Journal of Computer-Aided Molecular Design, 10 (1996) 372–396
ESCOM

J-CAMD 357

# Computational combinatorial ligand design: Application to human α-thrombin

Amedeo Caflisch

*Department of Biochemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland*

Received 26 April 1996
Accepted 11 June 1996

*Keywords:* Structure-based drug design; Thrombin; Combinatorial chemistry; Functional group; CCLD; Electrostatic screening; Desolvation; Finite-difference Poisson–Boltzmann technique

**Summary**

A new method is presented for computer-aided ligand design by combinatorial selection of fragments that bind favorably to a macromolecular target of known three-dimensional structure. Firstly, the multiple-copy simultaneous-search procedure (MCSS) is used to exhaustively search for optimal positions and orientations of functional groups on the surface of the macromolecule (enzyme or receptor fragment). The MCSS minima are then sorted according to an approximated binding free energy, whose solvation component is expressed as a sum of separate electrostatic and nonpolar contributions. The electrostatic solvation energy is calculated by the numerical solution of the linearized Poisson–Boltzmann equation, while the nonpolar contribution to the binding free energy is assumed to be proportional to the loss in solvent-accessible surface area. The program developed for computational combinatorial ligand design (CCLD) allows the fast and automatic generation of a multitude of highly diverse compounds, by connecting in a combinatorial fashion the functional groups in their minimized positions. The fragments are linked as two atoms may be either fused, or connected by a covalent bond or a small linker unit. To avoid the combinatorial explosion problem, pruning of the growing ligand is performed according to the average value of the approximated binding free energy of its fragments. The method is illustrated here by constructing candidate ligands for the active site of human α-thrombin. The MCSS minima with favorable binding free energy reproduce the interaction patterns of known inhibitors. Starting from these fragments, CCLD generates a set of compounds that are closely related to high-affinity thrombin inhibitors. In addition, putative ligands with novel binding motifs are suggested. Probable implications of the MCSS–CCLD approach for the evolving scenario of drug discovery are discussed.

## Introduction

Computer-aided structure-based ligand design is a complex and challenging area of research. It is concerned with the prediction of chemically reasonable compounds that are expected to bind strongly to key regions of biologically relevant molecules (e.g., enzymes, receptor fragments) of known three-dimensional structure so as to inhibit or alter their activity. Its complexity is documented in several successful cases where structure-based ligand-design efforts have led to the development of compounds that are currently in clinical trials [1–3]. Despite significant advances in molecular simulation methodologies over the last two decades [4–6] and the ever-decreasing price/performance ratio of computers, the prediction of binding affinities (even only qualitative) is still very difficult, if not impossible. Hence, the computational approach is often considered less mature than the experimental techniques involved in the drug-discovery process [1,7]. At the same time, the enhanced capabilities for the cloning and fast sequencing of both human and nonhuman genomes and refined gene technologies promise that an ever-increasing number of enzymes and receptors will become available as potential drug targets in the coming years. Moreover, the determination of the three-dimensional structure of these proteins or protein fragments will be facilitated by recent advances in nuclear magnetic resonance techniques [8,9] and homology modelling approaches [10–12]. Thus, new ideas and methods for computational approaches to the ligand-design problem are

needed. This constitutes a challenge for theoreticians, who would like to develop and use computational techniques not only to rationalize experimental data a posteriori, but also to make predictions, which might be utilized as viable alternatives to experimental structure determination.

The strategy we have chosen for computer-aided ligand design consists of three parts [13]. The first one is an efficient method for the exhaustive search of optimal positions and orientations of small and mainly rigid molecules or molecular fragments on the surface of a macromolecular target. To solve this problem, the multiple-copy simultaneous-search (MCSS) procedure was developed [14]. It is known from a multitude of crystal structures of enzyme–inhibitor complexes, that most, if not all, of the functional groups of ligands with high affinity and selectivity are involved in favorable interactions with the surrounding protein atoms [7,15,16]. Hence, it is evident that low-molecular-weight ligand molecules have only a minimal number of linkage elements not involved in favorable binding interactions.

Secondly, given a set of such positions and orientations for functional groups, it is necessary to find possible connections between these fragments to form putative ligands. Ideally, the linker units should be as small as possible if they are not involved in favorable interactions with the protein. The program CONNECT was developed to generate peptide leads from optimal positions of *N*-methylacetamide (NMA) groups and functional groups representing side chains by fusing atoms belonging to MCSS minima [17]. HOOK is another approach which was developed to retrieve, from a three-dimensional database, molecular skeletons that fit well into the protein binding region and make bonds to functional groups [18].

Thirdly, a method is needed to estimate which of the resulting candidate molecules are likely to have the highest affinity 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 treatment of the interactions, as well as a rigorous treatment of solvent and entropic effects. This can be applied only to a limited set of molecules.

A stepwise procedure is used because it is 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 binding free energy. By firstly docking functional groups and then connecting them to form candidate ligands, it is possible to search through a very large number of highly diverse molecules in a relatively short time.

A novel approach for addressing the second step is presented in this study. The MCSS minima are firstly sorted according to an approximated free energy of binding, whose solvation component is assumed to be the sum of electrostatic and nonpolar contributions [19]. For each protein–MCSS minimum complex the electrostatic contribution is calculated in the continuum dielectric approximation by the numerical solution of the linearized Poisson–Boltzmann (LPB) equation [20,21]. A new computational scheme is described for the efficient and accurate evaluation of the shielded electrostatic interaction between protein and bound fragment, and their electrostatic desolvation energies. The nonpolar solvation energy, which incorporates cavitation effects and solute–solvent dispersion interactions, is assumed to be proportional to the change in solvent-accessible surface area [22,23]. A program has been developed for computational combinatorial ligand design (CCLD). Starting from the MCSS minimum with the most favorable binding free energy, the ligand-generation algorithm proceeds in an iterative way by linking an additional fragment to the actual construct. Although CCLD performs an exhaustive search, it is very efficient because of the precomputation of a list of overlapping, i.e., mutually excluding, fragment pairs, and a list of bonding fragment pairs. The linker units are small (from 0 to 3 covalent bonds), since their function is to optimally connect two fragments without adding considerably to the molecular weight. Thus, the candidate ligands generated by CCLD have most of their groups involved in optimal interaction patterns with the surrounding protein atoms. A set of simple rules has been implemented to preferentially select linker units that result in molecules with few rotatable bonds and of accessible chemical synthesis. To avoid combinatorial explosion problems, the 'growth' of a ligand is stopped if the average value of the approximated binding free energy of its fragments exceeds an user-selected threshold value. In a typical run with in the order of 1000 MCSS minima, CCLD produces several thousands of compounds, which are then sorted by average free energy and clustered according to a similarity criterion based on the percentage of identical fragments.

This methodology was tested on human α-thrombin, a trypsin-like serine protease which fulfills a central role in both haemostasis and thrombosis [24]. This enzyme was selected for its intrinsic interest and for the wealth of structural information [15,25,26] and binding-affinity data available [3,24,27]. The vast majority of the MCSS minima with the lowest approximated binding free energy are involved in the same interaction patterns as those of the functional groups of high-affinity thrombin inhibitors. It is shown that the solvation correction is essential for a realistic ranking of the minimized positions of the functional groups. This represents a major improvement with respect to previous applications of MCSS to thrombin [13,28]. Using the MCSS minima with favorable binding free energy, CCLD generates a set of ligands with an aliphatic or aromatic group in S3, an aliphatic moiety in S2 and a positively charged functionality in S1. These are closely related to high-affinity active-site thrombin inhibitors. Moreover, several candidate ligands suggested by

CCLD show new binding motifs. The latter provide sources of inspiration for novel ligands and/or serve as indicators of viable modifications of known inhibitors.

Some aspects of the combinatorial design approach of CCLD are common to previously published works [17,18,29,30]; a comparison will be given in the Discussion. Furthermore, the field of computer-aided structure-based ligand design has been recently reviewed by several contributors [13,29,31].

## Methods

Firstly, the MCSS procedure as implemented in the present study is summarized. The continuum method used to evaluate the electrostatic contribution to the free energy is then outlined, with a detailed description of the approach used to decompose the electrostatic free energy into protein desolvation, ligand desolvation, and intermolecular electrostatic energy, as screened by the solvent. Finally, the program for the combinatorial generation of putative ligands is described.

### *Multiple copy simultaneous search*

The MCSS method [14,17] determines energetically favorable positions and orientations (local minima of the potential energy) of functional groups on the surface of a protein or receptor of known three-dimensional structure. In preparation for the use of CCLD, MCSS was applied to the thrombin active site with the structure taken from the complex with PPACK [15,25], D-Phe-Pro-Arg-$CH_2Cl$ (PDB code 1PPB), the archetypal thrombin inhibitor [32]. The side chain of $Trp^{148}$, which is part of the autolysis loop, and that of $Glu^{192}$ are exposed to solvent and assume different orientations in complexes with different inhibitors, depending on the crystallization conditions and on the inhibitor type [26]. They were mutated to alanine to avoid possible artificial positions of the fragments. The coordinates of the hydrogen atoms were generated with the HBUILD [33] option of the CHARMM program and subsequent minimization with fixed non-hydrogen atoms. For each of the functional groups listed in Table 1, 10 000 replicas were randomly distributed in a 9.0-Å sphere centered on the coordinates of the carbonyl carbon of the PPACK proline. To avoid excessive steric clashes between the atoms of the fragments and those of thrombin, a minimal distance of 2.0 Å (1.8 Å for groups with hydrogen atoms) was used as cutoff during the random-placement phase. The size of the sphere was chosen to cover the S3 to S2' pockets of thrombin (from $Ile^{174}$ to $Leu^{40}$); as a basis for comparison, the heavy atom most distant from the proline carbonyl carbon in PPACK is a nitrogen in the arginine guanidinium group at 8.03 Å. The functional groups used are

TABLE 1
FUNCTIONAL GROUPS USED FOR MCSS

| Group | Electrostatic solvation free energy[a] | CHARMM energy[b] | | No. of minima found | $\Delta G_{binding}$ [c] | | | | No. of minima with $\Delta G_{binding} < 0$ |
|---|---|---|---|---|---|---|---|---|---|
| | | Lowest | Highest | | Lowest | 2nd | 3rd | Highest | |
| **Nonpolar groups** | | | | | | | | | |
| propane | 0.0 | −7.1 | −1.6 | 84 | −9.2 | −9.1 | −8.6 | 23.0 | 54 |
| cyclopentane | 0.0 | −9.0 | −2.1 | 49 | −9.3 | −9.1 | −8.8 | 15.9 | 40 |
| cyclohexane | 0.0 | −9.5 | −2.5 | 42 | −10.0 | −8.7 | −8.7 | 23.8 | 31 |
| benzene | 0.0 | −11.4 | −4.4 | 32 | −9.9 | −9.5 | −9.4 | 18.7 | 25 |
| **Polar groups** | | | | | | | | | |
| methanol | −7.4 | −23.4 | −1.2 | 78 | −8.3 | −7.3 | −7.2 | 13.7 | 54 |
| 2-propanone | −5.3 | −18.6 | −1.7 | 57 | −8.2 | −7.3 | −7.0 | 18.3 | 35 |
| NMA | −9.1 | −28.8 | −3.0 | 125 | −9.8 | −9.6 | −9.5 | 18.8 | 83 |
| NDMA | −5.8 | −24.8 | 0.4 | 150 | −10.8 | −10.4 | −10.1 | 20.8 | 103 |
| pyrrole | −3.2 | −18.7 | −6.8 | 50 | −8.3 | −8.2 | −8.0 | 20.6 | 31 |
| imidazole | −6.0 | −22.2 | −1.9 | 104 | −10.9 | −10.6 | −10.5 | 19.1 | 76 |
| phenol | −6.8 | −22.5 | −6.8 | 108 | −11.4 | −11.0 | −10.6 | 17.2 | 78 |
| **Charged groups** | | | | | | | | | |
| methylammonium | −99.0 | −58.5 | −6.1 | 52 | −6.1 | −1.9 | −0.7 | 20.2 | 6 |
| methylguanidinium | −84.5 | −59.0 | −7.7 | 141 | −12.5 | −12.1 | −11.4 | 10.0 | 94 |
| pyrrolidine | −82.2 | −49.8 | −8.2 | 68 | −9.9 | −5.7 | −4.8 | 13.2 | 28 |
| 2-acylpyrrolidine | −78.1 | −39.9 | −8.7 | 145 | −11.4 | −11.1 | −9.6 | 17.4 | 64 |
| acetate ion | −71.5 | −42.0 | −6.9 | 29 | −7.5 | −6.9 | −4.7 | 9.7 | 5 |

All energy values are in kcal/mol.
[a] Calculated by numerical solution of the LPB equation.
[b] The CHARMM energy is the sum of intermolecular and intraligand energies.
[c] Calculated by use of Eq. 2.

small chemical fragments commonly found as substituents of larger organic molecules. To map both the hydrophilic and hydrophobic regions of the thrombin active site, charged (methylammonium, methylguanidinium, pyrrolidine, 2-acylpyrrolidine, acetate), polar (methanol, 2-propanone, *N*-methylacetamide, *N*,*N*-dimethylacetamide), aromatic (benzene, pyrrole, imidazole, phenol), and aliphatic (propane, cyclopentane, cyclohexane) groups were used (Table 1). Subsets of 500 randomly distributed replicas of the same group were simultaneously minimized in the force-field of the protein. The CHARMM [34,35] program was utilized for all minimizations performed in this work. For both the protein and the functional groups, the parameters from the polar hydrogen set (PARAM19) were used. Polar hydrogens are treated explicitly, whereas aliphatic and aromatic hydrogens are considered as part of the extended carbon atom to which they are bonded. This considerably simplifies the search procedure in that it reduces the number of atoms and eliminates torsional degrees of freedom (e.g., for the $CH_3$ of methanol). A classical version of the time-dependent Hartree (TDH) approximation [36] 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 m does not interact with replica n, for each m and n in the subset. Since the protein atoms are fixed, 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. The minimization began with 500 iterations of the steepest-descent algorithm, which provides a better performance than higher-order algorithms for very poor starting conformations where the gradient is large. The conjugate-gradient algorithm was then applied [34,37]. The positions were compared every 1000 steps to eliminate replicas converging toward a common minimum. The criteria used to characterize a common minimum were a deviation of 0.2 Å rms or less between two replicas and a decreasing rms deviation in the final 200 steps. A convergence criterion of 0.001 kcal $mol^{-1}$ $Å^{-1}$ for terminating the minimization was utilized. For a complete minimization, between $4 \times 10^3$ and $15 \times 10^3$ steps were usually required, depending on the size and complexity of the functional group.

The implementation of the MCSS procedure used in this study differs in three points from that in the original description [14]. Firstly, a distance-dependent dielectric function was used instead of the unit dielectric constant in the vacuum potential. This results in additional minima, since in the constant dielectric calculation used in previous studies [14,17], the strong vacuum Coulombic interaction yielded a smoother configurational space than the one with the distance-dependent dielectric function [13]. Secondly, nonbonding cutoffs of 5.0 Å and 6.0 Å were used for the first and second cycles of minimization, respectively, (each consisting of 1000 steps) to speed up the calculation, while a cutoff of 7.5 Å, was used for the remaining cycles. This corresponds to the default value for the CHARMM polar hydrogen parametrization, in which the nonbonding interactions are shifted by the use of a fourth-degree polynomial [34]. Finally, a UNIX shell script was developed to postprocess the MCSS minima and compute the loss in solvent-accessible surface area of both the protein and functional group and the electrostatic contribution to the free energy of binding (see below).

*Electrostatic solvation free energy*

Polarization of the solvent by the charges on the solute affects the electrostatic energy of a molecular assembly in two ways: (i) the interactions between solute partial charges are screened; and (ii) the solvent reaction field interacts directly with each solute charge (self energy). The continuum electrostatic free energy of solvation is the sum of the screening effect and the direct interaction of each solute charge with the solvent [21].

Studies have shown that the numerical solution of the linearized Poisson–Boltzmann (LPB) equation yields a good estimate of the electrostatic free energies of solvation in macromolecules [20,21,38,39]. The LPB differential equation is approximated by a set of finite-difference equations on a grid [40]. The latter are solved on a computer by iterative adjustments of the value of the potential at each grid point. In this study, the UHBD program [41–43] was utilized for solving the finite-difference LPB equation. The partial charges and atomic radii of the CHARMM polar hydrogen potential were used for the LPB calculations. It has been shown that the solvation free energy of models of polar and ionizable compounds, calculated with the finite-difference method and the CHARMM polar-hydrogen parameter set, agree well with experimental data [44].

UHBD places the charges on a grid according to a trilinear weighting method [45]. The solute dielectric constant was set to 1.0, which is consistent with the value used for the parametrization of the CHARMM charges. A dielectric constant of 78.5 was assigned to the continuum solvent medium. The surface of the low dielectric region was delimited by applying a solvent probe of 1.4-Å radius. Furthermore, the permittivity was linearly interpolated at the midpoints between grid points intersecting the dielectric boundary (dielectric boundary smoothing), since this reproduces the potential near the discontinuity region more accurately and has been shown to improve convergence [43,46]. Values of 298 K for the temperature, 100 mM for the ionic strength (corresponding to physiological conditions), and 2.0 Å for the Stern layer (ion-exclusion layer) were used.

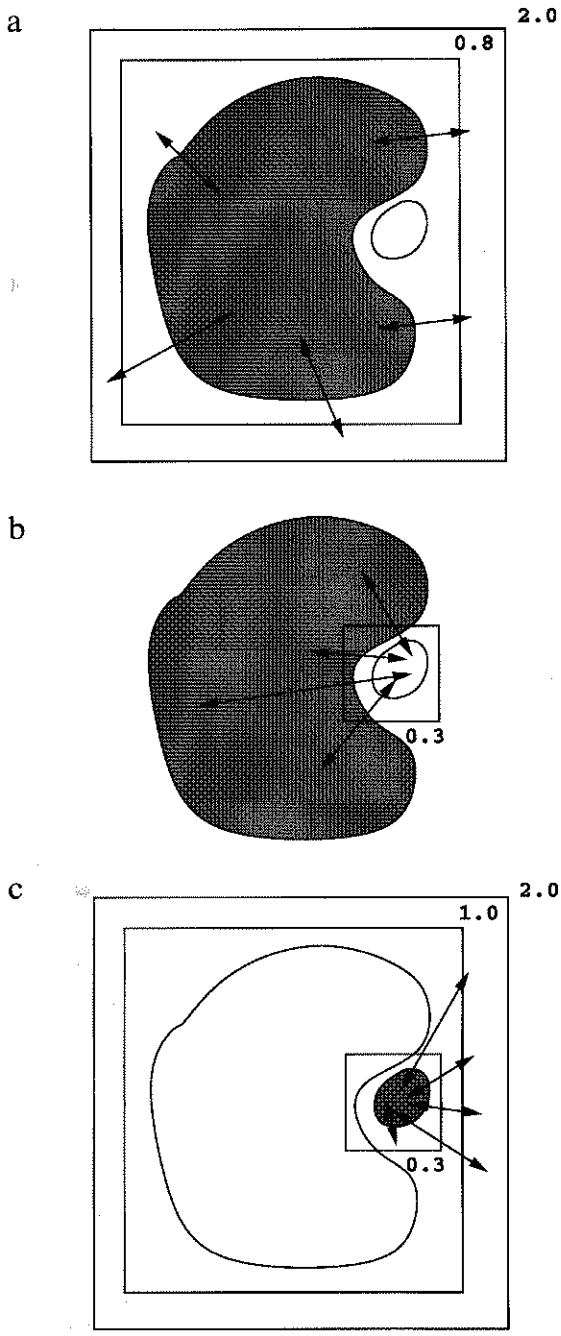


Fig. 1. Schematic representation of the set-up for the numerical solution of the LPB equation. Protein and ligand are represented by a large and a small shape, respectively. Each rectangle corresponds to the boundary of the grid on which the PB equation was solved numerically. The number close to a grid is the distance between grid points in Å. A rectangle enclosed in a larger one represents a focussed calculation whose boundary values were taken from the calculation done on the larger grid. Arrows symbolize electrostatic interactions, shaded shapes are charged while empty shapes are neutral. (a) Electrostatic solvation energy of the protein–uncharged-ligand complex. First grid, $48 \times 46 \times 45$; second grid, $94 \times 89 \times 88$. (b) Electrostatic free energy of interaction between protein and ligand. The field obtained from the focussed calculation in (a) was used to set the boundary potential of the third grid ($67 \times 67 \times 67$). (c) Electrostatic solvation energy of the ligand–uncharged-protein complex. First grid, $48 \times 46 \times 45$; second grid, $75 \times 71 \times 70$; third grid, $67 \times 67 \times 67$.

The scheme shown in Fig. 1 was used to calculate the three parts of the electrostatic contribution to the binding free energy, i.e., protein desolvation, shielded intermolecular interaction, and ligand desolvation. For the evaluation of the protein desolvation the protein atoms were charged, while the ligand was considered as a neutral region of low dielectric, which displaces the solvent (Fig. 1a). To set the boundary potential the molecular complex was considered as a single Debye–Hückel sphere of 20-Å radius and the protein net charge. Firstly, a grid of $48 \times 46 \times 45$ points and a grid spacing of 2.0 Å were used; this yields a layer of solvent (high dielectric constant) of at least 20 Å around the structure of the complex (low dielectric constant). The potential obtained from this calculation was used for the boundary potential of a second focussed [46,47] calculation, which was performed with a grid of $94 \times 89 \times 88$ points and a grid spacing of 0.8 Å (10-Å layer of solvent around the solute). Both these grids were centered on the rigid protein and are the same for all protein–MCSS minimum complexes. This dramatically reduces the error originating from the distribution of the charges on the grid points. The resulting potential was used to calculate the electrostatic solvation energy of the complex between the protein and uncharged ligand. For this purpose, the finite-difference approximation of the Coulombic interaction energy between charged atoms and the interaction energy of each atom with its own potential (this contribution arises from the discretization of the atom charges onto a grid) were subtracted from the total electrostatic energy of the system calculated by the finite-difference LPB technique [48]. For an interior dielectric of 1.0 this yields the same result as the usual (and computationally more expensive) method of performing two finite-difference calculations; the first one with the low-dielectric solute in a high-dielectric continuum and the second one with the low-dielectric solute in a vacuum (1.0-dielectric) continuum. To obtain the electrostatic desolvation energy of the protein, the solvation energy of the isolated protein (–4346.01 kcal/mol) was then subtracted from the solvation energy of the protein–uncharged-ligand complex. It is worth noting that even upon binding of a nonpolar functional group the protein experiences some electrostatic desolvation, especially if the nonpolar group binds in the vicinity of polar groups. Thus, the protein-desolvation term was calculated for the MCSS minima of all functional group types.

A third focussed calculation was then performed with a grid of $67 \times 67 \times 67$ points and a grid spacing of 0.3 Å centered on the MCSS minimum (Fig. 1b), the potential obtained from the previous focussed calculation was used for the boundary potential. The intermolecular electrostatic energy, as mediated by the solvent, was calculated by:

$$\Delta G_{elect}^{interm} = \sum_{j=1}^{N_m} q_j \phi_j \qquad (1)$$

where $N_m$ is the number of atoms in MCSS minimum m, $q_j$ is the charge of atom j on minimum m, and $\phi_j$ is the electrostatic potential generated by the protein at the location of atom j. No factor 1/2 appears, since the partial charges generating the electrostatic field reside on the atoms of the protein, while the charges $q_j$ belong to the MCSS minimum.

To calculate the desolvation of the ligand, partial charges were assigned to the atoms of the ligand, while the protein was considered as a neutral region of low dielectric constant. The LPB equation was firstly solved on the same grid used at the beginning of the protein-desolvation calculation, i.e., $48 \times 46 \times 45$ and 2.0-Å grid spacing (≥ 20-Å layer of solvent). This was followed by two focussed calculations; the first one on a grid of $75 \times 71 \times 70$ and 1.0-Å grid spacing centered on the protein (≥ 10-Å layer of solvent) and the second on a grid of $67 \times 67 \times 67$ points and a grid spacing of 0.3 Å centered on the MCSS minimum (Fig. 1c), i.e., the same grid used for the evaluation of the intermolecular electrostatic energy. The

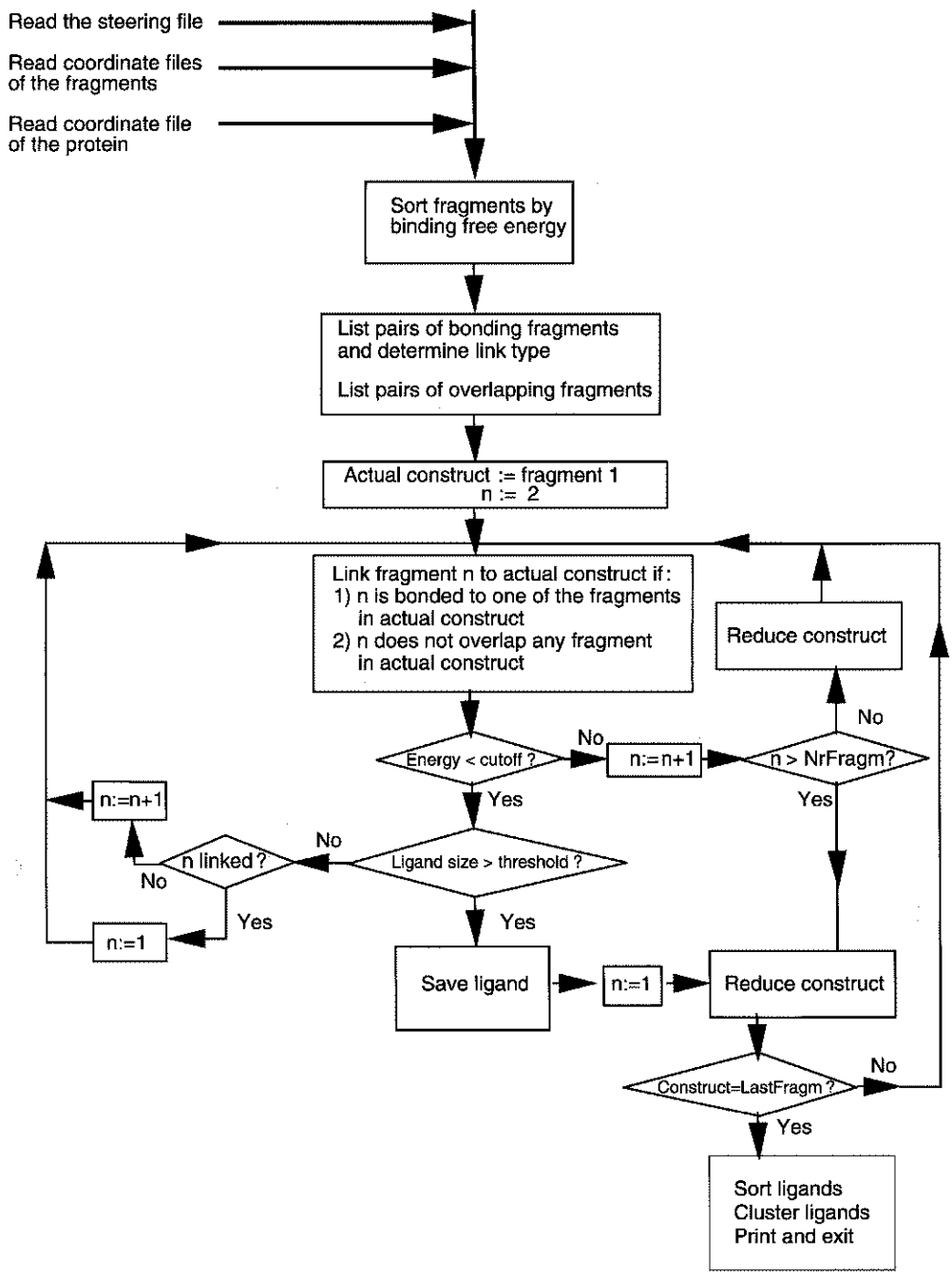


Fig. 2. Schematic representation of the CCLD program. Variable assignments are symbolized by ':='. Conditional statements are enclosed by diamonds (<>).

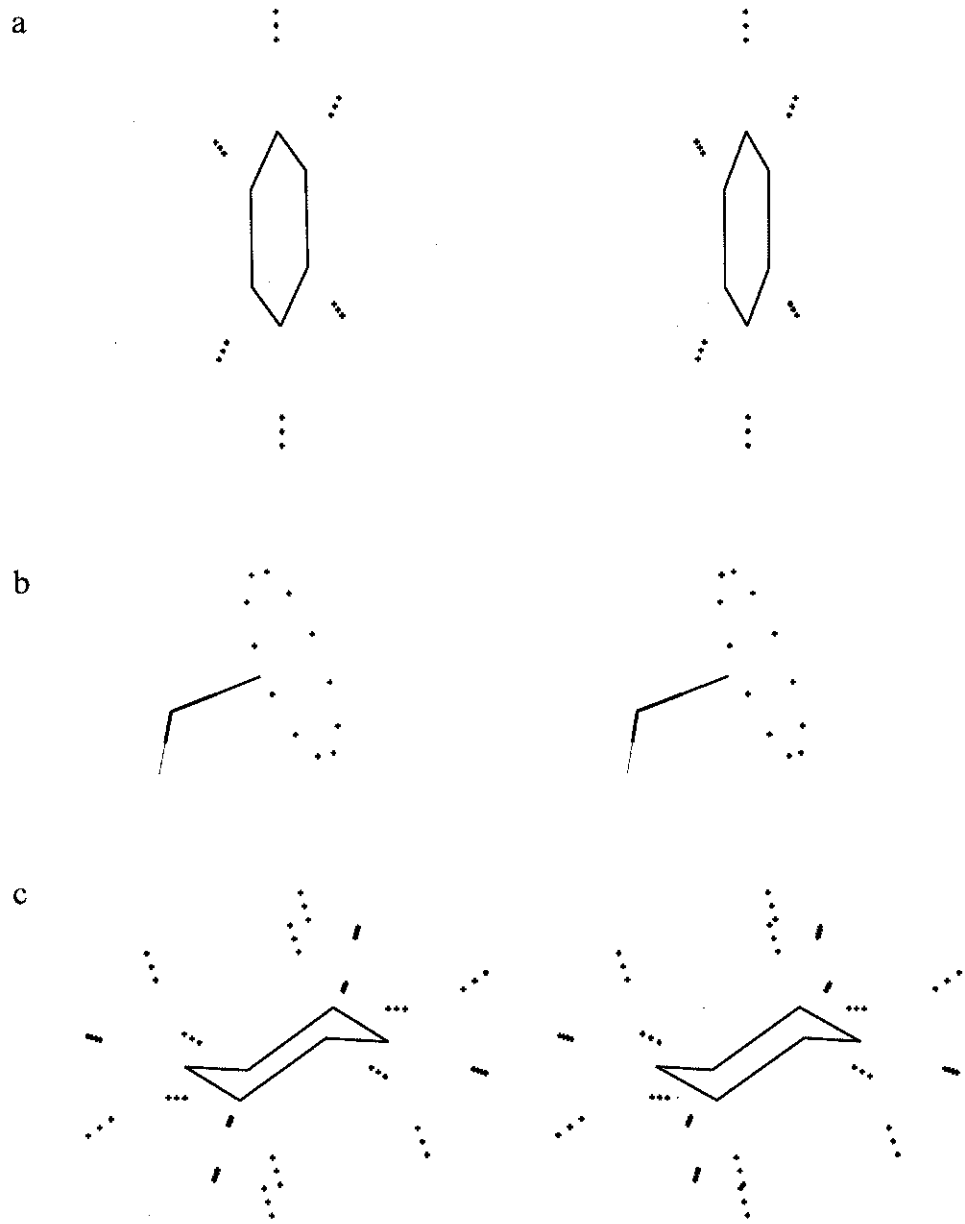


Fig. 3. Stereoviews of the linkage points generated by CCLD for the linkage atoms. (a) For an $sp^2$ atom (e.g., benzene carbons), three linkage points are defined on the plane at a distance of 1.3 Å, 1.5 Å, and 1.7 Å. (b) For an $sp^3$ atom (e.g., the carbon atom in methanol), 12 linkage points are distributed on each of two circles at an angle of 110° with respect to the C–O direction and a distance of 1.2 Å and 1.5 Å from the $sp^3$ carbon (only one circle of linkage points is shown in this picture for clarity sake). (c) For an $sp^3$ carbon connected to two heavy atoms in the fragment (e.g., the cyclohexane carbons), six linkage points are defined in a tetrahedral arrangement (three points for each vertex) at a distance of 1.3 Å, 1.5 Å, and 1.7 Å. In addition, four linkage points are distributed on the C–C–C plane at a distance of 1.26 Å, 1.32 Å, 1.38 Å, and 1.44 Å. These are used only for a conversion from $sp^3$ carbon to $sp^2$ nitrogen if the linker unit is a keto group.

subtraction scheme mentioned above [48] was used to compute the electrostatic solvation energy of the ligand–uncharged-protein complex. The electrostatic desolvation energy of the ligand was then computed by subtracting the solvation energy of the isolated ligand (see values in second column of Table 1) from the solvation energy of the ligand–uncharged-protein complex.

The sensitivity to the position of the complex within the grid was tested: the electrostatic desolvation energies of the protein differed by less than 0.6 kcal/mol (final grid spacing of 0.8 Å to avoid excessive memory requirements), while the intermolecular energies and the ligand-desolvation energies differed by less than 0.2 kcal/mol (final grid spacing of 0.3 Å for both).

### *Computational combinatorial ligand design*

#### *Overview*

The CCLD program requires as input atomic coordinates and partial charges of the protein atoms, as well

as the coordinates of the MCSS minima and the individual contributions to the free energy of binding. An additional file contains a number of control parameters and, for each functional group used for MCSS, a list of atoms which can be used for connection (linkage atoms). The following procedures are performed during a regular execution of CCLD (Fig. 2): (i) the MCSS minima are firstly sorted according to their approximated binding free energies; (ii) then, a list of bonding fragment pairs and a list of overlapping fragment pairs are generated; (iii) this is followed by the combinatorial generation of putative ligands; and (iv) finally, the ligands are sorted and clustered.

*Binding free energy estimate*

For every protein–MCSS minimum complex the binding free energy is approximated by use of the following equation:

$$\Delta G_{binding} = \Delta E^{fragm} + \Delta E_{vdW}^{interm} + \Delta G_{elect}^{interm} + \Delta G_{elect,desolv}^{protein} + k\Delta G_{elect,desolv}^{fragm} + \Delta G_{np}^{complex} \quad (2)$$

The first term on the right side represents the difference in energy of the fragment upon binding:

$$\Delta E^{fragm} = \Delta E_{bonding}^{fragm} + \Delta E_{vdW}^{fragm} + \Delta E_{elect}^{fragm} \quad (3)$$

and is a sum of the bonding (bonds, angles, and torsions) energy terms ($\Delta E_{bonding}^{fragm}$), the van der Waals interaction ($\Delta E_{vdW}^{fragm}$), and the vacuum Coulombic energy between atoms of the MCSS group ($\Delta E_{elect}^{fragm}$). The CHARMM force-field is used to compute $\Delta E^{fragm}$ and $\Delta E_{vdW}^{interm}$, which is the van der Waals interaction energy between the protein and fragment. The solvation free energy is expressed as a sum of separate electrostatic and nonpolar contributions [19,49]. The electrostatic contribution to the free energy of binding consists of shielded intermolecular interaction ($\Delta G_{elect}^{interm}$, see Eq. 1), protein desolvation ($\Delta G_{elect,desolv}^{protein}$), and desolvation of the fragment ($\Delta G_{elect,desolv}^{fragm}$). These energy values are calculated by solving the finite-difference LPB equation [48]. A scaling factor (k) for the electrostatic desolvation of the fragment is introduced to take into account the fact that when a fragment is part of a larger ligand, its desolvation is smaller. For all MCSS minima, a value of $k = 0.4$ was used. This is based on the comparison of the electrostatic desolvation energy upon binding of NMA and the dipeptide *N*-acyl-Gly-NH-$CH_3$ molecules to a macromolecular target (Caflisch, unpublished results).

On the basis of experimental data on alkane–water partition coefficients [22], the nonpolar contribution to the free energy of binding ($\Delta G_{np}^{complex}$) is assumed to be proportional to the loss in solvent-accessible surface area (A) [23]:

$$\Delta G_{np}^{complex} = \gamma \left(A^{complex} - (A_{isolated}^{protein} + A_{isolated}^{fragm})\right) \quad (4)$$

The constant $\gamma$, which may be interpreted as the vacuum–water microscopic surface tension, is assigned a value of 0.025 kcal/mol Å$^2$ [50]. For the structure of the complex and its isolated components, the total area, i.e., area of polar and nonpolar groups [6] is computed by the CHARMM implementation of the Lee–Richards algorithm [23] by using a probe sphere of 1.4-Å radius.

*Lists of bonding fragment pairs and overlapping fragment pairs*

The user has to specify for each functional group type which atoms are to be used for connection to other fragments. These will be called 'linkage atoms' henceforth. For each linkage atom, CCLD generates a set of possible linkage points (Fig. 3), i.e., points which will be used to determine the position and orientation of the link. All possible pairs of minimized positions are then analyzed and added to the list of bonding fragment pairs if they can be linked; otherwise, if two fragments have bad contacts they are added to the list of overlapping fragment pairs. A pair of bonding fragments may be connected by a linker unit, by a single covalent bond (1-bond), or by fusing two overlapping atoms belonging to different fragments (0-bond). The linker units are small, since their function is to optimally connect two fragments without adding considerably to the molecular weight. The following linker elements have been so far implemented: Keto and methylene (2-bond), amide and ethylene (3-bond). The user is free to choose minimal and maximal values for the distance (d) between linkage atoms for each connection type. In the application to thrombin the following values in Å were used: $d < 0.43$, 0-bond; $1.2 < d < 1.8$, 1-bond; $2.2 < d < 2.7$, 2-bond; $3.6 < d < 4.0$, 3-bond. More permissive values produce ligands with a larger degree of distortion. For 0-bonds and 1-bonds, the bonding angles are checked and the linkage points are not used. For 2-bonds, whenever the distance between linkage atom $a_1$ on fragment $f_1$ and linkage atom $a_2$ on fragment $f_2$ is in the user-specified range, the distance between all pairs of $a_1$ and $a_2$ linkage points is calculated; if it is smaller than a given cutoff value (1.4 Å in the present application), angle checking is performed and the two linkage points which result in the best geometry are used to determine the position of the additional carbon atom for the 2-bond between fragments $f_1$ and $f_2$. In addition, the position of the oxygen atom for an eventual keto-linker is determined. The Coulombic energy between the carbonyl group of the keto moiety (partial charges of +0.55e and −0.55e for the carbon and oxygen atom, respectively, as in the CHARMM PARAM19 force-field) and the protein atoms is then calculated with a constant dielectric value of 1.0 and a cutoff of 9.0 Å. A keto-link is preferred to a methylene group if its Coulombic interaction energy

TABLE 2
MINIMA OF NONPOLAR GROUPS

| Rank[a] | Rank[b] | Intermolecular vdWaals[c] | Desolvation | | $\Delta G_{binding}$ [f] | MCSS rank[g] | Site |
|---|---|---|---|---|---|---|---|
| | | | Nonpolar[d] | Elect[e] | | | |
| **Cyclopentane** | | | | | | | |
| **1** | 39 | −6.9 | −8.0 | 5.5 | −9.3 | 12 | S2 |
| 2 | 46 | −6.6 | −7.8 | 5.2 | −9.1 | 17 | S2 |
| 3 | 60 | −6.4 | −7.6 | 5.2 | −8.8 | 20 | S2 |
| **4** | 64 | −4.3 | −7.8 | 3.3 | −8.8 | 33 | S3–S2 |
| **5** | 91 | −7.1 | −7.7 | 6.4 | −8.4 | 9 | S3 |
| **6** | 109 | −2.2 | −6.1 | 0.2 | −8.1 | 46 | $Trp^{60D}$ |
| 7 | 111 | −2.1 | −6.1 | 0.2 | −8.0 | 49 | $Trp^{60D}$ |
| 8 | 112 | −2.2 | −6.2 | 0.3 | −8.0 | 48 | $Trp^{60D}$ |
| 9 | 115 | −2.2 | −6.2 | 0.5 | −8.0 | 45 | $Trp^{60D}$ |
| 10 | 116 | −2.2 | −6.1 | 0.4 | −7.9 | 47 | $Trp^{60D}$ |
| 30 | 563 | −8.5 | −7.7 | 12.9 | −3.3 | 3 | $Leu^{40}$ |
| 35 | 700 | −8.6 | −7.7 | 14.7 | −1.6 | 2 | $Leu^{40}$ |
| 38 | 756 | −8.4 | −7.7 | 15.2 | −0.9 | 4 | $Leu^{40}$ |
| **47** | 1174 | −9.0 | −8.0 | 26.1 | 9.2 | 1 | S1 |
| **48** | 1215 | −5.7 | −7.0 | 23.7 | 11.0 | 26 | S1' |
| **49** | 1260 | −7.0 | −8.1 | 30.9 | 15.9 | 11 | S1' |
| **Benzene** | | | | | | | |
| **1** | 22 | −5.0 | −7.5 | 2.7 | −9.9 | 28 | S3–S2 |
| **2** | 30 | −7.3 | −7.9 | 5.7 | −9.5 | 11 | S2 |
| **3** | 38 | −8.0 | −7.5 | 6.2 | −9.4 | 6 | S3 |
| 4 | 53 | −7.6 | −6.5 | 5.2 | −8.9 | 8 | $Trp^{148}Ala$ |
| **5** | 56 | −5.3 | −4.2 | 0.6 | −8.9 | 23 | $Trp^{60D}$ |
| 6 | 58 | −5.3 | −4.2 | 0.7 | −8.9 | 24 | $Trp^{60D}$ |
| 7 | 117 | −8.4 | −7.4 | 7.8 | −7.9 | 4 | $Trp^{148}Ala$ |
| 8 | 153 | −5.0 | −6.7 | 4.3 | −7.4 | 26 | S3–S2 |
| 9 | 187 | −5.0 | −4.5 | 2.7 | −6.9 | 25 | $Trp^{60D}$ $C^{\alpha}$, $C^{\beta}$ |
| 10 | 200 | −7.2 | −7.3 | 7.7 | −6.8 | 13 | $Glu^{192}Ala$ |
| 18 | 531 | −9.7 | −7.5 | 13.7 | −3.6 | 3 | $Leu^{40}$ |
| 20 | 558 | −7.4 | −7.3 | 11.4 | −3.3 | 9 | $Leu^{40}$ |
| 26 | 1020 | −11.4 | −7.7 | 23.5 | 4.4 | 2 | S1 |
| **28** | 1045 | −11.4 | −7.8 | 24.0 | 4.9 | 1 | S1 |
| **31** | 1250 | −7.7 | −7.8 | 29.4 | 14.0 | 7 | S1' |
| **32** | 1283 | −6.3 | −7.3 | 32.3 | 18.7 | 18 | S1' |

Energy values in kcal/mol are listed for the 10 cyclopentane and 10 benzene minima with the lowest binding free energy and for other minima discussed in the text.

[a] Ranked among the minima of the same functional group type according to binding free energy. Minima with rank in bold are shown in Figs. 4a and b.

[b] Ranked among all minima according to binding free energy.

[c] Calculated with CHARMM.

[d] Calculated by use of Eq. 4.

[e] Calculated by numerical solution of the LPB equation as shown in Fig. 1a.

[f] Calculated by use of Eq. 2, i.e., the sum of columns 3 to 5.

[g] Ranked among the minima of the same functional group type according to total CHARMM energy, i.e., the sum of intermolecular and intraligand energies.

with the protein is more favorable than −1.5 kcal/mol. This value is somewhat higher than the free energy of solvation of 2-propanone (experimental value of −3.85 kcal/mol [51] and a continuum dielectric value of −5.32 kcal/mol). As a 2-bond linker unit, a keto group is preferred over a methylene group, because it often results in an additional intermolecular hydrogen bond. In addition, if the linkage atom is an $sp^3$ carbon in a cycloalkane and the CO group is close to the plane of the ring, the $sp^3$ carbon is automatically converted into an $sp^2$ nitrogen, i.e., CCLD produces a cyclic secondary amide connection instead of a keto-linker unit, the former generally being of easier synthetic accessibility.

For the 3-bond, a procedure similar to that of the 2-bond is used. Angle checking is performed whenever a linkage point of $a_1$ is between 1.0 Å and 1.8 Å of a linkage point of $a_2$. An amide link is used if its Coulombic interaction energy with the protein is more favorable than

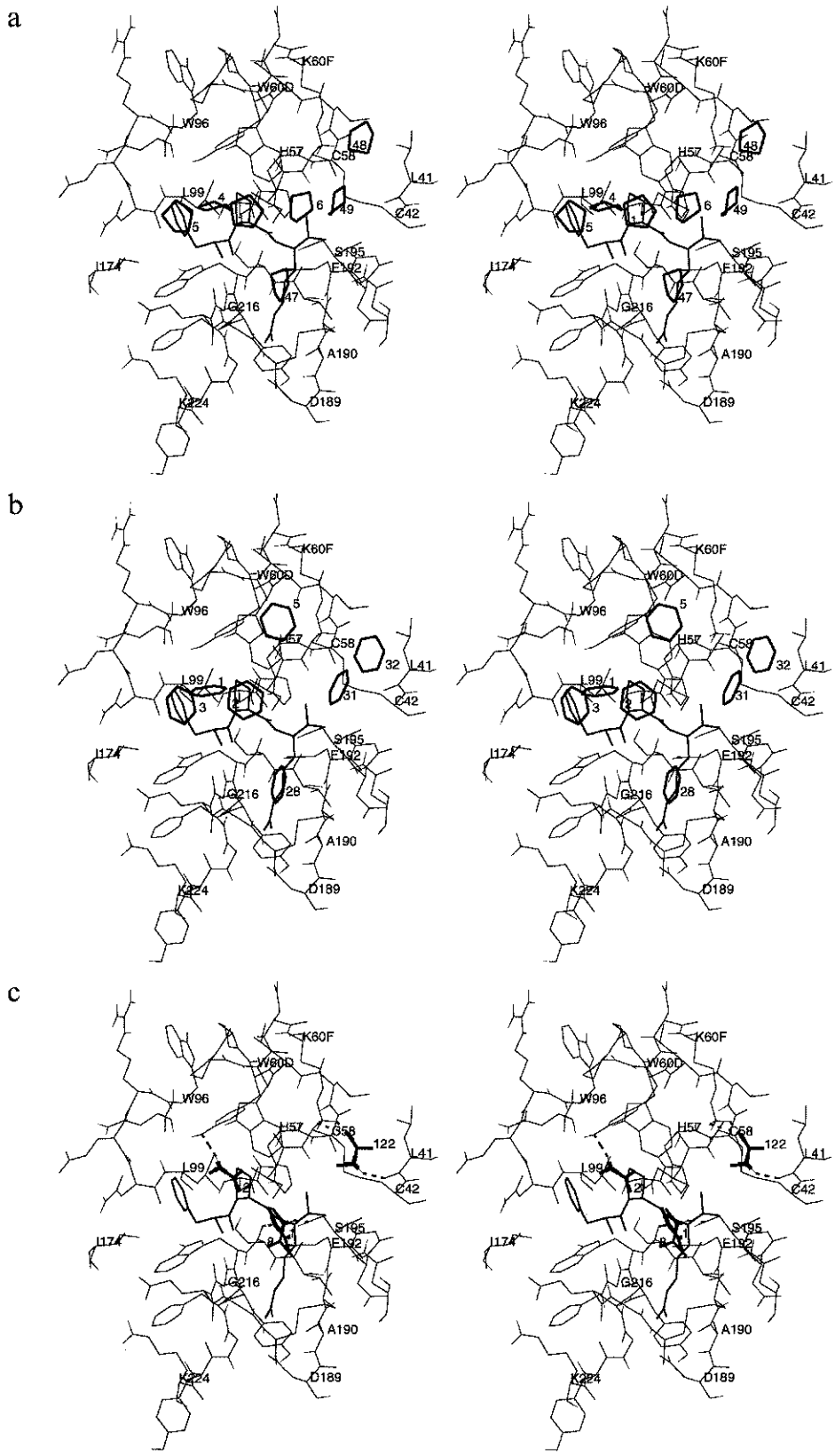


Fig. 4. Stereoviews of the MCSS minima (thick lines for heavy atoms and thin lines for polar hydrogens) in the thrombin active site (thin lines). The PPACK inhibitor is also shown (medium lines), though it was removed during the MCSS procedure. Some $C^{\alpha}$ atoms of thrombin are labeled. In the chymotrypsin numbering of Bode and co-workers [25], $Gly^{219}$ follows directly after $-Gly^{216}-Glu^{217}-$, i.e., there is no residue with number 218. The MCSS minima are labeled according to their binding free energy rank within minima of the same type. Hydrogen bonds between protein and MCSS minima are shown as dashed lines; (a) cyclopentane; (b) benzene; (c) *N*-methylacetamide (NMA).

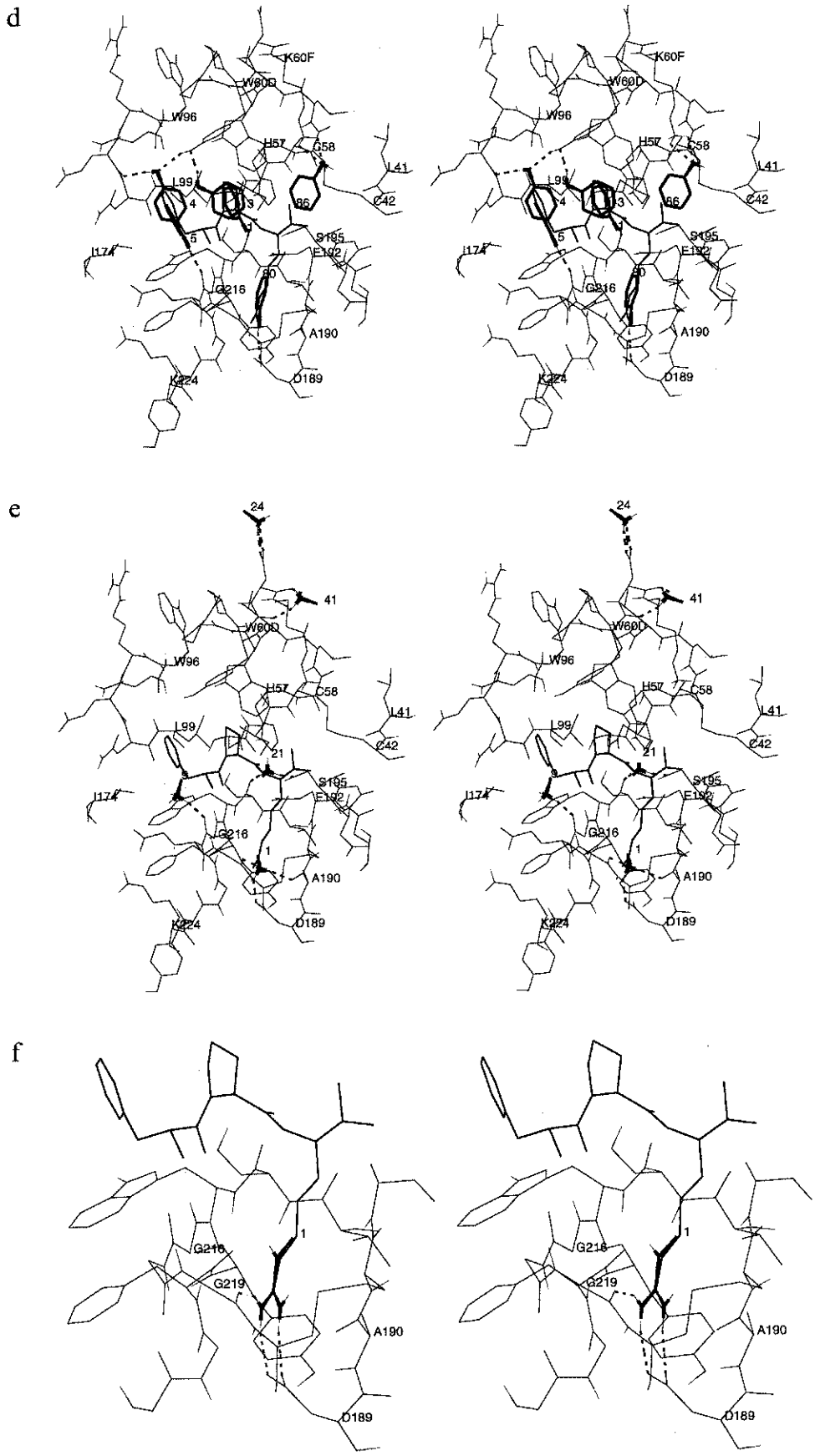


Fig. 4. (d) phenol; (e) methylammonium; (f) methylguanidinium minimum 1.

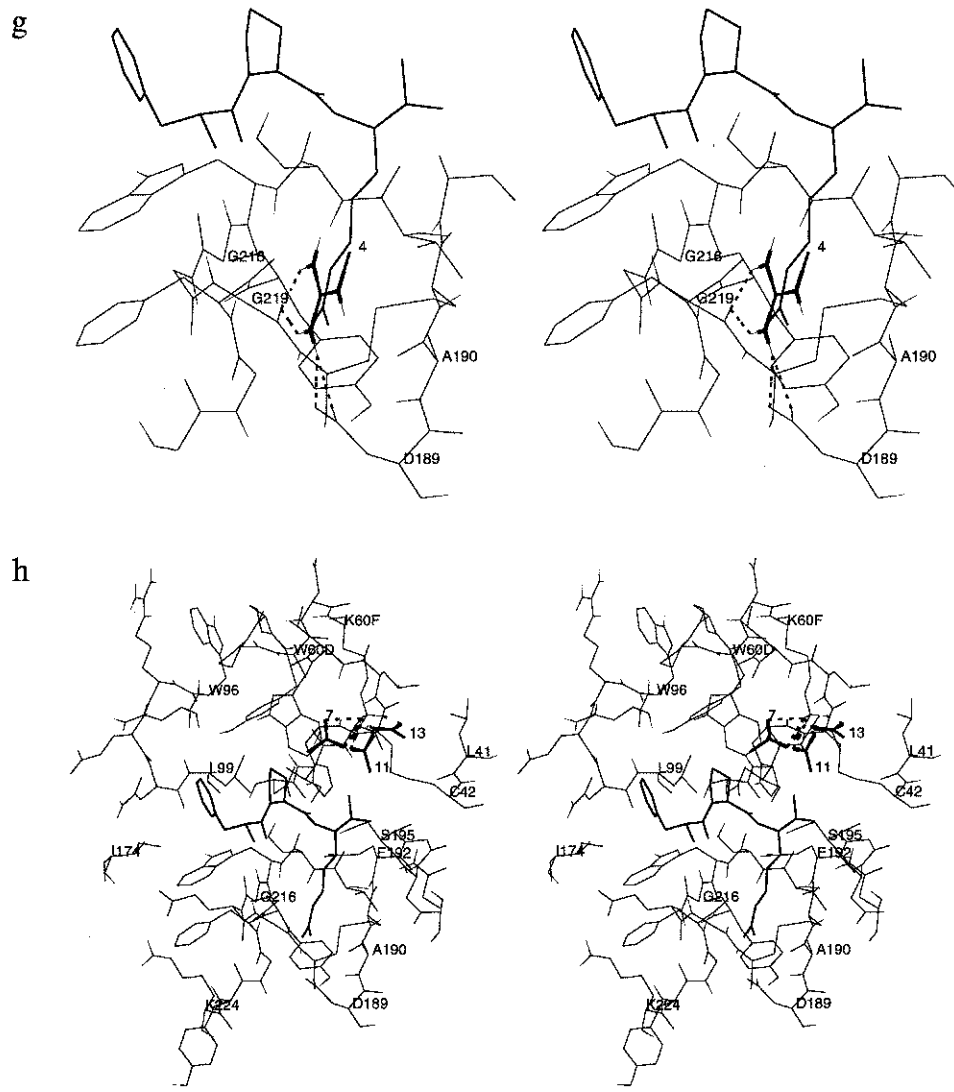


Fig. 4. (g) methylguanidinium minimum 4; (h) acetate ion.

–3.0 kcal/mol. Although this value is higher than the free energy of solvation of NMA (experimental value of –9.71 kcal/mol [51] and continuum dielectric value of –9.07 kcal/mol) it is chosen because an amide linker is more rigid and in most cases easier to synthesize than an ethylene linker. Furthermore, amide linkers are often involved in intermolecular hydrogen bonds.

The list of bonding fragment pairs and the list of overlapping fragment pairs are created only once before entering the combinatorial search (Fig. 2). The use of these lists results in a significant increase in the speed with which ligands are generated.

*Combinatorial ligand generation*

Starting from the MCSS minimum with the most favorable binding free energy, the ligand-generation algorithm proceeds in an iterative and exhaustive way by linking an additional fragment to the actual construct. Such an 'elongation' step is very fast, since it is sufficient to check that the new fragment may be connected to one of the fragments in the actual construct (by looking in the list of bonding fragment pairs), and that the new fragment does not overlap with any of the fragments in the actual construct (Fig. 2). The combinatorial explosion problem is kept under control by pruning, which is performed according to the average value of the free energy of binding of the fragments. Whenever the addition of a fragment to the growing ligand results in an average value of the binding free energy higher than a user-specified threshold, the construct is reduced by deletion of the latest added fragment (Fig. 2). A ligand with an energy below the threshold is saved if it is larger than a user-specified minimal size and if it is not a substructure of a ligand found previously. The energy of the linker elements is not taken into account, except for the two following cases: (i) firstly, methylene and ethylene linker

units are penalized by 3.0 and 4.5 kcal/mol, respectively, to bias the combinatorial selection algorithm towards ligands with a small number of flexible dihedrals; (ii) secondly, if the vacuum electrostatic interaction energy of keto- and amide-linker units with the protein is less favorable than their electrostatic solvation free energy, then these are penalized by the difference between the two energy values. For the solvation free energy, the values obtained by solution of the LPB are used, i.e., −5.32 kcal/mol and −9.07 kcal/mol for 2-propanone and NMA, respectively.

*Clustering of ligands*

After exiting the combinatorial search procedure, the ligands are sorted according to the sum of the free energy of binding of the fragments and eventual linker penalties divided by the number of fragments. A simple clustering procedure based on the degree of similarity between candidate ligands is then performed. Since these are coded as strings of integers, with each integer representing an MCSS minimum, an efficient procedure is implemented to check if two ligands have more than a user-specified percentage (p) of the fragments in common. The user is free to select the value of p. The ligand with the lowest average free energy is selected as representative of the first cluster and all the ligands with p% or more fragments in common with this representative are assigned exclusively to the first cluster. The procedure then iterates by selecting the next ligand, which is not already a member of any cluster, as the representative of a new cluster, until all ligands are either representatives or members of a cluster. The user can specify the number of ligands, whose coordinates have to be printed out in any case, even if they are not representative of a cluster. Otherwise, CCLD prints out only the coordinate files of the cluster representatives; in addition, an output file which contains information on each cluster is generated. This is particularly useful if the user wants to analyze a set of compounds which have one or more common binding motifs.

*Computation time*

All calculations were performed on SGI computers with R4400 central processor units (CPU). Each MCSS run (minimization of 10 000 replicas and calculation of the loss in solvent-accessible surface area for the minimized positions) required between 5 h (for methanol) and 30 h (for methylguanidinium) of CPU time. The evaluation of the three terms of the continuum electrostatic energy took about 17 min of CPU time for each thrombin–MCSS minimum complex. For the nonpolar fragments, the calculation of the electrostatic desolvation of the protein took about 7 min of CPU time. Hence, a total of 3.5 days on a four-processor SGI Challenge were required for the evaluation of the electrostatic contribution to the binding free energy of the 1314 MCSS minima. A CCLD run requires from 2–3 min (for 200 to 300 fragments) to less than 1 h CPU time (for about 1000 fragments).

## Results

*Thrombin functionality maps*

In presenting the MCSS results, both structural and energetic properties of the minima are analyzed. In addition, a detailed comparison of the functional group sites with the interaction patterns of known inhibitors is given for nonpolar, polar, and charged fragments.

*Nonpolar group minima*

Propane, cyclopentane, cyclohexane, and benzene minima are distributed over most of the apolar regions of the thrombin active site. Since their functionality maps and energy values are similar, only the minimized positions of cyclopentane and benzene are analyzed in detail.

*Cyclopentane* The energy values of the ten cyclopentane minima with the lowest free energy of binding are listed in Table 2; minima 1, and 4 to 6 are shown in Fig. 4a. Minima 1 to 3 overlap the PPACK proline side chain, minimum 4 is positioned between S3 and S2, and minimum 5 is close to the aromatic ring of the PPACK phenylalanine. Minima 6 to 10 are on the surface of thrombin and interact only with the six-membered ring of $Trp^{60D}$; they are positioned on the indole face opposite to the S2 pocket. This is consistent with the position of the hydroxyphenyl substituent of cyclotheonamide A (CtA) in its complex with thrombin [52]. Minima 1 to 5 have good van der Waals interactions with the hydrophobic S3 and S2 pockets of thrombin (values ranging from −7.1 to −4.3 kcal/mol) and pay a small penalty for the electrostatic desolvation of the protein (values ranging from 3.3 to 6.4 kcal/mol). Since minima 6 to 10 interact only with $Trp^{60D}$, they have a weaker van der Waals energy (from −2.2 to −2.1 kcal/mol) but have a negligible electrostatic desolvation penalty (from 0.2 to 0.5 kcal/mol). Also, minima 1 to 5 are more buried, hence their nonpolar desolvation term (from −8.0 to −7.6 kcal/mol) is more favorable than that of minima 6 to 10 (from −6.2 to −6.1 kcal/mol).

The MCSS ranking is different from the free energy ranking, since it does not take into account desolvation effects; it is based on the sum of the CHARMM intermolecular and intraligand energies. The latter are negligible for the nonpolar groups used in this study. Although the MCSS ranking is less significant, it is useful to analyze some of the minima with the lowest intermolecular energy and compare them with the most favorable free energy minima. The cyclopentane minimum with the lowest CHARMM energy (free energy minimum 47, see Table 2) overlaps the alkyl part of the arginine side chain

of PPACK in S1 (Fig. 4a). The penalty for the electrostatic desolvation of the protein is 26.1 kcal/mol, since it buries the solvent-accessible side of the peptide groups of residues 191-192 and 215-216, and partially buries the $Asp^{189}$ carboxyl group located at the bottom of the S1 pocket. Minima 2 to 4 (CHARMM ranking) make strong van der Waals interactions with the $Leu^{40}$ side chain in S2' (not shown), but partially desolvate the side chains of $Arg^{73}$ and $Gln^{151}$. According to binding free energy they rank 35, 30, and 38, respectively (Table 2). In the S1' pocket, the minimized positions of cyclopentane with the 11th and 26th lowest CHARMM energy are involved in favorable van der Waals interactions with the 42–58 disulfide bridge and the $Leu^{41}$ side chain, respectively (Fig. 4a). Both of these minima bury part of the primary amino group of the $Lys^{60F}$ side chain. Due to the high electrostatic desolvation penalty of 30.9 kcal/mol (CHARMM minimum 11) and 23.7 kcal/mol (CHARMM minimum 26), they have the worst binding free energy of the 49 cyclopentane minima (Table 2).

*Benzene* The energy values of the ten benzene minima with the lowest free energy of binding are listed in Table 2; benzene minima 1 to 3, and 5 are shown in Fig. 4b. The three best minima are in the S3 and S2 pockets and have strong van der Waals interactions (values ranging from −8.0 to −5.0 kcal/mol) and minor electrostatic desolvation of the protein (from 2.7 to 6.2 kcal/mol). Minimum 3 is close to the aromatic ring of the PPACK phenylalanine side chain. Minima 5 and 6 are involved in a face-to-face aromatic interaction with the solvent-exposed face of the indole ring of $Trp^{60D}$; their van der Waals interaction with the protein is −5.3 kcal/mol and the electrostatic protein-desolvation penalty is negligible (from 0.6 to 0.7 kcal/mol). As a basis of comparison, the hydroxyphenyl substituent of CtA is involved in edge-to-face rather than face-to-face interactions with the indole of $Trp^{60D}$, probably because of its intramolecular edge-to-face arrangement with the phenyl substituent [52]. Minima 4 and 7 occupy the position of the indole ring of $Trp^{148}$ and minimum 10 is very close to $Glu^{192}$, both of which side chains were mutated to alanine for the MCSS runs (see Methods). The two benzene minima with the lowest CHARMM energy are sandwiched between the amide groups of residues 215-216 and 191-192 in S1 and occupy the same position as the aromatic ring of benzamidine in the NAPAP–thrombin complex [26]. Similar results were found for methylbenzene in a previous work [13]. They partially desolvate the $Asp^{189}$ side chain; hence, their electrostatic contribution to protein desolvation is high (24.0 and 23.5 kcal/mol). This is not compensated by favorable electrostatic interactions between the aromatic ring and the amide planes, since the former does not bear any partial charge in the PARAM19 force-field. Hence, they have an unfavorable total free energy of binding (4.9 and 4.4 kcal/mol). The MCSS minima of benzene with the highest binding free energy, 31 and 32 (7 and 18 according to the CHARMM energy, respectively), bury part of the amino group of the $Lys^{60F}$ side chain (Fig. 4b).

*Cyclohexane* Minima 1 and 3 occupy the S2 and S3 pockets of thrombin, respectively, while minima 2, 4, and 5 are on the surface and interact only with the six-membered ring of $Trp^{60D}$, in the same orientation as the cyclopentane minima 6 to 10. Their individual energy contributions are analogous to those of the corresponding cyclopentane and benzene minima.

*Propane* Minima 1 and 9 occupy the S3 pocket, while minimized positions 2 to 8 are placed in the S2 subsite and minimum 10 is positioned between S3 and S2. Propane minimum 11 is close to the six-membered ring of $Trp^{60D}$ and matches the $C^{\alpha}$, $C^{\beta}$, and $C^{\gamma}$ atoms of the vinylogous tyrosine unit of CtA [52].

From the analysis of the thrombin functionality maps of the nonpolar groups it is evident that hydrophobic moieties prefer to bind to the S3 and S2 pockets. The solvent-exposed face of the $Trp^{60D}$ indole is another favorable site, though the intermolecular van der Waals interactions are much smaller. Binding to the S2' region is favored by interactions with the $Leu^{40}$ side chain, but implies a desolvation penalty because of the burial of part of the $Arg^{73}$ guanidinium and/or the $Gln^{151}$ side chain. The latter might be an artifact of the rigid protein structure used in the minimization, since the side chains of $Arg^{73}$ and $Gln^{151}$ are flexible enough to displace their polar groups towards a more exposed region. Binding to the neighbouring $Leu^{41}$ side chain in S1' is highly unfavorable because of the concomitant desolvation of $Lys^{60F}$.

*Polar group minima*

Polar neutral groups are scattered over all hydrophilic regions of the active site. The minima of *N*-methylacetamide and phenol will be discussed in detail, while those of methanol, 2-propanone, *N*,*N*-dimethylacetamide, pyrrole, and imidazole will be analyzed only briefly.

*N-methylacetamide (NMA)* The NH in the NMA minimum with the lowest free energy of binding is involved in the same hydrogen bond as the backbone NH of the arginine residue in PPACK, i.e., it donates to the carbonyl oxygen of residue 214 (Fig. 4c). The distance between the nitrogen atom in the NMA minimum 1 and the main-chain N atom of arginine in PPACK is 0.51 Å. Since the CO group of the NMA minimum 1 is not engaged in hydrogen bonds, most of the −4.5 kcal/mol of electrostatic interaction energy (Table 3) originates from the NH–214CO hydrogen bond. NMA minimum 2 occupies the S2 pocket and donates to the side-chain O atom of $Tyr^{60A}$. Since the geometry of this intermolecular hydrogen bond is not ideal, it has a weaker intermolecular interaction with the protein than minimum 1. On the other hand, it pays a smaller penalty in electrostatic desolvation of the protein (3.0 kcal/mol instead of 8.8

TABLE 3
MINIMA OF POLAR GROUPS

| Rank[a] | Rank[b] | Strain[c] | Intermolecular | | Desolvation | | | $\Delta G_{binding}$ [i] | MCSS rank[j] | Site and H-bond partners |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | vdWaals[d] | Elect[e] | Nonpolar[f] | Electrostatic | | | | |
| | | | | | | Protein[g] | Ligand[h] | | | |
| **NMA** | | | | | | | | | | |
| 1 | 24 | 0.0 | −8.2 | −4.5 | −8.2 | 8.8 | 2.3 | −9.8 | 38 | S1; 214CO |
| **2** | 26 | 0.0 | −5.8 | −2.0 | −6.4 | 3.0 | 1.6 | −9.6 | 88 | S2; $Tyr^{60A}$ $O^{\eta}$ |
| 3 | 28 | 0.0 | −6.7 | −5.3 | −7.2 | 7.6 | 2.0 | −9.5 | 26 | 148NH |
| 4 | 52 | 0.0 | −8.2 | −3.7 | −7.1 | 8.0 | 2.1 | −9.0 | 32 | 147NH |
| 5 | 54 | 0.0 | −7.0 | −5.0 | −7.3 | 8.3 | 2.1 | −8.9 | 29 | 148NH |
| 6 | 59 | 0.0 | −5.0 | −2.1 | −6.2 | 3.3 | 1.3 | −8.8 | 99 | S2; $Tyr^{60A}$ $O^{\eta}$ |
| 7 | 61 | 0.0 | −6.7 | −5.6 | −7.1 | 8.3 | 2.3 | −8.8 | 30 | 148NH |
| **8** | 74 | 0.0 | −8.1 | −3.9 | −8.2 | 9.6 | 2.0 | −8.6 | 41 | S1; $Ser^{195}$ $O^{\gamma}$ |
| 9 | 79 | 0.0 | −6.8 | −4.9 | −7.1 | 8.1 | 2.3 | −8.5 | 33 | 148NH |
| 10 | 85 | 0.0 | −8.0 | −3.9 | −8.2 | 9.4 | 2.2 | −8.4 | 36 | S1; 214CO |
| 98 | 1004 | 0.0 | −4.0 | −8.7 | −6.2 | 21.0 | 1.9 | 3.9 | 1 | surface; $Arg^{173}$ $N^{\eta}$ and $N^{\varepsilon}$, $Glu^{192}$ $O^{\varepsilon 1}$ |
| 99 | 1012 | 0.0 | −3.9 | −8.1 | −6.1 | 20.4 | 1.8 | 4.1 | 2 | surface; $Arg^{173}$ $N^{\eta}$ and $N^{\varepsilon}$, $Glu^{192}$ $O^{\varepsilon 1}$ |
| **122** | 1270 | 0.6 | −2.1 | −11.8 | −7.4 | 35.4 | 2.4 | 17.2 | 3 | S1'; 41CO, $Lys^{60F}$ $N^{\zeta}$ |
| **Phenol** | | | | | | | | | | |
| **1** | 3 | 0.2 | −7.1 | −4.0 | −8.0 | 5.5 | 1.9 | −11.4 | 41 | S2; 214CO |
| 2 | 7 | 0.3 | −8.0 | −5.9 | −7.4 | 7.6 | 2.6 | −11.0 | 12 | 145CO, 147NH |
| **3** | 10 | 0.1 | −7.2 | −2.4 | −7.9 | 5.3 | 1.5 | −10.6 | 48 | S2; $Tyr^{60A}$ $O^{\eta}$ |
| **4** | 13 | 0.1 | −9.1 | −3.2 | −7.8 | 7.5 | 2.0 | −10.4 | 20 | S3; 97CO, $Tyr^{60A}$ OH |
| **5** | 15 | 0.1 | −5.3 | −3.7 | −7.2 | 4.2 | 1.6 | −10.3 | 60 | S3; 216CO |
| 6 | 16 | 0.1 | −8.7 | −3.4 | −7.8 | 7.5 | 2.1 | −10.3 | 19 | S3; 97CO, $Tyr^{60A}$ OH |
| 7 | 23 | 0.1 | −7.9 | −2.0 | −8.0 | 6.5 | 1.5 | −9.9 | 44 | S3; $Tyr^{60A}$ $O^{\eta}$ |
| 8 | 29 | 0.3 | −7.7 | −4.7 | −7.7 | 8.3 | 2.0 | −9.5 | 24 | 148NH, $Thr^{147}$ $O^{\gamma 1}$ |
| 9 | 35 | 0.4 | −8.1 | −2.2 | −8.2 | 6.9 | 1.8 | −9.4 | 64 | S2; $His^{57}$ $N^{\varepsilon 2}$ |
| 10 | 40 | 0.1 | −7.3 | −2.8 | −8.2 | 7.1 | 1.9 | −9.2 | 67 | S2; $His^{57}$ $N^{\varepsilon 2}$ |
| **80** | 839 | 0.3 | −12.5 | −5.1 | −7.9 | 22.9 | 2.7 | 0.4 | 1 | S1; $Asp^{189}$ $O^{\delta 1}$ |
| **86** | 1000 | 0.3 | −8.2 | −6.8 | −8.3 | 24.6 | 2.2 | 3.8 | 2 | S1'; $Lys^{60F}$ $N^{\zeta}$ |
| 103 | 1216 | 0.6 | −5.8 | −8.9 | −8.0 | 31.4 | 1.6 | 11.1 | 3 | S1'; $Lys^{60F}$ $N^{\zeta}$ |

Energy values in kcal/mol are listed for the 10 MCSS minima of NMA and phenol with the lowest binding free energy and for the three with the lowest CHARMM energy.

[a] Ranked among the minima of the same functional group type according to binding free energy. Minima with rank in bold are shown in Figs. 4c and d.

[b] Ranked among all minima according to binding free energy.

[c] Sum of intraligand energy terms is calculated with CHARMM (Eq. 3).

[d] Calculated with CHARMM.

[e] Calculated by numerical solution of the LPB equation as explained in the text (Eq. 1) and in Fig. 1b.

[f] Calculated by use of Eq. 4.

[g] Calculated as shown in Fig. 1a.

[h] Calculated as shown in Fig. 1c. Values are scaled by k = 0.4 (see the text following Eq. 2 for the meaning of k).

[i] Calculated by use of Eq. 2, i.e., the sum of columns 3 to 8.

[j] Ranked among the minima of the same functional group type according to total CHARMM energy, i.e., the sum of intermolecular and intraligand energies.

kcal/mol for minimum 1). Of its two methyl groups, the one close to the carbonyl group is buried in the S2 pocket, while the N-methyl group is at the interface between the S3 and S2 pockets. NMA minimum 8 is close to minimum 1; its NH group donates to the side-chain O atom of $Ser^{195}$ instead of the 214CO (Fig. 4c). Minima 3, 4, 5, 7, and 9 are close to the autolysis loop and their position may have been affected by the $Trp^{148}$-to-Ala mutation. For each accessible main-chain polar group in the thrombin active site there are one or more NMA minima involved in hydrogen bonds with a favorable binding free energy. Minima 20, 21, 41 and 42 (not shown) donate to the CO group of $Gly^{216}$, minima 41 and 42 accept also from the NH group of $Gly^{219}$, and minima 58 and 81 donate to the CO group of residues 40 and 41, respectively. Minimum 65 overlaps the Phe-Pro amide of PPACK and acts as an acceptor for the NH group of $Gly^{216}$. The three NMA minima with the lowest

CHARMM energy bind to charged side chains on the surface of thrombin. They are ranked as 98, 99, and 122, respectively, according to the increasing binding free energy (Table 3). This is a consequence of their highly unfavorable electrostatic contribution to protein desolvation (values ranging from 21.0 to 35.4 kcal/mol).

*Phenol* Minima 1 and 3 have their aromatic ring in S2, while minima 4 and 5 occupy the S3 pocket (Fig. 4d and Table 3). The phenol minimum with the lowest free energy of binding acts as donor in a hydrogen bond with the main-chain CO group of residue 214, while minimum 3 donates to the hydroxyl O atom of $Tyr^{60A}$ (Fig. 4d). Minimum 4 donates to the main-chain CO group of residue 97 and accepts from the hydroxyl group of $Tyr^{60A}$, while minimum 5 acts as a donor to the main-chain CO group of $Gly^{216}$. These minima have strong van der Waals and electrostatic interactions with the protein and the electrostatic desolvation penalty of the protein is small (values ranging from 4.2 to 7.6 kcal/mol). Hence, they rank among the best 15 of the 1314 minima found in this work. On the other hand, the phenol minimum with the lowest CHARMM energy (no. 80 according to free energy ranking) makes a strong hydrogen bond with the $Asp^{189}$ side chain in S1 (–5.1 kcal/mol of electrostatic interaction energy) and is involved in very favorable van der Waals interactions with the S1 atoms (–12.5 kcal/mol) but has to pay a significant electrostatic desolvation penalty (22.9 kcal/mol for the protein and 2.7 kcal/mol for the phenol group). The same holds for the phenol minima with the second- and third-best CHARMM energy (no. 86 and 103, respectively), which accept from the side chain of $Lys^{60F}$ in S1'.

*Methanol* The minima of this small functional group are scattered over most of the active site. Some of them are almost completely buried in small cavities, e.g., at the bottom of the S1 pocket close to the main-chain NH group of $Glu^{217}$. Others have the methyl group exposed and might be linked to a larger molecular structure. These participate in hydrogen bonds with polar groups of thrombin which are used as hydrogen-bond partners by known active-site inhibitors. Numbers 11 and 24 donate to the carbonyl oxygen of residue 214 (binding free energy of –5.5 and –3.8 kcal/mol, respectively). Minimized positions 18 and 26 make two hydrogen bonds with the main-chain polar groups of $Gly^{216}$ (binding free energy of –4.4 and –3.6 kcal/mol, respectively).

*2-Propanone and N,N-dimethylacetamide (NDMA)* These groups have minima close to the autolysis loop and minima which accept from the main-chain NH group of $Gly^{216}$ and $Gly^{219}$. In addition, as for every polar group with a hydrogen-bond acceptor, there is a cluster of minima interacting with the $Lys^{60F}$ side chain with an unfavorable free energy of binding because of the electrostatic desolvation penalty of the protein.

*Pyrrole and imidazole* These have similar maps scattered around accessible hydrogen-bond acceptors of the thrombin active site. There are minimized positions of these groups in the S3 and S2 pockets and also minima donating to the $Ser^{195}$ hydroxyl oxygen, the main-chain CO group of residues 214 and 216 in S1, and 40 and 41 (only pyrrole minima) in S2'.

It is impossible to draw general conclusions about preferential thrombin sites for polar groups. These will depend on the particular arrangement of charges and on the radii of the atoms in the hydrophilic group. Also, the optimal position and orientation of a group having both polar and aromatic or hydrophobic character might be a compromise between strong hydrogen bonds and good van der Waals interactions (e.g., the phenol minima with the lowest free energy, Fig. 4d). It is important to note that there are several polar groups on the thrombin main chain that are involved in strong hydrogen bonds with minima of hydrophilic functional groups (favorable binding free energy). These are: 214CO, 216NH, 216CO, and 219NH in S1; 193NH and 195NH in the oxyanion hole; 41CO in S1'; 40CO in S2'; and 147NH and 148NH on the autolysis loop, whose exposure is dependent on crystallization conditions and inhibitor type. On the other hand, the results obtained in this study indicate that a charged side chain, which is partially or completely exposed to solvent, may not be an ideal partner for a polar group, because of unfavorable electrostatic desolvation effects.

*Charged group minima*

These tend to cluster close to side chains of opposite charge. They may also be found in the vicinity of polar and neutral groups, particularly if they can make more than one hydrogen bond.

*Methylammonium* The three minima with the lowest free energy of binding have the most favorable CHARMM energy and are located in the S1 pocket (Table 4). Minimum 1 is involved in hydrogen bonds with the $Asp^{189}$ $O^{\delta 2}$ atom (N–O distance of 2.6 Å), and the main-chain CO group of residues 190 and 219 (N–O distance of 2.7 and 2.8 Å, respectively; Fig. 4e). In the crystal structure of the complex between *N*-acetyl-D-Phe-Pro-boro-homoLys-OH and thrombin the homolysine side chain donates to both carboxylate oxygens of $Asp^{189}$ (2.9 and 3.0 Å) and participates in polar contacts with the backbone carbonyl oxygens of $Ala^{190}$ and $Gly^{219}$ (3.6 and 4.1 Å, respectively) [53]. Furthermore, there is a water molecule between the homoLys $NH_3^+$ and the carbonyl oxygen of $Phe^{227}$ [53]. The MCSS runs were performed without explicit solvent molecules; this may have affected the position of the methylammonium nitrogen of minimum 1, which is shifted towards the $O^{\delta 2}$ of $Asp^{189}$ instead of being located in a symmetrical position with respect to both carboxylate oxygens of $Asp^{189}$, as in the structure with the boronic acid inhibitor [53]. Methylammonium minima 2 to 4 are also involved in a salt bridge with

$Asp^{189}$ $O^{\delta 2}$ (not shown), but orient their methyl group in a small cavity below the main-chain NH and CO groups of $Glu^{217}$, so that they cannot be part of a longer ligand. Since they are not completely free to select the best orientation for optimization of the electrostatic interaction and since their methyl group does not shield them from solvent in S1, their electrostatic interaction energy with the protein (mainly with $Asp^{189}$) is between 15.1 and 18.7 kcal/mol less favorable than that of minimum 1 (Table 4). Methylammonium minima 5 and 6 are involved in a salt bridge with the $Glu^{146}$ side chain close to the autolysis loop (not shown). The $Glu^{146}$ side chain is partially exposed to solvent; hence, the electrostatic interaction energy is roughly a factor of four lower than that of minimum 1 (values of −43.7, −9.9 and −11.8 kcal/mol for minima 1, 5, and 6, respectively). The total free energy of binding of minima 5 and 6 is −0.5 and −0.1 kcal/mol, respectively. Of the 52 methylammonium minima found by MCSS only six have a favorable free energy of binding. Methylammonium minima 7 to 52 are involved either in hydrogen bonds with polar groups or make salt bridges with Asp or Glu side chains on the surface of thrombin. To show that electrostatic interactions at the protein surface do not contribute significantly to the binding free energy, methylammonium minima 24 and 41 are shown in Fig. 4e and their energies are listed in Table 4. Minimum 24 (no. 4 according to CHARMM energy) participates in a salt bridge with the carboxylate oxygens of $Asp^{60E}$, while minimum 41 (no. 5 according to the CHARMM energy) donates to the main-chain carbonyl oxygens of residues 60D and 60E. They have a CHARMM electrostatic energy (R dielectric constant) of −45.9 and −44.3 kcal/mol, respectively. Their shielded electrostatic energy (computed by the continuum approach) is −11.1 and −13.9 kcal/mol. This is not even enough to balance the total electrostatic desolvation penalty of 11.2 kcal/mol and 19.6 kcal/mol, respectively.

There is a minimized position of methylammonium

TABLE 4
MINIMA OF CHARGED GROUPS

| Rank | Rank | Strain | Intermolecular | | Desolvation | | | $\Delta G_{binding}$ | MCSS rank | Site and H-bond partners |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | vdWaals | Elect | Nonpolar | Electrostatic | | | | |
| | | | | | | Protein | Ligand | | | |
| **Methylammonium** | | | | | | | | | | |
| **1** | 260 | 0.6 | −0.3 | −43.7 | −5.0 | 17.7 | 24.6 | −6.1 | 1 | S1; $Asp^{189}$ $O^{\delta 2}$, 190CO and 219CO |
| 2 | 677 | 0.6 | −0.9 | −25.2 | −4.4 | 4.3 | 23.6 | −1.9 | 3 | S1; $Asp^{189}$ $O^{\delta 2}$, 219CO |
| 3 | 762 | 0.3 | 0.9 | −28.6 | −4.5 | 11.4 | 19.8 | −0.7 | 2 | S1; $Asp^{189}$ $O^{\delta 2}$, 219CO |
| **9** | 851 | 0.1 | −0.6 | −10.9 | −4.2 | 1.9 | 14.4 | 0.6 | 26 | S3; 216CO |
| **21** | 967 | 0.4 | 1.0 | −19.1 | −5.4 | 6.6 | 19.5 | 3.1 | 13 | S1; $Ser^{195}$ $O^{\gamma}$, 214CO |
| **24** | 975 | 1.1 | 4.3 | −11.1 | −2.2 | 4.0 | 7.2 | 3.4 | 4 | surface; $Asp^{60E}$ $O^{\delta 1}$ and $O^{\delta 2}$ |
| **41** | 1102 | 0.4 | 3.8 | −13.9 | −3.4 | 6.1 | 13.5 | 6.7 | 5 | surface; CO of 60D and 60E |
| **Methylguanidinium** | | | | | | | | | | |
| **1** | 1 | 0.7 | −6.6 | −40.9 | −6.8 | 20.9 | 20.2 | −12.5 | 1 | S1; $Asp^{189}$ $O^{\delta 1}$ and $O^{\delta 2}$, 219CO |
| 2 | 2 | 0.3 | −5.3 | −33.7 | −6.6 | 18.9 | 14.3 | −12.1 | 4 | S1; $Asp^{189}$ $O^{\delta 1}$ and $O^{\delta 2}$, 219CO |
| 3 | 5 | 0.7 | −6.9 | −43.1 | −6.9 | 22.0 | 22.8 | −11.4 | 2 | S1; $Asp^{189}$ $O^{\delta 1}$ and $O^{\delta 2}$, 219CO |
| **4** | 18 | 1.2 | −7.6 | −35.1 | −6.9 | 22.0 | 16.3 | −10.1 | 3 | S1; $Asp^{189}$ $O^{\delta 2}$, 219CO |
| 5 | 25 | 1.3 | −4.5 | −32.4 | −6.6 | 18.9 | 13.7 | −9.7 | 6 | S1; $Asp^{189}$ $O^{\delta 2}$, 219CO |
| 6 | 36 | 1.4 | −7.1 | −36.0 | −6.9 | 21.2 | 18.0 | −9.4 | 5 | S1; $Asp^{189}$ $O^{\delta 2}$, 219CO |
| **Acetate ion** | | | | | | | | | | |
| 1 | 144 | 0.0 | −8.7 | −11.0 | −5.7 | 6.9 | 11.0 | −7.5 | 7 | $Asn^{143}$, 147NH, 148NH |
| 2 | 190 | 0.0 | −9.0 | −10.0 | −5.9 | 7.0 | 11.1 | −6.9 | 6 | $Asn^{143}$, 147NH, 148NH |
| 3 | 380 | 0.0 | −9.4 | −6.4 | −5.9 | 7.2 | 9.9 | −4.7 | 9 | $Asn^{143}$, 147NH |
| 4 | 570 | 0.1 | −5.4 | −7.3 | −4.2 | 8.3 | 5.3 | −3.2 | 11 | $Asn^{143}$, $Thr^{147}$ OH |
| 5 | 628 | 0.0 | −2.9 | −3.5 | −2.9 | 3.5 | 3.2 | −2.5 | 15 | $Trp^{60D}$ $N^{\varepsilon 1}$ |
| **7** | 910 | 0.4 | 4.0 | −12.7 | −5.6 | 5.5 | 10.1 | 1.7 | 10 | S2–S1'; $Lys^{60F}$ |
| **11** | 1163 | 0.3 | −1.0 | −17.5 | −6.6 | 25.4 | 8.1 | 8.9 | 4 | S1'; $Lys^{60F}$ |
| **13** | 1184 | 0.1 | 0.6 | −19.2 | −5.4 | 25.2 | 8.1 | 9.5 | 3 | S1'; $Lys^{60F}$ |

Energy values in kcal/mol are listed for the minima of charged groups discussed in the text. Minima with rank in bold are shown in Figs. 4e–h. For heading explanation see caption of Table 3.