Comparing Performance of Computational Tools for Combinatorial Library Design

Erik Evensen,[†] John E. Eksterowicz,[‡]
Robert V. Stanton,[§] Connie Oshiro,[||]
Peter D. J. Grootenhuis,[⊥] and Erin K. Bradley*,[†]

*Deltagen Research Labs, 740 Bay Road,
Redwood City, CA 94063 and 4570 Executive Drive, Ste.
400, San Diego, California 92121*

Received November 22, 2002

Abstract: In using computational tools for library design it is necessary to understand the performance and limitations of available methods. This letter reports systematic comparisons of applying ligand-based and structure-based tools across therapeutic project-derived data sets. Included are assessments of performance in real-world iterative design applications and the utility of target structural information. The results suggest that combining screening and target structure information is robust; further, a well-designed screening library can compensate for lacking structural information.

Introduction. The introduction of parallel synthesis and high throughput screening in the medicinal chemistry setting has provided new challenges and opportunities for computational medicinal chemists. Design strategies need to assist chemists in optimizing the information gathered from screening libraries and capitalize on the larger amount of available data. In response to these changing needs the methods and measures by which computational tools are assessed must be updated.

Historically, computational tools may have been used to prioritize, in a single pass, compounds from a large static corporate collection so it was appropriate to assess tool performance in terms of recovery of known actives from a large, diverse pool. With the integration of computational methods in the iterative library design−synthesize−screen−analyze process, the methods are required to discern active compounds from congeneric sets and to capitalize on large amounts of sometimes-noisy screening data. In addition, the methods should assist in discovering novel or alternate chemotypes to provide back-up series in case a particular lead suffers intractable ADMET or PK issues.

This letter describes evaluating computational methods including proprietary in-house pharmacophore-based and structure-based ensemble and informative design tools. Virtual high throughput screening (docking) with a variety of scoring methods was chosen as an external benchmark. The methods were tested using data generated on internal discovery projects and a software framework to simulate using the tools in a project setting. Comparison metrics were chosen to reflect the demands on computational tools in modern drug discovery. These measures are cumulative enrichment, enrichment per round, and number of scaffolds discovered with active compounds. A single measure does not reflect adequately the potential impact of a given method on a therapeutic project; nor does the same metric have the same importance when considering a tool for different stages of a therapeutic project (e.g., lead discovery, evolution, or optimization). The metrics, therefore, must be considered together and combined with the insights of medicinal and computational chemists to understand the balance of project needs and how a particular method can meet those needs.

Data Sets. Two data sets comprising screening data generated for internal discovery projects focusing on different gene families were assembled. Both data sets were processed to remove duplicate compounds. Activities for the compounds were binned according to criteria described for the respective data sets below. To best simulate therapeutic project scenarios, each of the data sets was also split into “starting data” and “targeted source pool” from which compounds were selected.

The first data set was derived from data associated with a project targeted on cyclin-dependent kinase 2 (CDK2).[1,10] It contains 17550 compounds divided into the “starting” or “training” data made up of 13359 compounds from an informative screening library and the “targeted source pool” comprising 4191 molecules; these libraries contained 207 and 161 active compounds, respectively. Compounds with $IC_{50} < 25\ \mu M$ or enzyme inhibition >50% at 30 μM were considered active for this study. Among the target library compounds there were 22 scaffolds, as defined by the original medicinal chemistry team, of these 14 contained active compounds. This data set represents a situation in which data from a screening library is used as the starting point for model building. In addition, 23 structures of CDK2 complexed with a variety of ligands were available; enabling the comparison between using screening data only and using target structure information.

The second data set is from a serine protease inhibitor discovery project. The starting data set contains three lead compounds found in the literature and patents, and the targeted source pool consists of 8098 compounds synthesized in the course of the project. Of the synthetic compounds 352 were considered active, with 14 scaffolds exhibiting activity. The activity cutoffs for this data set were $K_i < 10\ \mu M$ or >50% enzyme inhibition at 25 μM. Three crystallographic structures of the target complexed with inhibitors were available. This data set allows comparison of how well various methods support lead hopping, with and without target structure information.

The distributions of pairwise Tanimoto similarities (1.0 − pairwise Tanimoto distance in Daylight fingerprint space[2]) between the active compounds in the training sets and the source pools for each data set are shown in Figure 1.

Methods. The studies presented form a retrospective analysis, meaning the methods were tested using compounds and data that already existed. This is the most

* To whom correspondence should be addressed.

[†] Present address: Sunesis Pharmaceuticals, 341 Oyster Pt. Blvd, South San Francisco, CA 94080, email: ee: ee@sunesis.com; ekb: ebradley@sunesis.com.

[‡] Present address: Celera Genomics, 180 Kimball Way, South San Francisco, CA, 94080, email: john.eksterowicz@celera.com.

[§] Present address: Pfizer Discovery Technology Center, 620 Memorial Dr., Cambridge, MA 02139, email: robert_stanton@cambridge.pfizer.com.

[||] Present address: Roche Palo Alto, 3431 Hillview Ave, Palo Alto, CA 94304.

[⊥] Present address: Vertex Pharmaceuticals Inc., 11010 Torreyana Rd., San Diego, CA 92121, email: Peter_Grootenhuis@sd.vrtx.com

10.1021/jm025618t CCC: $25.00
Published on Web 10/18/2003

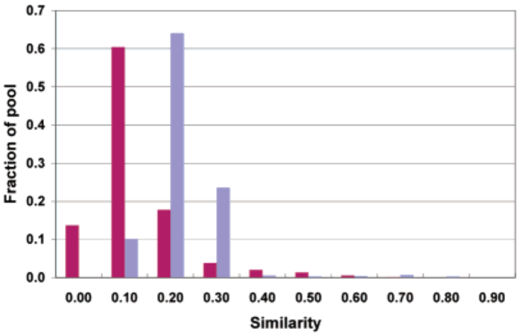

Figure 1. Histogram of distribution of pairwise Tanimoto similarities between active training set compounds and active source pool compounds. Purple bars represent CDK2 compounds and blue bars represent protease compounds.

realistic and consistent manner possible by which to compare the methods; nevertheless, it does present limitations in that the methods are able only to discover previously known compounds.

The performance of two ligand-based and two structure-based methods was compared to random selection. For all the methods except docking, an initial computational model was derived from the "starting data set", this was followed by five simulated iterations of the design−synthesize−screen−analyze cycle within the "targeted source pool".

For the internal ligand- and structure-based methods a full conformational model was generated for all compounds using the program CONAN,[3] and the presence of three-dimensional 2-, 3-, and 4-point pharmacophores with bin-widths ranging from 0.8 Å at short distances (minimum distance 2.2Å) to 3.0 Å at longer distances (maximum distance 24 Å) in any of a compound's conformers was encoded as binary strings or fingerprints.[4,5] Feature combinations could occupy adjacent bins simultaneously if their distance fell within 20% of a bin width of the bin boundary. These descriptors were used as the basis for developing ensemble models using machine learning[4,6,7] or performing informative design[8,9] in pharmacophore spaces derived either from the screening library (CDK2) or known active (protease) data.[10] In addition, informative design was performed in structure-derived pharmacophore spaces.[11] For all internal methods five rounds of compound selection and model or design space refinement were performed. For the CDK2 data set, a total of 1000 compounds were selected in batches of 500, 250, 100, 100, and 50. A total of 2500 protease compounds were selected in batches of 500 compounds per round.

To investigate whether virtual high-throughput screening[12−16] provided advantages in comparison to the proprietary methods, compounds were selected from the source pools by docking them to the available protein structures and evaluating their interaction energy scores using UCSF DOCK.[17] Partial charges were assigned using the Gasteiger method[18,19] and 10 CONAN conformers per stereoisomer of each ligand were docked rigidly to each target structure. The limited number of conformers was chosen to achieve reasonable throughput in the docking calculation in a moderate-sized Linux clusters (approximately 40 dual 1.4 GHz processor nodes); this constraint reflects the computational demands of the docking calculation in comparison to the ligand-based methods which required at most two processors of similar speed. In the case of the CDK2 data set, compounds were docked to only four structures[20−23] because of the large amount of CPU time required per target structure; the protease compounds were docked to all three available structures. The DOCK energy score, DOCK chem score, and DOCK vdw score scoring functions were applied both individually and in combination by taking a consensus of their results. In addition, consensus scores were calculated across the available protein structures.[24−26] Because there is not a straightforward, systematic method for incorporating assay data in the DOCK protocol and scoring functions, a single selection of a number of compounds equal to the number selected in the five rounds for each target was made using each scoring function and combinations thereof. The ranges and averages of enrichment and active scaffolds are reported for each target.

Finally, to establish a baseline for comparison, random selections matching the numbers picked in the experiments of the internal tools were made.

All methods were evaluated according to several parameters chosen to measure quantities important for therapeutic project success. To assess the ability of each method to help discover active compounds, enrichment was calculated by dividing the fraction of selected molecules that were active by the fraction of active compounds in the source pool. The enrichment provided by each method was calculated for the cumulative collection of molecules selected, as well as for the molecules selected in each round. The latter was used to ascertain whether the methods were capitalizing on the data generated.

It is important to note that depending on the source pool size and composition there is a broad range of possible enrichment values; therefore, the maximum possible enrichment is reported for each experiment. This quantity is calculated using the function described above with the assumption that all possible active compounds were selected; for example, if a pool contained 5000 compounds of which 100 were active, the maximum possible enrichment for a selection of 1000 compounds is 5. In contrast, other evaluations of computational methods often involve searching for a small number of actives in a large, diverse collection; typically this might involve searching for 100 active compounds using a 1000 compound selection from a 100000 compound database. The maximum expected enrichment for this type of selection is 100.

Each method was also evaluated for its ability to recover actives on multiple scaffolds. This parameter is important because it gives an indication of how successful a method is at discovering alternate chemotypes or lead hopping. Since compounds can fail for many reasons as they move through optimization and development having multiple backup series is critical.[27] The ability to identify and optimize chemically diverse scaffolds should enhance the chances of finding active compounds that have favorable pharmacokinetic profiles. Thus, the other metric used for comparing the methods is the number of scaffolds containing active compounds each method discovered.

Results. The cumulative enrichment and the number of scaffolds found by each method are summarized in Table 1. The results indicate that all computational

Table 1. Summary of Cumulative Enrichment and Active Scaffolds Recovered by the Computational Tools

	cumulative enrichment	no. of active scaffolds
CDK2		
ligand-based infodesign	1.61	10
ligand-based ensemble model	1.51	8
structure-based infodesign	1.41	12
DOCK(best)	1.74	13
DOCK(average)	1.34	10
DOCK(worst)	0.94	7
random	0.94	10
maximum	4.19	14
Protease		
ligand-based infodesign	2.18	13
ligand-based ensemble model	2.89	12
structure-based infodesign	1.58	12
DOCK(best)	2.07	11
DOCK(average)	1.54	9
DOCK(worst)	1.01	7
random	1.04	12
maximum	3.24	14

methods perform better than randomly selecting compounds. In other words, when looking for active compounds there is little downside to using any of these methods and there is a potentially large upside. It can also be seen that the methods providing the highest enrichment did not result in the largest number of active scaffolds. This suggests that there may be a tradeoff between obtaining higher cumulative enrichments and finding active scaffolds. This observation might be attributable to the specific compositions of the data sets and conditions of this study.

The random selection numbers in Table 1 reflect one trial. 1000 random selections in the CDK2 data set yielded an average of 10.7 ($\sigma = 1.4$) scaffolds, the minimum number of scaffolds recovered randomly was 6 and the maximum, 14. Average enrichment for the random trials was 1.0 ($\sigma = 0.14$). Thus, the enrichments achieved by the computational methods are generally significant. The active scaffold parameter is more difficult to interpret, except to note that finding more active scaffolds is desirable.

For docking, the best, worst, and average values for these parameters for the one-time selection of the required number of compounds are reported. The DOCK scoring scheme that yielded the best performance varied between the two targets. While taking a consensus of score among protein structures yielded the best performance for the CDK2 data set, the same method gave extremely poor results for the protease set. This exemplifies one of the difficulties of using DOCK in this setting: determining the optimal scoring function a priori is not straightforward. The average values achieved using DOCK are used for the purpose of comparison. In addition to overall enrichment, the per-round performance is shown in Figures 2 and 3, for the CDK2 and protease data sets, respectively. Since the DOCK score would not change based on new data, its performance is shown at a constant level for both the best and worst case.

Discussion. In examining these results it is important to consider that tradeoffs will be required in selecting the optimal tool for a given problem. In particular, for the parameters examined here, obtaining higher cumulative enrichments can come at the expense of finding fewer scaffolds. These trends need to be considered in the context of a particular project's needs: in early discovery it might be acceptable to sacrifice enrichment to find more scaffolds; conversely, in optimization enrichment would be emphasized.

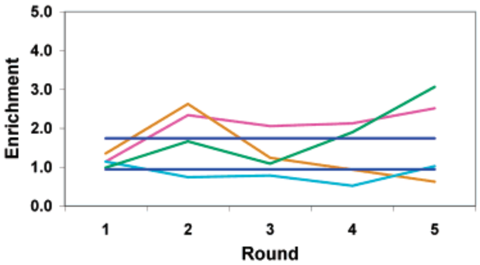

Figure 2. Plot of enrichment per round in CDK2 data set. Each line represents the enrichment in each selection cycle; the colors correspond to the methods as follows: ligand-based ensemble model: gold; ligand-based informative design: magenta; informative design in structure-derived space: green; the DOCK best and worst (blue) performance bounds are indicated by horizontal blue lines; one trial of random selection in each round is shown in turquoise.

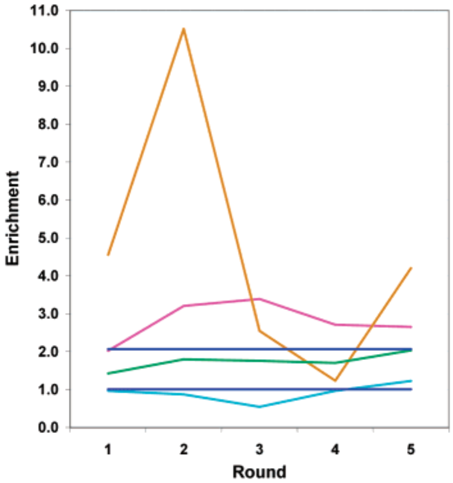

Figure 3. Plot of enrichment per round in Protease data set; color scheme is the same as Figure 2.

The ensemble model method produces an impressive cumulative enrichment for the protease data set and a cumulative enrichment comparable to the other methods (at the expense of active scaffolds) for the CDK2 data set. Examining the round-to-round enrichments for the ensemble models, however, reveals some potential deficiencies with this method. For both data sets, the ensemble models find many active compounds in the early rounds but the enrichment drops off in later rounds. This can be interpreted as the ensemble models being good at finding compounds similar to the known actives, but having to find novel chemotypes serendipitously.

This appears to be consistent with the distributions of Tanimoto similarities between training and source active compounds observed for the two data sets in Figure 1. The protease pool actives are relatively more similar to the training set actives; therefore, it is expected that the ensemble models derived from the training actives should be more able to discern active compounds in the source pool. The ensemble methods, therefore, are more suitable for projects where the pharmacophores of interest are more defined and the structure–activity relationships are better established, and where finding alternate chemical series has lower priority.

The informative design based methods tend to support the discovery of larger numbers of active scaffolds. This observation is not particularly surprising given that one of the goals of this method is to sample efficiently the available descriptor space. In addition, informative methods manage scaffold discovery without a large sacrifice in enrichment.

This study also provides comparisons involving structure-based methods; informative design in structure-derived pharmacophore spaces yields enrichments essentially identical to the average values for DOCK. The internal method, however, discovers more active scaffolds and provides a defined way in which to capitalize on both target structure information and screening data. Moreover, the structure-based informative design method should prove more consistent across all targets as it is not sensitive to scoring function choices.

There is no serious deficiency in the performance of the ligand-based methods with respect to the structure-based methods. Thus, applying these methods can compensate for the lack of target structure. Conversely, the results also show that in the absence of screening data or lead compounds target structure is useful for prioritizing compounds for synthesis.

It is not entirely appropriate to name a single method as being superior; nevertheless, several trends emerge in the data. First, methods utilizing informative design more consistently discover a larger number of scaffolds as well as increasing enrichment round-to-round. Second, while docking can deliver good performance, it is sensitive to the scoring function, the choice of which is often not apparent. Third, a well-designed screening library and follow-up strategy can deliver performance similar to that observed for methods leveraging target structural information. Finally, methods that can incorporate both target structural information and ligand screening data such as informative design in a structure-derived pharmacophore space can yield the most balanced performance.

Overall, applying any of the methods examined here generally improves the chances of finding active compounds and in no case decreases the number of active compounds discovered below that which would be found by random selection. Further, applying these methods often leads to discovering novel chemotypes that may be overlooked by traditional medicinal chemistry methods.

This letter reports a more rigorous approach to evaluating the impact of computational methods on medicinal chemistry projects. These comparisons realistically and critically examine the performance that can be expected from a variety of computational tools. By benchmarking methods in this manner, insights will be learned that can lead to improving existing methods. Finally, these studies can help develop valuable intuition into optimal applications of the respective methods.

Acknowledgment. The authors thank the therapeutic project teams including G. Beaton, A. Castellino, S. Cheng, J. Cohen, K. Gubernator, K. Jenkins, D. Kassel, M. Ramirez-Weinhouse, L. Robinson, D. Rogers, R. Rourick, E. Saiah, J. Srinivasan, M. Wayland, R. Xu, and J. Zhou for providing the data.

References

(1) Sielecki, T. M.; Boylan, J. F.; Benfield, P. A.; Trainor, G. L. Cyclin-dependent kinase inhibitors: useful targets in cell cycle regulation. *J. Med. Chem.* **2000**, *43*, 1−18.

(2) Daylight; Daylight Chemical Information Systems, Inc.; Mission Viejo, CA, 92691.

(3) Smellie, A.; Stanton, R. V.; Henne, R. M.; Teig, S. Conformational Analysis by Intersection: CONAN. *J. Comput. Chem.* **2003**, *24*, 10−20.

(4) Bradley, E. K.; Beroza, P.; Penzotti, J. E.; Grootenhuis, P. D.; Spellmeyer, D. C. et al. A rapid computational method for lead evolution: description and application to alpha(1)-adrenergic antagonists. *J. Med. Chem.* **2000**, *43*, 2770−2774.

(5) Beno, B. R.; Mason, J. S. The design of combinatorial libraries using properties and 3D pharmacophore fingerprints. *Drug Discov. Today* **2001**, *6*, 251−258.

(6) Penzotti, J. E.; Lamb, M. L.; Evensen, E.; Grootenhuis, P. D. A computational ensemble pharmacophore model for identifying substrates of P-glycoprotein. *J. Med. Chem.* **2002**, *45*, 1737−1740.

(7) Lanctot, J. K.; Putta, S.; Lemmen, C.; Greene, J. Using Ensembles to Classify Compounds for Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2003**, in press.

(8) Teig, S. L. Are informative libraries better? *J. Biomol. Screen* **1998**, *3*, 85−88.

(9) Miller, J. L.; Bradley, E. K.; Teig, S. L. Luddite: an information-theoretic Library Design Tool. *J. Chem. Inf. Comput. Sci.* **2002**, *43*, 47−54.

(10) Bradley, E. K.; Miller, J. L.; Saiah, E.; Grootenhuis, P. D. J. Informative Library Design as an Efficient Strategy to Identify and Optimize Leads: Application to Cyclin-Dependent Kinase 2 Antagonists. *J. Med. Chem.* **2003**, *46*, 4360−4364.

(11) Eksterowicz, J. E.; Evensen, E.; Lemmen, C.; Brady, G. P.; Lanctot, J. K. et al. Coupling structure-based design with combinatorial chemistry: application of active site derived pharmacophores with informative library design. *J. Mol. Graph Model* **2002**, *20*, 469−477.

(12) Schneider, G.; Bohm, H. J. Virtual screening and fast automated docking methods. *Drug Discov. Today* **2002**, *7*, 64−70.

(13) Gruneberg, S.; Stubbs, M. T.; Klebe, G. Successful virtual screening for novel inhibitors of human carbonic anhydrase: strategy and experimental confirmation. *J. Med. Chem.* **2002**, *45*, 3588−3602.

(14) Good, A. C.; Krystek, S. R.; Mason, J. S. High-throughput and virtual screening: core lead discovery technologies move towards integration. *Drug Discov. Today* **2000**, *5*, 61−69.

(15) Fox, S.; Farr-Jones, S.; Yund, M. A. High Throughput Screening for Drug Discovery: Continually Transitioning into New Technology. *J. Biomol. Screen* **1999**, *4*, 183−186.

(16) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43*, 4759−4767.

(17) Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aided Mol. Des.* **2001**, *15*, 411−428.

(18) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity: a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3210−3228.

(19) Marsili, M.; Gasteiger, J. π charge distribution form molecular topology and p orbital electronegativity. *Croat. Chem. Acta* **1980**, *52*, 601−614.

(20) Russo, A. A.; Jeffrey, P. D.; Patten, A. K.; Massague, J.; Pavletich, N. P. Crystal structure of the p27Kip1 cyclin-dependent-kinase inhibitor bound to the cyclin A-Cdk2 complex. *Nature* **1996**, *382*, 325−331.

(21) Meijer, L.; Thunnissen, A. M.; White, A. W.; Garnier, M.; Nikolic, M. et al. Inhibition of cyclin-dependent kinases, GSK-3beta and CK1 by hymenialdisine, a marine sponge constituent. *Chem. Biol.* **2000**, *7*, 51−63.

(22) Schulze-Gahmen, U.; De Bondt, H. L.; Kim, S. H. High-resolution crystal structures of human cyclin-dependent kinase 2 with and without ATP: bound waters and natural ligand as guides for inhibitor design. *J. Med. Chem.* **1996**, *39*, 4540−4546.

(23) Gray, N. S.; Wodicka, L.; Thunnissen, A. M.; Norman, T. C.; Kwon, S. et al. Exploiting chemical libraries, structure, and genomics in the search for kinase inhibitors. *Science* **1998**, *281*, 533−538.

(24) Paul, N.; Rognan, D. ConsDock: A new program for the consensus analysis of protein−ligand interactions. *Proteins* **2002**, *47*, 521−533.

(25) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **1999**, *42*, 5100−5109.

(26) Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus scoring for ligand/protein interactions. *J. Mol. Graph Model* **2002**, *20*, 281−295.

(27) Kennedy, T. Managing the drug discovery process. *Drug Discov. Today* **1997**, *2*, 436−444.

JM025618T