Protein Structures

DOI: 10.1002/ange.200602125

Protein Structure Similarity Clustering: Dynamic Treatment of PDB Structures Facilitates Clustering**

Bradley D. Charette, Richard G. MacDonald, Stefan Wetzel, David B. Berkowitz, and Herbert Waldmann**

Protein structure similarity clustering (PSSC)[1–3] is one of a number of potential guiding principles[4,5] that have been introduced to focus combinatorial-library design/protein targeting. PSSC clusters protein targets with similar ligand-binding cores in which little sequence or functional similarity is evident. Lead compounds for one member of the cluster then provide novel starting points in chemical space for ligand development for other members of the PSSC.

We describe herein a new clustering procedure that lends itself to ligand docking, molecular dynamics (MD), and the vector-alignment-search-tool (VAST)[6] algorithm. This MD-assisted approach offers an alternative to the serial structure retrieval/inspection steps used to generate the original protein structure similarity cluster, centered on Cdc25A.[2] Furthermore, in a particularly difficult de novo test, the mannose 6-phosphate/insulin-like growth factor II receptor (M6P/IGF2R), lends itself to clustering, with MD opening up connections to partners that the static crystal structure coordinates fail to find.

Ortholand and Ganesan[7] have described the disappointing early yield of clinical candidates from combinatorial chemistry/high-throughput screening as a "paradise lost" sentiment, probably attributable to unreasonable expectations and the need for maturation in approaches to combinatorial library design.[8] Both diversity-oriented synthesis (DOS) and biology-oriented synthesis (BIOS) have arisen in response to the this need. DOS utilizes more-complex, often fused polycyclic frameworks, richer in stereochemistry and frequently functionalizes them through annulation or coupling reactions.[9] Within BIOS,[10] biological relevance and prevalidation are employed as key arguments for the design of natural-product-derived/inspired[3,11] compound collections.

PSSC seeks to add value to such compound libraries by grouping targets into otherwise unrecognized clusters. In this way, a BIOS library, originally directed at enzymes of a specific functional or homology family, can be redirected at a PSSC partner protein and produce novel leads for this target. PSSC groups proteins based on a comparison of extracted "ligand-sensing cores", are independent of sequence or functional similarity. Importantly, core extraction is independent of domain or fold constraints and involves carving out a sphere around a key ligand or protein-based center in the binding pocket. This approach is complementary to a number of interesting, surface pattern recognition methods that compare binding pockets based upon exposed protein surface pseudocenters (CavBase,[12] SiteEngine[13]), triangles (pvSOAR),[14] cavities,[15] or residues (efSite,[16] PINTS[17]), and to the "all atom" approach exemplified by SitesBase.[18]

The PSSC approach begins in biological space and seeks to first identify a cluster of targets of pharmaceutical or bioorganic interest and then looks for an appropriate lead scaffold, drawing from binding data for one member of the cluster. The notion of a PSS cluster is related to the concept advanced by Sternberg and co-workers of "supersites", reflecting binding-site similarity in the absence of homology.[19] PSSC also resonates with interesting ideas put forth by Quinn and co-workers,[20] suggesting that structural motifs used to bind intermediates in enzymes along a biosynthetic pathway for a natural product might also be used to find (unrelated) target proteins for that same natural product.

The original identification of a PSS cluster[2] began from the single available X-ray crystal structure (1C25)[21] for the catalytic core of Cdc25A, a dual-specificity protein phosphatase (Figure 1). The Cdc isozymes may represent novel chemotherapeutic targets.[22] Initial structure similarity searches by using the combinatorial extension (CE)[23] and Dali-FSSP (families of structurally similar proteins)[24] algorithms connected Cdc25A with hydroxynitrile lyase (3YAS)[25] and methylene methanopterin dehydrogenase (1LU9),[26] respectively.

These initial cluster members indicated that the parent dual-specificity phosphatase could be related to proteins capable of promoting nucleophilic addition to carbonyl centers (perhaps presaging carboxyl esterase activity) and with nicotinamide-dependent dehydrogenase activity. Koch

[*] B. D. Charette, Prof. D. B. Berkowitz
Department of Chemistry
University of Nebraska
Lincoln, NE 68588 (USA)
Fax: (+1) 402-472-9402
E-mail: dbb@unlserve.unl.edu
S. Wetzel, Prof. H. Waldmann
Max-Planck-Institut für Molekulare Physiologie
Abteilung für Chemische Biologie
Otto-Hahn-Strasse 11, 44227 Dortmund (Germany)
and
Universität Dortmund
Fachbereich 3, Chemische Biologie
Fax: (+49) 231-133-2499
E-mail: herbert.waldmann@mpi-dortmund.mpg.de
Prof. R. G. MacDonald
Department of Biochemistry & Molecular Biology
University of Nebraska Medical Center
Omaha, NE 68198 (USA)

[**] This research was supported by the Max-Planck-Gesellschaft, the Deutsche Forschungsgemeinschaft, and the Fonds der Chemischen Industrie. D.B.B. thanks the Max-Planck-Gesellschaft (sabbatical stay), the American Heart Association (GIA 0650014Z), and NIH (CA 62034 and RR016544-01) for support. B.D.C. acknowledges a UCARE fellowship. R.G.M. acknowledges support from the NIH (CA 91885). We thank Dr. Ingrid R. Vetter for useful discussions. PDB = protein data bank.

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

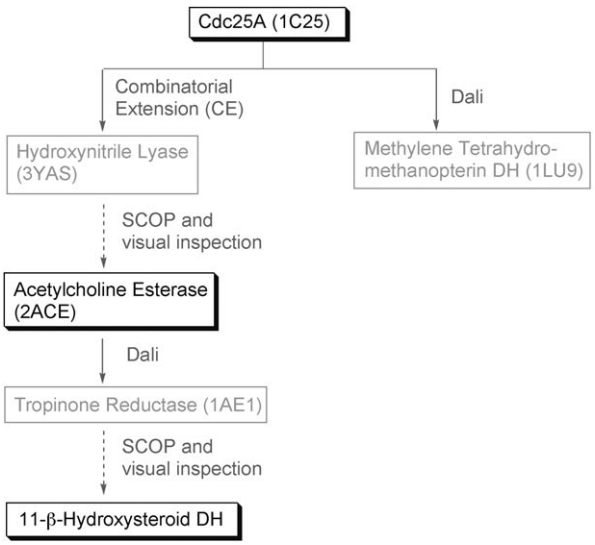

Figure 1. The pathway taken by Koch[2] to establish the original PSS cluster. Note the use of SCOP and intuition/manual inspection.

and co-workers next used SCOP (structural classification of proteins; scop@mrc-lmb.cam.ac.uk), a database from the MRC (The Medical Research Council, Cabridge, UK) that groups evolutionarily proteins based on conserved structural elements.[2,27] In this way, beginning from the 3YAS lead, structurally related proteins could be located, retrieved, DaliLite[4] or CE pairwise-aligned, and examined by visual inspection (Figure 1). This pathway led to acetylcholinesterase (AChE) as a PSS partner for Cdc25A (see Table 1).

A second iteration of this sequence was then carried out. This time, the Dali structural alignment identified tropinone reductase as an AChE partner. The idea that pyridine nucleotide-based dehydrogenases (DHs) might be in this PSSC had been foreshadowed by the initial 1LU9 hit. Indeed, manual inspection of the Rossmann fold superfamily, analogous to the earlier SCOP step, then led from tropinone reductase to 11β-hydroxysteroid dehydrogenase (11β-HSDH). Eventually, a 147-member library based on the sesterterpene Cdc25A inhibitor, dysidiolide,[28] also produced inhibitors for both AChE and 11β-HSDH (see the Supporting Information).[2]

Motivated by the success of this de novo application of PSSC to generate a target cluster that responds to a natural-product-inspired combinatorial library, we set out to examine the clustering protocol more closely. Namely, the initial formulation of this PSSC required two labor-intensive SCOP/superfamily searching steps (Figure 1). This meant the serial retrieval of a good number of PDB coordinate sets for cluster candidates and their evaluation for active-site structural homology with the catalytic core of Cdc25A. This was necessitated because available web-based structural comparison algorithms, including CE,[23] Dali,[24] and VAST,[6] all fail to find either AChE or 11β-HSDH as structural homologues when starting from 1C25, the only available structure for Cdc25A.

Given that such structure-comparison algorithms are usually limited to searching a static set of crystallographic coordinates in the PDB, we wondered whether introducing dynamics[29] would facilitate the structure-homology search. The notion is simply that the crystallographic snapshot of a protein that is deposited in the PDB is likely to be one of a number of conformations that an approaching ligand might encounter in solution. If one could begin from an array of conformations that Cdc25A likely samples in protein structure space, then a structural comparison might be able to identify PSSC partners that would not otherwise appear.

Toward this end, (−)-dysidiolide was initially docked with the 1C25 structure by using Autodock 3.0.[30] Then, a one-nanosecond MD simulation was carried out on the docked structure (Gromacs[31], Figure 2), collecting snapshots every picosecond. To analyze the variations in backbone conforma-

Table 1: MD simulation Cdc25 A with docked (−)-dysidiolide.

Time [ps][a]	Alignment length (AChE-1H22)[b]	RMSD [Å] (AChE-1H22)[b]	Alignment length (11β-HSDH-1XSE-A)[c]	RMSD [Å] (11β-HSDH-1XSE-A)[c]
0	88	6.0	85	4.2
100	90	6.3	90	4.5
200	90	5.9	81	3.6
300	70	3.4	81	3.5
400	95	5.7	80	3.6
500	86	3.7	96	4.8
600	86	4.0	88	4.3
700[d]	**91**	**4.3**	**90**	**3.9**
800	86	3.9	91	4.8
900	79	5.0	93	5.0
1000	83	4.5	92	5.4

[a] Representative conformers at 100-ps intervals in the MD simulation on the Cdc25A-dysidiolide docked structure. [b] Results of a DaliLite pairwise comparison with the 1H22 acetylcholinesterase crystal structure. [c] Results of a DaliLite pairwise comparison with the 11β-hydroxysteroid dehydrogenase structure 1XSE, molecule A in the unit cell. [d] Data in bold highlights the time region in which the simulation samples Cdc25A conformations that align especially well with these cluster partners.

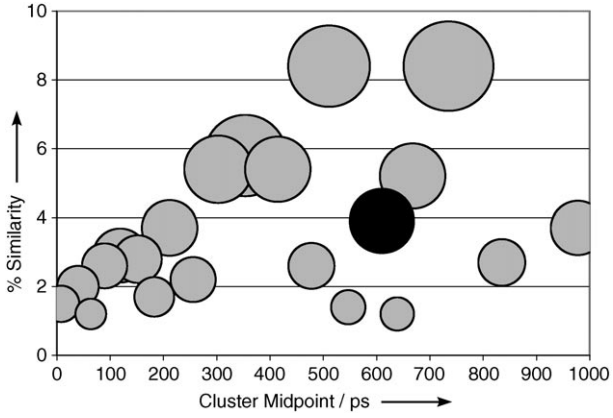

Figure 2. Cluster distribution of the 1000 conformers resulting from the MD simulation (Gromacs) on the dysidiolide-docked Cdc25A structure shown in Figure 3. All clusters with more than 10 members are shown as a function of the time of appearance of the midpoint conformer. Circle diameters are proportional to cluster size. The midpoint conformer of the black-shaded cluster is shown in Figure 3.

tion with time, along the simulation, pairwise comparisons with the structures of AChE (1H22)[32] and of (11β-HSDH-1 (1XSE-A)[33] were performed at 100-ps intervals (see Table 1 for results). In particular, one notices that in the 600–700-ps regime, the conformations of Cdc25A being sampled apparently show greater structural homology with these two PSSC partner proteins.

The 1000 Cdc25A conformations sampled were clustered by using Gromos[31] (Figure 2). Of a total of 60 conformational clusters found, 26 have more than 10 members. The three most populated clusters only begin to appear near the midpoint, probably reflecting a large conformational change associated with "flap closure" dominating the initial part of the simulation. This motion is quite apparent from an animation of the MD simulation (see the Supporting Information) and can be inferred from a comparison of early and late snapshots (Figure 3).

The midpoint conformers of each of the top ten clusters were arbitrarily sampled for their ability to find members of the original PSS cluster (Figure 1). By using the VAST algorithm, all 10 midpoint conformers find hydroxynitrile lyase (HNL) and four find methylene tetrahydromethanopterin DH (MTHMP DH), which were the only two PSS partners that were initially identifiable from the static 1C25 structure. Pleasingly, however, a VAST search initiated from conformer 668 (midpoint of the cluster shaded black in Figure 2) finds all five members of the Cdc25A cluster. More specifically, this conformer hits a significant fraction of the available structures for these proteins in the PDB: HNL (22/23), MTHMP DH (2/2), tropinone reductase (3/6), AChE (14/66), and 11β-HSDH (1/6).

VAST finds good alignment of conformer 668 with 11β-HSDH-1(1XEA) in the following regions (residues 37–56, 63–66, 69–77, 91–98, 100–112, and 124–133). If one follows the average root mean square deviation (RMSD) for C_α atoms across these residues, relative to their position in the starting conformer, there is a sharp increase from approximately 0.7–1 Å in the time regime of 600–800 ps. Consistent with this, the conformer at 700 ps also finds four of the five cluster members (see the Supporting Information). A structural overlay of conformer 668 with AChE and 11β-HSDH-1 is presented in Figure 3 (bottom).

In light of the success of MD to open up new structural alignments with VAST, we next set our sights on a de novo test of the protocol. In this regard, a particularly difficult case was chosen; namely, the cation-independent, M6P-IGF2R (see Figure 4).[34] This large 300-kDa transmembrane receptor contains 15 homologous extracytoplasmic repeats. Domains 3 and 9 are known to have a high affinity for M6P. The receptor binds M6P-functionalized proteins and IGF-II independently (domain 11). In addition to serving a range of housekeeping functions associated primarily with channeling proteins to the lysosomes, the receptor is known to internalize circulating IGF-II and hence has been labeled a cancer-suppressor gene.[35] Internalization appears to be accelerated by M6P-type ligand binding,[36] increasing interest in high-affinity ligands for this site.[37]

Only three PDB entries are available for the M6P-sensing core of the receptor; namely of a domain 1–3 fragment, one

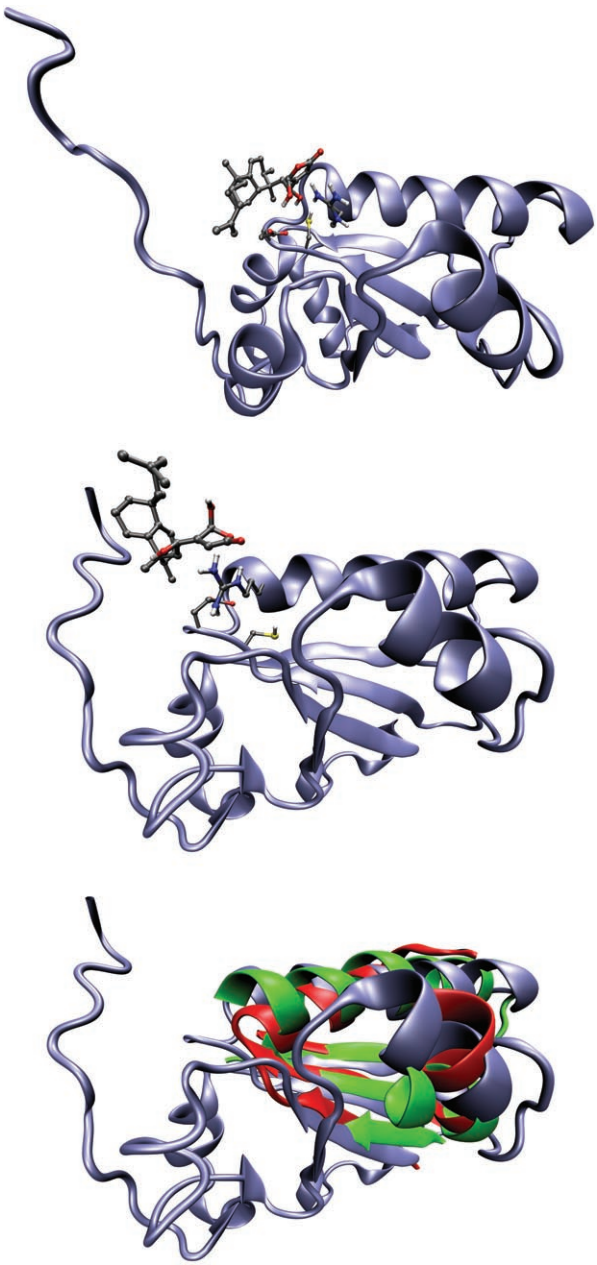

Figure 3. Top: Docked structure of Cdc25A (1C25) and dysidiolide. The highlighted residues are Cys430, Glu431, and Arg436. Middle: Conformer 668, the midpoint conformer of the black-shaded cluster shown in Figure 2 (see also Table 1 for evidence that MD samples particularly relevant conformational space for the Cdc25A/AChE/11HSD PSSC cluster in this time regime). Bottom: Structural overlay (DaliLite) of conformer 668 of Cdc25A following MD simulation (blue) with acetylcholinesterase (1H22, red) and 11β-HSD1 (1XSE, green). VAST finds this triplet only after the initial 1C25 coordinates are allowed to relax through MD.

unbound ligand (1Q25),[38] and two nearly identical structures with bound M6P (1SZ0 and 1SYO).[39] An initial VAST search of 1Q25 and 1SZ0 revealed few potential PSSC partners (three hits for each structure, see the Supporting Information). Therefore, MD simulations with 1000-ps intervals were performed on domains 1–3 of both the ligand-unbound (1Q25) and ligand-bound (1SZ0) structures. In the former

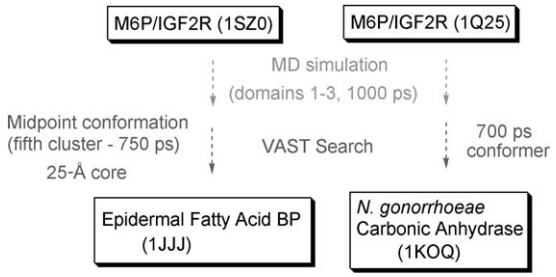

Figure 4. MD-assisted clustering procedure for the cation-indepedent mannose 6-phosphate/insulin-like growth factor II receptor. BP = binding protein.

case, conformers were arbitrarily sampled at 100-ps intervals and subjected directly to VAST searches. The 700-ps conformer identified the interesting carbonic anhydrase (CA) from *N. gonorrhoeae* (1KOQ)[40] as a potential cluster partner for this M6P receptor.

The ligand-bound conformers were clustered by domain 3, and VMD-extracted[41] 25-Å cores (centered on C5 of the M6P ligand) from midpoint conformers of the 10 most populated clusters were subjected to VAST searches. From the fifth cluster, a connection to epidermal fatty acid binding protein (E-FABP) emerged. E-FABP belongs to a family of lipid-binding proteins that are associated with fatty acid signaling, cell growth, and cell differentiation. E-FABP is known to be overexpressed in hyperproliferative skin diseases, such as psoriasis,[42] and also appears to be a cancer marker.[43] For these reasons, this target appears to be an excellent PSS cluster partner for the M6P-IGF2R. Neither *N. gonorrhoeae*, CA, nor E-FABP were identified by VAST searches without MD. An overlay of the ligand-bound receptor core (750-ps conformer) with E-FABP is presented in Figure 5.

In conclusion, we describe a new clustering procedure for PSSC that exploits MD and the VAST algorithm. This streamlined approach removes serial structure retrieval and visual alignment steps previously required for clustering (Cdc25A PSSC), and opens up new windows to structural neighbors not seen when dealing with fixed X-ray crystal structure coordinates (M6P/IGF2R PSSC). Our results suggest that, in general, by treating otherwise static PDB entries as an ensemble of conformers generated by MD, one can sample a broader, and potentially more relevant, swath of protein-structure space when searching for structural partners.

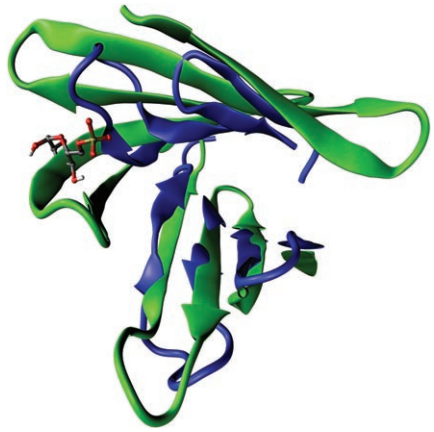

Figure 5. Overlay of the MD-relaxed (750 ps) M6P-IGF2R core (blue) with the human E-FABP (green). The 1SZ0-bound M6P ligand (position at 750 ps) is highlighted with a ball-and-stick model.

Received: May 27, 2006
Revised: August 25, 2006
Published online: October 31, 2006

Keywords: biological information · combinatorial chemistry · molecular dynamics · protein structures · target selection

[1] M. A. Koch, H. Waldmann, *Drug Discovery Today* **2005**, *10*, 471.
[2] M. A. Koch, L.-O. Wittenberg, S. Basu, D. A. Jeyaraj, E. Gourzoulidou, K. Reinecke, A. Odermatt, H. Waldmann, *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 16721.
[3] R. Breinbauer, I. R. Vetter, H. Waldmann, *Angew. Chem.* **2002**, *114*, 3002; *Angew. Chem. Int. Ed.* **2002**, *41*, 2878.
[4] L. Holm, J. Park, *Bioinformatics* **2000**, *16*, 566.
[5] J. Fejzo, C. A. Lepre, J. W. Peng, G. W. Bemis, Ajay, M. A. Murcko, J. M. Moore, *Chem. Biol.* **1999**, *6*, 755.
[6] J.-F. Gibrat, T. Madej, S. H. Bryant, *Curr. Opin. Struct. Biol.* **1996**, *6*, 377.
[7] J.-Y. Ortholand, A. Ganesan, *Curr. Opin. Chem. Biol.* **2004**, *8*, 271.
[8] C. Lipinski, A. Hopkins, *Nature* **2004**, *432*, 855.
[9] a) D. S. Tan, *Nat. Chem. Biol.* **2005**, *1*, 74; b) M. D. Burke, E. M. Berger, S. L. Schreiber, *J. Am. Chem. Soc.* **2004**, *126*, 14095; c) M. D. Burke, S. L. Schreiber, *Angew. Chem.* **2004**, *116*, 48; *Angew. Chem. Int. Ed.* **2004**, *43*, 46.
[10] A. Noeren-Mueller, I. Reis-Correa, Jr., H. Prinz, C. Rosenbaum, K. Saxena, H. J. Schwalbe, D. Vestweber, G. Cagna, S. Schunk, O. Schwarz, H. Schiewe, H. Waldmann, *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 10606.
[11] a) A. M. Boldi, *Curr. Opin. Chem. Biol.* **2004**, *8*, 281; b) D. J. Newman, G. M. Cragg, K. M. Snader, *J. Nat. Prod.* **2003**, *66*, 1022; c) M.-L. Lee, G. Schneider, *J. Comb. Chem.* **2001**, *3*, 284.
[12] a) G. Klebe, *Drug Discovery Today* **2006**, *11*, 580; b) N. Weskamp, D. Kuhn, E. Huellermeier, G. Klebe, *Bioinformatics* **2004**, *20*, 1522.
[13] A. Shulman-Peleg, R. Nussinov, H. J. Wolfson, *J. Mol. Biol.* **2004**, *339*, 607.
[14] T. A. Binkowski, L. Adamian, J. Liang, *J. Mol. Biol.* **2003**, *332*, 505.
[15] M. Stahl, C. Taroni, G. Schneider, F. Hoffmann, *Protein Eng.* **2000**, *13*, 83.
[16] K. Kinoshita, J. i. Furui, H. Nakamura, *J. Struct. Funct. Genomics* **2002**, *2*, 9.
[17] A. Stark, S. Sunyaev, R. B. Russell, *J. Mol. Biol.* **2003**, *326*, 1307.
[18] N. D. Gold, R. M. Jackson, *J. Mol. Biol.* **2006**, *355*, 1112.
[19] R. B. Russell, P. D. Sasieni, M. J. E. Sternberg, *J. Mol. Biol.* **1998**, *282*, 903.
[20] B. M. McArdle, M. R. Campitelli, R. J. Quinn, *J. Nat. Prod.* **2006**, *69*, 14.
[21] E. B. Fauman, J. P. Cogswell, B. Lovejoy, W. J. Rocque, W. Holmes, V. G. Montana, H. Piwnica-Worms, M. J. Rink, M. A. Saper, *Cell* **1998**, *93*, 617.
[22] M. A. Lyon, A. P. Ducruet, P. Wipf, J. S. Lazo, *Nat. Rev. Drug Discovery* **2002**, *1*, 961.
[23] a) I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **2001**, *29*, 228; b) I. N. Shindyalov, P. E. Bourne, *Protein Eng.* **1998**, *11*, 739.
[24] L. Holm, C. Sander, *Nucleic Acids Res.* **1996**, *24*, 206.

[25] J. Zuegg, K. Gruber, M. Gugganig, U. G. Wagner, C. Kratky, *Protein Sci.* **1999**, *8*, 1990.
[26] U. Ermler, C. H. Hagemeier, A. Roth, U. Demmer, W. Grabarse, E. Warkentin, J. A. Vorholt, *Structure* **2002**, *10*, 1127.
[27] a) A. Andreeva, D. Howorth, S. E. Brenner, T. J. P. Hubbard, C. Chothia, A. G. Murzin, *Nucleic Acids Res.* **2004**, *32*, D226; b) L. Lo Conte, S. E. Brenner, T. J. P. Hubbard, C. Chothia, A. G. Murzin, *Nucleic Acids Res.* **2002**, *30*, 264; c) A. G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia, *J. Mol. Biol.* **1995**, *247*, 536.
[28] S. P. Gunasekera, P. J. McCarthy, M. Kelly-Borges, E. Lobkovsky, J. Clardy, *J. Am. Chem. Soc.* **1996**, *118*, 8759.
[29] a) W. Sherman, T. Day, M. P. Jacobson, R. A. Friesner, R. Farid, *J. Med. Chem.* **2006**, *49*, 534; b) C. A. Sotriffer, O. Kraemer, G. Klebe, *Proteins Struct. Funct. Bioinf.* **2004**, *56*, 52.
[30] D. S. Goodsell, A. J. Olson, *Proteins Struct. Funct. Genet.* **1990**, *8*, 195.
[31] X. Daura, K. Gademann, B. Jaun, D. Seebach, W. F. VanGunsteren, A. E. Mark, *Angew. Chem.* **1999**, *111*, 249; *Angew. Chem. Int. Ed.* **1999**, *38*, 236.
[32] D. M. Wong, H. M. Greenblatt, H. Dvir, P. R. Carlier, Y.-F. Han, Y.-P. Pang, I. Silman, J. L. Sussman, *J. Am. Chem. Soc.* **2003**, *125*, 363.
[33] D. Ogg, B. Elleby, C. Norstroem, K. Stefansson, L. Abrahmsen, U. Oppermann, S. Svensson, *J. Biol. Chem.* **2005**, *280*, 3789.
[34] a) A. B. Hassan, *Am. J. Pathol.* **2003**, *162*, 3; b) J. C. Byrd, J. H. Y. Park, B. S. Schaffer, F. Garmroudi, R. G. MacDonald, *J. Biol. Chem.* **2000**, *275*, 18647; c) R. G. MacDonald, S. R. Pfeffer, L. Coussens, M. A. Tepper, C. M. Brocklebank, J. E. Mole, J. K. Anderson, E. Chen, M. P. Czech, A. Ullrich, *Science* **1988**, *239*, 1134.
[35] a) J. Li, G. G. Sahagian, *Oncogene* **2004**, *23*, 9359; b) D. B. O'Gorman, J. Weiss, A. Hettiaratchi, S. M. Firth, C. D. Scott, *Endocrinology* **2002**, *143*, 4287; c) G. R. Hankins, A. T. De Souza, R. C. Bentley, M. R. Patel, J. R. Marks, J. D. Iglehart, R. L. Jirtle, *Oncogene* **1996**, *12*, 2003.
[36] S. J. York, L. S. Arneson, W. T. Gregory, N. M. Dahms, S. Kornfeld, *J. Biol. Chem.* **1999**, *274*, 1164.
[37] D. B. Berkowitz, G. Maiti, B. D. Charette, C. D. Dreis, R. G. MacDonald, *Org. Lett.* **2004**, *6*, 4921.
[38] L. J. Olson, R. D. Yammani, N. M. Dahms, J.-J. P. Kim, *EMBO J.* **2004**, *23*, 2019.
[39] L. J. Olson, N. M. Dahms, J.-J. P. Kim, *J. Biol. Chem.* **2004**, *279*, 34000.
[40] S. Huang, Y. Xue, E. Sauer-Eriksson, L. Chirica, S. Lindskog, B.-H. Jonsson, *J. Mol. Biol.* **1998**, *283*, 301.
[41] W. Humphrey, A. Dalke, K. Schulten, *J. Mol. Graphics* **1996**, *14*, 33.
[42] G. Siegenthaler, R. Hotz, D. Chatellard-Gruaz, S. Jaconi, J. H. Saurat, *Biochem. Biophys. Res. Commun.* **1993**, *190*, 482.
[43] a) L. H. Gutierrez-Gonzalez, C. Ludwig, C. Hohoff, M. Rademacher, T. Hanhoff, H. Ruterjans, F. Spener, C. Lucke, *Biochem. J.* **2002**, *364*, 725; b) P. Sinha, G. Hutter, E. Kottgen, M. Dietel, D. Schadendorf, H. Lage, *Electrophoresis* **1999**, *20*, 2952.

