Combinatorial computational method gives new picomolar ligands for a known enzyme

Bartosz A. Grzybowski*, Alexey V. Ishchenko*†, Chu-Young Kim‡, George Topalov*, Robert Chapman*, David W. Christianson‡, George M. Whitesides*, and Eugene I. Shakhnovich*§

*Harvard University, Department of Chemistry and Chemical Biology, 12 Oxford Street, Cambridge, MA 02138; and ‡University of Pennsylvania, Department of Chemistry, 250 South 33rd Street, Philadelphia, PA 19104

Communicated by Roy Gordon, Harvard University, Cambridge, MA, December 14, 2001 (received for review November 22, 2001)

Combinatorial small molecule growth algorithm was used to design inhibitors for human carbonic anhydrase II. Two enantiomeric candidate molecules were predicted to bind with high potency (with R isomer binding stronger than S), but in two distinct conformations. The experiments verified that computational predictions concerning the binding affinities and the binding modes were correct for both isomers. The designed R isomer is the best-known inhibitor ($K_d \sim 30$ pM) of human carbonic anhydrase II.

The development of new drugs often depends on the identification of molecules ("leads") that have high affinities for specified macromolecular targets. Although tight binding is only one important characteristic of a drug (1), it is often used as a guide in initial stages of drug discovery. Two contrasting methods—combinatorial (2–4) and rational (5–8)—represent the extremes in strategies for discovery of high-affinity leads. Combinatorial methods make it possible to screen large numbers of potential candidates and do not require prior knowledge of the structure of the receptor molecule. Rational methods attempt to design high-affinity ligands based on knowledge of the atom-level structure of the receptor and of molecular interactions. In practice, both combinatorial and rational methods can efficiently identify relatively low-affinity leads; both are inefficient and unreliable in identifying high-affinity ($K_d \sim$ nM) ligands (9–12).

Here, we describe a computational methodology that combines combinatorial and rational strategies in the form of a computational system that is rapid enough to generate biased libraries of leads and accurate enough to give useful predictions of energetics and geometry. In its first experimental test, this method yielded a new ligand for a human carbonic anhydrase II (HCA) that has the highest known affinity for this enzyme ($K_d \sim 30$ pM). To our knowledge, it is the first time that a computational method has created a ligand that has the highest known affinity for a protein target.

Our method, called CombiSMoG for *combi*natorial *s*mall *mo*lecule *g*rowth, incorporates the philosophy of combinatorial synthesis into computational drug design and is based on two interrelated components: a knowledge-based potential and a Monte Carlo ligand growth algorithm. The knowledge-based potential (13) is derived from a set of 1,000 protein–ligand complexes, whose structures are deposited in the Protein Data Bank crystallographic database (14). In this potential, two atoms—one on the ligand and one on the protein—are said to be in contact if the distance between them is less than a specified cutoff value (usually 5 Å). The contacts are classified according to the constituent atom types, and the frequencies of their occurrences in the database are transformed into energies by means of a Boltzmann-like relation to give the scoring function used in CombiSMoG (15, 16). This potential has three main advantages over the commonly used force fields (17–20). (*i*) The binary definition of atom–atom interactions allows binding energies to be evaluated rapidly. (*ii*) These energies implicitly take into account the entropy of water desolvation after binding of a ligand and have the meaning of binding free energies. (*iii*) Because the potential is based on the large representative set of protein–ligand complexes, it provides a statistically significant description of interactions between proteins and small molecules. Unlike most semiempirical force fields used in drug design, the potential used in CombiSMoG is solidly based on statistical mechanics and does not have any arbitrary adjustable parameters (15, 16).

Ligands are generated in the active site of the target protein from 100 common organic groups (e.g., hydroxyl, carbonyl, furan, phenyl, etc.) deposited in the program's virtual combinatorial library. At each step of the growth algorithm, a random fragment is chosen from this library and attached to the part of the ligand already present in the active site (Fig. 1*A*). The energy of the newly formed adduct is evaluated by using the CombiSMoG potential, and the addition or rejection of the fragment is decided by a Boltzmann criterion (21), which biases the growth toward structures with low energy. The simplicity of the scoring function allows many candidate molecules (≈50,000 per day) to be generated quickly and, at the same time, evaluates their binding affinities accurately. The large number of available molecular fragments allows the algorithm to probe a range of structural types that is less constrained than that of experimental combinatorial methods, whereas the Monte Carlo growth method focuses the search toward strong binders and ensures that the ligands generate a sample of a large conformational space within the active site of the protein.

As a proof of principle for CombiSMoG, we used it to design new inhibitors for HCA metalloenzyme—a medically important (22) and structurally well defined protein for which a high quality x-ray structure (23) is available. We explored a family of ligands based on a well-characterized benzene sulfonamide moiety. We synthesized two of these molecules and compared their experimental binding affinities with those predicted computationally; we also obtained the crystallographic structures of the ligand–HCA complexes. Predicted and observed binding constants, and geometries, were in good agreement.

We chose the *para*-substituted benzene sulfonamide H_2NSO_2-C_6H_4-$CONH_2$ (BS) as the starting fragment for CombiSMoG design. The binding orientation of this moiety is well established, with the sulfonamide group (as the anion) coordinating to the zinc atom in the active site of HCA. This fragment has three advantages as a starting point for combinatorial simulations. (*i*) BS is only a moderately strong binder [$K_d = 120$ nM at 25°C and pH 7.5 (24)], and there is much room for improving binding affinities in the designed molecules. (*ii*) There are many well-

Abbreviations: HCA, human carbonic anhydrase II; CombiSMoG, combinatorial small molecule growth.

†Present address: Concurrent Pharmaceuticals, 1 Broadway, 14th Floor, Cambridge, MA 02142.

§To whom reprint requests should be addressed at the * address. E-mail: eugene@belok.harvard.edu.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

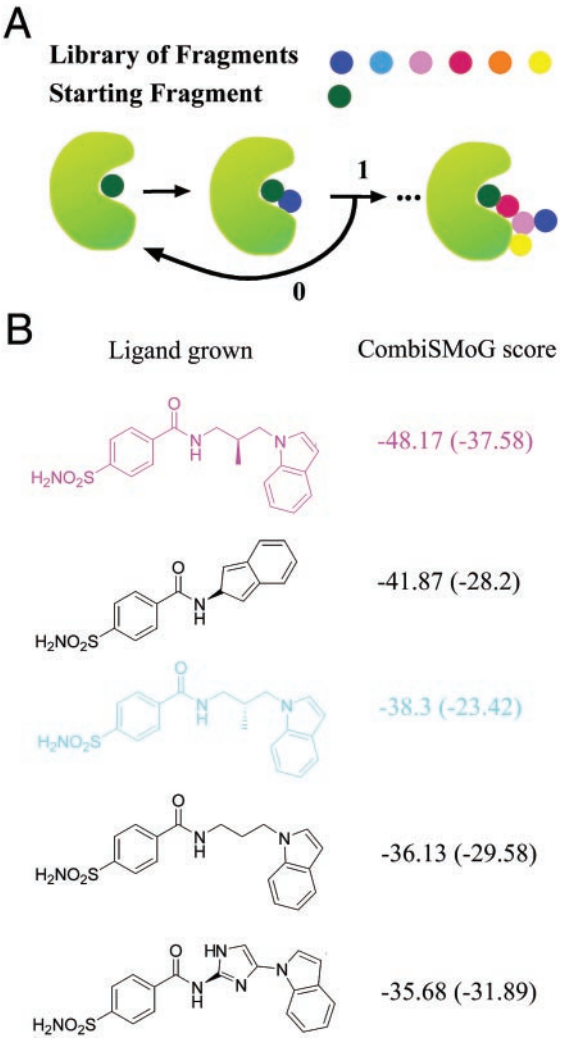

Fig. 1. (*A*) illustrates the principle of the CombiSMoG algorithm. The design begins with specifying a starting molecular fragment (dark green) within the binding region of the protein (light green); this fragment can be as small as a single hydrogen atom, or can consist of several heavy atoms. In the first step, a functional group (dark blue) from a diverse library of 100 common organic groups is joined by a single bond to the starting fragment. The new fragment is rotated around the newly formed bond in increments of 60°, and the conformation without steric clashes and with the lowest CombiSMoG score (which has the meaning of free energy) is chosen. The CombiSMoG score g per heavy atom of the newly formed molecule g_n is compared with that of the starting fragment g_s. If the difference in scores $\Delta g = g_n - g_s$ is less than zero, the newly formed molecule is always accepted, and if it is greater than zero, the probability of acceptance is proportional to $\exp(-\Delta g/T)$. The above sequence is repeated for each new fragment added to the currently accepted structure. The ligands are grown until a stop condition (usually a maximum number of heavy atoms) is matched. The structures and CombiSMoG scores of the five top-scoring inhibitors of HCA are shown in *B*. The scores of the structures minimized in CHARMM are in parentheses. The R and S stereoisomers that were subsequently synthesized and tested are colored violet and blue, respectively.

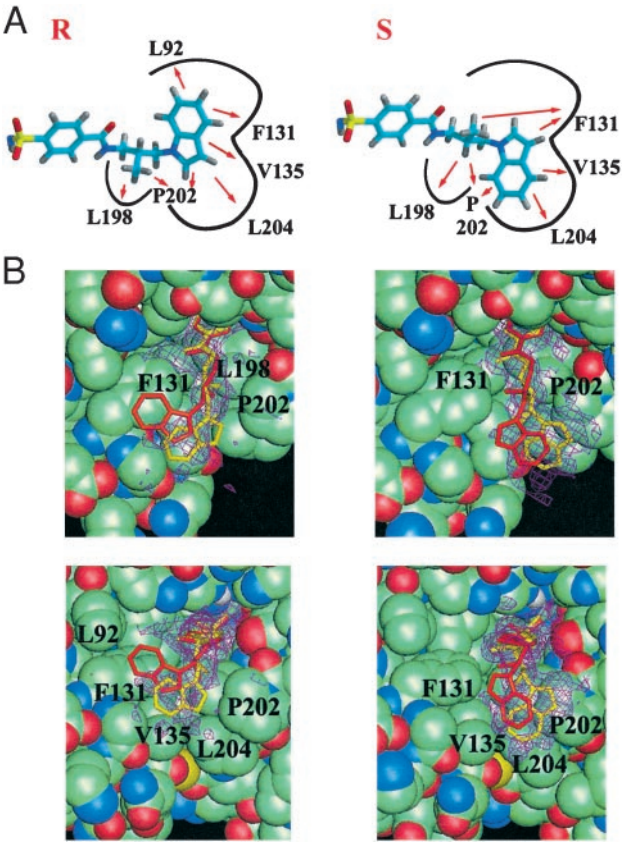

Fig. 2. *A* shows schematically the interactions of HCA II with the R (*Left*) and S (*Right*) stereoisomers grown by CombiSMoG. The surface of the protein is represented by a black curve, on which the approximate positions of the protein residues contacting the ligand are indicated. Three distinct binding pockets are separated by Pro-202 and Phe-131. The red arrows indicate the contacts between the ligand and the protein residues. The predicted and x-ray binding conformations of the R (*Left*) and S (*Right*) ligands are compared in *B*. The conformations predicted by CombiSMoG are colored red, and the x-ray difference electron density maps are shown in purple. These maps were calculated with Fourier coefficients $|Fo|-|Fc|$ and phases derived from the final model less the inhibitor and active-site solvent molecules; their contours are at 2 sigma. The fits to the electron density maps are shown in yellow. There are two different viewing angles: the top shows the contacts formed by methyl groups, and the bottom illustrates the interactions of the indole group with the protein. The protein residues making contacts with these groups are marked by letter codes.

characterized HCA inhibitors based on this moiety, against which we could calibrate the performance of CombiSMoG. (*iii*) By starting with the BS moiety, we avoid calculating interactions involving the zinc atom. These interactions involve quantum mechanical effects and cannot be accurately described by using the potential of CombiSMoG; our choice, therefore, simplifies the task of design.

The ligands were grown from one of the carboxamido hydrogens of benzene sulfonamide H_2NSO_2-C_6H_4-$CONH_2$. We specified the maximum ligand size as 30 heavy (i.e., nonhydrogen) atoms, and the number of ligands to be generated as 100,000. This analysis took ≈60 h on an Octane UNIX station. The several best ligands were further analyzed. Fig. 1*B* shows five top-scoring candidates generated by CombiSMoG. The structures of these molecules do not show any internal chemical incompatibilities and most are relatively easy to synthesize. To relieve conformational stress and Van der Waals clashes, we performed a local minimization of the five top-scoring ligand–HCA complexes by using the CHARMM force field (25), and recalculated the CombiSMoG scores of the candidates. The R-stereoisomer of the methyl-indole ligand (colored red in Fig. 1*B*) with the lowest score was the obvious candidate for synthesis. The indole group contacted a hydrophobic patch defined by Phe-131 and made favorable contacts with Phe-131 and Leu-92; the methyl group fitted nicely into another hydrophobic compartment where it interacted with Leu-198 and Pro-202 residues (Fig. 2*A*). An additional reason to test this compound was that its S stereoisomer (colored blue in Fig. 1*B*) also had a low

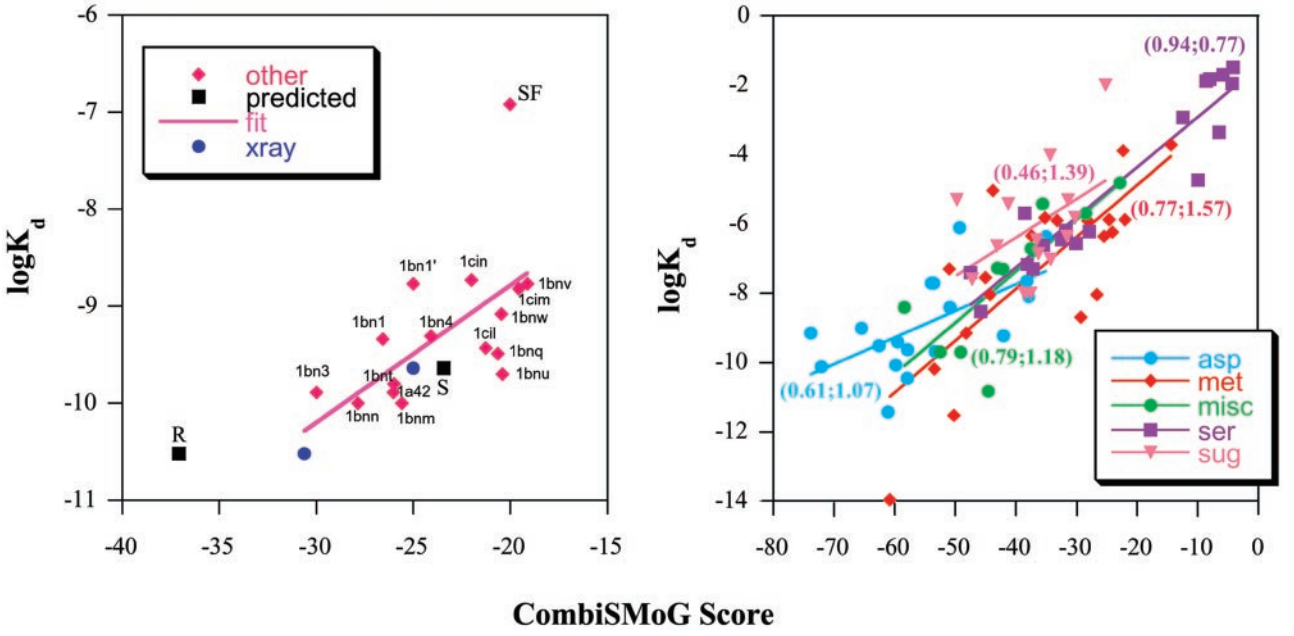

Fig. 3. (*Left*) Correlates the CombiSMoG scores with the logarithms of the binding constants K_d of all of the reported HCA inhibitors, for which binding assays were carried out at 37°C and pH = 7.5, and for which the x-ray structures of the HCA-inhibitor complexes were determined. The red solid diamonds correspond to inhibitors reported by others; the PDB access codes are given next to the markers. The black squares correspond to the structures of R and S inhibitors predicted by CombiSMoG, and the blue circles denote the x-ray structures of R and S. The line is a linear fit to data. Chemical structures of the ligands are given in the supporting information. (*Right*) This graph relates the CombiSMoG scores and the logarithms of the experimental binding constants of 80 ligands for five classes of proteins (asp, aspartic proteases; met, metalloproteins; ser, serine proteases; sug, sugar-binding proteins; misc, miscellaneous proteins). The list of the PDB access codes for all of the structures is included in the supporting information. The lines are linear fits to data. The correlation coefficients and the SDs from the linear fits for each family of proteins are given in parentheses next to the linear fits.

CombiSMoG score. In the S isomer, the methyl group was predicted to be in the same hydrophobic pocket as in R, but the indole moiety contacted another hydrophobic patch (Fig. 2*A*) defined by residues Phe-131, Val-135, Leu-204, and Pro-202. This pair of stereoisomers offered a test of the accuracy of the program in dealing with subtle structural differences.

The synthesis of the stereoisomers is described in supporting information, which is published on the PNAS web site, www.pnas.org. We measured the binding affinities of the isomers by using a competitive binding assay (26), in which the ligand to be tested displaces an inhibitor of known binding constant (dansylamide, K_d = 826 nM) from the binding pocket of the enzyme. The binding constants were measured to be K_d = 30 (±15) pM for the R stereoisomer and K_d = 230 (±45) pM for the S stereoisomer. To our knowledge, the R stereoisomer is the highest-affinity inhibitor of HCA II now known (27).

The crystal structures of the complexes were of good quality and showed good agreement between experimental and predicted binding modes. The overall structure of HCA in each of the enzyme-inhibitor complexes was effectively indistinguishable from that of the native enzyme, with the rms deviation of 258 C_α atoms of 0.19 Å in the R stereoisomer complex and 0.29 Å in the S. The electron density corresponding to the bound inhibitor molecule was well defined in both stereoisomers (Fig. 2).

The x-ray structures of the ligands were compared with the conformations predicted computationally (Fig. 2*B*). The position of the starting fragment is nearly identical for both R and S, with the nitrogen atom of the sulfonamide coordinating to zinc. Also, in both stereoisomers, the methyl groups are positioned in the pockets predicted by CombiSMoG; the distances between the predicted and observed locations of methyl carbons are 0.85 Å for the R stereoisomer and 2 Å for the S. The atoms of the aliphatic links between the starting fragment and the nitrogen atom of indole are well superimposed with an rms deviation (rmsd) of 0.52 Å in the R ligand and 0.88 Å in S. The indole groups in both stereoisomers contact the binding pockets predicted by the program. The R isomer is slightly displaced away from Phe-131 and toward Val-135 (rmsd = 2.84 Å). In the S ligand, the predicted orientation of indole is "flipped" compared with the x-ray structure, with the six-membered rings roughly aligned, but the five-membered rings pointing toward Phe-131 in the CombiSMoG prediction and toward Pro-202 in the x-ray; the rmsd of the indole positions in S is 3.15 Å.

Any computational algorithm for drug design is validated only if its predictions correlate with observed binding constants. Because the measured binding constant of the R-stereoisomer is ≈7 times smaller than that of the S isomer, we expect the CombiSMoG score of R to be lower than that of S. Indeed, for both predicted and x-ray structures, the scores for R are lower (−37.58 predicted; −30.5 x-ray) than for S (−23.42 predicted; −25.08 x-ray). Moreover, because the CombiSMoG scores have the meaning of free energies (13, 16), we also expect they should have a linear relationship to the logarithms of the binding constants (K_d). This relationship is observed for 20 ligands that have had their binding affinities measured under the same conditions as our R and S isomers, and for which the ligand–HCA x-ray structures are available (Fig. 3 *Left*). The ligands show roughly a linear dependence of logK_d on the computational score (correlation coefficient, 0.63; SD of logK_d from the linear fit, σ = 0.59).

We further verified that the logarithms of the experimental binding affinities correlate with the CombiSMoG scores for a diverse set of ligands for several structurally unrelated proteins (Fig. 3 *Right*). The correlation coefficients range from 0.46 for sugar-binding proteins to 0.94 for serine proteases, and the SDs from the linear fits are between 0.77 for serine proteases and 1.57 for metalloproteins. These results indicate that CombiSMoG reproduces—and thus can be expected to predict—the experimental binding affinities within a range of 1–2 orders of mag-

nitude. We note that the slopes and the intercepts of the linear fits can vary between the families of proteins, especially if the ligand–protein interactions involve quantum-mechanical effects for which the ''classical'' knowledge-based potential cannot account. For example, the inhibitors of HCA had lower values of $\log K_d$, by about 3 orders of magnitude, than ligands of other proteins with comparable CombiSMoG scores. We speculate that the quantum-mechanical nature of the interaction between the sulfonamide moiety of a ligand (common to all of the ligands in Fig. 3 *Left*) and the zinc atom in the active site of HCA is responsible for this effect.

CombiSMoG is a new computational tool for designing protein ligands that is based on the complementarity of the knowledge-based potential and the dynamic Monte Carlo algorithm for ligand growth. The simple coarse-grained potential allows rapid generation and evaluation of large numbers of structurally diverse—and, at the same time, energetically biased—virtual ligands. The Monte Carlo growth algorithm, in turn, benefits from searching a smoothed (coarse-grained) energy hypersurface without being ''jammed'' in too many local energy minima. Although many knowledge-based potentials and growth algorithms have been developed in recent years (28–30), CombiSMoG is the only computational method that uses these two components synergistically and permits realistic combinatorial chemistry, with usefully large numbers of ligands, to be carried out *in silico*. Our method was validated experimentally by designing and testing new potent ligands for a protein target. Strong correlations between CombiSMoG scores and binding constants of known ligands for a variety of structurally unrelated proteins suggest that our approach should also be successful in designing binders for proteins other than HCA. We believe that CombiSMoG will be able to generate nanomolar and subnanomolar ligands consistently, and that its predictions of binding affinity will have an uncertainty of approximately 1–2 orders of magnitude.

We thank Dr. Robert DeWitte for his help during the initial stage of the project. D.W.C. and E.I.C. were supported by National Institutes of Health Grants GM45614 and GM52126, respectively, and B.A.G. and G.M.W. were supported by the Department of Energy.

1. Panchagnula, R. & Thomas, N. S. (2000) *Int. J. Pharm.* **201,** 131–150.
2. Myers, P. L. (1997) *Curr. Opin. Biotech.* **8,** 701–707.
3. Lazo, J. S. & Wipf, P. (2000) *J. Pharmacol. Exp. Ther.* **293,** 705–709.
4. Li, J., Murray, C. W., Waszkowycz, B. & Young, S. C. (1998) *Drug Discovery Today* **3,** 105–112.
5. Babine, R. E. & Bender, S. L. (1997) *Chem. Rev.* **97,** 1359–1472.
6. Farber, G. K. (1999) *Pharmacol. Ther.* **84,** 327–332.
7. Kubinyi, H. (1997) *Drug Discovery Today* **2,** 457–467.
8. Kubinyi, H. (1997) *Drug Discovery Today* **2,** 538–546.
9. Martin, Y. C. (1997) *Perspect. Drug Discovery* **7/8,** 159–172.
10. Ajay, V. & Murcko, M. A. (1995) *J. Med. Chem.* **38,** 4953–4967.
11. Böhm, H.-J. & Klebe, G. (1996) *Angew. Chem. Int. Ed. Engl.* **35,** 2589–2614.
12. Böhm, H.-J., Banner, D. W. & Weber, L. (1999) *J. Comput. Aided Mol. Des.* **13,** 51–56.
13. DeWitte, R. S. & Shakhnovich, E. I. (1996) *J. Am. Chem. Soc.* **118,** 11733–11744.
14. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1997) *J. Mol. Biol.* **112,** 535–542.
15. Shimada, J., Ishchenko, A. V. & Shakhnovich, E. I. (2000) *Protein Sci.* **9,** 765–775.
16. Grzybowski, B. A., Ishchenko, A. V., DeWitte, R. S., Whitesides, G. M. & Shakhnovich, E. I. (2000) *J. Phys. Chem. B* **104,** 7293–7298.
17. Kollman, P. (1993) *Chem. Rev.* **7,** 2395–2417.
18. Böhm, H.-J. (1994) *J. Comput. Aided Mol. Des.* **8,** 243–256.
19. Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V. & Mee, R. P. (1997) *J. Comput. Aided Mol. Des.* **11,** 425–445.
20. Head, R. D., Smythe, M. L., Oprea, T. I., Waller, C. L., Green, S. M. & Marshall, G. R. (1996) *J. Am. Chem. Soc.* **118,** 3959–3969.
21. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953) *J. Chem. Phys.* **21,** 1087–1092.
22. Botrè, F., Gros, G. & Storey, B. T., eds. (1991) *Carbonic Anhydrase from Biochemistry and Genetics to Physiology and Clinical Medicine* (VCH, New York).
23. Eriksson, A. E., Jones, T. A. & Liljas, A. (1988) *Proteins Struct. Funct. Genet.* **4,** 274–282.
24. Jain, A., Whitesides, G. M., Alexander, R. S. & Christianson, D. W. (1994) *J. Med. Chem.* **37,** 2100–2105.
25. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, J. & Karplus, M. (1983) *J. Comput. Chem.* **4,** 187–217.
26. Chen, R. F. & Kernohan, J. C. (1967) *J. Biol. Chem.* **242,** 5813–5823.
27. Grunberg, S., Wendt, B. & Klebe, G. (2001) *Angew. Chem. Int. Ed. Engl.* **40,** 389–393.
28. Muegge, I. & Martin, Y. C. (1999) *J. Med. Chem.* **42,** 791–804.
29. Nobeli, I. Mitchell, J. B. O., Alex, A. & Thornton, J. M. (2001) *J. Comp. Chem.* **22,** 673–688.
30. Bohacek, R. S. & McMartin, C. (1995) *Comput. Aid. Mol. Des. ACS Sym. Ser.* **589,** 82–97.