Multiple Highly Diverse Structures Complementary to Enzyme Binding Sites: Results of Extensive Application of a *de Novo* Design Method Incorporating Combinatorial Growth

Regine S. Bohacek* and Colin McMartin

Contribution from the Research Department, Pharmaceuticals Division, Ciba-Geigy Corporation, Summit, New Jersey 07901

Received August 5, 1993. Revised Manuscript Received April 5, 1994®

Abstract: A computer program for *de novo* molecular design was used to explore the diversity of molecules complementary to the binding sites of enzymes. The program, GrowMol[1] (preliminary results presented at the XIIth International Symposium on Medicinal Chemistry, Basel, Switzerland, 1992), generates molecules with steric and chemical complementarity to the three-dimensional structure of a host binding site. The molecules are created in the host binding site one atom or functional group at a time. At each step the position and type of atom to be added is randomly selected from a set of possible values which are consistent with internal bond lengths and bond angle requirements as well as the spatial and chemical properties of the binding site. The selection is achieved using Boltzmann statistics to bias acceptance toward atoms which can form favorable interactions with the binding site. Rings are generated by connecting closely positioned nonbonded atoms. This process ensures that a highly diverse set of molecules is generated unbiased by knowledge of previous guest molecules. When applied to several enzyme binding sites, the program produced structures which were identical to or closely resembled known inhibitors of these enzymes. In addition, the program rapidly produced *tens of thousands of distinct molecules* which display a large variety of structural motifs. A detailed analysis of the potential diversity of thiol inhibitors of thermolysin was carried out. Twenty-two thousand structures were generated in the thermolysin binding site. Twelve thousand of these were unique. After energy minimization in the active site and rejection of structures with a high conformational strain energy, four thousand unique structures were found which had an estimated potency comparable to those of known inhibitors. To investigate the range of diversity of these structures, they were clustered into distinct families. A representative example from each cluster was selected, resulting in a set of approximately three hundred structures which were examined visually. The resulting set included structures which were substrate mimics, non-peptidic structures which satisfied the hydrogen-bonding requirements of the binding site in novel ways, a large variety of different groups filling the hydrophobic binding pockets, and various macrocycles and fused-ring structures which spanned adjacent binding pockets. Many of these structures differ considerably from each other and have little in common with known inhibitors. To identify structures most likely to bind well to the enzyme, a scoring algorithm based on complementarity was developed. The estimated potency determined with this algorithm correlated well with the experimentally determined values of known inhibitors. The results of this study clearly demonstrate that molecular structures generated by computer programs incorporating algorithms for *de novo* design can reveal a wealth of opportunities for the design of novel enzyme inhibitors.

Introduction

The availability of the structure of a receptor binding site at atomic resolution offers a unique opportunity for the design of novel drugs. Currently, methods used to design inhibitors complementary to such host binding sites are being revolutionized by advances in computer technology.

Traditionally the design of new enzyme inhibitors has been based on knowledge of molecules known to bind to the target enzyme. The enzyme's substrate, "designed by nature" to bind tightly to the enzyme, is often used as a starting point. Very potent transition-state mimics have been designed in this way.[2,3] However, inhibitors that mimic the substrate too closely may be undesirable as drugs. They may be susceptible to enzymatic degradation or too polar to be absorbed from the intestinal tract or to cross cell membranes. The discovery of small potent, orally bioavailable molecules that bind to a specific target is the goal of the medicinal chemist.

Since enzymes are "designed" to bind to the transition state of a substrate, how different can a molecule be from the substrate and still bind to the enzyme? One way in which this question has been explored is by systematic modification of substrates or known inhibitors. Biased by known compounds, this method is unlikely to completely sample the full range of structures compatible with an enzyme binding site. However, computational methods are now available which do not require knowledge of known inhibitors but use only the three-dimensional structure of the binding site as a design target. In this study we wanted to investigate what these computational methods could tell us about the structural diversity of molecules complementary to a known binding site.

Two different strategies for the discovery of new inhibitors complementary to a binding site have been described. One method takes molecules from data bases, docks them into an enzyme binding site, and evaluates them. This method explores the possibility that already known compounds, possibly unrelated to the substrate, may in fact be inhibitors of the enzyme. Examples of programs which have successfully applied this approach to drug design include DOCK[4] and CLIX.[5] A second, more recently developed approach, generates *de novo* molecular structures that fit into a three-dimensional representation of the binding site.

® Abstract published in *Advance ACS Abstracts*, June 1, 1994.

[1] Preliminary results presented at the XIIth International Symposium on Medicinal Chemistry, Basel, Switzerland, 1992.

[2] Wiley, R. A.; Rich, D. H. *Med. Res Rev.* **1993**, *13*, 327–384. Rich, D. H. In *Medicinal Chemistry 1992—Proceedings of the XIIth International Symposium on Medicinal Chemistry*; Testa, B., Kyburz, E., Fuhrer, W., Giger, R., Eds.; Verlag Helvitica Chimica Acta: Basel, Switzerland, 1993; pp 15–23.

[3] Bartlett, P. A.; Marlowe, C. K. *Biochemistry* **1983**, *22*, 4618–4624. Bartlett, P. A.; Marlowe, C. K. *Biochemistry* **1987**, *26*, 8553–8561.

[4] Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A. *J. Mol. Biol.* **1982**, *161*, 269–288. Leach, A. R.; Kuntz, I. D. *J. Comput. Chem.* **1992**, *13*, 730–748.

0002-7863/94/1516-5560$04.50/0

Not limited to structures registered in data bases, these methods have the potential to reveal a much larger number of structures which fit the binding site.

A recent article by Rotstein and Murcko[6] provides an excellent review of various methods which have been successfully applied to the *de novo* generation of molecules. Two fundamentally different *de novo* algorithms have been used. One of these docks molecular fragments, selected from a predefined library, into the binding site as isolated units and then connects them to form molecules. The programs GROW[7] and LUDI[8] are examples of the use of this strategy. A second approach grows molecules in the binding site an atom or a fragment at a time.[6,9] LEGEND,[9] which uses atoms, generates molecules which satisfy van der Waals, i.e. steric, requirements for filling the binding site. An additional program LORE is then used to select structures with good molecular mechanics energy and hydrogen bonding to the binding site. LEGEND was the first program reported in the literature which generates *de novo* novel, organic molecules in a binding site. GROUPBUILD[6] grows molecules by the addition of fragments from a library. At each step all the possible fragment additions are evaluated according to the molecular mechanics energy and one of the best is randomly chosen.

In this paper we describe our computer program, GrowMol, and its application to the generation of structures in enzyme binding sites. One of the significant ways in which GrowMol differs from other *de novo* programs is the way in which new atoms are selected to be added to the growing chain. GrowMol evaluates each new atom according to its chemical complementarity to the nearby enzyme atoms. We have found that molecular mechanics interaction energy does not correlate well with potency and that using hydrophobic and hydrogen bonding is a more successful predictor of the ligand atom type found near certain types of enzyme atoms.[10] A Boltzmann weighting factor is used to bias the probability of selection toward atoms with a high complementarity score. Atoms which are less complementary will also be chosen but with a low probability. This allows structures to be formed in which groups of atoms with favorable interactions can be connected by atoms which are less complementary. The use of Boltzmann statistics ensures that most of the final structures will be highly complementary to the binding site.

Another important distinguishing feature of the GrowMol program is the ability to connect a newly grown atom to a previously grown atom in the growing structure. In this way polycyclic rings and fused aromatic systems are generated. Such conformationally restricted structures are of special interest. Many known drugs are conformationally restricted molecules which fix the position of hydrophobic groups in the optimum positions for binding to the enzyme and prevent hydrophobic collapse in solution.[2] Conformationally restricted molecules have a positive effect on binding due to favorable entropic effects. The addition of these complex cyclic structures to the repertoire of grown structures increases the chance of finding groups of atoms which have high steric and chemical complementarity to the binding site.

The multiple choices available in each step of the growth process result in very large sets of unique structures due to a combinatorial effect. This program is, therefore, suitable for exploring the diversity of potential ligands for a host binding site.

An important part of the *de novo* design process is the evaluation and ranking of the resulting structures. At present, quite independently of the development of *de novo* growth programs, considerable research is being to devoted the prediction of binding free energies. Some successes have been achieved using free energy perturbation methods to compute the relative binding energy of two ligands.[11] However, this method is computer intensive and can be applied only to cases where there are small differences between the two inhibitors under investigation. Presently no methods are available which will accurately determine the binding free energy of novel ligands. We have developed a simple, approximate estimate of inhibitory potency (K_i) based on the number of favorable enzyme/ligand contacts and use this method in the present study.

GrowMol was applied to HIV protease, pepsin, and thermolysin. Structures generated in the binding site of thermolysin were subjected to a detailed investigation. Thermolysin is particularly useful for this type of study because of the variety of inhibitor, substrate, and crystallographic data available. The structures of native thermolysin and several thermolysin/inhibitor complexes have been determined by X-ray crystallography and are available from the Brookhaven Protein Data Bank. Furthermore, there is experimental evidence that the crystal structure of thermolysin is similar to the solution conformation.[12] Crystalline thermolysin is able to hydrolyze peptides.[13] Also significant for this study is the finding that the thermolysin binding site does not change appreciably when complexed with numerous different inhibitors. Thermolysin is a bacterial zinc metallo protease. A growing number of mammalian zinc metallo proteases have been discovered which are of interest as therapeutic targets. Examples include neutral endopeptidase (24.11) (NEP), angiotensin converting enzyme (ACE), stromelysin and collagenase, and, more recently, a number of toxins including tetanus toxin and botulinum neurotoxin.[14] Similarities have been shown between the location and function of several amino acids important for the mechanism in all of these proteases.[15] For these reasons, thermolysin has been frequently used as a model for zinc peptidases. The combination of structural and biological information and the widespread use of thermolysin as a model for therapeutic target enzymes make it ideal for a study of the diversity of potential inhibitors.

For this study only the S1′ and S2′ parts of the thermolysin binding site were used. This part of the binding site has the best defined pockets, and X-ray crystal structures of small inhibitors which only use this part of the binding site are available. Since thermolysin is a zinc metallo peptidase, known inhibitors display a variety of different zinc-chelating groups. This study focuses on thiols.

Although the S1′ and S2′ binding site of thermolysin is quite small, the number of diverse structures generated was very large. Consequently additional methods had to be developed for the selection of a smaller set of structures which represent the full range of diversity of molecules with good estimated binding energies. Visual inspection of this relatively small set was then used to analyze the different types of structures and their binding modes.

This paper describes the generation, evaluation, and analysis of a large, highly diverse set of structures which satisfy steric, hydrophobic, hydrogen bonding, and energetic requirements for binding to the catalytic site of thermolysin.

[5] Lawerence, M. C.; Davis, P. C. *Proteins: Struct., Funct., Genet.* **1992**, *12*, 31–41.

[6] Rotstein, S. H.; Murcko, M. A. *J. Med. Chem.* **1993**, *36*, 1700–1710.

[7] Moon, J. B.; Howe, W. J. *Proteins: Struct. Funct., Genet.* **1991**, *11*, 314–328.

[8] Bohm, H.-J. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 61–78.

[9] Nishibata, Y.; Itai, A. *Tetrahedron* **1991**, *47*, 8985–8990.

[10] Bohacek, R. S.; McMartin, C. *J. Med. Chem.* **1992**, *35*, 1671–1684.

[11] Wong, C. F.; McCammon, A. *J. Am. Chem. Soc.* **1986**, *108*, 3830–3832. Bash, P. A.; Singh, U. C.; Langridge, R.; Kollman, P. A. *Science* **1987**, *236*, 564. McCammon, J. A.; Harvey, S. C. *Dynamics of Proteins and Nucleic Acids*; Cambridge University Press: New York, 1987. van Gunsteren, W. F., Weiner, P. K., Eds.; *Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications*; Escom Science Publishers: Leiden, The Netherlands, 1989. Jortensen, W. L. *Acc. Chem. Res.* **1989**, *22*, 184.

[12] Horrocks, W. D., Jr.; Sudnick, D. R. *Acc. Chem. Res.* **1981**, *14*, 384.

[13] Holmes, M. A.; Matthews, B. W. *Biochemistry* **1981**, *20*, 6912–6920.

[14] Schiavo, G.; Shone, C. C.; Rossetto, O.; Frances, C. G.; Montecucco, A.; Montecucco, C. *J. Biol. Chem.* **1993**, *268*, 11516–11519.

[15] Vallee, B. L.; Auld, D. S. *Biochemistry* **1990**, *29*, 5647–5659.

Table 1. Classification of Binding Site Zones (Distances and Atom Types Used To Define the Zones of the Grid Representation of the Binding Site)

		range of distances (Å)		
binding site zone	code	lower	upper	binding site atom
Grid Map 1 (Used for Structure Generation)				
forbidden				
hydrogen bond acceptor	1	0.0	V_i[a]	oxygen, —N=
hydrogen bond donor	1	0.0	V_i	polar hydrogen
hydrophobic	1	0.0	$V_i + 0.9$	nonpolar atoms
contact				
hydrogen bond acceptor	6[b]	V_i	3.0	oxygen, —N=
	3	3.0	$V_i + 2.1$	
hydrogen bond donor	7[c]	V_i	2.6	polar hydrogen
	4	2.6	$V_i + 2.1$	
hydrophobic[d]	2	$V_i + 0.9$	$V_i + 2.1$	nonpolar atoms
neutral	5	$V_i + 2.1$	$V_i + 3.5$	all atoms
Grid Map 2 (Used for Structure Evaluation)				
forbidden				
hydrogen bond acceptor	1	0.0	1.4	oxygen, —N=
hydrogen bond donor	1	0.0	1.4	polar hydrogen
hydrophobic	1	0.0	V_i[a] + 0.9	nonpolar atoms
contact				
hydrogen bond acceptor	6[b]	1.4	3.0	oxygen, —N=
	3	3.0	$V_i + 2.1$	
hydrogen bond donor	7[c]	1.4	2.6	polar hydrogen
	4	2.6	$V_i + 2.1$	
hydrophobic[d]	2	$V_i + 0.9$	$V_i + 2.1$	nonpolar atoms
neutral	5	$V_i + 2.1$	$V_i + 3.5$	all atoms

[a] van der Waals radii of enzyme atoms: sp^3 carbon, 1.65; sp^2 carbon, 1.5; single-bond oxygen, 1.35; double-bond oxygen, 1.3; —N=, 1.75; sulfur, 2.55; hydrogen, 1.0. [b] Forbidden to ligand atoms except hydrogen bond donors. [c] Forbidden to ligand atoms except hydrogen bond acceptors. [d] Grid points in the hydrophobic zone must *not* be in either the hydrogen bond acceptor or donor zone.

Computational Methodology

The application of GrowMol involves four major steps: (1) preparation of a three-dimensional grid map of the binding site; (2) generation of the molecular structures; (3) estimation of potency; (4) evaluation of generated structures.

1. Construction of the Binding-Site Grid Map. The grid map is a three-dimensional representation of the volume enclosed by the binding site cavity. A regular array of grid points separated by 0.25 Å is constructed to enclose the entire binding site. At each grid point the enzyme binding site environment is tested and an appropriate integer is placed into the corresponding position of a three-dimensional matrix. Table 1 gives the distance cutoff values for each of the different enzyme binding site zones and the corresponding integer assigned to the matrix. For example, grid points which are too close to enzyme atoms are assigned a value of 1 to identify them as lying in a forbidden region. Values of 6 and 7 are assigned to grid points within hydrogen-bonding distance to enzyme atoms. These zones can be occupied only by inhibitor atoms forming hydrogen bonds with the enzyme. The value 6 is assigned to the binding-site hydrogen bond acceptor zone, and therefore, only ligand hydrogens are allowed in this area. The value 7 is for the binding-site hydrogen bond donor zone, where only ligand hydrogen bond acceptors are allowed. Zones 3 and 4 are hydrogen bond acceptor and donor zones but further removed from binding-site atoms than zones 6 and 7. In these zone all types of atoms are allowed but, because of the proximity to hydrogen bond donors and acceptors, ligand atoms able to form hydrogen bonds will be favored. All remaining grid points which are within nonbonded contact distance to enzyme atoms are assigned a value of 2. Grid points which are too far from any enzyme atoms to be considered to be in direct nonbonded contact but are within approximately 5 Å of an enzyme atom are in a neutral zone and are assigned a value of 5. The calculations are performed by a computer program call GRIDBOX.

The criteria used for classifying different zones of the enzyme binding site grid are based on those previously determined to be optimum for the peptide/peptide interfaces within proteins as well as for the binding site/ligand interface of a series of inhibitors bound to an enzyme.[10] This previous study had shown that the distance between a point on the Lee and Richards accessible surface of a protein binding site and the hydrogen bond donor and acceptor atoms of the binding site is a powerful predictor of the type of *ligand* atom to be found nearby. For the present application, this method of mapping binding-site specificity was extended so that it was not restricted to the accessible surface but included the entire volume of the binding-site cavity. Table 1 gives the distances used to define the various binding-site zones.

The cutoff distances used to construct the different zones of the grid map are somewhat less than ideal. This allows for the generation of structures which upon further optimization by energy minimization may adopt conformations that are spatially and energetically compatible to the binding site. After energy minimization, the structures are re-evaluated with a second grid map with more stringent cutoff distances for the hydrogen bond zones (grid map 2, Table 1). The relationship between the potency of a compound and its position in the grid map is discussed in section 3 below.

2. Molecular Generation. To begin the molecular generation, the user selects a root atom, i.e. a point were "growth" is initiated. The root atom can be any atom present in the binding-site structure from which a molecule is to be "grown". The atom can belong to the enzyme or can be an atom belonging to an inhibitor fragment. Next the computer program determines growth points for the root atom. Growth points indicate all the positions available for new atoms which might be added to the growing structure. The growth point for a given atom is obtained from a look-up table of rotational isomeric states appropriate for that atom.

All subsequent atoms are generated by choosing the position and atom type for a new atom in the following manner:

(1) One of the growth points is randomly selected.

(2) An atom or functional group is randomly chosen from the functional group library. The atoms and functional groups used in this study are sp^3 carbon, oxygen, nitrogen, negatively charged oxygen, hydrogen, carbonyl group, NH, benzene, and five-membered unsaturated rings. Hydrogens are bound only to nitrogen or to oxygen. Carbon atoms are generated without protons.

(3) The coordinates of the new atom are computed using the bond length, bond angles, and dihedral angles associated with the growth point selected in step 1. These are based on the geometrical parameters of the MM2 force field.[16] To create heterocyclic unsaturated five-membered rings, all possible rings containing up to two nitrogens are generated and one of these is selected using procedures 4 and 5 given below.

(4) The binding-site zone occupied by the new atom is identified. If the atom is in a forbidden zone, i.e., too close to a binding-site atom, the atom and the associated growth point are erased and the program returns to step 1. If the atom is in one of the allowed zones, a complementarity score is assigned which reflects the degree of complementarity between the new atom and the binding-site zone it occupies. For five- and six-membered rings, up to two atoms are permitted in the forbidden zone.

(5) A Metropolis sampling[17] criterion is used to decide if the new atom will be retained and added to the growing chain. The probability of retaining the new atom is given by the Boltzmann factor, BF = exp(−complementarity score/*RT*). The complementarity score (see below) is used as the energy in the Metropolis sampling criterion. A random number between 0 and 1 is generated, and if it is less than BF, the new atom is connected to the rest of the chain. The original growth point used to generate

[16] Allinger, N. L. *J. Am. Chem. Soc.* **1977**, *99*, 8127–8140.

[17] Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087–1092.

the new atom is erased, and the new growth points associated with the new atom are computed and stored. If the random number generated is greater than BF, the atom is erased, the growth point is restored, and the program returns to step 1.

The temperature used in the Boltzmann calculation is not intended to represent a temperature with physical meaning. The advantage of using this method for evaluating an atom is that the user can select different temperatures and thus allow the generation of structures with varying degrees of complementarity to the enzyme active site. For instance, higher temperatures will produce molecules which are less complementary but which may posses highly novel features. In this study a temperature of 300 K was used.

The *complementarity scores* used in this study are as follows: −10 for an atom occupying a complementarity zone (i.e. a carbon in a hydrophobic zone or hydrogen-bonding atoms in the appropriate acceptor or donor zone); 0 for any atom occupying the neutral zone; 10 for a mismatch between atom type and zone (i.e. oxygen in the hydrogen bond acceptor or hydrophobic zone or NH in the donor or hydrophobic zone). A smaller penalty, i.e. a score of 5, is used for carbon in a hydrogen bond acceptor or donor zone. This increases the probability of placing carbons in hydrophilic zones where they may be necessary as linkers of various sections of the chain. Atoms belonging to aromatic rings which are in the forbidden zone incur a score of 10 points each.

The complementarity scores are used to bias the selection of atoms to those which are highly complementary to the binding site. The scores can be changed by the user to control the relative amounts of hydrophobic and hydrogen-bonding contacts found in the ensemble of structures generated. In the present version of GrowMol, the complementarity scores used in the growth process are not intended to accurately represent contributions to the free energy of binding.

The user could select complementarity scores that exactly reflect contributions to free energy where sufficient data is available to obtain these parameters. In our experience, however, the selection of these scores must also take into account the fact that in some parts of the molecule *noncomplementary* atoms may be necessary to ensure continued growth into regions where high complementarity can be obtained.

(6) Once the user-specified number of atoms has been exceeded as a result of atoms added in one growth step, the molecule is saved. The user may also specify minimum numbers of hydrophobic contacts and hydrogen bonds which must be met in order for a molecule to be saved. However, in order to determine the number of unique structures that the program can generate, this option was not used in this study and no requirements for minimum enzyme/ligand interactions were given.

If all of the growth points have been used or the user-specified number of atoms or complementary contacts has not been reached, the molecule is not saved. In either case the program proceeds to step 7.

(7) The arrays are reinitialized, and the program returns to step 1 to begin the growth of a new molecule. When the user-specified number of molecular structures has been generated, the program stops.

In addition to the above rules, a number of further processes occur at each growth step.

(1) The distances between each new atom and all other atoms of the growing structure which are in nonbonded contact are computed. Nonbonded contacts are considered to occur between atoms separated by at least three atoms. If the distance is <1.0 Å, the atom is rejected. If the distance and bond angle are appropriate for bond formation and the atom satisfies the Metropolis criterion, a ring is closed by connecting the atoms. This allows spontaneous generation of cyclic systems.

(2) A few additional rules are used to ensure chemically realistic structures. For instance, each sp^2 atom must be connected to at least one other sp^2 atom, and chemically unstable groups such as peroxides are not allowed.

Table 2. Complementarity between Ligand and Enzyme in Thermolysin/Inhibitor Complexes Determined by X-ray Crystallography

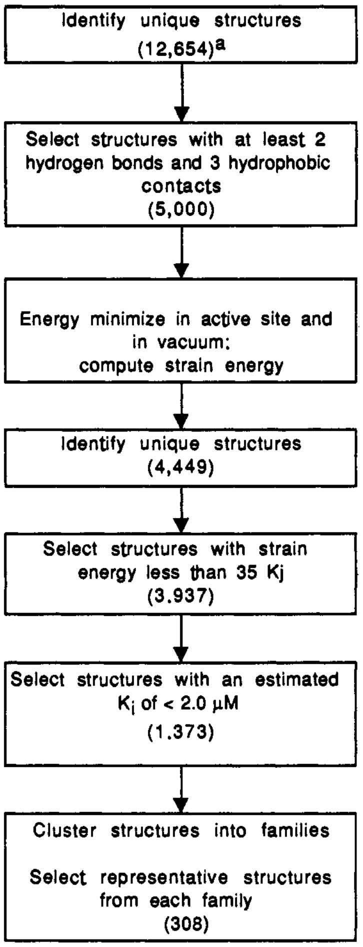
ligand[a]	potency[b] K_i, μM	hydrophobic contacts	hydrogen bonds[c]	estimated potency[d] K_i, μM	
				X-ray data	energy minimized
ZF^{P}LA	0.000 068	9	8	0.0001 8	0.0064
ZG^{P}LL	0.0091	7	6	0.0076	0.0031
phosphoramidon	0.028	5	6	0.053	0.0087
CLT	0.05	6	7	0.0082	0.0070
HONH-BAGN	0.43	5	4	0.032	0.049
BAG	0.75	5	3	0.78	0.78
P-Leu-NH_2	1.3[e]	2	5	2.0	0.89
THIO	1.8[f]	4	3	2.0	0.78
RETRO	2.3[f]	4	3	2.3	0.78
r^2 [g]	0.94				
p	0.0003	0.0049	0.0091		

[a] X-ray structures of the ligand/thermolysin complexes were used. Cbz = carbobenzoxy; ZF^{P}LA = Cbz-Phe^{P}-L-Leu-L-Ala;[23] ZG^{P}LL = CBZ-Gly^{P}-L-Leu-L-Leu;[19] phosphoramidon = *N*-[(α-L-rhamnopyranosyloxy)hydroxyphosphinyl]-L-Leu-L-Trp;[24] CLT = *N*-(1-carboxy-3-phenylpropyl)-L-Leu-L-Trp;[25] HONH-BAGN = HONH-(benzylmalonyl)-L-Ala-Gly-*p*-nitroanilide;[13] BAG = (2-benzyl-3-mercaptopropanoyl)-L-alanylglycinamide;[26] P-Leu-NH_2 = *N*-phosphoryl-L-leucinamide;[24] THIO = thiorphan (*N*-[(*S*)-2-(mercaptomethyl)-1-oxo-3-phenylpropyl]glycine);[27] RETRO = retrothiorphan ((((R)-1-(mercaptomethyl)-2-phenylethyl)amino)-3-oxopropanoic acid).[27] [b] From data compiled by Matthews.[20] [c] Bifurcated hydrogen bonds, i.e. between ligand carbonyl and Arg 203, are counted only once. [d] K_i computed using the equation resulting from the linear regression between the logarithm of potency, the number of hydrophobic contacts, and the number of hydrogen bonds. [e] Because there is a discrepancy between the data compiled by Matthews[20] and the original data, the original data of Powers et al.[21] is used. [f] See ref 22. [g] r^2 = correlation between the logarithm of potency, the number of hydrophobic contacts, and the number of hydrogen bonds. p = significance of the linear regression.

3. **Estimation of Potency.** A simple approximate estimate of potency, based on the number of favorable enzyme/inhibitor contacts, was used to evaluate the structures generated by GrowMol. The empirical equation used was derived from experimental structural and potency data in the following manner: nine of the ten[18] potent thermolysin inhibitors for which the binding mode in thermolysin has been determined by X-ray crystallography were placed in the grid representation of the active site of thermolysin. The number of hydrophobic contacts (i.e. the number of ligand carbons other than carbonyl carbons which occupy the hydrophobic zone) and the number of hydrogen bonds (defined as the number of ligand hydrogens in the hydrogen acceptor zone plus the number of ligand oxygens found in the hydrogen bond donor zone) of each inhibitor were determined. These numbers are obtained using a computer program called EVAL, which uses the grid map to determine the binding-site zone occupied by each inhibitor atom. Table 2 gives the values obtained for the thermolysin inhibitors in grid map 2 (the grid map used for evaluation of the GrowMol structures). A multiple linear regression was carried out to correlate the potency of each inhibitor with the number of hydrophobic contacts and the number of hydrogen bonds the inhibitor makes with the enzyme. The following results were obtained:

[18] One of the inhibitors, ZG^{P}(O)LL, was not used because it was designed to determine the loss in potency resulting from changing an NH (which forms a hydrogen bond with the enzyme) to an oxygen.[19] This change places two oxygens which cannot hydrogen bond together and results in a 1000-fold reduction in potency. Because of insufficient data, we do not score such unfavorable contacts in this analysis and, therefore, this inhibitor was not included in the analysis. For further details on the effect of including this inhibitor in this type of analysis see ref 10.

Table 3. Procedure Used for Identifying the Most Promising Structures out of a Set of 22 000 Structures Containing 17–25 Atoms Generated in the S1′ and S2′ Areas of the Thermolysin Binding Site[a]


Identify unique structures
(12,654)a
Select structures with at least 2 hydrogen bonds and 3 hydrophobic contacts
(5,000)
Energy minimize in active site and in vacuum:
compute strain energy
Identify unique structures
(4,449)
Select structures with strain energy less than 35 Kj
(3,937)
Select structures with an estimated Ki of < 2.0 µM
(1,373)
Cluster structures into families
Select representative structures from each family
(308)

[a] The number of structures after each step is given in brackets.

$$\begin{array}{lllll} \log(K_i) = & 3.16 & -0.42(\mathrm{PHOB}) & -0.39(\mathrm{HBOND}) & \\ p & 0.0003 & 0.0049 & 0.0091 & \\ n = 9 & \mathrm{sd} = 0.40 & p = 0.0002 & r^2 = 0.94 & \end{array}$$

where PHOB = the number of hydrophobic contacts and HBOND = the number of hydrogen bonds. This linear regression equation was used to convert the number of favorable enzyme/inhibitor contacts of GrowMol structures to an approximate K_i.

A similar analysis was carried out to check grid map 1. (Grid map 1 is used to generate the structures and is somewhat larger than the actual volume of the binding site to allow the inclusions of structures which may upon energy minimization fit into the binding–site cavity.) The correlation gave an $r^2 = 0.92$.

Since the computer-generated structures are energy minimized in the active site of the enzyme, it was necessary to test the scoring procedures on energy-minimized structures. The nine thermolysin inhibitors were subjected to energy minimization in the binding site of thermolysin using the GROWMIN[28] program. The complementary contacts were then determined for each minimized structure. The correlation between biological activity and complementarity for the energy-minimized structures using grid maps 1 and 2 gave r^2 values of 0.77 and 0.80, respectively. The correlation of the conformations of the inhibitors resulting from energy minimization is, thus, not as high as that for the conformations determined by X-ray crystallography. This is not entirely surprising, since the zones of the accessible surface upon which the grid boxes were based were parametrized using high-resolution X-ray diffraction data of proteins.[10]

We do not claim that we can predict potency using this simple method; however, these results indicate that there is a strong relationship between the number of atoms which occupy complementary zones of the grid maps and the biological activity of a molecule.

For the present application we do not including terms which measure numbers of noncomplementary interactions. Examples of such interactions would be ligand hydrogen bonding atoms found in an inappropriate binding-site zone, i.e. a ligand oxygen close to an enzyme oxygen or in a hydrophobic zone. These extra terms would undoubtedly improve the correlation, but the limited number of compounds available for establishing the regression does not justify the use of so many explanatory variables.

4. Evaluation of Structures. Although the structures are generated to be spatially and chemically complementary to the receptor binding site and to have low internal energy, they do not all form equally good interactions with the binding site. Methods were developed to evaluate each structure based on (1) favorable interactions the structure makes with the enzyme binding site atoms (expressed either as the number of hydrophobic contacts and the number of hydrogen bonds or as the estimated K_i) and (2) the internal molecular mechanics energy of the bound structure. A series of procedures was developed which uses these criteria to filter out the less desirable structures, eventually retaining a smaller set of only the most interesting structures which can be clustered in similar families and inspected visually.

The following procedures are used to identify the most promising structures (see Table 3):

(1) Duplicate structures were identified and rejected. Structures are considered to be identical if for each atom of one structure there is an atom in the other structure which has similar coordinates and is of the same atom type. The difference between the x, y, and z coordinates of the two atoms must be less than 0.5 Å.

(2) The number of hydrophobic interactions and the number of hydrogen bonds that a structure made with the binding site were counted. If the number of these interactions was less than a user-defined threshold, the structure was rejected.

(3) The remaining structures were subjected to energy minimization in the active site and in the absence of the active site. A Cartesian conjugate gradient minimizer, GROWMIN,[28] and the AMBER force field[29] parameters were used for these calculations. The difference between the energy of the bound conformation and the energy of the conformation minimized outside the binding sites was taken as an initial measure of the ligand strain energy.

(4) Any additional duplicate structures resulting from the movement of atoms during the energy minimization were identified and rejected.

(5) Structures with a ligand strain energy above 35 kJ were rejected. The strain energy is used as a way to detect structures which can *only* fit into the binding site with considerable distortion. Molecules which bind well to a binding site are unlikely to be highly strained. For example, active inhibitors such as the thermolysin inhibitors listed in Table 2 all have ligand strain energies under 25 kJ.

(6) Until this step, a broad definition of complementarity has been used, anticipating that the complementarity might change upon energy minimization in the active site. After minimization, a second, more stringent, set of definitions for complementarity (Table 1) is applied. In this step, the complementarity of the structures was re-evaluated and those with an estimated K_i of less then a user-specified threshold were rejected.

(7) The remaining structures were clustered into families based on similarity. Two structures were taken to be in the same family if 60% of the atoms of the larger structure were within 0.5 Å of an atom in the other structure.

(8) For each cluster, the structure with the lowest estimated

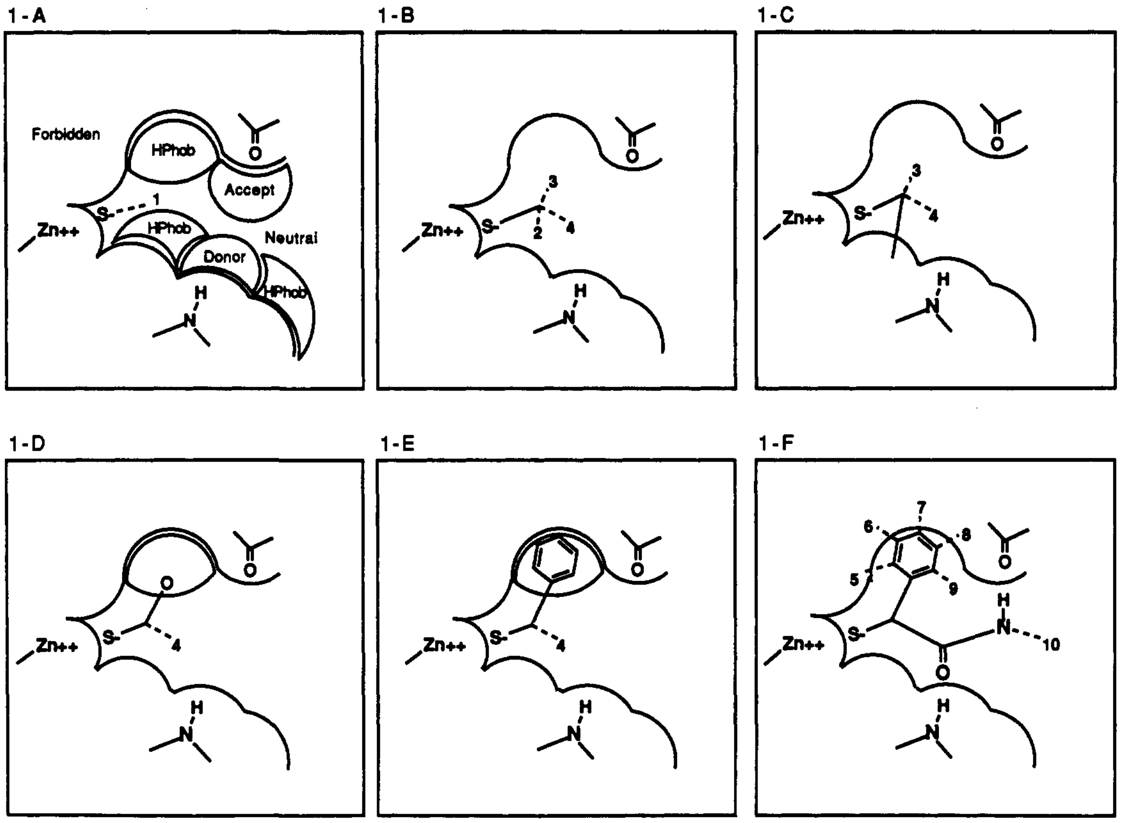


Figure 1. Cartoon representation of successive steps in the generation of a molecule in the binding site of an enzyme. Figure 1-A shows the accessible volume of the binding site. The volume of the binding site is divided into zones which reflect the chemical specificity of the enzyme (See Table 1). HPHOB is the hydrophobic zone and ACCEPT and DONOR are the hydrogen bonding acceptor and donor zones, respectively. Near the enzyme atoms is the forbidden zone. In the center is the neutral zone. Figure 1-Z also shows the root atom (sulfur). At this stage there is a single growth point connected to the root atom by a dotted line. In Figure 1-B, a carbon atom occupies the position specified by the first growth point. Three new growth points have been added to this atom. In Figure 1-C, position 2 has been randomly selected and a carbon atom, also randomly selected, has been generated. The atom is found to be in the forbidden zone of the binding site, and the atom and growth point are eliminated. In Figure 1-D, position 3 and an oxygen atom have been selected. This atom is in a hydrophobic zone and, therefore, not compatible. Thus there is only a low probability that this atom will be retained, and in this example it was eliminated. In Figure 1-E, a benzene ring has been generated in the hydrophobic zone and is retained. In Figure 1-F, the process continues until the binding site is filled or the user-specified number of atoms has been reached.

K_i was selected to represent that cluster. If multiple structures occur with the same K_i, the structure with the lowest ligand strain energy was selected. The representative structures were ranked according to K_i and combined into one file ready for visual inspection.

Application to Thermolysin

The S1′ and S2′ portions of the thermolysin binding site obtained from the structure of thermolysin bound with ZG^{P}LL[19] were used in this study. The structure, labeled 5TMN, was obtained from the Brookhaven Protein Data Bank.[30] The binding site included the following residues: Asn 111, Asn 112, Ala 113, Phe 114, Thr 129, Phe 130, Leu 133, Asp 138, Val 139, His 142, His 143, His 146, Tyr 157, Glu 166, Ile 171, Ile 188, Gly 189, Val 192, Tyr 193, Ser 201, Leu 202, Arg 203, Asp 226, and His 231.

The root atom chosen for this study is a sulfur atom which binds to the zinc of the enzyme. (See Figure 1 for a cartoon representation illustrating the steps involved in the generation of a structure in the binding site of a zinc metallo protease.) The three-dimensional structure of a well-known thiol zinc metallo protease inhibitor, thiorphan,[22] bound to thermolysin has been determined by X-ray crystallography and published.[27] However, the coordinates of the entire thermolysin/thiorphan complex have not been published. Therefore, thiorphan was minimized into the thermolysin active site[31] using the thermolysin structure from 5TMN. The sulfur root atom as well as the initial growth point, which defines the dihedral angle that an atom connected to the sulfur will adopt, was obtained from this thermolysin/thiorphan complex. In this study, we decided to probe the site using thiols, and therefore, a carbon was selected for the first guest atom.

Results

Generation of Unique Structures in Thermolysin. GrowMol generated structures in the thermolysin binding site cavity very

[19] Bartlett, P. A.; Marlowe, C. K. *Science* **1987**, *235*, 569–571. Tronrud, D. E.; Holden, H. M.; Matthews, B. W. *Science* **1987**, *235*, 571–574.

[20] Matthews, B. W. *Acc. Chem. Res.* **1988**, *21*, 333–340.

[21] Kam, C.; Nishino, N.; Powers, J. C. *Biochemistry* **1979**, *18*, 3032–3038.

[22] Benchetrit, T.; Fournie-Zaluski, M. C.; Roques, B. P. *Biochem. Biophys. Res. Commun.* **1987**, *147*, 1034–1040.

[23] Holden, H. M.; Tronrud, D. E.; Monzingo, A. F.; Weaver, L. H.; Matthews, B. W. *Biochemistry* **1987**, *26*, 8542–8552.

[24] Tronrud, D. E.; Monzingo, A. F.; Matthews, B. W. *Eur. J. Biochem.* **1986**, *157*, 261–268.

[25] Monzingo, A. F.; Matthews, B. W. *Biochemistry* **1984**, *23*, 5724–5729.

[26] Monzingo, A. F.; Matthews, B. W. *Biochemistry* **1982**, *21*, 3390–3394.

[27] Roderick, S. L.; Fournie-Zaluski, M. C.; Roques, B. P.; Matthews, B. W. *Biochemistry* **1989**, *28*, 1493–1497.

[28] GROWMIN, unpublished results, C. McMartin.

[29] Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Algona, G.; Profet, S.; Weiner, P. A. *J. Am. Chem. Soc.* **1984**, *106*, 765–784.

[30] Berstein, F. C.; Koetzle, G. J. B.; Williams, G. J. B.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, R.; Tasumi, M. *J. Mol. Biol.* **1977**, *112*, 535–542. Abol, E. E.; Bernstein, F. C.; Bryant, S. H.; Koetzle, T. F.; Weng, J. In *Crystallographic Database-Information Content, Software Systems, Scientific Applications*; Allen, F. H., Bergerhoff, G., Suievers, R., Eds.; Data Commission of the International Union of Crystallography: Bonn, Cambridge, Chester, 1987; pp 107–132.

[31] Guida, W. C.; Bohacek, R. S.; Erion, M. D. *J. Comput. Chem.* **1992**, *13*, 214–228.

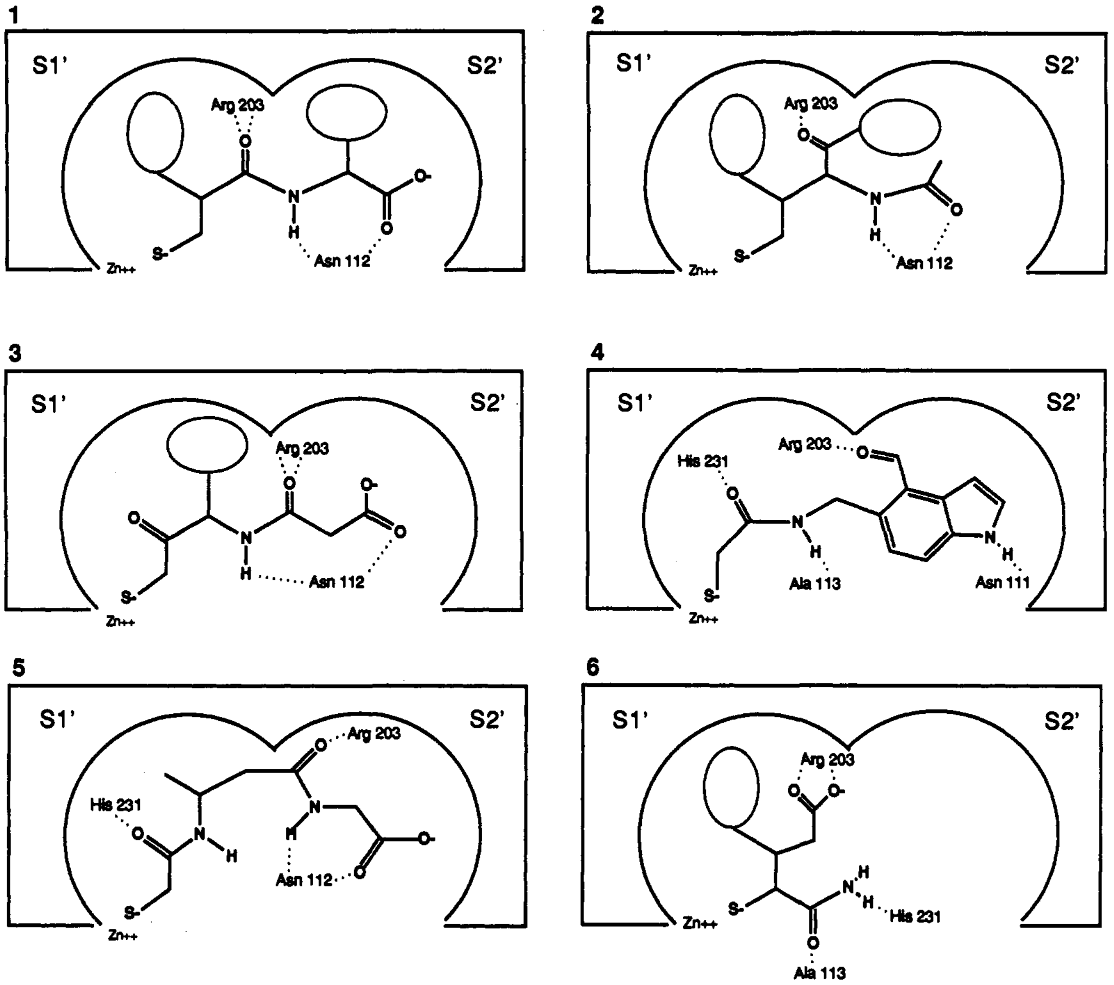


Figure 2. Different ways in which the backbone of grown structures extended through the binding site. The ellipses represent hydrophobic groups, e.g. branched alkane chains or cyclohexane or benzene rings. The dotted lines show hydrogen bonding to enzyme residues. The first example shows the binding mode proposed for the substrate. Most known inhibitors are based on this binding mode. The remaining examples show binding motifs which differ appreciably from this mode and provide novel solutions to the hydrogen-bonding requirements of the enzyme. The examples represent a small fraction of the total number of different binding modes found in this study.

Table 4. Number of Unique Structures Found When Sets of Molecules Containing Different Numbers of Atoms Were Generated in the Binding Site of Thermolysin

average no. of atoms	no. of structures	no. of unique structures	replication rate[a]	α[b]
12.0	50 000	3938	12.7	1.99
13.8	50 000	10 543	4.7	1.96
18.7	72 000	39 680	1.8	1.76

[a] The number of structures divided by the number of unique structures found. [b] Measure of diversity: number of unique structures = $\alpha^{(\text{number of atoms})}$.

rapidly. Structures with an average size of 14 atoms were generated at a rate of approximately 17 structures per minute using a VAX 6410 computer. It was, therefore, possible to first explore diversity by generating large test sets of structures in order to determine how many different structures GrowMol would actually generate. Secondly, by using the evaluation techniques, it was possible to reduce the large set of structures to smaller sets of only the most interesting structures which were analyzed individually.

We generated three large sets of structures: 50 000 with an average of 12.0 atoms, 50 000 with an average of 13.8 atoms, and 72 000 with an average of 18.7 atoms. The number of unique structures in each of these sets was determined and is shown in Table 4.

In the set of the smallest structures, each unique structure was found on average 12.7 times. This high replication rate indicates that most of the unique structures of this size that can be generated using GrowMol, as described here, have been found. For the third and largest set, the replication rate was only 1.8 although 72 000 structures were generated. Therefore, the number of unique structures will almost certainly underestimate the true diversity of molecules of this size.

Due to the combinatorial explosion, the number of unique structures might be expected to increase as an exponential function of the number of atoms:

$$N_{\text{unique}} = \alpha^{\text{Natom}}$$

This equation allows diversity to be described in terms of a single parameter, α. The value of α found for the first two sets was close to 2 (1.99, 1.96) whereas the third set has a lower value (1.76). This is probably due to insufficient sampling.

Another set of 7000 structures with 25–35 atoms was also generated in order to investigate some of the additional features which can appear in larger molecules. This set contained 4706 unique structures.

Diversity in Other Enzyme Binding Sites. A high degree of diversity was also found in structures generated in the binding sites of HIV protease and pepsin.[32] Structures were generated in the S2, S1, S1′, and S2′ sites of HIV protease. Only structures with at least five hydrophobic contacts and five hydrogen bonds were retained during the growth process. Of the 4893 structures

[32] Bohacek, R. S. Unpublished results.

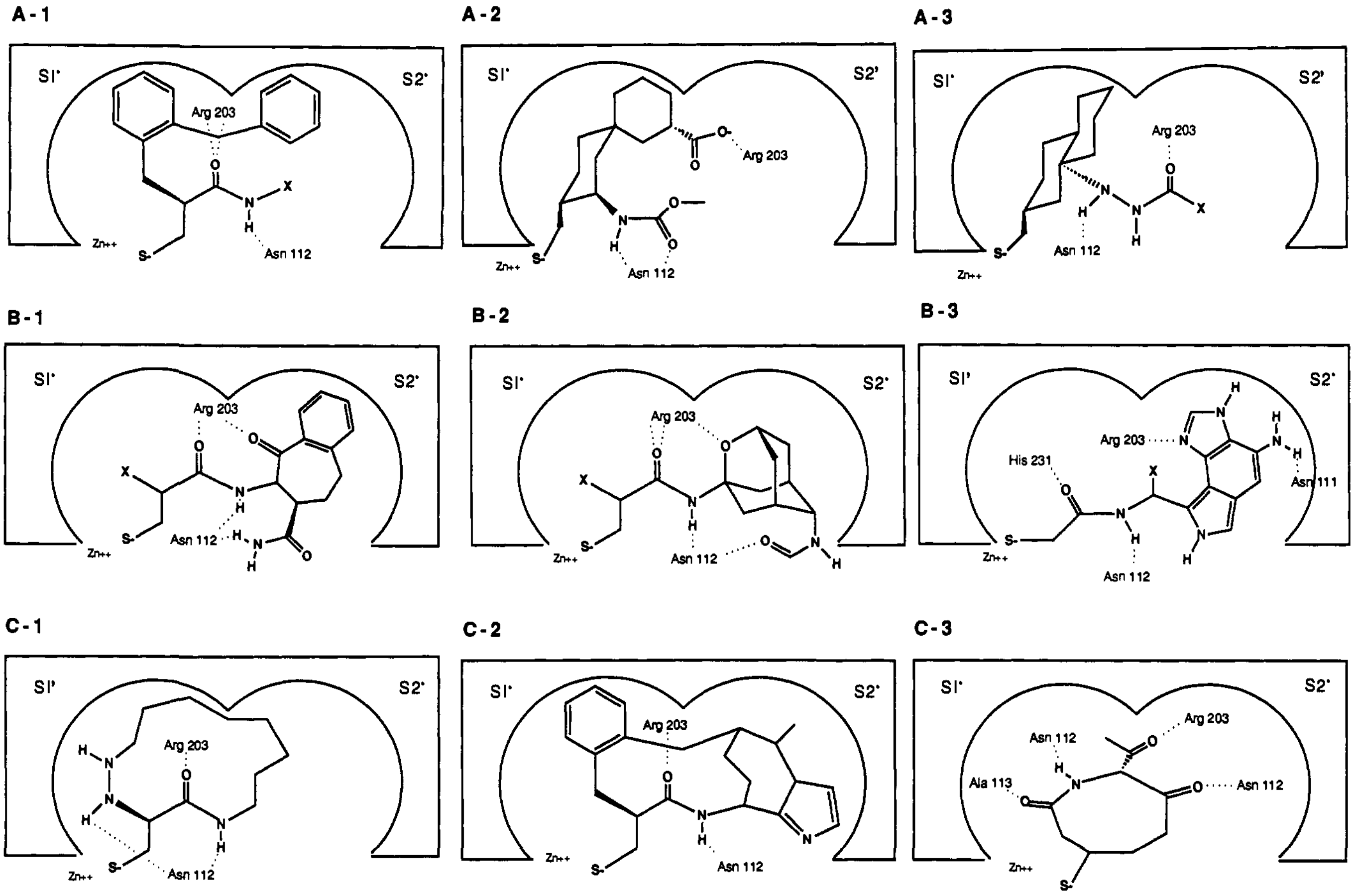


Figure 3. Examples of novel ways in which the S1′ and S2′ binding pockets of thermolysin can be occupied. Structures A-1–A-3 occupy the S1′ pocket; structures B-1–B-3 occupy the S2′ pocket; and the last structures are macrocycles which bridge the two pockets. These examples show the ability of the *de novo* growth algorithm to generate conformationally restricted structures which are chemically complementary and fill the complex volume of the binding site (also see Figure 4).

collected, 4855 were unique. The HIV protease structures ranged in size from 23 to 41 atoms. Structures were generated in the S3, S2, S1, and S1′ sites of pepsin. In this case only structures with at least four hydrogen bonds and four hydrophobic contacts were retained. The structures ranged in size from 19 to 59 atoms, and nearly all (4874) of the 4879 structures collected were unique. This shows that large numbers of unique structures can also be formed in these binding sites.

Evaluation of Structures. A set of 22 000 structures with an average size of 18.7 atoms generated in the binding site of thermolysin was evaluated as summarized in Table 3. Five thousand unique structures with at least three hydrophobic contacts and two hydrogen bonds were selected for energy minimization inside and outside of the binding site. After energy minimization, 551 duplicate structures were rejected, resulting in a set of 4449 unique structures. Of these, 3937 structures were found to have strain energies of less than 35 kJ. This set of structures was now suitable for further analysis (see sections 3 and 4 below). A much smaller set was obtained by selecting all structures with an estimated $K_i \leq 2$ μM and clustering these structures into distinct families. The structure with the lowest K_i was selected as a representative. (If multiple structures had the same K_i, then the structure with the lowest strain energy was chosen.) Three hundred and eight structures remained after this step. These were visually inspected and further classified (see sections 1 and 2 below).

Classification of Structures. The following properties were used to classify the structures: (1) the way in which the backbone chain extends through the binding site (we will call this the binding mode); (2) the functional groups which occupy a specific area of the enzyme binding site, e.g. one of the hydrophobic pockets; (3) the variation in the structures which have a constant feature, e.g. a phenyl in the S1′ pocket. In addition the structures were compared to known inhibitors to provide an experimental validation of the method.

1. **Backbone Motifs.** The structures in the present study were generated using the catalytic site and the S1′ and S2′ areas of thermolysin. Figure 2 shows a summary of the major binding modes of these structures. The binding mode in the first example in Figure 2 corresponds to the mode found experimentally for substrate mimics such as thiorphan. Of the 3937 structures, 1036 had an amide bond corresponding to that found in thiorphan, but only 135 of these had the carbon connected to the α carbon, as expected for a substrate mimic. The remaining figures show completely novel ways in which molecules occupied the active site. All of these motifs represent structures in low-energy conformations which form the specific hydrophobic and hydrogen-bonding interactions believed to be important for binding to thermolysin.

2. **Side-Chain Motifs.** The structures were analyzed according to the type of functional group found to occupy and S1′ and S2′ pockets and how they are connected to the rest of the molecule. Figure 3A shows some of the novel groups in S1′, and Figure 3B shows conformationally restricted groups in the S2′ area. In thermolysin the S1′ and S2′ pockets are not completely separated. Figure 3C shows a series of macrocycles, with rings containing between 8 and 13 bonds, which can span this part of the binding site. Figure 4 is a set of stereo diagrams showing how well the structures fill the binding site. Structures A-2, B-3, and C-2 are shown.

3. **Structures Sharing a Common Feature.** Another way to organize the data is to group together structures containing a common feature. This method is especially useful because it allows visual examination of the diversity of structures containing

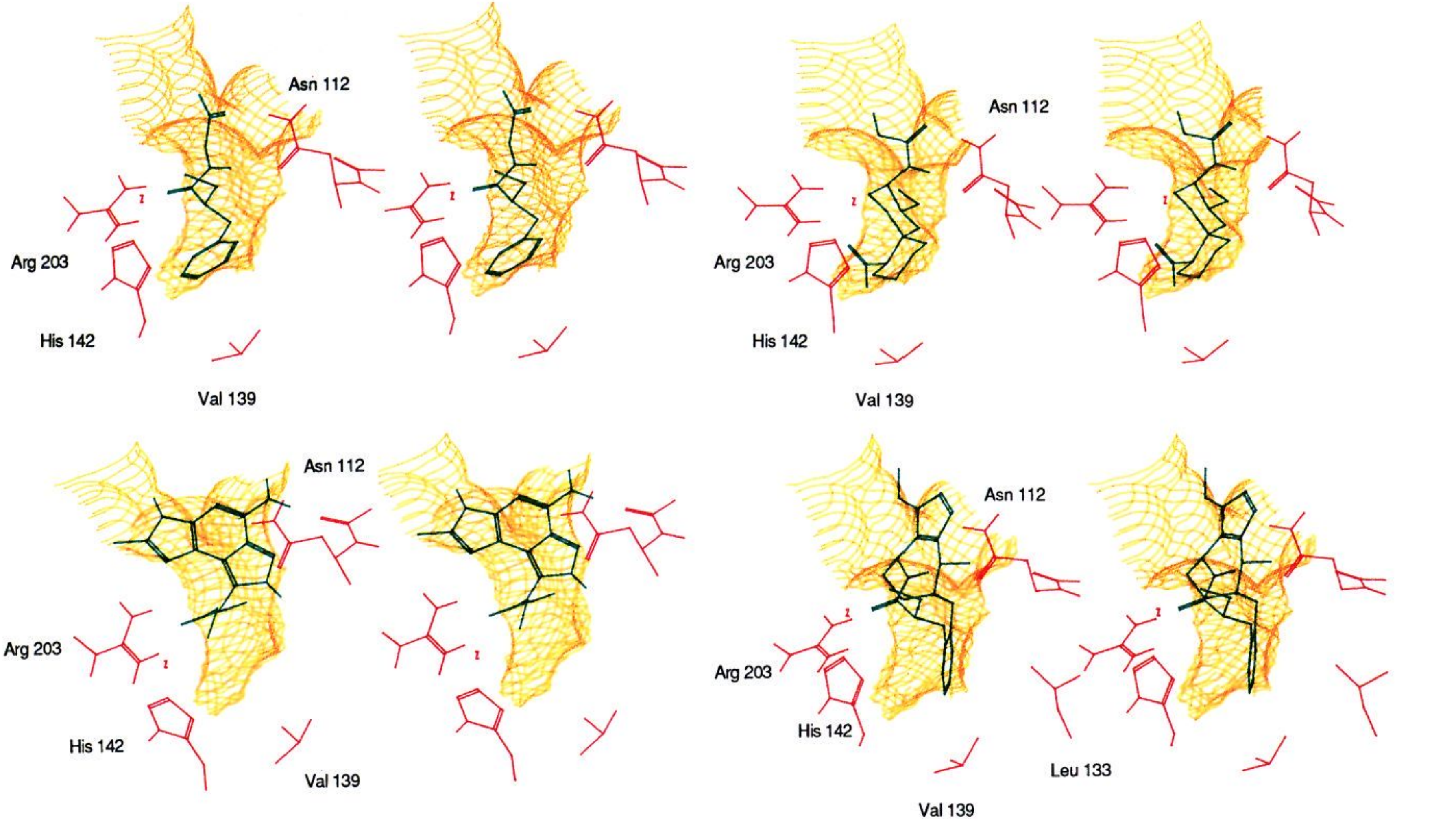


Figure 4. Stereorepresentations of four structures generated in the binding site of thermolysin. The yellow mesh represents the accessible surface of the binding site. Structure 4-1 (top left) is thiorphan, also shown schematically in Figure 6, Part A-1. Schematic diagrams of structures 4-2 (top right), 4-3 (bottom left), and 4-4 (bottom right) and their specific interactions with the enzyme are given in Figure 3, Parts A-2, B-3, and C-2, respectively.

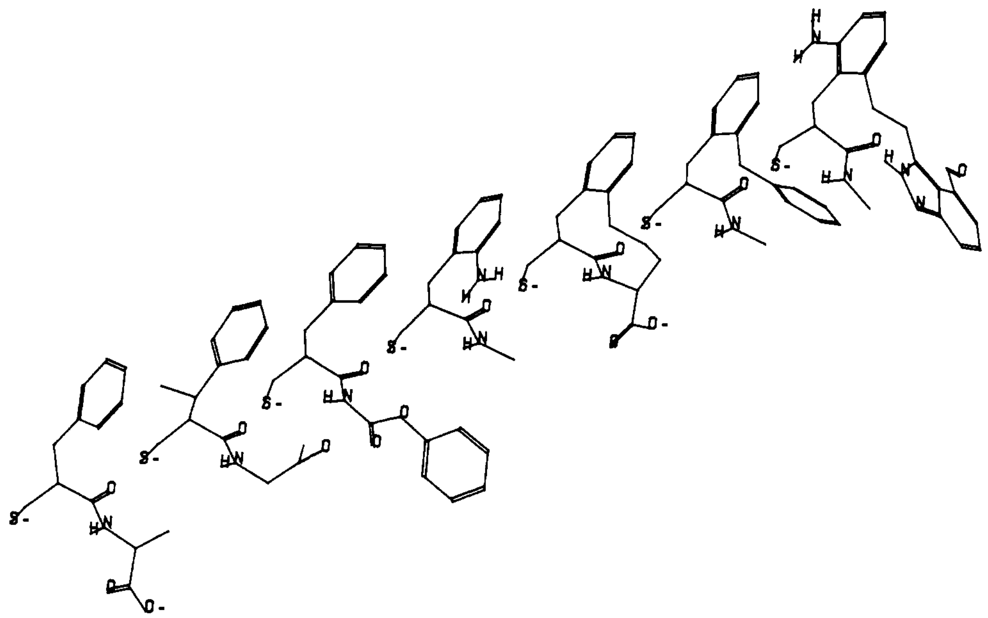


Figure 5. Examples of structures which have a common feature: a benzene ring in P1′ and the amide bond common to substrate-based inhibitors such as thiorphan. The structures are shown in the conformation in which they were generated. Each structure is offset 10 Å on the *x* axis. This makes it easy to look at a series of structures depicted in a single image. These examples show some of the variations found in the remaining parts of the structure. All the molecules are highly complementary to the binding site: all carbonyl, carboxyl, -N=, and NH groups shown form hydrogen bonds, and most of the nonpolar atoms occupy hydrophobic areas.

a specific feature of interest. Figure 5 shows an example where the common feature was a substructure consisting of the benzene ring and the amide bond of thiorphan. Figure 5 shows a highly diverse set of structures contain this fragment.

4. Validation by Comparison to Inhibitors of Known Potency. The ability of GrowMol combined with the evaluation procedure to produce potent inhibitors was assessed by comparing reported thiol inhibitors of thermolysin to those generated by GrowMol. A literature search was conducted which revealed fourteen thiol thermolysin inhibitors. Six of these compounds are α thiols.[33,34] At present, there is no experimental data available describing the binding mode of α thiols to thermolysin. It was suggested that α thiols bind to the zinc of thermolysin in a bidentate fashion.[33] In the present study, structures are grown from a sulfur bound to zinc in a monodentate manner and, therefore, the published α thiols cannot be used for comparison. The (*S*) isomer of retrothiorphan has a K_i of 94 μM and was, therefore, also excluded. To the list of the remaining seven inhibitors, we added an eighth compound, a potent benzofused macrocycle, CGS 26670.[35] Figure 6 shows a comparison between these compounds and similar structures generated by GrowMol.

GrowMol generated structures identical to (*S*)- and (*R*)-thiorphan as well as structures similar to five of the other known inhibitors. The final inhibitor had an experimental potency of 52 K_i, and it is, therefore, not suprising that it was not generated by GrowMol.

Figure 6 shows a comparison of experimental potencies and estimated potencies of the known inhibitors. The calculated potencies are in good agreement with the experimental values. Figure 6 also shows the value of energy minimization in the active site prior to evaluation. For each of the structures given, the estimated potency is improved after this optimization step.

Discussion

The results show that very large numbers of highly diverse molecules can be generated which are complementary to enzyme binding sites. It might be felt that this will offer so much choice that it will be difficult to decide which compounds should be synthesized. One approach might then be to use *de novo* programs to generate a small set of structures[6,9] for consideration for synthesis. We believe, however, to the contrary, that large sets of complementary structures offer an enormous opportunity for the rapid discovery of molecules with the full range of properties required for a successful therapeutic application. Such molecules will usually be required to be orally bioavailable and stable to metabolism in the body. In addition, the molecules must be synthetically accessible. It may also be advantageous to use the large set of structures to select molecules which show a specificity for one binding site versus another. In other situations it may be desirable to select structures which may act as dual inhibitors which can interact with two or more binding sites.

Our approach is, therefore, to generate large data bases of complementary structures and then to develop and apply various computational techniques to find the most suitable structures for a given purpose. In the future, it can be anticipated that it will be possible to generate nearly all the possible complementary structures (see below). When this stage has been reached, the critical issue will be how to select molecules which satisfy all the requirements for a given application. In this paper, we confine our attention to the selection of structures with low conformational strain energy, low estimated potency, and a diversity of binding motifs.

Diversity. The diversity can be considered to arise from two main sources. One of these involves permutation of atoms with similar bonding geometries. For example, an sp^3 carbon can be replaced by an sp^3 nitrogen or an ether oxygen without greatly perturbing the backbone structure. The other source of diversity involves replacement by an atom of different binding geometry, e.g. an sp^2 atom instead of an sp^3 atom. This results in molecules with totally different structural backbones. The combination of these two effects leads to the high degree of diversity found in this study.

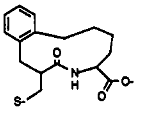
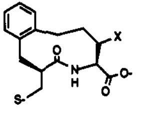
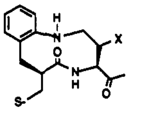
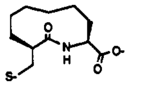
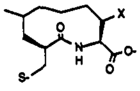
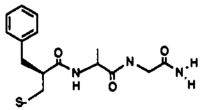
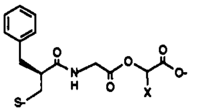
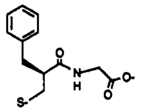
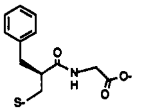
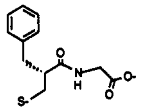
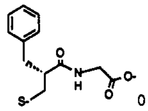
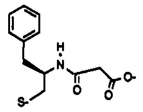
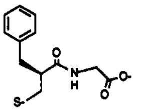
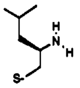
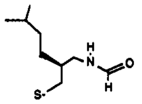
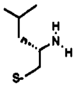
[33] Holmquist, R.; Vallee, B. L. *Proc. Natl. Acad. Sci. U.S.A.* **1979**, *76*, 6216–6220.

[34] Blumberg, S.; Tauber, Z. *Eur. J. Biochem.* **1983**, *136*, 151–154.

[35] Ksander, G.; Bohacek, R. S.; de Jesus, R.; Yuan, A.; Sakane, Y.; Berry, C.; Ghai, R.; Trapani, A. J. Manuscript in preparation.

[36] MacPherson, L. J.; Bayburt, E. K.; Capparelli, M. P.; Bohacek, R. S.; Clarke, F. H.; Ghai, R. D.; Sakane, Y.; Berry, C. J.; Peppard, J. V.; Simke, J. P.; Trapani, A. J. *J. Med. Chem.* **1993**, *36*, 3821–3828.

[37] Pickering, D. S.; Krishna, M. V.; Miller, D. c.; Chan, W. W. *Arch. Biochem. Biophys.* **1985**, *239*, 368–374.

Known Inhibitors	Experimental Potency K_i, µM	Estimated Potency K_i, µM	GROWMOL Structures	Estimated Potency K_i, µM	Estimated Potency K_i, µM Before Energy Minimization In Active Site
	0.019[a]	0.043		0.30	0.78
				0.30	2.0
	3.8[b]	5.4		0.30	2.0
	0.75[c]	0.78		0.12	0.89
	1.8[d]	0.11		0.11	5.3
	3.0[d]	0.77		0.77	97.8
	2.3[d]	0.77		0.11	5.3
	5.3[e]	85		12.3	32.4
	52[f]	224	none		

[a] See ref 35. [b] See ref 36. [c] See ref 13. [d] See ref 22. [e] This compound binds to thermolysin in a noncompetitive manner, and, therefore, cannot be compared with the rest of the potency data.[37] [f] See ref 37.

Figure 6. Comparison between thiol inhibitors of known potency and structures generated by GrowMol. Potency estimates based on complementarity to the binding site are shown for the known inhibitors as well as the GrowMol structures. To show the effect of optimization by energy minimization in the active site, the potency estimates *before* energy minimization are also given.

For thermolysin it was only possible to adequately sample the diversity using structures of limited molecular weight. The results suggest that for the S1′ and S2′ portions of the thermolysin binding site the number of structures may be related to the size of the structure by the approximate relationship

$$\text{number of unique structures} = 2^{(\text{number of atoms})}$$

This implies that for structures of 18 atoms there will be 262 144 unique molecules spatially and chemically compatible to this binding site. Even allowing for reduction of the number of unique structures upon energy minimization and further filtering based on complementarity, the number of possible ligands clearly remains very large. For an inhibitor with 30 atoms (ZF^{P}LA, the very potent thermolysin inhibitor, has 39 atoms—not counting hydrogens bound to carbon atoms),[3] approximately one billion complementary structures can be anticipated.

The large number of unique structures which can be obtained with this method is too large for appraisal by visual inspection. The selection of a greatly reduced set of structures is, therefore, essential. Even after energy minimization and the removal of structures with poor complementarity and high energies, the number of remaining structures was still too high.

Since some of the diversity is due to the fact that structures may occur which differ only slightly from each other, clustering based on similarity can be used to group members of these families together. Examining representatives from each cluster can then provide an overview of the range of significant diversity. The structure with the lowest estimated K_i and the lowest strain energy was chosen to represent the cluster. In this way the optimal binding features available in that family are sampled.

Examination of the structures in this way revealed a number of significantly different binding motifs. Thus it is clear that the diversity does not simply arise from minor variations in one major type of structure but is also due to structures which are almost entirely different from each other.

Finding such highly diverse structures which fit well into the binding site of thermolysin was surprising. Thermolysin is an endopeptidase with a high degree of specificity for peptides with a hydrophobic residue in position P1′, i.e. after the scissile bond. This specificity is due to a deep S1′ binding pocket. The peptide bond connecting the P1′ and P2′ residues is held in a close-fitting tunnel and is aligned by well-formed hydrogen bonds to the carbonyl oxygen and the amide hydrogen. Thus it might be expected that, as a result of the steric and hydrogen bonding requirements of the site, many of the grown structures with good complementarity scores would be substrate mimics. Out of 3937 complementary structures, 2858 were found that did not have an amide bond corresponding to that found in substrate mimics (see Figure 3–1). Many of these are non-peptidic. Preliminary experiments suggest that structures differing radically from the substrate can also be found in other enzymes such as HIV protease and pepsin.

Types of Structures. Significant numbers of conformationally restricted structures were found. These structures, which contained either macrocycles or fused rings, often had high hydrophobic contact scores. Thus conformationally restricted molecules can be constructed that are also highly complementary to the binding site of thermolysin.

As shown in Figure 6, GrowMol also generated structures identical or similar to a number of known thiol thermolysin inhibitors.

The overall shape of the structures is governed by the binding site as represented by the grid box. The grid box is somewhat larger than is actually allowed by the enzyme atoms so that structures which upon further optimization may fit well into the binding site are not missed. Therefore, we expect a fraction of the structures to have strain energies higher than our threshold. The fact that of the 4449 unique, energy-minimized structures 512 of the structures were rejected because of energy indicates that most of the structures, 88.5%, were indeed low-energy conformations and that only 11.5% exceeded the energy threshold, as expected.

Limitations. The results show that GrowMol can generate a large number of highly diverse structures. However, it is anticipated that the structures we found are a subset of a significantly larger set of structures complementary to the binding site. GrowMol in its present form uses rules which limit the number and type of structures which can be generated. Molecules are constructed with fixed bond lengths, bond angles, and dihedral angles corresponding only to the rotational isomeric states of each torsion bond. An additional limitation is the use of relatively few atoms and functional groups.

Refinements of the GrowMol program are underway which will add energetically realistic perturbations to all the geometrical parameters presently used. Since this will allow slightly different positions of all atoms including the root atom, the possibility ofdetecting additional molecules which can only fit the binding site in conformations slightly different from those obtained with the "standard geometry" presently used will be increased. With this refinement and a larger set of atoms and functional groups, it seems likely that computer programs using atom-based combinatorial growth algorithms will be able to generate most of the structures which are complementary to the three-dimensional structure of a host binding site.

In the present study a static model of the enzyme binding site was used. In the case of thermolysin, experimental evidence suggests that the binding site does not change significantly when ligands bind. Where necessary, minor changes in enzyme side chain positions can be readily treated by allowing the side chains to move during the energy minimization step. The box in which the structures are grown can also be adjusted to tolerate the generation of structures which only fit the binding site after structural optimization by energy minimization. Structures which do not fit after minimization can be rejected using the ligand strain energy. Where there is considerable movement of the enzyme binding site, it may advantageous to carry out several GrowMol runs with different binding-site geometries. In this case, also, portions of the binding site can be allowed to move during energy minimization.

The present paper uses complementarity and molecular mechanics energy to evaluate the structures. In the absence of accurate methods for predicting binding constants, we have applied an approximate estimate of potency based on complementarity. This is an empirical method based on a limited series of compounds and may not be accurate when applied to significantly differently structures. With the development of more adequate methods for predicting potency, it should be possible to use *de novo* computer programs to rapidly design novel, highly potent molecules targeted toward a specific binding site.

Conclusions

A combinatorial, *de novo* growth algorithm has proven useful for probing the diversity of potential ligands for an enzyme binding site.

The application of GrowMol to the thermolysin binding site showed that large numbers of different structures with steric and chemical complementarity could be formed. Similar results were obtained in preliminary studies using the HIV protease and pepsin binding sites. While the very large number of structures which can be generated might be considered to be a liability of the method, in fact the ability to reveal such a wide range of diversity makes the method very powerful. Large data bases of complementary structures can be created and then searched to reveal structures having the desired properties. Using chemical complementarity, molecular mechanics energies, and clustering of similar structures, it is possible to identify a set of highly diverse structures with properties believed to lead to good binding.

It would also be possible to apply other criteria to the selection of compounds from the data base of complementary structures. Compounds could be selected based on their similarity to known compounds with good oral bioavailability. An additional important criteria is likey to be synthetic accessibility.

Therefore, we anticipate that *de novo* growth methods will play a key role in the discovery of therapeutic agents based on the structure of the target binding site.

Acknowledgment. We thank our colleagues Drs. Frank Clark, Wayne Guida, and Jeff Watthey for the numerous suggestions and stimulating discussions. The constructive comments on the manuscript of Drs. Wayne Guida and Greg Paris are also gratefully acknowledged. We would also like to thank the reviewers for suggestions concerning validation of the method.