Reprint from

GENETIC ENGINEERING NEWS

Vol. 17 No. 20 November 15, 1997

Drug Discovery Development of a Universal Informer Library: Data Derived from the Training Set

By John Saunders, Ph.D., Peter L. Myers, Ph.D., Doug Barnum, Jonathan W. Greene, Ph.D., and Steven L. Teig

Once a target of therapeutic interest has been clearly identified and a suitable assay established, the drug discovery process requires compounds to be designed, synthesized, tested and the resulting data assessed. Where there is little prior information, design cannot play a useful role. As an alternative approach, pharmaceutical companies attempt to leverage the strengths of high-throughput screening by testing large libraries of compounds, attempts which are often limited by the diversity of the company's corporate collection and more recently prepared combinatorial libraries. However, this 'shotgun' approach only occasionally provides lead compounds worthy of a medicinal chemical lead optimization and fails to capture all of the screening data, including that of only modestly active compounds.

Predictive Hypotheses

Increasingly, there are many novel targets for which there are only limited prior information, which itself exists primarily as a result of such efforts as the Human Genome Project. For this situation, **CombiChem Inc.** (San Diego, CA) has developed a Universal Informer Library, which is a collection of about 10,000 molecules selected to produce information about almost any target against which they are screened, although it may not produce leads per se. The Universal Informer Library (UIL) consists of highly promiscuous molecules (those that bind to many different targets) and is intended to provide a few weak actives against the background of many, varied inactives. Using this data, predictive structure-activity hypotheses may be extracted which allows the drug discovery cycle *(see Figure 1)* to be initiated.

Such hypotheses are an explanation in binary code format of why the active molecules in the 10,000 set are indeed active while the remainder are inactive. This representation is devoid of chemical connectivity and is the format which is most efficiently manipulated computationally.

The virtual library, consisting of billions of novel, readily synthesizable compounds, is next searched for those that best match the current hypotheses, thereby selecting candidates to be synthesized in the first daughter libraries. It is likely that there will be several poorly focused hypotheses at this stage, so the purpose of the first round of chemical synthesis is to rapidly assess those hypotheses that will converge on highly active molecules from those that may simply be the result of inaccuracies, such as spurious screening data. Screening the first daughter libraries improves the data set and allows the remaining hypotheses to be refined so that after about six cycles, highly active and selective novel ligands are produced whose chemical structures may be unrelated to those of the initial hits. Convergence of more than one hypothesis is indicative of more than one binding site on the molecular target of interest.

Before constructing the complete UIL, a "training set," with about 2,400 members, was compiled so that the parameters which contribute to promiscuity (believed to be an essential feature in ultimate library design) could be refined as necessary. This set has been evaluated in 10 biologically diverse assays *(see Table 1, above)*.

Compounds were screened at a single final concentration of 30 μM, with duplicate observations per compound. Repeat assay was performed when these two measurements were outside an acceptable range (20% inhibition). As can be seen from a superficial analysis of the data returned from screening this set *(see Table*

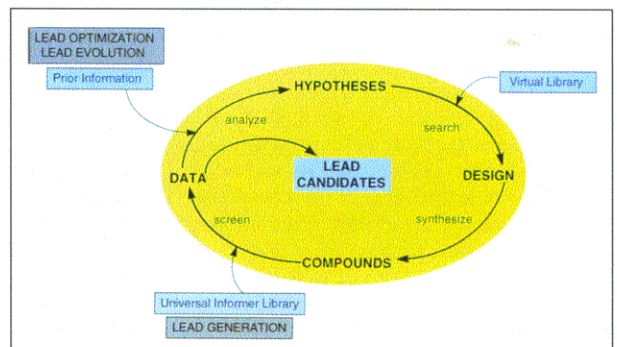

Figure 1. The drug discovery cycle.

ASSAY (at test conc. 30 μM)	NUMBER OF COMPOUNDS SHOWING INHIBITION > 90%	>50%
Dopamine Uptake	115	391
Sodium Channel	128	350
Dopamine^R $D_{4.4}$	122	218
Central Benzodiazepine^R	25	156
Phosphodiesterase PDE-II	9	70
EGF Tyrosine Kinase	13	33
Inducible NO-synthase i-NOS	4	19
Cathepsin-B	1	10
Glucagon-like Peptide^R GLP-1	0	6
Tumour Necrosis Factor^R TNF-α	0	1

Table 1. Ten biologically diverse assays used to evaluate the training set.

1), multiple hits were identified in every assay but one, although there was considerable variation in the total number. Clearly, in a library four times the size of the training set, hypothesis generation of at least eight of the targets would be sufficiently refined as to justify the combinatorial synthesis of daughter libraries.

i-NOS Results

First consider one of those assays that returned a smaller number of actives—i-NOS, for example. The profile for the top eleven hits is displayed in *Figure 2*. In the chart, each compound is represented by a bar color-coded for each of the ten assays *(see legend)*, with the height of each bar recording level of activity. We consider this chart to be close to ideal for an informer library with the correct level of promiscuity for the other targets. Note that each compound is not indiscriminate in that they show some level of activity in two or three other assays without uniformally hitting every target. In addition, the i-NOS hits do not systematically interact with another single target.

As is obvious from the chemical reaction catalyzed by i-NOS, simple guanidine derivatives, including arginine itself, should be expected to show affinity for the active site; four of the eleven hits are guanidines, leaving seven other interesting hits to establish an i-NOS hypothesis. Unfortunately, the enzymatic reaction proceeds in two steps, both NADPH dependent, so that at least three independent hypotheses could be required to define this target. This complication can clearly be handled by the methodology (see above) but may well require a larger database of hits and additional cycles to distinguish the competing binding modes. Overall, this state of affairs is not dissimilar to interpreting data from cell-based assays, where a cascade of multiple molecular targets within a single pathway. When the end-product is measured, it may contribute to an initial 'fuzzy' picture; indeed, cell- penetrability issues will further confound the process. In this situation, since there will appear to be many distinct hypotheses, a corresponding increase in the number of primary daughter libraries will have to be synthesized.

D4 Results

Next, consider the other extreme typified by the D4 results. Here, 218 compounds inhibited specific binding by greater than 50%, with 122 showing over 90% inhibition at the test concentration of 30 μM, and one can see that their activities are indistinguishable; to put these ligands in rank order, a repeat assay at a lower concentration will be required. Fortunately, the dopamine uptake site, known to have considerable cross-talk with dopamine receptors, may be used as a specificity gate (in the absence of D2 binding data) to limit the data set and to focus hypothesis generation toward D4 (or vice versa, as required). Restricting the list in this way (D4 > 80%; DAUP < 30% inhibition) leaves 15 'selective' compounds *(see Figure 3)* for rapid convergent-hypothesis generation although many of these have known affinity for closely related gpcr's.

As a consequence of this argument, it can be expected that more rapid progress toward specific hypotheses for a given molecular target will be achieved by evaluating the universal library simultaneously in a homologous target assay.

In summary, the training set has shown that the UIL approach to library design, will be validated for many molecular targets but that some prior knowledge about the nature of that target may facilitate the design process, as illustrated. Receptor assays present more binding opportunities to potential ligands, thereby increasing the apparent promiscuity of such ligands. Enzymes, on the other hand, are evidently more discerning but, because of the sensitivity of the assay system to highly functionalized compounds, may more readily yield false-positives. Finally, cell-based assays involving a read-out from a complex intracellular signaling pathway, or targets having multiple binding modes, will be approachable by the current methodology, albeit requiring a more intensive deconvolution of the compound hypotheses. Relevant selectivity and cell penetrability assays will assist this process. The Discovery Engine is already being put to work in collaborations with **Roche Biosciences**, **Sumitomo** and **Teijin** and is clearly applicable to such fields as agrochemical research.

John Saunders, Ph.D., is vp, medicinal chemistry, Peter Myers, Ph.D., is COO and chief scientific officer, Doug Barnum is senior member, design staff, Jonathan Greene, Ph.D., is director, design technology, and Steve Teig is vp, advanced technology, all at CombiChem in San Diego and Palo Alto, CA.

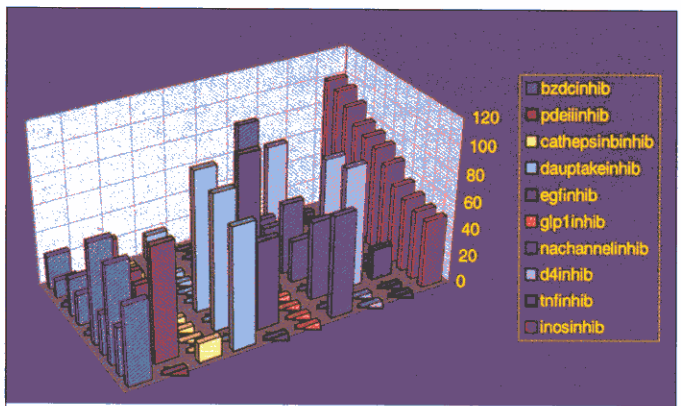

Figure 2.

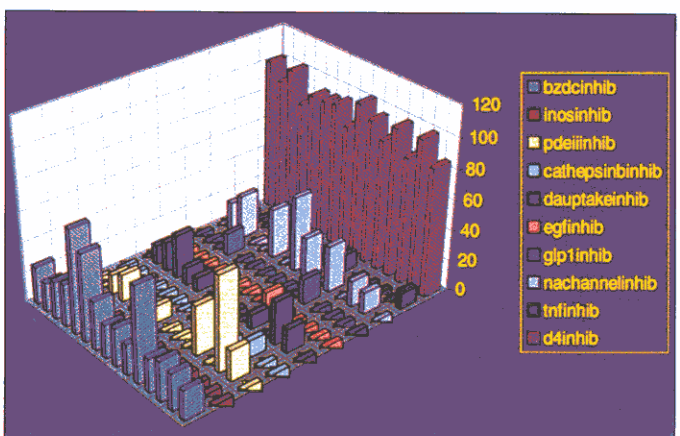

Figure 3.