Journal of Computer-Aided Molecular Design, **12:** 597–604, 1998.
KLUWER/ESCOM

CombiDOCK: Structure-based combinatorial docking and library design

Y. Sun*, T.J.A. Ewing, A.G. Skillman & I.D. Kuntz**
Department of Pharmaceutical Chemistry, University of California, San Francisco, CA 94143-0446, U.S.A.

Received 5 March 1998; Accepted 24 April 1998

Key words: combinatorial library design, molecular docking, structure-based drug design

Summary

We have developed a strategy for efficiently docking a large combinatorial library into a target receptor. For each scaffold orientation, all potential fragments are attached to the scaffold, their interactions with the receptor are individually scored and factorial combinations of fragments are constructed. To test its effectiveness, this approach is compared to two simple control algorithms. Our method is more efficient than the controls at selecting best scoring molecules and at selecting fragments for the construction of an exhaustive combinatorial library. We also carried out a retrospective analysis of the experimental results of a $10 \times 10 \times 10$ exhaustive combinatorial library. An enrichment factor of approximately 4 was found for identifying the compounds in the library that are active at 330 nM.

Introduction

One of the most exciting new developments in medicinal chemistry in recent years is combinatorial chemistry [1]. The modular display of functional groups allows a large number of compounds to be considered for synthesis. Coupled with automation technologies and high-throughput screening, combinatorial chemistry offers great potential for the discovery of drug leads. Nonetheless, even though billions of compounds can be proposed, it remains difficult to validate and assay such numbers of compounds. Typically, unless the library is based on oligomeric units, only very small subsets of fragments are selected for actual synthesis, in a process known as combinatorial library design. Thus, a critical challenge for computational chemistry is to select sets of fragments that have the best potential for the discovery of new leads for a given target.

The structure-based drug design method utilizes the information contained in receptor structures by analyzing how well potential lead compounds might bind to the receptors. This includes both the structure-based database screening and de novo ligand designs [2]. Since the number of protein structures available, computational methodologies, and computer resources are all improving at a rapid pace, it is inevitable that using the information of target structures in drug design will become increasingly important. A recent pioneer study of combining structure-based design and combinatorial chemistry yielded encouraging results [3]. In that study, calculations were based on fixed scaffold orientations and fragments were scored independently for each attachment site. Fixed scaffold orientations were reasonable in that case because experimental evidence supported limited orientational and conformational freedom for the scaffold. However, to be generally applicable to combinatorial library design, the structure-based design method has to be able to take into account the inter-dependency of fragments at different binding sites, without prior knowledge of the scaffold orientation. This inter-dependency requires one to deal with the large number of combinations produced by combinatorial chemistry. If all the combinations are created and examined individually, as in the traditional database screening approach, then

*Present address: Computer-Assisted Drug Design, Bristol-Myers Squibb Company, 5 Research Parkway, Wallingford, CT 06492, U.S.A.
**To whom correspondence should be addressed.

millions, even billions, of compounds will have to be screened. Such numbers are far beyond present-day computational resources. In this work we report a method that could be used to carry out efficient docking calculations for such large virtual combinatorial libraries.

In the second part of our study, we will use the combinatorial docking method to analyze the experimental results obtained in the previous study [3]. In that study, 1000 molecules from a $10 \times 10 \times 10$ exhaustive library were synthesized on solid support and assayed individually. These experimental data offer us a unique opportunity to test objectively how well our computational methods actually perform. We think this type of direct comparison over a large number of compounds will have wide implications for future work in the development of better scoring functions and in the design of experiments.

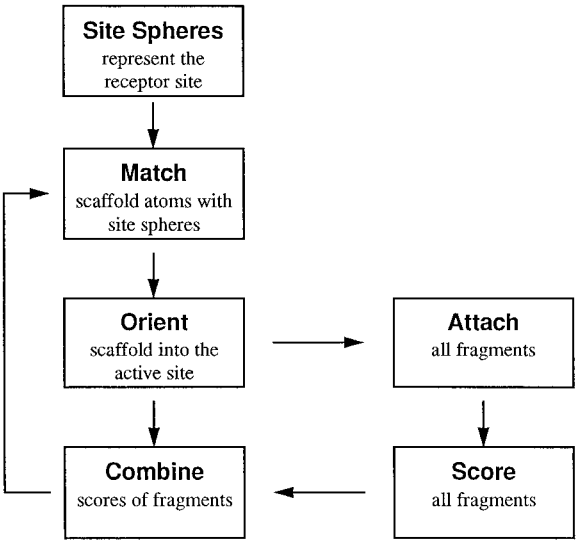

Figure 1. CombiDOCK algorithm.

Computational method

Dock

The basic DOCK algorithm has been described in detail elsewhere [4, 5]. Four steps are needed to carry out the calculation: (1) the negative image of the receptor active site is represented by a set of spheres; (2) internal distance matches between a subset of spheres and a subset of ligand atoms are searched; (3) for every match, the ligand is juxtapositioned onto the active site; (4) a score is calculated for the ligand in that orientation. For a single compound in a typical database screen against an enzyme target, the program might find 10 000 matches and generate an orientation for each. Approximately, 1000 of these orientations that do not bump into the receptor are finally scored, using force field or empirical functions to approximate interaction energies.

CombiDOCK

The combinatorial docking strategy is a simple variation of the basic DOCK algorithm (Figures 1 and 2). The site sphere generation is unchanged as step 1. For matching, only scaffold atoms are used instead of the entire ligand. At steps 3 and 4, once a scaffold is matched onto the active site, all fragments are attached individually for each site position, searching through multiple connecting torsions. Interaction scores are calculated for the scaffold and each attached fragment. As a final step, combinations of fragments are made and the best combinations are then checked for intramolecular clashes and saved if no clashes are found. It should be noted that this kind of fragment superposition algorithm has been tried previously for non-combinatorial problems, such as directed database searching and conformational searching [6].

Although all combinations of fragments are, in theory, examined, the strength of this method is that the combinatorial process is reduced to the simple numerical additions of the fragment scores at all sites. It is thus possible to use simple numerical techniques to speed up the combinatorial process. Specifically, after scoring all fragments at each scaffold orientation, the fragments are sorted according to their scores and the combination process can be terminated once it is determined that no combinations better than a user-defined limit can be found. In addition, the internal clash checks, which are computationally expensive, are only necessary for combinations that have good enough scores to be eventually saved.

Test cases

Part I. Combinatorial docking and library design

For our first test of the algorithm, we chose to dock a virtual library of benzodiazepine derivatives [7] to dihydrofolate reductase (DHFR). We selected the benzodiazepine library partly due to its historic role in combinatorial chemistry as one of the first non-oligomeric combinatorial libraries (Figure 3). 1,4-

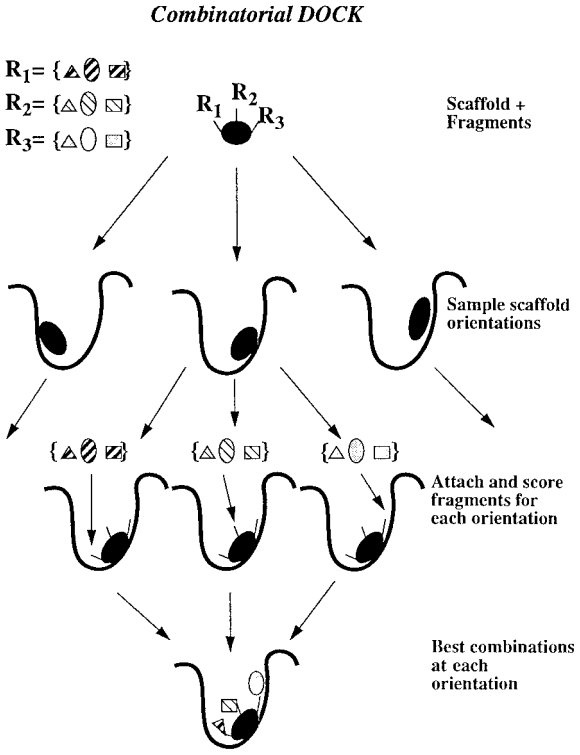

Figure 2. CombiDOCK illustration.

Benzodiazepine derivatives have been shown to have a wide range of bioactivities [8]. Because we do not have the crystallographic structures for the natural benzodiazepine receptors, we have chosen dihydrofolate reductase as the target for the benzodiazepine library. Since the main purpose of the study is to test the feasibility and efficiency of the combinatorial docking methodology, DHFR is a good target because of its large and deep binding pocket. This binding site provides an excellent test of the inter-dependency among fragments because the resulting 'combined' molecules must fit properly into the pocket. This point will be discussed below.

We used the Available Chemical Directory (ACD95.1) from MDL Information Systems (San Leandro, CA) and found 308 acid chlorides (R_1), 305 amino acids (R_2) and 404 alkylating agents (R_3) that satisfy the synthetic requirements for building the virtual library at the three attachment sites. The total number of all potential combinations is about 36 million ($308 \times 305 \times 404$). A newly developed program, Diversify [9], was used to prepare the fragments. Diversify incorporates routines from the Daylight Toolkit [10] to remove the leaving atoms on the fragments and to add tags identifying atoms connecting to the scaffold. The Rubicon program [10] was used to generate one three-dimensional conformation for each fragment and the results were saved as mol2 files [11], with the connecting atom information stored in the @<TRIPOS>SET field. Similarly, the scaffold, 1,4-benzodiazepine, was built and the connecting atom information was also identified. Complete molecules with proper chemical bonds between the scaffold and the fragments were constructed during the docking calculation.

The combiDOCK is adapted from a new version of DOCK, v. 4.0 [5]. The only new parameter required is the number of uniform torsional positions to be sampled for the connecting bond between the scaffold and each fragment. We searched six torsional positions in our tests. The regular single point DOCK force field scoring method was used with one modification. A positive score (penalty) of 0.5 kcal/mol was added for all non-hydrogen atoms of ligands. This modification was made to avoid the largest fragments always having the best scores. Potent yet small compounds are generally more desirable as leads.

As controls, two other methods were also tested: (1) random selection: fragments were randomly selected from all available candidates; (2) single fragment docking: in this strategy, fragments at different sites were assumed to be independent. Each fragment was attached to the benzodiazepine scaffold by itself and the resulting mono-substituted molecule was docked. The best scoring fragments for each site were then selected as the best candidates for the combinations.

Two steps are needed in docking and designing combinatorial fragment libraries. The first step is to find the best scoring compounds made from all possible combinations of potential fragments. If compounds are to be synthesized individually, no more library design is needed. A completely exhaustive combination approach, i.e. making all possible combinations from selected fragments at each site, however, is a more efficient experimental design for making an equal number of compounds. If exhaustive combination is desired, then fragments at each site have to be selected based on the results from the first step. We will show the results obtained at both steps.

At step one, i.e., finding the best scoring single molecules, the constraint of using similar amounts of computer time meant that only 20 fragments could be used for each site in the random selection method and single fragment method (still producing 8000

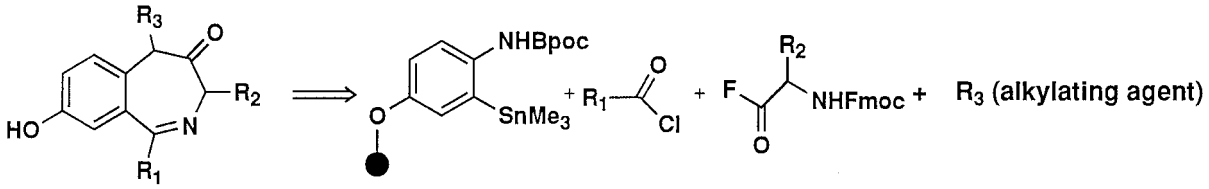

Figure 3. 1,4-Benzodiazepine combinatorial library.

combinations!). This time constraint also limited that only one or two conformations per molecule could be examined. The conformations were generated by randomly assigning the torsional angle connecting a fragment and the scaffold. To observe the dependency of the searching results on the number of conformations used, calculations were done for both one conformation per molecule and two conformations per molecule. Whenever fragments were added to the scaffold, intramolecular clashes were checked and molecules with internal clashes were removed, typically about 10% of all combinations.

At step two, i.e., constructing an exhaustive combinatorial library, the following procedures were used for the selection of fragments: (1) combinatorial docking: fragments were ranked and selected according to the frequencies they appeared in the top 1000 scoring combinations; (2) random selection: 10 fragments were selected randomly; (3) single fragment: fragments were ranked and selected according to docking scores of the mono-substituted compounds, i.e., compounds with one fragment attached to the scaffold.

Part II. Retrospective analysis of the experimental results of a combinatorial library

Structure-based library design has been used to design fragment libraries for a hydroxyethylamine scaffold [12] (figure 4) targeting cathepsin D, an aspartyl protease. There are three fragment attachment sites on the scaffold. In the previous study, 10 fragments were chosen for each site and incorporated in the final combinatorial synthesis [3]. The resulting 1000 molecules were assayed for activity at 1 μM, 330 nM, and 100 nM, with 67, 23, and 7 compounds having inhibition greater than 50% at each concentration respectively.

Our goal in this work is to analyze in more detail the experimental results for the compounds synthesized and assayed. This is a test for both the searching algorithm and the scoring function. Since only 10 fragments were finally used at each site, fragment conformations can be sampled more extensively than in the initial design process. A systematic dihedral searching method was used to generate fragment conformations. For torsions with rotational barriers below 2 kcal/mol, according to the AMBER force field [13], dihedral angles were sampled every 60°. When a double bond was involved, then only the trans and cis forms were used. The conformational searches generated a total of 282, 152, 225 molecular conformations for the 10 fragments at each site. We used the same scaffold conformation from the previous work [3], which was determined by matching the scaffold with the crystal structure of pepstatin in the complex with cathepsin D (1LYB) and torsional searching for the three undetermined dihedral angles [3]. All calculations were performed on SGI Indigo2 with R4000 CPU and 128 MB of memory.

Results and discussion

I. Combinatorial docking of the benzodiazepine library to DHFR

As mentioned, the first step is to find the best scoring compounds from combinations of all potential fragments. The distribution of scores for the top 500 scoring molecules found with each method, together with the CPU time used to search for them, are shown in Figure 5. Searching was limited, as described in the method section, for the random selection method and the single fragment method so that each approach was given roughly the same computer CPU time as the combinatorial docking. The average scores of the top 500 scoring compounds are -25.6, -18.1 and -15.7 for combinatorial docking, random selection, and single fragment methods, respectively. It is interesting that selecting compounds based on one fragment at a time (single fragment method) is even worse than a random selection. The reason for this is that the single fragment method assumes independence between fragments, and it picked out similar fragments at all three positions that dock very well into the binding pocket when studied as the mono-substituted scaffold. Once these fragments are put together in the

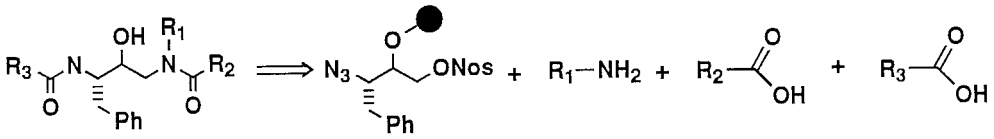

Figure 4. Hydroxyethylamine library.

same molecule, however, they interfere with one another. This often results in either inefficient docking, in which fragments interact with the target weakly, or, worse yet, one fragment bumps into the target and the combination must be discarded. Clearly, the combinatorial algorithm provides a much better chance of finding the best scoring molecules: the distribution of scores does not even overlap with the results from the control methods.

Having found the best scoring individual compounds, we next considered the design of an exhaustive combinatorial library. Here the goal is to select a small arbitrary number of fragments from all available fragments for each site to prepare the best library where the combinations are exhaustively made. In our test, we selected 10 fragments for each site yielding a $10 \times 10 \times 10$ format for a 1000-compound library.

Using the fragment selection method described previously, three $10 \times 10 \times 10$ libraries were constructed based on the results of the combinatorial docking and the two control algorithms. To compare these three libraries of $10 \times 10 \times 10$ molecules, 25 random conformations were generated for each combination, again by randomly assigning connecting torsions. It should be mentioned that even with 25 conformations docked for each molecule, the conformational search is still quite limited. Conformations that had internal clashes were discarded. For each molecule, i.e., each combination of fragments, the conformation with the best docking score was saved as the final score for the molecule. The distributions of the scores are shown in Figure 6. As in the first step (Figure 5), combinatorial dock performed best, and random selection is better than the single fragment approach. The average scores for the three libraries are -18.9, -11.2, and -6.7. However, there is now much more overlap among the docking method and the two control algorithms. The primary reason for this is that exhaustive combinations force the inclusion of many not-so-good combinations.

We caution, however, that this result does not prove that synthesizing the individual best scoring combinations is a better strategy than the exhaustive combinatorial approach. First, we do not believe that

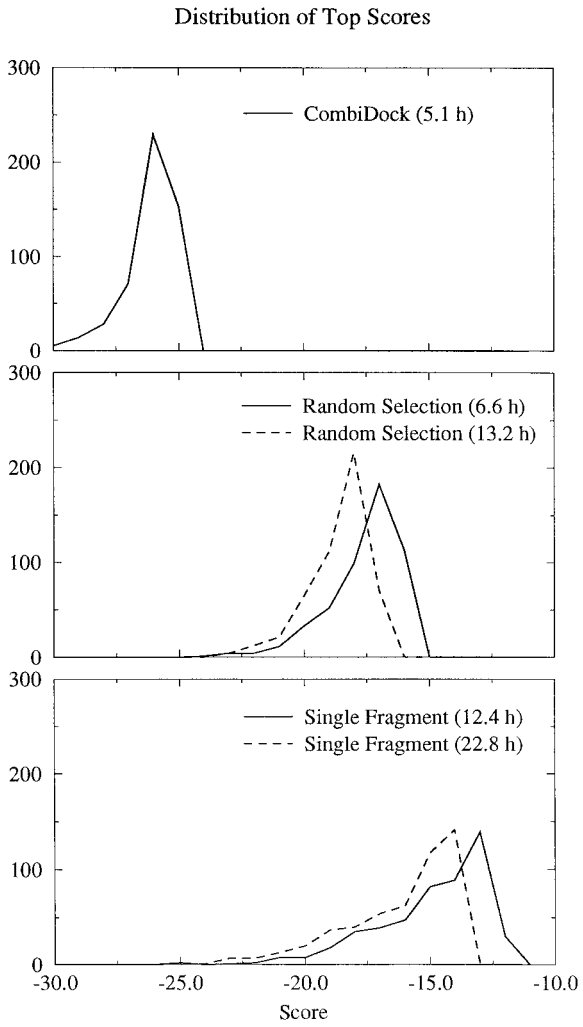

Figure 5. 1,4-Benzodiazepine combinatorial library. Distributions of the top 500 scoring molecules from three different searching methods. For the random selection method and the single fragment method, only one conformation per combined molecule was generated for the short runs (solid lines), and two conformations per combined molecule were generated for the long runs (dashed lines).

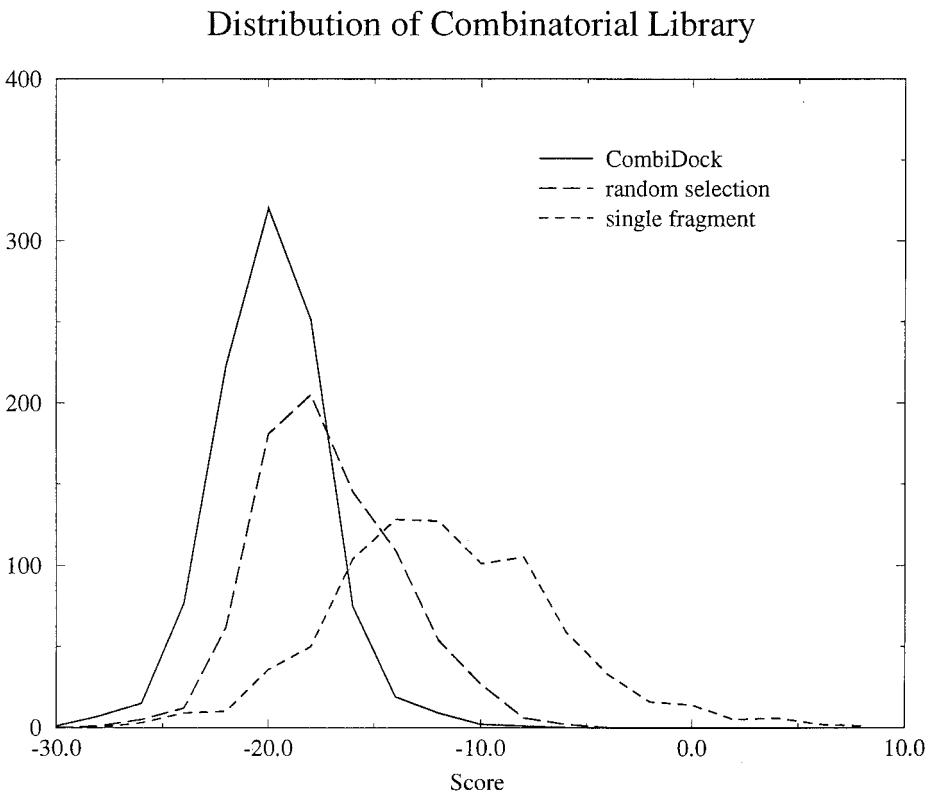

Figure 6. Distributions of the scores of the designed exhaustive combinatorial libraries.

the current scoring functions are reliable enough to unambiguously identify the best binding ligands. Second, we recognize the substantial increase in synthetic effort involved in making the individual best scoring compounds. In our test, we have attempted to separate the searching algorithm from the scoring function to demonstrate the efficiency of searching for a given scoring function. The reality, however, is that the quality of scoring functions is critical to the quality of predictions. The quality of predictions will in turn influence how the actual experiments should be optimally designed. It is interesting to note that the score for methotrexate, a potent inhibitor of DHFR, is only -24, while the best scoring compound from the library is -30 and more than 200 compounds from the library scored better than -24. Since this is unlikely to be true, it reveals the deficiency of the scoring method. Of course, on the other hand, the scoring method did rank methotrexate at about 200th among the total of 36 million combinatorial compounds of the library. It shows that the method of docking and scoring certainly has the ability to enrich the compound selection from a large molecular database or combinatorial library.

The scaffold-based combinatorial docking method presented here is suited for problems where the target structure is known but the ligand position is uncertain. Under these circumstances, the ligand orientations can be explored using docking procedures. However, the scaffold cannot be too flexible. If there are only a few low-energy conformations available for the scaffold, then these conformations could be treated independently and results from different conformations can be combined at the end. On the other hand, it would be difficult if the scaffold has many low-energy conformations, unless the bound conformation or the bound geometry of the backbone is restricted or known. This was the case for our earlier work using a hydroxyethylamine-based library targeting cathepsin D [3].

II. Retrospective analysis of the experimental results of a combinatorial library

As a further test of the combinatorial dock method and the scoring function we use, we examined the

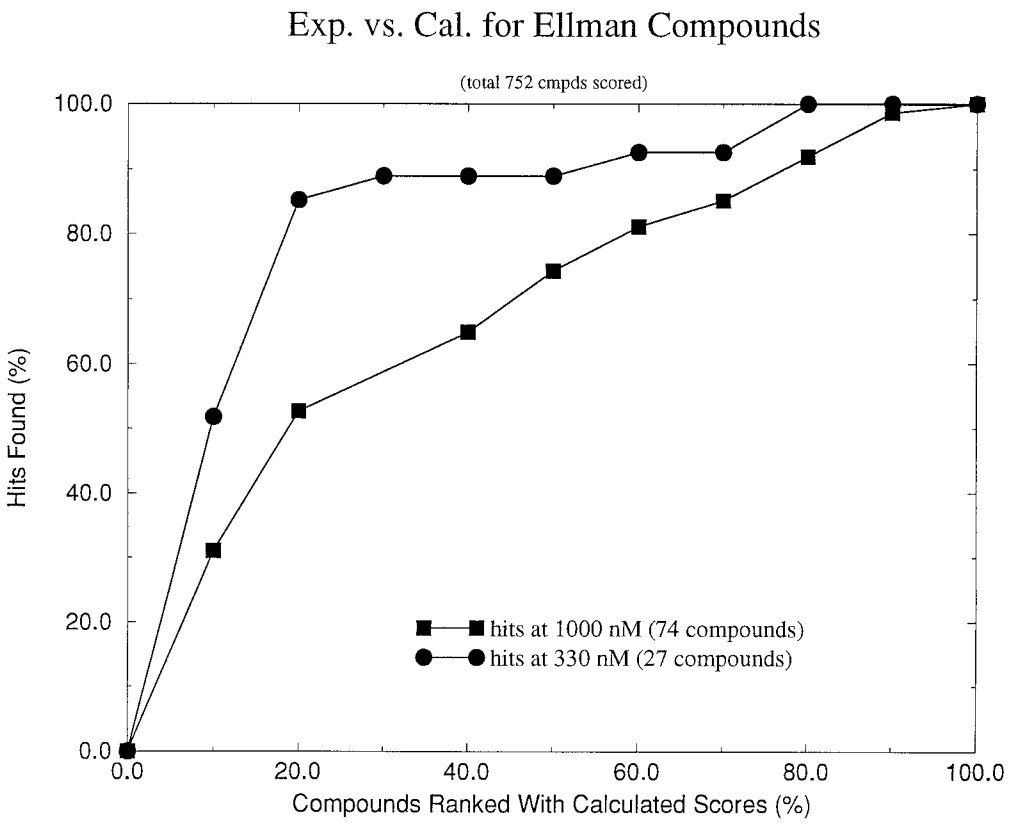

Figure 7. The relationship between the percentage of experimental hits found and the percentage of compounds ranked with calculated scores at selected ranking cutoff for the hydroxyethylamine library. The dashed line represents the expected results from a random ranking.

1000 hydroxyethylamine compounds synthesized and assayed in our earlier work. Because of a difference in the way the conformation searching was carried out, acceptable bound geometries were found for only 75% of the molecules (752/1000). The likely source of this problem is our limited ligand conformational searching and the neglect of receptor flexibility. Since our main goal here is to test how well our calculated scores relate to the experimental results, we decided to use only the 752 compounds that we could readily score. We use the enrichment factor (defined as the ratio of hit-rates from a calculation and from random) as a measure of the quality of the calculation. As shown in Figure 7, when the experimental results at 1 μM were used, the enrichment factor is about 2.5. When the data at 330 nM were used, however, the enrichment factor increases to about 4. A completely random ranking would result in an enrichment factor of 1. This calculation suggests that the present scoring function has utility in selecting the more potent ligands in spite of its many approximations [14].

Conclusions

We have implemented and tested a combinatorial docking strategy. We have shown that it is able to find better scoring combinatorial molecules than two control methods. When completely exhaustive combinations are required, fragments selected based on the results from the combinatorial docking also produced better scoring compounds. The combinatorial docking method is fast enough to allow using structure-based library design for general combinatorial chemistry problems when target structures are available. We have also analyzed the experimental results from a previous combinatorial library. An enrichment factor of 4 was obtained using the force-field-based scoring method.

Acknowledgements

This work was supported by NIH Grants GM39552 (C. Craik, principal investigator) and GM31497 (I.D.K.). We gratefully acknowledge additional sup-

port from Daiichi Corporation to Y.S. and NIH Training Grants to T.J.A.E. and A.G.S.

References

1. Thompson, L.A. and Ellman, J.A., Chem. Rev., 96 (1996) 550.
2. a. Kuntz, I.D., Science, 257 (1992) 1078.
   b. Kuntz, I.D., Meng, E.C. and Shoichet, B.K., Acc. Chem. Res., 27 (1994) 117.
   c. Eisen, M.B., Wiley, D.C., Karplus, M. and Hubbard, R.E., Proteins, 19 (1994) 199.
   d. Caflisch, A., J. Comput-Aided Mol. Design, 10 (1996) 372.
   e. Bohm, H.J., J. Comput.-Aided Mol. Design, 6 (1992) 593.
3. Kick, E.K., Roe, D.C., Skillman, A.G., Liu, G., Ewing, T.J.A., Sun, Y., Kuntz, I.D. and Ellman, J.A., Chem. Biol., 4 (1997) 297.
4. a. Kuntz, I.D., Blaney, J.M., Oatlet, S.J., Langridge, R. and Ferrin, T.E., J. Mol. Biol., 161 (1982) 269.
   b. Shoichet, B.K., Bodian, B.K. and Kuntz, I.D., J. Comput. Chem., 13 (1992) 380.
   c. Meng, E.C., Shoichet, B.K. and Kuntz, I.D., J. Comput. Chem., 13 (1992) 505.
5. Ewing, T.J.A. and Kuntz, I.D., J. Comput. Chem., 18 (1997) 1175.
6. a. Shoichet, B.K., et al., Science, 259 (1993) 1445.
   b. Schoichet, B.K., personal communication.
7. a. Bunin, B.A. and Ellman, J.A., J. Am. Chem. Soc., 114 (1992) 10997.
   b. Bunin, B.A., Plunkett, M.J. and Ellman, J.A., Proc. Natl. Acad. Sci. USA, 91 (1994) 4708.
   c. Plunkett, M.J. and Ellman, J.A., J. Am. Chem. Soc., 117 (1995) 3306.
8. a. Hsu, M.-C., et al., Science, 254 (1991) 1799.
   b. Chambers, M.S., et al., Biomed. Chem. Lett., 3 (1993) 1919.
   c. James, G.L., et al., Science, 260 (1993) 1937.
9. Skillman, A.G. and Kuntz, I.D., to be published.
10. Daylight Chemical Information Systems Inc., Santa Fe, NM.
11. SYBYL v. 6.2, Tripos Inc., St. Louis, MO.
12. Kick, E.K. and Ellman, J.A., J. Med. Chem., 38 (1995) 1427.
13. Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R. and Kollman, P.A., J. Am. Chem. Soc., 117 (1995) 5179.
14. Charifson, P. and Kuntz, I.D., In Charifson P. (Ed.) Practical Applications of Computer-Aided Drug Design, Marcel Dekker, New York, NY, 1997, pp. 1–37.