Computational combinatorial chemistry for de novo ligand design: Review and assessment*

Amedeo Caflisch[a,b] and Martin Karplus[a,c]

[a]*Department of Chemistry, Harvard University, Cambridge, MA 02138, U.S.A.*
[b]*Biochemisches Institut der Universität Zürich, CH-8057 Zürich, Switzerland*
[c]*Laboratoire de Chimie Biophysique, Institut Le Bel, Université Louis Pasteur, F-67000 Strasbourg, France*

Summary

Computational combinatorial chemistry divides the ligand design problem into three parts: the search for optimal positions and orientations of functional groups in the binding site, the connection of such optimally placed fragments to form candidate ligands, and the estimation of their binding constants. In this review, approaches to each of these problems are described. The present limitations of methodologies are indicated and efforts to improve them are outlined. Applications to HIV-1 aspartic proteinase, which is a target for the development of AIDS therapeutic agents, and human thrombin, a multifunctional enzyme that has a central role in both haemostasis and thrombosis, are presented. The relation between combinatorial methods for drug discovery on the computer and in the laboratory is addressed.

Introduction

The cloning and sequencing of the human genome promises that an ever increasing number of proteins will become available as potential drug targets in the coming years. X-ray crystal structures [1], nuclear magnetic resonance solution structures [2–5] and homology-modelling predictions [6–9] will provide the information necessary for structure-based design of novel therapeutic agents for the treatment of a variety of diseases. Computer-aided structure-based ligand design is concerned with the prediction of ligands that are expected to bind strongly to key regions of biologically important molecules (e.g., enzymes, macromolecular receptors) of known three-dimensional structure, so as to inhibit or alter their activity. An immense effort has been and continues to be dedicated to developing methods by which drug design (or, more properly, ligand design, since whether or not a ligand will result in a drug involves factors beyond the present concerns) can be made an automatic and rational process. Despite several successful case studies where structure-based ligand design efforts led to compounds which are currently in clinical trials [10], the field is still in its infancy, as is evident from recent reviews [11,12] as well as from

*Inquiries concerning the CHARMM, MCSS, and HOOK programs should be addressed to M. Karplus (e-mail: marci@tammy.harvard.edu). Inquiries concerning the program CONNECT should be addressed to A. Caflisch at Department of Biochemistry, University of Zürich, CH-8057 Zürich, Switzerland.

Perspectives in Drug Discovery and Design, Vol. 3, pp. 51–84
0928-2866/$ 6.00 + 1.00

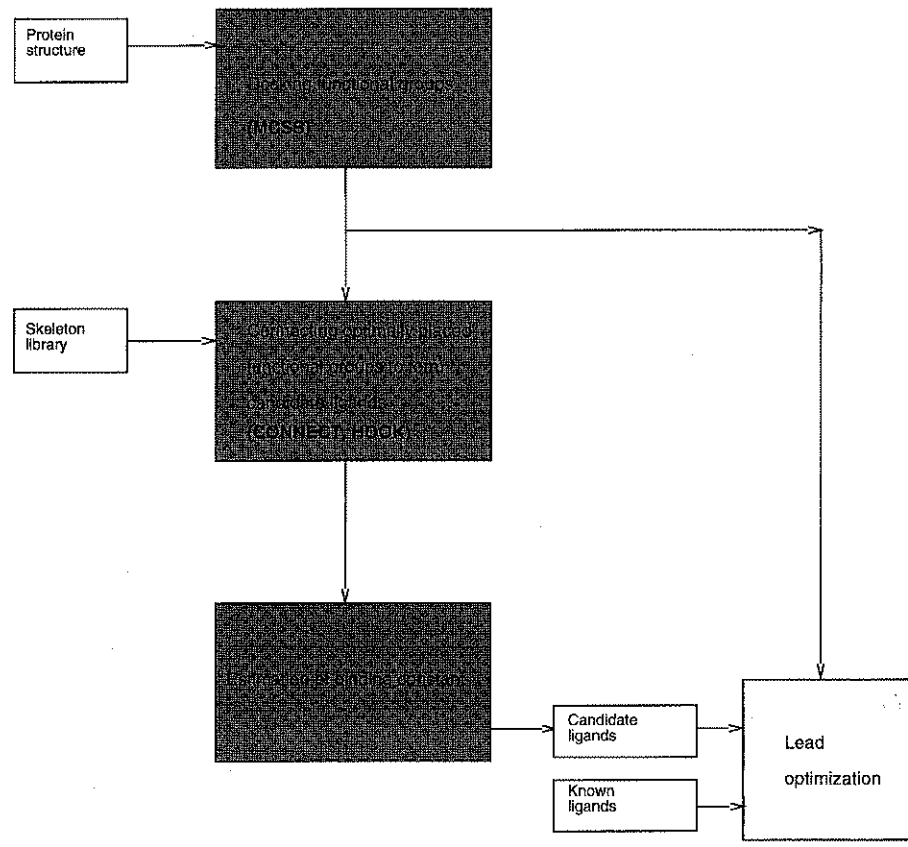


Fig. 1. Schema representing the strategy for computer-aided structure-based ligand design. The three steps of the strategy are depicted in the shaded rectangles; computer programs are in bold.

earlier evaluations [13,14]. Thus, new ideas and methods for approaching the ligand design problem are still needed.

The combinatorial strategy we have chosen for structure-based ligand design consists of three parts (Fig. 1). The first is an efficient method for the search of known binding sites (or, more generally, receptor surfaces) for positions that interact strongly with a range of functional groups. This is necessary for a ligand with high selectivity. To solve this problem, the Multiple Copy Simultaneous Search (MCSS) procedure was developed [15]. The importance of functional groups has been stressed recently in an assessment of binding modes of inhibitors for trypanosomal triosephosphate isomerase [16]. Also, the wealth of structural information available on thrombin and thrombin–ligand complexes indicates that both its active site and the fibrinogen recognition exosite can be divided into a number of pockets, each of which is filled by a preferred functional group in a variety of known inhibitors [17].

Second, given a set of such positions and orientations for functional groups, it is necessary to connect the functional groups to form putative ligands that are candidates for synthesis. Two approaches will be discussed. The program CONNECT was developed

to generate peptide leads from optimal positions of *N*-methylacetamide (NMA) groups and functional groups representing side chains [18]. This method can also be used to construct peptidomimetic compounds, e.g., oligo(N-substituted glycines), from *N*,*N*-dimethylacetamide fragments. HOOK is another approach, which was developed to find molecular skeletons from a three-dimensional database that fit well into the protein binding region and make bonds to functional groups [19].

Third, a method is needed to estimate which of the resulting candidate molecules are likely to have the strongest binding constants and can be synthesized without excessive effort. Evaluating the free energy of binding of the resulting candidates in the third step requires a more sophisticated and time-consuming analysis of the interactions as well as a treatment of solvent and entropic effects. This can be applied only to a limited set of molecules.

A stepwise combinatorial procedure is used because it is much more efficient than doing everything at once. It would take an inordinate amount of time to dock hundreds of thousands of ligands into the binding site and evaluate their energies. By first docking functional groups and then connecting them to form candidate ligands, it is possible to search through a very large number (10^{10} to 10^{15}) of molecules in a relatively short time. When a set of 100 to 1000 or so likely candidates have been selected, it is appropriate to make a more detailed analysis of their binding properties and routes to their synthesis, although the functional groups and linker fragments can be chosen with reference to the ease of synthesis (D. Joseph, P. Deslongchamps and M. Karplus, work in progress).

Several methods have been developed that are related to various aspects of the three-step approach. One is the GRID program [20–23], which is the best known and most widely used method for functional group searches. GRID determines favorable binding sites with an interaction based on an empirical energy function. It uses simple spherical ligands and has emphasized positional information, although recently specific hydrogen-bonding interactions have been included [22,23].

Another related approach is that embodied in the program LUDI [24–26]. It makes use of statistical data from small-molecule crystal structures to determine binding sites of molecular fragments, i.e., discrete positions on the binding site surface that are suitable to form hydrogen bonds and/or to fill hydrophobic sites of the receptor. Alternatively, it uses simple rules or the output of the program GRID to generate the interaction sites. Finally, the fragments fitted in the interaction sites are connected by linker groups. Although LUDI is very fast (i.e., it can be used interactively), the use of purely geometric rules to position fragments in the binding site may miss the optimal positions of functional groups. This is the case for polar groups, whose orientation in the binding site is often a compromise between optimal intermolecular hydrogen bonds and more delocalized electrostatic and van der Waals interactions [27].

The program DOCK [28,29] pioneered the use of databases and geometric criteria to select ligands which best complement the shape of the protein binding site. In its original version, DOCK neglected electrostatic interactions and was unable to handle ligand flexibility. Recently, the DOCK scoring function has been expanded to include a full intermolecular force field [30]. In addition, Leach and Kuntz have supplemented the program with a systematic search algorithm to introduce ligand flexibility [31].

Bartlett and co-workers developed CAVEAT [32], which uses databases of cyclic compounds and functional group vectors in a manner that is similar to what is done in

HOOK. However, CAVEAT searches through a database for molecules that link together specified functional groups without consideration of the fit between the binding region and the molecule. This is done in a postprocessing step.

Recently, the software package LEGO has been developed for de novo ligand design based on the combination of multiple fragment docking, automatic connection by small linker units (one to four atom chains), and searching of 3D databases for complementary molecules [91,94]. The LEGO method has been implemented within the MOLOC molecular modelling system [94], which allows the visualization of the functionality maps and interactive model building of the growing ligands. LEGO does not explicitly take into account solvation effects; thus, it is very efficient and can be run interactively and almost in real time. It is based on a force field which omits all hydrogen atoms and does not require partial charges but uses geometrical criteria for hydrogen bonds and was successfully tested by reproducing the structural aspects of 1589 compounds derived from the Cambridge Structural Database [95].

A different approach for fragment-based de novo ligand design involves the sequential buildup of a candidate ligand molecule. Moon and Howe developed GROW, which utilizes a template set of in vacuo generated amino acid conformations (constructed as *N*-acetyl-*N'*-methylamides) and iteratively pieces them together by amide end group superimposition in the presence of a model of the receptor [33]. Their growth algorithm corresponds to a tree search in which each library template is attached to the seed (or to each actual construct) and the search space is kept under control by pruning according to an energy evaluation, based on van der Waals, Coulombic, strain and desolvation terms. Although the GROW results depend strongly on both the seed position and the choice of templates, the method has been validated by reproducing the known binding orientations of inhibitors of both HIV-1 aspartic proteinase (HIV-1 PR) and rhizopuspepsin. A limitation was demonstrated in the prediction made for MVT-101 (*N*-acetyl-Thr-Ile-Nle-Ψ[CH_2-NH]-Nle-Gln-Arg-amide) in HIV-1 PR [34]. Only the minor orientation of MVT-101 was obtained, because the Ile-Nle peptide bond from the published crystal structure was employed as a seed position.

Rotstein and Murcko developed GroupBuild, a fragment-by-fragment ligand generator [35]. GroupBuild uses a library of common organic templates and a force field description of the nonbonding interactions between the ligand and the enzyme to build putative ligands that have chemically reasonable structures as well as steric and electrostatic properties which are complementary to the enzyme. To partially account for the hydrophobic effect, the difference in solvent-accessible surface area upon binding is calculated for heavy nonpolar atoms. Although no attempt is made to estimate the electrostatic contribution to the free energy of desolvation, GroupBuild was able to reproduce known binding motifs found in a variety of inhibitors for FKBP-12, human carbonic anhydrase, and HIV-1 PR.

A program similar to GroupBuild was recently described by Bohacek and McMartin [36]. It uses a Boltzmann weighting factor to bias the probability of selection of new atoms to be added to the growing chain towards those with a high complementarity score, based on rewarding carbons in hydrophobic regions or hydrogen-bonding atoms near appropriate partners, and penalizing mismatches between atom type and binding region.

Sequential fragment build-up procedures have several weaknesses. The main one is that they do not use information about critical binding regions and often fail to connect

distant binding pockets. Furthermore, the suggested compounds and their orientation in the binding site are affected by the choice and position of the seed fragment. To solve the latter problem, Moon and Howe [33] developed a preprocessing algorithm similar in spirit to MCSS. To keep the number of molecules within bounds, pruning is done based on energy calculations that could eliminate possible ligand candidates. So far molecules have been constructed with fixed bond lengths and bond angles, and dihedral angles corresponding only to the rotational isomeric states of each bond.

In the remainder of this review, the methodologies we have developed for structure-based ligand design are considered in more detail. Advantages and limitations of the approaches are discussed and ongoing developments to deal with such problems are described. Applications to HIV-1 PR and human thrombin are presented. It is shown that the methods can be used for lead optimization, as well as for de novo design. We conclude by indicating the advances in ligand design that may be expected in the near future. Methodologies for searching 3D databases to test pharmacophore hypotheses and select compounds for screening are not discussed here; they have been reviewed in Ref. 37.

Probing the binding site

Multiple Copy Simultaneous Search

The multiple copy simultaneous search (MCSS) method determines energetically favorable positions and orientations (local energy minima) of functional groups on the surface of a protein of known three-dimensional structure [15,18]. Functional groups are small chemical fragments commonly found as substituents of larger organic molecules. To investigate both the hydrophilic and hydrophobic character of the binding site, charged (e.g., acetate, methylammonium, methylguanidinium), polar (e.g., methanol, *N*-methylacetamide), aromatic (e.g., methylbenzene, naphthalene), and aliphatic (e.g., propane, isobutane, methylcyclohexane) groups are used. Additional functional groups can be introduced by the user.

The method is fully automated, although certain critical parameters can be adjusted to optimize it for specific applications. Several thousand replicas of a given group are randomly distributed inside a sphere whose radius is chosen sufficiently large to cover the entire region of interest. This can be a known binding site or the entire protein, if one wants to explore alternative binding pockets. The initial random distribution can also be performed inside a parallelepiped if the region of interest is elongated in one or two directions. A minimal distance can be given as input to avoid bad contacts between functional group atoms and protein atoms for the initial distribution. More sophisticated and less random initial seeding (use of intuitive chemical rules, layer distribution on a Connolly surface) are currently under investigation (A. Caflisch and C. Ehrhardt, work in progress). Preliminary results indicate that for monofunctional groups (e.g., acetate, methylammonium) and polyfunctional groups (e.g., *N*-methylacetamide) the more accurate seeding results in savings of CPU time of up to about 50% and 30%, respectively, to obtain the same minima as with a random initial distribution.

Subsets of between 500 and 3000 randomly distributed replicas of the same group are simultaneously minimized in the force field of the protein. A classical version of the time-dependent Hartree (TDH) approximation [38,39] is used to divide the system into two

parts, protein and functional group replicas, each of which feels the average field of the other. The interactions between the group replicas are omitted, i.e., replica i does not interact with replica j, for each i and j in the subset. In applications to the sialic acid binding site of the influenza coat protein hemagglutinin [15], HIV-1 PR [18], and thrombin, the protein was kept fixed. Hence, the TDH approximation is exact. The force on each replica consists of its internal forces and those due to the protein, which has a unique conformation and, therefore, generates a unique field. Protein flexibility is being taken into account in studies that use minimization and quenched molecular dynamics for a system consisting of multiple copies of protein side chains, as well as functional groups (C. Stultz and M. Karplus, work in progress).

The minimization begins with 500 iterations of the steepest descent algorithm, which has the important property that it tends to reach the nearest local minimum, i.e., the optimal position and orientation characterized by the smallest displacement from the starting point. It also provides a better performance than higher order algorithms for very poor starting conformations where the gradient is large. The conjugate gradient algorithm is then applied to optimize the functional group positions in the local minima [40,41]. The positions are compared every 1000 steps to eliminate replicas converging toward a common minimum. The criteria used to characterize a common minimum are an rms deviation of 0.2 Å or less between two replicas and a decreasing rms deviation in the final 200 steps. A convergence criterion of 0.001 kcal/mol Å for terminating the minimization is utilized. For a complete minimization, between 4×10^3 and 15×10^3 steps are usually required, depending on the size of the functional group. Further details concerning the methodology are given in Refs. 15 and 18.

Application to HIV-1 PR

Retroviral proteinases, which are members of the aspartic proteinase family, specifically process high-molecular-weight viral polyproteins into individual structural proteins and enzymes [42]. Mutation of the active site aspartate to asparagine in HIV-1 PR prevents processing of polypeptides [43], so that immature, noninfective virions result. Hence, specific inhibitors of HIV-1 PR would be candidates to serve as AIDS therapeutics.

As a test of the MCSS methodology, the functional group minima were compared with those corresponding to the inhibitor MVT-101. This is the reduced analog of a hexapeptide substrate, with the sulfur of the methionine side chain replaced by a methylene group in norleucine for synthetic simplification; it has an inhibition constant (K_i) of 780 nM [34]. A detailed description of the functionality maps in the binding site of the HIV-1 PR structure has been presented previously [18]. Both the native conformation [44] and the structure derived from the complex with the inhibitor MVT-101 [34] were used. When the MCSS minima were compared with the positions of corresponding moieties in the MVT-101 inhibitor [18], all of the backbone peptide groups of the inhibitor corresponded to one or more NMA minima in the MCSS functionality map (rms deviations of 2.1 Å or less). Also, all inhibitor side-chain positions were found by at least one appropriate functional group with an rms deviation of 2.4 Å or less. It is likely that some of the deviations are due to the fact that the binding of the complete inhibitor is not ideal in terms of individual functional group interactions; e.g., it was found that when the peptide backbone was divided into NMA groups that were minimized separately, they converged towards the

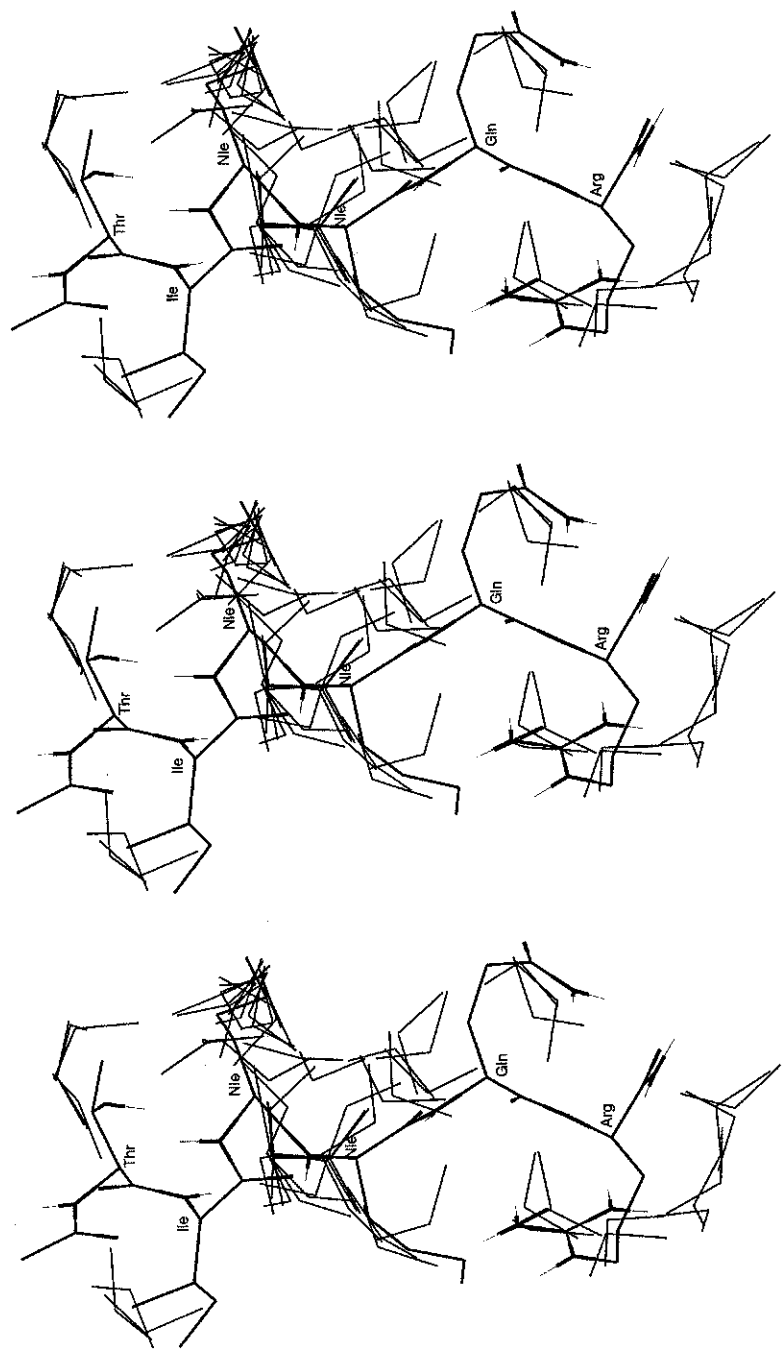
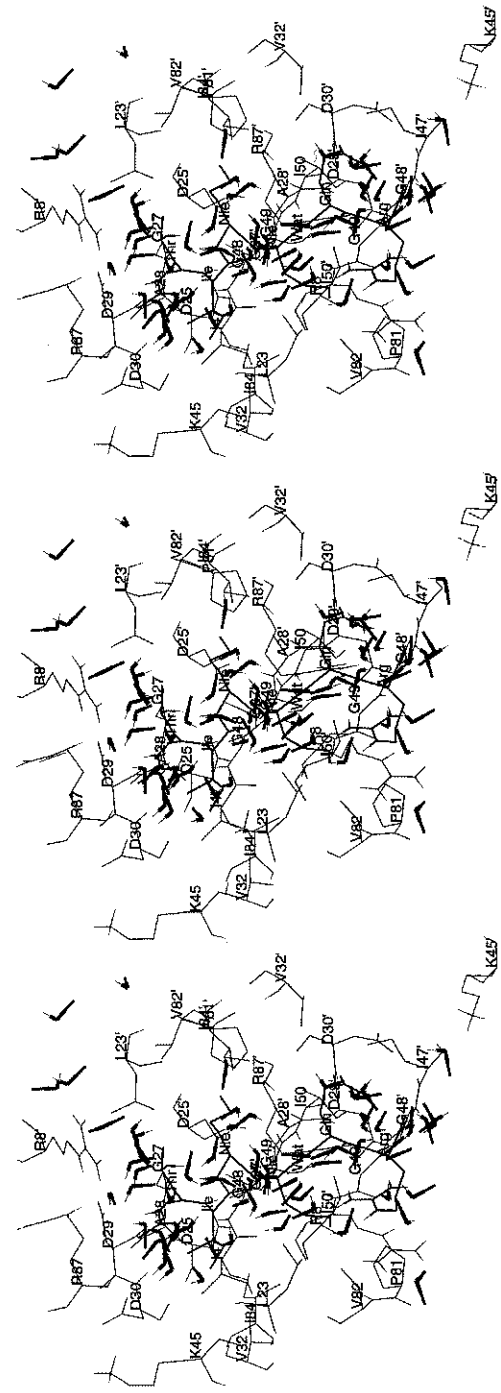


Fig. 2. Stereoview of the 79 MCSS-minimized positions of propane in the HIV-1 proteinase binding site, superimposed onto the structure of bound MVT-101 inhibitor. In all stereo figures, the view is cross-eyed for the left pair and wall-eyed for the right pair.



a

Fig. 3. (a) The 72 MCSS minima of methanol (thick lines indicate heavy atoms and thin lines indicate polar hydrogens) in the HIV-1 proteinase binding site (thin lines), shown with the MVT-101 inhibitor (medium lines).

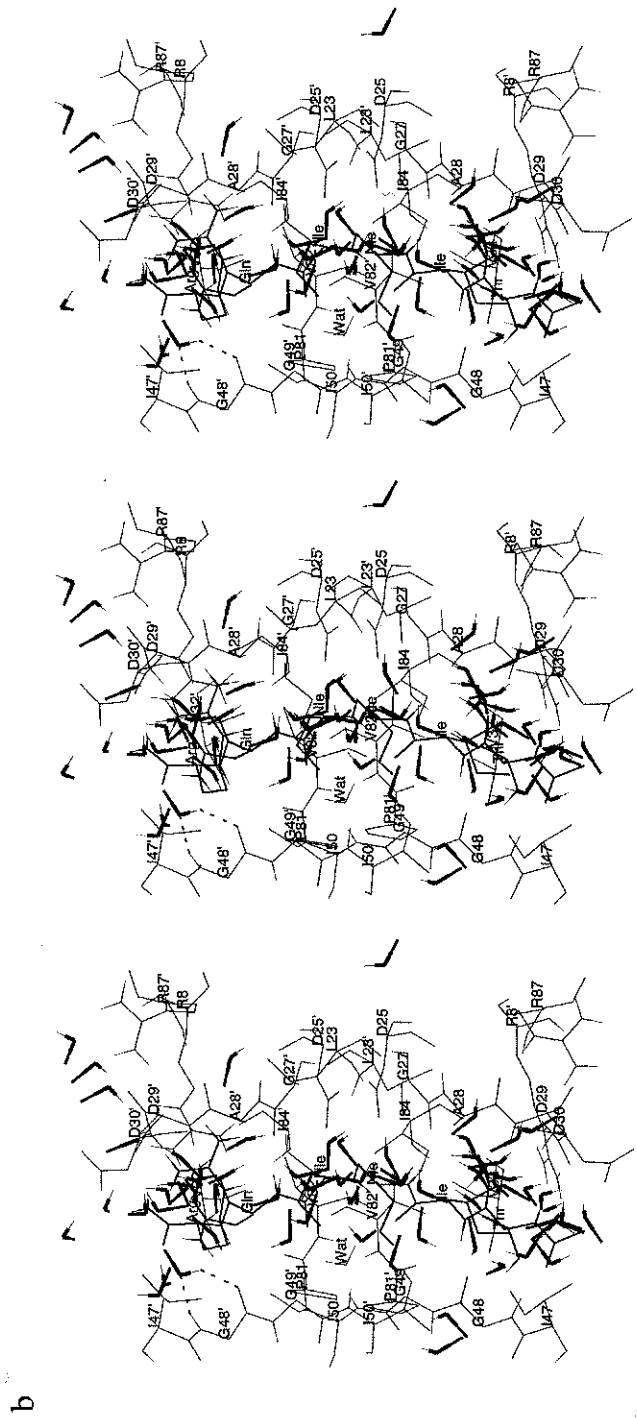


Fig. 3. (b) Same as in (a) after rotation by 120° around x, −60° around y, and 95° around z; the figure was rotated to better visualize the methanol minima discussed in the text. The hydrogen bonds between an interesting methanol minimum (see text) and the proteinase are indicated with dashed lines. The hydrogen bond criterion used was a donor–acceptor distance smaller than 2.6 Å.

nearest (or next-nearest) minima found by MCSS. The same occurred for the MVT-101 side chains. Moreover, MVT-101 is a relatively poor inhibitor compared with known subnanomolar inhibitors of HIV-1 PR [45,46].

Besides being a very useful first step for de novo design (as discussed in the following sections), the MCSS results can be used for lead optimization. Several aspects of the HIV-1 PR functionality maps suggest modifications of MVT-101 for the improvement of binding potency. The 79 propane minima obtained from MCSS in the HIV-1 PR structure from the complex with MVT-101 are shown in Fig. 2. The six propane minima characterized by the best interaction energy with the proteinase are located at the S2 and S2' subsites, and the two largest clusters of propane (Fig. 2) and isobutane (results not shown) functionalities are at the S1 and S1' subsites. This is in agreement with the requirement of hydrophobic residues at these subpockets in most of the sequences of HIV-1 PR substrates [47] and inhibitors [48]. From Fig. 2 it is clear that the S1 and S1' subsites can accommodate a more bulky hydrophobic side chain than the butyl group of norleucine. This suggests that modification of the MVT-101 norleucine residues (at P1 and P1') into cyclohexylalanine should result in tighter binding. Propane and methylcyclohexane have similar free energies of desolvation (transfer from dilute aqueous solution to the gas phase), i.e., the value for propane is 0.24 kcal/mol more favorable than that for methylcyclohexane [49]. Consequently, improved van der Waals / hydrophobic interactions with HIV-1 PR should lead to better binding.

Figures 3a and b show that methanol minima are distributed over a large part of the binding site and also at both open ends. They participate in interactions with most of the charged and polar side chains, as well as with many proteinase main-chain NH and CO groups. There are several minima for methanol that interact strongly with Asp^{25} and $Asp^{25'}$. This is in agreement with the interactions between the hydroxyl functionality (at P1–P1') and the catalytic aspartates present in most of the known HIV-1 PR–inhibitor complexes [48,50]. One methanol minimum is involved in hydrogen bonds to the main-chain NH and CO of residue 48' (see the upper left part of Fig. 3a). It could be connected to the C-terminus of MVT-101 by the replacement of the C-terminal -C=O moiety with a $-C-CH_2-OH$ group. Since the additional functional group interacts mainly with the main chain of the proteinase, missense mutations are likely to have a small effect on the additional potency of the modified ligand.

The 20 methylammonium minima can be grouped into eight clusters, the largest of which contains five minima in the vicinity of the catalytic residues Asp^{25} and $Asp^{25'}$, see Fig. 4. Although the desolvation of a charged group always involves a significant free energy cost, it was suggested that introduction of an amino group in the P1–P1' position (e.g., substitution of the main-chain methylene by a $CH_2NH_3^+$) could result in a tighter binding ligand [18]. This is consistent with a report of C_2 symmetrical penicillin-derived HIV-1 PR inhibitors having two secondary amino groups between P1 and P1' [51] and with the recent discovery of a series of aminodiol inhibitors [52].

Possible improvements in MCSS

The present version of MCSS does not take into account the effects of the solvent, i.e., all protein–ligand interactions are calculated with a vacuum potential [41]. This choice is based on the principle that fast methods are necessary to perform effective searches of the

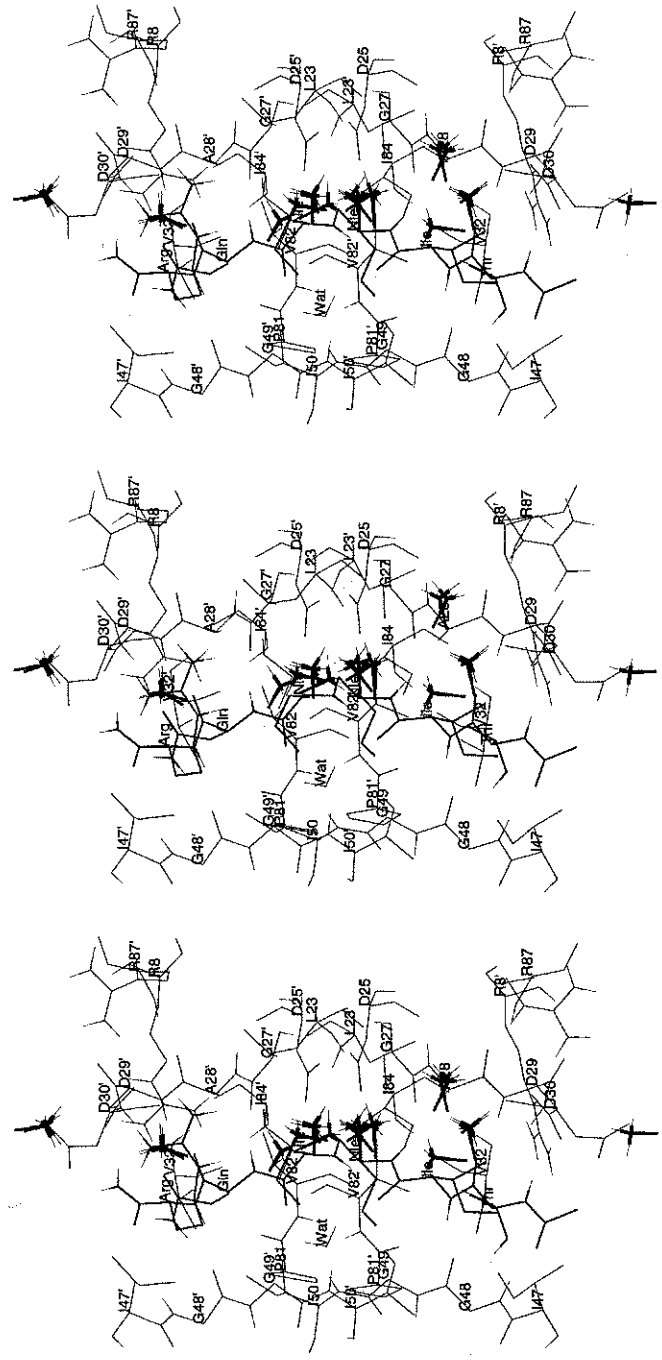


Fig. 4. Same as in Fig. 3b for the 20 MCSS minima of methylammonium.

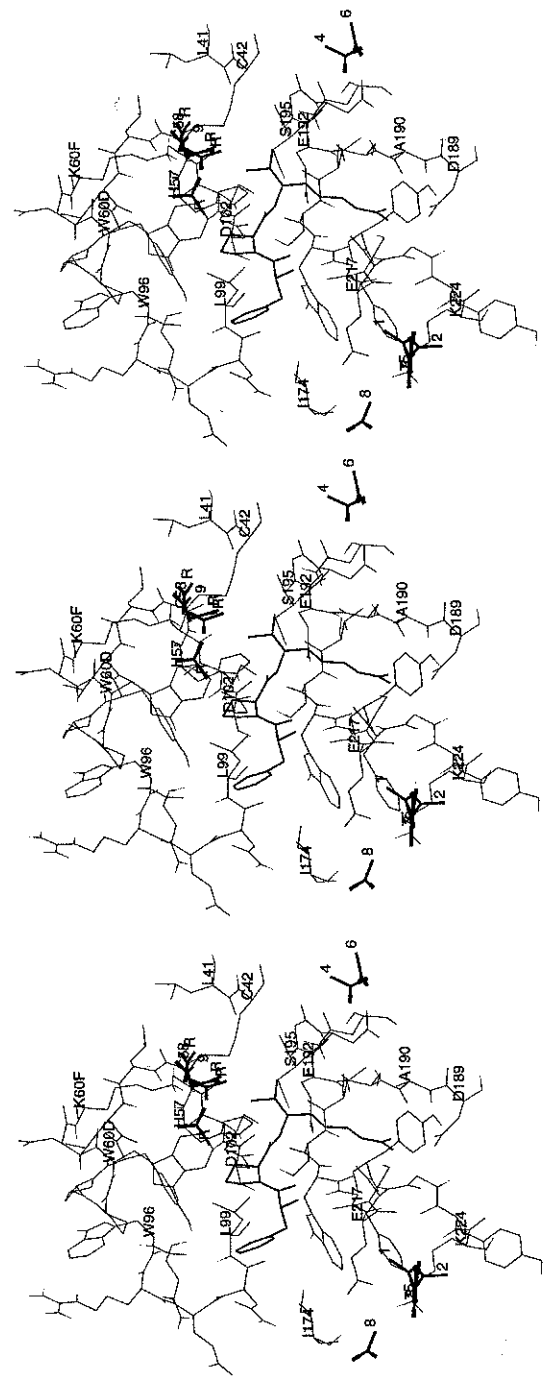


Fig. 5. Stereoview of the nine MCSS minima of acetate obtained with a constant dielectric (labeled with digits ranging from 1 to 9) and the four MCSS minima of acetate generated with a distance-dependent dielectric at S1′ (labeled with R) in the human thrombin binding site (thin lines). The PPACK inhibitor is also shown (medium lines), although it was removed during the MCSS procedure. Some C^{α} atoms of the binding site residues are labeled.

binding site and that good candidate ligands subsequently can be ranked in terms of their binding free energy (vide infra). A possible difficulty with this approach is that binding sites may be missed or mispositioned due to the lack of a solvation correction. Such effects are likely to be particularly important for relatively flat, solvent-exposed binding sites, such as that in SH3 molecules (R. Tan and M. Karplus, unpublished results). Even for protected binding sites, minimized positions of nonpolar fragments may be found in hydrophilic pockets because of the lack of an energy penalty for protein desolvation. A representative example is the cluster of propane minima overlapping the guanidinium group of MVT-101 (Fig. 2). These are close to the Leu^{23} side chain of HIV-1 PR, but partially desolvate the carboxyl group of $Asp^{29'}$.

In an MCSS study of thrombin (vide infra), it was observed that acetate minima tend to cluster in the vicinity of lysine and arginine side chains on the thrombin surface (Fig. 5). There is only one minimum in the binding site; it accepts a hydrogen bond from the primary ammonium of Lys^{60F}. As a simple test of the importance of electrostatic shielding, a distance-dependent dielectric function [53] was introduced instead of the unit dielectric constant in the vacuum potential. The overall shape of the map was found to be similar, but there are three more acetate minima interacting with the Lys^{60F} side chain. In the constant dielectric calculation, the strong Coulombic interaction yields a smoother configurational space than the one with the distance-dependent dielectric function. An alternative to the use of a dielectric function is to reduce exposed charges so as to mimic the effect of solvent shielding. This approach has been used successfully in docking studies and in an analysis of the reaction mechanism of prolyl isomerization by the immunophilin FKBP [54; A. Caflisch et al., unpublished results].

To account for the solvent contributions during the MCSS minimization of the functional group replicas, the CHARMm energy can be used, supplemented by a simplified continuum methodology. This allows one to calculate analytically the screened electrostatic interaction energy and the electrostatic self energy of a molecular system in solution by integrating the energy density of the electrostatic field [55,56].

An approach that takes into account specific water molecules between the ligand and the protein can be introduced into the MCSS procedure. A functional group consisting of an organic fragment and one or more water molecules is defined (A. Caflisch and M. Karplus, unpublished results). A well-characterized example of this method is the water bridging the two HIV-1 PR flaps and the CO groups at the inhibitor positions P1 and P2'. Such an approach does not include the effect of bulk solvent.

An important attribute of the MCSS methodology that is just beginning to be exploited is its utility for including flexibility in the binding surface [15]. This contrasts with other search methods (e.g., GRID). An effective way of doing this is to follow a standard MCSS run for a given functional group with a minimization or quenched dynamics simulation in which the side chains and parts of the main chain involved in the binding are replicated. In this way, optimum functional group positions can be obtained without excessive use of computational resources (C. Stultz and M. Karplus, work in progress).

Connecting optimally placed fragments

Two approaches to forming molecules from functional groups are discussed in this

section. One of these links together the functional groups themselves and the other uses compounds from a database to link them.

Design of peptides and nonpeptidic peptidomimetics

Although peptidic compounds are easily metabolized and often show poor pharmacological profiles, they offer several advantages as lead compounds with respect to nonpeptidic molecules: (i) they are easier to synthesize than most organic molecules; (ii) they are sufficiently flexible to fit even in a complex binding site (though this is a two-edged sword, since more rigid compounds undergo a smaller loss of conformational entropy upon binding); and (iii) their energetics can be analyzed by well-parametrized molecular mechanics force fields.

The MCSS positions and orientations for selected functional groups in the binding site can be connected directly to build candidate peptide ligands. A computer program (CONNECT) was developed to form the backbone from *N*-methylacetamide minima and to generate the side chains by attaching the various functional groups to the main chain [18]. Since no linker pieces are required and only MCSS minima are used, all connected fragments have optimal interactions with the protein. MCSS minima of any side-chain type are evaluated for attachment to each C^{α} atom of the NMA backbones. In addition, if an active peptide ligand sequence is known, but its bound conformation is not, CONNECT can be used as a docking algorithm to suggest one or more binding conformations for the peptide.

In constructing the peptide ligands, the functional groups are kept fixed in their respective minimized positions and a simple pseudoenergy function, $E_{ps} = E_b + E_{nb}$, is used to evaluate all possible ways of connecting the groups. Here E_b represents a 'bonding' interaction and is used to determine whether two groups have relative positions that permit them to be joined together; E_{nb} is a nonbonded interaction that eliminates interacting groups with bad steric contacts. Both pseudoenergy terms are quadratic and positive definite. The use of fixed functional group positions enables the rapid construction of plausible ligands. However, because the groups are fixed, a rather permissive pseudoenergy criterion is used in joining them together. Subsequent minimization serves to regularize the internal structure of the relatively small number of chosen ligands with satisfactory bonding geometry.

Given the pseudoenergy function, the construction of peptides is performed in four steps: main-chain generation, clustering of backbone structures, side-chain attachment, and final minimization. Since the function E_{ps} contains only positive contributions, it is possible to discard a partially built structure if its pseudoenergy E_{ps} is larger than a cutoff value. The fixed functional group positions, the simple energy function, and the use of a cutoff for eliminating incomplete ligands makes the calculation very fast. An algorithm of the branch-and-bound [57] type is used to generate main chains. Increased efficiency relative to an exhaustive search is achieved by pruning the search tree, whose interior nodes represent partial solutions, i.e., n (peptides) of an m (peptide) candidate ligand (where $n < m$). A branch is eliminated when it can be determined with certainty that its elongation will yield a peptide with a pseudoenergy larger than the cutoff value. This pruning criterion is guaranteed to find all optimal solutions, because it is based on evaluation of a pseudoenergy function that is always less than or equal to the true pseudoenergy; it consists of a subset of the terms in the complete pseudoenergy, all of which are

positive definite. It is feasible, for example, to evaluate on a single processor of an SGI 340GTX in less than 20 s all possible ways of building terminally blocked hexapeptide backbones from the 83 NMA minimized positions obtained by running MCSS in the HIV-1 PR binding site [18]. This would be an impossible task if all $83!/(83-7)!$ (about 2×10^{13}) complete hexapeptide main chains would have to be evaluated. After possible main chains have been constructed, a clustering procedure based on their rms deviations with respect to each other can be performed to reduce the number of putative peptide ligands for further study. Details of the clustering algorithm can be found in Ref. 18.

Side chains for the selected set of backbone representatives are constructed in a manner similar to that used for the main chain. For each backbone C^{α}, the MCSS-minimized position with the lowest value of the pseudoenergy after attachment is selected. To prioritize fragments for side-chain selection, several methods to estimate the free energy of desolvation of different functional groups are currently under investigation (E. Evensen and M. Karplus, work in progress). For deeply buried MCSS minima, the fastest approach is to retrieve from a look-up table the hydration free energies of the standard amino acid side chains [58] and a collection of non-ionic monofunctional and bifunctional compounds [49]. For partially exposed fragments one could multiply the ratio between their buried and total surface with the value retrieved from the look-up table.

The resulting peptides are optimized in the field of the protein by a conventional minimization algorithm (e.g., conjugate gradient) or by Monte Carlo Docking (MCD), a stochastic optimization scheme which combines the advantages of the Metropolis Monte Carlo method in global optimization and that of the conjugate gradient algorithm in local minimization [18,59]. Figure 6 shows seven NMA minima in the HIV-1 PR binding site which were selected by the CONNECT program as a terminally blocked hexaglycine ligand (before side-chain attachment), superimposed on the minimized structure. Most of the hydrogen bonds with the proteinase binding site are preserved upon minimization and additional ones are formed, e.g., from the NH of the third NMA minimum (the ordinal number refers to the order in the hexaglycine sequence and not to a ranking based on interaction energy) to the CO of Gly^{27} and from the NH of the seventh NMA minimum to the side chain of $Asp^{29'}$ (Fig. 6). Only one hydrogen bond is lost, from the NH of $Gly^{48'}$ to the CO of the seventh NMA minimum.

With the MCSS/CONNECT/MCD methodologies, the published structure of MVT-101 in HIV-1 PR was reproduced and the end-to-end flipped orientation of MVT-101 was predicted to be more stable [18]. This result was a posteriori found to be in agreement with recent high-resolution crystallographic data (M. Miller and A. Wlodawer, private communication).

The approach described in this section is also appropriate for the de novo design of polymeric ligands consisting of nonpeptidic peptidomimetic units, or for any other synthetic method that is based on linking together a relatively small number of different functional group units. Thus, it could be used to complement and analyze results from combinatorial libraries. The use of peptidomimetic units should lead to putative ligands with better metabolic stability and enhanced pharmacokinetic profiles. A set of N-substituted glycine units has been created, each bearing a nitrogen substituent similar to those of the natural α-amino acid side chains [60]. The polymerization of these monomers results in a class of compounds which was termed 'peptoids'. Recently, certain biological

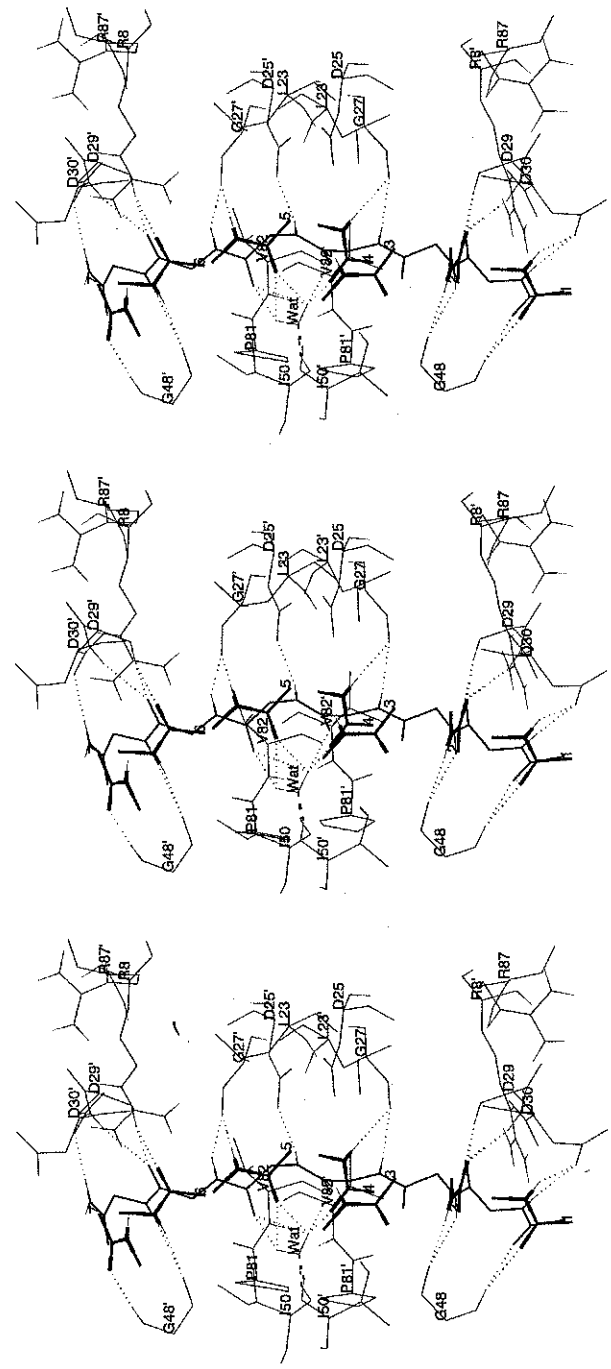


Fig. 6. A sequence of seven NMA minima (thick lines) selected by the program CONNECT in the complexed proteinase (thin lines). Medium lines indicate a minimized terminally blocked hexaglycine. Hydrogen bonds are dotted.

activities have been established for specific peptoid sequences [61]. MCSS can be used to probe the binding site of a given protein target for *N*,*N*-dimethylacetamide as a peptoid main-chain representative, and a variety of functional groups for the side chains. The resulting optimal positions and orientations can be used by CONNECT to generate peptoid ligands. Due to the efficient assembly of diverse peptoid libraries from readily available starting materials [62] and automated synthesis technology [63], the number of potentially available compounds is greater than the number that can be efficiently and accurately screened in solution-phase competition assays [61]. Hence, one of the goals of computer-aided peptoid ligand design is to generate suggestions which can be used for planning experiments, e.g., to limit the number of members in each monomer library to focus it towards functional groups that are predicted to bind well by theoretical methods. As an example, if many of the computer-designed peptoid ligands contain a hydrogen bond-accepting side chain at the N-terminal monomer, one could exclude all aliphatic and basic residues from the library used in the first synthesis step.

An example of the design of nonpeptidic ligands: Thrombin

The program HOOK [19] has been developed to connect a suitably chosen set of functional groups (MCSS minima and/or fragments from known inhibitors) and generate molecules that satisfy the chemical and steric characteristics of a macromolecular binding site. HOOK uses as input the 3D structure of a protein binding site, and a collection of functional group sites that reflect how and where putative ligands could be positioned to make favorable interactions with the protein. These functional groups are linked by searching through a precomputed database consisting of molecular skeletons. Ideal skeletons are relatively rigid hydrocarbon molecules with CH bonds that can be replaced by bonds to functional groups. In this way novel molecules are suggested that contain several functional groups in favorable positions. The linkage is introduced by fusing specified bonds in the skeleton (called 'hooks') with free CH_3-R bonds in two or more functional groups. The resulting molecule has its skeleton positioned in the binding site and the fit between a given molecule and the binding site is estimated by computing an 'overlap score' based on a simplified form for the van der Waals interactions. The program is designed in such a way that a large number of functional group sites and skeletons can be searched very rapidly. HOOK was first applied to construct ligands in the binding site of the hemagglutinin molecule of the influenza A virus and in the binding region of chloramphenicol acetyltransferase [19].

Recently, we have applied HOOK to the active site of human thrombin. In preparation for the use of HOOK, MCSS was used for the active site with the structure from the PPACK complex [17,64]. The active site of thrombin has both hydrophobic and polar character (Fig. 5). Its precleavage sites (S1–S3 subpockets) have been targeted for the development of small-molecular-weight noncovalent inhibitors [65,66]. The following MCSS groups were used: propane, methylcyclohexane, methylbenzene, methanol, *N*-methylacetamide, methylguanidinium and acetate (Table 1). All of these, except methylbenzene and methylcyclohexane, have been employed previously in a survey of the thrombin binding site and in a comparison of the functional group sites with the interactions of known inhibitors (P. Grootenhuis and M. Karplus, unpublished results).

Methylcyclohexane and methylbenzene minima are distributed over the entire binding

TABLE 1
MINIMA FOUND BY MCSS FOR THE THROMBIN BINDING SITE

Group	Lowest ligand strain	Highest ligand strain	Lowest energy[a]	Highest energy[a]	Cutoff energy[b]	No. accepted	No. discarded
Propane	0.0	0.1	−7.9	2.8	2.0	96	1
Methylcyclohexane	0.0	7.9	−10.6	10.0	1.7	329	11
Methylbenzene	0.0	2.0	−14.8	−3.0	−0.9	109	0
Methanol	0.0	1.1	−29.6	−1.5	−5.1	84	7
N-methylacetamide	−1.4[c]	7.9	−41.5	12.8	−9.7	99	23
Methylguanidinium	0.1	5.8	−112.9	−18.1	−37.5	84	36
Acetate	0.0	0.5	−86.3	2.2	−46.5	9	11
Acetate (R dielectric)	0.0	0.9	−86.1[d]	29.5	−46.5	12	38

All energy values are in kcal/mol. The energy of the isolated fragment minimized in vacuo is: propane, 0.0 kcal/mol; methylcyclohexane, −0.1 kcal/mol; methylbenzene, −0.2 kcal/mol; methanol, 0.0 kcal/mol; *N*-methylacetamide, −3.5 kcal/mol; methylguanidinium, 17.4 kcal/mol; acetate, 0.0 kcal/mol.

[a] Sum of ligand strain and interaction energy with the thrombin active site.

[b] For the nonionic compounds the energy cutoff corresponds to the free energy of desolvation, i.e., the free energy of transfer from dilute aqueous solution to the gas phase [49]. For methylguanidinium and acetate the energy cutoff corresponds to one-half the solvation enthalpy of the functional group [92]; the values were derived from Ref. 93.

[c] *cis*-NMA has lower internal energy (because of a more favorable Coulombic and no dihedral penalty).

[d] A distance-dependent dielectric was used during minimization, but the final energy (after minimization) was evaluated with $\varepsilon = 1$.

site from S3 to S2'. The intermolecular van der Waals energy of the 20 best methylcyclohexane minima ranges from −10.58 to −8.79 kcal/mol. Most of these minima occupy the S2 pocket and the region between the S2 and S1 pockets. To facilitate the analysis, they were clustered with a 2.6 Å rms criterion and the lowest energy minimum within each of the resulting nine clusters was chosen as the cluster representative. Figure 7 shows the nine representatives. The distribution of the 20 best methylcyclohexane minima looks similar to that for propane, except for the lack of methylcyclohexane minima at S3. The minimum overlapping the alkyl part of the arginine side chain in PPACK suggests that modification of arginine into cyclohexylamidine could result in good binding. The intermolecular van der Waals energy of the 20 best methylbenzene minima ranges from −14.84 to −9.48 kcal/mol, similar to that for methylcyclohexane. The 10 representatives are shown in Fig. 8a, along with minimum 24 which occupies the S3 subsite and has a van der Waals energy of −9.32 kcal/mol. The four lowest energy minima are in the S1 pocket; they overlap the alkyl part of the arginine side chain of PPACK, as in the propane and methylcyclohexane functionality map. Minimum 1 occupies the same region of the specificity pocket as the benzene ring of benzamidine in NAPAP (Fig. 8b).

Methanol and NMA minima are scattered over all polar regions of the binding site. Minimum 12, which donates to Asp^{189} on the bottom of the S1 pocket, suggests modification of the arginine side chain into an alkyl side chain with a terminal hydroxyl group (Fig. 9). This may result in a slightly reduced intermolecular interaction, since the $Arg-Asp^{189}$ salt bridge is enthalpically stronger than the alkyl hydroxyl-Asp^{189} hydrogen bond. However, this would be expected to be balanced, at least in part, by the fact that desolv-

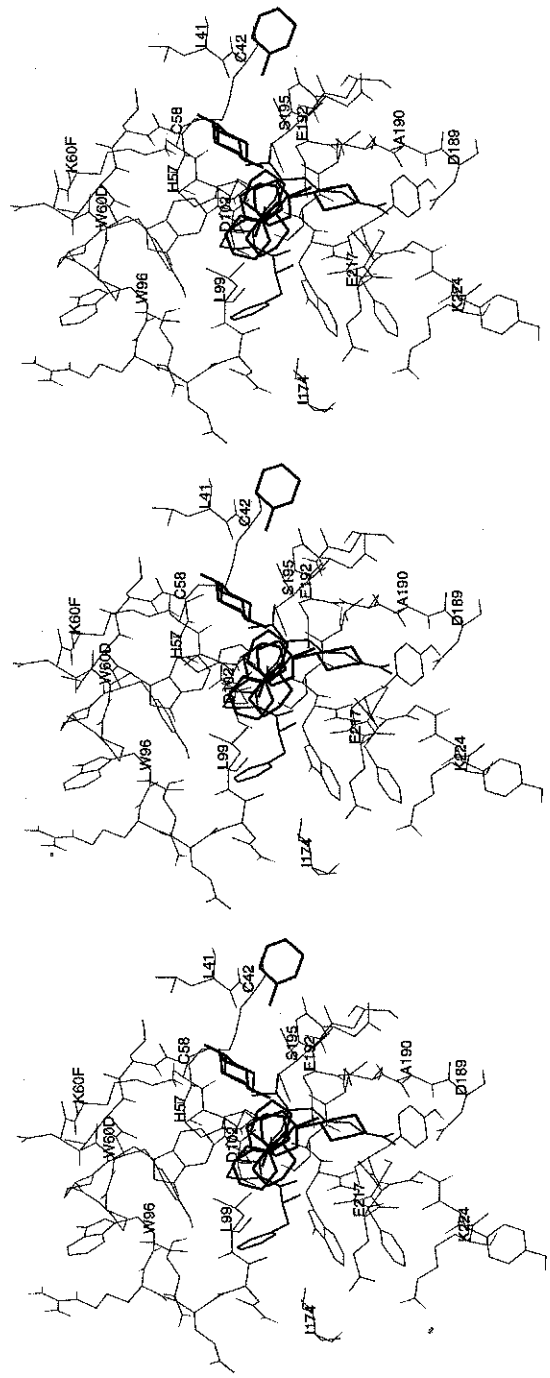
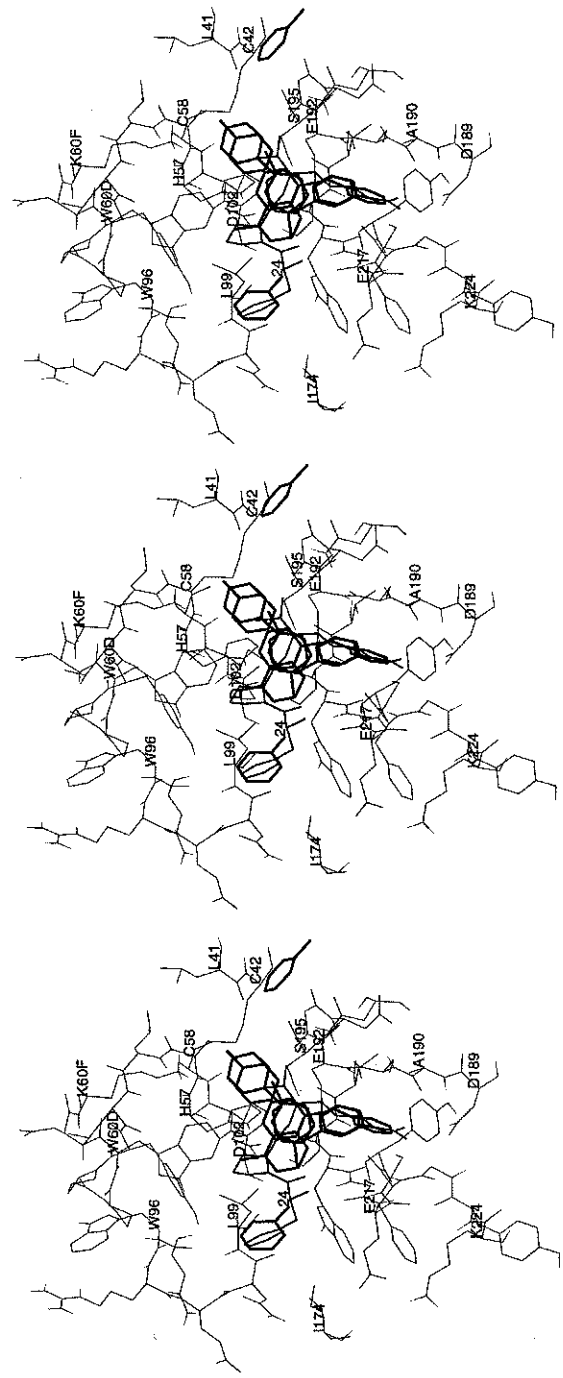


Fig. 7. Stereoview of the nine representatives of the 20 best MCSS minima of methylcyclohexane (thick lines) in the human thrombin binding site (thin lines). The PPACK inhibitor is also shown (medium lines); it was removed for MCSS.



a

Fig. 8. (a) Same as in Fig. 7 for the 10 representatives of the 20 best MCSS minima of methylbenzene. Methylbenzene minimum 24 is also shown with a label. (b) Stereoview of the methylbenzene minima 1 and 24 (thick lines) in the human thrombin binding site (thin lines) from the complex with PPACK. The NAPAP inhibitor is also shown (medium lines); its position was determined by superimposing the thrombin binding site residues of the thrombin–NAPAP complex onto the corresponding residues of the thrombin–PPACK complex.

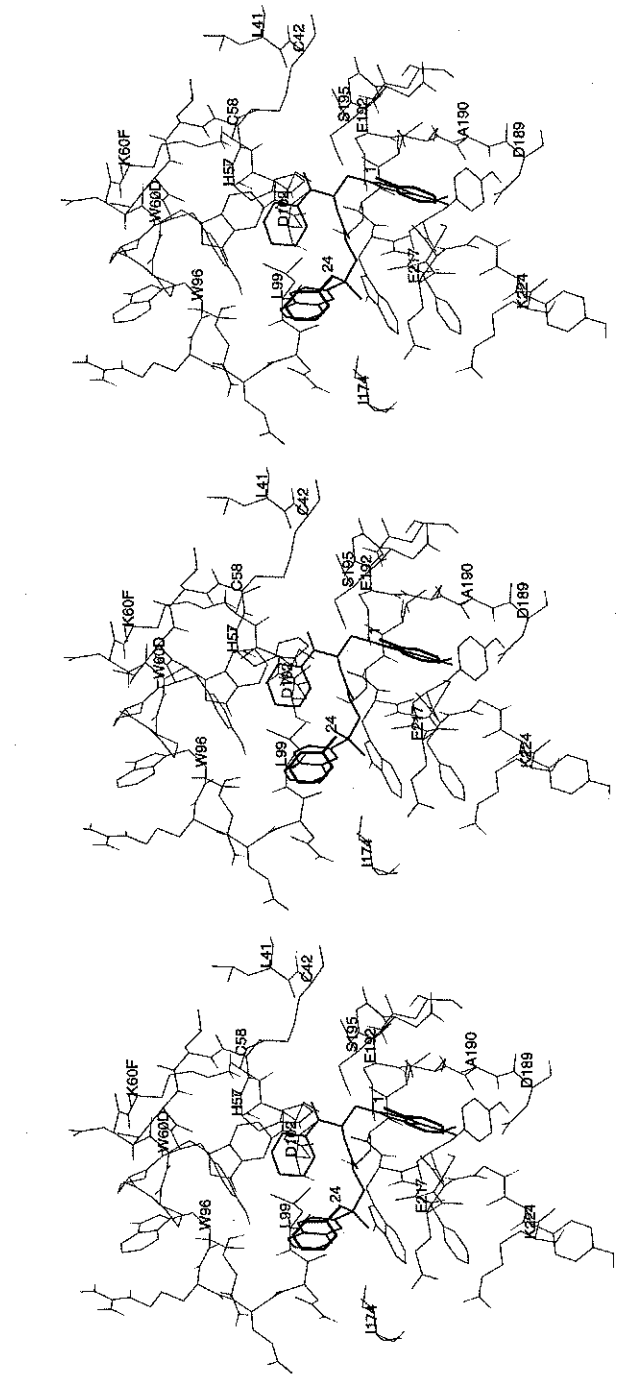


Fig. 8. (continued).

ation of an alkyl hydroxyl is less unfavorable than desolvation of an alkyl guanidinium. Such a modification should increase selectivity against trypsin and plasmin, since compared to these compounds, the S1 pocket is more hydrophobic in thrombin. This results from the Ala^{190} to serine and Glu^{192} to glutamine substitutions in both trypsin and plasmin, and the Val^{213} to threonine substitution in plasmin. In addition, substitution of the positively charged group (guanidinium) with a polar but uncharged (hydroxyl) moiety should facilitate intestinal resorption and diminish side effects [65].

To connect MCSS minima to form candidate molecules, 481 accepted functional group minima were used, including all the minima, except for those involving methylcyclohexane. A database was used composed of 100 small hydrocarbon skeletons constructed from a set of mono-, bi-, and tricyclic hydrocarbons with ring sizes of five and six carbons. The parameters of Eisen et al. [19] were employed, except for the rms overlap criterion between the skeleton hooks and the functional groups (0.3 Å instead of 0.5 Å), the minimal distance between functional group atoms and skeleton atoms (2.0 Å instead of 1.0 Å) and the distance range where the overlap score has its maximal value ($O_{low}=3.5$ instead of 3.2 Å; $O_{high}=4.5$ instead of 4.2 Å). Furthermore, only the hits with scores larger than 100 (instead of 50) were saved. The use of more restrictive parameters resulted in 4494 hits with reasonable geometry and no bad contacts with the protein binding site atoms. HOOK generated four compounds, each having five functional groups. Among the HOOK hits with more than three MCSS minima, there were several with a guanidinium group in S1 and/or a benzene in S3, while a few also had acetate at S1'.

Figure 10a shows the best candidate ligand proposed by HOOK after first sorting the 4494 hits by their number of hooked groups and then by overlap score. This molecule consists of a tricyclic hydrocarbon ring connecting five MCSS minima, i.e., methylbenzene minimum 31 in S3, methanol minimum 28, which makes two hydrogen bonds with the polar groups in Gly^{216}, propane minimum 28, which interacts with the side chain of Trp^{148}, methylguanidinium minimum 8 in S1, and *N*-methylacetamide minimum 35 between the Trp^{60D} and Ser^{195} side chains. In addition, one of the two outer five-membered rings of the skeleton is positioned in the S1 site. This structure has most of the noncovalent interactions between PPACK and the thrombin active site and appears to be a good ligand candidate. Another candidate ligand suggested by HOOK (number 7) consists of a tricyclic hydrocarbon ring with four MCSS minima as substituents (Fig. 10b). These are methylguanidinium minimum 5 in S1, methylbenzene minimum 17, which overlaps the proline side chain of PPACK in S2, and methanol minima 35 and 42, which donate to the side chain of Glu^{192} and the backbone CO of Cys^{191} (Fig. 10b).

A bicyclic hydrocarbon ring linking four MCSS minima is shown in Fig. 10c (HOOK candidate number 14). It has methylbenzene minimum 4, whose aromatic ring is 'sandwiched' between the two amide groups in the top half of S1 (backbone of residues 191–192 and 215–216) and it overlaps the alkyl part of the PPACK arginine side chain. The other MCSS minima are NMA 36, which donates to the His^{57} side chain and accepts from the NH of Gly^{193}, acetate 12 (obtained with a distance-dependent dielectric), which forms a salt bridge with Lys^{60F} at S1', and methylbenzene 82, which occupies part of the S2 and S3 sites. Figures 10a and 10c show that the S2 and S3 pockets are contiguous and suggest that the hydrophobic portions of these sites may be linked by an ethylene or phenyl ring, thereby increasing the rigidity of a candidate ligand having apolar side chains at S2 and S3.

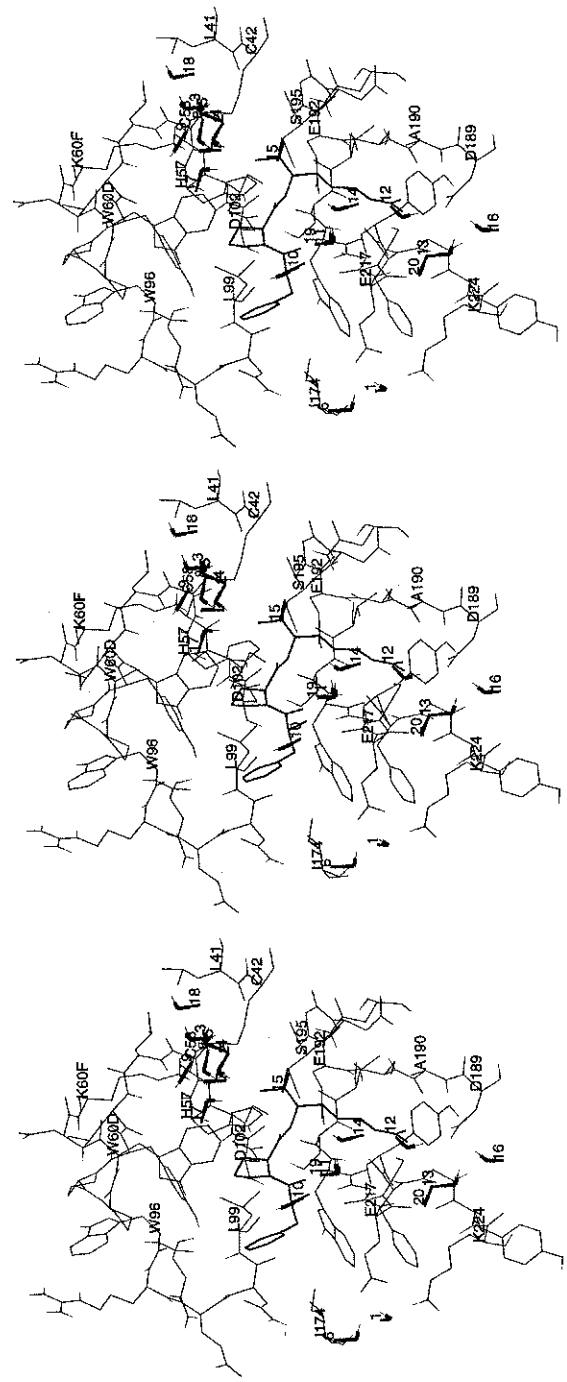
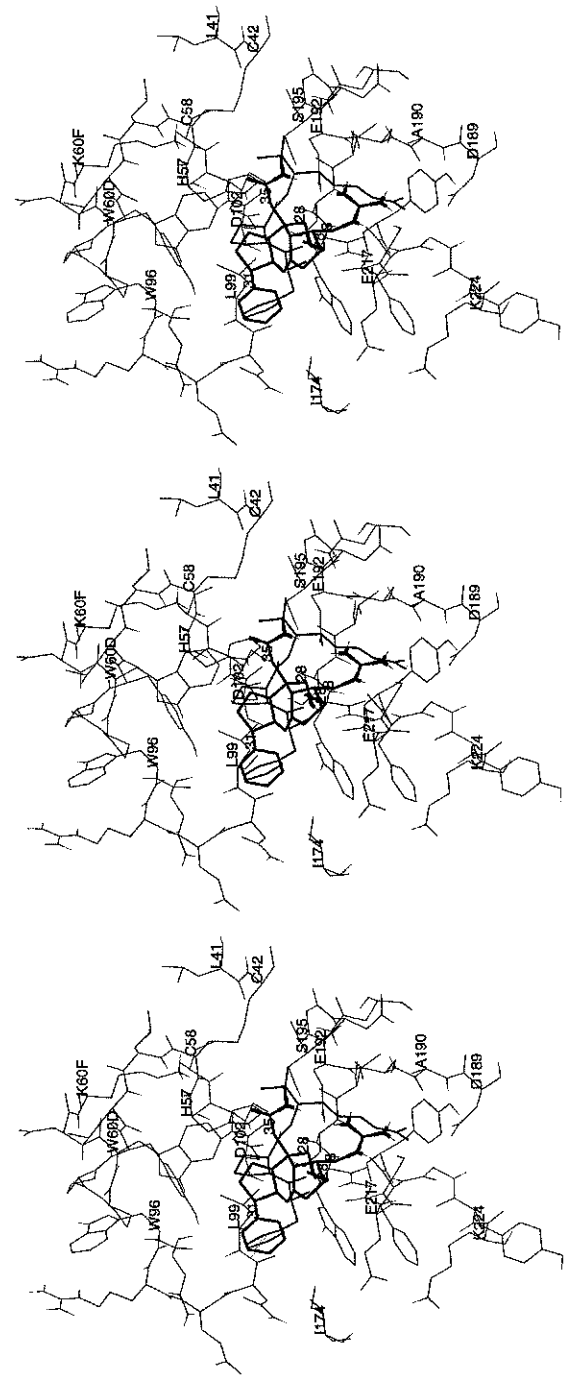


Fig. 9. Same as in Fig. 7 for the 20 best MCSS minima of methanol.



a

Fig. 10. HOOK candidate ligands numbers 1, 7 and 14 (a, b and c, respectively; thick lines) in the human thrombin binding site (thin lines). The MCSS minima selected by HOOK are labeled according to their CHARMM energy and are discussed in the text. The PPACK inhibitor is also shown (medium lines).

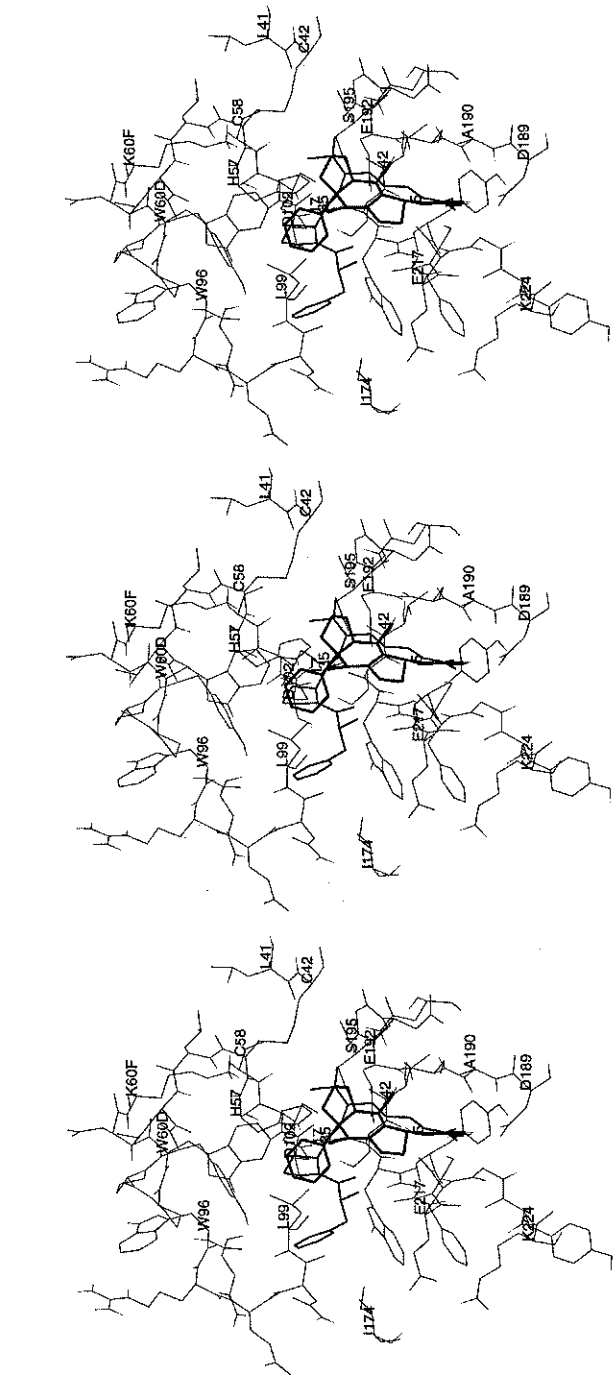


Fig. 10. (continued).

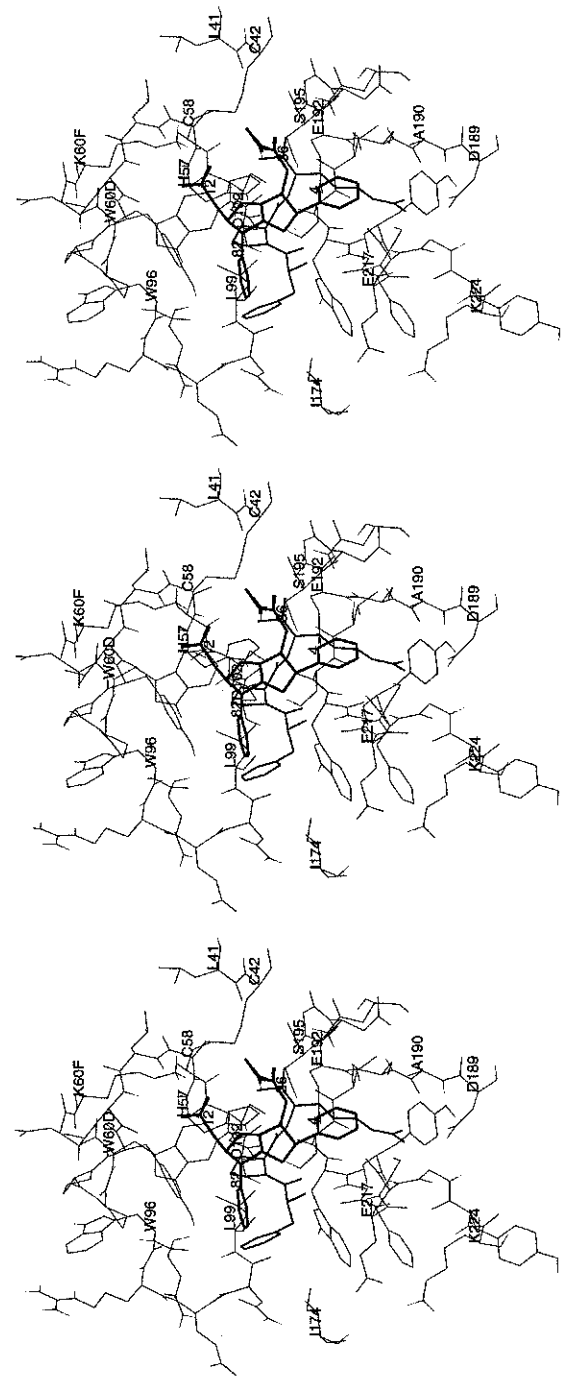


Fig. 10. (continued).

Limitations of CONNECT and HOOK

To circumvent the problems originating from the poor pharmacological properties of peptides, a more general version of the program CONNECT is currently being developed (A. Caflisch, unpublished results). This will allow the treatment of novel synthetic monomers which, when connected in a combinatorial fashion, could yield relatively low molecular weight polymeric compounds with enhanced pharmacological profiles. This computational approach will be ideally suited for complementing combinatorial synthesis methodologies [67,68].

The main limitation of HOOK and related approaches based on the connection of isolated fragments is that synthetic problems can arise in the compounds that are generated; in fact, the suggested molecules may not even be stable. In many cases, a skilled organic chemist can quickly choose the best candidates. Also, it would be possible to do postprocessing based on retrosynthetic programs, such as the LHASA program developed by E.J. Corey and co-workers at Harvard University [69]. As an alternative, known compounds in databases such as the Fine Chemicals Database [70] can be used in HOOK (R. Hubbard, private communication). Finally, the skeletons can be defined such that the allowed hooks are limited to the position where derivatization by functional groups is known to be possible.

HOOK and related database search methodologies are limited to finding candidate ligands that are derived from a set of known compounds. This is not a problem if the objective is to start with compounds from a proprietary database. For 'de novo' candidates, a method has been developed that creates connecting fragments from a distribution of carbon atoms in the binding site. This approach (Dynamic Ligand Design) employs a Monte Carlo optimization procedure with an appropriately chosen pseudoenergy function [71]. An alternative is to use genetic algorithms for constructing new molecules [72].

Estimation of the free energy of binding

Any computer-aided ligand design approach requires a methodology for the evaluation of the binding free energy of a protein–ligand complex. The compounds suggested by CONNECT, HOOK, and related approaches have to be ordered in terms of their expected binding constants, as well as other properties (e.g., solubility) to choose certain ones for synthesis and testing for their biological activity. This means that relatively simple approaches are needed that can be used for the evaluation of 100 to 1000 compounds. Methods with the required accuracy and speed are not yet available, but some progress is being made in their development. In many cases, the problem is simplified somewhat by the fact that an absolute binding constant is not needed, i.e., relative values of binding constants are sufficient for determining which modifications of a lead compound are the best candidates for further investigation.

The most reliable approach, particularly for small changes, is based on free energy simulations. Successful examples of applications of such simulations for the comparison of the binding constants of similar ligands are given in Refs. 73–75. Unfortunately, free energy simulations cannot be used for large-scale screening because they are too costly in terms of computer time. This has led to the introduction of other methods that are much