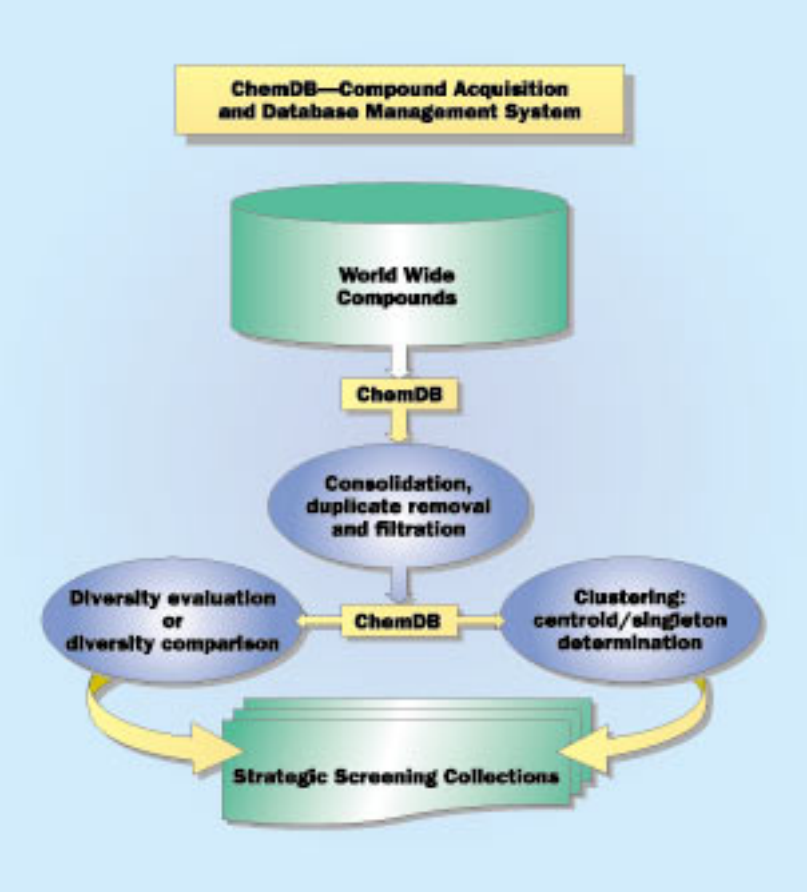
The satisfaction of initiating a drug discovery program with screening assays that incorporate newly discovered enzymes or receptors is an event that is regularly repeated in upstart biotech and pharmaceutical companies throughout the world.

Once the glow of that milestone passes, the daunting task of identifying modulators, preferably low molecular weight substances, to interact with the proteins of these assays, is the inexorable activity of the company drug discovery engine.



A flow chart explaining the compound acquisition and database management system.

The avenues for pursuing these goals are varied, but all flow through the screening tunnels that channel most agents to their unheralded fate. The prize is, of course, that unanticipated "hit" molecule that sets the direction of chemical synthesis activity, especially if the molecule already has features that are "drug-like." If not inherently possessing pharmacodynamic and pharmacokinetic properties found in acceptable therapeutics, this structure should at least lend itself to molecular modification so that these properties can be incorporated into an analog of this prized nugget of discovery research. One approach to

the identification of these hit molecules is the testing of pre-established chemically diverse compound libraries that claim to fill "pharmacophore space." Indeed, over the years a variety of these generic ligand collections have promised to provide the requisite number of compounds for probing the activity sites of pharmacological targets. While success stories are claimed with these proprietary and often expensive libraries, they do not routinely fulfill all target expectations, and hit molecules can be disappointing, if not chemically scary.

Another approach to these diversity libraries can be assembly through the acquisition of compounds from a vast array of commercial vendors where the expectation may be that a "diversity of vendors will result in a diversity of compounds." Of course, anyone familiar with the vendor arena knows that many compound sources are commonly shared, and the resulting overlap in library offerings can be quite high. Alternatively, one can shop for proprietary alternatives, but the range of pharmacophore space that these series probe is limited. These chemical pools are often produced through combinatorial methodology, and lack the template variability of agents present in the libraries of commercial vendors. And again, the cost of proprietary combinatorial libraries can be several fold that of vendor stock.

Chemical diversity management

Compounds available in the vast array of vendor libraries that by our estimation exceed two million agents do indeed have chemical diversity advantages. With the implementation of a proper strategy, it is these sources that can be a fruitful and cost-effective means of identifying reasonable chemical starting points for lead and ADMET optimization, the real challenge of the drug discovery odyssey. The cornerstone of a plan, in which chemical diversity is chosen as the expedient approach to identify pharmacological modulators, is the characterization of what is already present in a screening library, or what is being purchased in a diversity library. From that point, the goal should be to systematically add only those agents that augment the diversity of the collection in place. Avoidance of duplication is one key factor in future compound

selection. However, along with this prerequisite should be a chemical diversity assessment tool that significantly increases the likelihood that the resultant chemical library will possess agents capable of the desired protein interactions.

The TimTec Corp., Newark, Del., utilizes a particularly efficient software package called ChemDBsoft to accomplish the goal of assessing and increasing chemical diversity. The components of this approach allow for the characterization, evaluation and management of compounds so that the strategy of chemical diversity optimization can be a systematically controlled and efficient process. The characterization and analysis component utilizes descriptors, readily generated from the SD file of a library, to identify those members in a given collection that are most chemically diverse from (or similar to) each other. Each compound is characterized by a defined set of molecular fragment or structural-physicochemical descriptors. A 2D molecular fragment descriptor consists of a central atom and neighboring atoms connected to it within a predefined sphere size bonds between central and edge atoms. For each fragment, the complete connection table is stored, and the entire set of fragments with selected sphere size forms the fragment library. The frequency of occurrence for each fragment in the library is then calculated. In the case of the structural-physicochemical descriptors, each atom is characterized by partial charge, polarizability and H-bond donor/acceptor factors. With either of these characterization methods, fragment calculations are then performed to estimate the similarity of each molecule in a database with all others. While plain structural fragments are adequate for differentiating molecules based on chemical structure, the structural-physicochemical descriptors are better for separation of biologically active compounds because they inherently encode information relevant to ligand-receptor interaction. The results of compound structure comparisons with a given library can be displayed using Tanimoto, Euclidean or Cosine metrics with compounds posted in decreasing order of the highest calculated dissimilarity scores. A diversity profile plot of the addition of compounds from a new database to an existing one can reflect which compounds from the new

database will indeed add diversity to the original chemical pool, and provide an indication of how many compounds should be considered for purchase from it.

Thus, utilization of ChemDBsoft allows selection of the most acceptable chemically diverse molecules from a given collection, and the augmentation of an in-house library with only those compounds that add structural diversity to it. Future acquisitions can then be made with the assurance that a collection will increase its chemical diversity and be more effective at hit-finding with each additional compound. Moreover, chemical filters can also be used in conjunction with this package (Lipinski, functional groups, undesirable fragments, etc.) to remove molecules and structure types that would not provide a reasonable starting point for a lead optimization chemistry effort.

In summary, the ChemDBsoft package permits:

- Calculation of similarity indexes for any chemical compound with all compounds in a database.
- Determination of the complete similarity matrix for a compound database.
- Estimation of the diversity matrix of any compound database.
- Selection of the most diverse subset of compounds from any database.
- Selective import of a subset of dissimilar compounds from an external database.
- Removal of compounds with undesirable functionality, fragment or physicochemical properties.

Diversity clustering

Another efficient drug discovery screening strategy is to sample large library pools through the

selection of molecules that are representative of a group (cluster) within this library. In this component of ChemDB, Jarvis-Patrick clustering based on fingerprints (bit strings) is utilized. By identifying the agent with the highest degree of similarity (centroid) to members of these clusters, as well as those agents that are chemically unique (singletons), a sampling of a large chemical pool is possible without the cost of purchasing the majority of these molecules. This approach also has the advantage of having cluster-members immediately available, if a centroid shows activity of interest. However, there is considerable flexibility in creating the cluster-singleton distribution with this method. The selected diversity parameters—threshold of acceptable dissimilarity of a cluster, and number of nearest neighbors in common—affect amount and size of clusters and the accompanying number of singletons. With some iteration the cluster statistics can be manipulated to provide a distribution that will fulfill sampling goals and provide reasonable homogeneity in cluster groups and a manageable number of singletons. The ChemDB software package includes this sampling procedure and TimTec's access to databases of hundreds of thousand of compounds provides libraries that approach the gamut of commercially available chemical diversity. This clustering approach, coupled with diversity strategy, can be a powerful means of library development for even large, mature collections.

Conclusions

Drug discovery companies implement a variety of approaches to their hit/lead-identification challenge. Through utilization of adaptable and effective software like ChemDB from TimTec Inc., rapid determination of compound parameters and expedient deployment of strategically useful programs can improve upon the efficiency of this process. Coupled with access to hundreds of thousands of molecules in various commercial databases, diversity libraries can be developed to meet the challenges of lead identification in the most demanding of situations. Library clustering also adds an efficient and practical dimension to this strategy that can increase cost effectiveness in both library establishment and archive collection augmentation.

Chemical management software promises diversity optimization

—Robert J. Chorvat, PhD
Director of Medicinal Chemistry Business Development, TimTec Inc., Newark, Del.

Drug Discovery, Reed Business Information, Morris Plains, NJ

http://www.dddmag.com