16*, 151.
[38] H.J. Böhm, M. Stahl, *Curr. Opin. Chem. Biol.* **2000**, *4*, 283.
[39] T.J.A. Ewing, S. Makino, A.G. Skillman, I.D. Kuntz, *J. Comput-aided Mol. Des.* **2001**, *15*, 411.
[40] M. Rarey, B. Kramer, T. Lengauer, G. Klebe, *J. Mol. Biol.* **1996**, *261*, 470.
[41] V. Schnecke, L.A. Kuhn, *Perspect. Drug Discov. Des.* **2000**, *20*, 171.
[42] H. Claussen, C. Buning, M. Rarey, T. Lengauer, *J. Mol. Biol.* **2001**, *308*, 377.
[43] B. Kramer, G. Metz, M. Rarey, T. Lengauer, *Medicinal Chemistry Research* **1999**, *9*, 463.
[44] G. Jones, P. Willett, R.C. Glen, A.R. Leach, R. Taylor, *J. Mol. Biol.* **1997**, *267*, 727.
[45] C. Bissantz, G. Folkers, D. Rognan, *J. Med. Chem.* **2000**, *43*, 4759.
[46] H. Gohlke, G. Klebe, *Curr. Opin. Struct. Biol.* **2001**, *11*, 231.
[47] R.D. Clark, A. Strizhev, J.M. Leonard, J.F. Blake, J.B. Matthew, *J. Mol. Graph. Model.* **2002**, *20*, 281.
[48] A.R. Leach, R.A. Bryce, A.J. Robinson, *J. Mol. Graph. Model.* **526**, *18*, 358.
[49] S.B. Shuker, P.J. Hajduk, R.P. Meadows, S.W. Fesik, *Science* **1996**, *274*, 1531.
[50] E. Jacoby, J. Davies, M.J.J. Blommers, *Curr. Top. Med. Chem.* **2003**, *3*, 11.
[51] H.J. Bohm, *J. Comput-aided Mol. Des.* **1992**, *6*, 61.
[52] D.C. Roe, I.D. Kuntz, *J. Comput-aided Mol. Des.* **1995**, *9*, 269.
[53] V.J. Gillet, W. Newell, P. Mata, G. Myatt, S. Sike, Z. Zsoldos, A.P. Johnson, *J. Chem. Inf. Comp. Sci.* **1994**, *34*, 207.
[54] R.S. Dewitte, E.I. Shakhnovich, *J. Am. Chem. Soc.* **1996**, *118*, 11733.
[55] S. Makino, T.J.A. Ewing, I.D. Kuntz, *J. Comput-aided Mol. Des.* **1999**, *13*, 513.
[56] Y. Sun, T.J.A. Ewing, A.G. Skillman, I.D. Kuntz, *J. Comput-aided Mol. Des.* **1998**, *12*, 597.
[57] B.A. Grzybowski, A.V. Ishchenko, J. Shimada, E.I. Shakhnovich, *Accounts Chem. Res.* **2002**, *35*, 261.