Biol. Chem., Vol. 382, pp. 1365 – 1372, September 2001 · Copyright © by Walter de Gruyter · Berlin · New York

Fragment-Based Flexible Ligand Docking by Evolutionary Optimization

Nicolas Budin, Nicolas Majeux and Amedeo Caflisch*

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

* Corresponding author

A new computational approach for the efficient docking of flexible ligands in a rigid protein is presented. It exploits the binding modes of functional groups determined by an exhaustive search with solvation. The search in ligand conformational space is performed by a genetic algorithm whose scoring function approximates steric effects and intermolecular hydrogen bonds. Ligand conformations generated by the genetic algorithm are docked in the protein binding site by optimizing the fit of their fragments to optimal positions of chemically related functional groups. We show that the use of optimal binding modes of molecular fragments allows to dock known inhibitors with about ten rotatable bonds in the active site of the uncomplexed and complexed conformations of thrombin and HIV-1 protease.

Key words: Drug design / Functional group / Genetic algorithm / Hash table / SEED / Solvation.

Introduction

The ever increasing number of three-dimensional structures of pharmacologically relevant enzymes and receptors is spurring a strong interest in computer-aided approaches to design drugs using structural information. The correct prediction of the binding mode (docking) of a small molecule (ligand) in the binding site of a protein of known three-dimensional conformation is an important component of structure-based drug design (Kuntz, 1992). Computer-aided structure-based ligand docking requires an automatic procedure able to search the conformational space of the ligand and its position and orientation in the binding site, and a scoring function. The latter should be accurate enough to recognize in a reasonable amount of time the correct binding mode from all the putative modes. The docking problem is the first of two challenges in structure-based drug design, the second being the search in chemical space, *i. e.*, the virtual screening of compounds with high affinity and selectivity for a given protein target (Apostolakis and Caflisch, 1999).

Several computer programs for flexible ligand docking use descriptors of physico-chemical properties of both the protein binding site surface and the ligand (see Apostolakis and Caflisch, 1999, for a review). The essential element of DOCK, the archetypal docking program, is the representation of the shape of the binding site by a minimum set of spheres (Kuntz *et al.,* 1982; Wang *et al.,* 1999). To orient the ligand within the binding site, some of the ligand atoms are matched to the DOCK sphere centers. GOLD is a genetic algorithm approach that uses hydrogen bond donor and acceptor atoms in the binding site to position the ligand by a least square fitting procedure in order to form as many hydrogen bonds as possible (Jones *et al.,* 1995). FlexX uses descriptors which map hydrogen bond donors/acceptors and apolar molecular surfaces (Rarey *et al.,* 1996). The descriptors used in these programs take into account only local interactions and almost completely neglect electrostatic solvation effects and the hydrophobicity of the binding site, although these play a key role in binding (Davis and Teague, 1999; Scarsi *et al.,* 1999).

In this report, we present a fragment-based strategy for efficiently docking flexible ligands in the active site of a rigid protein. First, the most favorable positions and orientations of mainly rigid molecular fragments in the receptor binding site are determined according to an accurate binding energy which includes electrostatic solvation effects (Majeux *et al.,* 1999, 2001). The optimal binding modes of the fragments are then used as binding site descriptors to guide the placement of ligand conformations generated by a genetic algorithm. This approach has been implemented in the program FFLD (fragment-based flexible ligand docking) and is illustrated by docking known nanomolar inhibitors to both complexed and uncomplexed forms of thrombin and HIV-1 protease.

Results

The FFLD approach is illustrated by docking NAPAP (Figure 1A) and XK263 (Figure 1B) in thrombin and HIV-1 protease, respectively. For both proteins, the native and complexed conformations are used and the docking results are compared.

Docking of NAPAP in Thrombin

For cyclohexane and naphthalene the geometrical centers of the 10 best energy positions were selected for ligand placement. Only the 3 best energy positions of benzamidine were used since a significant energy gap is observed between the third and the fourth position.

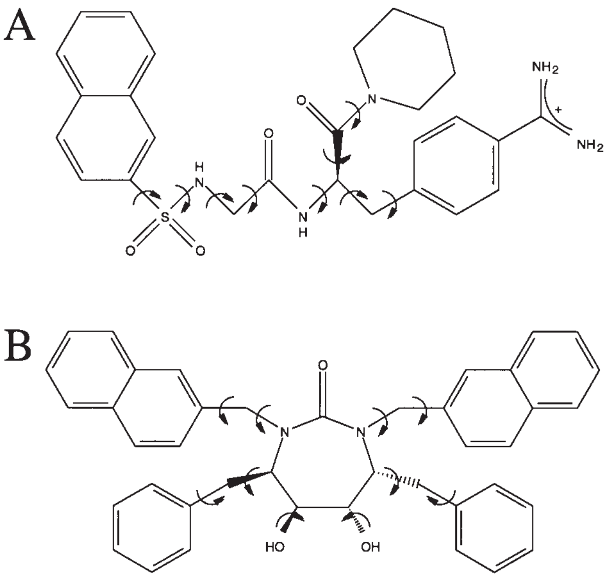

Fig. 1 Ligands Docked by FFLD.
(A) NAPAP and (B) XK263. Rotatable bonds are represented by circular arrows. They were considered as variable degrees of freedom during the docking.

Furthermore, using the 10 best geometrical centers of benzamidine does not affect significantly the docking results, but does however slightly increase the CPU time required by FFLD. The distribution of the fragment geometrical centers in 1HGT and 1DWD is similar: all benzamidine geometrical centers are located in the specificity pocket S1, whereas the cyclohexane and naphthalene positions are found in the hydrophobic regions of the active site, namely the S2 and S3 pockets and the upper part of S1 (Figure 2A,B). The distribution of fragment positions is similar to results obtained in a previous SEED application on thrombin (Majeux *et al.,* 1999). Minor differences are probably due to the value of the solute dielectric constant, 4.0 in this work and 1.0 in Majeux *et al.* (1999).

Five docking runs were performed in both 1HGT and 1DWD with different initial random number values to test the reproducibility of the genetic algorithm. In all 10 docking runs performed by FFLD the conformation with the best scoring energy is similar to its counterpart observed in the corresponding X-ray structure (Figure 2A,B). The benzamidine is correctly placed in S1 and its salt bridge with Asp189 is reproduced. Piperidine and naphthalene are located in the S2 and S3 pockets, respectively. The hydrogen bond between the NH of the NAPAP sulfonamide group and the carbonyl group of Gly216 is reproduced in two docking runs out of five performed on 1HGT, whereas it is not reproduced in the five NAPAP binding modes found by FFLD in 1DWD. The average heavy atom RMS deviation of the five docked NAPAP molecules from their minimized counterpart in 1DWD and 1HGT is 1.79 ± 0.30 Å and 1.15 ± 0.53 Å, respectively. It is interesting to note that similar results are obtained with both thrombin X-ray structures and the deviation is smaller when using the uncomplexed conformation. This is rather counterintuitive but might be related to the extremely short H-bond distance between the NH of the NAPAP sulfonamide and the carbonyl group of Gly216 (N···O distance of 2.27 Å) in the X-ray structure of the NAPAP-thrombin complex. Moreover, the complexed structure has a poorer crystallographic resolution (3.0 Å; Banner and Hadvary, 1991) than the native structure (2.2 Å; Skrzypczak-Jankun *et al.,* 1991).

Docking of XK263 in HIV-1 Protease

1HVR Since XK263 has two phenyl and two naphthyl substituents, the 15 best energy positions of benzene and naphthalene were used for docking XK263 in the active site of HIV-1 protease. The naphthalene geometrical centers are mainly distributed in the S2 and S2' pockets, whereas most of the benzene geometrical centers are located in S1', S2 and S2' (Figure 2D). The best energy binding modes resulting from 5 docking runs performed with FFLD are very similar to the minimized binding mode of XK263 (Figure 2D). The average RMS deviation of the docked ligands from the minimized inhibitor is 1.22 ± 0.12 Å. In all binding modes, the phenyl and naphthyl rings of the inhibitor pack in the hydrophobic S1/S1' and S2/S2' pockets, respectively. Moreover, all four hydrogen bonds between the inhibitor and the receptor are reproduced. The carbonyl group of the cyclic urea is hydrogen bonded to the NHs of isoleucines 50 and 50'. Furthermore, two hydrogen bonds with the catalytic aspartic acids are observed: in two runs out of five both hydroxyl groups of the central cyclic urea are hydrogen bonded to Asp25 whereas in the other runs each hydroxyl is hydrogen bonded to one of the two catalytic Asp residues.

3HVP The uncomplexed conformation of HIV-1 protease represents a challenge for FFLD since its active site is more open than the one in most of the known inhibitor-HIV-1 protease complexes. Binding of the inhibitor results in large motions of the flap regions (Miller *et al.,* 1989). Isoleucines 50 and 50', which are hydrogen bonded to the inhibitor in the complexed structure, are located at the tips of the flaps and undergo a displacement of up to 7 Å upon ligand binding. Moreover, the flap motion decreases the size of the active site cavity upon inhibitor binding. As expected, the geometrical centers distribution differs significantly from the one observed in the complexed form of the enzyme (Figure 2C,D). FFLD finds two different binding modes for XK263 in the native form of HIV-1 protease. Ten docking runs were therefore performed in order to increase the statistics. In 5 runs out of 10 the best energy binding mode is similar to the minimized position of the inhibitor, although one naphthyl ring occupies a different location (Figure 2C). The average RMS deviation of the five FFLD binding modes from the minimized inhibitor without taking into account the naph-

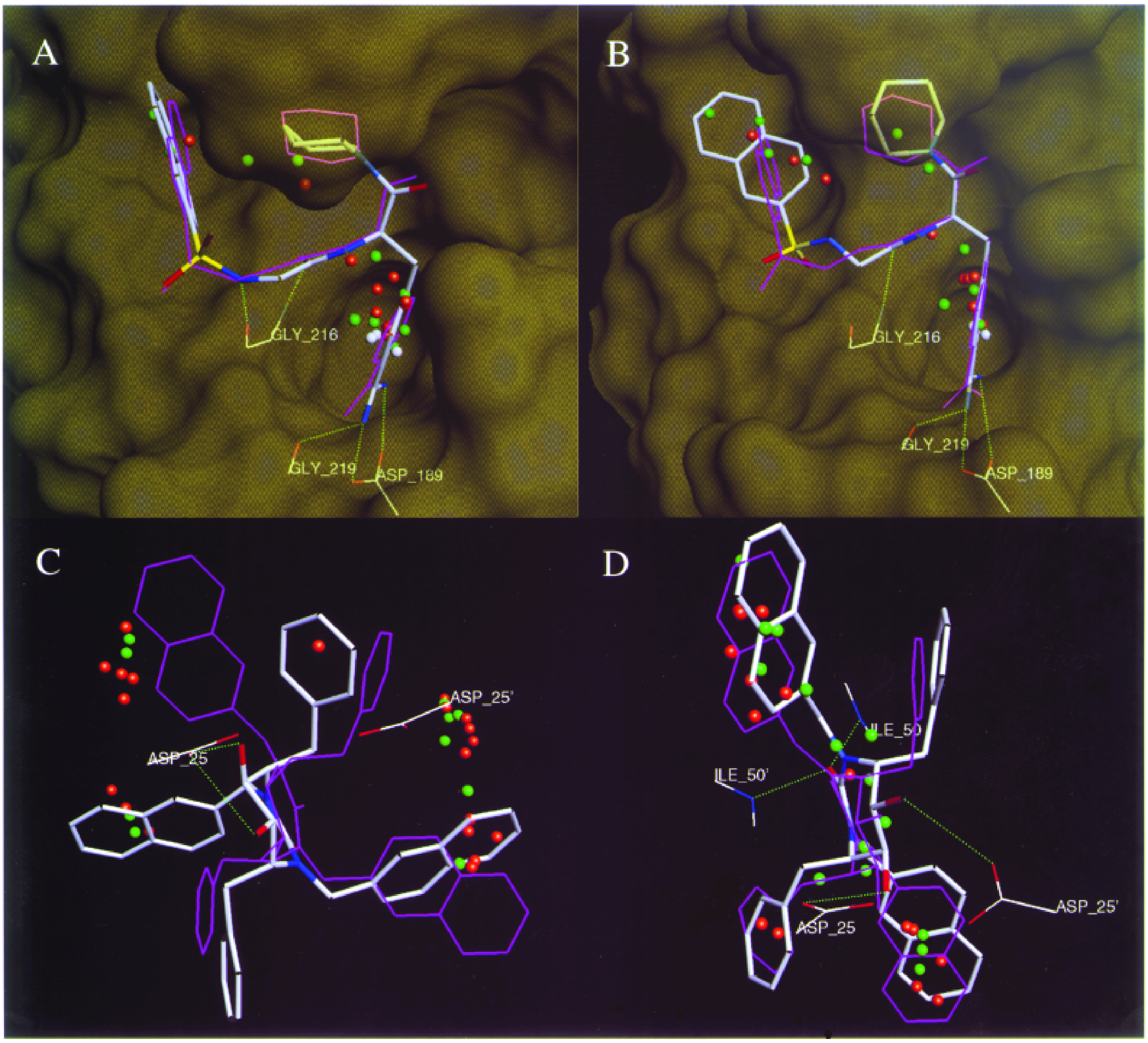

Fig. 2 FFLD Results on the NAPAP-Thrombin Complex (A,B) and the XK263-HIV-1 Protease Complex (C,D).
The minimized X-ray conformation of the inhibitor is shown in magenta. The structures docked by FFLD are shown by thick cylinder colored by atom type (carbon, white; nitrogen, blue; oxygen, red; sulfur, yellow; hydrogen atoms are not displayed). Some important protein side chains (see text) are also shown (atom type coloring). Intermolecular hydrogen bonds (donor-acceptor distance shorter than 3.5 Å) are indicated by green dotted lines and displayed only for the FFLD structures. (A) NAPAP docked to the uncomplexed structure of thrombin (1HGT) and (B) the structure from the thrombin-NAPAP complex (1DWD). (A–B) Geometrical centers of benzamidine, cyclohexane, and naphthalene used for NAPAP placement are indicated by white, green and orange spheres, respectively. (C) XK263 docked to the uncomplexed structure of HIV-1 protease (3HVP) and (D) the complexed form of HIV-1 protease (1HVR). (C–D) Geometrical centers of naphthalene and benzene used for XK263 placement are indicated by green and orange spheres, respectively.

thyl ring is 2.36 ± 0.27 Å. This is satisfactory considering the large displacement of the flaps. The best energy binding modes of the 5 other runs are located outside of the active site and can therefore be disregarded.

Discussion

We have presented FFLD, a flexible ligand docking approach based on a genetic algorithm search that uses a very efficient scoring function. The docking is performed by positioning and orienting the ligand in the rigid binding site according to the optimal binding modes of molecular fragments determined by an energy function that takes into account electrostatic solvation effects. This is the main advantage of FFLD with respect to the available docking programs which either neglect or crudely approximate solvation. One exception is a recent version of the DOCK program (Zou *et al.,* 1999) that uses an implicit solvent model, similar to the one developed by Scarsi *et al.* (1997). Another important difference is that FFLD uses optimally docked fragments to describe the protein binding site, whereas previous published programs employ descriptors determined by interactions of mainly local character. GOLD is based on the identification of hydrogen bonds between the ligand and the receptor and is expected to encounter difficulties for docking mainly hydrophobic compounds (Jones *et al.,* 1995). There is no such limitation in FFLD as the docking of a predominantly hydrophobic molecule can be driven by the optimal binding modes of apolar fragments determined by SEED. This has been shown in the FFLD docking of XK263 into HIV-1 proteinase. In FlexX, the algorithm for the placement of the base fragment exploits physicochemical properties which are assigned to every atom of the ligand and the receptor (Rarey *et al.,* 1996). This type of placement is based on short-range interactions and is expected to be less accurate in general than the SEED exhaustive fragment docking based on a force field energy. Another difference of FFLD with respect to FlexX is that the docking algorithm in FlexX is an incremental construction strategy which is based on the assumption that position and orientation of a partially grown ligand in the binding site will be also nearly optimal for the binding mode of the whole ligand. In FFLD such assumption is not made since the ligand is docked as a whole entity in the binding site of the receptor. Moreover, the FFLD approach includes implicitly solvent effects as the binding modes of the fragments are determined by an accurate energy function with electrostatic solvation (Majeux *et al.,* 1999, 2001).

The main limitation of FFLD is the assumption of a rigid protein. In cases where FFLD is used to determine the binding mode of a small set of compounds, the computational efficiency is of secondary importance. Several docking runs could therefore be performed and the best binding modes for each compound could be minimized in a (partially) flexible protein. Another limitation is the requirement of at least three anchoring fragments for the molecule to be docked. This might prevent the docking of poorly functionalized compounds for which the program SEED is probably more appropriate (Majeux *et al.,* 2001).

The decomposition of the ligand and the selection of the fragment functionality maps used for the ligand placement are performed manually in the current version of FFLD. It is planned to automatize them for the efficient high-throughput docking of large libraries of organic compounds. One important question concerns the number and diversity of SEED fragments necessary to fully represent the interacting moieties of a large collection of compounds. An interesting possibility would be to *a priori* consider synthetically accessible and toxicologically benign fragments by including in the SEED library common substituents of known drug molecules (Bemis and Murcko, 1999). In principle, the FFLD approach is suitable for computational high-throughput screening since the docking of a database of 10 000 compounds would require about two days on a cluster of ten PCs.

Materials and Methods

Overall Strategy

First, the program SEED is used to dock a library of mainly rigid functional groups into the binding site of the receptor (Majeux *et al.,* 1999, 2001). SEED determines functionality maps, *i. e.,* the most favorable binding modes of the fragment library (Figure 3A). Subsequently, three fragments of the ligand are chosen (Figure 3B), and for each of them the SEED functional group with the highest similarity to the fragment is selected. The corresponding functionality maps are then used as binding site descriptors to place the ligand by a least square fitting method (Kabsch, 1976) (Figure 3C). In the current version of FFLD the decomposition into fragments and the selection of SEED functional groups have not yet been automatized. The ligand docking is performed by a genetic algorithm that uses a fast scoring function. The genetic algorithm perturbations affect only the conformation of the ligand since its placement in the binding site is determined by the SEED functionality maps. The main advantage of this ap-

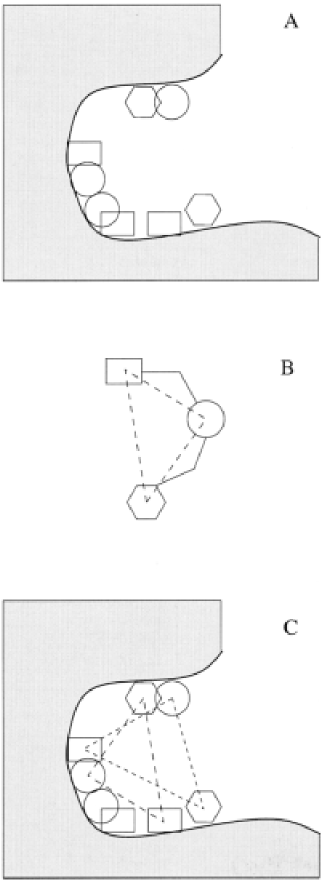

Fig. 3 Schematic Description of the FFLD Approach.
(A) SEED functionality maps are represented by three geometrical objects in a cartoon representation of the protein binding site. (B) Ligand conformation generated by the genetic algorithm. The three fragments used for ligand placement are schematized by a rectangle, a circle and an hexagon. The placement triangle defined by the geometrical centers of the ligand fragments is shown with dashed lines. (C) Two congruent triangles used for ligand placement are shown with dashed lines. Each triangle is defined by the geometrical centers of three functional group positions.

proach is that the position and orientation of the ligand in the binding site is determined by the best binding modes of small molecules previously docked using an accurate energy function with electrostatic solvation (Scarsi *et al.,* 1997). The scoring function used in FFLD to rank the ligand binding modes is based on van der Waals and hydrogen bond terms and does not include solvation for efficiency reasons and since solvation is taken into account during the docking of the functional groups.

SEED Functionality Maps

SEED is a computational approach for exhaustively determining favorable positions and orientations of small to medium-size molecular fragments in the binding site of a rigid protein and ranking them according to their binding energy with electrostatic solvation (Majeux *et al.,* 1999, 2001). The current SEED library consists of 70 mainly rigid fragments with between 7 and 31 atoms. It contains 17 apolar fragments (no hydrogen bond donors or acceptors), 39 polar and neutral compounds, and 14 fragments with one or two formal charges (Majeux *et al.,* 2001). Many of the molecular frameworks found frequently in known drugs (Bemis and Murcko, 1996) are included (*e. g.,* benzene, pyridine, naphthalene, 5-phenyl-1,4-benzodiazepine *etc.*) and some of them can be used for the synthesis of combinatorial libraries. Functional groups of mainly hydrophilic character are docked such that at least one hydrogen bond with the receptor is formed with close to optimal geometry. For this task, vectors on polar groups of the receptor and SEED fragments are defined automatically. Nonpolar molecules are docked in hydrophobic regions of the receptor (Scarsi *et al.,* 1999). For both polar and apolar fragments, the docking is exhaustive on a discrete space. The discretization originates from the finite number of preferred directions and rotations around them. The binding energy is the sum of electrostatic and van der Waals terms. The electrostatic contribution consists of screened intermolecular energy and receptor and fragment desolvation terms. It is evaluated efficiently by a numerical approach based on the continuum dielectric approximation in which the system is partitioned into solvent and solute regions that are assigned different dielectric constants (78.5 and 4.0, respectively, in the applications presented here). The binding modes determined by SEED are sorted according to binding energy and clustered by using a criterion based on distances between similar atom types. The functionality maps contain the best ranked positions of the first *n* clusters of each functional group where *n* can be specified by the user. Additional details about SEED and a complete description of the continuum electrostatic method have been presented in previous publications (Scarsi *et al.,* 1997, 1998; Majeux *et al.,* 1999, 2001; Scarsi and Caflisch, 1999). The SEED input parameters used for the test cases presented in this report are identical to those used in a previous study (Majeux *et al.,* 2001).

Hash Tables

For the three SEED fragments selected for ligand placement, the geometrical centers of their optimal binding modes are determined in a preprocessing step. To store this information for efficient placement of the ligand in the binding site, three hash tables are generated, one for each of the three pairs of SEED molecules. Each hash table stores in buckets of 0.5 Å all pairs of positions of its two fragments sorted according to the distance between geometrical centers. This allows efficient docking by fast access to the positions of pairs of functional groups whose distance falls within a given range. Distances shorter than 3.5 Å are not stored in the hash table since they correspond to conformations with internal sterical clashes.

Before generation of the hash tables, the fragment binding modes are clustered in SEED according to their position and orientation in the binding site using a criterion based on distances between similar atom types. An alternative would be to cluster the SEED binding modes according to the position of their geometrical centers instead of using all atoms. This would lead to an initial set of more heterogeneous values of the triplets of distances between fragments in the binding site, which might improve docking.

Genetic Algorithm

A genetic algorithm is a stochastic optimization method that mimics the process of natural evolution by manipulating a population of data structures called chromosomes (Goldberg, 1989; Davis, 1991). The genetic algorithm used in FFLD searches the regions of conformational space of the ligand that allow it to fit in the binding site. Each chromosome contains so-called genes that encode the values of the angles of rotation around the rotatable bonds of the ligand. Since covalent bond lengths and angles are kept rigid, a chromosome of N genes fully specifies the conformation of a molecule with N rotatable bonds. Starting from an initial randomly generated population of chromosomes, the genetic algorithm repeatedly applies two mutually exclusive genetic operators, one-point crossover and mutation, which yield new chromosomes (children) that replace appropriate members (parents) of the population. Both operators require parent chromosomes that are randomly selected from the existing population with a bias toward the fittest. At each cycle of genetic algorithm reproduction, the population of ligand conformations is docked in the binding site using the hash tables. A given ligand conformation can have several different locations in the binding site and is therefore assigned the score of its best binding mode. On the other hand, a ligand conformation is assigned a very high score when, according to the SEED functionality maps, it is not possible to place it in the binding site. The emphasis on the survival of the fittest introduces an evolutionary pressure into the algorithm and ensures that over time the population should move toward the conformation(s) representing the minimum of the scoring function, which approximates the binding energy (see below). For each conformational optimization by the genetic algorithm, a population of 100 chromosomes was used and 200 cycles were performed.

Docking of the Ligand Conformations Generated by the Genetic Algorithm

For each ligand conformation in the genetic algorithm population, the three distances between the geometrical centers of three ligand fragments are evaluated. The triplet of distances defines a triangle, called henceforth placement triangle, that is used to orient the ligand conformation within the binding site of the receptor. In the next step, the information stored in the three hash tables is used to determine all triangles of SEED functional groups that are congruent to the placement triangle. For each side of the placement triangle, pairs of geometrical centers are searched in buckets ranging from $n - m$ to $n + m$ in the corresponding hash table, where *n* is the index of the bucket whose distance range encompasses the length of the side of the placement triangle and *m* is initially equal to 0. A congruent triangle is defined by three pairs of geometrical centers, and each of the three pairs shares its geometrical centers with the two others (Figure 3C). If no congruent triangle is found, the value of *m* is incremented until either at least one congruent triangle is found, or *m* is larger than a user-defined value. In the applications presented here, a value of $m = 6$ was chosen which gives a tolerance of ± 3 Å. The ligand conformation is finally positioned in the binding site by matching the placement triangle to a congruent triangle of SEED functional groups using an algorithm that minimizes the square of the distance between the vertices of the two triangles (adapted from Kabsch, 1976). As mentioned above, for a given ligand conformation several congruent triangles can be extracted from the hash tables depending on the distribution of geometrical centers.

Scoring Function

The scoring function used to rank the ligand binding modes is

$$\Delta E_{\text{total}} = E_{\text{vdW}}^{\text{ligand}} + E_{\text{polar}}^{\text{inter}} + E_{\text{vdW}}^{\text{inter}} \tag{1}$$

The first term ($E_{\text{vdW}}^{\text{ligand}}$) is the van der Waals intraligand energy, which is needed to prevent steric clashes among atoms of the ligand. It is described as the sum of an attractive dispersion and a steep repulsion term with the 6 – 12 Lennard-Jones model, which is calculated explicitly for each pair of ligand atoms separated by at least three covalent bonds. Energy minima and optimal interatomic distances for the van der Waals energy were taken from the CHARMm22 parameter set (MSI Inc.). The interaction energy between ligand and receptor is described by the last two terms in equation (1), where $E_{\text{polar}}^{\text{inter}}$ and $E_{\text{vdW}}^{\text{inter}}$ are the polar and van der Waals energy contributions, respectively. These are detailed in the two following subsections.

Hydrogen Bonds and Unfavorable Polar Contacts

The ligand-receptor polar interaction term ($E_{\text{polar}}^{\text{inter}}$) is

$$E_{\text{polar}}^{\text{inter}} = n_{\text{HB}}^{\text{inter}} E_{\text{HB}} + n_{\text{UP}}^{\text{inter}} E_{\text{UP}} \tag{2}$$

where $n_{\text{HB}}^{\text{inter}}$and $n_{\text{UP}}^{\text{inter}}$ are the number of hydrogen bonds and the number of unfavorable polar contacts, respectively. E_{HB} and E_{UP} are approximated by constant values. Following criteria are used for the definition of a hydrogen bond: a distance between the acceptor and the hydrogen atom shorter than 2.5 Å, and a donor-H···acceptor angle larger than 130°. An additional check for clashes is performed between the donor hydrogen and eventual hydrogen(s) covalently bound to the acceptor atom. The lists of donor and acceptor atoms in the ligand and in the receptor are determined at the beginning of the program to improve calculation efficiency. Receptor atoms involved in an intra-receptor hydrogen bond can be excluded from the receptor list. An interaction between two donors or two acceptors is considered an unfavorable polar contact if the interatomic distance is smaller than 2.8 Å. The score values used for E_{HB} and E_{UP} are – 3.0 and +3.0 kcal/mol, respectively. Experimental investigations have suggested that a neutral hydrogen bond contributes between – 0.5 and – 1.5 kcal/mol to the binding energy and up to – 4.7 kcal/mol in the case of a salt bridge (Davis and Teague, 1999). As a compromise a score value of – 3.0 kcal/mol is used for E_{HB}. A somewhat arbitrary value of +3.0 kcal/mol is assigned to E_{UP}.

Soft Core van der Waals

Since the protein is rigid, its van der Waals potential is mapped on look-up tables to speed-up the calculation of $E_{\text{vdW}}^{\text{inter}}$ (Majeux *et al.,* 2001). The asymptotic behavior of the Lennard-Jones repulsive part is however inappropriate for flexible ligand docking because it penalizes binding modes with small atomic interpenetrations with the protein surface even if they are very close to energy minima. Furthermore, the repulsive part is not able to differentiate between steric clashes at the surface and in the interior of the protein. A soft core van der Waals, which is compatible with the look-up table mapping, is used in FFLD to address these limitations.

The volume occupied by the protein is first divided into two distinct regions (Figure 4A). The first region represents a 1 Å layer below the protein molecular surface. The remaining protein volume is assigned to the inner region. A three-dimensional grid (grid spacing of 0.3 Å) is used to discretize the volume of the protein binding site. Each grid point contains the information (atom type and coordinates) of its closest protein atom. Upon ligand placement the program first checks for each ligand atom whether it is located within the inner region of the protein in which case its contribution $E_{\text{vdW}}^{\text{inter}}$ is a somewhat arbitrary penalty of 150 kcal/mol (severe clash). Otherwise, the data of the protein atom closest to the ligand atom are extracted from the grid and the van der Waals interaction energy is calculated. The interaction energy is linearized if its value is higher than a cutoff (Figure 4B) in which case the contributions from the remaining protein atoms are neglected. Otherwise, the ligand atom van der Waals energy contribution is calculated by trilinear interpolation using the two protein look-up tables (repulsive and attractive terms), which take into account the contribution of all receptor atoms.

System Setup

The following X-ray structures were downloaded from the PDB database (Berman *et al.,* 2000): uncomplexed thrombin (code 1HGT; Skrzypczak-Jankun *et al.,* 1991), thrombin complexed with Nα-(2-naphthyl-sulfonyl-glycyl)-(DL)-*p*-amidinophenylalanyl-piperidine (NAPAP, Figure 1A, code 1DWD; Banner and Hadvary, 1991), human immunodeficiency virus type 1 (HIV-1) protease (code 3HVP; Wlodawer *et al.,* 1989), and HIV-1 protease complexed with the cyclic-urea inhibitor XK263 (Figure 1B, code 1HVR; Lam *et al.,* 1994). The water molecules and the inhibitor (if present) were removed. Hydrogen atoms were added and minimized with the CHARMM program (Brooks *et al.,* 1983).

The NAPAP fragments used for ligand placement in the active site of thrombin are benzamidine, piperidine and naphthalene. The corresponding functional groups docked by SEED in 1HGT and 1DWD are benzamidine, cyclohexane and naphthalene, respectively. The 9 rotatable bonds of NAPAP were flexible during docking. The fragments used for the placement of XK263 in the active site of HIV-1 protease are benzene and naphthalene (the latter twice). These functional groups were docked by SEED in the active site of both complexed (1HVR) and native (3HVP) enzymes. The 10 rotatable bonds of XK263 were flexible during docking.

For the FFLD docking, the binding site was defined as the smallest parallelepiped which encompasses the residues with one or more atoms within a distance cutoff from the inhibitor. For 1HGT, 1DWD and 1HVR the cutoff distance was 5 Å whereas for 3HVP a value of 6 Å was used to take into account the large displacements of part of the protein upon binding (see below). As reference structures for the root mean square (RMS) deviation calculation, the inhibitor position in 1HGT and 3HVP was obtained from the corresponding complexed X-ray structure by superposing the C_α atoms of the two protein conformations and CHARMM minimization with a distance dependent dielectric function [$\in(r) = 4r$]. For consistency reasons, the same minimization protocol was applied to the X-ray inhibitors in 1DWD and 1HVR.

Computation Times

All calculations were carried out on a single 800 MHz Pentium III processor. One FFLD docking run of NAPAP into thrombin took between 2 and 3 minutes, while the docking of XK263 in the complexed and uncomplexed conformation of HIV-1 protease required 2 and 4 minutes, respectively. The variance is due to differences in the hash tables which originate from different distributions of the SEED geometric centers. These timings do not in-

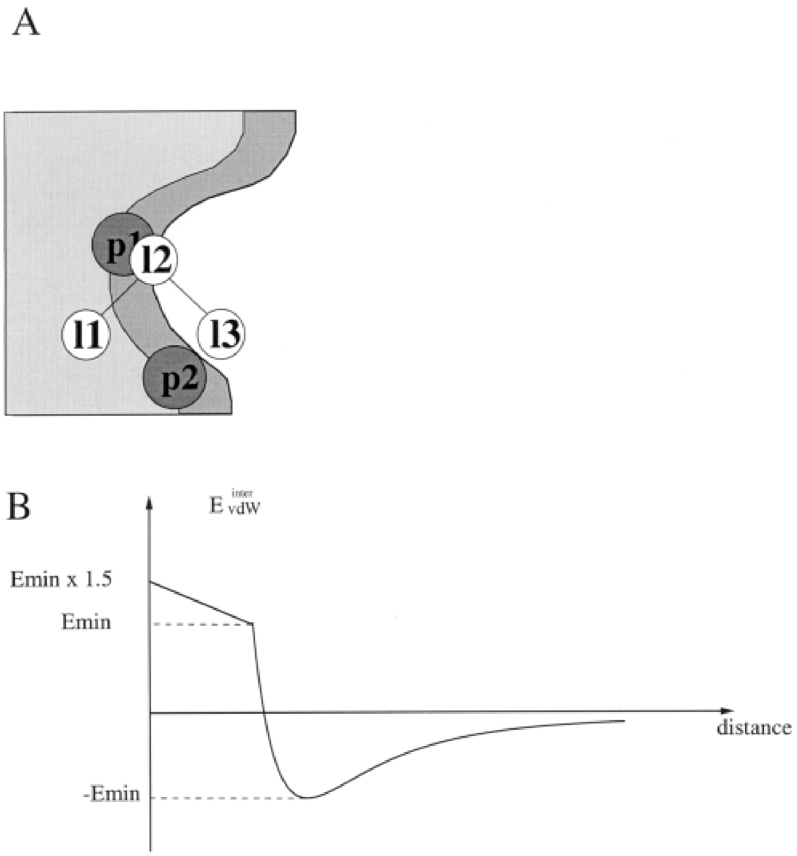

Fig. 4 Evaluation of van der Waals Energy between Ligand and Protein, E_{vdW}^{inter}.
(A) Two regions are defined within the volume occupied by the protein binding site. These are the inner region (light grey) and a thin layer below the molecular surface (dark grey). The former exemplifies a severe clash while the latter a minor overlap. Two atoms of the protein are represented by dark grey circles labeled p1 and p2. The ligand is simplified by three atoms represented by circles labeled from l1 to l3. Atom l1 is located within the innermost protein volume (light grey region) and is assigned a clash penalty of 150 kcal/mol. Since atom l2 clashes into the thin layer below the molecular surface, only its linearized Lennard-Jones potential with its closest protein atom (p1) is calculated. For atom l3 the full van der Waals energy with the protein is evaluated. (B) Soft core van der Waals. The x and y axes are the interatomic distance and the interaction energy, respectively. The van der Waals interaction energy minimum is –Emin. The procedure used for the linearization is adapted from Figure 5 of Gehlhaar *et al.* (1995). For a given pair of atoms, the linearization energy cutoff corresponds to the minima of their van der Waals energy multiplied by – 1. The slope is such that for a zero interatomic distance the linearized interaction energy is equal to the cutoff value multiplied by 1.5.

clude the docking of the fragment library by SEED (Majeux *et al.,* 2001), which has to be performed only once for a given protein target.

Acknowledgements

We thank Drs. J. Apostolakis, C. Ehrhardt and C. Tenette-Souaille for useful discussions. The FFLD program for PC running the Linux operating system is available from A. Caflisch. This work was supported by Novartis Pharma Inc. and the Swiss National Science Foundation (grant no. 31-53604.98).

References

Apostolakis, J., and Caflisch, A. (1999). Computational ligand design. Comb. Chem. High Throughput Screen. *2*, 91 – 104.

Banner, D. W., and Hadvary, P. (1991). Crystallographic analysis at 3.0 Å resolution of the binding to human thrombin of four active site-directed inhibitors. J. Biol. Chem. *266*, 20085 – 20093.

Bemis, G., and Murcko, M. A. (1996). The properties of known drugs. 1. Molecular frameworks. J. Med. Chem. *39*, 2887 – 2893.

Bemis, G., and Murcko, M. A. (1999). Properties of known drugs. 2. Side chains. J. Med. Chem. *42*, 5095 – 5099.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. Nucleic Acids Res. *28*, 235 – 242.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. (1983). CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J. Comput. Chem. *4*, 187 – 217.

Davis, A. M., and Teague, S. J. (1999). Hydrogen bonding, hydrophobic interactions, and failure of the rigid receptor hypothesis. Angew. Chem. Int. Ed. *38*, 736 – 749.

Davis, L. (1991). Handbook of Genetic Algorithms (New York, USA: Van Nostrand Reinhold).

Gehlhaar, D. K., Verkhivker, G. M., Rejto, P. A., Sherman, C. J., Fogel, D. B., Fogel, L. J., and Freer, S. T. (1995). Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. Chem. Biol. *2*, 317–324.

Goldberg, D. E. (1989). Genetic Algorithms in Search Optimization and Machine Learning (Reading, USA: Addison-Wesley).

Jones, G., Willett, P., and Glen, R. C. (1995). Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. J. Mol. Biol. *245*, 43–53.

Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. Acta Cryst. *A32*, 922–923.

Kuntz, I. D. (1992). Structure-based strategies for drug design and discovery. Science *257*, 1078–1082.

Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R., and Ferrin, T. E. (1982). A geometric approach to macromolecule-ligand interactions. J. Mol. Biol. *161*, 269–288.

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. (1994). Rational design of potent bioavailable nonpeptide cyclic ureas as HIV protease inhibitors. Science *263*, 380–383.

Majeux, N., Scarsi, M., Apostolakis, J., Ehrhardt, C., and Caflisch, A. (1999). Exhaustive docking of molecular fragments on protein binding sites with electrostatic solvation. Proteins: Struct. Funct. Genet. *37*, 88–105.

Majeux, N., Scarsi, M., and Caflisch, A. (2001). Efficient electrostatic solvation model for protein-fragment docking. Proteins: Struct. Funct. Genet. *42*, 256–268.

Miller, M., Schneider, J., Sathyanarayana, B. K., Toth, M. V., Marshall, G. R., Clawson, L., Selk, L., Kent, S. B. H., and Wlodawer, A. (1989). Structure of complex of synthetic HIV-1 protease with a substrate-based inhibitor at 2.3 Å resolution. Science *246*, 1149–1152.

Rarey, M., Kramer, B., Lengauer, T., and Klebe, G. (1996). A fast flexible docking method using an incremental construction algorithm. J. Mol. Biol. *261*, 470–489.

Scarsi, M., and Caflisch, A. (1999). Comment on the validation of continuum electrostatics models. J. Comput. Chem. *14*, 1533–1536.

Scarsi, M., Apostolakis, J., and Caflisch, A. (1997). Continuum electrostatic energies of macromolecules in aqueous solutions. J. Phys. Chem. *A101*, 8098–8106.

Scarsi, M., Apostolakis, J., and Caflisch, A. (1998). Comparison of a GB solvation model with explicit solvent simulations: potentials of mean force and conformational preferences of alanine dipeptide and 1,2-dichloroethane. J. Phys. Chem. *B102*, 3637–3641.

Scarsi, M., Majeux, N., and Caflisch, A. (1999). Hydrophobicity at the surface of proteins. Proteins: Struct. Funct. Genet. *37*, 565–575.

Skrzypczak-Jankun, E., Carperos, V. E., Ravichandran, K. G., Tulinsky, A., Westbrook, M., and Maraganore, J. M. (1991). Structure of the hirugen and hirulog 1 complexes of α-thrombin. J. Mol. Biol. *221*, 1379–1384.

Wang, J., Kollman, P. A., and Kuntz, I. D. (1999). Flexible ligand docking: a multistep strategy approach. Proteins: Struct. Funct. Genet. *36*, 1–19.

Wlodawer, A., Miller, M., Jaskolski, M., Sathyanarayana, B. K., Baldwin, E., Weber, I. T., Selk, L., Clawson, L., Schneider, J., and Kent, S. B. H. (1989). Conserved folding in retroviral proteases: crystal structure of a synthetic HIV-1 protease. Science *245*, 616–621.

Zou, X., Sun, Y., and Kuntz, I. D. (1999). Inclusion of solvation in ligand binding free energy calculations using the generalized-Born model. J. Am. Chem. Soc. *121*, 8033–8043.

Received April 2, 2001; accepted July 5, 2001