Pacific Symposium on Biocomputing 5:552-562 (2000)

Compound Acquisition Strategies

James B. Dunbar Jr.
Parke Davis Pharmaceutical Research Division,
Warner Lambert Company
2800 Plymouth Road
Ann Arbor, Michigan 48015, USA

Compound acquisition has always been a very important component in the discovery and development of new biologically active entities. With the rapid advances in high throughput screening coupled with the ever-decreasing time requirements for the discovery phase, the number and quality of compounds screened is of great importance. This presentation will discuss some of the techniques and processes that can be used in compound acquisition.

1 Introduction

1.1 Overall process

Compounds are acquired usually by purchase, but they can also be acquired by some exchange mechanism. In either case, the goal is to obtain compounds that are different from those compounds already in the in-house collection and have at the same time characteristics that are appropriate for the field of interest. In this introduction, I will touch briefly on two time-consuming issues and spend some additional time on the "appropriate characteristics" issues.

In the selection process it is usually very straightforward to find compounds that are different, but defining the filters to remove inappropriate compounds is much more complicated and time consuming. I consider it absolutely critical to have in place a methodology to remove inappropriate compounds before undergoing any acquisition of actual compounds that will be used in any form of biological testing. It is, in my opinion, very much like the "garbage in - garbage out" concept, but here it is more likely "garbage in - garbage out hundreds of times" complicating the analysis of everything that compound is involved in.

Another time-consuming process is the contractual agreements between the parties in the acquisition. Unless the purchase is from a chemical supplier under a nonpatented, non-exclusive use basis, it has been my experience that the process of arriving at a mutually agreeable contract is the single most time consuming part of each *individual* acquisition. I have found it useful to have a set of generic contracts for purchase or exchange that can be used early on in the negotiations as a means to efficient discussions with a concrete proposal of how one could write the contracts. Typically the contracts would also address items such as the method of shipment and even containers and barcodes that would be used.

1.2 Filtering

As stated above, I consider the filtering process crucial to long term, effective compound management and data analysis processes. As a start, one should be able to filter on an acceptable range of calculated physical properties. Software is readily available from a number of sources to provide this type of functionality. Typically this would include some estimate of the partition coefficient between water and an organic solvent; examples would include CLogP[1] or MlogP[2]. The fragment based approaches such as CLogP while being more accurate, may not have some of the fragments necessary to describe all of the compounds of interest. The atom-based approaches will at least provide some estimate for all of the compounds. Estimates of volume such as calculated molar refractivity (CMR [1]) and molecular weight are also used.

Another metric which has gained popularity is whether a compound passes or fails the "Pfizer Rule of 5"[3]. This will provide an estimate as to whether a compound may or may not have pharmacokinetic problems. This can be applied as a hard cut in an acquisition process, thereby obtaining only those predicted "okay" because we are acquiring compounds for future use in an unspecified area. This is in contrast to work in a lead optimization process where the compound synthesized may be for answering specific questions related to mechanistic or other aspects of therapeutic discovery and not envisioned as becoming a clinical candidate. Under these circumstances, the "Rule of 5" need not be as stringently applied.

A very important and initially the most time consuming step is to determine how to identify compounds which have unwanted functionality. This involves two parts, one is deciding on what software package to use and the second is to determine what is considered unwanted functionality and cast that information into a query or queries for the chosen software. The software, in general, must be able to run on the order of 100 to 200 queries on several hundred thousand compounds in a reasonable amount of time, i.e. in an over night job in a scriptable fashion.

For example, Parke-Davis has developed over the course of several years a set of filters to remove unwanted functionality that use the Daylight[1] toolkit. These are a series of 268 SMARTS targets in a file representing the unwanted functionality, unwanted fragments of molecules. What is defined as unwanted depends upon the situation, field of interest, and can include reactive species, intercalators, toxic compounds, compounds too similar to those in literature of a given field, etc, etc. This set is then used to partition the input dataset into two files, those that have passed through the filters and those that match a SMARTS target and therefore fail. Only those compounds, which have passed the filters, would be taken to the next set of analysis tools. The philosophy of this method is to remove compounds that have specifically unwanted functionality, rather than try to define what is wanted and remove the rest.

1.3 Sources of Compounds

The availability of compounds that would be suitable for screening has dramatically increased in the past few years with the advent of combinatorial chemistry. An entire industry has grown out of this technology. Companies can supply peptides, peptidometic compounds, and small molecules in tremendous numbers. Here one would be interested in whether or not the compounds are on an exclusive or semi-exclusive basis. Quality control issues also come into play, because of the possibility of having reactive starting materials rather than the final product in the wells. One must also consider what legal restrictions, software and/or hardware aids that will be associated with the analoging process when following up a hit from the library. Another question is the coverage in chemistry space of the individual libraries and how will it merge with existing libraries. Coverage issues are very subjective and probably very specific to the existing corporate culture.

Additional sources of compounds for purchase keep appearing on a regular basis. Companies have been formed to collect and sell compounds gleaned from academic institutions, both in the US and abroad. Individual academic groups are providing compounds for sale. Companies that have historically provided fine chemicals for use in synthesis are also advertising compound collections for screening applications. Yet another source of compounds is an exchange between two companies. With the increasing use of high throughput screening, there are now a wider variety of places to acquire compounds for the screening efforts.

2 Techniques Used in Acquisitions

2.1 Example acquisition

Although the focus will be on compound acquisition for the purpose of increasing the molecular diversity of in-house databases, it can also be applied to the selection of electronic databases of compounds. In either case, the assessment will probably be based on the comparison of electronic databases of structures. Generally the goal is to obtain compounds or structures that differ from those held. This precludes redundant testing (by actual screening or database queries) and maintains efficiency while increasing diversity. But just being different is not sufficient; also the compounds to be acquired should be relevant to the area of research. For example, the acquisition of only those compounds that are not already available in-house, yet possess a high probability of pharmaceutical relevance. What constitutes pharmaceutical relevance is an area of intense interest, but as yet without any definitive answers.

An example of a clustering process applied to structure selection has been described by Shemetulskis; *et al* [4] Here additional electronic structures were to be selected for use in drug design applications. The goal was to select approximately 100,000 broadly representative structures from a CAST3D [5] database, comprised of roughly 400,000 structures. At the time this study was performed, analysis of several hundreds of thousands of compounds by clustering techniques remained computationally challenging. Due to these computational constraints, the process was broken down into two stages.

In the first step, the Daylight Clustering Package was used to reduce the CAST3D dataset to a more manageable size. A near-neighbor calculation was done on the full dataset from CAS. The Jarvis-Patrick [6] clustering technique with the Daylight default fingerprints was used to analyze the data. The need-versus-near level was chosen to provide the requisite number of cluster "centroids" and singletons, which were then extracted to form a smaller, but representative dataset of compounds. The cluster centroid is defined in the Daylight Cluster Package as the structure in the cluster "which explains the most variance of the cluster"; the one nearest the centroid. A more detailed description is provided in the manual.

The second step was to compare the more-manageable CAST3D dataset with the in-house database. To facilitate the comparisons, the structures from both databases were labeled as to origin, then combined, and the nearest-neighbor calculation redone. The data was clustered using the Jarvis-Patrick technique, and then the resulting clusters were examined for label content. Clusters containing only CAS structures and CAS singletons were given the highest priority for acquisition. Clusters that contained only a very small percentage of in-house structures were assigned a reduced priority, and those clusters that contained a relatively high percentage of in-house structures were discarded.

In an acquisition analysis where actual sample is to be obtained, the selection process should be subjected to more rigorous criteria than that used in the CAST3D selection. For example, no effort was made to remove reactive entries. In the following examples, both compound purchase and compound exchange will be explored as a means of acquisition.

If one is to acquire compounds through an exchange mechanism, a prerequisite is to determine what subset of your organization's database is available for exchange. This can be accomplished by numerous means so that the requirements of both organizations can be met. Many criteria may be used in this decision, such as: Is there enough sample available? Can the compound be exchanged without additional entanglements, for example was the compound synthesized in-house or was it acquired from another source? Criteria based on the date of synthesis can be used, such as only exchanging those compounds that have not been synthesized within the last five years. If all other criteria have been met, has the compound shown up as an active in any current assay? If such criteria are used, the set to be exchanged must be re-evaluated on a regular basis.

The next step in the acquisition process is to pre-screen the acquisition dataset if at all possible. This will allow clustering into two sets; those to be analyzed in detail and those not to be considered for acquisition based on simple filtering techniques such as unwanted reactive functionality. This presupposes that the prescreening characteristics are known, or that these characteristics can be calculated based on existing information. If not, then one could apply a more generalized prescreening step such as the centroid method of Turner *et al* [7] to determine if the dataset as a whole will increase the diversity of the in-house compound set. In this method, a single number that describes the internal similarity of the possible acquisition dataset can be used to compare with another single number from the in-house dataset. This will provide a rough estimate of increase or decrease of diversity in the in-house dataset if it were to be combined with the acquisition set. This method can also be applied to single compounds. It is also very useful in compound exchange processes where minimal information release is also a consideration. The centroid method could also be used if one were faced with several independent sources and was limited to choosing only one for acquisition due to budgetary considerations. Another mechanism for minimal information screening is to initially exchange only fingerprint information, which, if of a hashed type and folded [6], is virtually impossible to convert back into the exact original structure. One can compare the acquisition dataset to the in-house dataset and make a decision on whether to proceed further and actually obtain a subset of structures for more precise selection of compounds.

If structures are available, several kinds of prescreening are possible. One can filter on calculated physical or topological properties. Acceptable ranges of molecular weights, calculated logP's, and calculated molecular refractivity's can be established. In ClogP filters, one could reject those whose ClogP fell outside of a given range, retaining those within the range along with those for which a reliable value could not be calculated. Compounds with unwanted chemical functionality, such as reactive functionality, metals, and pure hydrocarbons, can be removed. Another example would be to discard the classes of compounds that were found to be promiscuous in biological screening; that is, those compounds that show up in numerous mass screening assays. This can be extended to compounds with undesirable pharmacological properties and compounds predicted to be teratogenic, mutagenic, or carcinogenic. Care must be taken in designing these filtering mechanisms so as to suit the entire organization's needs now and, if possible, in the future. A "junk compound" may turn out to be the best initial hit from a different screen.

Once the prescreening steps have been accomplished, more rigorous comparisons of the possible acquisition structures and those of the in-house dataset are done. As with the CAST3D acquisition, one very common mechanism is to cluster the data based on the Tanimoto coefficient (Tc)[1]. The most common mechanism for arriving at the Tc is via comparison of the fingerprints of the structures in the acquisition set with those of an in-house set. Brown *et al.* [8,9], have published two reports examining the various fingerprinting techniques, noting which of the various fingerprint types were best suited to their study. They also discussed at what level of similarity one would expect the two structures to have the same biological action. Based on the work by Brown and colleagues, there is now a comparison of several fingerprinting techniques and a mechanism for a better understanding of what is available. Brown *et.al.*, found that the Maccs [10] fingerprint technique, based on turning on/off a bit according to the existence or not of a pre-defined fragment, worked best for them. Daylight fingerprints (fully hashed) and Tripos [11] fingerprints (hashed into particular regions) also produced acceptable results.

Once two databases are fingerprinted, a fingerprint comparison calculation is run. Brown, *et al.*[9], determined that a Tc of 0.85 or greater resulted in selecting a high percentage of compounds that had the same biological action. Tripos [12,13] has published a method for selecting compounds for synthesis by combinatorial or parallel methods to produce a library of diverse compounds for lead discovery. They have also concluded that 2D fingerprints are a valid metric for use in compound selection and that the cutoff is also about 0.85 similarity. Therefore, based on these studies, one would then choose compounds that had less than 0.85 Tc (similarity) from any compound already held in-house.

The acquisition database can then be sorted into an ordered list, whose primary key is the increasing value of Tc with the reference database (in-house) and the secondary key being that of the Tc of compounds within the acquisition database itself. This will allow one to select the compounds that are the most different from the in-house set *and at the same time* select compounds that are the most diverse regarding the acquisition set. This helps to preclude the situation where the compounds to be acquired are very different from the in-house dataset, but with a large percentage of them being very similar to each other.

One could then make a determination as to how dissimilar a compound within the acquisition set must be, relative to another compound within the acquisition set, for final selection. (Typically only the duplicates are removed. If one selected the compounds such that each compound also had at least one very similar compound, there would be a possibility for some immediate structure-activity relationship data and some limited, initial confirmation that the class of compound is a real hit and not a false positive.) An additional benefit of an ordered list is that the most different compounds can be requested first, to obtain the most internally-varied set early on in the compound exchange process.

To recap this example, the dataset is first either prescreened to filter out unwanted compounds, or a "blind" (without the *actual structures*) evaluation of the acquisition database is done relative to the in-house database. This is followed by the Tc selection process and the generation of an ordered list based on the Tc of the acquisition set to the in-house set as the primary key and the secondary sort key being the Tc internal to the acquisition set. From this ordered list, the final set for acquisition is selected.

2.2 Other Techniques

With the advent of combinatorial chemistry and the subsequent large numbers of compounds that are possible, selection of broadly representative sets of compounds with a particular chosen chemistry should improve the efficiency of high volume screening. The question is how to choose that set. One aspect of the debate is whether product diversity is better selected in product space or reactant space. While the problem of selecting in reactant space is more tractable; I and others believe that it is best to select from product space, whether the set is for lead generation or lead optimization.

One example of the combination of cluster-based selection in product space and actual synthesis is the Tripos/Panlabs[14] Optiverse library of compounds. Since the Optiverse library is commercially available, it could provide one example, of many commercially available compound libraries, that can be tracked for performance over time in a number of assays in a number of different hands. Some of the selection mechanism for this library has been published. Patterson[12] *et al.* and Cramer[13] *et al.*, have described work on validating descriptors for neighborhood behavior in the area of lead generation. Neighborhood behavior refers to the concept that compounds in the same cluster (neighborhood) should have the same action. Cramer [15] *et al have* recently published a companion paper on the Tripos proprietary ChemSpace [14] technique which utilizes the descriptors and neighborhood analysis previously described.

Once identified, the appropriate descriptors could be used in a cluster analysis to select molecules for subsequent testing or synthesis. In their studies, the main focus is on the construction of lead generation libraries from combinatorial synthesis, specifically to the Optiverse library. For this purpose they have determined that 2D sidechain fingerprints in combination with topomeric molecular fields performed the best. Topomeric fields are steric fields generated for a molecule whose conformation is chosen by a rule-based method. 2D fingerprints of the whole molecule were found to be only somewhat useful for this purpose. Other descriptors such as connectivity indices, logP, molar refractivity and random numbers were not useful at all.

While these descriptors seem to be appropriate for combinatorial libraries, the applicability to libraries of dissimilar compounds typically found in pharmaceutical companies is not so straightforward. There is the alignment problem for the topomeric fields and the question of what constitutes a sidechain in two unrelated molecules. This difference is highlighted in the paper by Matter and Lassen [16]. They have found that 2D fingerprints alone are very useful for analyzing global diversity in a database such as the IndexChemicus (IC93), while the 2D fingerprints and 3D molecular shape are good for local similarity such as that found in combinatorial libraries. In the analysis of the IndexChemicus, the authors report that 2D fingerprints worked well with both hierarchical cluster analysis and with maximum dissimilarity techniques. In another paper by Brown and Martin [17] examining the information content of fingerprints, it was found that MACCS structural keys appears to encode a great deal of information relevant to the interactions found in ligand-receptor binding - "Our results suggest that by making an appropriate choice of structural descriptor, and of clustering method where applicable, the ligand-receptor binding forces can be accounted for without having to explicitly code them in the descriptor."[17]

Many other approaches can be used in diversity assessments for the selection of compounds for high volume screening or combinatorial libraries. The one by Tripos utilizes sidechain fingerprints in combination with a field-based technique. Another approach is the three- or four-point pharmacophore analysis as exemplified in the Chem-X [18] software. Here the molecules are described by the hydrogen bond donor atoms, hydrogen bond acceptor atoms, hydrophobic centers, positively-charged centers, and aromatic ring centers. The molecule is then subjected to a conformational expansion and the bits are set in the pharmacophore key. The subset of molecules that have the most of the overall pharmacophore set are then chosen. Additional criteria such as flexibility, number of pharmacophores contained in a molecule and overlap of pharmacophores are used to further prune the selection.

Yet another type of approach is the use of Pearlman's BCUT [19] values. This approach utilizes three classes of matrices. One class has diagonal elements based on the atomic charge; the second matrix diagonal is a representation of the polarizability, and the third, based on H-bond abilities. The off-diagonal elements can be composed of topological information of 2D connectivity and/or 3D information. These matrices are then used to identify a "chemistry space" as defined by the library. This chemistry space can be used to identify areas where there are only a few compounds or none at all. This would provide a mechanism not only of selecting a broadly representative set of compounds for assay purposes, but also a compound acquisition mechanism to fill in the areas which are void or poorly populated. This can be used in a "blind" (structureless) analysis by exchanging the voids files and examining the results in the "fill" format. This "fill" format will provide a ranked order of the compounds that add to the diversity. In using the voids file, a few compounds that fill cells that are different from the in-house collection, but are similar to each other may provide some trends and limited conformation of the class of hit. Paul Menard *et.al*. [20] have utilized DVS [19] and their non-linear binning extensions to DVS in conjunction with other methods to analyze and compare databases of compound collections.

Parke-Davis has made extensive use of the DVS[19] suite of techniques in the comparison and analysis of compound databases, particularly of combinatorial databases. In terms of selecting a diverse set of molecules from a database or combination of databases, DVS selects a structurally diverse set of molecules without any observed bias in molecular weight, ClogP[1], number of rotatable bonds, CMR [1], and heteroatom composition. DVS is being used as the primary selection tool for a current compound exchange program with another company.

Molecular Simulations also provide a suite of tools for library comparison and diversity analysis. As an example, C^2-LibCompare [21] and C^2-Diversity [21] can be used for the selection of libraries and individual compounds. I have included in the reference the MSI web site [21] where these and other modules are completely described. These two modules will pull in parts of several other modules for use, such as the C2-QSAR+[21] and Catalyst/SHAPE [21] modules for descriptors. In addition to those modules already discussed, Tripos also provides techniques for comparing and selecting compound sets, examples include the Selector [14] and ChemEnlighten [14] modules along with other clustering and analysis routines for use in the selection and filtering of compounds. Here again the Tripos web site[14] is provided in the reference for detailed information on these and other modules.

3 Conclusion

Compound acquisition is a very important task associated with chemical discovery; providing additional screening possibilities and potential speedups in the discovery process. The "better" the selections, the more useful the compound collection will be; yet care must be taken to obtain compounds that are relevant to the field of interest and minimize the number of compounds that are predicted to be problematic. In this manner, one can maximize the efficiency of time and minimize the expense involved in discovery process now and in the future.

Acknowledgements

I would like to acknowledge the considerable number of people who have contributed to the compound acquisition process that I have discussed. Dr. Bonnie Dunbar has written several custom programs for use in the various stages of compound acquisition and Mark Duffield for the SMARTS filtering code. The medicinal chemists who have, over the years, reviewed structures and provided feedback on the construction and suitability of SMARTS filters: Dr. Christopher Bigge, Dr. Alexander Bridges, Dr. Robert Kennedy, Dr. Michael Rafferty, Dr. Bruce Roth, Dr. B. K. Trivedi, and Dr. Lawrence Wise. I would like to thank Dr. Christine Humblet for critical reviews and scientific discussions on the acquisition processes. I would like to thank Prof. Peter Willett for his substantial collaboration in Parke-Davis' first compound exchange and the scientific discussions and contributions in the subsequent years. I would like to thank Prof. Robert Pearlman for the many scientific discussions over the years regarding science in general and compound acquisition in particular. John Brussolo, Tammy Grant, and Lori Emery have helped in the selection of PD compounds available for exchange and handled the exchange and registration of the physical samples going out of and into Parke-Davis.

References

1 Daylight Chemical Information Software, Daylight Chemical Information Inc., 27401 Los Altos, Suite #370, Mission Viejo, CA 92691. info@daylight.com http://www.daylight.com/

2 I. Moriguchi, S. Hirono, I. Nakagome, and H. Hirano, Chem. Pharm. Bull. 42(4), 976 (1994)

3 C. Lipinski, F. Lombardo, B. Dominy, and P. Feeny, Adv. Drug Delivery Rev, 23(1-3), 3 (1997)

4 N. Shemetulskis, J. Dunbar Jr., B. Dunbar, D. Moreland, and C. Humblet, J. Comput.-Aid. Mol. Design, 9, 407 (1995)

5 Chemical Abstract Service, 2540 Olentangy River Road, P.O. Box 3012, Columbus, Ohio, USA, 43210-0012.

6 R. Jarvis and E. Patrick, IEEE Trans. Comput., C-22, 1025 (1973)

7 D. Turner, S. Tyrell, and P. Willett, J. Chem. Inf. Comput. Sci., 37 18 (1997)

8 R. Brown and Y. Martin, J. Chem. Inf. Comput. Sci.,36, 572 (1996)

9 R. Brown, M. Bures, and Y. Martin, Similarity and cluster analysis applied to molecular diversity. American Chemical Society Meeting, Anaheim, CA., USA, COMP3, (1995)

10 Maccs II, Molecular Design Ltd., 14600 Catalina St., San Leandro, CA.,USA, 94577

11 Unity manual, Tripos Inc., 1699 S. Hanley Road, Suite 303, St. Louis, MO., USA, 63114

12 D. Patterson, R. Cramer, M. Ferguson, R. Clark, and L. Weinberger, J. Med. Chem., 39, 3049 (1996)

13 R. Cramer, R. Clark, D. Patterson, and M. Ferguson, J. Med. Chem., 39, 3060 (1996)

14 Tripos Inc., 1699 S. Hanley Road, Suite 303, St. Louis, MO., USA, 63114 http://www.tripos.com/ and Panlabs Inc., 11804 North Creek Parkway South, Bothell, WA, USA, 98011.

15 R. Cramer, D. Patterson, R. Clark, F Soltanshahi, and M. Lawless, J. Chem. Inf. Comput. Sci. 38, 1010, (1998)

16 H. Matter and D. Lassen, Chim. Oggi, 14, 9, (1996)

17 R. Brown and Y. Martin, J. Chem. Inf. Comput. Sci., 37, 1 (1997)

18 K. Davies and C. Briant, Combinatorial Chemistry Library Design using Pharmacophore Diversity, Network Science, http://www.awod.com/netsci/Science/Combichem/feature05.html

19 R. Pearlman and K. Smith, Perspectives Drug Discovery Design, 9, 339, (1998)

20 P. Menard, J. Mason, I. Morize, and S. Bauerschmidt, J. Chem. Inf. Comput. Sci., 38(6), 1204, (1998)

21 Molecular Simulations Inc., 9685 Scranton Road, San Diego, CA, USA 92121
http://www.msi.com