A Novel Search Engine for Virtual Screening of Very Large Databases

David Vidal,[†] Michael Thormann,*[,‡] and Miquel Pons*[,†,§]

Laboratory of Biomolecular NMR, Barcelona Biomedical Research Institute, Parc Científic de Barcelona,
Josep Samitier, 1-5 08028 Barcelona, Spain, Origenis GmbH, Kopernikusweg 1, 82152 Martinsried, Germany,
and Departament de Química Orgànica, Universitat de Barcelona, Martí i Franquès, 1-11,
08028- Barcelona, Spain

Received October 18, 2005

Virtual screening of large chemical databases using the structure of the receptor can be computationally very demanding. We present a novel strategy that combines exhaustive similarity searches directly in SMILES format with the docking of flexible ligands, whose 3D structure is generated on the fly from the SMILES representation. Our strategy makes use of the recently developed LINGO tools to extract implicit chemical information from SMILES strings and integrates LINGO similarities into a pseudo-evolutionary algorithm. The algorithm represents a combination of a fast target-independent similarity method with a slower but information richer target-focused method. A virtual search of FactorXa ligands provided 62% of the potential hits after docking only 6.5% of a database of nearly 1 million molecules. The set of solutions showed good diversity, indicating that the method shows good scaffold hopping capabilities.

INTRODUCTION

Drug discovery and development is an interdisciplinary, expensive, and time-consuming process, in which pharmaceutical companies can spend, on average, $880 million and 15 years of research time[1,2] until a drug finally enters the market. The identification of small-molecule modulators of protein function, called hits, is commonly the starting point for any drug discovery project.[3] Computational methodologies can offer significant advantages at this stage, with an estimated $130 million and 0.8 years average reduction in the cost of drug discovery.

Virtual screening methodologies are widely used as part of many discovery programs to select a limited number of promising compounds from the huge number of possibilities present in chemical libraries of available or virtual products. The aim of virtual screening is to reduce the costs associated with HTS (high-throughput screening) methods[4] by reducing the size of the libraries that will be experimentally synthesized and screened while increasing the number of hits. Virtual screening can be combined with the prediction of molecular properties to eliminate, at an early stage, those compounds that have unfavorable ADME (absorption, distribution, metabolism, and excretion) properties.[5]

There are two main classes of virtual screening methods: *similarity-based methods* and *receptor-based methods*. The former class is based on the similarity principle or neighborhood behavior,[6,7] which states that structurally similar molecules have a higher probability of presenting similar activities. Medicinal chemistry relies on the concept of bioisosterism, in which similar substructures may be interchanged while maintaining some degree of activity.[8] There are many different measures of similarity, but in general, they tend to be not too computationally demanding and are favored for the screening of large databases. However, similarity-based methods tend to find results that are closely related to known active products and perform poorly for the identification of new scaffolds. Receptor-based methods require some knowledge about the structure of the receptor, or at least of its active site, and are based on modeling molecular recognition. In these methods, the aim is to find molecules that, in a proper orientation and conformation, achieve the highest degree of complementarity between the ligand and the protein.[9] Receptor-based methods are computationally much more demanding than similarity-based methods and are usually restricted to small databases.

Corporate and public libraries often contain several million compounds and already exceed the capacity of current virtual high-throughput screening methodologies. With the widespread use of combinatorial chemistry and virtual chemistry tools, the size of the libraries is expected to grow exponentially. There is clearly a need for new, very efficient screening methodologies suitable for the screening of very large databases. In our view, the most promising avenues include the application of learning strategies[10,11] and the combination of similarity-based and receptor-based virtual search methods.[1] Efficient handling of chemical information will be crucial as the size of the databases increases.

Even with the fastest current docking algorithms and supercomputers, an exhaustive screening of a database of millions of compounds remains a formidable task. A realistic approach to high-throughput receptor-based virtual screening requires restricting the number of compounds to be actually analyzed to a target-focused small fraction of the database. In addition, data storage and handling requirements of 3D structures needed by docking algorithms constitute, per se, a challenge when the size of the database is very large. The SMILES format is highly compressed and especially indicated to store and manage large databases.[12,13]

[†] Parc Científic de Barcelona.
[‡] Origenis GmbH.
[§] Universitat de Barcelona.
* Corresponding authors. E-mail: mpons@ub.edu (M.P.) and michael.thormann@origenis.com (M.T.).

10.1021/ci050458q CCC: $33.50
Published on Web 02/03/2006

Here, we describe a new algorithm that optimizes the selection of compounds that will be docked from a database in SMILES format, to maximize the number of hits in a given small fraction of the total database. The outline of this paper is the following. First, an overview of the complete process will be presented, with a brief introduction to genetic algorithms. In Section 2, ADME filtering of the database and automatic generation and parametrization of the 3D structures of ligands from SMILES representations will be described. Ligand scoring is based on docking calculations using Autodock 3.0, although the selection strategy could be easily adapted to any other scoring method. Details of our scoring protocol are given in Section 3. In Section 4, we describe, in detail, the strategy used to select the sets of molecules that form each generation in our evolutionary algorithm. Parameter optimization is described in Section 5, and an example of virtual screening for ligands of FactorXa is presented in Section 6.

RESULTS AND DISCUSSION

1. Methodological Overview. Two key points define our protocol: (i) databases are stored and processed directly using SMILES representations and (ii) only a selected, hit-enriched fraction of the database is actually used for receptor docking.

Compound selection is performed dynamically, on the basis of the computed similarity to some of the highest scoring molecules previously found during the search. Very efficient similarity calculations can be carried out directly on SMILES representations using our recently described LINGOsim program.[16] LINGO-based tools are also used to predict molecular properties directly from SMILES representations and to remove compounds for which unfavorable ADME properties are predicted.

Compound selection can be viewed as a target-focused reduction process of the chemical space represented in the form of the static compound database. Similarity measures are assumed to provide target-independent structural neighbors. In combination with a target-dependent scoring function, they can provide an optimal path through the chemical space that is especially rich in potential ligands for that particular target. Scoring functions provide a receptor−ligand interaction energy assumed to be lowest for the best ligands. Genetic algorithms (GAs) have repeatedly proved their efficiency and robustness as optimization tools in very large, multidimensional, continuous spaces. However, the chemical space is formed by discrete molecules for which a generally applicable genetic representation can hardly be found. For this reason, we have developed a pseudo-genetic algorithm adapted to our selection process. This search engine retains the advantages of conventional GAs and is applicable to problems where no suitable genetic representations can be formulated.

The performance of the algorithm depends on a number of free parameters. Since these parameters mutually interact in a complex way, their optimization was carried out using a classical genetic algorithm.

Finally, the method was validated in a virtual screening of an initial library of more than 2 million compounds aimed at finding Factor Xa hits, that is, ligands with low binding energies. The number and the diversity of the hits thus identified were compared with the results of a random search of the same database.

Genetic Algorithms. Genetic algorithms[14] are used at three different levels in our methods, and a brief introduction seems appropriate for the general reader; however those knowledgeable in this field may skip the rest of this section. Genetic algorithms are optimization methods inspired in the process of biological evolution. An “individual” represents a possible solution to the problem being considered. In a docking problem, each “individual” represents a different combination of position and conformation of the ligand. In a database search, the “individuals” represent molecules present in the database. In an algorithm optimization problem, “individuals” represent a set of specific values adopted by the parameters that control the algorithm.

A set of individuals form a population. Populations evolve according to a set of rules that describe which individuals are maintained or eliminated and how to select new individuals. Evolution is governed by some “fitness” scoring. Best fit individuals have a higher probability of being selected and their traits being transmitted to the next generation. In classical genetic algorithms, an individual is composed of different “genes” or independent traits that can be combined in different ways. For example, the position and the orientation of a molecule or the dihedral angle describing the rotation around a particular bond are “genes” in a docking problem. New individuals may be created from existing ones by taking each particular gene from a different parent, a process known as “recombination”. For some particular problems, such as identifying the best ligand in a database, the complete individual can be considered a single gene and no proper recombination is possible. Random modification of a particular gene is a different way of creating a new individual, and the process is known as “mutation”.

Fitness is evaluated for the individual as a whole, and therefore, interactions between genes are taken into account. Genetic algorithms are nondeterministic, and the subject of the evolution is the ensemble of the population, which can be characterized by the statistical distribution of selected properties.

From an initial random population, genetic algorithms drive the evolution through several generations. A shift in the property distribution occurs through a selection mechanism that biases the probability of transfer of particular characters to the next generations. The probability that the best possible solution is present in a population increases during the evolution.

In classical genetic algorithms, new individuals are generated by crossover from an almost infinite number of virtual possibilities by choosing new combinations of the genes that are present in the population after a selection process. Variability in the gene population is steadily maintained by random mutations. In our database-searching approach, the evolving population is a small subset of a finite population and the algorithm selects the individuals that will be added to the subset at each generation in order to bias the property distribution in the desired direction. The selection is made on the basis of similarity to the members of the subset that display the desired characteristics.

Chemical similarity is an elusive concept, and there are numerous ways of defining similarity measures. The pairwise comparison of similarities, measured by different methods, of a collection of molecules to a given target may show poor correlation. However, all of them may show good neighborhood behavior in the sense that more similar molecules have a higher probability of displaying similar properties than less similar pairs.

This probabilistic interpretation of similarity suggests that similarity-based searches should be carried out using nondeterministic methods, such as the GA used in this work.

The efficiency of the algorithm depends on the size of the populations, the number of generations or the proper choice of the genetic operators (e.g., recombination and mutation), and their probabilities. The complexity of the problem increases with the number of degrees of freedom or "genes", and in general, more generations or larger populations will be required to find the best solution with a given probability.

2. Database Preparation and SMILES-to-PDB Conversion. A database was prepared by combining several commercially available databases and converted to SMILES format (see Methods). The initial database contained more than 2 million entries. An initial filtering stage was used to remove duplicate entries and those molecules with unfavorable ADME properties, according to criteria inspired by Lipinski's rule of five:[15] molecular weight between 100 and 450 Daltons, log P less than (or equal to) 5, number of hydrogen-bond acceptors less than (or equal to) 10, hydrogen-bond donors less than (or equal to) 5, number of rotatable bonds less than 9, and surface polar area less than (or equal to) 150 Å^2. Molecular descriptors, including log P values, were derived directly from the SMILES representations using our recently developed LINGO-QSPR methods.[16] The final database used in this work contained 913 763 compounds in SMILES format.

3D coordinates were generated on the fly, *only for those molecules that were to be docked*, from the SMILES format in a completely automated protocol. The procedure comprises different stages. First, the protonation state of ionizable groups at pH 7.4 is determined by using the *cxcalc* module of the *JChem* software,[17] which calculates the major microspecies at a given pH, thus giving the corresponding modified SMILES codes. Next, 3D coordinates and charges are calculated with the MAB algorithm by using the *Msmab* module of the *Moloc* software.[18] Finally, rotatable bonds are identified and compounds are converted to the desired final PDBQ (PDB + charges) molecular format by using the *autotors* module of *Autodock*. The *autotors* module is specifically designed to analyze peptides, so the output PDBQ file needs to be modified by a homemade program which explores the molecular graph in order to explicitly block those torsions that have high barriers, such as those around $C_{sp2}-C_{sp2}$ bonds.

3. Scoring. Scoring is based on a modified version of Autodock 3.0 that uses an enhanced genetic algorithm to find the minimum energy solution to the docking problem defined by the position, orientation, and conformation of the ligand with respect to a rigid protein structure.[19] Autodock 3.0 is a very fast, grid-based docking method that uses a genetic algorithm as a global optimizer combined with energy minimization as a local search method. Our previously reported enhanced genetic algorithm and multiple docking protocol makes it a very robust search method for ligands with several rotatable bonds.[20] Each rotatable bond adds an additional degree of freedom and decreases the probability of finding the solution of lowest energy after a certain number of generations. The probability of finding the global solution can be increased by running multiple evolutions. Thus, the number of runs used per compound in the docking calculation is set as a function of the number of the rotatable bonds of the molecule. Thus, for molecules with less than two allowed torsions, two runs are calculated; between three and five, four runs are done; and between six and eight, six runs are required. Molecules with more than eight rotatable bonds are directly excluded by the application of the initial filters.

The standard Autodock force field gives acceptable structures for tight complexes. In our experience, however, sometimes nonoptimal complexes are predicted in which polar ligand groups are being buried in hydrophobic pockets of the protein. This is the result of not considering the penalty associated with removing a polar group from the solvent. We have extended the volume-based method of Stouten[19] already used by Autodock 3.0 for carbon atoms to all polar atoms in the ligand. Our implementation leaves unchanged the energies of tight complexes but increases the energy of unrealistic structures.

After the docking calculations, the output file is automatically analyzed, and for each compound, the lowest energy conformation is written to the list of results.

4. An Evolutionary Algorithm to Select Compounds from the Database. On a single CPU, only a fraction of a large database can actually be screened through docking calculations with a receptor structure. To compensate for this limitation, it is important that the selected subset is enriched in potential hits. However, since this information is not usually known for a new target, the search is started using a population of molecules (individuals) chosen randomly from the database. Molecules are ranked using a scoring function, as described in Section 3. The ranked list is used as input to select the new population from a database, in a way reminiscent of a standard genetic algorithm but with major differences: (i) No offspring are generated by parents; the new generation is instead ***adopted***: candidate molecules are not created using the information contained in a previous population but are selected from a fixed list according to their similarity. (ii) Standard genetic operators cannot be applied since no genetic representation for the database exists: while recombination is not possible, mutation can be assimilated to random selection from the database. And finally, (iii) compartmentalization is accounted for by following a number of alternative, independent solutions.

How is the information from one generation used to drive the selection of a new population? We make use of the "neighborhood principle" on which medicinal chemistry is based: similar compounds should display similar properties, and new generations are selected from a similarity matrix calculation of the complete database to a set of molecules (that we call "parents") that include the best ranked molecules in previous generations.

The choice of a proper similarity measure is a key point of our algorithm. Ideally, similarity calculations should be performed *directly from the SMILES representations, the initial representation of our static database*, avoiding the time-consuming need to generate 3D structures of the complete database. They should display reasonable neighborhood behavior; that is, a set of molecules displaying high similarity to a reference compound should be enriched in

Table 1. Definition of MPA Parameters

MPA parameters	description
population	total number of individuals per generation
PARENTS_select	number of individuals selected as parents based on fitness
PARENTS_random	number of individuals randomly selected as parents
CHILD_select	number of individuals to be adopted by similarity to each parent
CHILD_random	number of individuals to be randomly selected from the database
AGED_select	number of generations an adopted individual is eligible as parent
AGED_random	number of generations a randomly selected individual is eligible as parent
PROBchild	number of top similarity ranked individuals in the sets from which children will be adopted by each parent

compounds showing properties similar to those of the reference one. Finally, similarity matrix calculations have to be very fast, since the complete database has to be compared to each of the parent molecules. We have recently introduced LINGOsim,[16] a method that has all the desired characteristics with a throughput of 75 000 comparisons per second on a Pentium 4 Xeon 3.0 GHz machine with 2 GB of RAM. LINGOsim provides a good discrimination between bioisoster pairs and random pairs, showing that a LINGO-based similarity measure captures biologically relevant similarities.

The chemical space of the database is explored by a combination of global exploration random walks and more intensive local searches around previously found promising candidates. This is a time-honored popular strategy for people picking (edible) mushrooms, a popular and rewarding pastime in several regions in Europe. For this reason, we refer to the evolutionary **M**assive **P**rocessing **A**lgorithm described as the “**M**ushroom **P**icker’s **A**lgorithm” or MPA.

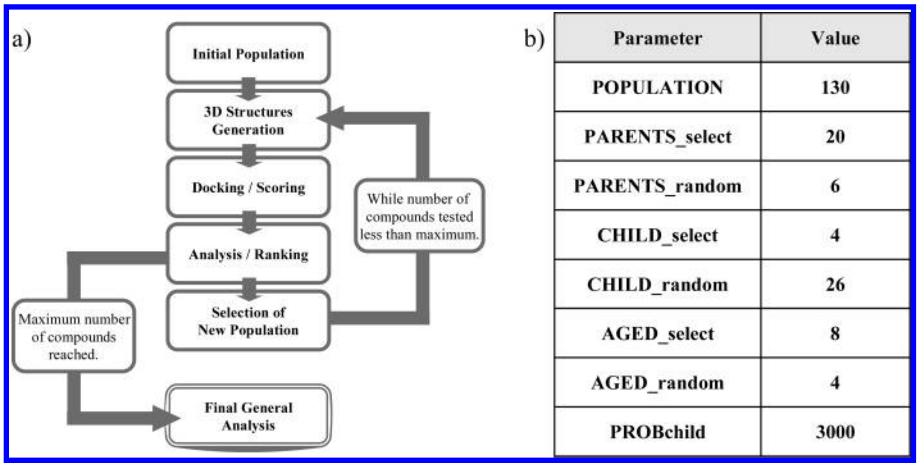
At each generation, a certain number of molecules are chosen as “parents”. These will be used to select new individuals (“children”) from a list of similar molecules present in the database. A number of parents, given by the parameter *PARENTS_select*, are chosen from the top-ranked allowed compounds in a list that contains the set of molecules explored in previous generations. To avoid repeatedly exploring the same region of chemical space, individuals are only allowed to become parents for a certain number of generations after they have been first selected. This is controlled by the *AGED_select* and *AGED_random* parameters that apply to individuals chosen by similarity to a previous compound or by random selection, respectively. Additionally, a number (*PARENTS_random*) of individuals coming from random selection are also chosen as parents, to explore promising lines started by random selection, thus allowing a separated short evolution.

For each parent, a fixed number (*CHILD_select*) of children are selected or “adopted”. The process is as follows: First, molecules in the database are ranked by similarity to the parent, using LINGOsim. A number (*PROBchild*) of top individuals in the list that have not been previously selected become eligible to be selected as children. The final choice is made from this shortlist on the basis of a pseudo-random number generator biased toward the selection of higher-ranked compounds.

At each generation, a number (*CHILD_random*) of individuals are randomly selected directly from the complete database. This ensures a continuous diversity input analogous to that introduced by mutations in biological evolution. The total number of molecules is fixed by the parameter *POPULATION*. The main parameters that control the MPA algorithm are summarized in Table 1.

5. Parameter Optimization. The performance of MPA is a complex function of many different parameters and was optimized by using a classical genetic algorithm. To this end, each parameter was considered a gene that could take values within reasonable predefined ranges (see the Supporting Information). Each particular implementation of MPA, with certain values for each of its parameters, was considered an individual. A population of 30 individuals (different MPA variants) was allowed to evolve for 24 generations. Scoring was done by counting the number of molecules with an energy of binding to a test protein below a given threshold that were present in a selection of 6000 molecules out of a database of 60 000 with a given variant of MPA. Previously, an exhaustive search of the complete database (data not shown) gave the total number of hits present (180). The best parameter set for MPA identified 80 out of the 180 hits (45%) after exploring just 10% of the database. This represents more than a 4.4-fold enrichment with respect to the starting parameter set. Although the choice of the parameters may be problem-dependent up to a certain extent, we decided to test this parameter set in the exploration of a large database against a protein not used for its optimization. This should provide an estimation of the robustness of the optimized parameters. The chosen protein was the serine protease Factor Xa. The optimized MPA parameters and an outline of MPA are presented in Figure 1.

6. MPA Screening of FactorXa. To test its performance and efficiency, MPA was tested by applying it to the filtered database containing 913 763 compounds. Factor Xa (PDB code: 1EZQ),[21] an enzyme involved in the coagulation pathways and of great interest for the development of new antithrombotics, was selected as the target.[22] The maximum number of compounds to be docked was restricted to 60 000 (6.5% of the database). For comparison reasons, the same number of compounds were randomly selected and independently docked and analyzed. In both cases, compounds with binding energies equal to or lower than -10.5 kcal mol^{-1} were taken as “*hits*”. Figure 2 shows the evolution of the average binding energy of the population, the best 10% and the worst 10%, as well as the binding energies of individual molecules that are below -10.5 kcal mol^{-1}. A comparison between a random search and a MPA search shows that the average energies are around 1.0 kcal mol^{-1} lower for MPA. We can also compare the average binding energy of the best 10% of the ligands in the populations sampled using the two protocols. The lowest MPA average is nearly 2.0 kcal mol^{-1} lower than the best one obtained with a random search. Finally, if we consider binding


a)
Initial Population
3D Structures Generation
Docking / Scoring
Analysis / Ranking
Selection of New Population
While number of compounds tested less than maximum.
Maximum number of compounds reached.
Final General Analysis
b)

Parameter	Value
POPULATION	130
PARENTS_select	20
PARENTS_random	6
CHILD_select	4
CHILD_random	26
AGED_select	8
AGED_random	4
PROBchild	3000

Figure 1. (a) MPA general scheme. This iterative process is repeated until the maximum number of compounds has been docked. (b) Final parameters for MPA optimized by a genetic algorithm.

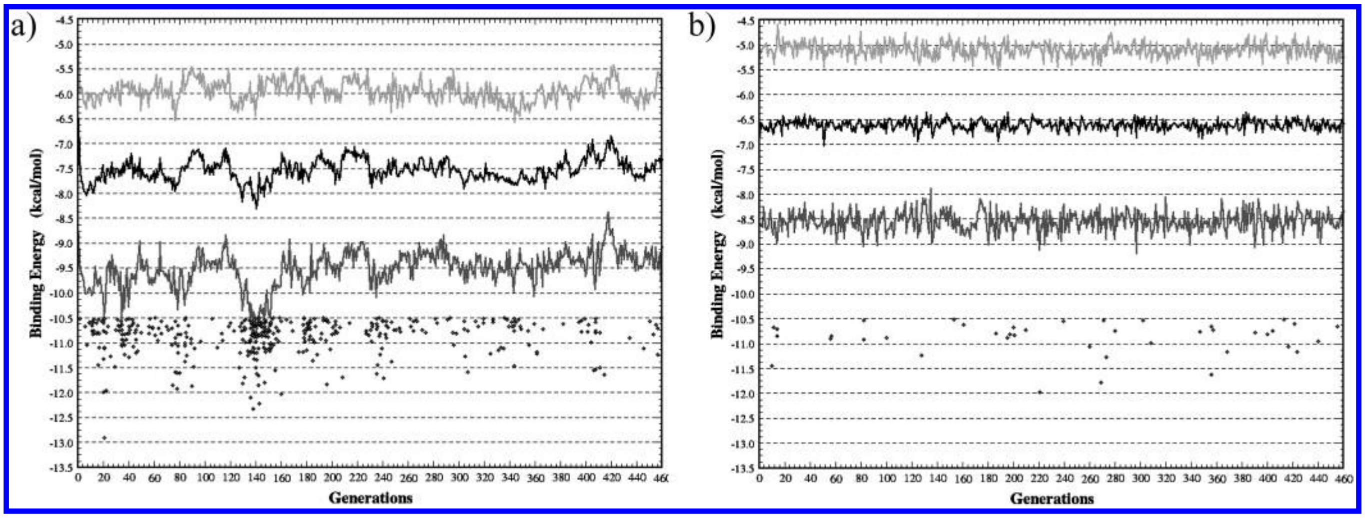


Figure 2. Analysis of populations generated by two different search protocols: (a) MPA and (b) random selection. Average binding energies (black), the best 10% (dark gray), the worst 10% (light gray), and individuals with binding energies below −10.5 kcal/mol defined as “hits” (dots).

energies lower than −10.5 kcal mol^{-1} as “hits”, 392 hits were found by using MPA while only 41 of them were identified by random selection of the database. This is nearly a 10-fold enrichment when using MPA (Figure 2).

The performance of MPA can also be assessed by comparing the frequency distributions of calculated binding energies of the two sets of ca. 60 000 molecules selected either using MPA or randomly. Both distributions are normal with mean values and standard deviations of −7.29 ± 1.30 kcal mol^{-1} and −6.40 ± 1.23 kcal mol^{-1}, respectively.

An analysis of variance considering one factor shows that the difference of the means is statistically significant to a level better than 99%. Alternatively, one can analyze the proportion of hits in the two populations, and the difference is significant with an error lower than 0.1%. A plot of the distributions is presented in Figure 3.

The 913 763-compound library contained 258 compounds from a focused library for FactorXa (Inte:ligand).[23] A total of 196 (76%) of them were found among the 60 000 compounds selected by MPA, while only 20 were present in the random selection. This represents, again, a nearly 10-fold enrichment in the compounds present in the focused library.

Table 2. Comparison of CPU Times Used by MPA and Random Screenings

	MPA	random
computation time (per generation)		
docking	50 min	50 min
3D structures generation	2 min	2 min
selection of new population	9 min	
equal number of compounds tested		
time (days)/compounds	15.3/60 000	13.1/60 000
hits found	392	41
enrichment	9.6	
equal computational time invested		
time (days)/compounds	13.1/51 300	13.1/60 000
hits found	368	41
enrichment	9.0	

From the results of random selection, the number of “*hits*” in the whole library is estimated to be 630. Therefore, a MPA search of only 6.5% of database (60 000 out of 913 763 compounds) provided 62% of the possible “*hits*” (392 of 630).

The average CPU times of the different tasks involved in MPA are presented in Table 2. To equalize CPU usage, the number of docked molecules in MPA has to be reduced. In

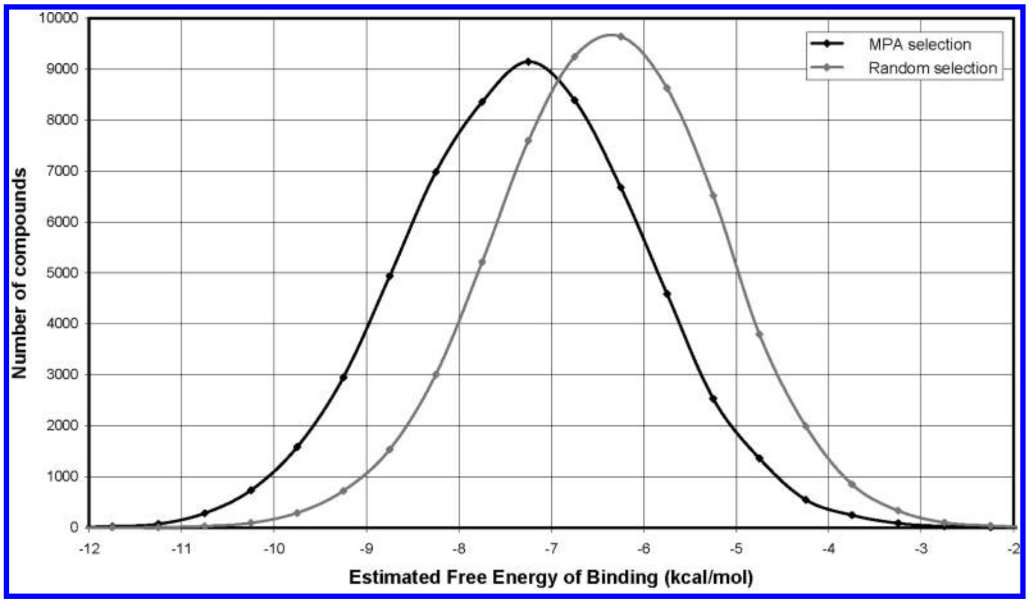


Figure 3. Histogram of calculated binding energies of two populations of 60 000 compounds extracted from a database of 913 763 compounds using MPA or a random choice. Points mark the population and the center of bins with a step size of 0.5 kcal mol^{-1}.

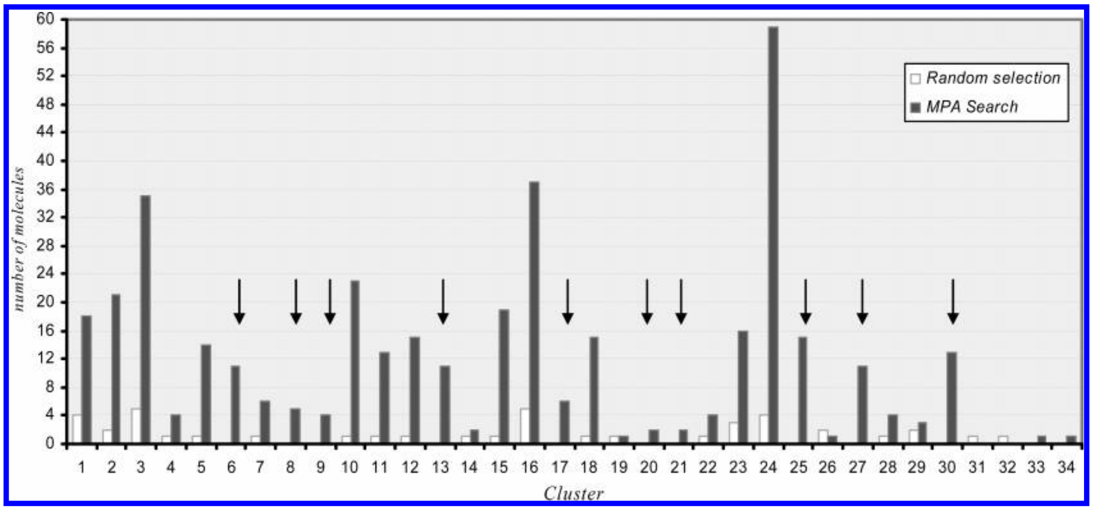


Figure 4. Cluster analysis of hits found using MPA and random selection. Black arrows denote clusters with more than one molecule that were identified only by MPA.

this case, the coverage of the database and the proportion of hits with respect to the estimated total are reduced to 5.6% and 58%, respectively, in MPA, but a higher than 10-fold enrichment is achieved. Compared with a random search, the enhancement factor of MPA is 9-fold greater.

The selection algorithm uses only 15% of the total time, a remarkably low overhead, considering that it includes, for each generation, the calculation of the similarity matrix of 26 parents to the complete database of nearly 1 million compounds. This is only feasible employing LINGOsim to compute similarities directly from SMILES representations.

The handling of SMILES strings instead of 3D structures until the docking stage represents a large savings in computational resources in itself: converting the 913 763-compound database would have required 9.77 days of CPU time, while the time spent to convert the 60 000 compounds to be docked was only 0.5 days. Although converted 3D structures could be reused in other searches of the same database, databases of commercially available compounds are being continuously updated. Obviously, MPA allows the dynamic update of the database during runtime. An additional savings in resources comes from the initial filters that reduced by more than 50% the size of the raw database of ca. 2 million compounds and eliminated compounds that would not have been acceptable solutions from a pharmacological point of view. Accurate property estimations for 2 million compounds made with LINGO-QSPR require only around 16 s.

Diversity Analysis. It may be argued that the use of similarity criteria in virtual screening has the risk of restricting the search to a limited number of scaffolds. The diversity of the hits obtained by random search and MPA were compared using a hierarchical clustering algorithm. Results of the two searches were combined providing a nonredundant list of 414 molecules. Cross similarities were calculated using LINGOsim. A graphical representation of the similarity matrix is presented as Supporting Information. Analysis with the program Cluster 3.0[24] identified 34 different clusters. Figure 4 shows the number of molecules in each cluster found by MPA and a random search. MPA identified representatives in 32 clusters (94%), as compared with 22 clusters (65%) found in a random search. A total of

12 clusters were identified only in the MPA search, while only two, single molecule, clusters were found in the random search and not by MPA. The three clusters that are more populated in MPA and random searches are the same. These results strongly suggest that MPA faithfully samples the variability present in the database and displays an excellent scaffold hopping behavior.

The presence of several representatives of each cluster increases the probability of finding the most active compounds experimentally. Moreover, the resulting target-focused library is information-rich, and the structural redundancy will serve to gain insight into structure–activity relationships.

The virtual search of FactorXa ligands affords the expected bezamidine-based compounds as well as a number of different chemical structures. Their biological significance is still being evaluated, but it is irrelevant in the present context since it would only reflect the accuracy of the particular target structure and docking protocol used, while MPA is a general search strategy applicable to any docking protocol and scoring function.

The generality of these results is supported by the fact that the MPA parameters used in the search of ligands for FactorXa had been optimized in a completely different system. In addition, work in progress in our laboratory includes studies on two different proteins (a SH2 domain and a low molecular weight tyrosine phosphatase), and the computational results are being tested by experimental methods (NMR, surface plasmon resonance, or enzyme inhibition assays). In the first example, 41 molecules have been selected from a database of nearly 1 million compounds, and we have already identified eight compounds that bind better than the peptide with the native sequence. Our preliminary results in the phosphatase example also show promising results: 20 representative molecules of 24 clusters have been selected from the same large database, and four of them, belonging to four different clusters, have K_I values in the range 50–800 μM. The complete studies will be presented in a forthcoming article.

CONCLUDING REMARKS

The size of chemical databases grows steadily and requires the introduction of new algorithms to efficiently handle these resources in the search for bioactive compounds. Recent efforts seek the involvement of the academic community in the initial stages of drug discovery by facilitating access to large public databases of potential ligands. In this work, we have introduced MPA, a highly efficient virtual search protocol that, with very limited computational resources, is able to handle very large databases using a combination of docking and similarity searches. The protocol makes extensive use of our recently introduced LINGO tools, which allow the extraction of the implicit chemical information present in SMILES representations. The performance is demonstrated with a search of a database of nearly 1 million compounds using a scoring scheme based on Autodock 3.0, but the protocol could be easily adapted to different scoring methods, including experimental data from biological assays. MPA has been tested in several proteins, and the results have been checked experimentally in some cases. MPA achieves a nearly 10-fold better performance than a random search using the same overall computational effort.

METHODS

Virtual Database Preparation. Several virtual compound databases, SD format, were acquired from different commercial suppliers (Peakdale, Bionet, IFLab, Maybridge, Timtec, IBScreen, HTS compounds, Pharmeks, Specs, Akos, Chemical Block, ChemDiv, ChemStar, MedChemLabs, and TOSLab) and converted into SMILES by using the *standardizer* module of the *JChem* software. Compounds containing elements other than C, N, S, O, H, P, and halides were not included in the final SMILES database.

Computational Resources. All calculations were carried out on one Pentium 4 computer running at 2.8 GHz and with 512 MB of RAM.

LINGO-Based Tools. The extraction of chemical information from SMILES strings has been reported elsewhere.[16] LINGO is the name given to substrings, usually of four characters, that can be derived from a SMILES representation. Similarities between SMILES strings (LINGOsim) are computed from the set of LINGOs that are derived from each of the two strings. LINGOsim shows good discrimination between bioisoster and random pairs. LINGO-based property predictions are based on models derived using partial-least-squares methods from appropriate training sets. Predictions of log *P* and log *S* are comparable to the best available methods, but they are 2 orders of magnitude faster.

ACKNOWLEDGMENT

This work was partially supported by funds from the Spanish Ministerio de Educación y Ciencia to M.P. (BIO2004-5436 and GEN2003-20642-C09-04). We thank Dr. Yolanda Pérez and Mr. Jascha Blobel for their help in the screenings of the SH2 domain and tyrosine phosphatase. D.V. is a recipient of a predoctoral fellowship from the Spanish Ministerio de Ciencia y Tecnología.

Supporting Information Available: A description of LINGO and LINGOsim and a table of value ranges of the variables included in the GA process. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Seifert, M. H. J.; Wolf, K.; Vitt, D. Virtual high-throughput in silico screening. *BIOSILICO* **2003**, *1*, 143–149.

(2) Branson, K. M.; Smith, B. J. The role of virtual screening in computer aided structure-based drug design. *Aust. J. Chem.* **2004**, *57*, 1029–1037.

(3) Bleicher, K. H.; Böhm, H. J.; Müller, K.; Alanine, A. I. Hit and lead generation: Beyond high-throughput screening. *Nat. Rev. Drug Discovery* **2003**, *2*, 369–378.

(4) Mestres, J. Virtual screening: a real screening complement to high-throughput screening. *Biochem. Soc. Trans.* **2002**, *30*, 797–799.

(5) Butina, D.; Segall, M. D.; Frankcombe, K. Predicting ADME properties in silico: Methods and models. *Drug Discovery Today* **2002**, *7*, S83–88.

(6) Barbosa, F.; Horvath, D. Molecular similarity and property similarity. *Curr. Top. Med. Chem.* **2004**, *4*, 589–600.

(7) Shanmugasundaram, V.; Maggiora, G. M.; Lajiness, M. S. Hit-Directed Nearest-Neighbor Searching. *J. Med. Chem.* **2005**, *48*, 240–248.

(8) Bender, A.; Glen, R. C. Molecular similarity: A key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204–3218.

(9) Kitchen, D. B.; Decornez, H.; Furr, J.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nat. Rev. Drug Discovery* **2004**, *3*, 935–949.

(10) Lengauer, T.; Lemmen, C.; Rarey, M.; Zimmermann, M. Novel technologies for virtual screening. *Drug Discovery Today* **2004**, *1*, 27−34.

(11) Warmuth, M. K.; Liao, J.; Rätsch, G.; Mathieson, M.; Putta, S.; Lemmen, C. Active Learning with support vector machines in the drug discovery process. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 667−673.

(12) Weininger, D. SMILES: a Chemical Language and Information System. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(13) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.

(14) Jones, G. Genetic and Evolutionary Algorithms. *Encyclopedia of Computational Chemistry*; Wiley: Chichester, U. K., 1998.

(15) 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.

(16) Vidal, D.; Thormann, M.; Pons M. LINGO, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities. *J. Chem. Inf. Model.* **2005**, *45*, 386−393.

(17) ChemAxon Ltd., Budapest, Hungary. http://www.chemaxon.com (accessed Mar 2005).

(18) Gerber, P. R. MOLOC − A Molecular Design Software Suite. http://www.moloc.ch (accessed Jan 2005).

(19) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639−1662.

(20) Thormann, M.; Pons, M. Massive docking of flexible ligands using environmental niches in parallelized genetic algorithms. *J. Comput. Chem.* **2001**, *22*, 1971−1982.

(21) Davie, E. W.; Fujikawa, K.; Kisiel W. The coagulation cascade: Initiation, maintenance, and regulation. *Biochemistry* **1991**, *30*, 10363−10370.

(22) Maignan, S.; Guilloteau, J. P.; Pouzieux, S.; Choi-Sledeski, Y. M.; Becker, M. R.; Klein, S. I.; Ewing, W. R.; Pauls, H. W.; Spada, A. P.; Mikol, V. Crystal structures of human Factor Xa complexed with potent inhibitors. *J. Med. Chem.* **2000**, *43*, 3226−3232.

(23) inte:Ligand. http://www.inteligand.com/ilibdiverse/samplelibs.shtml (accessed Aug 2005).

(24) Cluster 3.0, command line version. http://bonsai.ims.u-tokyo.ac.jp/∼mdehoon/software/cluster (accessed Jul 2005).

CI050458Q