

LEA3D: A Computer-Aided Ligand Design for Structure-Based Drug Design

Dominique Douguet,^{*,†} H el ene Munier-Lehmann,^{‡,§} Gilles Labesse,[†] and Sylvie Pochet^{||}

Centre de Biochimie Structurale (CNRS UMR 5048, INSERM UMR U554), Facult e de Pharmacie, Universit e Montpellier I, 15, avenue Charles Flahault, 34060 Montpellier Cedex, France, and Unit e de Chimie Organique (URA CNRS 2128) and Laboratoire de Chimie Structurale des Macromol ecules (URA CNRS 2185), Institut Pasteur, 28 rue du Dr Roux, 75724 Paris Cedex 15, France

Received September 22, 2004

We present an improved version of the program LEA developed to design organic molecules. Rational drug design involves finding solutions to large combinatorial problems for which an exhaustive search is impractical. Genetic algorithms provide a tool for the investigation of such problems. New software, called LEA3D, is now able to conceive organic molecules by combining 3D fragments. Fragments were extracted from both biological compounds and known drugs. A fitness function guides the search process in optimizing the molecules toward an optimal value of the properties. The fitness function is built up by combining several independent property evaluations, including the score provided by the FlexX docking program. One application in *de novo* drug design is described. The example makes use of the structure of *Mycobacterium tuberculosis* thymidine monophosphate kinase to generate analogues of one of its natural substrates. Among 22 tested compounds, 17 show inhibitory activity in the micromolar range.

Introduction

The process of designing new molecules possessing desired physical, chemical, and biological properties is an important and difficult problem in the chemical, material, and pharmaceutical industries. The traditional approaches involve a laborious and expensive trial-and-error procedure. To create new molecules with the desired profiles, hundreds of molecules may be synthesized and tested in many biological test systems with the objective of finding a class that is suitable for development as a drug. Alternatively, experimental and virtual screenings might be performed to mine huge libraries of molecules in order to identify lead compounds.¹ Usually, such procedures identify micromolar ligands that have to be improved.

Computer-aided drug design (CADD) is currently developed to increase the efficiency of the drug discovery process. It can form a valuable partnership with experiments by providing estimates when experimental approaches are difficult or expensive and by coordinating the experimental data available. These *in silico* methods encompass pharmacophoric model identification or QSAR (quantitative structure–activity relationship) analysis. But these can only be applied to the “forward” problem, which requires the computation of physical, chemical, and biological properties from the molecular structure of known active molecules.² Indeed, much less attention has been paid to the “inverse” problem, which requires the identification of the appropriate molecular structure given the desired physicochemical properties.

On the contrary, *de novo* drug design programs can identify novel molecular structures that are predicted to fit the active site of a target protein. In fact, some of them already have been applied to the lead generation for actual drug targets.^{3–7} For example, evolutionary *de novo* design by the program TOPAS led to the identification of a new molecular scaffold that served as a lead structure candidate for a novel Kv1.5 blocking agent.⁸ In another example, the program LUDI was used to generate ideas for the *de novo* design of new FKBP-12 ligands based on the core structure of FK506.⁹ To achieve the *de novo* lead compounds, several computer programs or methodologies have been proposed, including LEGEND,¹⁰ LUDI,¹¹ SPROUT,¹² HOOK,¹³ GrowMol,¹⁴ PRO-LIGAND,¹⁵ CONCERTS,¹⁶ and LeapFrog.¹⁷ Compared to virtual screening by docking of known molecules in compound databases, *de novo* ligand design programs can generate novel active structures fitting the active site of the target protein while efficiently searching the whole chemistry space. But one has to solve a huge combinatorial and nonlinear structure–property correlation problem for which an exhaustive search is impractical.

Among structure-based *de novo* drug design programs, LeapFrog,¹⁷ ADAPT,¹⁸ LigBuilder,¹⁹ PRO-LIGAND,¹⁵ TOPAS,⁸ and the one by Blaney et al.²⁰ apply genetic algorithms (GA) to the design of new molecular structures. The essence of a GA lies in allowing a dynamically evolving population of molecules to be gradually improved by competing for the best performance or fitness. Many studies have shown that GAs are able to locate optimal solutions for many desired target constraints and are able to discover a diverse population of near-optimal solutions (e.g., several new families of structures).²¹

De novo design programs try to assemble molecules by using some building blocks, which could be either

* Corresponding author. Phone (33) 4 67 04 38 42. Fax (33) 4 67 52 96 23. E-mail: douguet@cbs.cnrs.fr.

† Centre de Biochimie Structurale.

‡ Laboratoire de Chimie Structurale des Macromol ecules, Institut Pasteur.

§ Present address: Unit e de Chimie Organique, Institut Pasteur.

|| Unit e de Chimie Organique, Institut Pasteur.

atoms or chemical fragments. Our previous version, LEA, used atom-based construction by using the SMILES line notation.^{22,23} Usage of single atoms as building blocks will give the maximum diversity, because all organic structures can be generated by assembling atoms. However, according to our own experience in using LEA, atom-based construction suffers from generating unreasonable structures. In addition, more steps are required to build up the whole molecule through an atom by atom process. We have developed a new fragment-based algorithm: LEA3D.

For this purpose, we selected 7621 drugs from the Comprehensive Medicinal Chemistry database (CMC from MDL Information Systems, May 2003) and 7282 compounds from the LIGAND database of KEGG (<http://www.genome.ad.jp/dbget/ligand.html>). The last one has been added because the hit rates in high-throughput screens determined for natural product collections are often dramatically higher than the rates found for large classical libraries.^{24,25} We computationally dissected them into rings and acyclic parts. This resulted in approximately 8000 fragments that led potentially to a virtual size library of 10^{14} (assuming four as the maximum number of building blocks a molecule can combine). Despite the huge size of the virtual library, it is still smaller than the “virtual chemistry space” evaluated to 10^{60} within a molecular weight (MW) of less than 500.²⁶ However, the space to search remains immense and a GA avoids the trap of fully enumerating the virtual library.

The new version of the program LEA, LEA3D, takes advantage of the new fragment library to design molecules. It is hoped that LEA3D designs are both chemically feasible and have favorable druglike properties, since the fragments were originally obtained from known bioactive molecules. Furthermore, LEA3D can build up molecules under the structural constraints of the target protein in order to fill optimally the active site of the receptor. The protein–ligand binding interactions are evaluated by using the FlexX docking program.²⁷ The derived score guides the design of the putative ligands of the following generation (optimization). FlexX is also a fragment-based docking method that divides the ligand into separate portions and uses an incremental construction algorithm. Usually the fragments selected by FlexX are the same as the building blocks used to create the molecule by LEA3D. Furthermore, combining two fragment-based approaches for designing and scoring molecules is an advantage. FlexX allows the user to dock a molecule (or a fragment) onto a reference substructure, molecule, or fragment already placed into the active site. In this way, we can both constrain a crucial fragment into the molecule designed by LEA3D (to maintain the affinity with the receptor) and request FlexX to superimpose it onto the reference within the receptor. The calculation of the docking score is highly accelerated and the docking placement may be improved.

To evaluate the new method, LEA3D was applied to the binding site of the thymidine monophosphate kinase of *Mycobacterium tuberculosis* (TMPKmt). When the structure of the target protein is available, the process of lead generation and optimization can be profoundly influenced and speeded, particularly when the three-

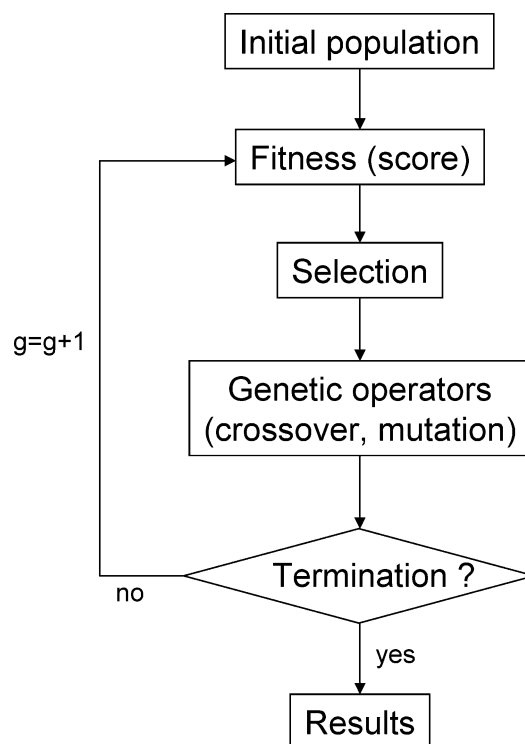


Figure 1. General flowchart for a genetic algorithm. An initial population of candidate solutions is generated, usually, by random process. The fitness of each candidate is evaluated via a fitness function (or score), which takes as input a candidate solution and returns a numeric score. Selection criteria are applied to choose candidates on the basis of their fitness score for breeding. Breeding functions (crossover and mutation) are applied to produce new solutions, which replace the parent solutions. The cycle (or generation g) continues until convergence criteria is met (usually, solutions are no more improved).

dimensional structure of the protein–ligand complex is available, as is the case for TMPKmt. Results of the virtual screening will be presented and the biological evaluation of the selected molecules will be discussed.

Methods

Genetic Algorithm. GAs are inspired by natural selection in evolution.²⁸ GAs approach the optimum of a given function in the same way nature selects the individual fittest to the environment. The genetic algorithm vocabulary is adopted from natural genetics. Populations and generations consist of a set of individuals, also referred to as chromosomes or genotypes. A population of individuals evolves through selection of various mutations of the individuals and recombination between individuals. A GA is usually implemented using the following procedure: evaluate the fitness of all the individuals in the population; create a new population by performing operations such as crossover (fitness-proportionate reproduction) and mutation on the individuals whose fitness has just been measured; discard the old population and iterate using the new population. One cycle of this loop is referred to as a generation. The first generation (generation 0) usually operates on a population of randomly generated individuals (see Figure 1). From there, the genetic operations, in concert with the fitness measure, operate to improve the population. The GA uses a blind search strategy, requiring no knowledge of the properties of the

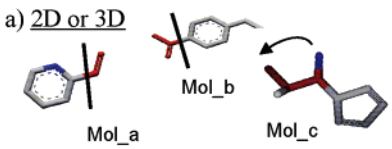
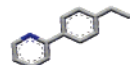

Molecular representation	Genetic operators (crossover, mutation)	
a) <u>2D or 3D</u> 	Result of the crossover operation between Mol_a and Mol_b 	Result of the 'cyclisation' mutation operation on Mol_c 
b) <u>SMILES line notation:</u> Mol_a <chem>N1=CC=CC=C1C=O</chem> Mol_b <chem>C1(C([O-])[O-])=CC=C(CC)C=C1</chem> Mol_c <chem>C1=CC=CC1C(N)C(C)C</chem>	<chem>N1=CC=CC=C1C1=CC=C(CC)C=C1</chem>	<chem>C1=CC=CC1C1=NC=CC=C1C</chem>
c) <u>Fragment-based notation:</u> Mol_a 106-2 Mol_b 37-40 Mol_c 56	106-40	Mutation by 'cyclisation' is prohibited

Figure 2. Examples of molecules with their 3D, SMILES line, and fragment-based representation. Mol_a and Mol_b are combined to produce one offspring by the mean of a crossover operation. Mol_c is modified by the “cyclization” mutation operator. (a) 3D representation of the three molecules. The lines which cut Mol_a and Mol_b represent the position where the molecule will be split into two portions before the recombination. The arrow above Mol_c indicates where the cyclization will occur. Results of the crossover and mutation operators are depicted. (b) SMILES line representation of the three molecules. Distinction is made between skeleton and branched groups in brackets. SMILES strings are manipulated (rewritten, segmented) with respect to chemical and SMILES line notation rules. Several, but fastidious, string analyses and combinations lead to the resulting chemical structures. (c) Fragment-based representation is a linear one. Although combinations are more restrict, generated molecules are various and the space of structures is large enough for our purpose. The cyclization mutation operator is ineffective in this case, since we are not able to modify a fragment.

function to be optimized, thus enabling the algorithm to be applied to a variety of optimization problems from robot behavior²⁹ to drug design.^{30–32}

The Molecular Representation. LEA and LEA3D use a GA to explore the molecular structure space. The character string of the chromosome is used to encode the values for different parameters being optimized as well as to encode a molecule. LEA and LEA3D differ significantly in this way. In LEA, we used a concept of molecular assembly encoded in SMILES line notation, which is useful for communication between chemistry softwares and storage. However, its manipulation by crossover and mutation operators is very fastidious and the generated molecules are often difficult to synthesize. In LEA3D, the molecular assembly is a linear combination of one to five fragments and each fragment is represented by a number (see Figure 2). Although fragment-based notation cannot code for all organic structures, its covering space of structures is still huge. Moreover, it speeds up the ligand construction and limits the number of unrealistic molecules. Each generated molecule is optimized by using the software CORINA that provides one to four conformers (user-defined number). This conformer search is crucial for ring systems, which are not flexible during a FlexX docking calculation.

Fragment Library. Molecular fragments found in successful drugs are more likely to be “druglike” in a general sense (usually satisfying the “rule-of-five” devised by Lipinski and co-workers and not possessing reactive functional groups) than random molecules.³³ Our fragment library has been constructed upon the

8474 drugs from the CMC and 7434 “biomolecules” from the LIGAND database of KEGG. Our first task was to identify and remove nonorganic molecules and to remove the counterion of salts from databases. Then, three-dimensional structures were generated by CORINA and only structures converted to 3D were kept.³⁴ Thereafter, the CMC database had 7621 remaining entries and the KEGG database had 7282 remaining entries. Then, molecules have been dissociated into single rings, fused rings, and acyclic parts (see Figure 3 and Table 1).

Comparisons of the two databases show that drugs are segmented into more fragments than biomolecules: 7621 drug molecules give 50 617 fragments (664%) and 7282 biomolecules give 41 208 fragments (565%). On average, a drug molecule is composed of 6.6 fragments (1.3 single ring, 0.5 fused ring, and 4.8 acyclic groups) and a biomolecule is composed of 5.6 fragments (0.9 single ring, 0.5 fused ring, and 4.2 acyclic groups).

Both databases show some fragment redundancy: only 5274 (10.4%) of 50 617 are unique for drug molecules and 4142 (10%) of 41 208 are unique for biomolecules. The proportion of unique fused ring system in the “biofragment” database is larger (45% of 4142) than in the drug fragment database (34% of 5274). Conversely, the proportion of unique acyclic groups in the drug fragment database is larger (47% of 5274) than in biofragment database (36% of 4142).

Analysis of fragment families showed that the “bio-database” contains more unique fused ring systems (three fused rings and above, 16.4% of 7282) than the drug database (12.5% of the 7621). On the contrary,

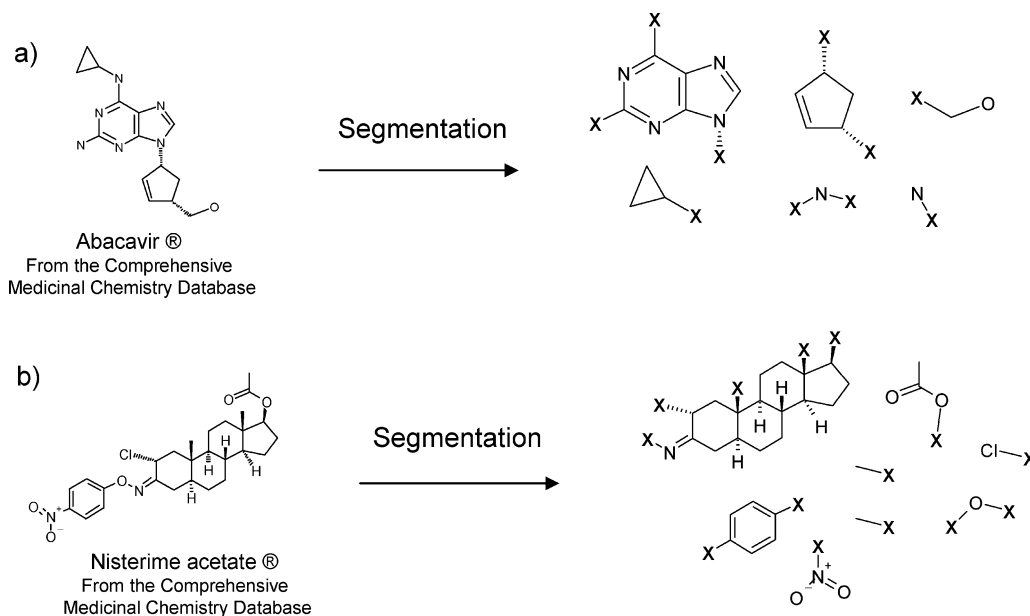


Figure 3. Two examples of segmentation. Molecules are dissociated into ring systems, fused-ring systems, and acyclic parts. Generated fragments keep the substitution pattern of the original molecule by replacing substituents by “X” dummy atoms.

Table 1. 3D Fragment Database Was Extracted from Both 7621 Known Drugs (CMC) and 7282 Biological Compounds (KEGG)

fragment classes	CMC	KEGG	fragments		
			distinct ^a	e ^b	f ^c
single ring					
three atoms	24 (0.31%)	21 (0.28%)	29	6	123
four atoms	23 (0.3%)	7 (0.09%)	26	6	115
five atoms	294 (3.8%)	225 (3%)	411	84	2138*
unsaturated, five atoms	104 (1.3%)	59 (0.8%)	112	28	424
six atoms	414 (5.4%)	375 (5.1%)	647	92	5354*
aromatic, six atoms	80 (1%)	51 (0.7%)	85	9	793
seven atoms	21 (0.27%)	13 (0.17%)	30	11	142
eight atoms	5 (0.06%)	4 (0.05%)	8	3	24
two fused rings	166 (2.2%)	135 (1.8%)	262	40	1517*
six and five atoms	191 (2.5%)	151 (2.5%)	305	71	1396*
aromatic, six and five atoms	137 (1.8%)	60 (0.82%)	152	38	544*
six and six atoms	279 (3.6%)	287 (3.9%)	485	69	3423*
aromatic, six and six atoms	83 (1%)	51 (0.7%)	104	11	730*
three or four fused rings	795 (10.4%)	832 (11.4%)	1446	179	12988*
five and more fused rings	162 (2.1%)	371 (5%)	495	25	6126*
acyclic parts					
15–20 atoms	1342 (17.6%)	699 (9.6%)	1665	1028	1631**
1–15 atoms	142 (1.8%)	105 (1.4%)	166	96	165*
>20 atoms	1012 (13.2%)	696 (9.5%)	1558	966	1936**
total number of fragments	5274	4142	7986	2762	39569

^a Refers to distinct fragments after merging CMC and KEGG distinct fragments. ^b e is a fragment that can only bind to a single fragment (only one substitution point). ^c f is a fragment that has more than one substitution point. Due to the combinatorial explosion of pairs (Figure 4), we sometimes set the maximum number of combinations at 5 (*) or at 10 (**) for each fragment.

unique fused two-ring systems are more numerous in the drug database, probably reflecting a medicinal chemistry bias (7.5% versus 5.3% in biodatabase) except for the 6–6 atoms nonaromatic system (3.6% in drugs and 3.9% in biomolecules).

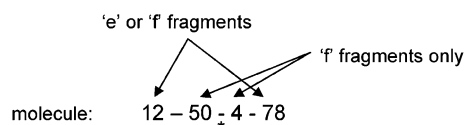
Moreover, the drug database contains more unique single ring systems (12.6% of 7621) than the biodatabase (10.3% of 7282). Finally, the drug database contains more unique acyclic groups than the biodatabase (32.7% and 20.5% respectively).

Broadly, both CMC and KEGG databases highly contribute to the final, merged fragment database. Most of their fragments are different. Drug molecules provide more diversity in single ring and acyclic fragments and biomolecules provide more diversity in fused ring systems (three and above). We obtained 2778 different

scaffolds (ring family) from the CMC drugs and 2642 from the KEGG biomolecules.

Our final fragment library contains 7986 categorized fragments stored in MDL sdf (Standard Derwent File) format upon one conformation. Each fragment keeps information about its original substitutions by possessing an “X” dummy atom at these positions (Figure 3). In this way, we respect the substitution pattern of the fragment and, if necessary, we can retrieve the original, parent molecule from CMC and KEGG databases.

For the purpose of the linear combination of fragments, we distinguished fragments that can be substituted at least twice from those which can be substituted only once (see Figure 4a). Thus, a fragment is either classified “e” (end) or “f” type. An e fragment is a fragment with only one substitution point (only one X

a) Fragment-based notation:b) The tag combination:

A 'f' fragment with
3 substitution points:



6 combinations
including the
symmetric ones

'left' tag	'right' tag
5	2
2	5
5	3
3	5
2	3
3	2

Figure 4. Fragment characteristics. (a) In a fragment-based representation, molecules are a combination of "e" fragments and "f" fragments. An e fragment is only used as the first or the last fragment to be combined. These fragments had only one substituent in their original molecule. They possess only one "X" dummy atom. An f fragment may be present at any position in the linear representation. An f fragment possess a "left" tag and a "right" tag, which indicate how to combine this fragment with another one. The tag indicates the number of the atom which bonds the X dummy atom. (*) The right tag of a defined fragment (here number 50) will be bonded to the left tag of the following fragment (here number 4) and so on. (b) Our linear representation forces us to generate combinatorial pairs of tags as well as the symmetric ones.

dummy atom) and f fragments are the others (multiple substitution points). In the context of a linear combination, fragments must possess one (e fragment) or two tags (f fragment) that indicate which atoms have to be bonded with another fragment. The "left" and "right" tags are the number of the atoms that bonds the previous X dummy atom. If a fragment possesses more than one substitution point, then a combination of pairs of tags is generated including the symmetric one (Figure 4b). This information is stored in the data block of the sdf file of the fragment. Thus, one fragment can be registered more than once but with different tags. Permanent tags are important to get reproducible molecules. Finally, we obtained a fragment library of 39 569 building blocks (see Table 1).

Fitness Function. There are many different classes of information that can play an important role in focusing the search space: first, the knowledge about known compounds that interact with the receptor and which exhibit some of the needed molecular properties and, second, the knowledge about the receptor structure. When the structure of the target protein is known, receptor-based computational methods can be employed. This information helps to streamline the search in the huge molecule space. Currently, LEA3D handles the output of the docking program FlexX (version 1.13.1) in order to perform a structure-based drug design process. This is another significant difference with the previous program LEA. By using LEA3D, FlexX output solutions are filtered in order to extract solutions that possess a docking score better than the user-defined one. Then, only one representative per binding mode is selected. The one with the best docking score and which is in the user-defined vicinity of the active site is kept. The user may also define several protein-ligand interaction constraints that the molecule must have with the receptor. In this way, we can focus our search on molecules that possess these crucial interactions. The

Molecular properties:

Polar surface area *

Volume *

Area *

Molecular Weight

LogP §

Molar Refractivity §

Radius of gyration: globularity index

Inertial axis along x, y and z

Length: longest distance between atoms in the molecule

Number of hydrogen bond donors

Number of Hydrogen bond acceptors

5 rules of Lipinski

searchable functions:

acid, ester, carbamate, amide, amide-ter, aldehyde, keto, amine, amine1, amine2, amine3, alcohol, alcohol1, alcohol2, alcohol3 and ether

Figure 5. List of available molecular properties. *Calculation using the program NSC V2.0 by Eisenhaber F.⁵² §Calculation using Ghose and Crippen atomic contribution method.⁵³ Amine1 refers to a primary amine, amine2 refers to a secondary amine, amine3 refers to a tertiary amine, and amine refers to any type of amine. Alcohol1 refers to a primary alcohol, alcohol2 refers to a secondary alcohol, alcohol3 refers to a tertiary alcohol, and alcohol refers to any type of alcohol.

docking score is defined as follow:

$$\text{Score}_{i,(p=\text{docking})} = - \frac{\text{FlexX_Score}_i}{60} \quad (1)$$

where p is the docking property and -60 is the optimal value of the FlexX score (the aim). FlexX_Score_i is negative or set at 0. The lower the score, the better the docking score.

The global fitness value of a molecule is given by combining its single scores of various molecular properties (molecular weight, number of atoms, surface area, presence of a defined function, etc.; see Figure 5) and FlexX docking. The composite fitness, in percentage terms, for an individual *i* generated by LEA3D is

$$\text{Score}_i = \sum_p W_p \times \text{Score}_{i,p} \quad (2)$$

W_p is the user-defined weight applied to the property *p* (the sum is taken over all the selected properties). The weights depend on the relative importance of the properties the user considered. But an alternative would set these weights on the basis of a previous QSAR analysis if a linear correlation exists.

Protein Specificity. Docking on related proteins of the target receptor may be necessary in order to predict potential side-effects.³⁵ For this purpose, LEA3D offers a structure-based drug design on many proteins (up to 10). Logic operators are applied for searching specific molecules (operator NOT), for creating several niches in the population (operator OR) or for creating molecules which have to bind two targets at the same time (operator AND).

Selection Method. The first generation operates on a population of randomly generated or user-defined molecules. Molecules are then evaluated by fitness

$$SV_i = \text{Score}_{i(\text{std})} * (\text{Max}-\text{Min})/2 + (\text{Max}+\text{Min})/2$$

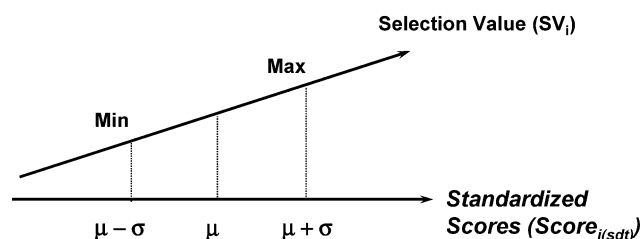


Figure 6. The fitness scaling is made upon standardized scores. The mean value of scores will have a selection value equal to the medium of the range [Min – Max]. This scaling lowers the difference between scores and gives less fitted candidates a higher chance to be selected.

function. For creating the next-generation population, molecules are selected from the mating pool of the current population using roulette wheel selection. This is a method for selecting the parents most fit to breed the next generation. The standardized score ($\text{Score}_{i(\text{std})}$) is scaling in a selection value (SV_i), which is the sector size of the roulette wheel (Figure 6). The fitness of each individual in the population is modified during the course of the run by fitness scaling. The scaling process helps to maintain competition by shrinking the differences between scores.

As the run goes on, the selection pressure increases in order to force the convergence near the end of the optimization. The interval called “Range” is the difference between “Max” and “Min”, two user-defined values that are set for the scaling process (see Figure 6). Range is increased along the run by the following equation

$$\text{Range} = \text{Range} + [\text{Progress} \times \text{Range}] \quad (3)$$

with

$$\text{Progress} = 1 - [g_{\text{max}} - g] / g_{\text{max}} \quad (4)$$

where Progress is set from 0.0 to 1.0 (0.0 at generation 0 and 1.0 at the generation g_{max}). g_{max} is the user-defined maximum number of generations in the run and g is the current number of generation. In the present calculation, Range will be multiplied by 2 from generation 0 to generation g_{max} .

Generating A New Population. A new population is created by performing crossover and mutation operations on the selected parent molecules.

Crossover selection is determined by a user-defined probability as well as the mutation one. Crossover is a mechanism for recombining two molecules to form two new molecules that preserve some of the characteristics of their parents. Crossover is applied every generation. Two types of crossover are performed: the one-point crossover combines a terminal portion (one or more building blocks) of a molecule with a similar terminal portion from another molecule. The two-point crossover involves the excision of an internal portion of a molecule. Then, the excised portion is inserted into a molecule that has a similar removed region. Thus, this operator may only be applied on molecules that combine at least three building blocks (Figure 7).

The second genetic operator is mutation. The present operator possesses five variations. Their selection is determined by a defined probability. This combination

of mutations allows an extremely variable series of molecular structures to be evolved: permutation, deletion, addition, and substitution. The mutation operator is applied after the crossover process. Either, it mutates the two new combined (breeding) molecules or it mutates the two parents if crossover failed (Figure 7b).

Parent selection and operator processes are applied until the whole new population contains, at least, the same number of molecules as the parent one. At maximum, $2n$ molecules are created if n is the number of molecules in the parent population. Then, the generation replacement involves a total or partial selection of the breeding candidates that replace the old parent population. However, an elitism strategy is applied. The best molecule in the parent generation is copied to the new generation without alteration.³⁶ This ensures that the best solution is never lost.

Termination. The algorithm is terminated after a prespecified number of generations. The fitness convergence criteria is ensured by the selection pressure discussed above. A parametrization study similar to the previous one performed with LEA showed that 100 generations along with a population of 40 individuals, an elitism strategy, and a fitness scale of 2–5 are optimal to efficiently optimize our molecules and thereby avoid a premature convergence.²²

Results

TMPKmt. TMPK represents a promising target for developing new antituberculosis drugs.^{37,38} It belongs to the NMPK family and is responsible for the reversible phosphorylation of deoxythymidine-5'-monophosphate (dTMP) to deoxythymidine-5'-diphosphate (dTDP) using ATP as its preferred phosphoryl donor. The dTMP binding site was the target for ligand design in this project (see Figure 8a,b). The holo structure of the protein with the dTMP has been determined by X-ray crystallography (PDB structure 1G3U).³⁷ The structure is composed of nine α -helices surrounding a five-stranded β -sheet core. The active site includes the usual P-loop (a phosphate binding loop) and LID region (a highly flexible stretch of residues covering the ATP binding site).

A previous analysis helped us to understand what would be the key interactions between the substrate and the active site.^{38–43} The thymine base of dTMP is deeply buried by interacting with a conserved Arg74 and with Asn100 as well as with a water molecule (HOH1002). The latter is interesting, since TMPK of other organisms possesses a glutamine that mimics this water molecule. The active site of the TMPKmt is slightly bigger. The sugar–phosphate moiety is less buried in the active site and interacts with Arg95, Asp9, Tyr39, a magnesium ion, and a water molecule. We assume that the main interactions occur at the back of the active site.

TMPKmt Representation. The active site was defined as a sphere of a radius of 10 Å around the substrate dTMP. Prior to starting the drug design process, it is necessary to assess that the representation of the active site is correct and that FlexX can reproduce the correct binding mode observed in the crystallographic complex. As seen in Figure 8c, the predicted binding mode from FlexX has a root-mean-square of 0.8 Å with the crystallographic structure. The main differ-

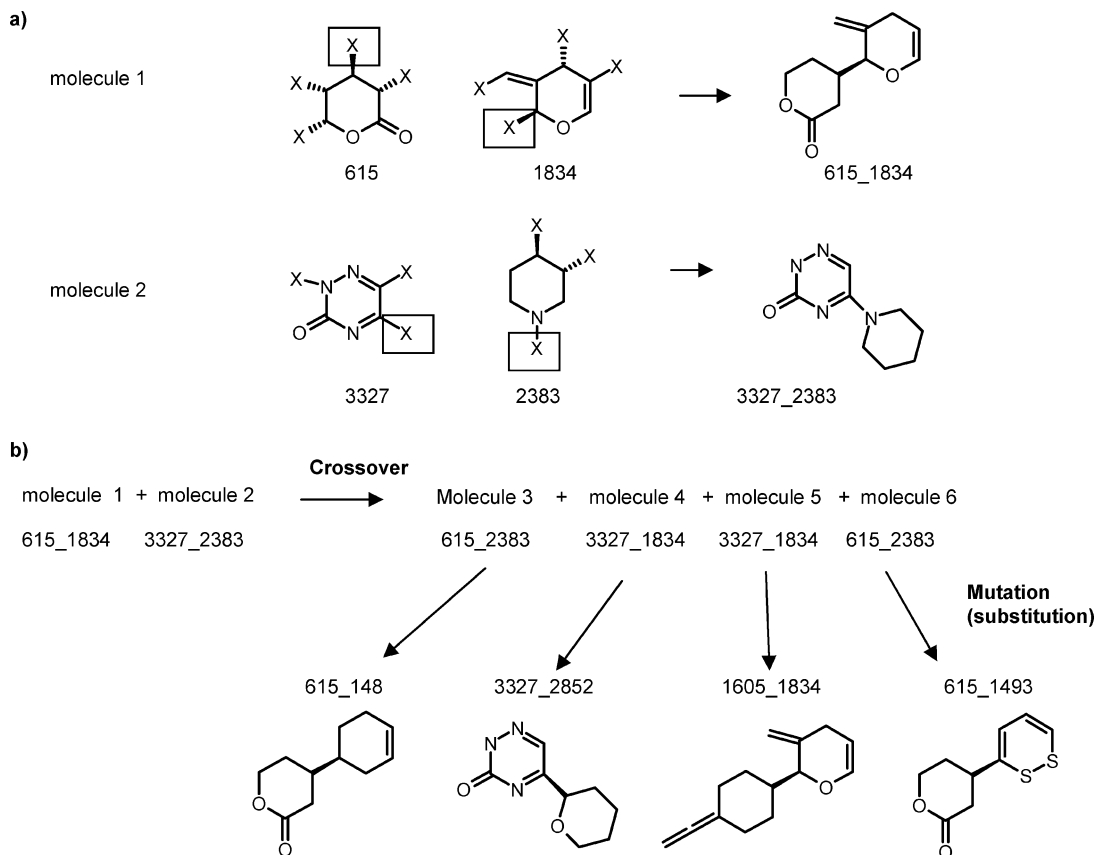


Figure 7. Molecule generation. (a) Creation of two molecules composed by two fragments (referred by a number). Each fragment possesses one or more X dummy atoms that refer to potential substitution points. However, in our fragment library, each fragment is associated with preselected tags in order to get reproducible molecules (X atoms with a square). These preselected X dummy atom are linked to create the molecule. During this step, all X dummy atoms are replaced by hydrogens, but the data block of the sdf file of the created molecule still contains the number of the unsubstituted positions (other X dummy atoms). (b) Crossovers and mutations on two molecules. The population size is two; thus, the crossover operation must be applied twice. In the present figure, a one-point crossover combines one fragment from molecule 1 to one fragment from molecule 2 (molecule 3 and molecule 5). The remaining fragments are also combined (molecule 4 and molecule 6). Four new molecules are generated. Then, mutations are applied on each new molecule and lead to four new final molecules. However, only the two best molecules (according to the user-defined fitness function) will be kept and will replace the parent population (molecule 1 and 2).

ences take place at the flexible phosphate part of the molecule. However, to obtain a correct representation, we have to keep one water molecule (HOH1002 in PDB structure 1G3U), which interacts with the substrate and Tyr165. Hydrogens were added on this water molecule in order to correctly form a hydrogen bond with the oxygen O2 of the substrate. The magnesium ion (MG300 in PDB structure 1G3U) is also important to maintain the flexible phosphate part close to the crystallographic position, but its presence is not essential for the correct binding mode of the thymine fragment.

Search for Ribose Substitute. In the search of thymine analogues, we focused on the replacement of the sugar-phosphate moiety. We decided to keep the thymine base moiety to take advantage of the conservation of the key interactions with the substrate observed in the complex. We constrained each molecule created by LEA3D to contain a thymine building block. Moreover, the thymine building block was also the main base fragment that FlexX uses to start the construction of the whole molecule. Thereafter, we asked FlexX to superimpose the created molecules onto the crystallographic reference of the thymine base by subgraph matching. This has the advantage of speeding up the docking calculation. We omitted the magnesium ion in order to have a bigger active site and because it is

apparent that it is bound after the binding of dTMP.^{44,45} However, the water molecule 1002 was still used to interact with the thymine base.

LEA3D was undertaken to search for analogues with a new sugar part using the FlexX docking program as the fitness function alone (no molecular properties requested).

Runs were set for 100 generations and the population size was 10 molecules per generation. The scaling process was set with the value of 2 for Min and a value of 5 for Max. The selection pressure was used. The number of fragments a molecule may combine was unconstrained but, by default, is limiting to 5. Each molecule was optimized, and three ring conformers were accepted when they exist. More than three conformers would slow the calculation. Several runs have been carried out. Solutions were visually inspected and removed when they interacted spuriously or when the chemical accessibility was difficult.

We identified that the sugar part of the dTMP can be replaced by a substituted benzyl group (Chart 1 and Figure 9). Compound 1 (Chart 1) possesses a benzyl group substituted by a 3-propionamide chain. The alkyl chain makes an elbow in order that the carboxamide group interacts with Asp9 and Tyr39, two residues

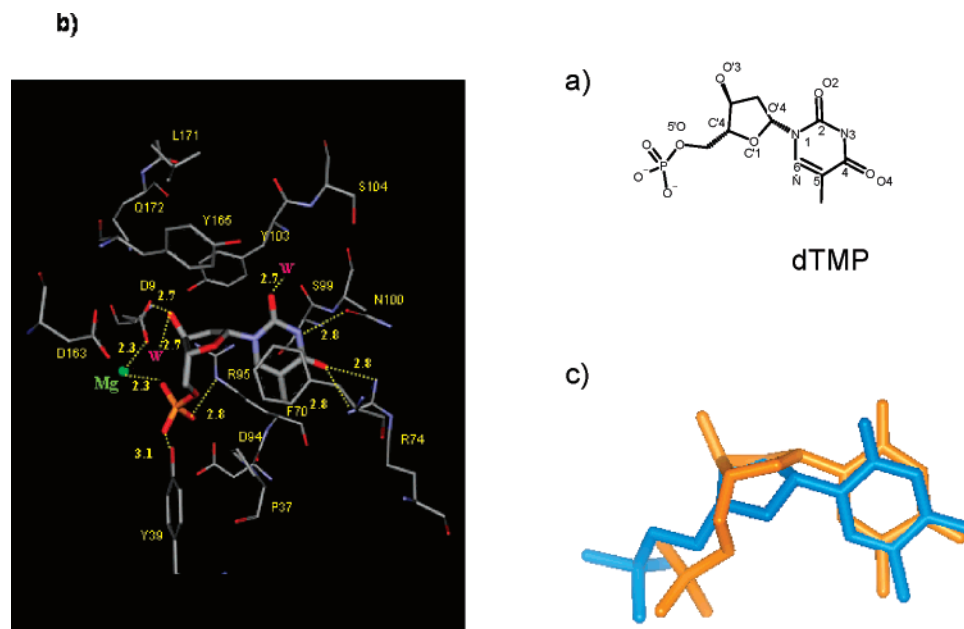
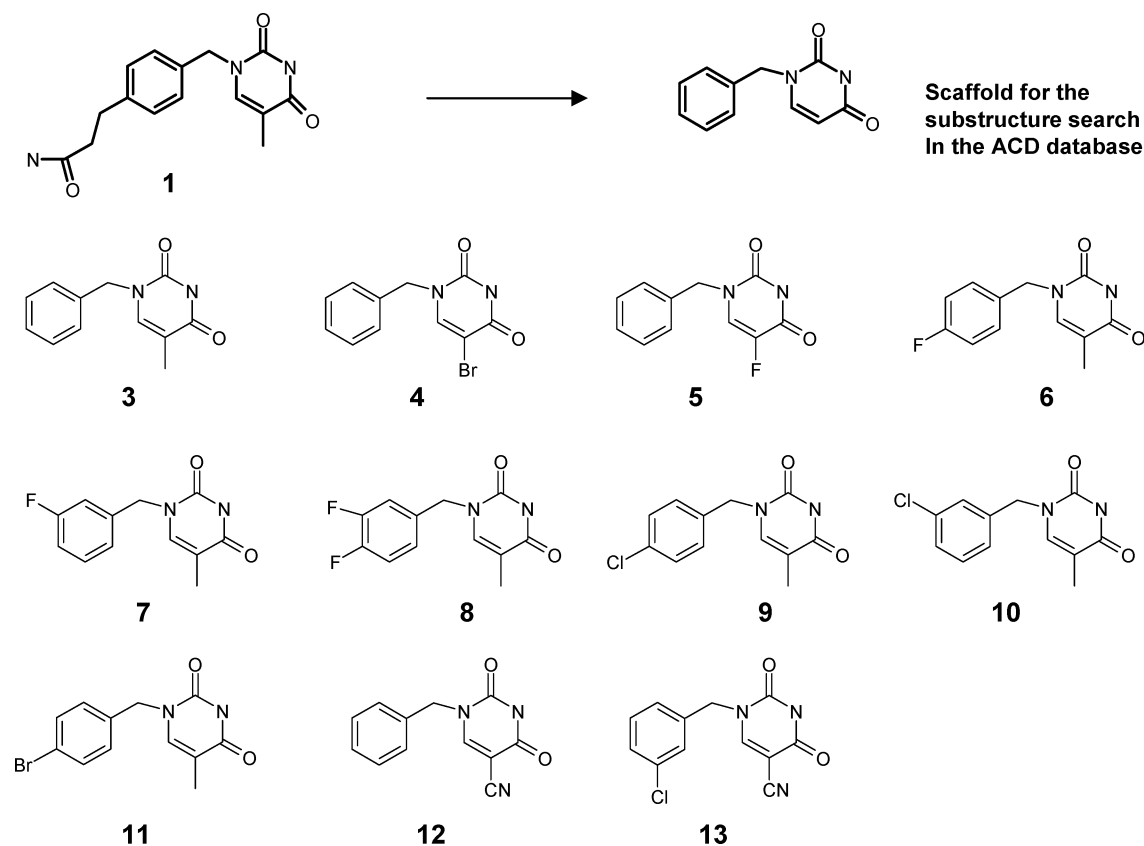


Figure 8. Active site of the TMPKmt. (a) Structure of the substrate dTMP with its numbering scheme. (b) Active site of TMPKmt in complex with its substrate dTMP (PDB structure 1G3U). Two water molecules (W in magenta) are included as well as the magnesium ion (green ball). The hydrogen-bond network is depicted in yellow. It involves the anchoring of the base moiety by 4 hydrogen bonds with Arg74 (two bonds), Asn100, and a water molecule. In addition, the thymine base is able to form a significant π -stack with the benzene ring of Phe70. The 5-methyl group of dTMP has been proved crucial for the activity. It correctly fills the cavity and probably endows the correct orientation of the sugar moiety of the dTMP molecule. The 3'-hydroxyl of the sugar moiety is interacting with Asp9 and a water molecule. The 5'-O-phosphate group interacts with Arg95, Tyr 39, and a water molecule as well as coordinating the magnesium ion. (c) Superposition of the crystallographic dTMP (in orange) with the predicted binding mode (in blue) using FlexX docking program. The rms between the two structures is 0.8 Å.

Chart 1. Structures of Molecules That Were Selected and Purchased (**12** and **13**) or Synthesized (**3–11**)



involved into the reaction pathway of TMPKmt. The FlexX score of **1** is close to the dTMP one (Table 2).

Prior to the synthesis of **1**, we undertook a substructure search of easily accessible analogues in the ACD

(Available Chemicals Directory 2003.1) commercial database (MDL Information Systems) (Chart 1). The substructure search was carried out using the combination of the two scaffolds (uracyl base and benzyl group).

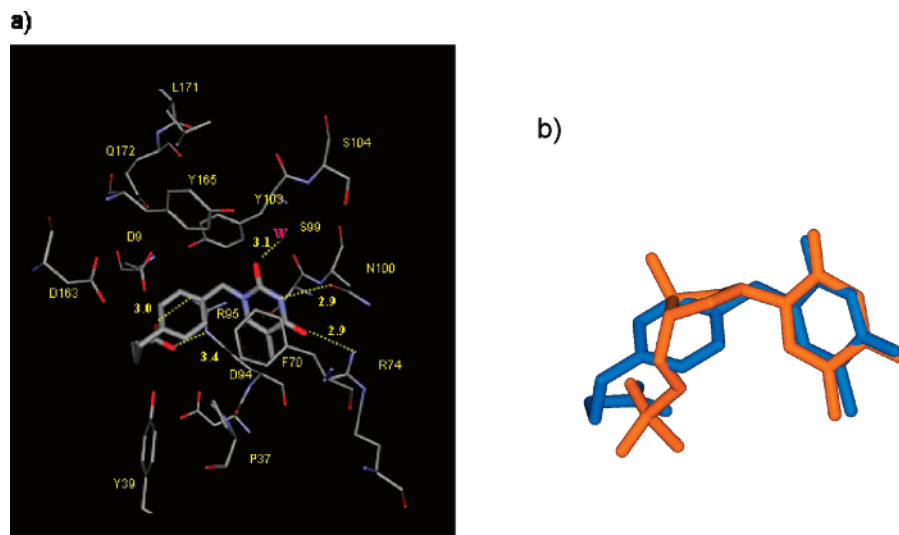


Figure 9. (a) The predicted binding mode of molecule **20**. This molecule is able to form five hydrogen bonds as well as a significant π -stack with the benzene ring of Phe70. Compared with Figure 8, the thymine base loses one hydrogen-bond interaction with Arg74. The carboxylate group is able to interact with Arg95 by balancing its negative charge. (c) Superposition between the substrate dTMP and molecule **20**.

Table 2. FlexX Scores and Inhibitory Potencies of Compounds 1–22

compd	TMPKmt		compd	TMPKmt	
	FlexX score ^a	K_i (μ M)		FlexX score ^a	K_i (μ M)
dTMP	-31 (-37 ^b)	$K_m = 4.5$	11	-31	38
dT	-29 (-30 ^b)	27	12	-31	980
1	-36	110	13	-30	810
2	-32	68	14	-38	240
3	-31	75	15	-32	NI ^c
4	-31	44	16	-33	NI
5	-31	NI	17	-31	NI
6	-31	45	18	-32	NI
7	-31	90	19	-31	265
8	-31	67	20	-39	16.5
9	-31	50	21	-38	12.3
10	-30	44	22	-51	32

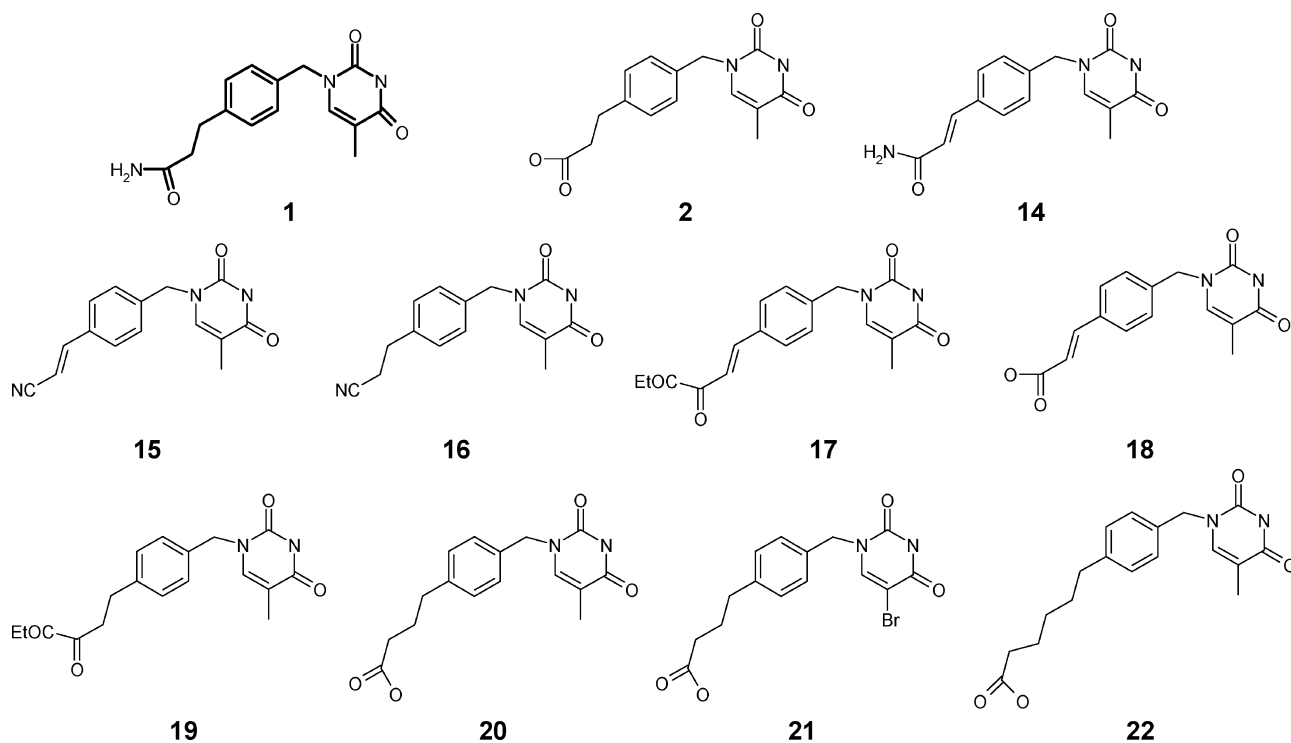
^a The water molecule HOH1002 (in PDB file 1G3U) is included during the docking by FlexX. ^b Refers for cases in which the magnesium ion is also included during the docking process. ^c NI: no inhibition detected.

A set of 133 commercially available compounds was retrieved for virtual screening using the FlexX docking program. Among them, two compounds (**12** and **13**) were purchased and nine compounds (**3**–**11**) were synthesized according to classical methodology (synthesis will be described elsewhere) in order to possess a substituent (usually a methyl) at position 5 on the uracyl base. The aim was to assess whether the sugar portion can be replaced by a benzyl moiety without detrimental effect. The inhibitory potency of these compounds was tested on TMPKmt by enzymatic assay (Table 2). Except compound **5**, they are all inhibitors of TMPKmt with K_i values ranging from 38 to 980 μ M. Compound **11** shows an unexpected efficiency for a rather nude molecule. Results confirm that a methyl or bromine at position 5 on the pyrimidine moiety improves the affinity of the molecule as previously observed (**3** and **4**).⁴⁶ Introduction of bromine in position 4 of the benzyl ring enhances the affinity of the compound **11**, probably by a better steric fit of this group. Finally, we observed that the FlexX scores do not correlate with the measured activity, including its failure to highlight differences between

substituents such as fluorine, bromine, and methyl (**5**, **4**, and **3** respectively). Although the affinities of these compounds were weaker than the one of the natural substrate dTMP, these results were promising enough to follow up the synthesis of derivatives of the previously designed **1**.

Ten molecules were derived and synthesized upon the structure of **1** (Chart 2 and Table 2). Compound **1** has a weak inhibitory potency of 110 μ M. The inhibition is improved with a 3-propionic acid chain (**2**, $K_i = 68 \mu$ M). The asset of **2** upon **1** is probably the carboxylate charge able to balance the positively charged environment of the LID. A previous crystallographic study shows that the Asp 9 is correctly ordered only after the binding of the magnesium ion.⁴⁵ But, **1** probably prevents the binding of this ion and, thus, probably prevents the predicted interaction with Asp9.

The best hit among this series is compound **21** with a K_i of 12.3 μ M. The potency of the molecules **20** (16.5 μ M) and **21** are better than the one of dT (27 μ M). Compounds **20** and **21** are predicted to balance the Arg95 positive charge as shown by the substrate dTMP (see Figure 9). As mentioned above, a 5-bromine (**21**) improves the affinity. So far, the best length of the spacer arm between the benzyl and the carboxylic acid is apparently three carbons (**20**, $K_i = 16.5 \mu$ M and FlexX score = -39) compared to two carbons (**2**, $K_i = 68 \mu$ M and FlexX score = -32) or five carbons (**22**, $K_i = 32 \mu$ M and FlexX score = -51). FlexX scores correctly reflect the difference of activity for **20** and **2** but not for **20** and **22**. Nevertheless, we noticed that the predicted binding mode of **22** was spurious despite its high score. Apparently, the spacer arm is too long to correctly promote an interaction between the carboxylate function and Arg95. On the contrary, a molecule that would possess four carbons as spacer arm is predicted to have a higher affinity (FlexX score = -45). The predicted binding mode is close for **20**, **21**, and the four-carbon compound. The differences are the width of the elbow and the hydrogen-bond capability between the carboxylate function and Arg95. Two hydrogen bonds are predicted between the carboxylate of the four-carbon derivative

Chart 2. Molecules Designed by LEA3D

^a Eleven structures derived from **1** have been synthesized (**2**, **14–22**).

and Arg95, one hydrogen bond between the carboxylate of **20** and Arg95, and none between the carboxylate of **2** and Arg95. The synthesis of the four-carbon derivative is currently in progress.

Discussion

We have described LEA3D, a new structure-based drug design program aimed at identifying novel structures that are predicted to fit the active site of a target protein. As illustrated in the described example, the association of LEA3D and the FlexX docking program permits the identification of a new family of thymine derivatives as inhibitors of TMPKmt. The methodology is particularly suitable for rapid identification of chemically accessible, druglike fragments that can be rapidly translated into potent lead compounds. Currently, we used a building block library that contains approximately 8000 fragments. But, if the user has some special requests, the building block library could be extended by the addition of new fragments. This can be easily done due to the organization of the library in a simple and open form. This design protocol, when coupled with a commercial database such as ACD, chemical intuition with regard to chemical stability, and synthetic ease, can lead to a great number of novel ligand candidates. A manual optimization may lead to a final generation of molecules.

Our strategy combining *de novo* drug design, substructure searches of the ACD, virtual screening, and experimental methods such as chemistry and enzymatic assays is close to the SHAPES method, developed at Vertex Pharmaceuticals.³³ The SHAPES strategy uses nuclear magnetic resonance (NMR) screening of a library of small druglike molecules with various complementary methods such as virtual screening, high throughput-screening (HTS), and combinatorial chem-

istry. Their strategy successfully produced submicromolar classes of compounds for the Jnk3 MAP Kinase.

By using LEA3D, it is also possible to combine prior knowledge concerning a particular binding site (for example, known binding mode) with the ligand design protocol by requesting a specific interaction known to exist or by selecting directions of growth through the use of partially grown molecules as restart fragments. One advantage of the design methodology presented in this paper is the ability to take into account for specificity by combining multiple target-docking scores as the fitness scoring function. This multiple target docking, at one time, is faster than the sequential docking process. In the present case, we undertook further LEA3D design upon the substructure of molecule **20**. We hope to improve its affinity for TMPKmt by growing and by promoting additional interactions, and we hope to find some selectivity versus the human homologue (TMPKh, PDB structure 1E2F).^{45,47} Some new analogues have yet to be designed (structures not shown). They possess interesting interactions with TMPKmt that are close to those observed for the dTMP and they are predicted not to bind the TMPKh by promoting steric hindrances and by interfering with a salt bridge described as crucial for the activity of this protein⁴⁶ (see Figure 10). These results are still predictions, but we are currently pursuing this line of development.

Improvement of the evaluation of the energy of interaction would be another crucial issue, to provide a quantitative score directly related to the activity. In the present case, optimizing the substituents on the benzyl moiety would need a more precise scoring function to estimate binding affinities of different side chains and, thereby, better rank the FlexX docking solutions. Indeed, FlexX can identify structures that can bind into the TMPKmt active site, but it presents some difficulties

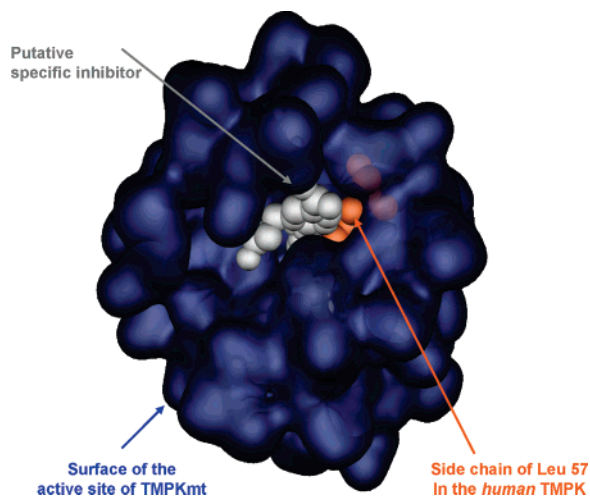


Figure 10. TMPKmt specificity versus TMPKh. The blue surface represents the surface of the TMPKmt active site. The side chain of the Leu 77 (in orange) in the TMPKh could create a steric hindrance with a second-generation molecule (in white).

to properly rank the solutions according to their inhibition potencies. A previous study on known ligands of TMPKmt reveals that FlexX succeeds in highlighting the more putative binders but without a correct rank (data not shown). These difficulties are well-known for scoring functions used in protein–small molecule docking methods.^{48–50} Potential extensions to *de novo* ligand design using LEA3D may include an improvement of the prediction of the docking score either by using a physical-based scoring function such as CHARMM⁵¹ or by using a consensus score resulting from the usage of several scoring functions or maybe different docking programs.

Experimental Section

Fragment Library. Molecules from the CMC and KEGG databases (sdf file format) have been dissociated into fragments that are classified into 18 families (eight single ring families, seven fused-ring families, and three acyclic families; Table 1). The fragment library generation is performed as follows.

First, salts of the database are dissociated. Then, three-dimensional structures are generated by the program CORINA (only one conformer). Molecules containing heavy metals or without 3D coordinates are discarded. Additional constraints may be added if needed: no phosphorus atom, a limited number of asymmetric carbons, possessing the “rule-of-five” of Lipinski or not, containing a “reactive group” (for example, aldehyde, isocyanate, nitrosamine, α -halocarbonyl, thiosulfate, hydrazine, oxime ester, epoxyde, aziridine). Finally, carboxylate and phosphate groups are ionized.

A program has been developed to dissect each molecule (sdf format). First, single ring systems are detected and duplicates are removed. Then, fused ring systems are identified: an association of single ring systems that share a bond. Rings are fused and removed from the single ring system list. Duplicates are also discarded. Single ring systems up to 10 atoms are identified. Above the limit of 10 atoms, single ring systems are excluded. In the present study, the CMC database contains only 86 molecules that are out of our limit. Acyclic parts encompass linkers and terminal substituents containing more than two atoms. Rings, fused rings, and acyclic parts are then classified in many classes (Table 1) depending on the aromaticity and the number of atoms. Finally, duplicates in each class are removed.

In Table 1, the number of unique fragments generated is given for each class and each database with a percentage

related to the number of molecules in the original database (7621 and 7282 molecules for CMC and KEGG, respectively). Finally, 5274 and 4142 distinct fragments have been generated upon the CMC and KEGG databases, respectively. The union of the two fragment libraries results into a final fragment library that contains 7986 unique fragments that have been categorized into two subfamilies: the “e” fragments and the “f” fragments.

For the purpose of the linear combination of fragments, we distinguished fragments that can be substituted at least twice from those that can be substituted only once (see Figure 4a). Thus, a fragment is either classified e (end) or f type. An e fragment is a fragment with only one substitution point (only one “X” dummy atom), and f fragments are the others (multiple substitution points). In the context of a linear combination, fragments must possess one (e fragment) or two tags (f fragment) that indicate which atoms have to be bonded with another fragment. The “left” and “right” tags correspond to the number of the atom that bonds the previous “X” dummy atom. If a fragment possesses more than one substitution point, then a combination of pairs of tags is generated including the symmetric one (Figure 4b). Due to the combinatorial explosion of pairs, we sometimes set the maximum number of combinations at 5 (referenced by * in Table 1) or at 10 (referenced by ** in Table 1) for each fragment. Combinations are sequentially generated following the number of the atoms in increasing order, without any distance criteria; the 5th or 10th first generated combinations are kept, and then the symmetric combinations are added. The tag information is stored in the data block of the sdf file of the fragment. Thus, one fragment can be registered more than once but with different tags. Permanent tags are important to get reproducible molecules. Finally, we obtained a fragment library of 39 569 building blocks (see Table 1).

TMPKmt in Vitro Assays. TMPKmt assays were done using the coupled spectrophotometric assay described by Blondin et al.⁵⁴ at 334 nm in an Eppendorf ECOM 6122 photometer. The reaction medium (0.5 mL final volume) contained 50 mM Tris-HCl pH 7.4, 50 mM KCl, 2 mM MgCl₂, 0.2 mM NADH, 1 mM phosphoenol pyruvate, and 2 units each of lactate dehydrogenase, pyruvate kinase, and nucleoside diphosphate kinase. The concentrations of ATP and dTMP were kept constant at 0.5 and 0.05 mM, respectively, whereas the concentrations of compounds varied between 0.1 and 2.9 mM. K_i values were calculated⁴⁶ by assuming the different compounds to be competitive inhibitors.

Acknowledgment. We thank I. Lascu for the kind gift of nucleoside diphosphate kinase. Also, Finly Philip and Guillaume Drin are thanked for their useful comments on the manuscript. This work was sponsored by the Institut Pasteur, CNRS, INSERM, and Ministère de la recherche (ACI–Molécules et Cible Thérapeutiques).

References

- (1) Cohen-Gonsaud, M.; Catherinot, V.; Labesse, G.; Douguet, D. From molecular modeling to drug design. *Practical Bioinformatics*; Bujnicki, J. M., Ed.; Springer-Verlag: Berlin, 2004; pp 35–72.
- (2) Venkatasubramanian, V.; Chan, K.; Caruthers, J. M. Evolutionary Design of Molecules with Desired Properties Using the Genetic Algorithm. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 188–195.
- (3) Parrill, A. L. Evolutionary and genetic methods in drug design. *Drug Discovery Today* **1996**, *1*, 514–521.
- (4) Schneider, G.; Bohm, H. J. Virtual screening and fast automated docking methods. *Drug Discovery Today* **2002**, *7*, 64–70.
- (5) Honma, T. Recent advances in *de novo* design strategy for practical lead identification. *Med. Res. Rev.* **2003**, *23*, 606–632.
- (6) Ooms, F. Molecular modeling and computer aided drug design. Examples of their applications in medicinal chemistry. *Curr. Med. Chem.* **2000**, *7*, 141–158.
- (7) Schneider, G.; Schrod, W.; Wallukat, G.; Muller, J.; Nissen, E.; et al. Peptide design by artificial neural networks and computer-based evolutionary search. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 12179–12184.

- (8) Schneider, G.; Lee, M. L.; Stahl, M.; Schneider, P. De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 487–494.
- (9) Babine, R. E.; Bleckman, T. M.; Kissinger, C. R.; Showalter, R.; Pelletier, L. A.; et al. Design, synthesis and X-ray crystallographic studies of novel FKBP-12 ligands. *Bioorg. Med. Chem. Lett.* **1995**, *3*, 1719–1724.
- (10) Nishibata, Y.; Itai, A. Automatic creation of drug candidate structures based on receptor structure. Starting point for artificial lead generation. *Tetrahedron* **1991**, *47*, 8885–8990.
- (11) Bohm, H. J. The computer program LUDI: A new method for the de novo design of enzyme inhibitors. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 61–78.
- (12) Gillet, V.; Johnson, A. P.; Mata, P.; Sike, S.; Williams, P. SPROUT: A program for structure generation. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 127–153.
- (13) Eisen, M. B.; Wiley, D. C.; Karplus, M.; Hubbard, R. E. HOOK: A program for finding novel molecular architectures that satisfy the chemical and steric requirements of a macromolecule binding site. *Proteins* **1994**, *19*, 199–221.
- (14) Bohacek, R. S.; McMartin, C. Multiple highly diverse structures complementary to enzyme binding sites: Results of extensive application of de novo design method incorporating combinatorial growth. *J. Am. Chem. Soc.* **1994**, *116*, 5560–5571.
- (15) Clark, D. E.; Frenkel, D.; Levy, S. A.; Li, J.; Murray, C. W.; et al. PRO-LIGAND: An approach to de novo molecular design. 1. Application to the design of organic molecules. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 13–32.
- (16) Pearlman, D. A.; Murcko, M. A. CONCERTS: Dynamic connection of fragments as an approach to de novo ligand design. *J. Med. Chem.* **1996**, *39*, 1651–1663.
- (17) *LeapFrog*, 6.8 ed.; Tripos, Inc.: St Louis, MO.
- (18) Pegg, S. C.; Haresco, J. J.; Kuntz, I. D. A genetic algorithm for structure-based de novo design. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 911–933.
- (19) Wang, R.; Gao, Y.; Lai, L. A Multi-Purpose Program for Structure-Based Drug Design. *J. Mol. Model.* **2000**, *6*, 498–516.
- (20) Blaney, J. M.; Dixon, J. S.; Weininger, D. Molecular Graphics Society Meeting on Binding Sites: Characterising and Satisfying Steric and Chemical Restraints; U.K., 1993.
- (21) Venkatasubramanian, V.; Chan, K.; Caruthers, J. M. Computer Aided Molecular Design Using Genetic Algorithms. *Comput. Chem. Eng.* **1994**, *18*, 833–844.
- (22) Douguet, D.; Thoreau, E.; Grassy, G. A genetic algorithm for the automated generation of small organic molecules: Drug design using an evolutionary algorithm. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 449–466.
- (23) Weininger, D. SMILES. 3. depict: Graphical depiction of chemical structures. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 237–243.
- (24) Breinbauer, R.; Vetter, I. R.; Waldmann, H. From protein domains to drug candidates—natural products as guiding principles in the design and synthesis of compound libraries. *Angew. Chem. Int. Ed.* **2002**, *41*, 2879–2890.
- (25) Cragg, C. M.; Newman, D. J.; Snader, K. M. For an outstanding analysis of the role of natural products in pharmaceuticals. *J. Nat. Prod.* **1997**, *60*, 52–60.
- (26) Bohacek, R. S.; McMartin, C.; Guida, W. C. The art and practice of structure-based drug design: A molecular modeling perspective. *Med. Res. Rev.* **1996**, *16*, 3–50.
- (27) Rarey, M.; Wefing, S.; Lengauer, T. Placement of medium-sized molecular fragments into active sites of proteins. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 41–54.
- (28) Holland, J. *Adaptation in natural and artificial systems*; University of Michigan Press: Ann Arbor, MI, 1976.
- (29) Murray, A.; Louis, S. J. Design strategies for evolutionary robots. *Proceedings of the Third Golden West International Conference on Intelligent Systems*; Kluwer Academic Press: Norwell, MA, 1995; pp 609–616.
- (30) Forrest, S. Genetic Algorithms: Principles of Natural Selection Applied to Computation. *Science* **1993**, *261*, 872–878.
- (31) Goldberg, D. E. *Genetic Algorithms in Search, Optimization, and Machine Learning*; Addison-Wesley: Reading, MA, 1989.
- (32) Fogel, L. J.; Owens, A. J.; Walsh, M. J. *Artificial Intelligence through Simulated Evolution*; John Wiley & Sons: New York, 1966.
- (33) Lepre, C. A.; Peng, J.; Fejzo, J.; Abdul-Manan, N.; Pocas, J.; et al. Applications of SHAPES screening in drug discovery. *Comb. Chem. High Throughput Screen* **2002**, *5*, 583–590.
- (34) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic Generation of 3D-Atomic Coordinates for Organic Molecules. *Tetrahedron Comput. Method* **1990**, *3*, 537–547.
- (35) Rockey, W. M.; Elcock, A. H. Progress toward virtual screening for drug side effects. *Proteins* **2002**, *48*, 664–671.
- (36) De Jong, K. A. An Analysis of the Behavior of a Class of Genetic Adaptive Systems. In *Department Computer and Communication Sciences*; University of Michigan: Ann Arbor, MI, 1975.
- (37) Li de la Sierra, I.; Munier-Lehmann, H.; Gilles, A. M.; Barzu, O.; Delarue, M. X-ray structure of TMP kinase from *Mycobacterium tuberculosis* complexed with TMP at 1.95 Å resolution. *J. Mol. Biol.* **2001**, *311*, 87–100.
- (38) Pochet, S.; Dugue, L.; Douguet, D.; Labesse, G.; Munier-Lehmann, H. Nucleoside Analogues as Inhibitors of Thymidylate Kinases: Possible Therapeutic Applications. *ChemBioChem* **2002**, *3*, 108–110.
- (39) Van Rompaey, P.; Nauwelaerts, K.; Vanheusden, V.; Rozenski, J.; Munier-Lehmann, H.; et al. *Mycobacterium tuberculosis* Thymidine Monophosphate Kinase Inhibitors: Biological Evaluation and Conformational Analysis of 2'- and 3'-Modified Thymidine Analogues. *Eur. J. Org. Chem.* **2003**, 2911–2918.
- (40) Vanheusden, V.; Munier-Lehmann, H.; Froeyen, M.; Dugue, L.; Heyerick, A.; et al. 3'-C-branched-chain-substituted nucleosides and nucleotides as potent inhibitors of *Mycobacterium tuberculosis* thymidine monophosphate kinase. *J. Med. Chem.* **2003**, *46*, 3811–3821.
- (41) Vanheusden, V.; Munier-Lehmann, H.; Pochet, S.; Herdewijn, P.; Van Calenberg, S. Synthesis and evaluation of thymidine-5'- Δ -monophosphate analogues as inhibitors of *Mycobacterium tuberculosis* thymidylate kinase. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 2695–2698.
- (42) Munier-Lehmann, H.; Pochet, S.; Dugue, L.; Dutruel, O.; Labesse, G.; et al. Design of *Mycobacterium tuberculosis* thymidine monophosphate kinase inhibitors. *Nucleosides Nucleotides Nucleic Acids* **2003**, *22*, 801–804.
- (43) Munier-Lehmann, H.; Chaffotte, A.; Pochet, S.; Labesse, G. Thymidylate kinase of *Mycobacterium tuberculosis*: A chimera sharing properties common to eukaryotic and bacterial enzymes. *Protein Sci* **2001**, *10*, 1195–1205.
- (44) Haouz, A.; Vanheusden, V.; Munier-Lehmann, H.; Froeyen, M.; Herdewijn, P.; et al. Enzymatic and structural analysis of inhibitors designed against *Mycobacterium tuberculosis* thymidylate kinase. New insights into the phosphoryl transfer mechanism. *J. Biol. Chem.* **2003**, *278*, 4963–4971.
- (45) Fioravanti, E.; Haouz, A.; Ursby, T.; Munier-Lehmann, H.; Delarue, M. et al. *Mycobacterium tuberculosis* thymidylate kinase: Structural studies of intermediates along the reaction pathway. *J. Mol. Biol.* **2003**, *327*, 1077–1092.
- (46) Pochet, S.; Dugue, L.; Labesse, G.; Delepierre, M.; Munier-Lehmann, H. Comparative study of purine and pyrimidine nucleoside analogues acting on the thymidylate kinases of *Mycobacterium tuberculosis* and of humans. *Chembiochem* **2003**, *4*, 742–747.
- (47) Ostermann, N.; Segura-Pena, D.; Meier, C.; Veit, T.; Monnerjahn, C.; et al. Structures of human thymidylate kinase in complex with prodrugs: Implications for the structure-based design of novel compounds. *Biochemistry* **2003**, *42*, 2568–2577.
- (48) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43*, 4759–4767.
- (49) Schulz-Gasch, T.; Stahl, M. Binding site characteristics in structure-based virtual screening: Evaluation of current docking tools. *J. Mol. Model. (Online)* **2003**, *9*, 47–57.
- (50) Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. Evaluation of docking performance: Comparative data on docking algorithms. *J. Med. Chem.* **2004**, *47*, 558–565.
- (51) Brooks, B. R.; Brucoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; et al. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187–217.
- (52) Eisenhaber, F.; Argos, P. Improved strategy in analytic surface calculation for molecular systems: Handling singularities and computational efficiency. *J. Comput. Chem.* **1993**, *11*, 1272–1280.
- (53) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic Physicochemical Parameters for Three-Dimensional Structure Directed Quantitative Structure–Activity Relationships. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163–172.
- (54) Blondin, C.; Serina, L.; Wiesmuller, L.; Gilles, A. M.; Barzu, O. Improved spectrophotometric assay of nucleoside monophosphate kinase activity using the pyruvate kinase/lactate dehydrogenase coupling system. *Anal. Biochem.* **1994**, *22*.

JM0492296