Acc. Chem. Res. **2002,** *35,* 261−269

ARTICLES

From Knowledge-Based Potentials to Combinatorial Lead Design in Silico

BARTOSZ A. GRZYBOWSKI,*,† ALEXEY V. ISHCHENKO,*,† JUN SHIMADA,‡ AND EUGENE I. SHAKHNOVICH†,‡

Concurrent Pharmaceuticals, 1 Broadway, 14th floor, Cambridge, Massachusetts 02142, and Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138

Received December 13, 2001

ABSTRACT

Computational methods are becoming increasingly used in the drug discovery process. In this Account, we review a novel computational method for lead discovery. This method, called CombiSMoG for "combinatorial small molecule growth", is based on two components: a fast and accurate knowledge-based scoring function used to predict binding affinities of protein−ligand complexes, and a Monte Carlo combinatorial growth algorithm that generates large numbers of low-free-energy ligands in the binding site of a protein. We illustrate the advantages of the method by describing its application in the design of picomolar inhibitors for human carbonic anhydrase.

Identification of high-affinity binders ("leads") for protein targets is the first step in the process of drug discovery and development and is usually a laborious and costly enterprise[1]. In some cases, up to several thousands of compounds need to be synthesized before a single lead is found, and even then, there is no guarantee that it would have the toxicological and pharmacological properties required of a drug.

It was recognized in the 60s, that computer-based methods can be of help in the discovery of leads and can potentially eliminate chemical synthesis and screening of many irrelevant compounds.[2−4] An ideal computational method for lead discovery should be able to generate structurally diverse leads rapidly and should give the estimates of binding affinities that would correlate with experimental values. The first requirement, that is, generation of chemical diversity in silico, is easily achieved using existing computational resources and algorithms: putative ligands can be either extracted from large databases of compounds,[5,6] or they can be "grown" computationally by joining molecular fragments[7−9] (or atoms[10]) stored in the computer's memory. The second prerequisite, accurate prediction of binding affinities (or, equivalently, binding free energies), has proven to be a much more difficult task. Because of the multitude of energetic and entropic factors involved, the thermodynamics of binding cannot be analytically modeled without first simplifying the problem. Computational methods that attempt to design leads vary in the nature and in the degree of the simplifying assumptions they use.[11]

In the first part of this Account, we categorize and briefly discuss the existing computational approaches to lead design. We then focus on a class of the so-called knowledge-based methods that have their roots in protein-folding studies but have recently become a promising strategy for designing and evaluating leads. We describe in detail a knowledge-based algorithm we developed, called CombiSMoG for "combinatorial small molecule growth", that incorporates the philosophy of combinatorial synthesis into computational drug design, and can rapidly generate large numbers of biased libraries of virtual leads and give accurate estimates of their binding affinities. Application of CombiSMoG to design extremely potent inhibitors for human carbonic anhydrase described in the last part of this Account illustrates the strengths and the limitations of our approach.

1. State of the Art in Computational Methods for Lead Discovery (CMLD). The CMLD can be divided into two broad categories: those that do not require the knowledge of the biological target, and those that do. The first class of methods aims at correlating structural features of a series of known compounds with their biological activities and derives from these correlations multiparameter empirical equations that are subsequently used to guide the design of new leads. Early quantitative structure−activity relationship (QSAR) methods related biological activity to the presence (or absence) of functional groups in a series of structurally related compounds (Free−Wilson model[12]), or to the physicochemical properties (lipophilicity, electronic properties) of the compounds in the training set (Hansch analysis[13]). More recently, three-dimensional

Bartosz A. Grzybowski received his A.M. from Yale University and his Ph.D. from Harvard University (with G. M. Whitesides). He is currently Director of Research at Concurrent Pharmaceuticals. His scientific interests include self-organization in multibody chemical and biochemical systems, theory of molecular recognition, and drug design.

Alexey V. Ishchenko received his A.M. from Kharkiv State University, Ukraine (with O. N. Kalugin) and his Ph.D. from Harvard University (with E. I. Shakhnovich) and is currently a Computational Chemist at Concurrent Pharmaceuticals. His research focuses on scoring functions and computational ligand design.

Jun Shimada received his A.M. and is about to complete a Ph.D. program with E. I. Shakhnovich at Harvard University. His research includes knowledge-based methods, computational ligand design, and protein folding.

Eugene I. Shakhnovich is Professor in the Department of Chemistry and Chemical Biology, Harvard University, and a cofounder of Concurrent Pharmaceuticals. His primary interests are protein folding, evolution and design; protein−ligand interactions and drug design; structural genomics and theoretical studies of complex (nonbiological) systems, such as polymers, spin glasses, and structural glasses.

* Address correspondence to either author. E-mails: bgrzybowski@concurrentpharma.com (Bartosz Grzybowski) and aishchenko@concurrentpharma.com (Alexey Ishchenko).

† Concurrent Pharmaceuticals.

‡ Harvard University.

10.1021/ar970146b CCC: $22.00
Published on Web 04/26/2002

QSAR methods have been developed[14,15] in which chemically related molecules of known activities are superimposed, placed on a 3D grid, and a discretized activity "field" around them is constructed by calculating their interactions with imaginary probe atoms placed in the nodes of the grid. Although QSAR models reproduce binding affinities of ligands in many training sets more accurately than other methods, and although they scored spectacular success in some applications (e.g., a priori prediction of binding orientation of dihydrofolate to dihydrofolate reductase[16]), they have three major shortcomings: (i) there must already exist many leads for the target under study to allow development of the structure−activity relationships; (ii) the equations are parametrized for one target and do not apply to another, so that they are not transferable; and (iii) the QSAR methods are of only limited use in understanding the nature of protein−ligand interactions and thermodynamics of binding and place emphasis on efficiency of lead design.

Structure-based (SB) approaches[17−19] overcome many of the limitations of QSAR, albeit at the expense of knowing the three-dimensional structure of the target. These methods aspire to develop a *general* theoretical description of the protein−ligand interactions that would enable an a priori design of new leads for an arbitrary biological target. At the heart of every SB method lies the so-called scoring function (also referred to as the force field or potential), that is, a mathematical function whose values are (or, at least in principle, should be) proportional to the binding affinities of the leads. A good scoring function should be able to give reliable estimates of binding affinities of structurally diverse leads for different protein targets. There are three major classes of scoring functions[20] that meet this criterion: (1) empirical all-atom force fields, (2) "master equation," and (3) knowledge-based (KB) functions. The KB scoring functions will be discussed in detail in Section 2.

Empirical force fields (e.g., OPLS,[21] AMBER,[22] CHARMM,[23] and others) account for all atom−atom interactions, which they calculate by summing bond, angle, dihedral, electrostatic, and van der Waals terms, and are parametrized against ab initio calculations, or structural, dynamic, and thermodynamic properties of small molecules or peptides. These force fields do not per se provide the free energies of binding, but rather they provide the energies of protein−ligand interaction in a given conformation. The free energies are usually calculated by thermodynamic integration or free-energy perturbation methods.[24,25] Such calculations are in many cases accurate, but they are also computationally costly and preclude the possibility of screening large numbers of potential leads. Recently, there has been considerable interest in developing faster, approximate free-energy methods,[26,27] that would estimate binding free energy from the average interaction energies of the ligand bound to the protein and unbound in the solvent. The accuracy of these methods is, however, yet to be determined.

In the master equation methods (e.g., LUDI,[28] VALIDATE,[29] Eldridge, et al.[30]), the binding free energy is arbitrarily decomposed into various enthalpic and entropic terms (e.g., hydrogen bonding, number of flexible bonds in ligands, and area of lipophilic contact) that are represented by simple functional forms incorporating free parameters. The parameters are obtained by optimization procedures maximizing the correlation between the computed and experimental binding free energies of a set of complexes with known structures and binding constants. These methods are fast and give relatively high correlations for diverse training sets. It is unclear, however, how accurately these approaches predict the binding affinities in complexes that are not included in these sets.

2. Knowledge-Based Potentials (KBP). Knowledge-based methods[31] derive free energies of molecular interactions from structural information contained in databases of known protein−ligand complexes. The first KBPs were developed as early as the 70s[32] in the context of protein-folding studies and related the frequencies of occurrence of particular structural features X in globular proteins to their "effective" free energies via a Boltzmann-like relationship: $N_X \sim \exp(-F_X/kT)$, where by "effective free energy" we mean the energy of interaction between X and the rest of the protein averaged over all solvent configurations; this exponential dependence has been substantiated by several experimental studies.[33]

Rapid advances in protein crystallography and NMR techniques during the past decade allowed determination of large numbers of three-dimensional structures of protein−ligand complexes and, consequently, construction of meaningful KBPs describing protein−ligand interactions (SMoG,[34] PMF,[35] BLEEP,[36] Drugscore,[37] and others). All existing methods for deriving such potentials relate the occurrences extracted from databases to energies in a Boltzmann-like fashion (Figure 1); in only a very few instances, however, has the applicability of this exponential relationship been discussed[38,39] or justified.[40] Since the statistical−mechanical arguments used in protein folding are *not* transferable to protein−ligand interactions, the relevance of Boltzmann-like law to protein−ligand complexes requires a separate derivation. A sketch of such a derivation is given in the next section.

A. Boltzmann-Like Statistics of Interatomic Contacts. Consider a database of χ protein−ligand complexes. Two atoms, σ_p of the protein and σ_l of the ligand, are said to form an intermolecular contact (σ_p,σ_l), if their separation is smaller or equal to a cutoff distance R_c; the total number of contacts in the database is denoted $N\chi$. We wish to relate the free energy $F(\sigma_p,\sigma_l)$ of a contact (σ_p,σ_l) to the probability of its occurrence in the database $P(\sigma_p,\sigma_l) = N(\sigma_p,\sigma_l)/N_\chi$, where $N(\sigma_p,\sigma_l)$ is the number of (σ_p,σ_l) contacts in the database. We select a set of complexes that have approximately equal protein−ligand interaction free energies F_I. Let the number of complexes with the same F_I be denoted N_{FI}, and the number of structures within this set with interaction (σ_p,σ_l) present be $N_{FI}(\sigma_p,\sigma_l) = N_{FI}(\sum_{ij\neq pl}F(\sigma_i,\sigma_j) \approx F_I - F(\sigma_p,\sigma_l))$. The right-hand side of the last equation expresses the fact that $N_{FI}(\sigma_p,\sigma_l)$ is equal to the number of complexes in which the sum of interactions other than (σ_p,σ_l) is equal to $F_I - F(\sigma_p,\sigma_l)$. Now three major

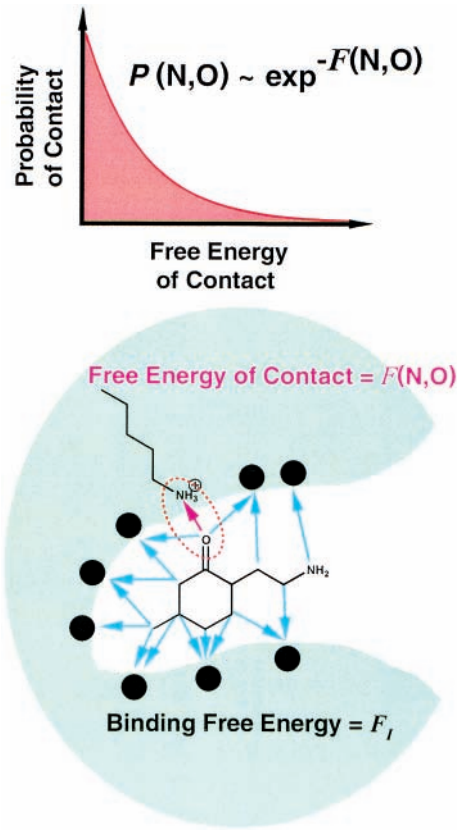

FIGURE 1. Origin of the Boltzmann-like statistics in protein−ligand complexes. The protein−ligand interaction free energy F_I is the sum of the free energies of all pairwise contacts (indicated by blue arrows). A particular contact (σ_p,σ_l) (represented by the magenta arrow) has free energy F(N,O). The number of ways in which the remaining contacts can be arranged to give a stable complex is proportional to the frequency of observing contact (σ_p,σ_l) in the database, and decreases exponentially with increasing F(N,O).

assumptions are made: (i) that the database is large and structurally diverse; (ii) that no single contact dominates the interaction free energy, that is, $F_I \gg F(\sigma_p,\sigma_l)$; and (iii) that the total free energy can be expressed as a sum of pairwise energies (eq 1).

$$F_I = \sum_{\sigma_p} \sum_{\sigma_l} N(\sigma_p,\sigma_l) F(\sigma_p,\sigma_l) \qquad (1)$$

If the first condition is fulfilled, $N_{FI}(F_I - F(\sigma_p,\sigma_l))$ can be assumed to be proportional to the number of ways in which the contacts (σ_p,σ_l), $(i,j \neq p,l)$, can be arranged to give a stable complex with contact (σ_p,σ_l) present and having total interaction free energy F_I: $\Omega_{FI}(F_I - F(\sigma_p,\sigma_l))$. The second supposition allows expansion of the logarithm of $\Omega_{FI}(F_I - F(\sigma_p,\sigma_l))$ about F_I, as in eq 2. Because the derivative is only a slowly varying function of free energy ($\sim$d ln $n!$/d$n \approx \ln n$, where n is the number of atoms in a ligand, it is well approximated as a constant, β. After rearranging, and noting that $\Omega_{FI}(F_I - F(\sigma_p,\sigma_l))/\Omega_{FI}(F_I)$ is proportional to the probability of observing contact (σ_p,σ_l) in the set of structures having interaction free energy F_I, we obtain $P((\sigma_p,\sigma_l)|F_I) \propto \exp(-\beta F(\sigma_p,\sigma_l))$. Summation over all values of interaction energies leads to the exponential relationship between the probability $P(\sigma_p,\sigma_l)$ of observing contact (σ_p,σ_l) in the database and the free energy of this contact $F(\sigma_p,\sigma_l)$ (eq 3; P_{ref} is a constant that will be discussed in detail in the next section).

$$\ln \Omega_{FI}(F_I - F(\sigma_p,\sigma_l)) = \ln \Omega_{FI}(F_I) - \frac{\mathrm{d} \ln \Omega_{FI}(F_I)}{\mathrm{d}F}\Big|_{F=FI} \cdot F(\sigma_p,\sigma_l) \qquad (2)$$

$$P((\sigma_p,\sigma_l)) = \sum_{F_I} P(F_I) p((\sigma_p,\sigma_l)|F_I) = P_{ref} \exp(-\beta F(\sigma_p,\sigma_l)) \qquad (3)$$

The derivation shows that the KB Boltzmann-like statistics is applicable only if the database from which it is obtained meets certain criteria. Overlooking these criteria may lead to meaningless potentials:

(i) If a database is not large or structurally diverse, the proportionality between N's and Ω's cannot be assumed. Small numbers of complexes having a particular feature will lead to overestimating the energies of these features.

(ii) If interactions between the ligand and the protein are dominated by one (or few) contacts, $F_I \sim F(\sigma_p,\sigma_l)$ and the Taylor expansion is not valid. The energetics of such complexes is not well-described by KB methods.

Specifically, our studies showed that a dataset of $\sim$300−400 complexes chosen randomly from the Protein Data Bank, was large and diverse enough to derive a satisfactory KBP.[41] The majority of the complexes in this set did not have an energetically dominant interaction (i.e., an interaction whose magnitude would exceed $\sim$10% of the total binding free energy). In metalloprotease complexes, however, metal−ligand interaction accounted for $\sim$50% of binding free energy; our scoring function did not reproduce accurately the affinities of several metalloprotease ligands.

B. Normalization of Probabilities and the Importance of the Reference State. Determination of the constant of proportionality P_{ref} in eq 3 is necessary for deriving a meaningful KBP. The value of P_{ref} must be such that the probabilities P derived from the database are properly normalized and that the zero of free energy (reference state) this constant defines is physically justified.

In our definition, the hypothetical reference state is a purely random mixture of connected protein and connected ligand atoms that do not interact, that is, $F_{ref}(\sigma_p,\sigma_l) = 0$ for all (σ_p,σ_l) contact pairs, given preserved-atom-type composition and connectivity. In the early version of our potential,[34,42] SMoG96, we chose a simple approximation for the reference state probability as average of contact probabilities: $P_{ref} = \langle P(\sigma_p,\sigma_l)\rangle_{(\sigma_p,\sigma_l)}$.

Although SMoG96 performed satisfactorily in several test cases, we noticed that it handled hydrophobic interactions better than polar ones. We reasoned that this bias might be a consequence of statistical effects present in a database that are independent of energetic effects. For example, if the database contains many complexes in which protein binding sites are deep and hydrophobic, the statistics of observed contacts will mirror the distribution of σ_p and will be skewed toward hydrophobic

contacts, resulting in apparent strong attraction between nonpolar atoms. To account for such nonenergetic effects (or, equivalently, to account for the *imperfection* of the database; cf. discussion in Section B), we redefined the reference state probabilities[43] as a product of the normalization constant C and the nonenergetic contribution to the observed probabilities $S(\sigma_p,\sigma_l)$. Normalization of probabilities gave eq 4, in which we retained, in contrast to SMoG96 function, the average free energy as a parameter.

$$\ln C = \left\langle \ln \frac{P(\sigma_p,\sigma_l)}{S(\sigma_p,\sigma_l)} \right\rangle_{(\sigma_p,\sigma_l)} + \langle F(\sigma_p,\sigma_l)\rangle_{(\sigma_p,\sigma_l)} \qquad (4)$$

Finally, because the nonenergetic term S depends on the numbers of protein and ligand atoms in the database that give rise to a particular type of contact, we approximated it as $S(\sigma_p,\sigma_l) = N(\sigma_l)^{\alpha}N(\sigma_p)^{\beta}$, where $0 < \alpha,\beta < 1$ were parameters. The last two equations specified the reference state of the SMoG2001 potential[41,43] that is used in the CombiSMoG method.[44]

Our methodology uses a simple, "coarse-grained" definition of contacts based on two distance intervals (0−3.5 and 3.5−4.5 Å) over which the statistics are collected separately. Recently, more elaborate, distance-dependent ("smooth-grained") potentials have been constructed, for example, the PMF function of Muegge and Martin.[35] In deriving such potentials, care must be taken to ensure that there are enough contacts in all volume elements to permit a KB approach. It is interesting to note that smooth-graining does not necessarily improve the potential; in predicting binding affinities, PMF performs better than SMoG96, but slightly worse than SMoG2001.[41] Since coarse-grained SMoG2001 and smooth-grained PMF are derived from similar number of complexes, it seems that it is the proper definition of the reference state and not the geometrical features of the potential that is a crucial feature of a good KBP.

C. From Probabilities to a Working Potential. Before the formulas for contact free energies can be applied to construct an actual potential, one needs to (i) define the scoring function, (ii) specify the database from which the probabilities are derived, and (iii) identify the α, β, and $\langle F(\sigma_p,\sigma_l)\rangle$ parameters. The first two tasks are relatively easy. In SMoG, we defined the scoring function F as a sum of pairwise interaction energies (eq 5; $\Delta(\sigma_p,\sigma_l)$ is 1 if atoms σ_p and σ_l are in contact and is 0 otherwise).

$$F = \sum_p \sum_l F(\sigma_p,\sigma_l)\Delta(p,l) \qquad (5)$$

The Brookhaven Protein Database (PDB) was an obvious source of 3D structures, and we extracted from it 750 structurally diverse complexes. We classified the atoms on the protein and on the ligand into 14 types according to the element type, hybridization, partial charge, or donor/acceptor properties. After specifying R_c at 4.5 Å (corresponding to the second coordination shell of water around an atom), the database statistics were collected, and the contact energies were derived. We briefly mention that the choice of the cutoff radius ensures that the influence of solvent (water) on the contact distribution is implicitly taken into account.

Determination of the parameters of the model required a separate self-consistent optimization procedure that is discussed in detail elsewhere.[43] In short, a set of proteins was chosen from the PDB, and ligands for these proteins were created *computationally*. The generated complexes were divided into two groups: a "toy" database and a "test" database. An arbitrary form of a coarse-grained KB potential was chosen ("true potential"), and scores of the ligands in both groups were calculated according to this potential. Next, the contact energies were back-extracted from the toy database via a Boltzmann-like relationship with SMoG2001 reference state and specified values of α, β, and $\langle F(\sigma_p,\sigma_l)\rangle$ parameters. The obtained "derived potential" was then used to recalculate the scores of the ligands in the test database. If the formulation by which the derived potential was extracted is valid, the derived scores should correlate well with the true scores of the test database complexes, and the optimal values of α, β, and $\langle F(\sigma_p,\sigma_l)\rangle$ should be such that they maximize this correlation. We found that the optimal values were $\alpha = \beta = 0.9$ and $\langle F(\sigma_p,\sigma_l)\rangle = 0$.

D. Comparison with Other Methods. A coarse-grained potential, such as SMoG, is ultimately a two-dimensional array containing the contact free energies between atoms of different types. Because SMoG and other similar methods do not have to evaluate arithmetic functions or perform ensemble averages to calculate binding free energies, they are several times faster than "master equation" methods and orders-of-magnitude faster than free-energy perturbation methods.

Knowledge-based potentials have one potential advantage over empirical scoring functions. There are only a small number of complexes for which both the structure and the binding constant are publicly available. Since many of these complexes have similar ligands, it is unclear whether there is sufficient diversity to adequately represent all features of the binding process when deriving an empirical scoring function. In contrast, only structural information is necessary for deriving knowledge-based parameters, meaning that a larger and more diverse set of complexes can be utilized. For this reason, knowledge-based potentials may be more transferable than empirical scoring functions to complexes not found in the training database.

We compared the accuracy of SMoG2001 to that of a well-known empirical scoring function (LUDI) and several other KBPs (PMF and DrugScore) in our recent work.[41] The performance of SMoG2001 (measured by the correlation coefficients between computed and experimental binding affinities of protein−ligand complexes) was superior to that of LUDI and PMF and comparable to DrugScore.

The free-energy perturbation methods are certainly more accurate than knowledge-based ones, but their scope is somewhat limited, since they are unable to rank binding affinities of structurally diverse ligands. These methods seem to be most useful in optimizing existing

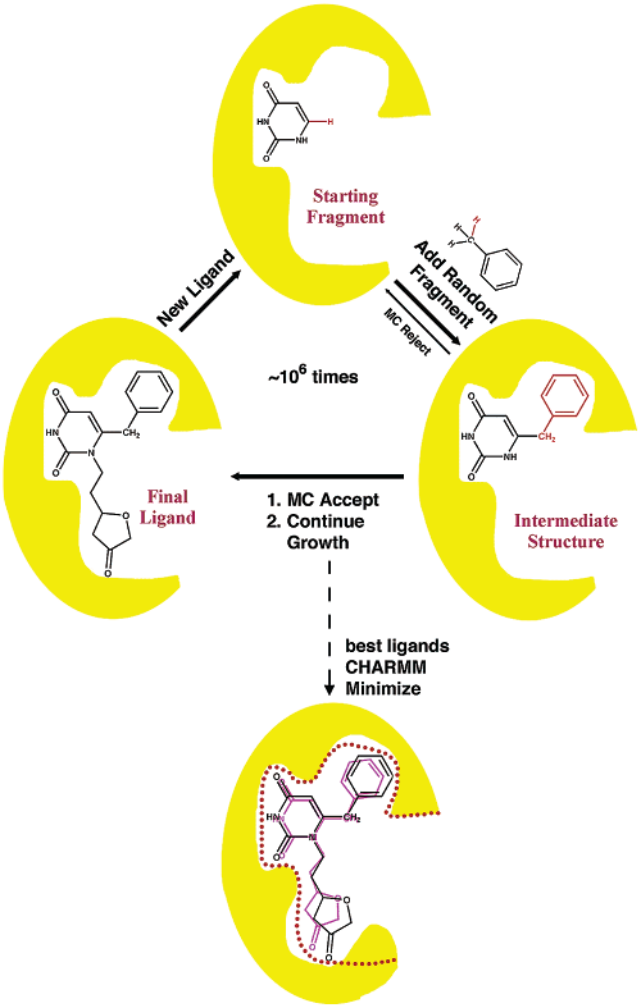

FIGURE 2. CombiSMoG growth algorithm.

leads by introducing minor structural changes, rather than in finding leads de novo.

3. KBP + Growth Algorithm = CombiSMoG. A scoring function can calculate binding free energy of a ligand in an existing complex but it cannot, obviously, create new ligands. What is needed for a complete lead-discovery method is an algorithm for lead generation. There are two types of such algorithms:[45] (i) docking of known organic molecules and (ii) de novo design that generates new chemical structures in the binding site of a protein. The latter is usually accomplished in two ways: either by placing chemical fragments in energetically favorable regions and connecting them by linking groups to form a molecule, or by growing the molecule in the binding site by sequentially adding chemical fragments. Several comprehensive reviews of the lead-generation methods have been recently published[5−10,45], and they will not be discussed in detail here. Instead, we will focus on the ligand growth algorithm used in conjunction with SMoG2001 scoring function, and will place emphasis on the complementarities of these two components.

A. Dynamic Monte Carlo Growth Algorithm. Figure 2 illustrates the procedure to generate ligands in the active site of the protein. Depending on a particular application, the molecule can be grown either completely de novo or from a starting fragment. In the first case, a hydrogen molecule is placed at a random location within the binding site and serves as a starting fragment. At each step of the growth, a random fragment is chosen from a diverse library of common chemical groups (e.g., phenyl, carbonyl). The library contains ~100 such groups, but can be restricted if required. The chosen fragment is added to the already grown part of the molecule by creating a bond ("link") between two randomly chosen heavy atoms. This newly added moiety is rotated around the link in increments of usually 5° within energetically low torsional space, and the SMoG2001 scores are calculated for each conformation. The conformation with the lowest score is retained, and its score is compared to that of the molecule before addition of the fragment. If the addition lowers the score of the ligand, it is accepted; if the score increases, the addition is accepted conditionally with probability proportional to $\exp(-\beta\Delta F)$, that is, the acceptance of the addition is described by the Metropolis criterion.[46] The acceptance of "high-free-energy" fragments helps overcome free-energy barriers and permits growing ligands that join "low-free-energy" pockets in the binding site. The fragments are added and evaluated until the grown ligand reaches the prespecified size (typically, 30−40 heavy atoms). The cycle is repeated many times to generate, typically, $\sim 10^5-10^6$ ligands, from among which the top-ranking ones (~50) are selected for further analysis and are subjected to local energy minimization using the all-atom CHARMM force field.[23] This minimization relieves the conformational strain and possible van der Waals clashes in the generated ligands.

B. The Scoring Function and the Growth Algorithm Work in Unison. Every de novo method for lead design requires generation of many candidate ligands. If only few ligands are created, it is highly probable that their scores will not represent the lowest possible values (and, thus, optimal designs) for a given target; the chances of identifying high-affinity binders increase with an increasing number of putative designs. The Monte Carlo (MC) growth algorithm we use not only searches the structural space efficiently, but also biases the search toward low-free-energy (high-affinity) binders. This bias ensures that the irrelevant, high-free-energy ligands are not screened, and results in substantial savings in the CPU time. Even with the MC growth, however, the number of molecules that need to be created and evaluated to obtain a statistically significant sample of the structural space is still very large ($>\sim 10^5$) and requires the use of a fast scoring function. SMoG certainly meets this demand, because it allows evaluation of large numbers of candidate molecules in short times (~100 000/day on Octane UNIX workstation). In other words, the fast SMoG potential is an ideal one to use in conjunction with the MC ligand generation method.

The opposite is also true: the MC growth benefits from searching the free energy hypersurface defined by the SMoG potential. Because of the binary definition of contacts, this hypersurface does not have too many deep

and narrow free energy minima that would "trap" the ligand growth. If the potential were made less coarse-grained (like the distance-dependent PMF scoring function), one would expect the growth algorithm to be frustrated by local free energy minima, and consequently, the structural space would not be searched efficiently.

C. Combinatorial Lead Design in Silico. The SMoG2001 potential and the MC ligand growth algorithm are the two components of the CombiSMoG lead design package. The designation of this method as "combinatorial" derives from its ability to probe a range of structural types that is comparable to and less constrained than that of experimental combinatorial methods. Indeed, the library of 100 molecular fragments creates a structural space of $\sim M^{100}$ compounds that can potentially be evaluated (M is the number of fragments per ligand). CombiSMoG can function either as a de novo combinatorial method or as a tool for optimizing existing molecules when these molecules are used as starting fragments. The major limitation of CombiSMoG as compared to wet combichem is that the MC growth method does not guarantee the synthesizability of the generated molecules. The synthetic feasibility of the candidates has to be determined by an organic/medicinal chemist. We are currently working on developing algorithms that would prevent formation of hard-to-make bonds during the ligand generation.

In summary, to design new leads using CombiSMoG, one has to (i) specify a starting position/fragment for ligand growth, (ii) generate large numbers of candidate molecules, and (iii) examine the top-scoring ligands for their synthetic feasibility and structural integrity. What remains to be established is whether CombiSMoG's binders are equally, and to what extent, potent in silico and in vitro.

4. CombiSMoG at Work. As a proof-of-principle for CombiSMoG, we used it to design new inhibitors for human carbonic anhydrase II (HCA) metalloenzyme[44], a medically important, and structurally well-defined protein that does not undergo marked conformational changes upon ligand binding (CombiSMoG does not account for protein rearrangement).

We chose the para-substituted benzene sulfonamide $H_2NSO_2-C_6H_4-CONH_2$ (BS) as a starting fragment for CombiSMoG design. The binding orientation of this moiety is well-established, with the sulfonamide group (as an 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) It is a relatively weak binder (K_d = 120 nM at 25 °C and pH 7.5)[44] so that binding affinities of the designed molecules could be significantly enhanced. (ii) By starting with the BS moiety, we avoid calculating interactions involving the zinc atom. These interactions involve quantum mechanical effects that are poorly described by CombiSMoG potential (cf. Section 2A). (iii) There are many well-studied sulfonamide-based inhibitors of HCA[47] against which we could calibrate CombiSMoG's performance.

The growth algorithm generated 100 000 candidate ligands, from which the top 20 were chosen for further analysis. Interestingly, the best and the fourth-best ligands were stereoisomers of *N*-(3-indol-1-yl-2-methyl-propyl)-4-sulfamoyl-benzamide (Figure 3A), differing in the chirality of the carbon atom β to the indole ring. These molecules did not show any internal chemical incompatibilities, and were deemed to be relatively easy to synthesize. On the basis of the correlation between the CombiSMoG scores of known HCA ligands and their experimental binding affinities, we expected the *R* isomer to bind with approximately picomolar activity, whereas the *S* isomer should be a subnanomolar binder. The two isomers were predicted to bind in different orientations. In the *R* isomer, CombiSMoG placed the indole group in the hydrophobic pocket defined by Phe131 and Leu92, but in the *S* isomer, the indole moiety was predicted to contact the hydrophobic patch defined by Phe131, Val135, Leu204, and Pro202 (Figure 3A). This pair of stereoisomers offered a challenging test of the accuracy of the program in dealing with subtle structural differences.

Both compounds were synthesized, their binding affinities were measured, and the X-ray structures of the complexes were obtained. The predicted and the observed binding constants and binding orientations were in excellent agreement (Figure 3B). The binding constants were K_d = 30 (±15) pM for the *R* stereoisomer and K_d = 230 (±45) pM for the *S* stereoisomer. The positions of the atoms in predicted conformations were superimposed with those of X-ray structures with RMSD <1 Å, except for the methyl carbon in the *S* stereoisomer (2 Å) and the indole group (2.84 Å in *R*, 3.15 Å in *S*). To our knowledge, the *R* stereoisomer is the highest-affinity inhibitor of HCA II now known.[44,47]

CombiSMoG would be further validated if it were shown that it consistently designs potent ligands for a variety of protein targets. Several experimental projects are under way in our laboratory, but we will be able to report their outcome only when the syntheses and assays are completed. Unfortunately, those syntheses and assays are much more time-consuming than CombiSMoG design. In the absence of the necessary experimental data, one can use the existing protein−ligand complexes to study the correlation between CombiSMoG's scores and the experimentally observed binding affinities: if the CombiSMoG scores have the meaning of true binding free energies, they should have a linear relationship to the logarithms of the binding constants (K_d). Such a linear relationship would indicate that CombiSMoG is capable of predicting binding affinities accurately, irrespective of the target.

Figure 4 shows the correlation between the logarithms of experimental binding constants and CombiSMoG scores of 119 complexes from eight structurally diverse subsets of proteins. Within each subset, the correlation coefficients range from r = 0.19 for endothiapepsin complexes to r = 0.84 for serine proteases, and the standard deviations from the linear fits are between σ = 1.0 for endothiapepsin complexes and σ = 1.7 for metalloproteases. These results indicate that CombiSMoG reproduces and potentially predicts the experimental binding affinities within a subset

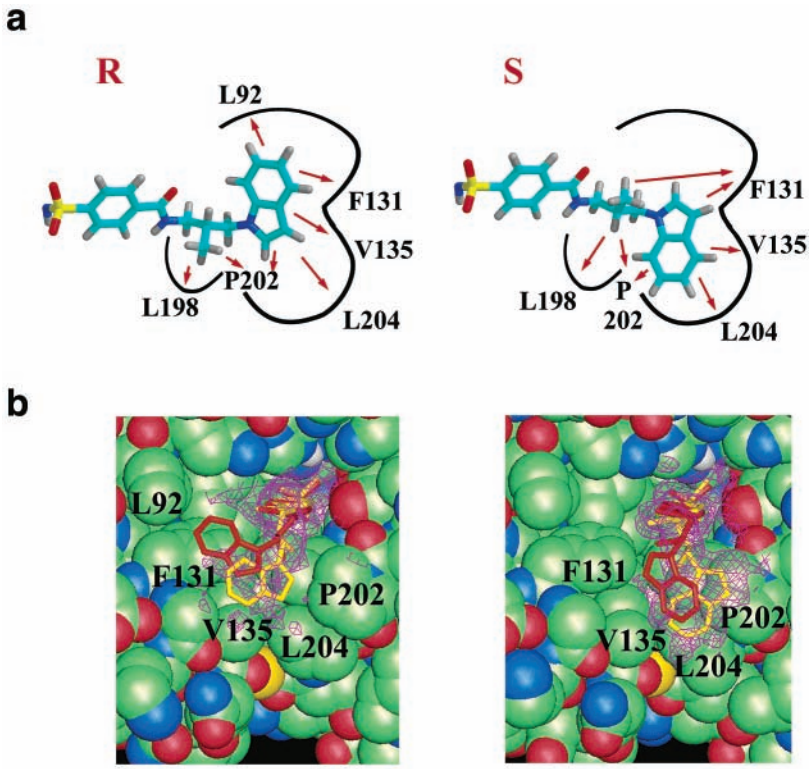

FIGURE 3. (a) Schematic representation of the interactions of the HCA with the *R* (left) and the *S* (right) stereoisomers of the *N*-(3-indol-1-yl-2-methyl-propyl)-4-sulfamoyl-benzamide ligand grown by CombiSMoG. The surface of the protein is represented by a black curve. Red arrows indicate the contacts between protein residues and ligand atoms. The predicted (red) and X-ray (yellow) binding conformations of the *R* and *S* ligands are shown in part b. The X-ray difference electron density maps at 2σ are colored purple.

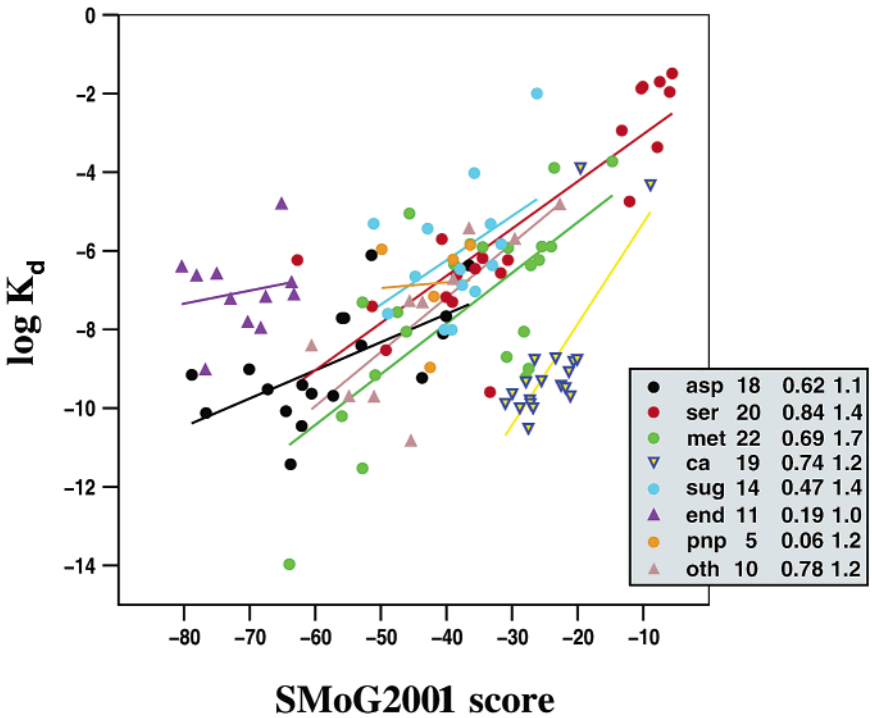

FIGURE 4. CombiSMoG scores and the logarithms of the experimental binding constants for the testing set of 119 complexes from eight subsets of proteins (aspartic proteases "asp", serine proteases "ser", metalloproteases "met", carbonic anhydrase "ca," sugar-binding proteins "sug", endothiopepsin "end", purine nucleoside phosphorylase "pnp" and other proteins "oth"). The insert contains the following numbers for each subset: the number of complexes, the correlation coefficient, and the standard deviation from the linear fit.

of similar proteins with the accuracy of roughly 1.5 orders of magnitude. We note, however, that the slopes and the intercepts of the linear fits vary between the subsets of proteins, especially if the ligand-protein interactions

involve quantum-mechanical effects (sulfonamide-based ligands of HCA family) or if the ligands are large and have many rotatable bonds (peptidic ligands of endothiopepsin, END). When the HCA and END subsets are included, the overall correlation coefficient is rather low ($r = 0.44$ and $\sigma = 2.1$) but increases to 0.77 (and σ decreases to 1.5) when they are omitted. We also note that for the set of 77 complexes used for validation of another widely used knowledge-based scoring function, PMF[35] (almost all of which are included in our test set), our function gives a slightly better correlation coefficient[41] (0.68 by SMoG2001 versus 0.61 by PMF). From these observations, we conclude that when designing medium-sized ligands whose binding to the target does not involve quantum mechanical interactions, CombiSMoG should predict binding affinities with the accuracy of 1.5−2 orders of magnitude.

Conclusions

CombiSMoG is a unique computational tool for designing protein ligands that combines an accurate SMoG2001 knowledge-based potential and a flexibile, "combinatorial" MC growth algorithm. In its first experimental test, CombiSMoG generated a very potent inhibitor of HCA and correctly predicted its binding orientation; to our knowledge, it is the first time that a computational method created a lead that had an affinity higher than any known ligand for a given protein target. Despite its initial success, CombiSMoG needs further improvement. The knowledge-based potential can be made more accurate by including more complexes in the training database. Paradoxically, we are most interested in complexes of *poor* binders, since they provide better statistics of unfavorable contacts (unfortunately, structures of low-affinity ligands are rarely published, and the database buildup is rather slow). We are also working on developing algorithms that would account at the stage of growth for conformational changes in the protein upon binding; for many targets, rational lead design is impossible without taking these changes into account. Finally, we are working on making the growth algorithm more "chemical" by adding to it simple synthetic rules that preclude formation of synthetically impossible bonds during ligand growth.

References

(1) Rami Reddy, M.; Parrill, A. L. *Rational Drug Design: Novel Methodology and Practical Applications*; Rami Reddy, M., Parrill, A. L., Ed.; American Chemical Society: Washington, DC, 1999; Vol. 719, pp 1−11.

(2) Hansch, C. A quantitative approach to biochemical structure−activity. *Acc. Chem. Res.* **1969**, *2*, 232−239.

(3) Topliss, J. G. Utilization of operational schemes for analogue synthesis in drug design. *J. Med. Chem.* **1972**, *15*, 1006−1011.

(4) Redl, G.; Cramer, R. D.; Berkoff, C. E. Quantitative drug design. *Chem. Soc. Rev.* **1974**, *3*, 273−292.

(5) Ewing, T. J. A.; Kuntz, I. D. Critical evaluation of search algorithms for automated molecular docking and database screening. *J. Comput. Chem.* **1997**, *18*, 1175−1189.

(6) Fontain, E. Application of genetic algorithms in the field of constitutional similarity. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 748−752

(7) Moon, J. B.; Howe, W. J. Computer design of bioactive molecules−a method for receptor-based *de novo* ligand design. *Proteins*, **1991**, *11*, 314−328.

(8) Böhm, H.-J. Site-directed structure generation by fragment-joining. *Perspect. Drug Discuss. Des.* **1995**, *3*, 21−33.

(9) Bohacek, R. S.; McMartin C. Modern computational chemistry and drug discovery: structure generating programs. *Curr. Opin. Chem. Biol.* **1997**, *1*, 157−161.

(10) Bohacek, R. S.; McMartin C. *De novo* design of highly diverse structures complementary to enzyme binding sites−application to thermolysin. *Comput.-Aid. Mol. Design;* American Chemical Society: Washington, DC, 1995; Vol. 589, pp 82−97.

(11) Ajay; Murcko M. A. Computational methods to predict binding free energy in ligand−receptor complexes. *J. Med. Chem.* **1995**, *38*, 4953−4967.

(12) Free, S. M.; Wilson, J. W. Mathematical contribution to structure−activity studies. *J. Med. Chem.* **1964**, *7*, 395−399.

(13) Hansch, C.; Fujita, T. Rho-sigma-pi analysis. Method for correlation of biological activity + chemical structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616−1626.

(14) Kubinyi, H. QSAR and 3D QSAR in drug design. 1. methodology. *Drug Des. Today* **1997**, *2*, 457−467.

(15) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular-field analysis (ComFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(16) Bystroff, C.; Oatley, S. J.; Kraut, J. Crystal structures of *Escherichia coli* dihydrofolate reductase − the NADP+ holoenzyme and the folate-NADP+ ternary complex − substrate binding and a model for the transition state. *Biochemistry* **1990**, *29*, 3263−3277.

(17) Böhm, H.-J.; Klebe, G. What can we learn from molecular recognition in protein−ligand complexes for the design of new drugs. *Angew. Chem., Int. Ed. Engl.* **1996**, *35*, 2588−2614.

(18) Babine, R. E.; Bender, S. L. Molecular recognition of protein−ligand complexes: applications to drug design. *Chem. Rev.* **1997**, *97*, 1359−1472.

(19) Klebe, G. Recent developments in structure-based drug design. *J. Mol. Med.* **2000**, *78*, 269−281.

(20) Gohlke, H.; Klebe, G. Statistical potentials and scoring functions applied to protein−ligand binding. *Curr. Opin. Struct. Biol.* **2001**, *11*, 231−235.

(21) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.

(22) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A second generation force-field for the simulation of proteins, nucleic acids and organic molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.

(23) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, J.; Karplus, M. CHARMM − a program for macromolecular energy, minimization and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187−217.

(24) Kollman, P. A. Free energy calculations − applications to chemical and biological phenomena. *Chem. Rev.* **1993**, *7*, 2395−2417.

(25) Kollman, P. A. Advances and continuing challenges in achieving realistic and predictive simulations of the properties of organic and biological molecules. *Acc. Chem. Res.* **1996**, *29*, 461−469.

(26) Aqvist, J.; Medina, C.; Samuelsson, J.-E. New method for predicting binding affinity in computer-aided drug design. *Prot. Eng.* **1994**, *7*, 385−391.

(27) Viswanadhan, V. N.; Reddy, M. R.; Wlodaver, A.; Varney, M. D.; Weinstein, J. N. An approach to rapid estimation of relative binding affinities of enzyme inhibitors: Application to peptidomimetic inhibitors of the human immunodeficiency virus type 1 protease. *J. Med. Chem.* **1996**, *39*, 705−712.

(28) Böhm, H.-J. The development of a simple empirical scoring function to estimate the binding constant for a protein−ligand complex of known three-dimensional structure. *J. Comput. Aid. Mol. Des.* **1994**, *8*, 243−256.

(29) Head, R. D.; Smythe, M. L.; Oprea, T. I.; Waller, C. L.; Green, S. M.; Marshall, G. R. VALIDATE: A new method for the receptor-based prediction of binding affinities of novel ligands. *J. Am. Chem. Soc.* **1996**, *118,* 3959−3969.

(30) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput-Aided Mol. Des.* **1997**, *11,* 425−445.

(31) Jernigan, R. L.; Bahar, I. Structure-derived potentials and protein simulations. *Curr. Opin. Struct. Biol.* **1996**, *6*, 195−209.

(32) Tanaka, S.; Scheraga, H. A. Statistical mechanical treatment of protein conformation. 1. Conformational properties of amino acids in proteins. *Macromol.* **1976**, *9*, 945−950.

(33) Finkelstein, A. V.; Badretdinov, A. Y.; Gutin, A. M. Why do protein architecture have Boltzmann-like statistics? *Proteins* **1995***, 23,* 142−150 and references therein.
(34) DeWitte, R. S.; Shakhnovich, E. I. SMoG: de novo design method based on simple, fast and accurate free energy estimates. 1. Methodology and supporting evidence. *J. Am. Chem. Soc.* **1996**, *118*, 11733−11744.
(35) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein−ligand interactions: a simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791−804.
(36) Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Thornton, J. M. BLEEP − potential of mean force describing protein−ligand interactions: I. Generating potential. *J. Comput. Chem.* **1999***, 20*, 1165−1176.
(37) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein−ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.
(38) Burgi, H. B. Can statistical analysis of structural parameters from different crystal environments lead to quantitative energy relationships. *Acta Crystallogr.* **1988**, *B44*, 445−448.
(39) Ben-Naim, A. Statistical potentials extracted from protein structures: Are these meaningful potentials? *J. Chem. Phys.* **1997**, 107, 3698−3706.
(40) Grzybowski, B. A.; Ishchenko, A. V.; DeWitte, R. S.; Whitesides, G. M.; Shakhnovich, E. I. Development of a knowledge-based potential for crystals of small organic molecules: calculation of energy surfaces for C═O...H−N hydrogen bonds. *J. Phys. Chem. B* **2000**, *104*, 7293−7298.
(41) Ishchenko, A. V.; Shakhnovich, E. I. SMoG2001—an improved knowledge-based scoring function for protein−ligand interactions. *J. Med. Chem.*, in press.
(42) DeWitte, R. S.; Ishchenko, A. V.; Shakhnovich, E. I. SMoG: de novo design method based on simple, fast and accurate free energy estimates. 2. Case studies in molecular design. *J. Am. Chem. Soc.* **1997**, *119*, 4608−4617.
(43) Shimada, J.; Ishchenko, A. V.; Shakhnovich, E. I. Analysis of knowledge-based protein−ligand potentials using a self-consistent method. *Protein Sci.* **2000**, *9*, 765−775.
(44) Grzybowski, B. A.; Ishchenko, A. V.; Kim, C.-Y.; Topalov, G.; Chapman, R.; Christianson, D. W.; Whitesides, G. M.; Shakhnovich, E. I. Combinatorial computational method gives new picomolar ligands for a known enzyme. *Proc. Nacl. Acad. Sci.* **2002**, *99*, 1270−1273 and references therein.
(45) Joseph-McCarthy, D. Computational approaches to structure-based ligand design. *Pharmacol. Ther.* **1999**, *84*, 179−191.
(46) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953***, 21*, 1087−1092.
(47) (a) Gao, J.; Cheng, X.; Chen, R.; Sigal, G. B.; Bruce, J. E.; Schwartz, B. L.; Hofstadler, S. A.; Anderson, G. A.; Smith, R. D.; Whitesides, G. M. Screening Derivatized Peptide Libraries for Tight Binding Inhibitors to Carbonic Anhydrase II by Electrospray Ionization-Mass Spectrometry**.** *J. Med. Chem.* **1996**, *39*, 1949−1955. (b) Burbaum, J. J.; Ohlmeyer M. H. J.; Reader, J. C.; Henderson, I.; Dillard, L. W.; Li G.; Randle, T. L.; Sigal, N. H.; Chelsky, D.; Baldwin, J. J. A paradigm for drug discovery employing encoded combinatorial libraries. *Proc. Natl. Acad. Sci.* **1995**, *92*, 6027−6031. (c) Supuran C. T.; Briganti, F.; Tilli, S.; Chegwidden, W. R.; Scozzafava, A. Carbonic anhydrase inhibitors: Sulfonamides as antitumor agents? *Bioorg. Med. Chem.* **2001**, *9*, 703−714.

AR970146B