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 4×10^3 and 15×10^3 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.

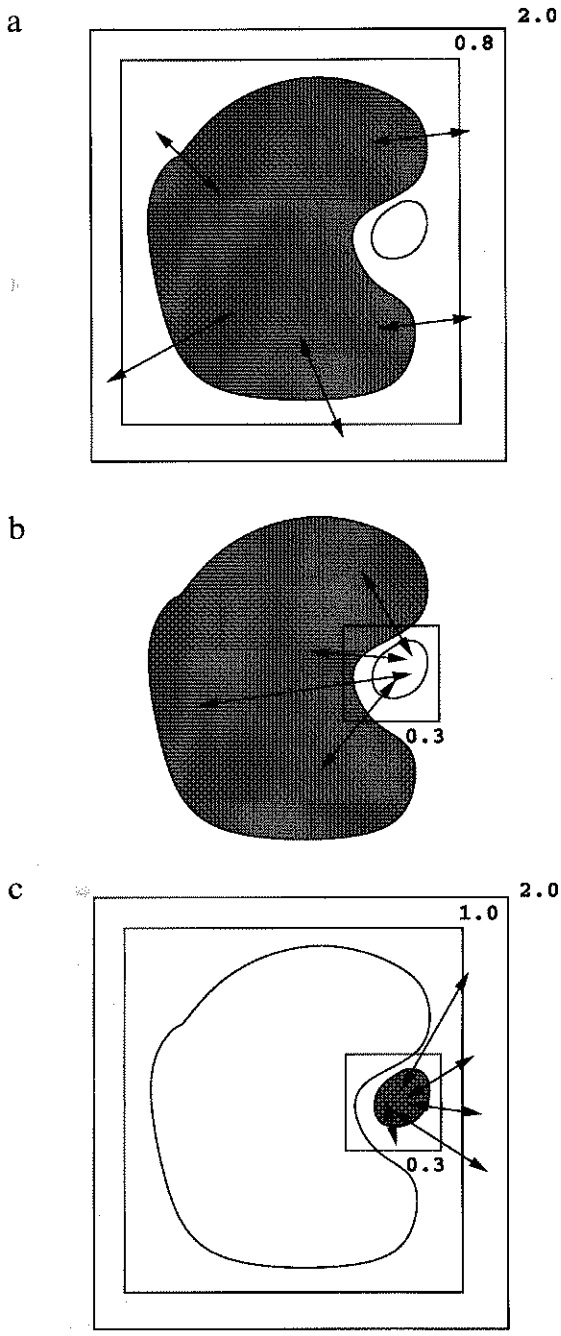

Fig. 1. Schematic representation of the set-up for the numerical solution of the LPB equation. Protein and ligand are represented by a large and a small shape, respectively. Each rectangle corresponds to the boundary of the grid on which the PB equation was solved numerically. The number close to a grid 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 ϕ_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

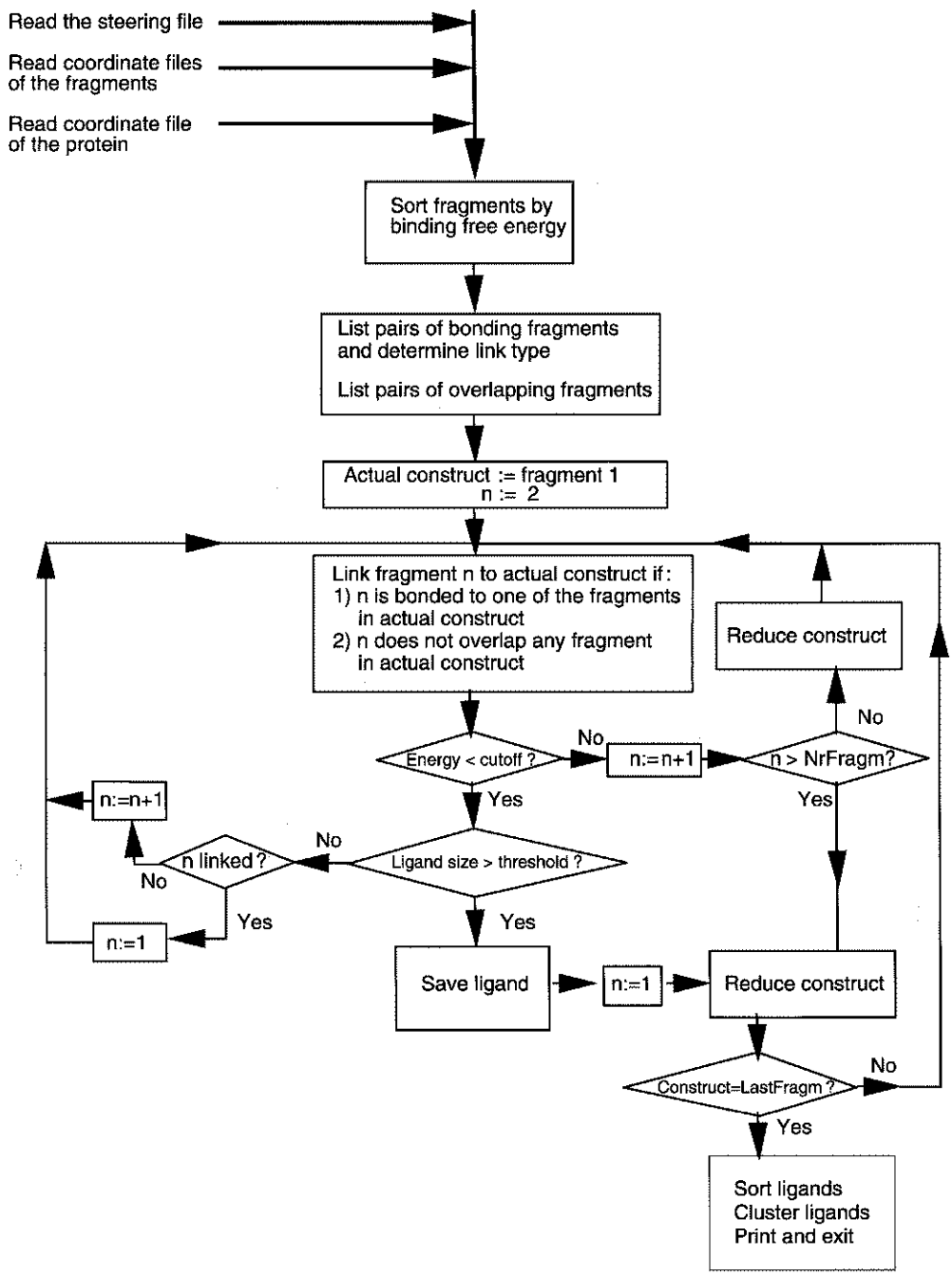

Fig. 2. Schematic representation of the CCLD program. Variable assignments are symbolized by ':='. Conditional statements are enclosed by diamonds (<>).

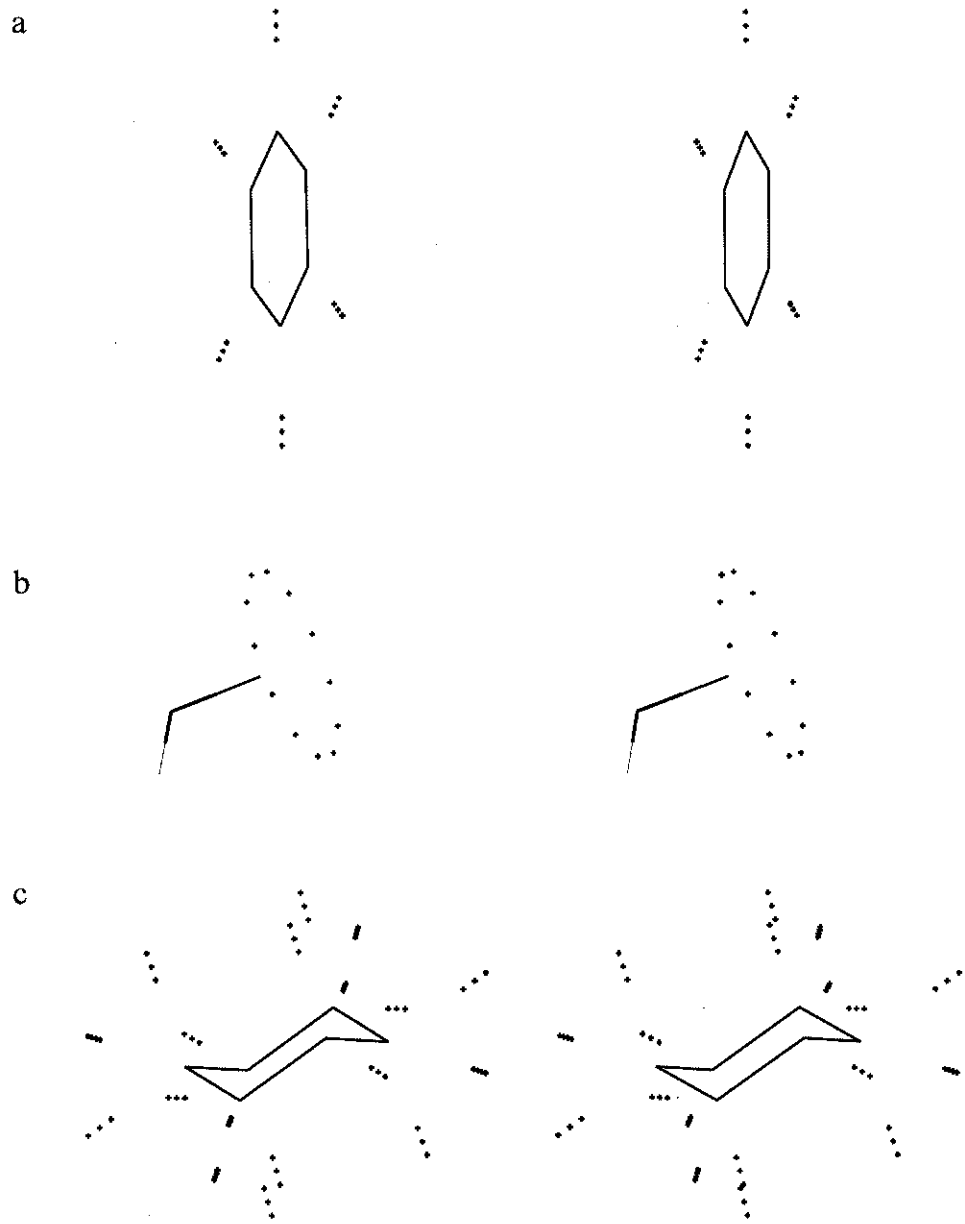

Fig. 3. Stereoviews of the linkage points generated by CCLD for the linkage atoms. (a) For an sp^2 atom (e.g., benzene carbons), three linkage points are defined on the plane at a distance of 1.3 Å, 1.5 Å, and 1.7 Å. (b) For an sp^3 atom (e.g., the carbon atom in methanol), 12 linkage points are distributed 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 (ΔE_{vdW}^{fragm}), and the vacuum Coulombic energy between atoms of the MCSS group (ΔE_{elect}^{fragm}). The CHARMM force-field is used to compute ΔE^{fragm} and ΔE_{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 γ, 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^{α}, C^{β}
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

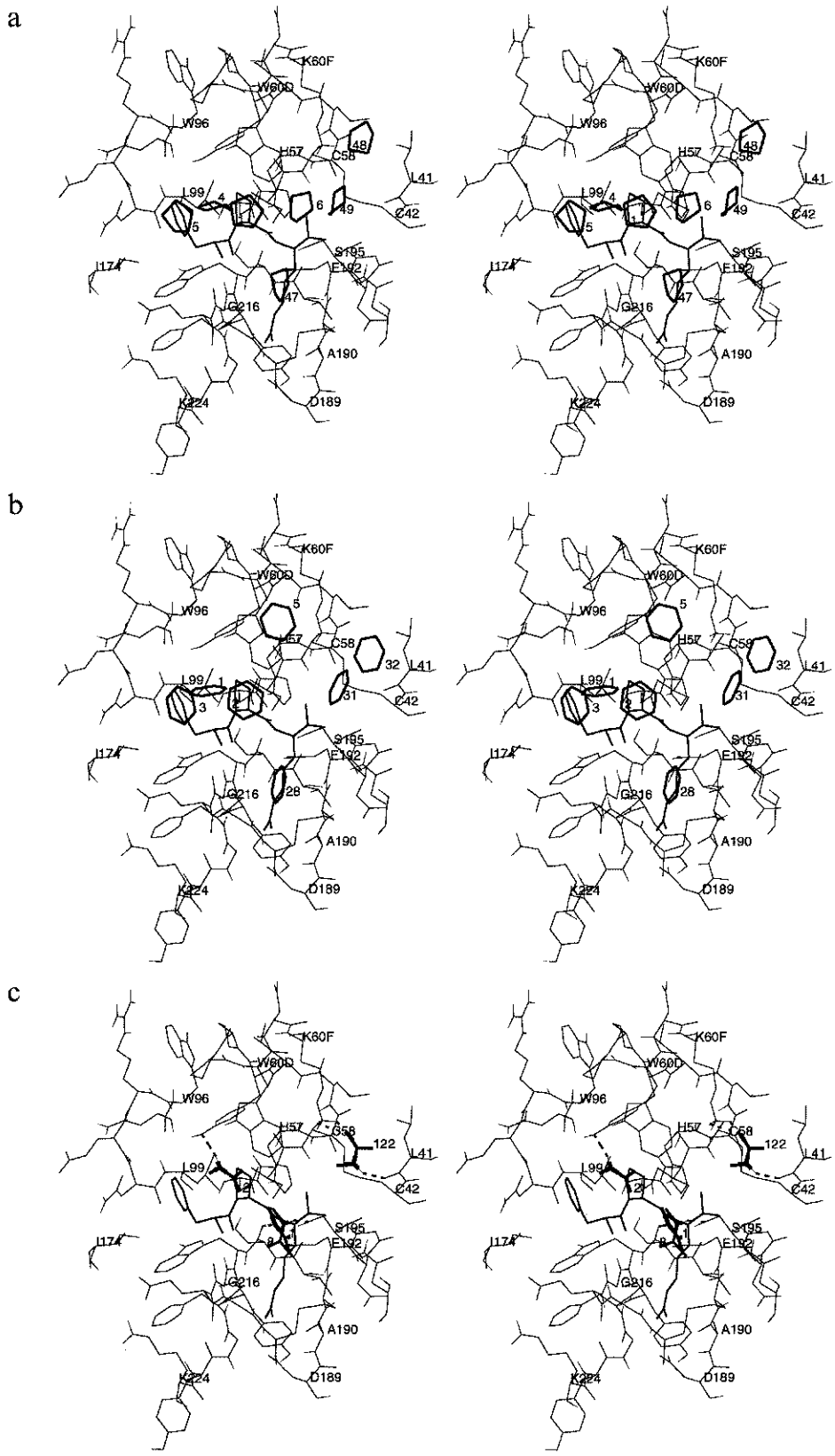

Fig. 4. Stereoviews of the MCSS minima (thick lines for heavy atoms and thin lines for polar hydrogens) in the thrombin active site (thin lines). The PPACK inhibitor is also shown (medium lines), though it was removed during the MCSS procedure. Some C^{α} atoms of thrombin are labeled. In the chymotrypsin 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).

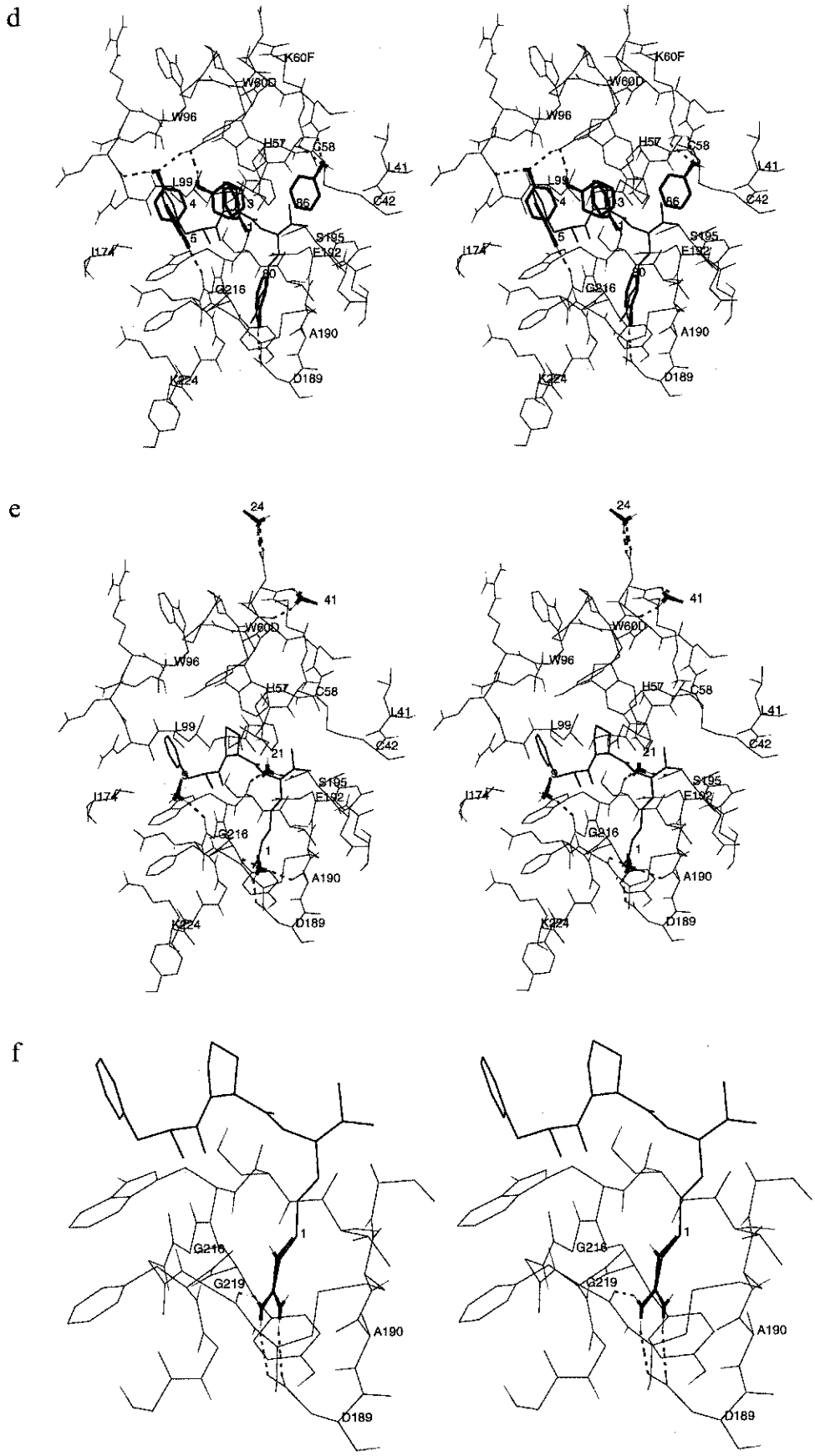

Fig. 4. (d) phenol; (e) methylammonium; (f) methylguanidinium minimum 1.

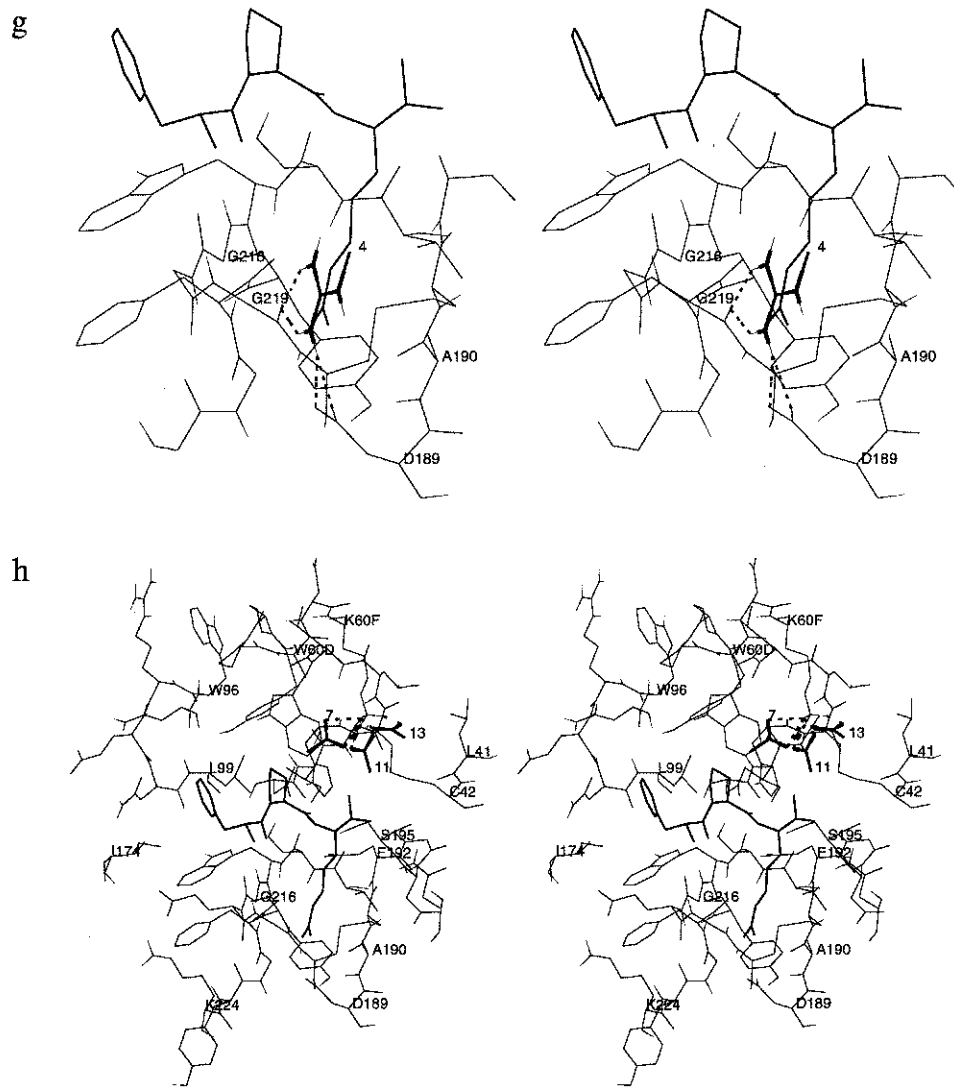

Fig. 4. (g) methylguanidinium minimum 4; (h) acetate ion.

–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^{α}, C^{β}, and C^{γ} 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^{η}
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^{η}
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^{γ}
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^{η} and N^{ε}, 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^{η} and N^{ε}, 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^{ζ}
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^{η}
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^{η}
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^{ζ}
103	1216	0.6	−5.8	−8.9	−8.0	31.4	1.6	11.1	3	S1'; Lys^{60F} N^{ζ}

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^{γ}, 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.

(minimum 9) which acts as a donor to the main-chain carbonyl oxygen of Gly^{216}, even though its position and orientation differ from that of the N-terminal amino group in PPACK (Fig. 4e). The nitrogen atom of minimum 21 is at a distance of 0.52 Å from the main-chain N atom of the arginine in PPACK and donates to the CO group of residue 214 and the Ser^{195} O^{γ} (Fig. 4e). That both minima 9 and 21 have a slightly unfavorable free energy of binding is due to the high desolvation penalty of the methylammonium group, which is only partially counterbalanced by the intermolecular electrostatic interactions (Table 4).

Methylguanidinium The six minima with the lowest free energy of binding have the most favorable CHARMM energy and are located in the S1 pocket (Table 4). The guanidino moiety of minima 1 to 3 forms a bidentate hydrogen-bonding interaction with the side chain of Asp^{189} (Fig. 4f). The heavy atoms of minimum 1 are within 0.2 Å of the corresponding atoms in the PPACK arginine. The terminal nitrogens N^{η} and $N^{\eta 1}$ are roughly equidistant from the carboxylic oxygens of Asp^{189} (2.73 and 2.76 Å, respectively). The N^{η} nitrogen is additionally hydrogen-bonded to the main-chain carbonyl of Gly^{219} (2.63 Å). Methylguanidinium minimum 3 (not shown) has the same orientation as minimum 1 but is rotated by 180° degrees around the N^{ε}–C^{ζ} axis so that the orientation of the methyl group and the hydrogen on N^{ε} are swapped. Minimum 2 (not shown) has its N^{ε} and N^{η} involved in the bidentate interaction with Asp^{189} (instead of N^{η} and $N^{\eta 1}$ as in minima 1 and 3). Minimum 4 displays another hydrogen-bond arrangement (Fig. 4g), where its N^{η} donates to the Asp^{189} carboxylate oxygens (2.60 and 3.42 Å), and both N^{η} and $N^{\eta 1}$ are involved in hydrogen bonds with the backbone CO of Gly^{219} (2.64 and 2.70 Å). A similar binding mode for the guanidino group has recently been found in a retro-binding peptide inhibitor of thrombin, Phe-*allo*-Thr-Phe-O-CH_3 acylated at its N-terminus with 4-guanidino butanoic acid [54]. Methylguanidinium minima 5 and 6 have the same orientation and interactions with the protein as minimum 4, despite slightly different orientations of the methyl group (not shown). The methylguanidinium minimized positions 1 to 6 have similar energy contributions; in particular the values of the intermolecular van der Waals energy and nonpolar desolvation energy do not vary significantly. Minima 1 and 3 have a more favorable electrostatic interaction and higher ligand desolvation penalty than 2, and 4 to 6 (Table 4), due to their optimal arrangement of intermolecular hydrogen bonds (bidentate interaction between the N^{η}, $N^{\eta 1}$, and the Asp^{189} carboxylate oxygens). There are other methylguanidinium minima with a favorable binding free energy, which interact with polar and charged groups on the thrombin active site; e.g., minimum 9 donates to the backbone CO group of residue 214, minimum 10 and 12 to the CO group of Gly^{216}, and minimum 15 to the 214CO and the hydroxyl oxygen of Ser^{195}.

Pyrrolidine The map of pyrrolidine is similar to that of methylammonium. Pyrrolidine minima 1 and 5 are involved in a salt bridge with Asp^{189} in S1, minima 2 and 3 donate an hydrogen bond to the CO group of Gly^{216} and have their carbon atoms in the S3 pocket (though not completely buried). Minima 4 and 6 are located between S3 and S2 and act as donors to the hydroxyl oxygen of Tyr^{60A}, while minimum 9 participates in hydrogen bonds with the CO of residue 214 and the hydroxyl oxygen of Ser^{195}.

2-Acylpyrrolidine The three minima with the lowest free energy are among the best of all the minima found in this work (Table 1). They have a total binding free energy of –11.4, –11.1, and –9.6 kcal/mol. The ring of minima 1 and 3 overlaps the PPACK proline in S2, minimum 1 has the same main-chain orientation as in PPACK, while minimum 3 is oriented in the opposite direction. Both minima have the CO group involved in a hydrogen bond with the backbone NH of Gly^{216}. Minimum 2 donates to the $N^{\varepsilon 2}$ of His^{57} and has its CO group oriented towards the oxyanion hole.

Acetate ion There are five minima with favorable binding free energy; four of these are close to the autolysis loop and accept from the Asn^{143} side chain and other polar groups close to it, while minimum 5 accepts from the $N^{\varepsilon 1}$ of Trp^{60D} (Table 4). Minima 7, 11, and 13 participate in hydrogen bonds with the Lys^{60F} side chain (Fig. 4h) but their total free energy of binding is unfavorable. For minimum 7, which has the methyl group in S2, the unfavorable binding free energy originates from the poor van der Waals interactions (4.0 kcal/mol) and the fact that the salt bridge is solvent-exposed. Minima 11 and 13 have stronger electrostatic interaction with the Lys^{60F} side chain (–17.5 and –19.2 kcal/mol, respectively) than minimum 7 (–12.7 kcal/mol) because of solvent displacement by their methyl group. Yet, their total energy is less favorable because of the high protein-desolvation penalty (electrostatic contribution of 25.4 and 25.2 kcal/mol for minimum 11 and 13, respectively).

Two conclusions can be drawn from the analysis of the minimized positions of the charged functional groups. Firstly, the minima with the lowest binding free energy have optimal hydrogen bonds with the Asp^{189} side chain. Since the Asp^{189} side chain is more buried than the Lys^{60F} one, the minima of positively charged groups interacting with the former have a more favorable binding free energy than those of the acetate minima close to the latter. This is due to reduced shielding of the charge–charge interaction and the smaller desolvation of the carboxylate oxygens of Asp^{189} than of the amino group in Lys^{60F} (Table 4). Secondly, polar groups on the protein surface may not be ideal partners for a charged functional group, because the high desolvation penalty might not be com-

pletely compensated for by the favorable electrostatic interaction energy. This finding is analogous to the results for the polar functional groups, i.e., binding of one of these to a charged and partially exposed side chain of the protein may result in an unfavorable total binding free energy.

De novo design of thrombin ligands

The 875 MCSS minima with a binding free energy lower than 1.0 kcal/mol were used in the CCLD run. As a pruning criterion in CCLD, a threshold of –6.5 kcal/mol was utilized for the energy of the ligand (value averaged over the fragments). With the choice of linkage parameters listed in the section on bonding fragment pairs, CCLD generated 6865 candidate ligands which were clustered in 691 clusters by a 65% similarity criterion. Among the best 300 ligands a large majority had a methylguanidinium minimum in S1, a hydrophobic fragment in S2, and an aliphatic or aromatic group in S3. These moieties were connected by a variety of minima and/or linkers. A representative example (ligand **I**) is shown in Fig. 5. It consists of the imidazole minimum 8 (only the free energy rank will be used henceforth) in S3, phenol 10 in S2, NMA 100 and pyrrolidine 20 in S1, 2-acylpyrrolidine 412 close to the backbone of residues 216 to 219, phenol 68 close to the autolysis loop and cyclohexane 73 in contact with the six-membered ring of Trp^{60D} (Fig. 5a). These MCSS minima are connected by a 1-bond (minima 10 and 100), as well as methylene, ethylene, amide, and keto linkers. Since the cyclohexane minimum is close to a cluster of benzene minima that are involved in face-to-face interactions with Trp^{60D} (see Fig. 4b), the cyclohexyl substituent was replaced by a phenyl group. To regularize the structure, a conjugate gradient minimization was then carried out. The thrombin structure was kept rigid, except for the Tyr^{60A}–Lys^{60F} loop and the resulting conformation of the complex is shown in Fig. 5c. Minimization of the same compound with cyclohexane instead of benzene generated an almost identical conformation (not shown). Only the amide linker between pyrrolidine 412 and phenol 68 moves on minimization; the corresponding amide plane rotates by about 90°. The remaining functional groups undergo minor displacement to relieve some minor strain and to improve their interactions with the thrombin active site. The Tyr^{60A}–Lys^{60F} loop moves as a rigid body towards compound **I**; this is consistent with crystallographic data for complexes of thrombin with different inhibitors (see Figs. 3 and 4 of Banner and Hadvary [26]). Compound **I** is involved in the same interactions with thrombin as PPACK. Furthermore, it forms five additional intermolecular hydrogen bonds; the hydroxyphenyl hydrogen in S2 donates to the hydroxyl oxygen of Tyr^{60A}, the 2-acylpyrrolidine carbonyl oxygen accepts from the NH group of Gly^{219}, and the hydroxyl group on the hydroxyphenyl moiety close to the autolysis loop donates to the hydroxyl oxygen of Thr^{147} and accepts from the main-chain NH group of Thr^{147} and from the side chain of Asn^{143} (Fig. 5c). Moreover, the face-to-face interaction between the benzene ring and the indole of Trp^{60D} may also contribute to affinity, although not as much as the interactions in the S3 to S1 pockets [52].

In another CCLD run an additional fragment was used. This consisted of the N and C^{α} atoms of D-Phe in PPACK, oriented in the same way as in the crystal structure of the PPACK–thrombin complex. As mentioned in the previous section, there is a methylammonium minimum (number 9) which donates to the CO group of Gly^{216}, but its position and orientation are different from the corresponding group in PPACK (see Fig. 4e). This CCLD run used the 300 fragments with the lowest binding free energy and an energy cutoff of –6.5 kcal/mol. It generated 157 putative ligands containing the two-atomic fragment from PPACK. Again, a large majority had a minimized methylguanidinium in S1, minima of hydrophobic fragments in S2, and aliphatic or aromatic group minima in S3. An interesting candidate ligand (**II**) is shown in Fig. 6. It consists of the imidazole minimum 8 in S3 (same minimum used in compound **I**), phenol 225 between S3 and S2, cyclopentane 39 in S2, 2-propanone 288 at the entrance of S1, and methylguanidinium 256 in S1 (Fig. 6a). Cyclopentane 39 is connected to the PPACK N-terminal ammonium group through a keto linker. Thereby, CCLD automatically mutates the closest cyclopentane carbon in an sp^2 nitrogen, which results in a secondary amide connection, i.e., a proline side chain in S2 (Fig. 6b). Compound **II** was then minimized in the rigid thrombin structure and the resulting complex is shown in Fig. 6c. Apart from the hydrogen bond between the main-chain NH group of the PPACK arginine and the CO group of residue 214, all noncovalent interactions between PPACK and thrombin are also present in the complex between ligand **II** and thrombin. In addition, the OH group on the hydroxyphenyl substituent donates to the hydroxyl oxygen of Tyr^{60A} and accepts from the indole NH moiety of Trp^{60D}, and there is a favorable edge-to-face interaction between the hydroxyphenyl group and the Trp^{60D} indole. Moreover, the intraligand hydrogen bond between the N-terminal amino and the proline carbonyl oxygen may result in additional stabilization energy of the ligand.

A CCLD option has been implemented to select the MCSS minima having one or more atoms within a sphere whose center and radius may be specified by the user. This option was used in a CCLD run to construct small molecular structures consisting of two or more MCSS minima from a sphere of radius 7.0 Å centered on the S1'-S2' sites. A total of 48 small molecules was generated and most of them had one of two major binding motifs;

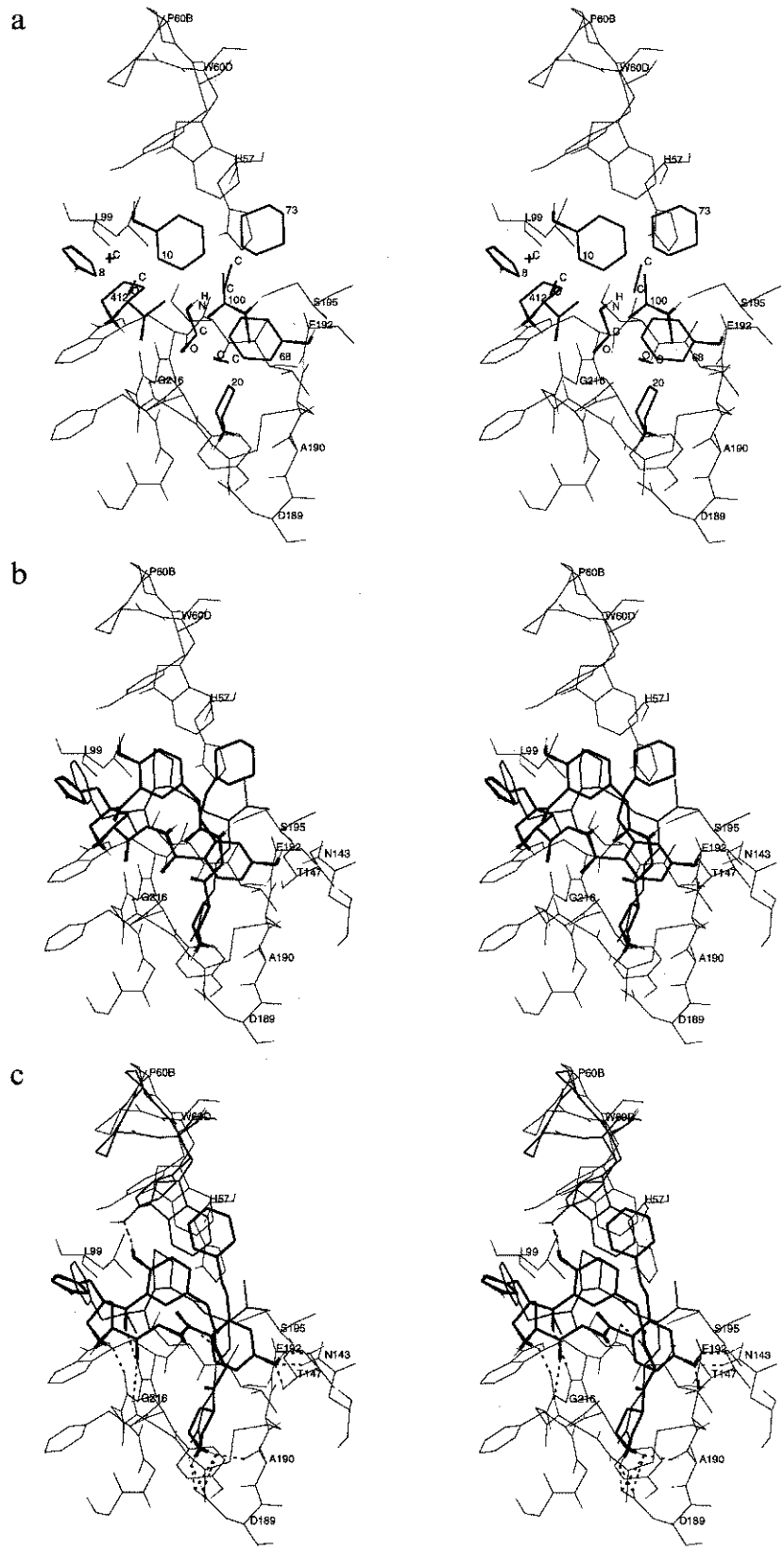

Fig. 5. Stereoviews of the MCSS minima (thick lines for heavy atoms and thin lines for polar hydrogens) selected by CCLD to generate the putative ligand **I** in the structure of the thrombin active site (thin lines). (a) Before connection. The labels on MCSS minima represent their rank among the 1314 MCSS minima, those on the linker units indicate atomtype. (b) After connection. PPACK is shown in medium lines. (c) After minimization in the thrombin structure, which was kept rigid except for the Tyr^{60A}–Trp^{60D} loop, whose minimized conformation is shown in medium lines. PPACK is shown in medium lines; dashed lines represent hydrogen bonds.

an aromatic fragment with its plane parallel to the amide plane of residues 192–193 or a hydrophobic group in contact with the Leu^{40} side chain. A variety of polar groups connected to these fragments was involved in hydrogen bonds with the backbone CO group of Leu^{40} and Leu^{41} and/or the side chain of Asn^{143} and Thr^{147}. These molecules are interesting candidates for the eventual elongation of a nonprime inhibitor into the S1'-S2' subpockets.

From these representative examples and from visual analysis of a series of cluster representatives, it is clear that CCLD produces molecular structures which fill hydrophobic pockets with nonpolar functional groups and have most of the polar groups involved in hydrogen bonds with hydrophilic groups of the protein. Moreover, the compounds generated by CCLD are rather rigid and usually do not have more than three stereocenters.

Experimental validation of two candidate thrombin inhibitors designed with the help of the MCSS–CCLD approach is currently going on at Sandoz Pharma in Basel (C. Ehrhardt and A. Caflisch, unpublished results).

Discussion

The CCLD program has been developed for the de novo design of candidate ligands for enzymes or receptors of known three-dimensional structure by a combinatorial search strategy. It exploits the functionality maps of the MCSS procedure, which provides optimal positions and orientations of small fragment molecules on the surface of a protein. A distinctive feature of the present approach is the evaluation of an approximated free energy of binding for each protein–MCSS minimum complex. For this purpose, the solvation free energy is assumed to be the sum of electrostatic and nonpolar contributions; the former is calculated by numerical solution of the LPB equation, while the latter is assumed to be proportional to the change in total solvent-accessible surface area.

In a typical CCLD run all possible ways of building molecules consisting of 4 to 7 MCSS minimized positions are evaluated, i.e., in the order of 10^{12} to 10^{17} compounds from a set of 50 to 300 MCSS-minimum–linker combinations in each subpocket. There are two features that allow CCLD to carry out such a search in less than 1 h of CPU time on a current workstation. Firstly, before beginning the combinatorial search, a list of bonding fragment pairs and a list of overlapping fragment pairs are generated. At the same time, for each pair of bonding fragments, the type of bond and the coordinates of the atoms of the linker unit are precomputed and stored. Secondly, the combinatorial growing process is kept under control by pruning, which is performed according to the average value of the approximated binding free energy of the fragments in the ligand.

The MCSS–CCLD approach has some features similar to LEGO [55], which is a recently developed tool for de novo ligand design. LEGO does not take into account explicitly 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 [56].

Some aspects of the CCLD program are similar to existing computational techniques based on combinatorial methodologies for structure-based ligand design. The computer program HOOK [18] places molecular 'skeletons' from a database into the protein binding region by making bonds between a carbon atom on the skeleton and a carbon on the MCSS minima. Because of the carbon–carbon connection, which yields a large number of stereocenters, the candidate ligands produced by HOOK are often difficult to synthesize. In addition, the skeletons are placed into the binding site by using a crude estimate of the intermolecular van der Waals energy and no evaluation of the electrostatic interaction between the skeleton and the protein atoms is performed. Hence, skeletons large enough to link functional groups distributed over a significant portion of the binding site add considerably to the molecular weight, but do not necessarily contribute to binding interactions. CCLD uses more connection types than HOOK, so that the resulting molecules have more diverse chemistry. In addition, amide and keto groups, which often yield additional intermolecular hydrogen bonds, are selected more often than ethylene and methylene linkers. Moreover, for each pair of fragments linked by an amide or keto linker, the electrostatic interaction between the linker unit and the protein binding site is precomputed and a 'growing' ligand is penalized if this energy is less favorable than the solvation free energy of NMA or 2-propanone, respectively. A different approach for fragment-based de novo ligand design involves the sequential build-up of a candidate ligand molecule. Rotstein and Murcko developed GROUPBUILD, a fragment-by-fragment ligand generator [29]. 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, and have 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. No attempt is made to estimate the electrostatic contribution to the free energy of desolvation. A program similar to GROUPBUILD was recently described by Bohacek and McMartin [30]. 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 comple-

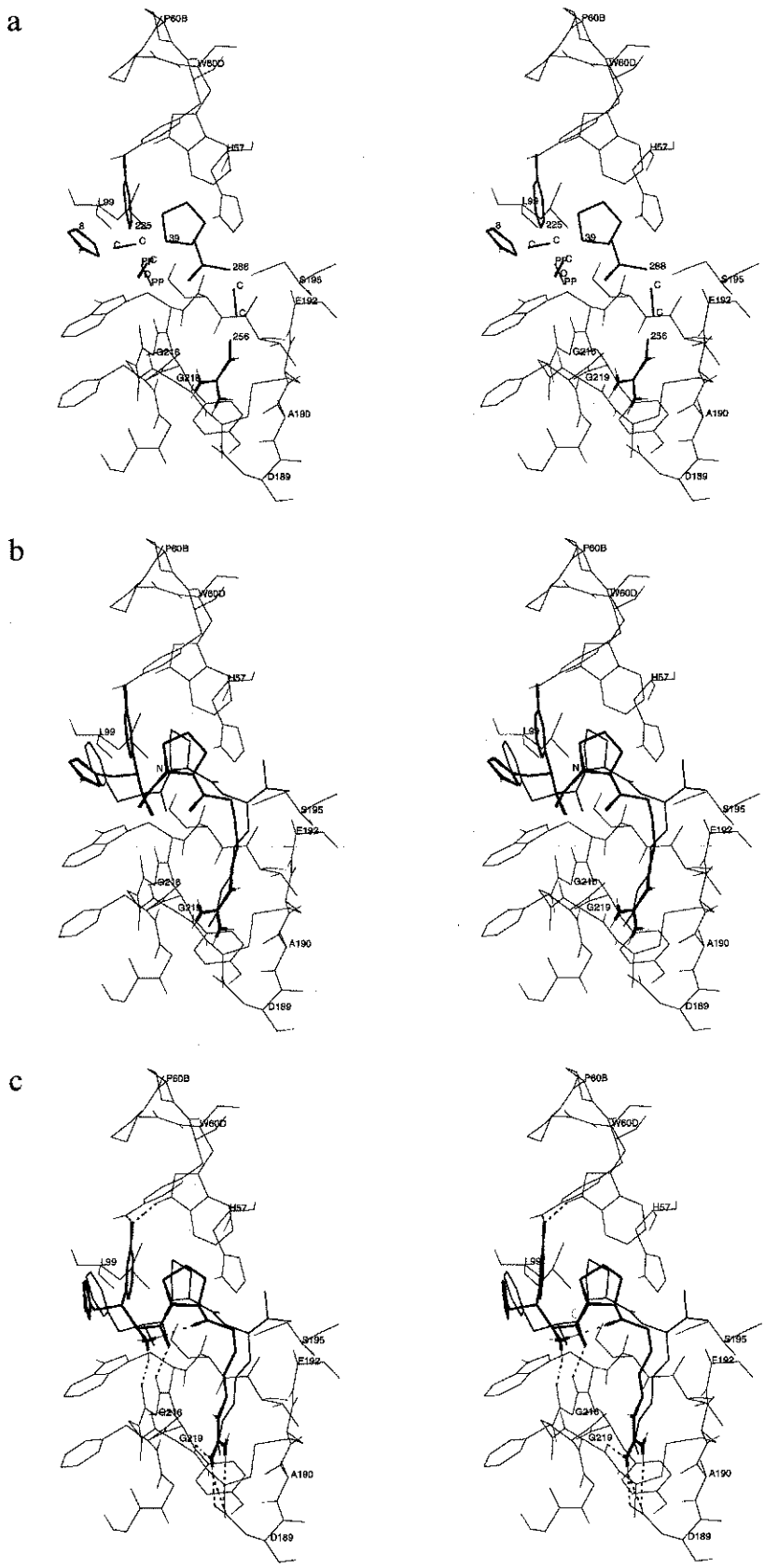

Fig. 6. Stereoviews of the MCSS minima (thick lines for heavy atoms and thin lines for polar hydrogens) selected by CCLD to generate the putative ligand II in the structure of the thrombin active site (thin lines). (a) The two-atomic segment-labeled PP corresponds to the N and C^{α} atoms of D-Phe in PPACK; (b) the label N indicates the cyclopentane atom which was mutated by CCLD from an sp^3 carbon into an sp^2 nitrogen; (c) after minimization in the rigid thrombin structure.

mentarity score, by means of rewarding carbons in hydrophobic regions or hydrogen-bonding atoms near appropriate partners, and penalizing mismatches between atom type and binding region. The main disadvantage of sequential fragment build-up procedures is that they do not use informations 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 keep the number of molecules within bounds, pruning is carried out based on energy estimates which do not take solvation effects into account and which might eventually eliminate possible ligand candidates. Also, so far only molecules with fixed bond lengths, bond angles, and dihedral angles corresponding only to the rotational isomeric states of each torsion bond have been constructed. CCLD does not suffer of any of these limitations.

The approach discussed here for computer-aided structure-based ligand design was tested on the thrombin active site. The MCSS methodology was applied to 16 functional group types and 1314 minimized positions were generated. These were sorted according to approximated binding free energy. Four major binding motifs may be identified from the analysis of the 100 MCSS minima with the lowest binding free energy. The first is the salt bridge between positively charged functional groups and the Asp^{189} side chain at the bottom of the S1 pocket. The remaining three involve nonpolar fragments in the S3 pocket, S2 pocket, or on the solvent-exposed surface of the Trp^{60D} indole. This is in agreement with the binding modes of known thrombin inhibitors [25,26,52,54]. A CCLD run, which utilized the 875 MCSS minima with favorable free energy of binding, generated several ligands showing the same interaction patterns as those of the PPACK–thrombin complex. Furthermore, larger ligands containing both a 'core' similar to PPACK and additional intermolecular hydrogen bonds and/or van der Waals interactions were generated. PPACK itself was not suggested since the methylammonium minimum interacting with the CO of Gly^{216} has a different orientation from that of the N-terminal amino group of PPACK (Fig. 4e).

In addition to the validation of the MCSS and CCLD procedures and the demonstration of the importance of the approximated free energy of binding for a realistic ranking of the MCSS minima, a number of results of the present study might be of interest for the design of small molecular weight and active-site-directed thrombin inhibitors. Firstly, although there are MCSS minima in both the nonprime and prime parts of the thrombin active site, those with the lowest free energy of binding are located in the nonprime subsites. In particular, the S3 and S2 pockets, due to their hydrophobic character, are ideal binding regions for nonpolar moieties, while the deep channel in S1 with Asp^{189} at its bottom is an optimal site for positively charged functional groups. This is consistent with known experimental data and could have been obtained by a visual analysis of the crystal structures of thrombin–inhibitor complexes. What cannot be deduced from a superficial analysis of the structural data is the fact that the binding affinity of an acetate ion for the Lys^{60F} side chain in S1' is much less favorable (positive binding free energy, see Table 4) than the binding affinity of a methylammonium ion for the Asp^{189} side chain in S1. This is a consequence of the shielding of the electrostatic interaction between charged partners in S1', which is much higher than in S1 (Table 4) because of the larger degree of solvent exposure of the former. It is important to note that an eventual systematic error originating from the choice of partial charges and atomic radii for the carboxylic and ammonium ions cancels out in this comparison, since the same sets of parameters are involved in the salt bridges in S1 and S1'.

Secondly, several modifications of known thrombin inhibitors can be deduced from the analysis of the candidate ligands produced by CCLD. Among these, ligands **I** and **II** suggest that the substitution of D-His for D-Phe in PPACK-derived inhibitors may result in equal or more favorable binding strength.

Thirdly, the functionality maps of nonpolar groups indicate that the S3 and S2 pockets are contiguous (Figs. 4a,b). Thus, a polycyclic aromatic or aliphatic compound, which fits in this region, might be an interesting starting point for derivatization, in agreement with the recent discovery of a novel nanomolar thrombin inhibitor consisting of a cyclic template at S2–S3 [57].

Fourthly, a series of CCLD ligands, e.g. **I**, suggest that by tethering a PPACK-type inhibitor on the N-terminal side one may achieve improved binding strength due to additional interactions with: (i) the NH group of Gly^{219}, which in the thrombin–hirudin complex [58] is involved in a hydrogen bond with the backbone carbonyl of hirudin Tyr^{3}; (ii) the Asn^{143} and Thr^{147} side chains; (iii) the Thr^{147} backbone NH; and (iv) the solvent-exposed surface of the Trp^{60D} indole. This approach might be worth investigating, although the approximated binding free energy of the MCSS minima indicate that interactions with these groups are expected to be weaker than those of hydrophobic moieties in S3–S2 and positively charged functionalities in S1.

Finally, a CCLD run, which used only the MCSS minima on the S1'–S2' sites, produced a set of small molecules showing two major interaction patterns which might be relevant for the design of extensions to the prime part of the thrombin active site.

In the present application of the MCSS–CCLD approach the thrombin structure was kept rigid. This approximation is acceptable for thrombin, which, apart from a relatively small rigid-body motion of the Tyr^{60A}–Trp^{60D} loop, assumes the same conformation in complexes

with different inhibitors [15,26]. To extend the MCSS–CCLD approach to proteins with flexible binding regions one might postprocess the MCSS minima by further cycles of minimization or molecular dynamics in the flexible protein. Finally, no attempt has been made in this study to solve the third step of our ligand design approach, i.e., the estimation of the binding constants of candidate ligands. Preliminary results for a series of HIV-1 aspartic proteinase inhibitors [13] suggest that the technique used for the evaluation of the solvation free energy of the MCSS minima, i.e., decomposition into electrostatic and nonpolar contributions, might be an efficient and accurate approach.

As recently emphasized in an excellent review article [59], an important shift of paradigm is taking place in the methodologies for drug discovery. The dual approach of 'rational design and random screening' is being replaced by a new scenario, which may be defined as 'random design and rational screening'. Although large-scale random-screening procedures may now be performed in a reasonable amount of time due to recent advances in robotics, miniaturization, and hybridization of well-developed techniques like solid-phase synthesis and photolithography [60,61], there is a trend towards more focussed screening based on elements of 'rational' design, e.g., mechanism- or structure-based criteria [59,62]. At the same time, the implementation of random and/or combinatorial techniques for computer-aided design allows the generation of whole sets of candidate ligands, which reveal patterns of preferred intermolecular interactions. These are more helpful for the experienced modeller than the 'rational' design of a single putative ligand. In this perspective, the present approach for computer-aided ligand design based on the combinatorial selection of optimally docked fragments is expected to be an important element in the new paradigm for structure-based drug discovery.

Acknowledgements

The development of the CCLD program was inspired by the enthusiasm and constructive criticisms of Dr. C. Ehrhardt (Sandoz Pharma AG, Basel). I gratefully acknowledge J. Apostolakis, Prof. M. Karplus, Prof. A. Plückthun, and Dr. A. Widmer for helpful discussions and for interesting comments on the manuscript. I thank Prof. J.A. McCammon for providing the UHBD program (v. 4.1), which was used for all finite-difference Poisson–Boltzmann calculations. The calculations were performed on an SGI Indigo2 and a four-processor SGI Challenge. The CCLD program, and the UNIX, CHARMm, and UHBD scripts, which were used to compute the approximated binding free energy of the MCSS minima, are available from the author. The CHARMM code within the program QUANTA (v. 4.0, MSI Inc.) was used for some of the minimization performed in this work (minimization of the complete ligands). This work was supported by the Swiss National Science Foundation (grant nr. 3100-043423.95), the Swiss Federal Office of Public Health (Nationales Aids-Forschungs-Programm, grant nr. 3139-043652.95), and the EMDO Stiftung, Zürich.

References

1 Greer, J., Erickson, J.W., Baldwin, J.J. and Varney, M.D., J. Med. Chem., 37 (1994) 1035.
2 Lam, P.Y.S., Jadhav, P.K., Eyermann, C.J., Hodge, C.N., Ru, Y., Bacheler, L.T., Meek, J.L., Otto, M.J., Rayner, M.M., Wong, Y.N., Chang, C.H., Weber, P.C., Jackson, D.A., Sharpe, T.R. and Erickson-Viitanen, S.K., Science, 264 (1994) 380.
3 Hilpert, K., Ackermann, J., Banner, D.W., Gast, A., Gubernator, K., Hadvary, P., Labler, L., Müller, K., Schmid, G., Tschopp, T. and Van de Waterbeemd, H., J. Med. Chem., 37 (1994) 3889.
4 Karplus, M. and Petsko, G.A., Nature, 347 (1990) 631.
5 Van Gunsteren, W.F. and Berendsen, H.J.C., Angew. Chem. Int. Ed. Engl., 29 (1990) 992.
6 Honig, B. and Nicholls, A., Science, 268 (1995) 1144.
7 Appelt, K., Perspect. Drug Discov. Design, 1 (1993) 23.
8 Clore, G.M. and Gronenborn, A.M., Science, 252 (1991) 1390.
9 Fesik, S.W., J. Med. Chem., 34 (1991) 2937.
10 Greer, J., Proteins Struct. Funct. Genet., 7 (1990) 317.
11 Havel, T.F., J. Mol. Simul., 10 (1993) 175.
12 Šali, A. and Blundell, T.L., J. Mol. Biol., 234 (1993) 779.
13 Caflisch, A. and Karplus, M., Perspect. Drug Discov. Design, 3 (1995) 51.
14 Miranker, A. and Karplus, M., Proteins Struct. Funct. Genet., 11 (1991) 29.
15 Stubbs, M.T. and Bode, W., Perspect. Drug Discov. Design, 1 (1993) 431.
16 Noble, M.E.M., Verlinde, C.L.M.J., Groendijk, H., Kalk, K.H., Wierenga, R.K. and Hol, W.G.J., J. Med. Chem., 34 (1991) 2709.
17 Caflisch, A., Miranker, A. and Karplus, M., J. Med. Chem., 36 (1993) 2142.
18 Eisen, M.B., Wiley, D.C., Karplus, M. and Hubbard, R.E., Proteins Struct. Funct. Genet., 19 (1994) 199.
19 Sitkoff, D., Sharp, K.A. and Honig, B., J. Phys. Chem., 98 (1994) 1978.
20 Warwicker, J. and Watson, H.C., J. Mol. Biol., 157 (1982) 671.
21 Gilson, M.K. and Honig, B.H., Proteins Struct. Funct. Genet., 4 (1988) 7.
22 Hermann, R.B., J. Phys. Chem., 76 (1972) 2754.
23 Lee, B. and Richards, F.M., J. Mol. Biol., 55 (1971) 379.
24 Tapparelli, C., Metternich, R., Ehrhardt, C. and Cook, N.S., Trends Pharmacol. Sci., 14 (1993) 366.
25 Bode, W., Mayr, I., Baumann, U., Huber, R., Stone, S.R. and Hofsteenge, J., EMBO J., 8 (1989) 3467.
26 Banner, D.W. and Hadvary, P., J. Biol. Chem., 266 (1991) 20085.
27 Lyle, T.A., Perspect. Drug Discov. Design, 1 (1993) 453.
28 Grootenhuis, P.D.J. and Karplus, M., J. Comput.-Aided Mol. Design, 10 (1996) 1.
29 Rotstein, S.H. and Murcko, M.A., J. Med. Chem., 36 (1993) 1700.
30 Bohacek, R.S. and McMartin, C., J. Am. Chem. Soc., 116 (1994) 5560.
31 Kuntz, I.D., Science, 257 (1992) 1078.
32 Kettner, C. and Shaw, E., Thromb. Res., 14 (1979) 969.

33 Brünger, A. and Karplus, M., Proteins Struct. Funct. Genet., 4 (1988) 148.
34 Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M., J. Comput. Chem., 4 (1983) 187.
35 Elber, R. and Karplus, M., J. Am. Chem. Soc., 112 (1990) 9161.
36 Dirac, P.A.M., Proc. Cambridge Phil. Soc., 26 (1930) 376.
37 Hestenes, M.R. and Stiefel, E., J. Res. N.B.S., 49 (1952) 409.
38 Bashford, D. and Karplus, M., Biochemistry, 29 (1990) 10219.
39 Davis, M.E., Madura, J.D., Luty, B.A. and McCammon, J.A., Comput. Phys. Commun., 62 (1991) 187.
40 Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P., Numerical Recipes in Fortran, Cambridge University Press, Cambridge, U.K., 1992.
41 Davis, M.E. and McCammon, J.A., J. Comput. Chem., 10 (1989) 386.
42 Davis, M.E. and McCammon, J.A., J. Comput. Chem., 11 (1990) 401.
43 Davis, M.E. and McCammon, J.A., J. Comput. Chem., 12 (1991) 909.
44 Lim, C., Bashford, D. and Karplus, M., J. Phys. Chem., 95 (1991) 5610.
45 Edmonds, D.T., Rogers, N.K. and Sternberg, M.J.E., Mol. Phys., 52 (1984) 1487.
46 Mohan, V., Davis, M.E., McCammon, J.A. and Pettitt, B.M., J. Phys. Chem., 96 (1992) 6428.
47 Gilson, M.K., Sharp, K.A. and Honig, B.H., J. Comput. Chem., 9 (1988) 327.
48 Luty, B.A., Davis, M.E. and McCammon, J.A., J. Comput. Chem., 13 (1992) 768.
49 Still, W.C., Tempczyk, A., Hawley, R.C. and Hendrickson, T., J. Am. Chem. Soc., 112 (1990) 6127.
50 Chothia, C., Nature, 248 (1974) 338.
51 Cabani, S., Gianni, P., Mollica, V. and Lepori, L., J. Solution Chem., 10 (1981) 563.
52 Maryanoff, B.E., Qiu, X., Padmanabhan, K.P., Tulinsky, A., Almond Jr., H.R., Andrade-Gordon, P., Greco, M.N., Kauffman, J.A., Nicolaou, K.C., Liu, A., Brungs, P. and Fusetani, N., Proc. Natl. Acad. Sci. USA, 90 (1993) 8048.
53 Weber, P.C., Lee, S.L., Lewandowski, F.A., Schadt, M.C., Chang, C.H. and Kettner, C.A., Biochemistry, 34 (1995) 3750.
54 Tabernero, L., Chang, C.Y., Ohringer, S., Lau, W.F., Iwanowicz, E.J., Han, W.C., Wang, T.C., Seiler, S.M., Roberts, D.G.M. and Sack, J.S., J. Mol. Biol., 246 (1995) 14.
55 Gubernator, K., Broger, C., Bur, D., Doran, D.M., Gerber, P.R., Müller, K. and Schaumann, T.M., In Hermann, E.C. and Franke, R. (Eds.) Computer-Aided Drug Design in Industrial Research, Springer, Berlin, Germany, 1995, pp. 61–77.
56 Gerber, P.R. and Müller, K., J. Comput.-Aided Mol. Design, 9 (1995) 251.
57 Obst, U., Gramlich, V., Diederich, F., Weber, L. and Banner, D.W., Angew. Chem., 107 (1995) 1874.
58 Rydel, T.J., Tulinsky, A., Bode, W. and Huber, R., J. Mol. Biol., 221 (1991) 583.
59 Müller, K., In Schwartz, T.W., Hjorth, S.A. and Sandholm Kastrup, J., (Eds.) Structure and Function of 7TM Receptors, Munksgaard, Copenhagen, Denmark, 1996, pp. 414–421.
60 Gallop, M.A., Barrett, R.W., Dower, W.J., Fodor, S.P.A. and Gordon, E.M., J. Med. Chem., 37 (1994) 1233.
61 Gordon, E.M., Barrett, R.W., Dower, W.J., Fodor, S.P.A. and Gallop, M.A., J. Med. Chem., 37 (1994) 1385.
62 Weber, L., Wallbaum, S., Broger, C. and Gubernator, K., Angew. Chem. Int. Ed. Engl., 34 (1995) 2280.