SMoG: de Novo Design Method Based on Simple, Fast, and Accurate Free Energy Estimates. 2. Case Studies in Molecular Design

Robert S. DeWitte, Alexey V. Ishchenko, and Eugene I. Shakhnovich*

Contribution from the Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138

Received October 23, 1996[⊗]

Abstract: In this paper, we summarize three ligand design studies performed using the program SMoG, which was developed in our lab. The aim of this presentation is to communicate through examples the potential of this method: the richness of the molecules that can be developed and the ease with which they are found. In particular, we present suggestions for ligands to Src SH3 domain (specificity pocket and LP site) and CD4.

Introduction

SMoG[1] is an automatic computational tool that generates candidate leads to protein binding sites. This method is based on a coarse-grained knowledge-based potential extracted from crystal structures of small molecule−protein complexes. This approach derives its success from the way in which appropriate levels of coarse graining serve to incorporate various solvation phenomena into what is effectively an estimation of the binding free energy. SMoG is an acronym for *small molecule growth*, so-named for the procedure with which ligand candidates are generated. In summary, ligands are formed by joining functional groups together with standard bond lengths and angles (and optimally selected torsions). The algorithm randomly accepts functional group additions with a thermal (Boltzmann) probability according to the metropolis Monte Carlo method.

Although the details of the methodology have been presented in the first paper in this series, some of the above description should be explained. In particular, coarse graining is a procedure commonly used in statistical mechanics to allow one to focus attention on events at an intermediate length scale so that one can deduce general trends without being overwhelmed by the variations in the most minute details. Formally, it amounts to an averaging of the interaction potential within cells of some carefully chosen size. The size of these cells should correspond to some physical distances in the system so that one can be assured of the essence of the details subsumed in the averaged potential. In our example of ligand−protein interactions, the length scale was chosen to correspond roughly to the distance over which a molecule can induce order within aqueous solvent (5 Å). In so doing, the course-graining approach integrates the solvent entropy terms of ligand binding into its potential surface. Practially speaking, the coarse graining defines a rather large radius of contact between atoms of a protein and a ligand. Thus, when examining a database of crystal structures of protein−ligand complexes or evaluating the binding interactions in the course of design, atoms within 5 Å of each other are considered to be in contact.

A knowledge-based potential is a set of interaction parameters that measure the contribution of various types of contacts to the free energy estimate. These parameters are derived from a database of structures by collecting statistics on the frequencies with which contacts are formed between all the various atom types. In combination with coarse graining, then, the knowledge-based potential provides a system for estimating binding free energies based on physical statistical inference.

Another benefit of coarse graining is that the potential surface is smoothed by the local averaging. This allows the space of possible molecules to be searched very efficiently by a Monte Carlo growth algorithm. In this method, as a functional group is being considered for addition to the growing molecule, its affect on the binding free energy estimate is determined. If the free energy decreases, the functional group is accepted as part of the molecule. If the free energy increases, the probability that the functional group is accepted is given by $\exp(-\delta G/T)$ where T is an algorithmic temperature. This metropolis criterion ensures that small increases in the binding free energy are far more likely to be accepted than large increases. Functional group additions that increase the binding free energy estimate are not strictly forbidden since their inclusion may be necessary to allow a different, and very beneficial, functional group to be added at a later stage in molecular growth.

The upshot of this integrated approach is a very efficient biasing routine, such that individual optimized ligands are generated in a few seconds on a modest workstation (100 MHz Pentium PC running Linux, for instance). With this efficiency, it is possible to generate thousands of candidate leads, recording only those with the lowest (most negative) free energy estimates. These molecules are then subjected to screening according to criteria such as ease of synthesis, solubility, chemical intuition, hydrogen bond formation, and empirical measures of interaction enthalpy. At present we use Quanta/CHARMM[2,3] (PARAM19) to obtain some of this information, although a long term goal is to develop a fully integrated ligand design suite that affords seamless analysis.

In this paper we present, case by case, a series of three design efforts, each of which illustrate the design process. From our experience with this approach, a general methodology has

* Address correspondence to the following authore: E. I. Shakhnovich, Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street Cambridge, MA 02138; phone (617) 495-4130; fax (617) 496-5948; email eugene@diamond.harvard.edu.

[⊗] Abstract published in *Advance ACS Abstracts,* May 1, 1997.

(1) DeWitte, R. S.; Shakhnovich, E. I. *J. Am. Chem. Soc.* **1996**, *118*, 11733−11744.

(2) Molecular Simulations Incorporated: 16 New England Executive Park, Burlington, MA 01803-5297, 1986.

(3) Brooks, B. R.; Bruccoleri, R.E.; Olafson, B.D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187−217.

S0002-7863(96)03689-X CCC: $14.00

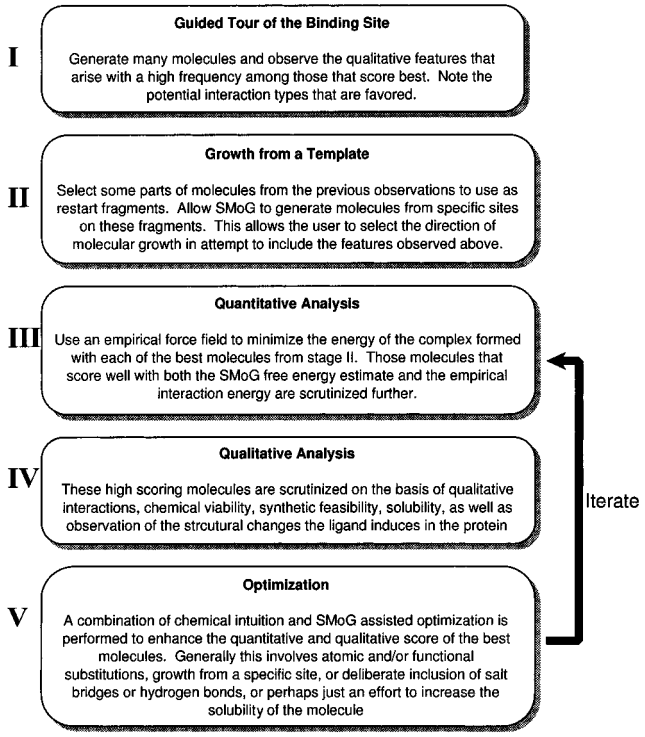

Figure 1. The stages of ligand design.

emerged, which yields consistent progress throughout the design process. However, every design effort has its own unique challenges, and those that follow are no exception. Here, the flexibility of the SMoG ligand design process is a definite advantage.

Methods

Figure 1 provides a general outline for ligand design using the SMoG approach.

At stage I, it is helpful to get an appreciation for the binding site (for instance, its shape and types of intermolecular interactions it may support) by allowing SMoG to generate a large number of molecules there. Our approach has been to generate 1000 molecules and record the structures of the best 50. This computation is generally completed in a few hours. By viewing the molecules that fit into the space of the binding site, and form complementary chemical interactions, particularly paying attention to those molecular details that arise frequently, the chemist gains an immediate understanding of what molecular scaffolds are likely to be fruitful leads to follow. For example, one may observe that several of the high-scoring molecules involve specific hydrogen bonds, attained through a specific functional group in a specific orientation. Alternatively, one may learn that the presence of lipophilic groups in a certain region are responsible for the high score of several other ligands. This "consensus based" qualitative understanding forms the basis for further work with SMoG.

Whereas the high score of the molecules in stage I is usually due to the presence of one well-placed molecular fragment, the desire at stage II is to build molecules that combine several of the positive features observed in stage I. This is done by selecting a few representative molecules from stage I, removing the parts of the molecule that are not important, and using the remaining strongly interacting molecular fragment as a *restart fragment*. SMoG has the ability to continue growth from any molecule provided as input, by selecting hydrogen atoms on it as points of further growth. Furthermore, the user may determine at which of these hydrogen atoms growth is allowed. For each moiety from stage I that is used as a restart fragment, a new line of molecules can be generated (generally another 1000 of which the best 50 are recorded). Each of these molecules will contain the tailored interactions as well as a variety of other positive features. Thus, at the end of stage II, there will arise from each line of molecules a small number of candidates that incorporate several qualitative as well as quantitative interactions. At this point, a large part of the combinatorial problem of ligand generation has been overcome (*i.e.*, the architecture of the molecule has been decided). Depending on the situation at hand, it may be desirable to repeat stage II once more with some lines of molecules to optimize the collection of interacting functional groups even further. We have not found this to be necessary in the studies reported in this paper.

Table 1. Quantitative Analysis of the First-Generation CD4 Candidates Shown in Figure 2

molecule	SMoG score per heavy atom	CHARMM interaction energy (kcal)
8	−26.2	−82.3
17	−30.0	−80.9
32	−28.5	−53.6
33	−36.3	−70.6
35	−26.8	−80.8
41	−45.7	−99.0
45	−26.9	−59.8

Table 2. Quantitative Analysis of the Second-Generation CD4 Candidates Shown in Figure 2[a]

molecule	SMoG score per heavy atom	CHARMM energies (kcal)		
		interaction	strain	net
41	−45.7	−99	14.4	−84.3
41b	−47.9	−139	20.0	−119.0
41c	−45.1	−128	20.7	−107.3
41d	−46.1	−120	33.2	−86.8
41e	−50.6	−112	36.6	−75.4
41f	−50.5	−116	20.1	−95.9
41g	−49.4	−82	15.0	−67.0
41h	−49.9	−119	22.5	−96.5
41i	−48.0	−86	20.5	−65.6

[a] The strain energy is calculated as the difference in internal energy between the bound conformation and the conformation resulting from gas phase minimization to convergence using the adapted-basis Newton−Raphson method. The net CHARMM energy is the interaction energy plus the strain energy.

Generally, after stage II, one might have a few dozen structures to consider in the subsequent stages. At this point, it is important to determine which of these molecules to focus on in more detail during the subsequent design stages. It is important to realize that SMoG does not include an intramolecular interaction potential in its growth. Therefore, one should relax the slight strains that the molecules are carrying through minimization of the protein−ligand complex with an empirical potential (for example, CHARMM). The empirical interaction energy of these relaxed complexes (especially the electrostatic component) is another useful measure of the quality of a molecule because the SMoG design potential does not explicitly account for electrostatic interactions between the molecule and the protein and, thus, slightly undervalues hydrogen bond and salt bridge formation. Conversely, since hydrophobic interactions are largely solvent entropy effects, empirical calculations of interaction enthalpies undervalue the contribution of nonpolar interactions to binding free energy. Thus, the two measures of interaction strength are somewhat complementary. Hence, the molecules that one should continue to focus on for the remainder of the design stages are those which have low CHARMM and SMoG energies. At present, we are moving to include explicit terms to handle electrostatic interaction events.

At stage IV, the remaining molecules (perhaps a dozen) need to be scrutinized qualitatively with the goal of optimization in mind, rather than exclusion. The criteria with which to judge the molecules include chemical stability, ease of synthesis, internal strain energy, strain induced in the protein, and solubility. One should also determine if subsequent growth or manual optimization can introduce more hydrogen bonds, or capitalize on other features of the binding pocket, such as stacking with delocalized π-bonding systems. It is clear from our experience that a few molecules will emerge as having greater potential than the others because of the nature of the interactions they incorporate presently as well as features that suggest either simple manual changes leading to improvement or directions in which automatic growth may

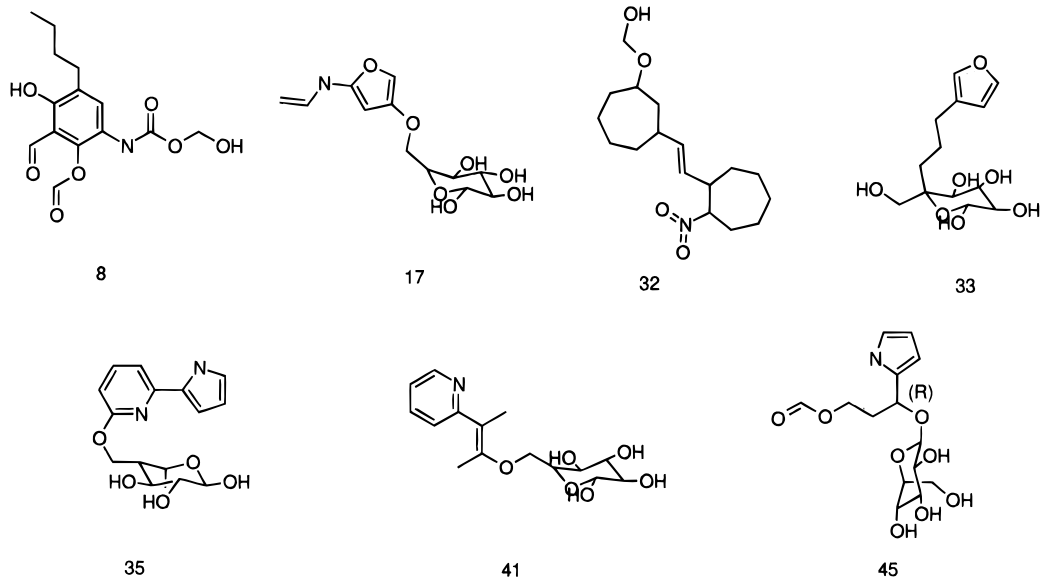

Figure 2. First-generation molecules as ligand candidates for CD4. On the basis of their SMoG and CHARMM interaction energies, these are the best 7 of 1000 molecules generated in the binding site. Notice the similarity in the structures: most of them contain a hydrogen-bonding fragment and a hydrophobic fragment in the same relative orientation. This consensus points to the fact that optimal ligands should incorporate these features. Table 1 describes the quantitative analysis of these molecules. On the basis of these data, molecule **41** was chosen as the parent of the second generation.

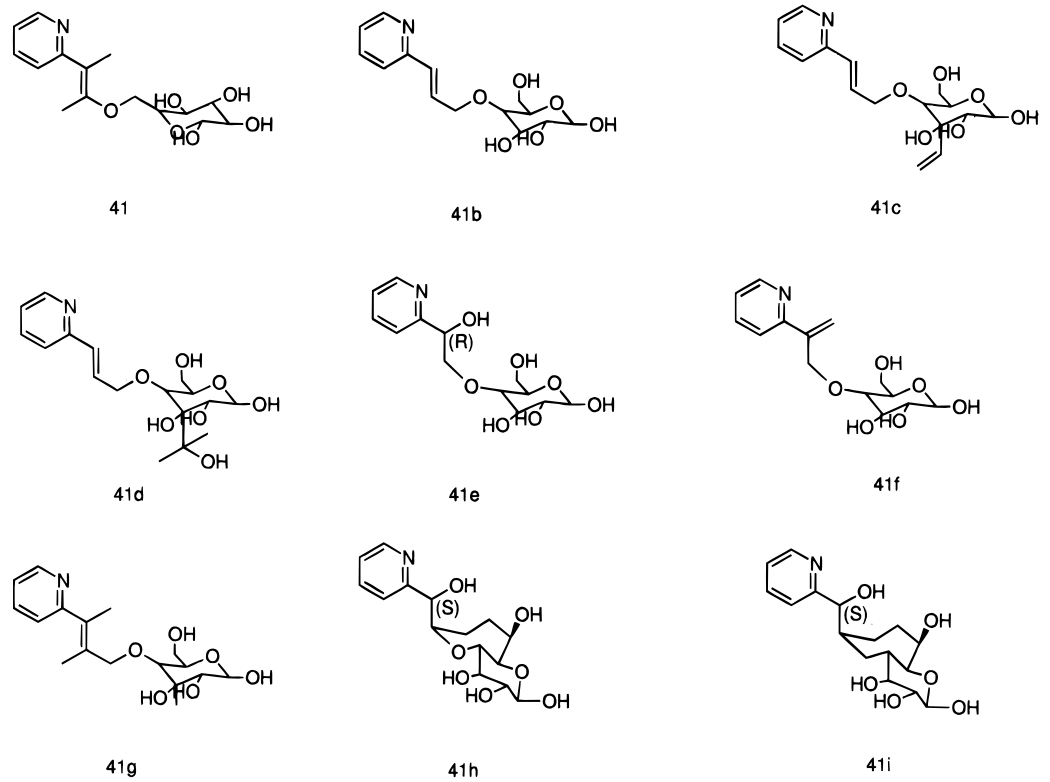

Figure 3. Second-generation molecules as ligand candidates for CD4. Molecule **41** is shown in the upper left. By manually altering the point of attachment of the sugar-like ring, molecule **41b** was formed, which had an improved π-stacking interaction with Phe 43. Molecules **41c**, **41d**, and **41g** derive from **41b** through SMoG-generated ring substituents. Molecule **41e** results from **41b** through shortening and saturating the flexible chain connecting to the pyridine group, which also improved the π-stacking. Molecule **41f** follows from **41e** via manual alteration suggested by the geometry of the binding site. Molecule **41h** derives from **41e** through the addition of a bridge from the flexible chain to the sugar-like ring which preserved the binding conformation of the molecule, thereby enhancing its rigidity. Molecule **41i** differs from **41h** through manual substitution of carbon for the oxygen atom on the seven-membered ring; this substitution weakens the π-stacking due to its affects on various angles in molecule **41i**. Table 2 describes the quantitative and qualitative analysis of these molecules.

enhance the binding interactions (using the whole current molecule as a restart fragment and allowing growth at only one or perhaps a few select hydrogen atoms).

In stage V, the modifications suggested in stage IV are introduced to the few select molecules that have the most potential, yielding yet another generation of structures which should be scrutinized quantitatively and qualitatively in stages III and IV.

In the process of designing a molecule that is likely to be a strong binding ligand, stages III−V may need to be iterated several times until a candidate is found which is qualitatively sound and scores among

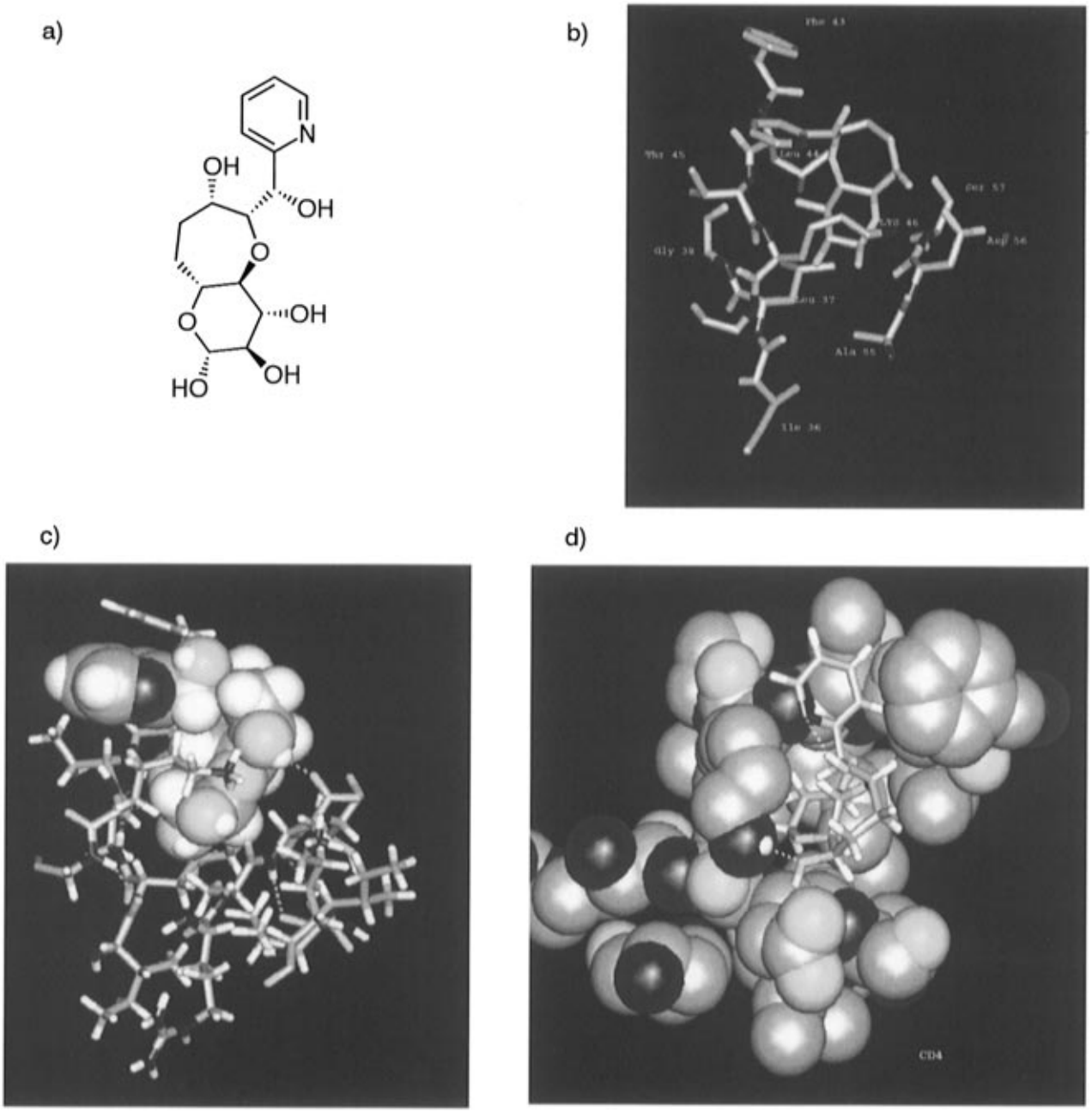

Figure 4. A candidate ligand for the Phe 43 binding pocket of CD4. This molecule is able to form five hydrogen bonds (four intermolecular and one intramolecular) as well as a significant π-stack with the benzene ring of Phe 43. (a) Molecular structure of the candidate: note the rigid structure. (b) Licorice diagram of the ligand in the binding site showing the residues with which a strong ligand should make interactions. (c) The ligand shown as a space-filling model. Notice the π-stacking with Phe 43. (d) Another view, this time with the protein as a space-filling model.

the best molecules according to SMoG and CHARMM. As the process converges to a ligand, one may wish to use other modeling tools to analyze the molecules and enhance decision making. These may include conformational analysis to ensure that the binding mode of the molecule is not a highly strained conformer and molecular dynamics simulation in solvent to observe the stability of the predicted complex.

Results

Below, we shall describe three case studies in design. The examples shown were chosen to be representative of the design process and the resulting lead candidates, rather than suggestions of the ultimate lead candidate. As such, several molecules, particularly those in the first generation, may contain undesirable features. Such candidates may either be eliminated from consideration or optimized through further iteration in design. Although each of these cases involved several distinct family lines of molecules, each leading to different lead candidates, we focus on one line in each case for the sake of clarity and brevity.

CD4. The CD4 protein is an immunoglobin-family transmembrane receptor expressed in helper T-cells.[4] It participates in contact between the T-cells and antigen-presenting cells by binding to the nonpolymorphic part of the class II major histocompatibility complex (MHC-II) protein, which is followed by the activation of the bound Lck kinase which leads to downstream activation events in T-cells. The human immunodeficiency virus (HIV) gains entry into a T-cell binding protein gp120 to the CD4 receptor. This gp120 binding site in the vicinity of Phe 43 of CD4 was the target for ligand design in this project (see Figure 4b).

(4) Bour, S.; Gelezunias, R.; Wainberg, M. A. *Microbiol. Rev.* **1995**, *59*, 63−93.

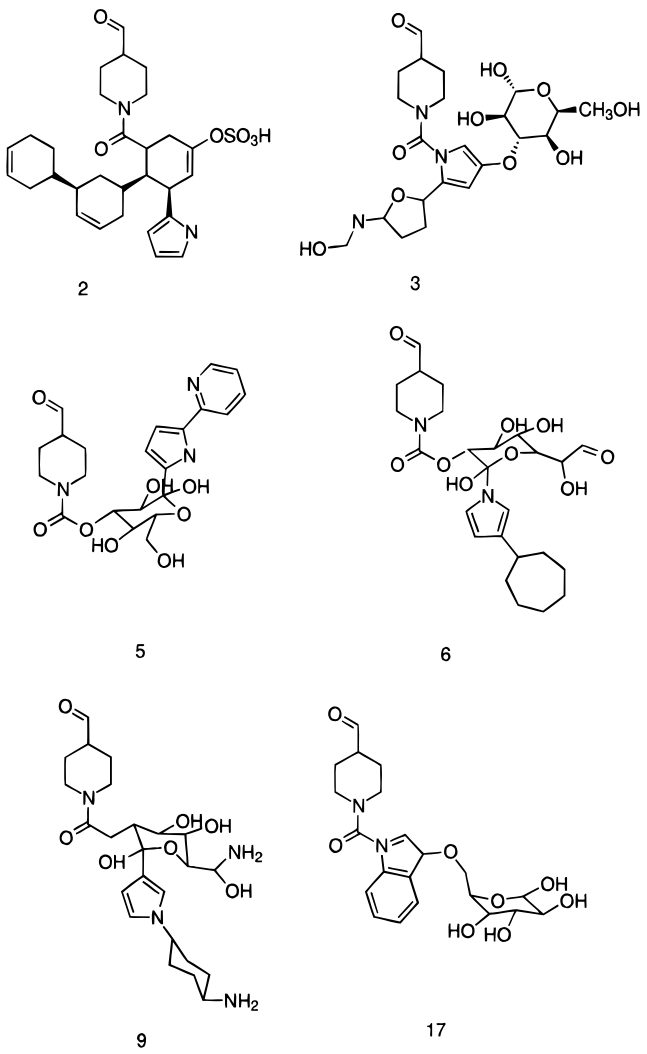

Figure 5. First-generation molecules as ligand candidates for the specificity pocket of Src SH3 domain. These are the best 6 of 1000 molecules generated in the binding site, using the experimental monomer as the restart fragment. These molecules provide a guided tour of the binding site, suggesting that a glucose-like ring be used to form hydrogen bonds with residues in the RT loop of the pocket and that unsaturated ring systems be used to make hydrophobic contact with the tryptophan and tyrosine residues in the binding pocket. Molecule **3** scores well quantitatively (see Table 3) and also provides suggestions for improved hydrophobic interactions with Tyr 55 and Trp 42. For these reasons, it became the focus of further attention in the design process.

Among the possible interactions that arose in stage I of the design process, it was apparent that $\pi-\pi$ interaction with the phenyl ring of Phe 43 was important, as well as the formation of hydrogen bonds in the narrow pocket bounded by Lys 46 and Asp 56. After one pass through the five stages, the first generation of molecules was evident. These are shown in Figure 2, where one can see the common elements of a hydrogen-bonding core and a hydrophobic moiety in the same relative orientation in most molecules. Qualitative features, as well as the data in Table 1, led to the selection of molecule **41** for further attention. Figure 3 and Table 2 describe the evolution, through SMoG-assisted optimization and manual editing suggested by chemical intuition, from molecule **41** to the best candidate, molecule **41h**. In this and all other cases, manual editing was performed using the 3D Molecular Editor facility of Quanta and subsequent minimization of energy was done with CHARMM.

Table 3. Quantitative Analysis of the First-Generation SH3 Specificity Pocket Candidates Shown in Figure 5

molecule	SMoG Score per heavy atom	CHARMM interaction energy (kcal)
2	−32.7	−117.9
3	−32.9	−59.4
5	−23.9	−107.2
6	−29.4	−62.9
9	−30.0	−68.4
17	−33.2	−46.3

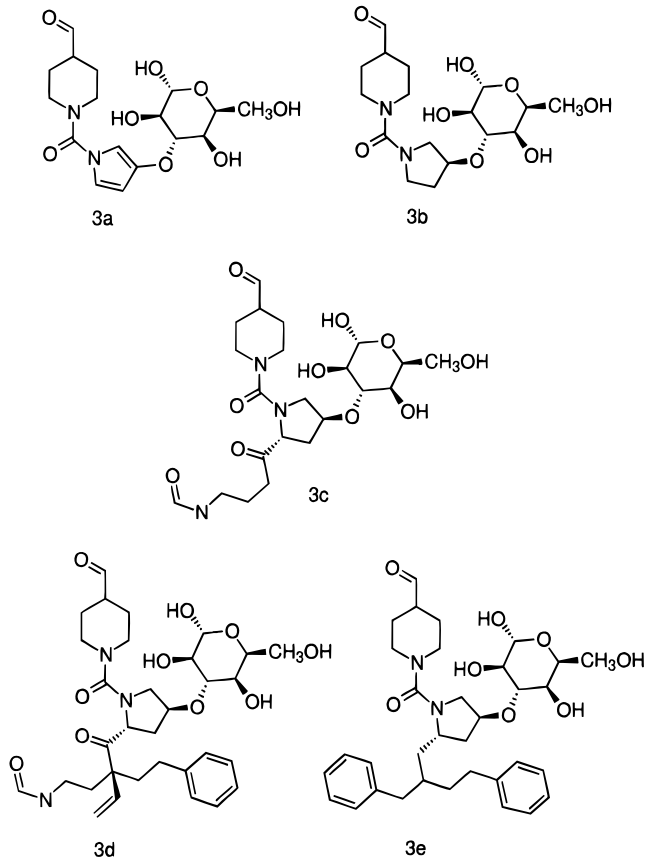

Figure 6. Second- and third-generation molecules as ligand candidates for the specificity pocket of Src SH3 domain. Molecule **3a** is derived from molecule **3** of Figure 6 by removing one substituent from the pyrrole ring. This molecule has considerable strain energy (see Table 4) which was relieved by saturating the five-membered ring in such a way that the conformation of the glucose was altered as little as possible, forming molecule **3b**. After minimization of the complex structure with CHARMM, **3b** was used as the restart fragment, with only the H atoms on the central five-membered ring being eligible attachment points. The best-scoring candidate, **3c**, was used as a restart fragment, leading to the SMoG best third-generation lead candidate, molecule **3d**, whose phenyl ring forms a π-stacking configuration with Tyr 55. Molecule **3e** was derived from **3d** by manual alteration after noting that the arrangement of the terminal amide group could form part of a phenyl group that made a partial π-stack with Trp 42. Also, the joining chain was made more flexible by the elimination of one carbonyl group, converting the carbon from sp^2 to sp^3, thus reducing internal strain energy.

Figure 4 shows the three-dimensional structure of molecule **41h** in the gp120 binding site of CD4. The interactions present include partial π-stacking with Phe 43, as well as four intermolecular hydrogen bonds with Lys 46 and Asp 56 and one intramolecular hydrogen bond which stabilizes the orientation of the pyridine group. The seven-membered fused-ring bridge gives this molecule a great deal of rigidity in its bound conformation.

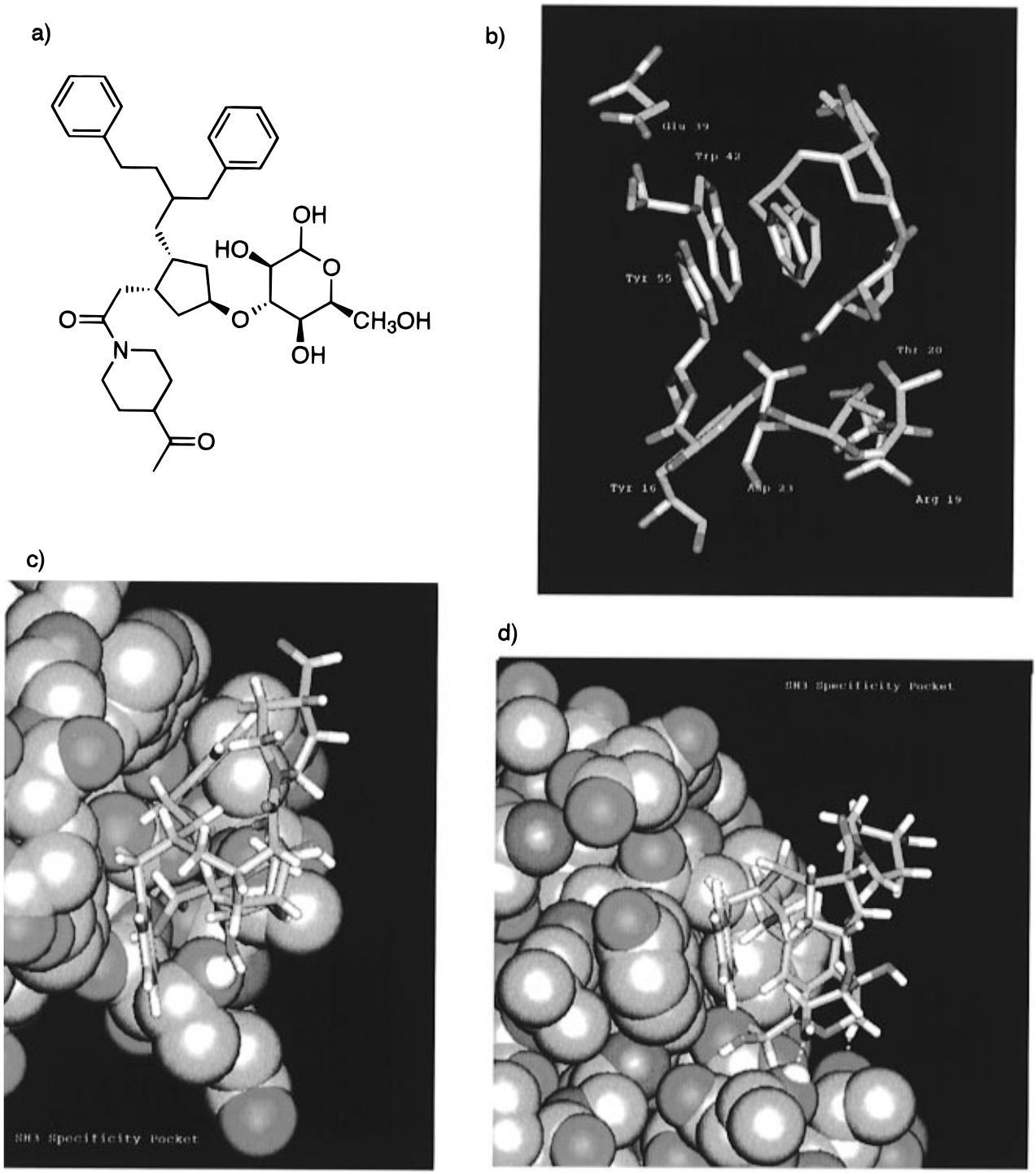

Figure 7. A candidate ligand for the specificity pocket of Src SH3 domain (Molecule **3e** from Figure 5). This molecule is able to form three hydrogen bonds as well as a significant π-stack with both Tyr 55 and Trp 42. (a) Molecular structure of the candidate. (b) Licorice diagram of the ligand in the binding site showing the residues with which a strong ligand should make interactions. (c) A space-filling model. Notice the π-stacking with Tyr 55 and Trp 42. (d) Another view.

SH3. The Src-homology-3 (SH3) domain is a conserved domain found in the variety of intracellular signal transduction mediators (PI3K, Grb2, Crk, etc.) and participates in the diversity of protein−protein interactions mediating the signal pathway eventually leading to the cell responses such as cell growth, differentiation, and migration.[5] The irregularities in these processes may contibute to the cause of several common diseases, thus making it important to consider the SH3 domain as candidate for therapeutic intervention.

Two classes of polyproline helix peptide were recently found to bind the Src SH3 domain[6,7](class I RXLPPLP and class II LPPLPXR). They are accommodated in three pockets formed by conserved residues: the specificity pocket occupied by arginine, directing the peptide orientation, and two LP pockets each occupied by an LP pair. The following two design efforts focus on one LP pocket and the specificity pocket.

Specificity Pocket. Combinatorially synthesized small molecule ligands attached to the N-terminus of "biasing element" PLPPLP (occupying two LP pockets, part of class I peptide

(5) Pawson, T. *Nature* **1995**, *373*, 573−580.
(6) Chen, J. K.; Lane, W. S.; Brauer, A. W.; Tanaka, A.; Schreiber, S. *J. Am. Chem. Soc.* **1993**, *115*, 12591−12592.
(7) Feng, S.; Chen, J. K.; Yu, H.; Simon, J. A.; Schreiber, S. L. *Science* **1994**, *266*, 1241−1247.

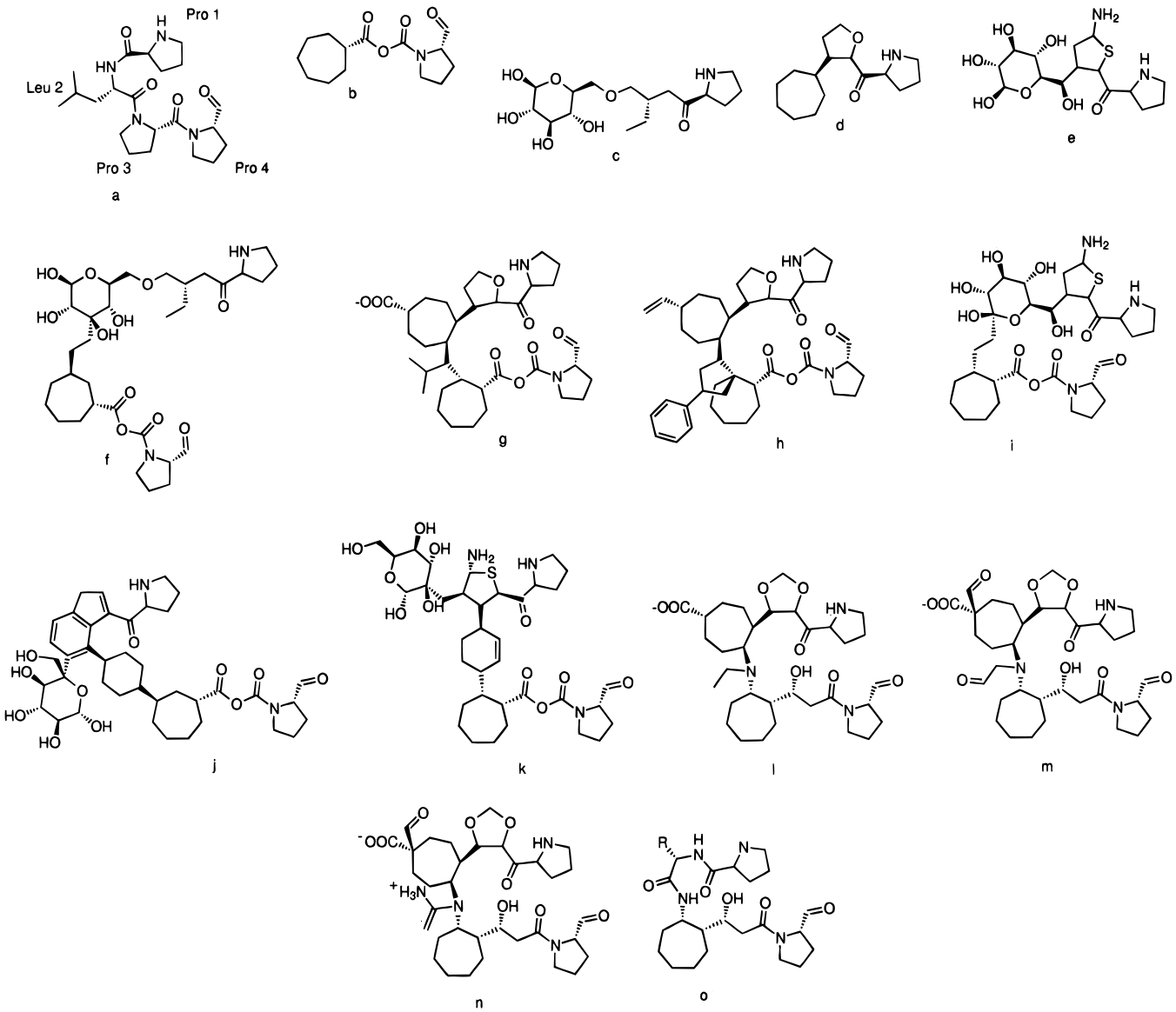

Figure 8. First- (b−e), second- (f−k), and third- (l−n) generation molecules as ligand candidates for the LP pocket of Src SH3 domain. The peptide PLPP that occupies the LP pocket is represented by a. The novel peptoid molecule is represented by o; various side chains, R, are shown in Figure 9.

Table 4. Quantitative Analysis of the Second-Generation SH3 Specificity Pocket Candidates Shown in Figure 6[a]

molecule	SMoG score per heavy atom	CHARMM energies (kcal)		
		interaction	strain	net
3a	−27.3	−54.6	32.2	−21.6
3b	−27.1	−51.4	18.5	−32.9
3c	−33.4	−77.2	27.4	−49.8
3d	−34.7	−59.7	24.4	−35.3
3e	−37.8	−57.9	19.0	−38.0

[a] The strain energy is calculated as the difference in internal energy between the bound conformation and the conformation resulting from gas phase minimization to convergence using the adapted-basis Newton−Raphson method. The net CHARMM energy is the interaction energy plus the strain energy.

Table 5. Quantitative Analysis of the First Three Generations of SH3 LP Pocket Candidates Shown in Figure 8

molecule	SMoG score per heavy atom	CHARMM interaction energy (kcal)
8a	−19.6	−20.7
8b	−35.6	−8.9
8c	−39.4	−23.6
8d	−34.3	−8.2
8e	−33.8	−23.0
8f	−36.6	−28.6
8g	−38.2	−53.8
8h	−47.9	−45.8
8i	−31.6	−52.7
8j	−35.7	−46.5
8k	−34.1	−53.2
8l	−37.4	−42.9
8m	−39.5	−52.2
8n	−38.8	−66.2
8o (R = H)	−28.6	−28.3

without Arg and with X = P) were recently shown to bind to the specificity pocket.[8] The assay revealed an extremely strong selection for the first monomer attached to the N-terminal proline. Since the acylated monomer provided the opportunity for growth into the pocket, it was used as a restart fragment, and SMoG was used to grow ligands into the specificity pocket by insisting that the growth proceed only from the acyl H atom on this monomer, thus preserving the peptide-like nature of the molecule.

After stage I, it was apparent that two characteristics of high-

(8) Combs, A. P.; Kapoor, T. M.; Feng, S.; Chen, J. K.; Daude-Snow, L. F.; Schreiber, S. L. *J. Am. Chem. Soc.* **1996**, *118*, 287−288.

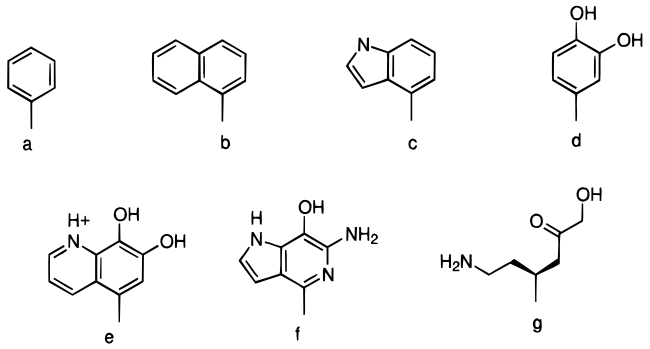

Figure 9. Various side chains for the molecule **8o**, as candidate ligands for the LP pocket.

Table 6. Quantitative Analysis of the Fourth-Generation SH3 LP Pocket Candidates Shown in Figure 9[a]

molecule	SMoG score per heavy atom	CHARMM energies (kcal)		
		interaction	strain	net
9a	−34.0	−25.4	30.8	+5.4
9b	−40.3	−27.6	4.9	−22.7
9c	−38.7	−25.4	34.9	+9.5
9d	−34.4	−41.1	12.3	−28.8
9e	−37.2	−46.8	12.1	−34.7
9f	−38.4	−52.3	6.3	−46.0
9g	−36.4	−52.9	31.2	−21.7

[a] The strain energy is calculated as the difference in internal energy between the bound conformation with Pro 1 and Pro 4 fixed, and the conformation resulting from gas phase minimization to convergence using the adapted-basis Newton−Raphson method, also holding Pro 1 and Pro 4 fixed. In this sense, the strain energy is the energy difference upon binding the portion of the helical-substituted biasing element in consideration to the protein. The net CHARMM energy is the interaction energy plus the strain energy.

scoring molecules were of special importance. First, the formation of a large amount of hydrophobic contacts with Tyr 55 and Trp 42. Second, the formation of hydrogen bonds with the donors and acceptors on Asp 23 and Thr 20 (see Figure 7b). The first-generation ligands are shown in Figure 5. After stage III, one basic template was selected for further optimization (molecule **3**), in which a sugar group made the hydrogen-bonding interactions and the remaining scaffold left a rich potential for enhancing the hydrophobic interactions. This selection was based predominantly on opportunities for enhancement and SMoG score, rather than the CHARMM score, which, though strong, was far weaker than other first-generation candidates (see Table 3). Some non-essential functionality was removed to prepare molecule **3a** (see Figure 6). Saturation of the pyrrole group led to molecule **3b**, the restart fragment for subsequent design, whose internal strain energy was greatly reduced (see Table 4). By using a few hydrogen atoms on this molecule as sites for potential growth, the SMoG algorithm then guided ligand design through two generations of optimization as described in the caption to Figure 6. The resulting molecule (**3e**), shown in Figure 7, is able to form two π-stacking interactions and three hydrogen bonds with the protein.

SH3 Domain LP Pocket. The design effort for the LP pocket faced additional challenges from the desire to replace L,P in position 2,3 of the biasing element with a mimetic. The new ligand should make amide bonds with the proline residues (1 and 4 of biasing element) at each boundary of the pocket, a goal which severely constrained the geometry of the molecules that would be reasonable structures.

This challenge was addressed in SMoG by using prolines 1,4 as restart fragments (see Figure 8a), such that molecular growth proceeded inward toward the pocket from each bounding proline.

The following are the results of stage I design: In place of Pro 3, SMoG demonstrated a strong preference for a seven-membered hydrophobic ring (Figure 8b) grown from Pro 4 which makes hydrophobic contacts with Tyr 52, Arg 11, Tyr 8, and Pro 19 side chains. In place of Leu 2, SMoG suggested several candidates grown from Pro 1, the best of which are shown in Figure 8c−e. These first generation molecules revealed that in the region where Pro 3 was bound, the preference is mainly for hydrophobic fragments whereas the Leu 2 site prefers fragments which make both hydrophobic contacts (with Trp 34) and hydrogen-bonding interactions (with residues Asn 51 and Ser 50). This last feature is absent in the purely hydrophobic leucine side chain.

In order to combine the above segments, several linkers were built both manually and by SMoG using the above segments as restart fragments. The most appropriate of these (*i.e.*, those that allowed covalent attachment without inducing much conformational strain energy) were used to join each pair of Leu site and Pro site fragments. The slight strains induced by linking were reduced with CHARMM minimization. The second generation molecules (*i.e.*, the best linked molecules) are shown in Figure 8f−k. Each of these show rather strong SMoG and CHARMM scores (see Table 5).

Qualitative analysis of these molecules indicated that the phenyl ring of molecule **8h** induces rather large deformation of the protein structure in the vicinity of the specificity pocket. Therefore, it was discarded. Those candidates containing glucose-like groups were also discarded in order to avoid molecules containing unnatural sugars.

This left only **8g** for subsequent study. Manual optimization with respect to synthetic feasibility led to the third-generation molecules **8l**−**n** which also contained some additional hydrogen-bonding groups, contributing to stronger binding energies (see CHARMM interaction energy in Table 5).

At this point, however, synthesis of the resulting Leu substitute would still be rather involved. Thus, in order to design the simplest molecule possible from the point of view of synthesis, while maintaining strong interactions, the seven-membered ring in place of Leu3 was removed. It became apparent that an amino acid linker would provide the ideal joint between Pro 1 and the Pro 3 substitute (ideal in the sense of synthetic ease and linkage without inducing conformational stress; moreover, the side chain direction points into the Leu 2 site). It should be noted that this observation was the direct result of examining candidate molecules which had been suggested by SMoG. The side chain of this amino acid was built by SMoG with growth restricted to position R on molecule **8o**. Several aromatic side chains were generated (Figure 9a−c) and further optimized manually by inserting various hydrogen-bonding fragments, yielding the fourth-generation molecules. Quantitative analysis of these final ligand candidates is shown in Table 6.

Molecule **9d** is shown in Figure 10 as an example of the best LP pocket ligand candidates designed using SMoG (molecules **9d**−**g**).

Discussion

In these few examples, we have shown that ligand design with SMoG as an automated de novo procedure has some distinct advantages. In particular, this automated approach allows the chemist to quickly assess the most critical interactions needed to build a strong binding ligand *and* simultaneously suggests chemical species that can meet these needs. Furthermore, the method provides a quantitative score which is related directly to the binding free energy,[1] so that any adjustments

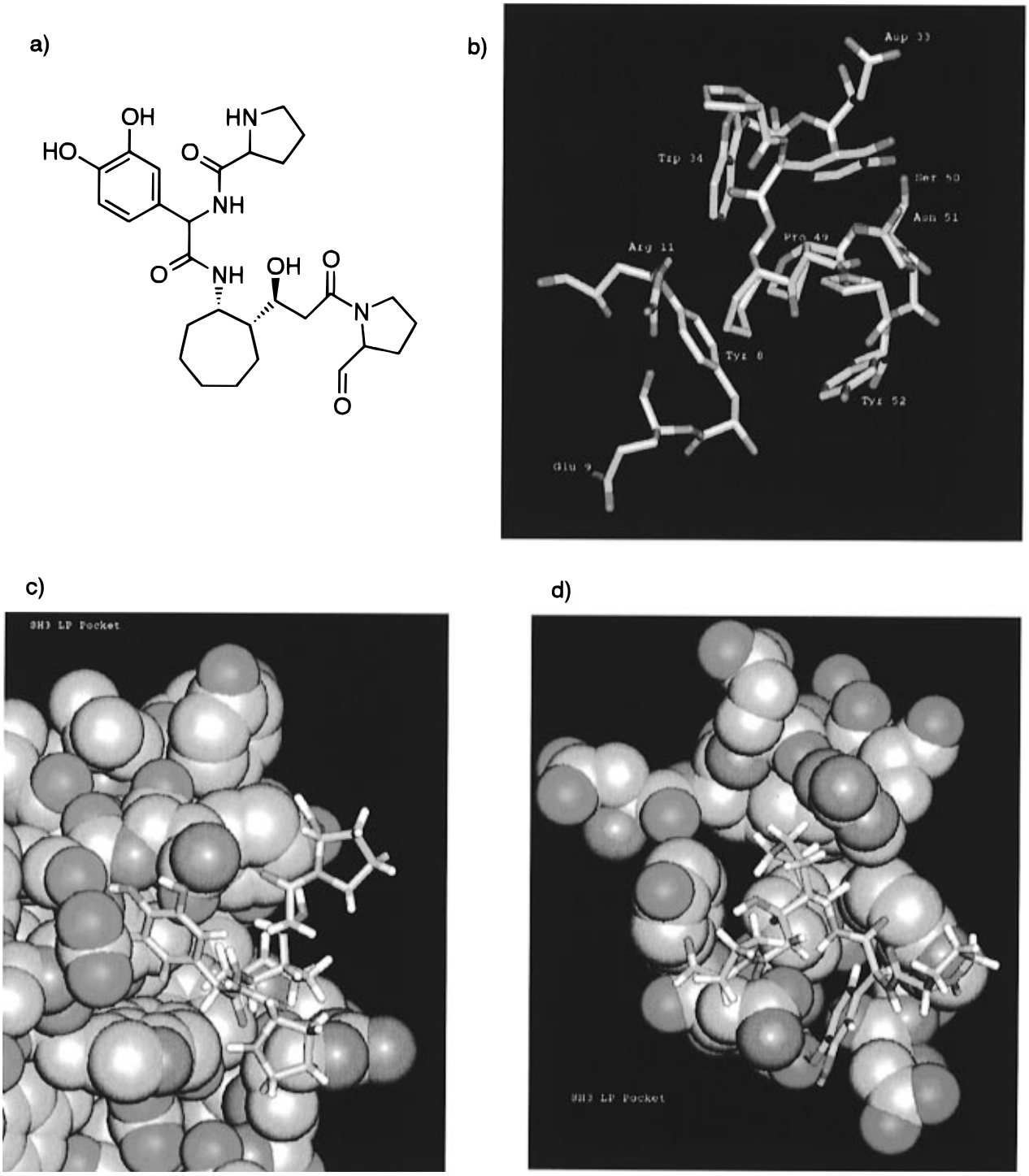

Figure 10. A candidate ligand for the LP pocket of Src SH3 domain. This molecule is able to form three hydrogen bonds and significant hydrophobic and electrostatic complementarity while bridging the bounding proline residues (also shown) of the biasing element. (a) Molecular structure of the candidate. (b) Licorice diagram of the ligand in the binding site showing the residues with which a strong ligand should make interactions. (c) A space-filling model. (d) Another view.

made to a molecule for the sake of qualitative improvement can be scrutinized quantitatively and *vice versa*. This protocol, when coupled with chemical intuition as regards chemical stability and synthetic ease, can lead to very many exciting novel ligand candidates with just a few weeks of effort.

Also, the diversity of the strategies presented here demonstrate the range of applicability of this method. In this regard, it is possible to combine prior knowledge concerning a particular binding site (biochemical information, natural and synthetic ligands) with the ligand design protocol, for instance by selecting directions of growth through the use of partially grown molecules as restart fragments or by modeling in a specific interaction known to exist in the natural ligands for the binding site (for example, a salt bridge or strong $\pi-\pi$ interactions) and using this modeled molecular fragment as a restart fragment. In principle, one can also build ligands that bind in two adjacent pockets on a protein surface, through careful mediation of the growth process using restart fragments.

One advantage of the design methodology proposed in this paper is the ability to improve the qualitative features (size, shape, location, connectivity, synthetic feasibility, π-stacking, hydrogen bonds, salt bridges, etc.) of the molecules without reducing their quantitative scores. Because of this, the quality of the ligands that can be generated is simply a product of effort,

insight, and intuition on the part of the user. However, the insight and intuition are only needed as analysis tools, since SMoG continuously provides suggested alterations and extensions of molecules that form excellent chemical and spatial complementarity with the protein binding site. In this sense, SMoG overcomes the otherwise intractable combinatorial task of generating optimal molecular scaffolds for scrutiny and optimization.

Careful examination of the tables of SMoG scores and CHARMM interaction energies reveals that those subtle molecular features that were added manually to take advantage of a hydrogen-bonding opportunity are not reflected very strongly in the SMoG score, but are reflected in the CHARMM interaction energy (particularly the electrostatic component). Conversely those attributes which correspond to increased hydrophobic interaction are reflected in the SMoG score, but not in the CHARMM interaction energy. This evidence supports using both measures of interaction energy equally, since their weaknesses and strengths are complementary. It also implies that the accuracy of the SMoG prediciton of binding free energy may be improved by adding specific terms to the form of the interaction potential that reflect electrostatic interactions such as hydrogen bonds and salt bridges. This work is currently in progress and is scheduled to appear as the third paper in this series.

In the role for which it has been designed, SMoG provides several advantages over other popular design methods. These include simple efficiency (each molecule taking just seconds on a personal computer), generating and evaluating whole molecules rather than separate fragments which later need to be linked, and most importantly, documented correlation between the scoring method and free energies of binding. Moreover, as these examples have indicated, the flexibility of the design protocol using SMoG allows one to combine chemical intuition, specific synthetic strategies, and prior understanding of the binding site into a strategy that continually improves the quantitative interaction score.

The SMoG limitations include those implied in the simple methods with which chemical geometry is handled: interfragment bond lengths and angles are all assumed to be standard and unvarying; the protein structure is considered fixed; steric repulsions are either on or off, depending on a simple distance test. Other limitations are implementation dependent, and the program has been designed to allow flexibility in the choice of operating conditions. For example, smaller-angle steps can be chosen to perform calculations more carefully, lower temperatures can be chosen, and the fragment library can be expanded.

Of course, as is the case with any design method, the crucial test of the SMoG merit will include the synthesis and measurement of the binding constant of a candidate ligand that was the direct result of SMoG design. We are currently pursuing this line of development vigorously.

Potential extensions to de novo ligand design using SMoG include monomer design in a second-generation combinatorial chemistry experiment, where some lead compound was elucidated in the first library whose structure has been determined. In this application, SMoG could be used to tailor a monomer for any positions where very weak consensus was observed in the original library. Also, present work is aimed at generalizing the approach to allow computational library testing and design.

We are confident that this approach, which is unique in many aspects, including the nature and source of the interaction potential and the growth algorithm, has much to offer the medicinal chemistry community because of its efficiency and the reliability of its scoring method. Moreover, as this brief account demonstrates, the approach to designing ligands is extremely flexible and fruitful.

Acknowledgment. We thank Stuart Schreiber and Sibo Feng for the structural information concerning the SH3 ligand design and Peter Pallai for CD4. Also, Jim Morken's extremely helpful advice on synthetic feasibility is much appreciated. This work was sponsored by the generous support of the Packard foundation. R.S.D. also acknowledges support from the Natural Sciences and Engineering Research Council (Canada).

Supporting Information Available: A listing of the atom types and g_{ij} interaction parameters (2 pages). See any current masthead page for ordering and Internet access instructions.

JA963689+