DOI: 10.1002/anie.200703323

Drug Design

KNOBLE: A Knowledge-Based Approach for the Design and Synthesis of Readily Accessible Small-Molecule Chemical Probes To Test Protein Binding**

*Christof Gerlach, Martin Münzel, Bernhard Baum, Hans-Dieter Gerber, Tobias Craan, Wibke E. Diederich, and Gerhard Klebe**

In memory of Jörg Stürzebecher

At the beginning of the last decade, potent high-throughput technologies stimulated the development of combinatorial chemistry as a productive source of candidate molecules suited for screening biological activity against putative drug targets. However, the sheer increase of readily available test compounds has not dramatically enhanced the efficiency of lead discovery.[1–3]

As a consequence, compound selection for testing is moving increasingly from "optimally diverse" sets to targeted compound libraries. As proteins occur in families characterized by conserved binding motifs, they can be addressed by ligand skeletons exhibiting binding epitopes complementary to the commonly exposed properties of the protein family. Accordingly, efficient design of targeted compound libraries starts with a central privileged skeleton that specifically addresses the exposed binding motif of a protein family.[4] Using appropriate substituents or side chains of the central privileged skeleton, individual members of a focused library are equipped with the desired selectivity towards distinct members of a protein target family.[5–8]

Herein, we present the novel strategy KNOBLE (KNOwledge-Based Ligand Enumeration) to assemble a focused combinatorial library using a structure-based approach. In all steps of design, the synthetic accessibility of the starting materials and products is considered. To test the concept, library members are synthesized and subjected to biological testing. Our goal is the discovery by simple chemistry of lead compounds that are potent enough to allow determination of the crystal structure together with the target protein. This structure can subsequently serve as a starting point to embark upon in-depth structure-based ligand optimization.

[*] Dr. C. Gerlach, M. Münzel, B. Baum, H.-D. Gerber, T. Craan, Dr. W. E. Diederich, Prof. Dr. G. Klebe
Institut für Pharmazeutische Chemie
Philipps-Universität Marburg
Marbacher Weg 6, 35032 Marburg (Germany)
Fax: (+49) 6421-282-8994
E-mail: klebe@mailer.uni-marburg.de

[**] The DFG-Graduiertenkolleg "Proteinfunktion auf atomarer Ebene" (C.G.) and the DPhG "Stiftung zur Förderung des wissenschaftlichen Nachwuchses" (W.E.D.) are gratefully acknowledged for their financial support.

Supporting information for this article is available on the WWW under http://www.angewandte.org or from the author.

The design of an appropriately focused combinatorial library starts with the selection of a central fragment equipped with linker groups allowing substitutions by standard synthetic chemistry. Putative building blocks are selected and virtually assembled in a computer simulation. The actual retrieval of suitable side chains is guided by the shape and the physicochemical properties of the binding pocket of the target protein. Usually protein binding pockets, particularly those in proteases, can be split into a set of distinct subpockets, which can be examined separately. Once such a subpocket is characterized and extracted, its description according to the pseudocenter concept can be used for cavity comparisons in the Cavbase approach.[10–13]

In Cavbase, binding pockets are automatically detected, and their physicochemical properties are encoded by five generic descriptors that describe the molecular recognition features in terms of hydrogen bonding and hydrophobic interactions. More than 130000 cavities have been stored, and their occupants are available through a hyperlink to the database Relibase.[14] Both databases store structural information about crystallographically determined protein–ligand complexes deposited with the protein databank (PDB).[15] Cavbase maps a predefined query cavity against the entire set of stored entries. The result is a list of cavities ranked in order of decreasing similarity to the query cavity.

With respect to the design of a targeted compound library, a Cavbase similarity search returns not only matching subcavities with similar properties but, most importantly, the chemical structures of the molecules or molecular fragments occupying these subcavities. The chemical structures of the retrieved occupants are subjected to a search in chemical catalogues to see whether they or their derivatives are commercially available. The composition of the returned chemicals is subsequently used as a guideline to design synthetic routes to build a targeted compound library following combinatorial chemistry principles.

As a case study to probe this new strategy, we selected the family of serine proteases. These enzymes cleave polypeptide chains with varying sequence selectivity. Along the reaction pathway, a covalently attached acyl-enzyme intermediate is formed, and the N-terminal part of the cleaved peptide chain remains bound to the protein. Consequently, serine proteases exhibit well-established recognition pockets on the unprimed side to accommodate the amino acid side chains of the N-terminal part of the peptidic substrate next to the cleavage

site. Furthermore, all serine proteases bind their substrates on the unprimed side through hydrogen bonds to the peptide backbone.

We picked thrombin as a reference for this case study. This enzyme exposes three well-defined binding subpockets to recognize the N-terminal substrate. Using these cavities as queries for a Cavbase search revealed about 4000 hits in each case. For the S1 subpocket, 646 cavities from other serine proteases were matched; for the S3 pocket, 550 such cavities were found. In case of the S2 pocket, only 274 entries from other serine proteases were retrieved, thus indicating the structural uniqueness of S2 in thrombin. This selectivity-determining pocket is rather tight and spatially restricted by the thrombin-specific 60 loop.

Closer analysis of the matching S3 pockets highlights the hydrophobic character of this site, which is strongly determined by a highly conserved tryptophan residue forming the floor of the pocket (e.g. acrosin, cathepsin G, factor VIIa, factor IXa, factor Xa, factor XIa, chymase, trypsin, C1s-elastase, tryptase, t-plasminogen activator, protein C, and granzyme A). Similarly, the search retrieved S1 pockets from, for example, cathepsin G, factor VIIa, factor IXa, factor XIa, trypsin, tryptase, protein C, factor Xa, t-plasminogen activator, enteropeptidase, and urokinase.

Subsequently, the actual occupants were extracted in terms of ligand fragments. As a first attempt, we decided to design a compound library for thrombin. Thus, in each step, a comparison with the thrombin reference subpocket was performed to retrieve only fragments that potentially exhibit a similar interaction pattern with the target protein. Evaluating the occupants of the S2 pocket matches highlights the unique shape of this pocket in thrombin: among the 100 best-ranked solutions, 98 examples are found in other thrombin structures.

Ligand fragments with an isopropyl group, a small aliphatic or heteroaliphatic ring such as a pyrrolidine moiety, or partially aromatic building blocks were preferentially detected (Figure 1).

In thrombin and many other members of the trypsin-like serine proteases, the S1 subpocket is dominated by the aspartate 189 residue, which is embedded in a highly conserved hydrophobic environment. The comparison of binding pockets reveals ligand building blocks with comparable pysicochemical properties, while the exact degree of similarity in the amino acid sequence of the surrounding protein is much less important. Nevertheless, besides the popular basic residues derived from benzamidine, guanidine, or aminopyridine building blocks, halogen-substituted aromatic moieties were also suggested; these interact with a conserved tyrosine residue at the floor of the S1 pocket (Figure 1). The most pronounced chemical diversity of occupants is proposed by the Cavbase search for the S3 pocket. For this pocket, possible side chains originate from other members of the serine protease family (Figure 1).

Thus, we see the strength of this approach, as not only well-known thrombin inhibitors are used to create the combinatorial library. Cavbase provides easy access to an entire class of proteins, and simple "hopping" between subpockets and ligand occupants across a protein family is easily possible.

Once a particular ligand fragment had been selected as a putative building block for the construction of a combinatorial library, a search in the Sigma–Aldrich catalogue was performed using the program JME.[16] For the different subpockets, these searches retrieved possible reaction components, which are given in the Supporting Information. The list of commercially available starting materials was then evaluated with respect to attached functional groups that are suited to connect the individual building blocks by a generally applicable synthetic route. In the synthesis design, we decided to focus on ester, amide, and sulfonamide bond formation as well as on nucleophilic substitution reactions. Three fragments, (*S*)-prolinol, (*S*)-2-(hydroxymethyl)piperidine, and (*S*)-valinol, were finally selected as central moieties to address the S2 pocket (Scheme 1).

To further diversify the central P2 moiety, a glycine spacer was attached between P2 and P3. The functional groups of this additional spacer are appropriate to address the non-specific backbone recognition site exposed by all serine proteases. We chose ester bond formation to connect the central moiety with a group propitious to address the S1 pocket. Finally, to place a suitable building block into S3, an amide bond was formed using either the amino functionality in the P2 central moiety or the glycine spacer. Alternatively, bond formation through nucleophilic substitution was envisaged. In Scheme 1, carboxylic acid derivatives suited to address the S1 pocket are listed together with carboxylic and sulfonic acid derivatives as well as the chloro derivatives to address the S3 pocket. The selected P1 to P3 building blocks were assembled on the computer following Scheme 2 and docked into the binding pocket of thrombin using the combinatorial module of FlexX.[17,18] Two libraries, with and without the glycine spacer, were assembled, each comprising 507 possible members. For each docking run, the 10 best solutions were stored and rescored using DrugScoreCSD.[19] Visual inspection of the docking solutions was performed using the graphical interface to evaluate the generated docking modes in terms of per-atom scoring contributions.[20]

For synthetic simplicity we decided to use the 3-chlorobenzyl and 4-cyanophenyl portion from the list of best-scored solutions to address the S1 pocket (apart from ligands exhibiting an amidino or guanidino portion at this site, which involve more synthesis steps). Selection of synthesis candidates for S2 and S3 was guided by synthetic feasibility and reactivity differences of the starting materials. The finally selected 2-pyridylacetic acid, *p*-fluorobenzoic acid, and the *tert*-butyloxy carbonyl moiety for S3 showed up in several of the best-scored hits. From the list of best-scored derivatives, compounds **2**–**6** were selected, synthesized (Scheme 3), and subjected to a photometric enzyme kinetics assay.[21] They showed inhibition in the micromolar range (see Scheme 3 and the Supporting Information). Subsequently, we succeeded in diffusing **5** into crystals of thrombin, from which the complex structure could be determined (Figure 2). This structure can serve as a starting point to embark upon structure-based lead optimization.

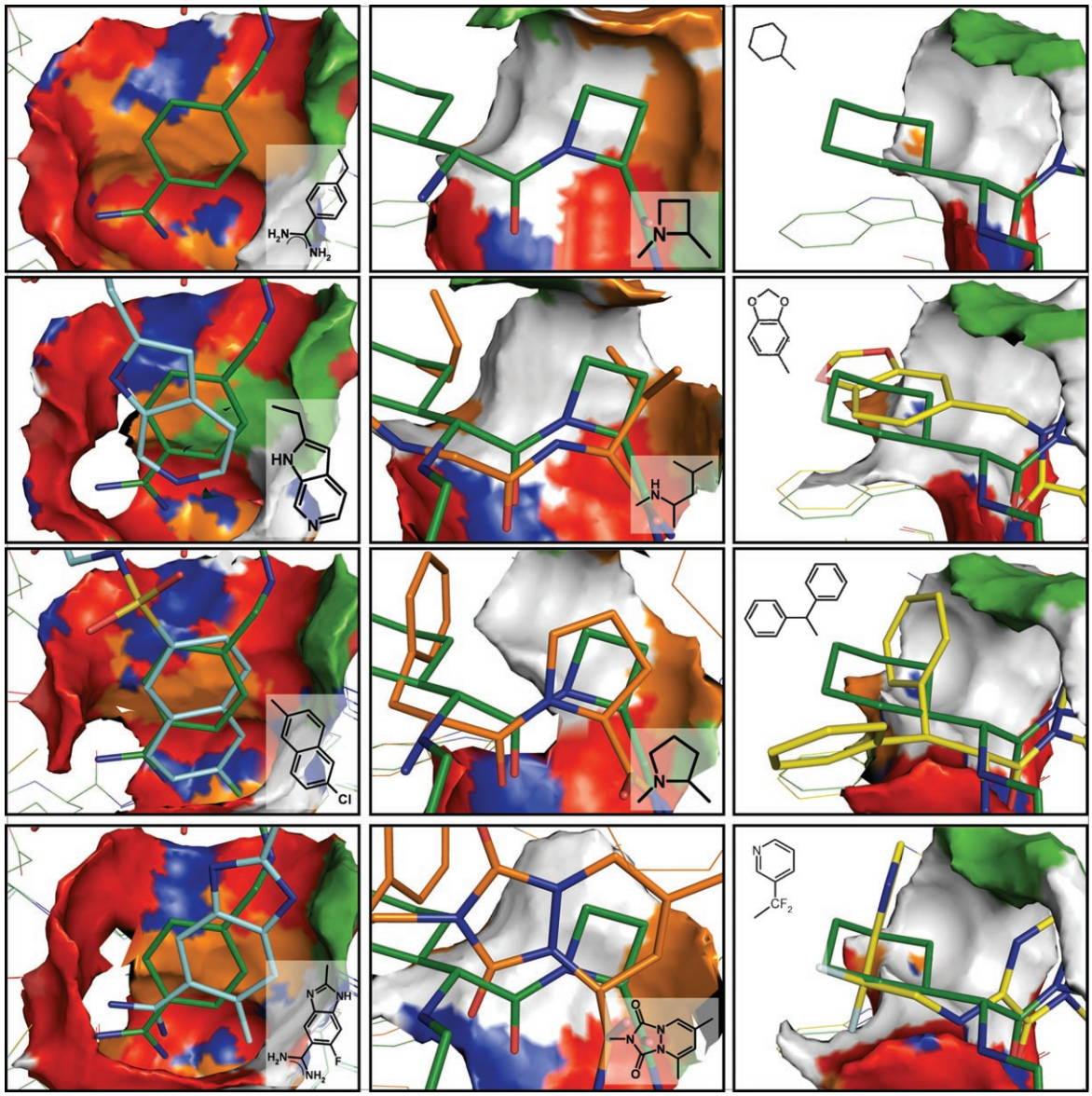
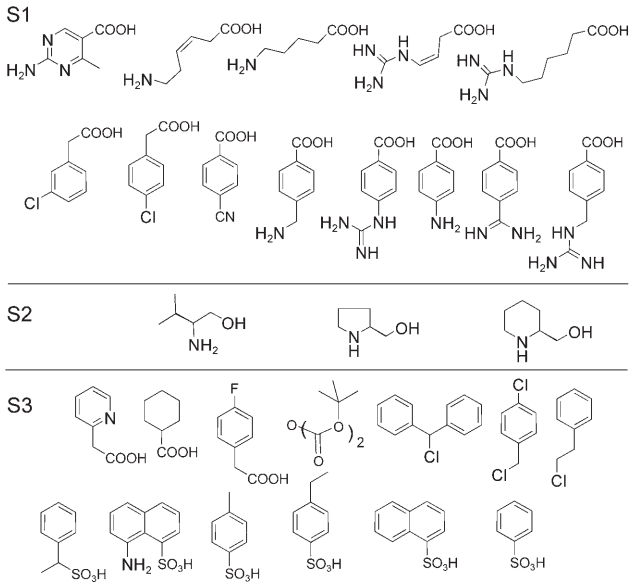


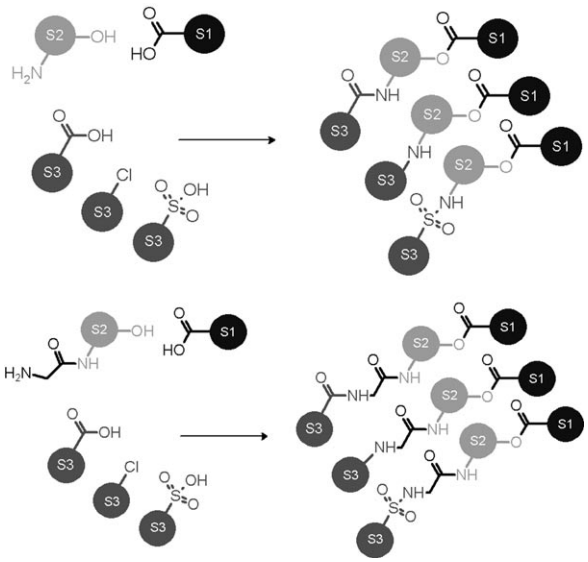
Figure 1. Fragments found by subcavity matching. For comparison, the crystallographically determined binding mode of melagatran (green, PDB code 1k22) in thrombin[9] is shown in the upper row for the S1 (left), S2 (center), and S3 (right) pocket together with three examples retrieved from occupants in other proteins (second to fourth rows). These protein surfaces display similar physicochemical properties (blue: H-bond donor, red: H-bond acceptor, green: donor or acceptor, white and orange: areas for interaction with aliphatic and aromatic groups, respectively). In the first row, the physicochemical properties for the melagatran complex are shown. In the following rows, the physicochemical properties of the subcavities in the other proteins matched by the algorithm are displayed. The adopted binding geometry (cyan, orange, yellow) and chemical formulas of the fragments occupying the corresponding subpockets are given.

As thrombin is a well-established target, the next optimization steps appear evident, that is, comparing the present complex with that of melagatran, a potent thrombin inhibitor. Firstly, the ester oxygen atom is not ideal, as it lacks the donor facility to form an H-bond to the neighboring serine residue. Melagatran uses its amide NH group at this position to establish such a contact. Secondly, our ester chain linking the P1 and P2 occupant appears to be too long and should be reduced by one member. In **5**, the ester carbonyl oxygen atom squeezes into a hydrogen bond to Gly216. This interaction prevents the glycine portion of **5** from forming the usual dual-ladder hydrogen-bond contact. In the case of melagatran, this H-bond network is well-established. Accordingly, the next synthetic step is to replace the four-membered P1–P2 ester linkage with a three-membered amide linker. This change will create additional space to allow the glycine building block to participate in the dual-ladder contact to Gly216.

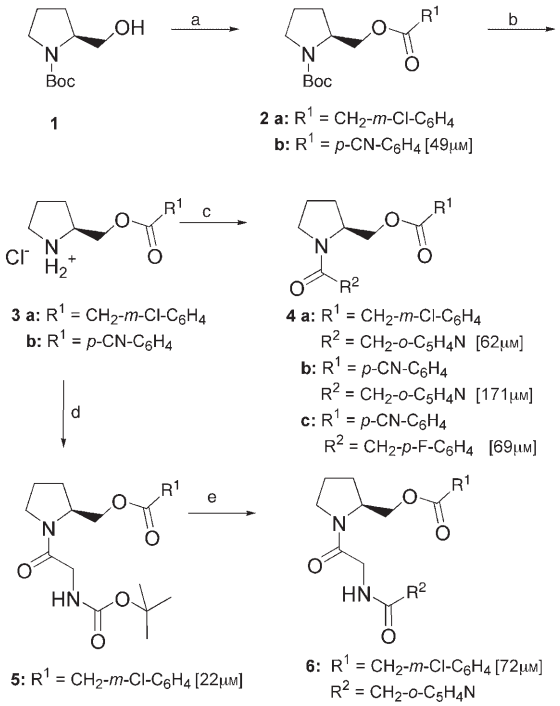
The present study has been performed to demonstrate the feasibility of the outlined knowledge-based approach to combinatorial ligand design. It exploits the available structural information regarding experimentally characterized binding geometries of molecular building blocks of ligands hosted in subpockets of other proteins. These latter proteins have to show pronounced similarity to subpockets present in



Scheme 1. Chemical formulas of the building blocks for the three subpockets.



Scheme 2. Synthetic route for the assembly of the virtual combinatorial library by means of FlexX^C.



Scheme 3. Synthesis of compounds **2**–**6**. K_i values for **2b**, **4a**, **4b**, **4c**, **5**, and **6** for thrombin inhibition are given in square brackets. Reagents and reaction conditions are given in the Supporting Information.

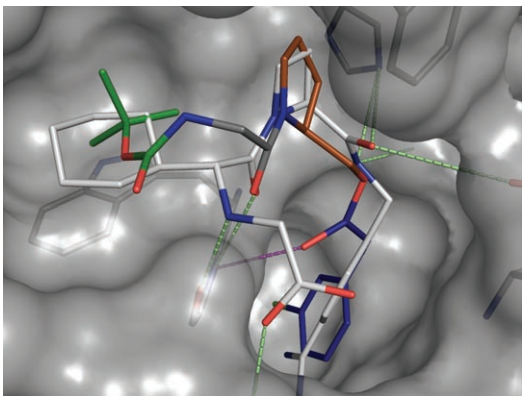


Figure 2. Superposition of the crystal structure of **5** (P1: 3-chlorobenzyl portion in blue; P2: prolinol moiety in brown, glycine spacer in dark gray; P3: *tert*-butyl moiety in green) with melagatran (extracted from PDB code 1k22). Melagatran (gray) forms by its P1–P2 amide-bond H-bonds (light green) to Ser195 and Glu192. Furthermore, it involves Gly216 in a double hydrogen bond. In contrast, in compound **5**, the formation of an optimal H-bond network between protein and bound ligand is perturbed by interaction between the ester carbonyl oxygen atom and the Gly216 NH group (magenta).

the target protein for which a putative ligand is to be developed. We recently applied Cavbase subpocket matching to select the most promising side chains of non-natural amino acids in the design and synthesis of peptidomimetic SARS protease inhibitors.[12]

Computational approaches are frequently accused of suggesting synthesis candidates that are difficult to make or much to elaborate as first candidates to validate a vague design hypothesis. In contrast, the concept presented herein tries to consider at a very early stage the synthetic feasibility of the assembled molecules. Computational design and combinatorial docking are consulted to select the most prospective candidates from the multiplicity of possible library members. Their synthesis is straightforward and is based on commercially available starting materials.

Once likely lead candidates are detected, their metabolic stability must be taken into account. For example, ester bonds used to connect individual building blocks are accepted for reasons of synthetic accessibility but must be removed owing to their likely metabolic instability. Nevertheless, the major

advantage of this approach is that it is based on binding geometries actually found and confirmed in experimentally determined crystal structures.

Received: July 24, 2007
Published online: October 23, 2007

Keywords: combinatorial chemistry · drug design · ligand design · thrombin

[1] E. J. Martin, R. E. Critchlow, *J. Comb. Chem.* **1999**, *1*, 32.
[2] A. R. Leach, M. M. Hann, *Drug Discovery Today* **2000**, *5*, 326.
[3] C. G. Wermuth, *Drug Discovery Today* **2006**, *11*, 160.
[4] G. Müller, *Drug Discovery Today* **2003**, *8*, 681.
[5] H. J. Böhm, D. W. Banner, L. Weber, *J. Comput.-Aided Mol. Des.* **1999**, *13*, 51.
[6] K. Illgen, T. Enderle, C. Broger, L. Weber, *Chem. Biol.* **2000**, *7*, 433.
[7] G. Schneider, M. L. Lee, M. Stahl, P. Schneider, *J. Comput.-Aided Mol. Des.* **2000**, *14*, 487.
[8] L. Weber, S. Wallbaum, C. Broger, K. Gubernator, *Angew. Chem.* **1995**, *107*, 2452; *Angew. Chem. Int. Ed. Engl.* **1995**, *34*, 2280.
[9] F. Dullweber, M. T. Stubbs, D. Musil, J. Stürzebecher, G. Klebe, *J. Mol. Biol.* **2001**, *313*, 593.
[10] S. Schmitt, D. Kuhn, G. Klebe, *J. Mol. Biol.* **2002**, *323*, 387.
[11] N. Weskamp, D. Kuhn, E. Hüllermeier, G. Klebe, *Bioinformatics* **2004**, *20*, 1522.
[12] S. I. Al-Gharabli, S. T. A. Shah, S. Weik, M. F. Schmidt, J. R. Mestres, D. Kuhn, G. Klebe, R. Hilgenfeld, J. Rademann, *ChemBioChem* **2006**, *7*, 1048.
[13] D. Kuhn, N. Weskamp, S. Schmitt, E. Hüllermeier, G. Klebe, *J. Mol. Biol.* **2006**, *359*, 1023.
[14] M. Hendlich, A. Bergner, J. Günther, G. Klebe, *J. Mol. Biol.* **2003**, *326*, 607.
[15] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **2000**, *28*, 235.
[16] P. Ertl, Novartis Institutes for BioMedical Research Basel, Switzerland, **2004**.
[17] M. Rarey, B. Kramer, T. Lengauer, G. Klebe, *J. Mol. Biol.* **1996**, *261*, 470.
[18] M. Rarey, T. Lengauer, *Perspect. Drug Discovery Des.* **2000**, *20*, 63.
[19] H. F. Velec, H. Gohlke, G. Klebe, *J. Med. Chem.* **2005**, *48*, 6296.
[20] P. Block, N. Weskamp, A. Wolf, G. Klebe, *Proteins Struct. Funct. Genet.* **2007**, *68*, 170.
[21] J. Stürzebecher, U. Stürzebecher, H. Vieweg, G. Wagner, J. Hauptmann, F. Markwardt, *Thromb. Res.* **1989**, *54*, 245.

