Design and analysis of HTS libraries

This review of the key elements in the design and analysis of HTS libraries presents analysis techniques to quantify the composition of an HTS library and highlights methods for establishing structural relationships between compounds in the screening collection, which can be used in conjunction with the screening data to expand hit selection. These techniques can be further used to expand the diversity and chemical space coverage of current screening collections while maintaining desirable property distributions.

High-throughput screening (HTS) continues to be a valuable component of early-stage drug discovery. Appropriate planning and library design can shorten hit identification and validation times, thereby decreasing costs and increasing possible patent lifetime. The library design strategy that provides the best chance to identify valid hits and produce successful hit-to-lead investigations incorporates collections of diverse compounds alongside clusters of similar compounds with desirable physical properties.

Research in HTS can be roughly divided into two areas. The first area pertains to developments in the experimental screening process, such as automation, miniaturisation, and the development of new evaluation methodologies, such as NMR techniques or calorimetric data to assess the results. A second area of research pertains to the design of the chemical composition of a HTS library, subsequent use, analysis of the biological results, and exploration of hits.

The selection and analysis of compounds in screening collections has been demonstrated to be an important component in the generation of quality HTS hit data as measured by enrichment metrics. The following guidelines and techniques can be used to set up a HTS library or to expand the diversity and chemical space coverage of current screening collections while maintaining desirable property distributions.

Guidelines in HTS library assembly

In assembling a screening library, several key guidelines are required to make the compound collection applicable to a wide range of biological targets and suitable for general-purpose HTS campaigns. Compounds are selected to represent chemically diverse space, to possess desirable physical properties, and to ensure the library also covers biological space.

A comprehensive, cost-effective screening library can be generated from a diverse aggregate of external compound acquisitions, internal collections, and the custom synthesis of small libraries of structurally related compounds. A good final distribution of properties within the library is easily implemented with current computational techniques employing filters to monitor *drug-like* and *lead-like* properties.

High-throughput screening libraries are not static entities and therefore need to be continually expanded and updated to meet future screening needs for new biological targets. Unique, newly available libraries from both internal and external sources may be used to augment the diversity and development of the main HTS library. The computational filters used to select external compounds are useful tools applied to internal compound libraries in order to maintain the appropriate balance of diversity and desirable properties in the final library.

Proper library design is crucial, as it can shorten hit identification and validation times. In order to reduce hit validation times, HTS libraries incorporate many small *clusters* of similar compounds. Computational *clustering* by chemical similarity or scaffold, performed prior to or in parallel to screening, allows rapid analysis and selection of promising lead series suitable for follow-up by medicinal chemistry teams. Results of scaffold and molecular similarity based clustering aid in the identification of false negatives through the application of traditional statistical techniques during follow-up of the primary screening hits. Borderline hits and false negatives can be statistically confirmed by comparing measured activities within a cluster. Additionally, increasing the number of clusters improves the chances that good potential lead series are found. Hit priority can be established on membership in possible lead series clusters and frequent hitters rapidly deprioritised.

Techniques in HTS library assembly and analysis

The necessary characteristics for building a useful screening collection are illustrated by describing the design and assembly of AMRI's own Diverse AMRI Screening Library (DASL). DASL is dynamic and comprises 120,000 compounds, providing a diverse representation of available chemical space. The library is constituted from three main sources of compounds: a *diversity selection* of 70,000 commercially available samples from more than 3 million currently available; 10,000 diverse samples manually selected by medicinal chemists from commercial collections and filtered to be particularly lead-like; a 40,000 compounds set selected as small *clusters* from 400,000 samples of AMRI's internal combinatorial libraries which have been recognised as novel compounds and not otherwise commercially available.

Application of computational analysis of chemical descriptors during commercial compound selection is an important means of initial filtering to ensure that compounds identified for inclusion meet both *drug-like* and *lead-like* (MW $<$ 550, cLogP $<$ 5.5 and rotatable bonds $<$ 10) parameters. A screening library need not exclusively contain *drug-like* compounds, however, all series of compounds in the library should be amenable to chemical modifications that improve drug-like properties without extreme effort. DASL was designed to produce *lead-like* and *drug-like* hits. The calculated chemical, physical and structural compound descriptors of DASL reveals about 90 per cent of the compounds satisfy all but one of Lipinski *drug-like* rules, 84 per cent satisfy the Oprea *lead-like* criteria, and a further 6 per cent can be classified as being *fragment-like*. More than half the compounds in the library have molecular weights below 400 and cLogP $<$ 4.

A screening library should extensively sample ring systems and scaffolds for detecting possible lead series in order to maintain *chemical diversity* and *chemical space* coverage. Scaffold analysis of a library ensures the chemical diversity results from changes in the core of the selected sample and from differences at chemical substitution points. A scaffold analysis of DASL shows that it contains almost 7,000 single-ring systems and about 70,000 two- and three-interconnected ring systems (Fig 1). By identifying related scaffolds and ring systems with various substituents, hits from a

Level	Number of Scaffolds		
	DASL	2010 Addition	
	Unique	New	Common
5	1933	136	0
4	13091	2463	19
3	35182	4793	604
2	35145	2090	1364
1	6970	47	262
ALL (1-7)	97071	9529	2249

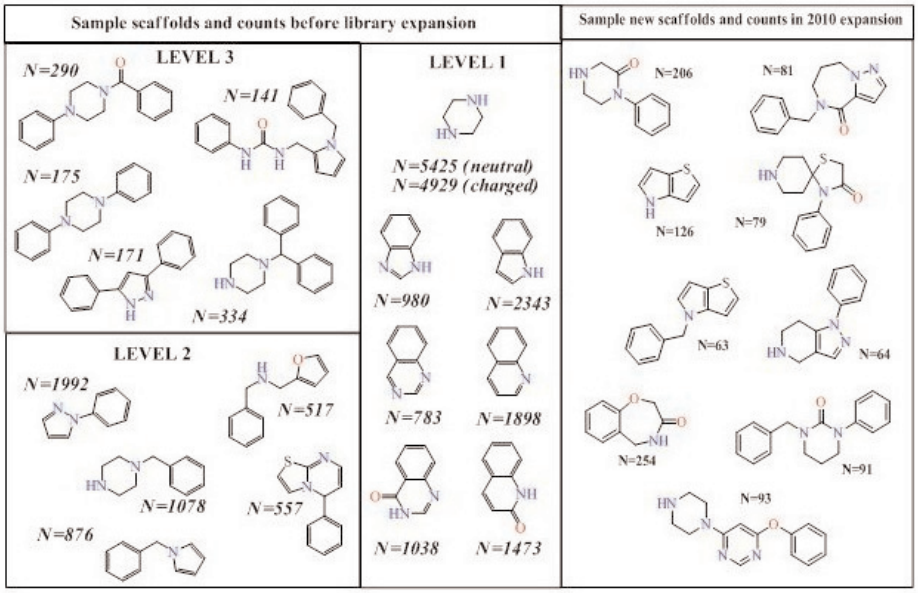

Fig 1. A scaffold analysis indicated that DASL contains almost 7,000 single-ring systems and about 70,000 two- and three-ring systems. In 2010, DASL was supplemented with the addition of more than 12,500 compounds from the AMRI Hungary compound repository. Sample scaffolds are shown for the most common substructures in DASL (left) and the most common new scaffolds from the compounds added to the screening library.

screening library can more rapidly be elaborated during lead optimisation.

Establishing the coverage of a screening library over known *biological space* can be estimated by chemical similarity calculations to a database of biological data. The *biological space* coverage of DASL was calculated by similarity to more than 300,000 compounds with associated biological data in BindingDB (Fig 2). These calculations show that DASL spans a significant number of currently known biologically active chemotypes with 32 per cent of the library showing similarity to known inhibitors of common biological targets. The large number of unclassified structures represents expected opportunities for the discovery of future hits for either known or novel targets.

Libraries used in HTS need to be *dynamically* expanded and updated to meet future screening needs against novel biological targets. Small sets of additional compounds with carefully selected properties continue to enhance library collections. For example, in 2010 DASL was supplemented with the addition of more than 12,500 compounds selected by scaffold analysis from the AMRI Hungary compound repository. Compounds that comprised new chemotypes were added and no further additions were made to chemotypes that were already well-represented (Fig 1). This addition resulted in a total of 9,529 chemotypes that are now accessible in DASL, an increase of about 2,000 chemotypes (compounds with single or two interconnected rings at the core).

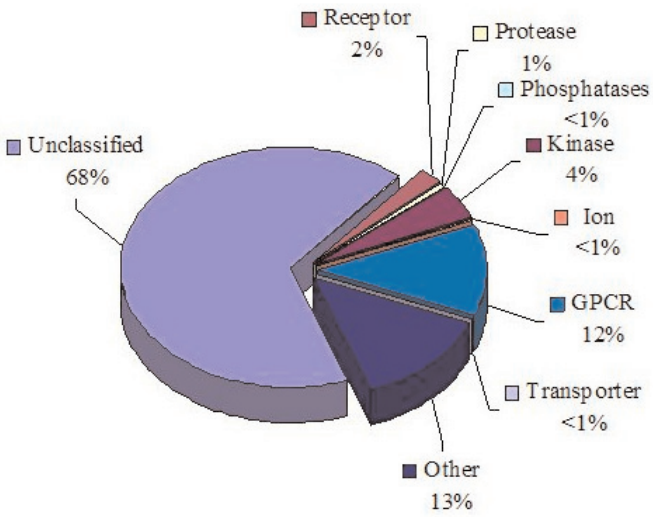

Fig 2. Calculations pertaining to DASL over known biological space indicate that the library spans a significant number of currently known biologically active chemotypes.

Facilitating hit-to-lead exploration

Analogues of hits should be readily accessible to further shorten hit-to-lead time. Many analogues of DASL compounds are readily available as part of AMRI's much larger combinatorial chemistry collection, or can be identified in an up-to-date database of available commercial libraries. The accessibility of these compounds assists in hit validation and provides a quick structure activity relationship (SAR) snapshot critical for shortening the hit-to-lead analysis. Concurrent virtual screening can be performed on internal and external databases in order to identify related analogues for testing. Multiple methods including similarity searching, pharmacophore searching, and ligand-protein docking are routinely used to establish the SAR relationship.

DASL is offered for access in conjunction with AMRI's lead discovery services which embrace in vitro biology from cell culture and protein production to assay development and validation into HTS. Hit-to-lead development of hits is facilitated within the framework of AMRI drug discovery services, which encompasses chemistry and ADMET services.

ACKNOWLEDGEMENT

The authors would like to thank Brian T. Gregg for his assistance in manuscript preparation and helpful discussion.

REFERENCES

For a complete list of references contact the authors.

Further information
Hélène Y. Decornez
Albany Molecular Research, Inc.
Email: helene.decornez@amriglobal.com

Bryan C. Duffy
Albany Molecular Research, Inc.
Email: bryan.duffy@amriglobal.com

Douglas B. Kitchen
Albany Molecular Research, Inc.
Email: douglas.kitchen@amriglobal.com

Web: www.amriglobal.com