SYNOPSIS: SYNthesize and OPtimize System in Silico

H. Maarten Vinkers,*,† Marc R. de Jonge,† Frederik F. D. Daeyaert,† Jan Heeres,† Lucien M. H. Koymans,† Joop H. van Lenthe,‡ Paul J. Lewi,† Henk Timmerman,§ Koen Van Aken,† and Paul A. J. Janssen†

Center for Molecular Design, Janssen Pharmaceutica N.V., Antwerpsesteenweg 37, B-2350 Vosselaar, Belgium, Theoretical Chemistry Group, Debye Institute, Utrecht University, Padualaan 14, 3584 CH Utrecht, The Netherlands, and Leiden/Amsterdam Center for Drug Research (LACDR), Division of Medicinal Chemistry, Department of Pharmacology, Faculty of Chemistry, Vrije Universiteit, De Boelelaan 1083, 1081 HV Amsterdam, The Netherlands

Received February 13, 2003

We present a de novo design program called SYNOPSIS, that includes a synthesis route for each generated molecule. SYNOPSIS designs novel molecules by starting from a database of available molecules and simulating organic synthesis steps. This way of generating molecules imposes synthetic accessibility on the molecules. In addition to a starting database, a fitness function is needed that calculates the value of a desired property for an arbitrary molecule. The values obtained from this function guide the design process in optimizing the molecules toward an optimal value of the calculated property. Two applications are described. The first uses an electric dipole moment calculation to generate molecules possessing a strong dipole moment. The second makes use of the three-dimensional structure of a viral enzyme in order to generate high affinity ligands. Twenty eight compounds designed with the program resulted in 18 synthesized and tested compounds, 10 of which showed HIV inhibitory activity in vitro.

Introduction

In the chemical and pharmaceutical industry a continuous demand exists for novel molecules with specific physical or biological properties. Traditionally these molecules are found either by accidental observation of an interesting characteristic or by testing many molecules, from natural sources or man-made, for the desired properties. Computational methods provide a complementary strategy in finding such molecules. Progress in fundamental understanding of physical, chemical, and biological systems along with ever-increasing computer power have brought these methods within everyday use.

A prerequisite for in vitro drug testing is to have a test whose outcome correlates significantly with some clinical effect (e.g., inhibition of the D_2-receptor in vitro and antipsychotic effect in man). Most extant approaches followed in the pharmaceutical industry try to find drug candidates by subjecting a large number of molecules to such tests. The expectation is that some of them will show up as active. This approach, called high throughput screening (HTS), generally yields only very few active molecules. Considering the vast number of all possible molecules, one should not be discouraged by this apparently poor success-rate in a limited sample; the ratio of active molecules to possible molecules is substantial. But performing a high throughput screen demands quite some resources and is possible only once given a supply of compounds and a test. Thus there is ample room for a directed search approach. Nevertheless, high throughput screening is a convenient starting point if little is known about the target system and an automated test for relevant properties is available.

Over the past decade computational methods have been developed and applied to design catalysts,[1,2] polymers,[3,4] proteins,[5−7] and drugs. Computational methods are particularly abundant in the latter.[8−12] Frequently, they benefit from knowledge about the function and three-dimensional structure of the molecular target(s) involved and are often referred to as 'structure-based drug design' methods. Structure-based drug design has become an established tool to find new leads or to optimize existing ones.[13] It is frequently used to guide the human designer who wants to modify an existing molecule in order to improve its characteristics and also has been implemented in automated methods for novel drug design. The latter are commonly referred to as 'de novo' design methods. The requirements—and restrictions—on the molecules that are designed depend on the purpose the molecules are designed for. To be acceptable as a drug, a molecule should normally not contain chemically reactive groups (except bactericides and antineoplastics) nor radioactive atoms (except radiotherapeutic agents). To be apt for oral administration its molar mass should be below 500 g/mol.[14]

A common structure-based computational design strategy is to construct a molecule directly in the binding site of the target protein. The quality of the designed molecules is evaluated with an interaction energy calculation or a pharmacophore model. The building process commences with an anchor fragment which is incrementally refined.[15−21] Alternatively, fragments are put independently at favorably interacting spots in the binding site and subsequently linked to a single molecule.[22−24] Another strategy[25−28] is to fill the binding site with generic atoms and to progress toward a molecule by specification of elements and bonds. A drawback of all these building strategies is that the resulting conformation of a molecule will almost always be so high in energy that it does not occur in that form

* To whom correspondence should be addressed. Tel: +(32)-14442294, fax: +(32)14410503, e-mail: mvinkers@janbe.jnj.com.

† Janssen Pharmaceutica N.V.

‡ Utrecht University.

§ Vrije Universiteit.

10.1021/jm030809x CCC: $25.00
Published on Web 05/24/2003

under natural circumstances. Furthermore the building procedure determines the orientation locally which does not necessarily yield the optimal fit for the whole molecule. To overcome these problems, we have developed a computer program called SYNOPSIS, that separates building from evaluating. This has the added benefit that our program can easily be adapted to all kinds of evaluation functions.

The properties of the designs are rarely experimentally verified. Generally, the quality of the designed molecules is estimated from the filling of the binding site, from presence of pharmacophore elements or by comparison with—or superimposition on—known compounds.[15,17,27,28] Occasionally verification proceeds by enriching the design space with known active patterns and analyzing their retrieval.[24] Synthesizing a molecule is the first step toward experimental determination of its properties. We will define the ease of synthesis with 'synthesizability', used in the sense that the more synthesizable a molecule is, the less effort will be required for the actual synthesis in terms of availability and cost of starting materials, number and yields of synthesis steps, time and apparatus required, etc.. When experimental verification of the designed molecules is a requirement, synthesizability of the generated molecules is an important issue. Incorporating synthesizability at an early stage in the method is however not an absolute necessity. One can consider synthesizability of the designs afterward, utilizing heuristic rules[29] or a retrosynthetic program. An example of a case where no specific measures were taken to ensure synthetic viability of the ligands, but where the designs were made and tested anyway, is given in Holloway.[30] Nevertheless, ease of synthesis is a desirable property: it is convenient for method validation, speed of screening, and ultimately for commercial reasons. A recent method[31] aims at generating synthesizable designs by employing a fragmentation and combining scheme based on cleavage and formation of bonds according to chemical reactions. SYNOPSIS enforces synthesizability from the onset by starting from available compounds and exclusively employing chemical reactions to create new molecules.

Method

SYNOPSIS requires three components: a database of existing molecules, a set of chemical reactions, and a fitness function. For each entry in the starting database the constituent atoms and bonds are specified. In the applications described, a subset of the ACD[32] was used as initial database. The subset was obtained to meet the restrictions for common medicinal chemicals by excluding any molecule which met one or more of the following criteria:

- contains any element other than carbon, nitrogen, oxygen, sulfur, fluorine, chlorine, bromine, and iodine
- contains nonnatural occurring isotopes
- is a radical

Compounds with more than an arbitrarily chosen number of 13 non-hydrogen atoms were also excluded, to prevent SYNOPSIS from putting much effort in synthesis with already large molecules. The final subset used consists of 32 287 molecules.

The implementation of the organic synthesis steps is based on a functional group approach. SYNOPSIS will determine the functional groups present in a molecule to decide which reactions are possible for that molecule. Currently 70 different reaction types have been implemented (available from the author upon request). Given the initial database and the reaction set, we have determined the number of molecules that can be synthesized in one step from the starting materials by exhaustively trying each reaction on each of the molecules from the initial database. This resulted in the generation of 373 174 909 new molecules.

The decision to allow a particular reaction is based on just a part of the molecule. If this approach is applied without any regard to the reactivity of the groups involved many erroneous synthesis steps will result. To prevent this an estimate of reactivity is implemented by means of additional rules for acceptance of a reaction. Depending on the type of potential error the following cases are distinguished:

1. More local structure than that contained in a functional group description is necessary for a particular reaction to take place. For example, an NH_2 moiety can be oxidized to an NO_2 moiety, but not when it is part of an $N-NH_2$ moiety. By defining a functional group that includes more atoms, in this example by requiring a $C-NH_2$ group, the feasibility of the reaction is preserved.

2. Other functional groups hinder the intended reaction to take place. Therefore some reactions are only performed in the absence of specific functional groups. For instance, an $H-N-C{=}O$ moiety can be reduced to an $H-N-C$ moiety only if there is no $C{=}S$ moiety present elsewhere in the molecule.

3. A functional group is present more than once. If it is possible to determine a difference in reactivity, the most reactive instance of the functional group will be used. For example, in a coupling reaction between an NH_2 moiety and a halogen atom, an aliphatic halogen atom is preferred to an aromatic one. Otherwise, the reacting group is randomly chosen.

4. A functional group's chemical reactivity may be too low. For instance, whether an aromatic halogen atom is reactive enough to be used in a nucleophilic coupling reaction depends on the other substituents of the aromatic system. There is no problem in performing the reaction if the aromatic system contains electron-withdrawing substituents in the appropriate places, e.g., an *o*-NO_2 moiety. On the other hand, when electron donating groups are present, e.g., an *o*-NH_2 moiety, the reaction will be difficult to impossible, depending on the strength of the attacking nucleophile. It is impractical to implement all electron-withdrawing and -donating effects of both position and nature of functional groups. Currently, SYNOPSIS only considers aromatic halogen atoms without examining the other substituents of the system.

We attempted to account for substituent effects on an aromatic halogen's reactivity with a quantum chemical approach, based on the assumption that a more positive charge on the carbon bound to the halogen implies higher reactivity in nucleophilic coupling reactions. We tested this hypothesis by comparing partial charges from a distributed multipole analysis[33] to observed rate constants. Hartree−Fock wave functions for the dis-

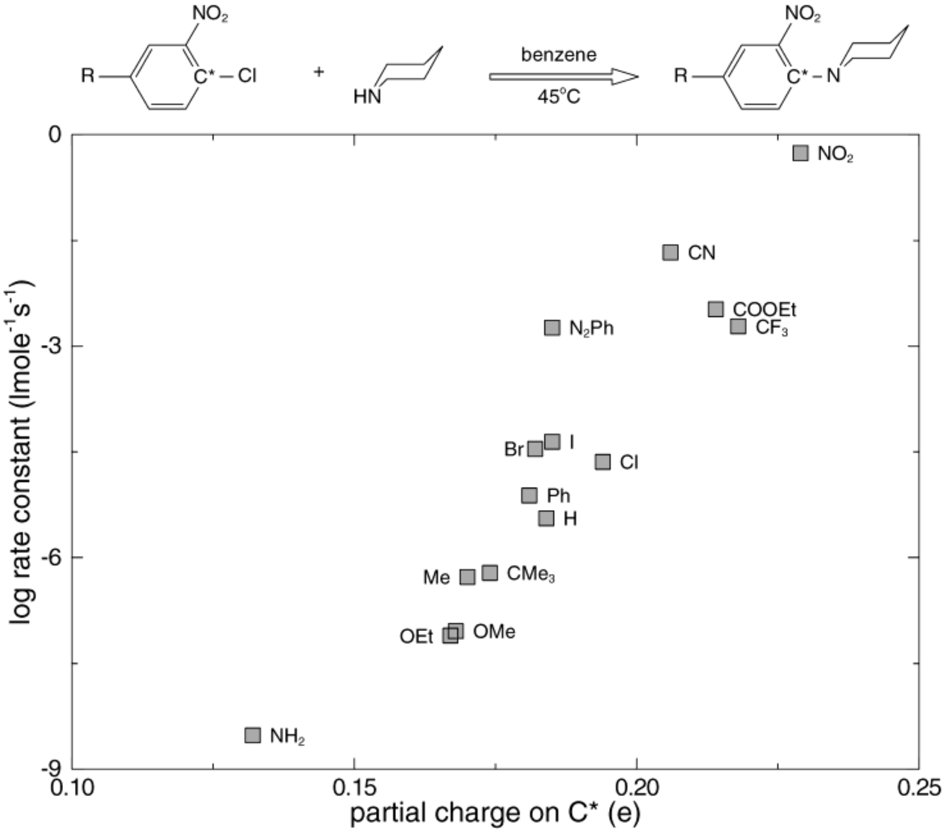

Figure 1. Log rate constant (taken from Miller[53]) versus partial charge on C* for some nucleophilic aromatic substitution reactions.

tributed multipole analysis were computed with the GAMESS−UK package[34] using a 6-31G basis set.

Figure 1 confirms that the computed partial charges on the carbon atom do indeed follow the reactivity. The partial charge on the carbon is only one of the factors determining the reactivity, ignoring different reaction conditions, steric effects of the substituents, and the strength of the attacking nucleophile. However, it may be used to estimate a threshold value for reactivity. If the computed partial charge exceeds this threshold the aromatic halogen is deemed to be reactive enough. Alternatively, if more aromatic halogens are present, the partial charge could be used to choose the more reactive one. For practical purposes the quantum mechanical calculations proved to be too time-consuming, i.e., calculation time is increased beyond the point where it is more economical to accept a certain percentage of suggestions that are not feasible. Without quantum mechanical calculations, SYNOPSIS is able to propose synthesis routes of which 64% was carried out with success in the laboratory.

In addition to a starting database and a reaction set, a fitness function is needed that calculates a property of interest for an arbitrary molecule. Exploiting this function SYNOPSIS will optimize novel molecules for the property. SYNOPSIS scores the generated molecules according to the fitness function provided. During the run, SYNOPSIS is increasingly driven to choose molecules with high computed values as reactants. The algorithm that drives the generation of the molecules toward increasingly better ones contains elements from simulated annealing optimization[35,36] and from genetic algorithm optimization.[37,38] A Metropolis[35] type of function selects the molecule that is used to generate a new one. The algorithm resembles a genetic algorithm in that new offspring is produced from a set of molecules depending on the fitness of the molecules.

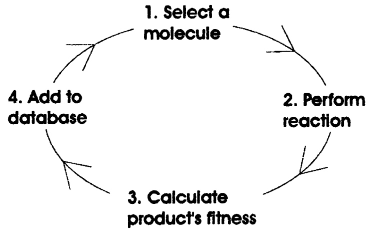

Figure 2. Steps constituting one iteration.

The algorithm is iterative: it performs the same series of steps over and over again exploiting the results from previous steps to get successively closer to a desired result. The steps forming one iteration are depicted in Figure 2.

Step 1: Select a Molecule. A molecule from the database is selected to reproduce by applying a weighted probabilistic function:

$$P_{\mathrm{i}} = \frac{e^{-(Q_{\mathrm{b}} - Q_{\mathrm{i}})/c}}{\sum_{j=1}^{n} e^{-(Q_{\mathrm{b}} - Q_{\mathrm{j}})/c}} \qquad (1)$$

where P_{i} is the probability that the ith molecule is selected, Q_{b} denotes the fitness of the current best molecule, Q_{i} denotes the fitness of the ith molecule, c is a cooling parameter regulating the extent of greediness in the selection, and n is the current population size. This simulated annealing step constitutes the selection pressure. Equation 1 shows that the probability of selecting a molecule depends on the difference between its fitness and the fitness of the current best molecule and also on the current value of the cooling parameter. Equation 1 is used to assign a probability to each molecule whereafter one is selected accordingly. The scores of the molecules present in the initial database are set to a minimum value instead of subjecting the molecules to the fitness function in order to save time. Initially all molecules are equally eligible to become selected because they all have a score equal to Q_b: in the early stages selection is essentially random. As the procedure continues Q_b increases: the chance for a less fit molecule to be selected will decrease. The cooling parameter is initialized and decreased in such a way that increasingly more fit molecules are selected. Since Q_b increases and c decreases, it gets exceedingly difficult for a less fit molecule to become selected: only molecules with fitness close or equal to Q_b will still be selected in the final stages.

Step 2: Perform Reaction. A new molecule is created from the selected one by performing a reaction with it. The reaction according to which the molecule will be transformed is randomly chosen from those that are possible for that particular molecule. If no reaction is possible, another molecule is selected. If the chosen reaction requires two reactants, a suitable partner is randomly picked from the database. If the reaction product about to be generated is already in the database, the procedure will revert to Step 1. Because in genetic terms the decoding of the genotype to phenotype occurs at the reaction level (in an abstract sense the gene of a molecule consists of starting materials and reaction steps), the change in the molecule brought about can be considerable. This is fine in early stages of the run where a broad sampling of the chemical space is wanted. In later stages the algorithm's behavior to optimize molecules is enhanced by a backtracking operator that generates analogues of the fittest molecules. The operator is applied at regular intervals after a certain number of iterations have been reached. It picks a molecule out of the 25 highest scoring molecules that were created from two reactants. The operator changes one of the reactants to a functionally similar one and performs the same reaction to generate a molecule as input for the next step. At this moment, only synthesis properties are used in the selection of another reactant. A more restrictive selection, e.g., by requiring a certain level of similarity, is currently under consideration.

Step 3: Calculate Product's Fitness. The reaction product's score is obtained by subjecting it to the function that calculates the property of interest. In genetic terms this constitutes the fitness function. The fitness function must produce higher values for better molecules. Two examples of fitness functions can be found in the application section.

Step 4: Add to Database. The molecule and the information concerning its reactant(s) and its calculated fitness value are added to the database.

For every generated molecule, the following information is available: its structure, its fitness, and a synthesis route. The synthesis route is by nature of the program composed of a series of steps starting from existing molecules. In building the synthesis route, it is possible to have SYNOPSIS check the presence of the intermediates encountered against a second database of existing materials. Intermediates involved in a synthesis route may already exist, so this check limits the number of steps in the final synthesis route to those that are actually needed.

SYNOPSIS is written in ANSI C and has been successfully compiled and run on SGI IRIX 6.5 and Redhat Linux 7.3. Without backtracking, it takes less than 100 milliseconds to generate a new molecule. In a typical application the rate-limiting step is the evaluation time for the fitness function. Because the interprocess communication consists of only one number, the fitness, the speedup of a run is linear up to hundreds of processors.

Results and Discussion

As a first application SYNOPSIS is used in conjunction with an electric dipole moment computation as fitness function. The computation subjects the molecule, whose dipole moment is to be computed, to a conformational analysis using an in-house developed force field.[39] This force field uses the conjugate gradient minimizer[40,41] as implemented in the TINKER package[42] and a truncated Newton minimizer[43] from the netlib repository.[44] The functional form and parameter set are derived from MMFF94s.[45] The parameter set is extended with respect to the potential types as well as the force constants, to allow for calculation of a broader range of molecules and to maintain compatibility with CVFF[46] parametrized molecules. The AM1 Hamiltonian[47] of MOPAC[48] is used to calculate the dipole moments of the low-energy conformers. The final dipole moment of the molecule is calculated as the sum of the dipole moments of the conformers times a pseudo-Boltzmann weight. The weights are distributed according to the computed energy of the conformers by applying the following function:

$$w_i = 2^{E_g - E_i} \tag{2}$$

where w_i is the weight of the ith conformer, E_g denotes the energy of the lowest energy conformer and E_i denotes the energy of the ith conformer. The weights are normalized after calculation.

These pseudo-Boltzmann weights were used to make the computation more robust with regard to errors in the force field derived energies of the conformers. Since the calculation time increases exponentially with the number of rotatable bonds present in the molecule, the dipole moment computation was set to reject any molecule with more than six torsions. This imposes an effective limit on the size of the generated molecules, because the creation of larger molecules from the initial database will generally be accompanied with an increase in the number of torsions for the created molecule. This has the effect of limiting the achievable dipole moments.

To assess the efficiency of the bias procedure over randomly searching, we ran SYNOPSIS with the same random seed and the selection step (eq 1) set to pure random. A plot of the averaged dipole moment of the top 25 and the highest dipole moment in time is given for both runs in Figure 3. From this figure it is clear that random searching is less efficient. The largest dipole moment molecule found in the simulated anneal-

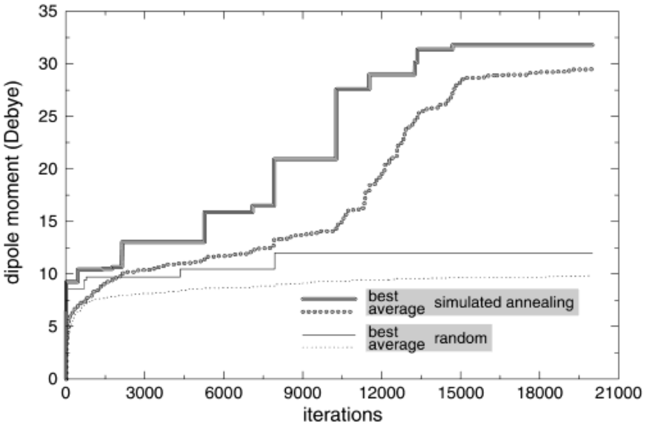

Figure 3. Averaged and best dipole moment of the top 25 in time for random selection and simulated annealing selection.

ing run together with its synthesis route is depicted in Figure 4. The molecule has a computed dipole moment of 31.8 D.

In a second application a computation of the affinity of a putative ligand to a protein binding site is used as the fitness function. The protein binding site used in this application is the nonnucleoside binding pocket of the protein reverse transcriptase from the human immunodeficiency Virus 1 (HIV-RT). The inhibitory strength of a ligand is expressed as an IC_{50} value, which is defined as that concentration of the ligand that gives a 50% protection against HIV-induced cytopathogenicity. The CC_{50} value is the 50% cytotoxic concentration, which is that concentration of the ligand that causes half the cells to die. These values are measured spectrophotometrically based upon the reduction of yellow-colored 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide by mitochondrial dehydrogenases of metabolically active cells to a blue formazan in HIV- and mock-infected MT-4 cell cultures.[49]

The fitness function is an activity computation based on a benchmark set of 34 highly active ligands. The fitness function yields the pIC_{50} value for an arbitrary molecule from its computed binding energy to HIV-RT. The computation involves the docking of all conformers up to 4 kcal/mol from a genetic algorithm based conformation analysis. The docking is done with an in-house written algorithm, that uses a combined Monte Carlo and simulated annealing search on a grid. It computes the docking energy as the sum of the van der Waals, Coulomb and hydrogen bond interactions between the ligand and the protein. The protein is kept rigid during the docking of the set of low energy conformers, while the smoothness of the potential energy function is decreased from a 4−8 to a 6−12 potential. The energetically most favorable complex is, after minimization, used to compute the pIC_{50} value. This value is computed from the sum of nonbonded interaction energies between the molecule and a set of relevant residues. The set of relevant residues was obtained by determining the best correlating set in a linear fit to the experimentally observed pIC_{50} values of the 34 highly active compounds, which had an r^2 of 0.96 (data not shown). This computation takes on average 60 min per molecule per processor. While this computation was used as fitness function to drive the generation of molecules in SYNOPSIS, an alternative model for calculating an arbitrary molecule's pIC_{50} value against HIV was developed. This involved a much larger set of ligands with a much larger spread in experimentally determined IC_{50} values. The steps of the computation remain essentially the same, except that not just the energetically most favorable complex is taken into account, but a range of complexes. The extent to which the different complexes contribute to the activity is given by the Boltzmann weights derived from the total interaction energy between the molecule and the target protein. Using multiple complexes roughly quadruples the calculation time. From a benchmark of 2021 molecules with known experimental IC_{50} values, 1521 were used to determine the set of relevant residues and 500 to validate the method. From this validation set 67% is correctly predicted, where correctly means to within plus or minus one log unit of the experimentally observed pIC_{50} value. The 165 molecules from the validation set that are not correctly predicted can be subdivided in a set of 46 false negatives, i.e., predicted pIC_{50} value more than one log unit lower than the experimental value, and a set of 119 false positives, i.e., predicted pIC_{50} value more than 1 log unit higher than the experimental value. The more elaborate computation was not used in the generation of the designed molecules, it was used a posteriori to compare the predicted values for the final designed molecules; these results will also be given. This computation will be referred to as ‘Model 2’ and the former as ‘Model 1’.

SYNOPSIS was run a number of times with different random seeds. From these runs molecules out of the top 25 were selected to be synthesized. The candidates were selected based on the following considerations: candidates must be chemically diverse and different from known nonnucleoside reverse transcriptase inhibitors, suggested synthesis route involves only a few steps (preferably just one step), and the suggested synthesis steps are deemed feasible by an organic chemist. This application resulted in the selection of 28 different designs and the effective synthesis of 18 molecules. The IC_{50} and CC_{50} values of the 18 molecules whose synthesis succeeded were experimentally determined. Table 1 gives an overview of the results; the structures of the molecules are shown in Figure 5.

When the second column in Table 1 states *app*, the synthesis route followed is approximately the same as the route followed by SYNOPSIS. If the suggestion from SYNOPSIS was substituting a chlorine atom and in practice this was done on a fluorine atom, that would count as approximately the same. Also when the suggested route involved the coupling of a functional group A on reactant 1 and a functional group B on reactant 2 and the actual synthesis route followed proceeded by coupling functional group A on reactant 2 and functional group B on reactant 1, *app* is stated. In some cases a different synthesis route was followed altogether. This might be because a starting material or an analogue thereof was not available, or—more often—that the suggested synthesis route was deemed not to be the best in terms of simplicity, price, or chance of success by the synthesis laboratory. When the third column in Table 1 indicates *chem* or *comp*, the synthesis did succeed, but the actual synthesized molecule is not exactly the same as the designed one, but a closely related analogue. The

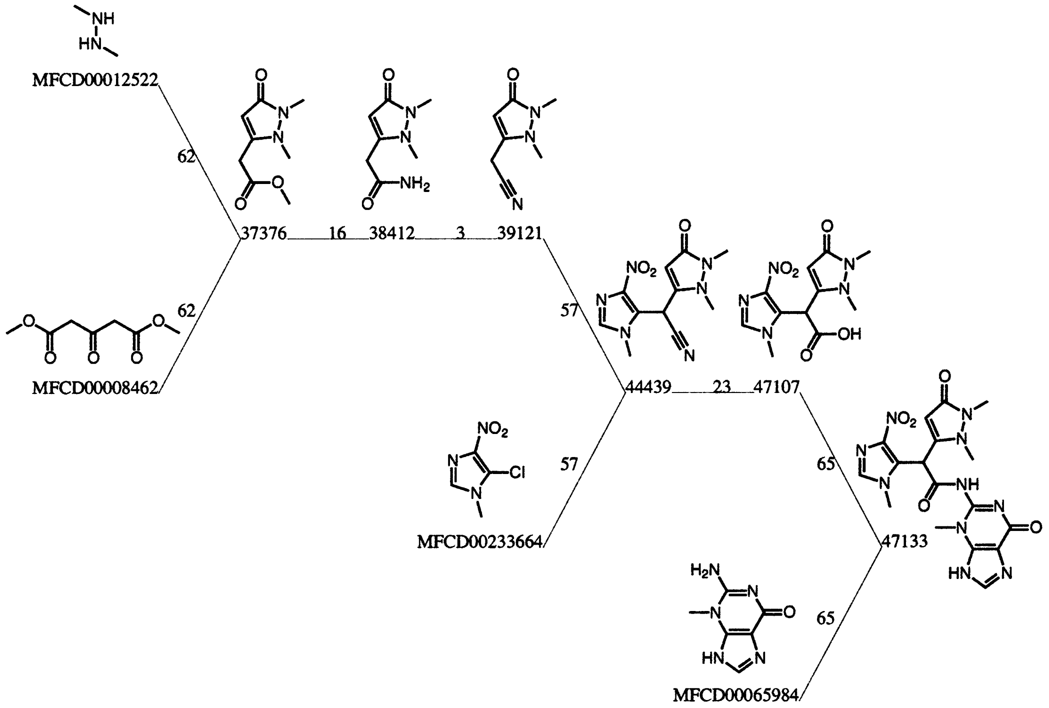

Figure 4. Largest dipole moment molecule found and the synthesis route taken. The number below a molecule is its index in the database. An identifier starting with MFCD points to an entry in the ACD. At the center of each line the reaction step number is indicated.

Table 1. Experimental Results of the Designed Inhibitors in Chronological Order[a]

molecule no.	route followed	synthesis succeeded	model 1 pIC_{50}	model 2 pIC_{50}	observed pIC_{50}	observed pCC_{50}	class
1	yes	yes	5.6	5.0	4.9	<4.0	active
2	app	chem	6.4	7.3	<4.4	4.4	toxic
3	yes	no	8.4	4.5			
4	yes	yes	8.2	5.3	4.1	<4.0	active
5	yes	yes	8.4	7.5	<4.3	4.3	toxic
6	app	yes	8.2	6.2	4.6	<4.0	active
7	yes	yes	8.3	5.5	4.8	<4.3	active
8	yes	yes	8.0	7.5	<4.0	<4.0	inactive
9	yes	chem	8.0	3.4	<4.0	<4.0	inactive
10	no	yes	8.1	5.7	4.5	<4.0	active
11	yes	yes	9.6	5.3	<4.0	<4.0	inactive
12	app	no	8.2	6.9			
13	yes	yes	8.1	9.3	<5.1	5.1	toxic
14	yes	yes	8.4	8.0	<4.3	4.3	toxic
15	app	no	8.2	4.2			
16	app	yes	8.1	6.1	4.4	<4.0	active
17	app	no	8.0	7.4			
18	no	no	8.7	5.6			
19	yes	chem	9.8	5.9	<4.0	<4.0	inactive
20	yes	no	8.1	8.3			
21	app	no	8.0	7.2			
22	app	no	8.2	6.6			
23	app	comp	8.8	7.0	7.0	<4.0	active
24	app	no	8.2	6.8			
25	yes	comp	8.5	5.9	5.8	4.3	active
26	yes	no	8.0	4.9			
27	no	chem	8.7	8.0	5.2	<4.0	active
28	no	comp	9.3	7.8	5.6	<4.0	active

[a] The corresponding molecular structures are depicted in Figure 5. In the second column, *app* means the actual synthesis route was slightly modified. In the third column, *chem* means designed molecule was slightly modified for reasons of synthesis and *comp* means designed molecule was slightly modified after additional computations. The next two columns show the outcome of the pIC_{50} computation. Model 1 was used in the design process and Model 2 serves for comparison. The observed pIC_{50} and pCC_{50} values were determined in a cellular HIV inhibition assay.[49] The final column indicates *active* if the pIC_{50} is higher than 4 and higher than the pCC_{50}, *toxic* if the pCC_{50} is higher than 4 and higher than the pIC_{50}, and *inactive* if both the pCC_{50} and the pIC_{50} are less than 4.

label *chem* signifies that the decision to synthesize an analogue instead of the original molecule resulted from synthetic chemical considerations (e.g., enabling cheap or readily available starting materials instead of expensive or rare ones). The label *comp* means that the decision to synthesize an analogue sprung from reasons of computational origin (traditional optimization with computational chemistry leading to improved variants). The column headed 'Model 1' gives the results of the pIC_{50} computation that was used as fitness function. The result in cases where an analogue was made because of synthetic chemical considerations applies to the original designed molecule. The column headed 'Model 2' gives the results of the extended computation

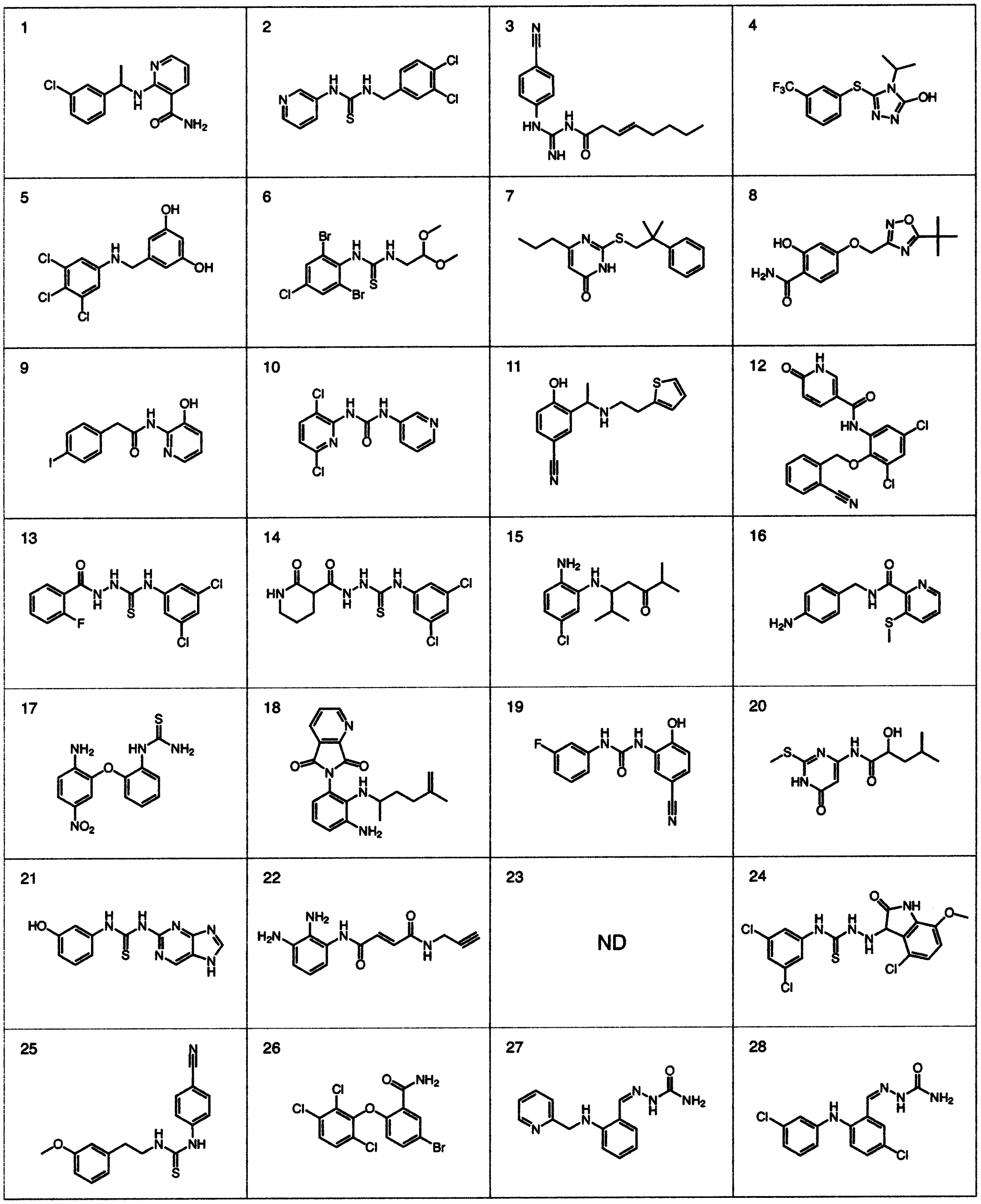

Figure 5. Structures of the molecules from Table 1. The structure of compound **23** is not disclosed.

as described earlier. The next two columns give the experimentally determined values for the pIC_{50} and pCC_{50}. The final column classifies a molecule as *active* if the experimental pIC_{50} is higher than 4 and higher than the pCC_{50}, *toxic* if the pCC_{50} is higher than 4 and higher than the pIC_{50}, and *inactive* if both the pIC_{50} and pCC_{50} are less than 4.

From Table 1 and Figure 6 it can be seen that 28 designed molecules resulted in 18 synthesized molecules. The set of 10 molecules that could not be synthesized in a reasonable timespan despite the expectations of an organic chemist otherwise amounts to 36%. From the set of 18 synthesized and experimentally tested molecules, 10 showed inhibitory activity, four were cytotoxic, and four molecules were inactive.

So 56% of the synthesized molecules proved to be active in vitro below the 100 μM level, which compares favorably to results from a typical HTS experiment[50] and also to results from a biased HTS experiment.[51] The activity of the four molecules which caused the cells to die can be classified as unknown. If one had a computational model predicting cellular toxicity, one could include this in the fitness function to prevent creation of such compounds.

The fact that 22% of the molecules were inactive, illustrates that the pIC_{50} computation is not perfect. Furthermore all tested compounds had high computed pIC_{50} values and showed only weak to moderate activity in the experiment. The extended computation ('Model 2') performs better in calculating the experimental pIC_{50} value, although even this computation is not as reliable as one would wish. A more rigorous test of this model would be to repeat the design process with this computation as fitness function. A number of reasons can be thought of to explain the discrepancies between the computed and experimental values. First of all, the pIC_{50} computation is set up using only measurably active molecules. The molecules in the benchmark may

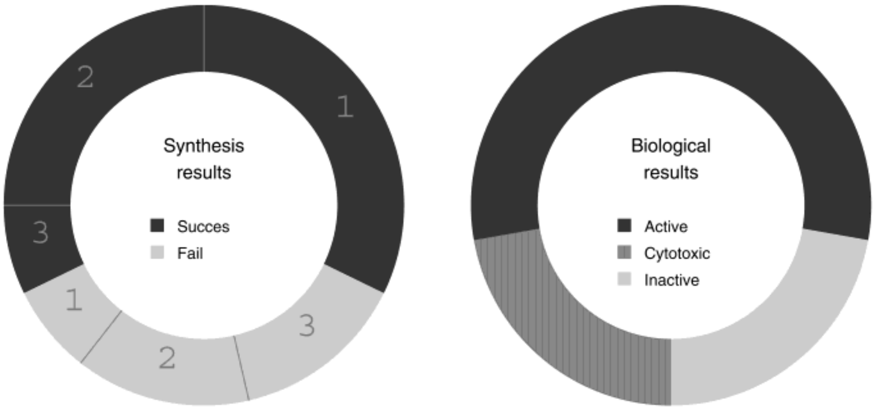

Figure 6. An overview of the experimental results. The numbers in the left pie indicate the number of steps involved in the synthesis.

have some features necessary to account for their activity in common. If that is the case, there is no need for the model to incorporate these features to reproduce the activities. An indication of the presence of this effect can be found in the better performance of the extended computation, where not only highly active but also weakly active molecules were included in the benchmark. One could consider including completely inactive molecules in setting up the pIC_{50} computation; however, this might obscure matters since the cause of the inactivity is unkown. Experimentally, the IC_{50} value is a whole cell measurement. Some molecules possess inhibitory activity which does not show up in the assay because the molecule never reaches the interior of the cell. Since cell penetration is not part of the model as is, these molecules will turn out as a false positive in the computation. For a few of the false positives from the benchmark this phenomenon has been confirmed by comparison of whole cell and enzyme activities. The cell penetration uncertainty can be avoided by direct optimization of binding constants instead of IC_{50} values, at the cost of having to resolve any problems with cell penetration later on. The pIC_{50} computation assumes that the inhibitory activity of a molecule results from binding to the nonnucleoside binding pocket of HIV-RT. Consequently, the activity of a molecule that binds to a different site of the HIV-RT protein or a different protein altogether cannot be expected to be accurately computed. If the molecule binds to a protein without inhibiting HIV, its activity most likely would be overestimated. Conversely, if the alternative binding of the molecule does inhibit HIV, its activity will be underestimated. Examination of the false negatives from the benchmark molecules showed at least two ligands that are known nucleoside inhibitors of HIV, highly active but deriving their activity from a different mechanism. An incorrect activity computation would also result if a molecule is broken down by any of the components in the assay, whether reagents or cell enzymes. A last source of error in the computation of a molecule's pIC_{50} relates to chiral compounds. When confronted with a chiral compound, the computation will assess the best binding stereoisomer automatically and use that one to calculate the activity. The designed molecules that are chiral, seven in total, were synthesized as racemic mixtures. Depending on the activity of the other stereoisomer, the error in the computed activity will be between 0 and 0.3.

Despite the noted limitations, a range of simple molecules is generated with an extremely high proportion of active lead[52] compounds compared to other methods of lead finding or generating. This demonstrates the merits of SYNOPSIS in drug discovery.

Conclusions

We have developed a computer program SYNOPSIS. This program, provided with a method to calculate a property of interest, generates synthetically feasible molecules with as much of the desired property as possible while remaining within synthetic constraints. We have used SYNOPSIS in conjunction with a computation of pIC_{50} values for putative ligands binding to HIV reverse transcriptase. This has proven its value in computational drug design: 18 of the 28 designed molecules could readily be synthesized, and 10 of the synthesized molecules showed HIV inhibitory activity in vitro.

References

(1) Labinger, J. A. Tuning into better catalysts. *Science* **1995**, *269*, 1833−1833.

(2) Thomas, J. M. Design, synthesis, and in situ characterization of new solid catalysts. *Angew. Chem., Int. Ed.* **1999**, *38*, 3588−3628.

(3) Venkatasubramanian, V.; Chan, K.; Caruthers, J. M. Computer-aided molecular design using genetic algorithms. *Comput. Chem. Eng.* **1994**, *18*, 833−844.

(4) Ulmer, II, C. W.; Smith, D. A.; Sumpter, B. G.; Noid, D. I. Computational neural networks and the rational design of polymeric materials: The next generation polycarbonates. *Comput. Theor. Polymer Sci.* **1998**, *8*, 311−321.

(5) Dahiyat, B. I.; Sarisky, C. A.; Mayo, S. L. De novo protein design: Towards fully automated sequence selection. *J. Mol. Biol.* **1997**, *273*, 789−796.

(6) Hellinga, H. W. Rational protein design: Combining theory and experiment. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 10015−10017.

(7) Street, A. G.; Mayo, S. L. Computational protein design. *Structure* **1999**, *7*, 105−109.

(8) Joseph-McCarthy, D. Computational approaches to structure-based ligand design. *Pharmacol. Ther.* **1999**, *84*, 179−191.

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

(10) Lunney, E. A. Structure-based drug design begins a new era. *Med. Chem. Res.* **1998**, *8*, 352−361.

(11) Murcko, M. A.; Caron, P. R.; Charifson, P. S. Structure-based drug design. *Annu. Rep. Med. Chem.* **1999**, *34*, 297−306.

(12) Shoichet, B. K.; Bussiere, D. E. The role of macromolecular crystallography and structure for drug discovery: Advances and caveats. *Curr. Opin. Drug Discov. Devel.* **2000**, *3*, 408−422.

(13) 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.

(14) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(15) Bohacek, R. S.; McMartin, C. Multiple highly diverse structures complementary to enzyme binding sites: Results of extensive application of a de novo design method incorporating combinatorial growth. *J. Am. Chem. Soc.* **1994**, *116*, 5560−5571.

(16) DeWitte, R. S.; Shakhnovich, E. I. SMoG: De novo design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence. *J. Am. Chem. Soc.* **1996**, *118*, 11733−11744.

(17) Gillet, V. J.; Johnson, A. P.; Mata, P.; Sike, S. Automated structure design in 3D. *Tetrahedron Comput. Methodol.* **1990**, *3*, 681−696.

(18) Nishibata, Y.; Itai, A. Automatic creation of drug candidate structures based on receptor structure. Starting point for artificial lead generation. *Tetrahedron* **1991**, *47*, 8985−8990.

(19) Rotstein, S. H.; Murcko, M. A. GenStar: A method for de novo drug design. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 23−43.

(20) Rotstein, S. H.; Murcko, M. A. GroupBuild: A fragment-based method for de novo drug design. *J. Med. Chem.* **1993**, *36*, 1700−1710.

(21) Moon, J. B.; Howe, W. J. Computer design of bioactive molecules: A method for receptor-based de novo ligand design. *Proteins: Struct., Funct., Genet.* **1991**, *11*, 314−328.

(22) Böhm, 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.

(23) Miranker, A.; Karplus, M. An automated method for dynamic ligand design. *Proteins: Struct., Funct., Genet.* **1995**, *23*, 472−490.

(24) Tschinke, V.; Cohen, N. C. The NEWLEAD program: A new method for the design of candidate structures from pharmacophoric hypotheses. *J. Med. Chem.* **1993**, *36*, 3863−3870.

(25) Gehlhaar, D. K.; Moerder, K. E.; Zichi, D.; Sherman, C. J.; Ogden, R. C.; Freer, S. T. De novo design of enzyme inhibitors by Monte Carlo ligand generation. *J. Med. Chem.* **1995**, *38*, 466−472.

(26) Lewis, R. A.; Dean, P. M. Automated site-directed drug design: The formation of molecular templates in primary structure generation. *Proc. R. Soc. London, Ser. B* **1989**, *236*, 141−162.

(27) Pearlman, D. A.; Murcko, M. A. CONCEPTS: New dynamic algorithm for de novo drug suggestion. *J. Comput. Chem.* **1993**, *14*, 1184−1193.

(28) Pearlman, D. A.; Murcko, M. A. CONCERTS: Dynamic collection of fragments as an approach to de novo ligand design. *J. Med. Chem.* **1996**, *39*, 1651−1663.

(29) Gillet, V. J.; Myatt, G.; Zsoldos, Z.; Johnson, A. P. SPROUT, HIPPO and CAESA: Tools for de novo structure generation and estimation of synthetic accessibility. *Perspect. Drug Discov. Des.* **1995**, *3*, 34−50.

(30) Holloway, M. K. A priori prediction of ligand affinity by energy minimization. *Perspect. Drug Discov. Des.* **1998**, *9/10/11*, 63−84.

(31) Schneider, G. 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.

(32) *Available Chemicals Directory*; Molecular Design Limited Information Systems, San Leandro, CA, 1999.

(33) Stone, A. J. Distributed multipole analysis; or how to describe a molecular charge distribution. *Chem. Phys. Lett.* **1981**, *83*, 233−239.

(34) Guest, M. F.; van Lenthe, J. H.; Kendrick, J.; Schoffel, K.; Sherwood, P. *GAMESS-UK*; CCLRC Daresbury Laboratory: Daresbury, UK, 1999.

(35) Metropolis, N.; Rosenbluth, A.; Rosenbluth, M.; Teller, A.; Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087−1092.

(36) Kirkpatrick, S.; Gelatt, C. D. J.; Vecchi, M. P. Optimization by simulated annealing. *Science* **1983**, *220*, 671−680.

(37) Holland, J. *Adaptation in natural and artificial systems;* MIT Press: Cambridge, MA, 1992.

(38) Clark, D. E.; Westhead, D. R. Evolutionary algorithms in computer-aided molecular design. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 337−358.

(39) de Jonge, M. R. *CMDFF*, in preparation.

(40) Luenberger, D. G. *Linear and nonlinear programming*; Addison-Wesley: Reading, MA, 1984.

(41) Nash, S. G.; Sofer, A. *Linear and nonlinear programming*; McGraw-Hill: New York, 1996.

(42) Ponder, J. W. *TINKER: Software tools for molecular design*; Washington University, School of Medicine: St. Louis, MO, 1999.

(43) Nash, S. G. Newton-type minimization via the Lanczos method. *SIAM J. Numer. Anal.* **1984**, *21*, 770−778.

(44) http://www.netlib.org.

(45) Halgren, T. A. MMFF VI. MMFF94s option for energy minimization studies. *J. Comput. Chem.* **1999**, *20*, 720−729.

(46) Dauber-Osguthorpe, P.; Roberts, V. A.; Osguthorpe, D. J.; Wolff, J.; Genest, M.; Hagler, A. T. Structure and energetics of ligand binding to proteins: E. coli dihydrofolate reductase-trimethoprim, a drug-receptor system. *Proteins: Struct., Funct., Genet.* **1988**, *4*, 31−47.

(47) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.

(48) Stewart, J. J. P. *MOPAC*; Quantum Chemistry Program Exchange (QCPE #455); Indiana University: Bloomington, IN, 1990.

(49) Pauwels, R.; Balzarini, J.; Baba, M.; Snoeck, R.; Schols, D.; Herdewijn, P.; Desmyter, J.; De Clercq, E. Rapid and automated tetrazolium-based colorimetric assay for the detection of anti-HIV compounds. *J. Virol. Methods* **1988**, *20*, 309−321.

(50) Teague, S. J.; Davis, A. M.; Leeson, P. D.; Oprea, T. The design of leadlike combinatorial libraries. *Angew. Chem., Int. Ed.* **1999**, *38*, 3743−3748.

(51) Böhm, H. J.; Boehringer, M.; Bur, D.; Gmuender, H.; Huber, W.; Klaus, W.; Kostrewa, D.; Kuehne, H.; Luebbers, T.; Meunier-Keller, N.; Mueller, F. Novel inhibitors of DNA Gyrase: 3D structure based biased needle screening, hit validation by biophysical methods, and 3D guided optimization. A promising alternative to random screening. *J. Med. Chem.* **2000**, *43*, 2664−2674.

(52) Hann, M. A.; Leach, A. R.; Harper, G. Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 856−864.

(53) Miller, J. *Aromatic nucleophilic substitution;* Elsevier: Amsterdam: The Netherlands, 1968.

JM030809X