Brief Articles

Informative Library Design as an Efficient Strategy to Identify and Optimize Leads: Application to Cyclin-Dependent Kinase 2 Antagonists

Erin K. Bradley,*,† Jennifer L. Miller,‡ Eddine Saiah,‖ and Peter D. J. Grootenhuis§

Deltagen Research Laboratories, 740 Bay Road, Redwood City, California 94063

Received October 23, 2002

The application of an informative, iterative library design strategy is presented for lead identification and optimization. The computational algorithm underlying informative design systematically uses data from both active and inactive compounds and maximizes the information gained from subsequent design−synthesis−screening cycles. Retrospective analysis of a released dataset of 17 550 compounds and corresponding cyclin-dependent kinase-2 activities showed that informative library design yields significant enrichments of active compounds and efficiently discovers novel chemotypes in comparison with commonly used diversity−similarity protocols.

Introduction

Although once perceived by some as a threat to molecular modeling and design, the advance of combinatorial and parallel chemistry approaches is now seen as a powerful ally of the computational chemist. No longer is there a need to select one or a few compounds based on a particular design method; typically ideas can be tested by libraries of hundreds of compounds. This is essential since many molecular modeling approaches are qualitatively predictive at best. In fact, for most methods, extrapolation to chemotypes beyond the set of compounds used to derive the model remains a significant problem in the day-to-day practice of the (computational) medicinal chemist.

The art and science of computational library design have been reviewed extensively.[1−3] The method and objective of the design can vary dramatically depending on the stage and nature of the therapeutic project. In the early phases, library designs may be aimed at finding chemically diverse hits that can be optimized to leads. Once hits have been identified, library designs may focus on quickly analoging around particular structural moieties to increase understanding of structure−activity relationships. Library design methods have also been used to compare, design, or supplement screening libraries. Different considerations can be included in the designs such as the drug-likeness, and one typically applies filters to remove the "swill".[4,5]

This paper describes the application of a novel strategy called *informative library design*.[6] The goal of informative design is to use molecules to "interrogate" the target receptor and determine what chemical features are required for activity. Each molecule, given its conformational flexibility, is able to ask many questions. The informative design method composes the library in such a way that a maximum number of conclusions can be drawn from the "answers" (assay results), as shown in Figure 1. This is accomplished by maximizing the Shannon entropy of the library (vide infra), which is described elsewhere.[7,8] Note that getting enhanced hit rates is not the a priori objective of an informative design strategy. The method involves both data and model generation, taking advantage of the iterative design−synthesis−screening cycle. After several rounds of this cycle, the model for activity (i.e., features required for binding) converges and may be applied for designing combinatorial libraries based on novel scaffolds or selecting compounds for testing from other sources.

This paper illustrates the utility of an informative design strategy with a retrospective analysis using compounds that were assayed during the course of a cyclin-dependent kinase 2 (CDK2) antagonists[9,10] project. We approximated the real-life situation of a drug discovery project by simulating multiple rounds of design−synthesis−screening and then measured performance by monitoring enrichment of active compounds, model convergence, and the ability to discover different active chemotypes. The latter is important, since the ability to identify and optimize activity on chemically diverse scaffolds greatly enhances the chances of finding active compounds that have a favorable pharmacokinetic profile. The informative design strategy has previously been shown to select many different chemotypes when the target's crystal structure information is used.[8] Here, we demonstrate the utility and advantages of this approach in the general case (i.e., unknown target structure). Finally, we compared this strategy with a more conventional diversity−similarity strategy that is used frequently in the pharmaceutical industry.

* To whom correspondence should be addressed: Tel: 650-266-3675. Fax: 650-266-3501. E-mail: ebradley@sunesis.com.
† Current address: Sunesis Pharmaceuticals, Inc., 341 Oyster Point Blvd., South San Francisco, CA 94080.
‡ Current address: Consultant, 935 College Drive, Menlo Park, CA 94025.
‖ Current address: Kémia, 9390 Towne Center Drive, San Diego, CA 92121.
§ Current address: Vertex Pharmaceuticals, Inc., 11010 Torreyana Road, San Diego, CA 92121.

10.1021/jm020472j CCC: $25.00
Published on Web 08/30/2003

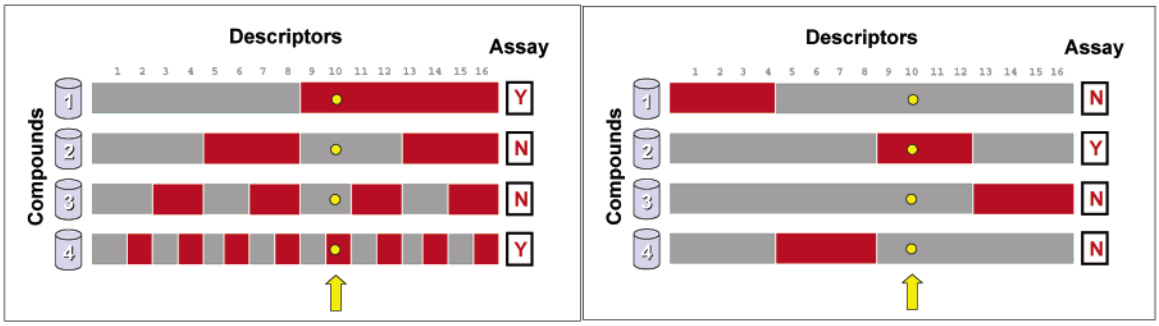

Figure 1. Schematic comparison of informative (left) and diversity (right) library design. In both figures, molecules are represented as rows, descriptors (ex: chemical feature) as columns. In the Assay column, "Y" corresponds to active. Each design method has selected four molecules to test a different portion of the "design space". Informative design uses molecules to "interrogate" a target receptor and determine which chemical features are required for activity. Each molecule is a "question", with the red areas indicating which features the molecule contains and the assay result is the "answer". Informative design selects molecules to maximize the difference between the patterns in the descriptor columns (codes). Unique codes enable the identification and retention of important features when the compounds are assayed. In this example, the assay result is "YNNY". Since every possible assay outcome corresponds to one code, feature 10 (shown by yellow dots) can be identified as a determinant of activity. In contrast, diversity methods select molecules to maximize the difference between the patterns in the molecules. The diverse design does find in active compound, but the assay code ('NYNN') has four equivalent explanations for activity: features 9−12.

Round	Informative	Diversity-Similarity	Random
0	1000 cmpds Informative Design	1000 cmpds Diversity Design	1000 cmpds Random sample
1	500 cmpds Informative Design	500 cmpds Similarity Design	
2	200 cmpds Informative Design	200 cmpds Similarity Design	
3	200 cmpds Informative Design	200 cmpds Similarity Design	
4	100 cmpds Similarity Design	100 cmpds Similarity Design	1000 cmpds Random sample

Figure 2. Protocols for methods comparison. Round 0 represents the design of a general screening collection that is used at the start of every project. A set of 1000 compounds was selected using each method from the general screening source pool of 13 359 compounds (gray boxes). In Rounds 1−3, small "libraries" of 200−500 compounds were selected from the targeted source pool of 4191 compounds. These rounds represent the typical SAR expansion/potency optimization that is done based on the initial screening results. Finally, in Round 4, similarity searching was applied to both protocols. In the case of the informative design, the similarity was calculated in the refined descriptor space generated in the previous rounds.

Methods

Protocol for Retrospective Computational Analysis. This study simulates the iterative nature of molecular selection/synthesis and information flow in a real-life therapeutic project setting. To this end, we distinguish an "early stage" (lead generation/identification) and a "late stage" (lead optimization/SAR expansion). In the early stage relatively large numbers of molecules are selected from a chemically diverse corporate or public compound collections. The late stage involves selecting a smaller numbers of compounds and operates on combinatorial collections characterized by large numbers of potential compounds on a limited set of scaffolds, targeted for the receptor of interest. To apply comparisons across different methods, five rounds of selection were defined, each having restrictions on the number of compounds that could be selected and the source pool of potential compounds. The protocols for each method are shown in Figure 2.

Compound Source Pool/Data Set. See Supporting Information. In the course of the CDK2 project, 17 550 compounds were screened for CDK2 antagonistic activity. Compounds with an IC_{50} lower than 25 μM or inhibition larger than 50% at 10 μM were considered active. The data set was split into two subsets for the retrospective study: general screening pool and targeted screening pool. The first pool consisted of 13 359 chemically diverse compounds from a general screening library, containing 207 actives. The targeted pool contained 161 actives and consisted of 951 compounds selected from the ACD by the project chemists, using standard ISIS[11] similarity searching, and 3240 combinatorial synthetic compounds on more than 22 unique chemical scaffolds, with actives on 14 out of the 22 scaffolds. These compounds were originally selected for synthesis using a variety of techniques, including Daylight fingerprint similarity,[12] medicinal chemistry intuition, informative library design, and similarity in pharmacophore signatures.

Informative Library Design Strategy. Whole molecule 3D pharmacophore-based descriptors were used,[13,14] with six feature types: hydrogen-bond acceptors and donors, hydrophobes, negative and positive charges, and aromatic groups.[15] The resulting potential "pharmacophore space" contained 3.4 million pharmacophores. Each molecule was encoded as a "molecular signature", a bit string recording the presence or absence of each pharmacophore in any of its conformations.[13,14,16]

The "design space" for each round of informative design consisted of a subset of the total "pharmacophore space". In Round 0, the design space was the subset of all pharmacophores displayed by at least 10 molecules in the general screening pool, ~1.8 million. For each of the subsequent rounds (1−4), the design space was determined by the activity data known at that point. Each pharmacophore was evaluated separately for its usefulness on the basis of ability to discriminate between actives and inactives in the existing data set using the "discrimination ratio" (DR), defined as $DR = [(A_{pharm}/A_{total})/(I_{pharm}/I_{total})]$. Where A_{pharm} and I_{pharm} denote the number of active and inactive molecules containing the pharmacophore, respectively, and A_{total} and I_{total} denote the total number of active and inactive molecules in the data set. The DR is a function of both active and inactive molecules and the total size of the existing data set, measuring the tradeoff between true positives and false positives. In each round of design, pharmacophores with a DR greater than 10 were retained as the "design space". A DR of 10 is equivalent to matching 1% actives and 0.1% inactives. Thus the design space is targeted for the receptor being assayed, since the retained pharmacophores differentiate between active and inactive molecules.

In rounds 0−3, compounds were selected using the informative design technique (Figure 2).[6,7] In round 4, the active molecules from the previous rounds were used as queries, and the Tanimoto similarity[17] was calculated between each query

Table 1. Cumulative Results for Different Computational Selection Methods

	cumulative enrichment[a]	fraction of actives recovered	active scaffolds recovered
General Screening Library: One Round, 1000 out of 13 359 Compounds Selected			
informative	2.00	0.15	
diversity	0.65	0.05	
random[b]	0.99	0.07	
maximum[c]	13.60	1.00	
Targeted Screening Library: Four Rounds, 1000 out of 4191 Compounds Selected			
informative	2.56	0.63	11
similarity	1.31	0.32	7
random[b]	1.03	0.24	10
maximum[c]	4.19	1.00	14

[a] Enrichment = (no. selected actives/no. selected)/(no. pool actives/no. pool) [b] Average of 10 random selections. [c] Maximum attainable, corresponds to recovery all actives in selected set.

molecule and compounds remaining in the source pool using the final 82K pharmacophore bit strings.

Diversity/Similarity-Based Library Design Strategy. Whole molecule 2D descriptors called MACCS keys[18] were used as implemented in the MOE platform from Chemical Computing Group.[19] This allowed both diversity design and similarity searching in the same descriptor space. The design space for each round of diversity and similarity design was the complete set of MACCS keys.

The selection method for Round 0 was the default molecular diversity-based subset selection method implemented in MOE. For each of the subsequent rounds (1−4), the design method was similarity searching using MACCS keys. Two variations of similarity selection criterion were tested: (1) Retain all molecules with a Tanimoto similarity greater than some threshold to any previously identified active. (2) Retain the topmost similar compounds to each of the previously identified actives. The first method performed the best, and those results are reported for comparison.

Random Control. To determine if either of the design methodologies produced significant improvements, we performed 10 random selections of compounds. The average results are reported.

Results and Discussion

The informative library design method differs from more traditional diversity−similarity procedures in its initial goal of building/optimizing a computational model, which is subsequently applied to identify actives on multiple chemical scaffolds. However, both selection methods have the same long-term objective: identifying the best compounds using the least amount of time and resources. Thus we compared the methods by holding the "resources" (compound source pools and number of compounds selected) constant and monitored performance with standard therapeutic project metrics: fraction of actives recovered, enrichment for activity in the designed library versus the source pool, and number of active scaffolds identified. The results are collated in Table 1. Enrichment is difficult to compare across different publications because it depends on the source pool composition and the selection size. So, Table 1 also includes the "maximum" enrichment, which corresponds to finding all the actives.

Table 2. Per Round Results on Model Refinement and Enrichment[a]

	informative		diversity−similarity	random[b]
round	pharmacophores	enrichment	enrichment	enrichment
0	1800 K	2.00	0.65	0.99
1	178 K	1.35	0.86	
2	111 K	3.69	1.14	
3	82 K	3.56	2.30	
4	82 K	7.71	1.56	1.03

[a] Enrichment = (no. selected actives/no. selected)/(no. pool actives/no. pool). [b] Average of 10 random selections.

Since the informative library design strategy is aimed at refinement of the computational model, we monitored the size of pharmacophore design space in each round (Table 2). The four rounds of informative design refined the model from 1.8 M to 82 K pharmacophores, at which point the final model was used for similarity search. When applied in Round 4, this final model yielded an enrichment of 7.7. It can also be observed in Table 2 that, although informative design does not initially aim to produce a library enriched for activity, there is an increase in library enrichment for each subsequent round.

Diversity design, like informative design, does not aim to identify actives, but rather to sample the chemical space. So it is not surprising that the initial screening library selection resulted in an enrichment less than random sampling, yielding only 10 out of 207 possible active leads. However, the observation that the similarity design method did not consistently produce libraries

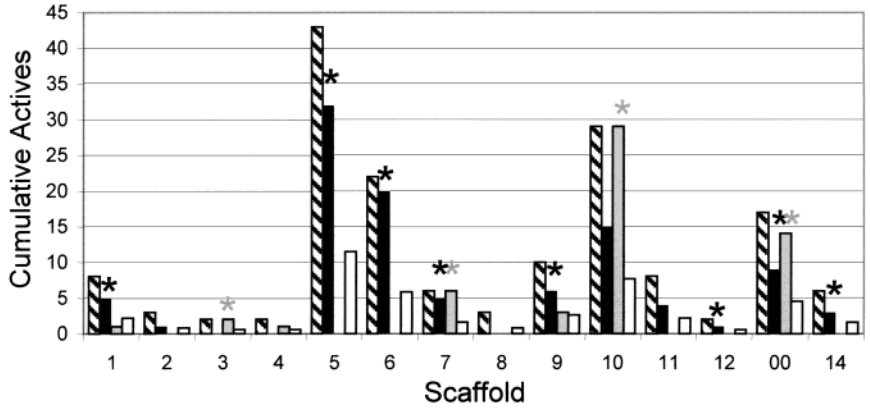

Figure 3. Active scaffold recovery. Histograms shows cumulative number of actives recovered on each of the 14 active scaffold classifications in the targeted source pool after four rounds of selection: source pool (black hashed bar), informative design (solid black bar), similarity design (solid gray bar), and random selection (white bar). An asterisk above the bar indicates that the most potent compound on the scaffold was retrieved by the corresponding method. Scaffold_00 contains a diverse set that did not easily fall into chemotype classifications.

enriched for activity is somewhat surprising. Although it has been noted recently that similar molecules do not necessarily have similar activities,[20,21] the nearest-neighbor principle is often employed as the technique-of-choice in the hit-to-lead process. The first round of similarity searching produced a lower than random enrichment, but subsequent rounds did result in enrichments of up a factor of 2.

For both the general and targeted screening selections, informative design provides higher enrichments than diversity−similarity or random control. In particular, for the targeted screening library selections, the cumulative enrichment of 2.56, observed for informative design, is more than 60% of the maximum attainable enrichment of 4.19. At all points, iterative informative design performed better than both random selection and diversity−similarity. Multiple rounds of informative design, followed by a final similarity search in the refined descriptor space, recover 60% of the actives in the source pool, having "synthesized" only 25% of the potential compounds in the source pool.

Figure 3 shows the overall performance of the different design protocols with respect to their ability to identify and optimize novel scaffolds. When informative library design was used, 11 of the 14 active chemical series were identified. Of those series, the most potent compound was recovered in eight scaffolds, and the second most potent in the other two cases. When similarity searching followed diversity design, only 7 of the 14 active chemical series were identified. This latter result is not surprising since similarity-based selections tend to pick chemically related compounds, which is arguably not the best strategy for discovering all the active scaffolds. Even random selection outperforms the similarity strategy for the same reason.

Figure 3 also hints at the difficulty of using similarity searching in combinatorial libraries. Similarity searching retrieved more than 90% the actives on scaffolds 3, 7, 10, and 13. However, in those cases, similarity searching retrieved the majority of the inactives as well. Thus the source of the lower enrichments for this approach is its tendency to retrieves all the compounds from a scaffold once an active is identified.

Conclusions and Perspective

There are many possible variables in any comparison of computational selection strategies (e.g., descriptor type, selection method, scoring function, fraction selected, etc.). Answering all those questions is beyond the scope of this study. We hope that others will use the supplemental data provided to expand these initial comparisons. Here, our goal was to compare informative design to a standard commercially available selection protocol (diversity−similarity), under realistically simulated conditions. Since the descriptors were not held constant, we cannot say conclusively that the improved performance was related to the 3D descriptors, the selection method, or a combination of both. However, the overall result is in agreement with another comparative study on the performance of active-site derived pharmacophore models in combination with informative design.[8] In that case, using the same compound and activity data, informative design performed better than docking or similarity protocols using a different set of descriptors. It appears that using either receptor-based or ligand-based information, the strength of informative design is in refining the model from all possible features/interactions, down to only those critical for binding. While this cannot be demonstrated unambiguously without testing all possible variables (e.g., 2D MACCS keys used with informative design), the overall conclusions are unlikely to change. This retrospective study has clearly demonstrated that informative design used with 3D pharmacophores significantly outperforms a commonly used diversity−similarity selection using chemical substructure descriptors.

Acknowledgment. The authors would like to thank our teammates, A. Castellino, S. Cheng, J. Cohen, K. Gubernator, K. Jenkins, D. Kassel, M. Ramirez-Weinhouse, L. Robinson, D. Rogers, R. Rourick, J. Srinivasan, M. Wayland, and R. Xu for their contributions to the CDK2 project data set, and C. Schachtele and associates at IMM in Freiberg for performing the assays.

Supporting Information Available: Information for compounds used for computational method performance comparisons. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Spellmeyer, D. C.; Grootenhuis, P. D. J. Recent Developments in Molecular Diversity: Computational Approaches to Combinatorial Chemistry. *Annu. Rep. Med. Chem. Rev.* **1999**, *34*, 287−296.
(2) Beno, B. R.; Mason, J. S. The Design of Combinatorial Libraries Using Properties and 3D Pharmacophore Fingerprints. *Drug Discov. Today* **2001**, *6*, 251−258.
(3) Willett, P. Chemoinformatics − Similarity and Diversity in Chemical Libraries. *Curr. Opin. Biotechnol.* **2000**, *11*, 85−88.
(4) Clark, D. E.; Pickett, S. D. Computational Methods for the Prediction of 'Drug-likeness'. *Drug Discovery Today* **2000**, *5*, 49−58 and references cited.
(5) Blake, J. F. Chemoinformatics − Predicting the Physicochemical Properties of 'Drug-like' Molecules. *Curr. Opin. Biotechnol.* **2000**, *11*, 104−107.
(6) Teig, S. L. Are informative libraries better? *J. Biomol. Screening* **1998**, *3*, 85−88.
(7) Miller, J. L.; Bradley, E. K.; Teig, S. L. Luddite: an information-theoretic Library Design Tool. *J. Chem. Inf. Comput. Sci.* **2003**, *1*, 47−54.
(8) Eksterowicz, J. E.; Evensen, E.; Lemmen, C.; Brady, G. P.; Lanctot, J. K.; Bradley, E. K.; Saiah, E.; Robinson, L. A.; Grootenhuis, P. D. J.; Blaney, J. M. 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.
(9) 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.
(10) Knockaert, M.; Greengard, P.; Meijer, L. Pharmacological inhibitors of cyclin-dependent kinases. *Trends Pharm. Sci.* **2002**, *9*, 417−425.
(11) ISIS Base 2.2.1, MDL Information Systems, Inc, San Leandro, CA.
(12) Daylight, 1995, Daylight Chemical Information Systems, Santa Fe, NM.
(13) Bradley, E. K.; Beroza, P.; Penzotti, J. E.; Grootenhuis, P. D. J.; Spellmeyer, D. C.; Miller, J. L. A Rapid Computational Method for Lead Evolution: Description and Application to α1-Adrenergic Antagonists. *J. Med. Chem.* **2000**, *43*, 2770−2774.
(14) Penzotti, J. E.; Lamb, M. L.; Evensen, E.; Grootenhuis, P. D. J. A Computational Ensemble Pharmacophore Model for Identifying Substrates of P-Glycoprotein. *J. Med. Chem.* **2002**, *45*, 1737−1740.
(15) Greene, J.; Kahn, S.; Savoj, H.; Sprague, P.; Teig, S. Chemical Functional Queries for 3D Database Search. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1297−1308
(16) Smellie, A.; Stanton, R. V.; Teig, S. L. Conformational Analysis by Intersection: CONAN. *J. Comput. Chem.* **2003**, *24*, 10−20.

(17) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(18) Hanzen, G.; et al. *MACCSII Facilities Guide and Reference*, Molecular Design Limited; San Leandro, CA.
(19) *Molecular Operating Environment (MOE 2000.02)*; Chemical Computing Group Inc: Montreal, Quebec, Canada.
(20) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.
(21) Kubinyi, H. Similarity and Dissimilarity − a Medicinal Chemists View. *Perspect. Drug Discov. Des.* **1998**, *11*, 225−252.

JM020472J