Herman Verheij and Bonnie L. Robeson

The Role of Chemical Diversity in Drug Discovery

Abstract

The role of chemical diversity in the drug discovery process needs to be assessed. In recent years the emphasis has been on genomics and biology. Although combinatorial chemistry has increased the number of compounds for screening, this has not increased the number of new scaffolds and new chemical entities. There are still only a few hundred major categories of structural types of compounds marketed as drugs. Today, more companies are analyzing their compound repositories for diversity. There are several methods used to achieve diversity. This article compares the methods of determining diversity and the diversity of compound libraries assembled by various methods. **SPECS** (Rijswijk, The Netherlands) acquires compounds from global sources and shows that this method increases diversity and increases the probability of increasing the hit rate.

Key Words

Chemical diversity, drug discovery process, compound repositories, compound library, global sources, synthetic, organic chemists

Drug discovery continues to be a challenging area. The established large pharmaceutical companies invest millions of dollars to screen tens of thousands of compounds to identify lead compounds. Hundreds of biotechnology companies have emerged using the information garnered from genomics and proteomics. All of these firms have the same goal: to develop treatments for diseases that plague humankind.

In recent years, the emphasis in research has been on developing assays to make use of the increasing information on human genes and biochemical pathways. Less attention has been directed toward the chemistry of the compounds screened.

Why has the number of drugs reaching clinical trial decreased in recent years? Lack of a better understanding of biology is not to blame. The large pharmaceutical companies, in most cases, have inventories of hundreds of thousands of compounds synthesized throughout the years as well as numerous medicinal chemists employed to synthesize even more. In recent years, combinatorial chemistry has promised to provide exponential numbers of compounds for screening, thus increasing the probability of identifying drug candidates. Where are these leads? Why are the pipelines not overflowing with candidate drugs?

> "ALTHOUGH COMBINATORIAL CHEMISTRY HAS INCREASED THE NUMBER OF COMPOUNDS FOR SCREENING, THIS HAS NOT INCREASED THE NUMBER OF NEW SCAFFOLDS AND NEW CHEMICAL ENTITIES."

Combinatorial chemistry and parallel synthesis, although it produces voluminous numbers of compounds, is expanding on a chemical theme using known scaffolds and common substituents. This approach has not provided the exciting leads anticipated.

Chemists were asked to estimate the numbers of different ring structures they had synthesized; the results are shown in *Figure 1*. The green line is what chemists think they make, the blue line is the analysis of synthetic compounds cataloged by large commercial databases, and the red line is the compounds listed in the Comprehensive Medicinal Chemistry (CMC) database. The conclusion is that the organic synthetic chemists are not as creative and diverse as they thought. The figure demon-

Dr. Herman Verheij *is Executive, ChemInformatics,* ***SPECS and BioSPECS B.V.,*** *Fleminglaan 16, 2289 CP Rijswijk, The Netherlands; tel.: +31 70 319 0077; fax: +31 70 319 0011; e-mail: herman.verheij@specs.net. He earned his Ph.D. degree in physical organic chemistry, in the group of Prof. Dr. Jan Verhoeven (Chairman of the European Photochemical Association, and laureate of the coveted Dutch Snellius Medal). He held a postdoc position with Prof. Ernst Sudholter (physical organic chemistry) at Wageningen University (The Netherlands). Dr. Verheij currently leads the ChemInformatics group, which is responsible for ChemInformatical analyses, prediction of compound properties, and software and knowledge base development.* ***Dr. Bonnie L. Robeson*** *is Director of North American Operations,* ***SPECS and BioSPECS B.V.,*** *Columbia, MD, U.S.A. She has worked in the drug discovery and development area for more than 16 years. Before her current position, she was Principal Investigator on a National Cancer Institute contract involving cancer and AIDS drug discovery. Dr. Robeson earned a Ph.D. in biochemistry and a degree in business administration from The Johns Hopkins University (JHU, Baltimore, MD), where she is also an adjunct professor at JHU in the School of Professional Studies in Business and Education.*

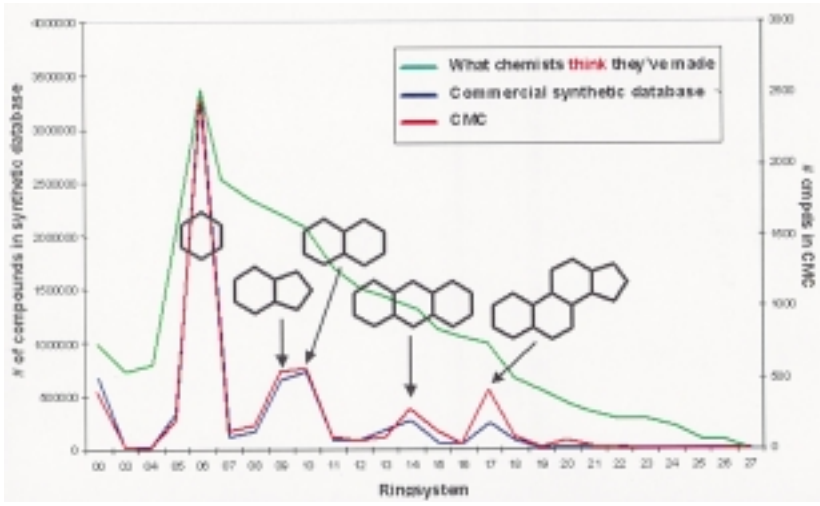

Figure 1 *Drugs and ring systems.*

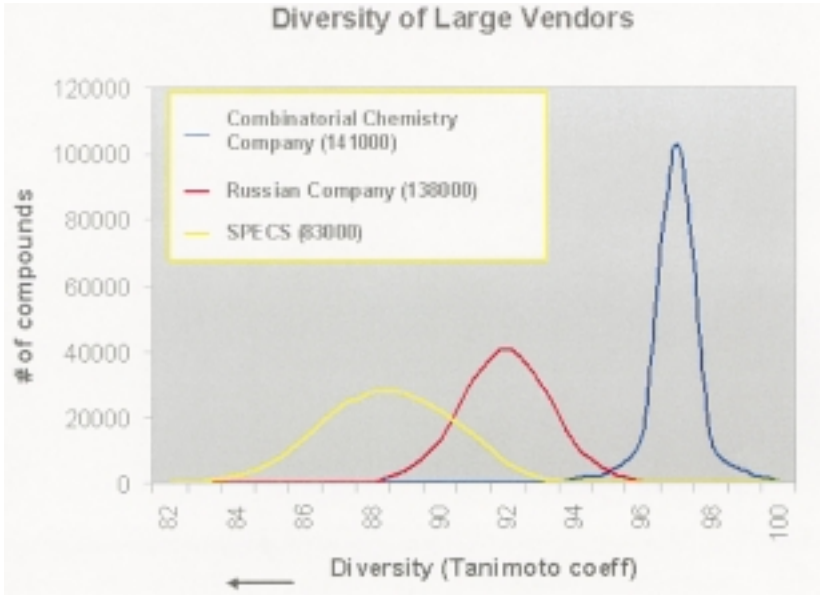

Figure 2 *Analysis of chemical diversity from three sources.*

strates that the compounds fall out in the common ring systems of six-membered rings and common six–six and six–five membered fused rings. There is another peak at the 17-membered ring area due to the large number of both natural and synthesized steroid ring systems.

Has quantity of compounds replaced quality of compounds in the screening programs? Now is the time for chemistry to take the lead in lead discovery programs. The pharmaceutical companies have invested a great deal of money and time on computational programs that predict the best characteristics of screening compounds and select based on the variables. Again, how far has this mind set advanced drug discovery? This line of questioning is not to discredit cheminformatics, but to determine the best use of these tools.

Let us look at the role of chemical diversity in the drug discovery process. There are a few hundred major categories of structural types of compounds marketed as drugs. Many companies analyze their compound repositories for diversity. There are several methods used to achieve diversity. In a compound library, one way is to acquire compounds from global sources. Organic chemists and medicinal chemists select an area on which to concentrate their research, and their students often follow in the same or a closely related area. Synthetic organic chemists in China and Australia have compounds that are vastly different from those in the U.S.

This concept is demonstrated in *Figure 2*. A large pharmaceutical company examined three sets of compounds for diversity—a combinatorial library, a collection from a limited geographical area, and a set of compounds assembled using the global acquisition approach. The greatest diversity is found in the sourcing of compounds from the greatest geographic diversity. Here, the wider the peak and the further to the left the peak, the more chemical diversity. As expected, the combinatorial chemistry (blue line) approach provides the least diversity because a small number of scaffolds are reacted with a limited number of functional groups. The compounds collected (red line) in a defined geographical area are somewhat more diverse, again demonstrating that compound diversity is restricted. The compound collection

> "THE LEAD RATE IS MORE THAN DOUBLE THAT OF A RANDOM SELECTION, PROVING THAT A LINK EXISTS BETWEEN GEOGRAPHICAL (CHEMICAL) DIVERSITY AND BIOLOGICAL ACTIVITY."

(yellow) with the most diversity is acquired from global sourcing of compounds. It is for this reason that the sourcing strategy of **SPECS** (Columbia, MD) focuses on collecting compounds from all over the world.

Figure 3 demonstrates the chemical diversity of libraries sourced using two different strategies. Two 10,000-compound selections are compared: one from global acquisition (**SPECS and BioSPECS B.V.,** [Rijswijk, The Netherlands]) (red markers) and the other from a supplier from the former Soviet Union.

The bulk of the molecules from both sets are evenly distributed over chemical space, but the largest differences can be found at the outer edges of the plot: at the upper-right and upper-left quadrant of the plot, large numbers of compounds (belonging to various different chemical classes) can be found that are present in the **SPECS** set, but

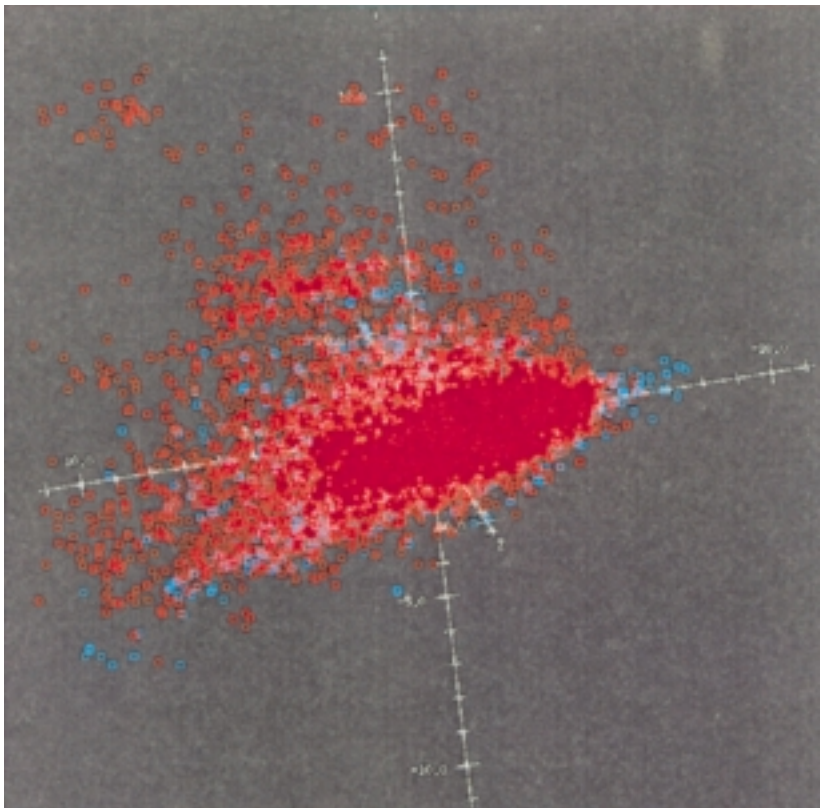

Figure 3 *Principal component analysis of two sets of 10,000 compounds. Two 10,000 compound sets displayed in chemical space (as defined by the first three principal components, upon a principal component analysis of 26 SAR descriptors). Red markers:* ***SPECS and BioSPECS*** *compounds; blue markers: compounds from a supplier from the former Soviet Union.*

Table 1

Results of assaying selected sets of compounds

Selection method	No. of compounds tested	No. of leads	Lead rate (%)
Random selection	10,000	6	0.06
Substructure search	8000	6	0.08
Targeted selection	4000	5	0.13
World Diversity Set*	10,000	14	0.14

*The World Diversity Set (WDS) is a collection of 10,000 compounds selected from around the world with lead-like features. The compound library is available from **SPECS and BioSPECS** for high-throughput screening.

cannot be found in the other set. Indeed, most of the molecules contain compounds that can be obtained from at least two, but sometimes from up to 10, different suppliers of screening compounds. The interesting part of a vendor's collection, therefore, is that which is unique to that vendor. The figure clearly demonstrates that a worldwide sourcing strategy leads to an increase in the number of such unique structures and structure classes, thus linking geographical diversity with chemical diversity.

It has always been the assumption that chemical diversity will lead to diversity in biological activity. Although this does sound logical, evidence on biological activity proves this point. Unfortunately, for a long time it was not possible to obtain such figures, simply because the life science industry does not disclose the results of assaying to third parties. **SPECS** has been able to establish collaborations with pharmaceutical and biotechnology companies to which it delivers compounds free of charge; in return it receives the assay results.

This enables a comparison of an entire array of selection methods, ranging from simple substructure-based selections to elaborate software-driven selections based on predicted biological activity. Also, the types of subsets of compounds from which these selections are made can be varied. The latter method was employed to investigate whether there is indeed a link between geographical diversity of compounds and their biological activity in two radically different therapeutic areas.

Table 1 displays the results for various selection methods. Note that the results for the two different therapeutic areas are combined so that one can obtain a broader general view of the differences that are a result of the selection methods.

The random selection taken from the compound collection serves as a kind of baseline for a nonbiased selection, and reflects the general biological activity of the compounds. As expected, after a substructure search (for substructures belonging to chemical classes known to be active in the therapeutic areas under investigation) the lead rate improves, but not by a large margin. A significant increase in activity, however, is found when a targeted selection is made, i.e., when using a software program that predicts biological activity, based on a training set of known biologically active molecules. However, this is a statistical process (the program predicts the chance [0–100%] that a molecule will be active in a certain therapeutic area); thus more testing is needed to establish the exact increase in lead rate yielded by the program. (The average lead rate will probably be between 0.10 and 0.20, but lead rates of 0.40 have also been achieved using a targeted approach.) Finally, the last selection, the World Diversity Set, is a selection of molecules belonging primarily to the compound classes found in the chemical space edges shown in Figure 3 (i.e., the diversity of this set is enriched).

The lead rate is more than double that of a random selection, proving that a link exists between geographical (chemical) diversity and biological activity. The exact chemical reasons are still being investigated for this apparent link, and, although a possible explanation exists, larger numbers of compounds are being assayed in order to confirm this finding.