Proceedings of the 50th ASMS Conference on Mass Spectrometry and Allied Topics, Orlando, FL, June 2-6, 2002

Targeted Analytical Evaluation of a Large Small Molecule Chemical Database

James A. Kelley[1], Christopher C. Lai[1], Joseph J. Barchi[1], Marc C. Nicklaus[1], Johannes H. Voigt[1], Lynne Anderson[1], Nancy M. Malinowski[1] and Nouri Neamati[2]

[1]Laboratory of Medicinal Chemistry, National Cancer Institute, NCI-Frederick, Frederick, MD 21702
[2]School of Pharmacy, University of Southern California, Los Angeles, CA 90089

Introduction

The National Cancer Institute (NCI) Chemical Database consists of a collection of over half a million unique structures which have been assembled over the past 45 years from a wide variety of sources[1]. Since little constraint was placed on the types of structures obtained, this database represents an extraordinarily eclectic collection of organic compounds. Almost every natural element in the periodic table is represented. The open or publically available portion of this database contains over 250,000 structures, of which approximately 60% are accessible as actual compounds. This makes the Open NCI Chemical Database the largest freely available, public domain chemical database in existence. A recent comparison of eight large chemical databases has also shown that the Open NCI Database contains by far the greatest number of structures that are unique to it, with approximately 200,000 structures not found in any of the other analyzed databases [2]. Additionally, a data enhanced version of the Open NCI Database, that is accessible by a Web-based, graphical user interface (http://cactus.nci.nih.gov), has recently been implemented to allow full and rapid searching by numerous criteria [3]. The Open NCI Database is thus a unique and valuable resource for testing, for drug development and for chemical information applications.

Our laboratory has used this database extensively for 3-dimensional pharmacophore searching to identify potential compounds for subsequent screening against various molecular targets [4]. The validity of this approach for identifying new lead compounds for drug development depends on the selected compounds possessing the stated structures. Since no quality testing is performed by NCI on any supplied sample because of the prohibitive expense for a database of this size, we undertook a structural and purity evaluation of several sets of compounds selected through pharmacophore searching using HIV-1 integrase as a target. An NMR and mass spectral analysis of these compound sets (n = 166) indicated 40% were not the indicated structure or had unacceptable purity. Since we wondered whether this result was characteristic of the database as a whole or unique to the structures selected by pharmacophore searching, we decided to undertake a targeted analytical evaluation of the structural integrity and purity of compounds in the Open NCI Database.

Methods

Database structures were randomized and a clustering analysis was conducted sequentially with randomly selected compounds (cluster seeds) using a similarity index of 0.55. A minimum cluster set size of 150 was chosen to identify 298 clusters which comprised 85.8% of the open NCI database. Previously analyzed compounds represented members of 72 clusters. For the remaining 226 clusters, the cluster seed, or the next similar compound, from each set was obtained for analytical evaluation. Positive and/or negative FAB/MS in appropriate matrices and 400 mHz ^{1}H NMR were employed for the initial structural analysis. Compounds were ranked as good, questionable or unacceptable based on spectral information. A 50-compound subset of the cluster seeds was also selected for future analysis by MALDI/MS. For those compounds identified as homogeneous yet of wrong structure, additional structural information will be obtained by accurate mass measurement and ^{13}C NMR analysis in order to assign a correct structure.

Results

Partial to complete analysis of 296 cluster seeds or their equivalents has been completed. Of these compounds, 58.4% are ranked as good, 10.5 % as questionable, and 31.1% as unacceptable. Designation as good indicates the structure is consistent with spectral data and the apparent purity is 80% or better. Ranking as questionable signifies the desired structure appears to be present but that the sample contains appreciable (>20%) contaminants or decomposition products. Classification as

unacceptable denotes there is no spectral evidence for the presence of the desired compound or contaminants exceed 50% of the sample. A good correlation is observed between mass spectral and NMR data for structural integrity and sample homogeneity, although neither technique alone is applicable to all compounds. No obvious correlation between compound ranking and molecular weight (Figure 1) or NSC number (database accession order)(Figure 2) has been observed. The combination of mass spectrometric and ^{1}H NMR analysis has also allowed the tentative assignment of a corrected structure in several instances where the sample has been homogeneous (Figure 3).

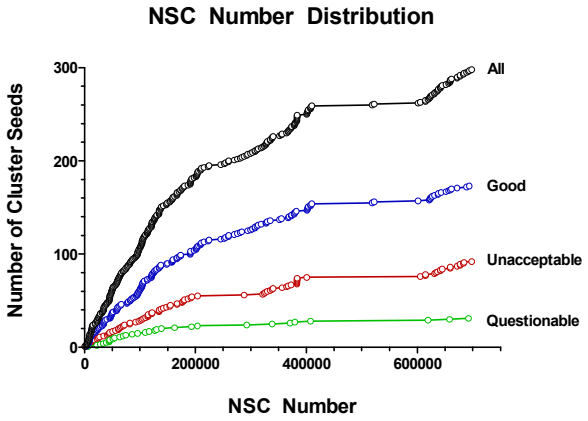

Figure 1. Compound distribution with respect to database accession order.

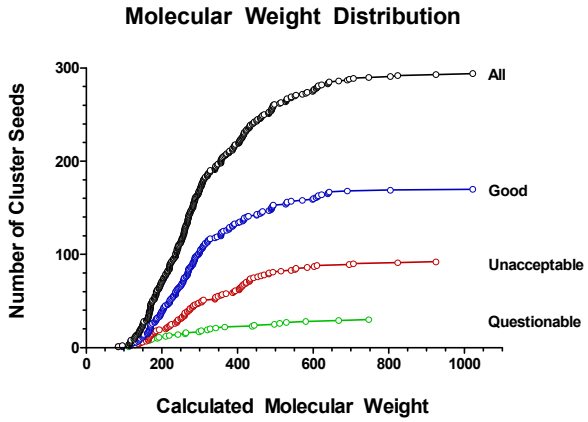

Figure 2. Compound distribution by molecular weight of database structure.

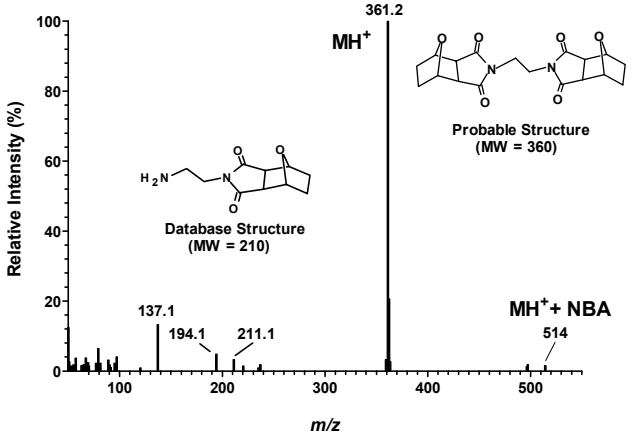

Figure 3. Scan-averaged, positive ion FAB spectrum of NSC 191878 representing a cluster with 214 members. The matrix background has been computer subtracted. The proposed structure has two axes of symmetry which correspond to the observed 1:1:1:2 proton signal measured by NMR.

Conclusion

NMR and MS analysis combined with the appropriate clustering techniques provides a manageable means of evaluating the quality (□5%) of a representative sample of large chemical database.

References

1. Milne, G.W.A.; Nicklaus, M.C.; Driscoll, J.S.; Wang. S. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1219-1224.
2. Voigt, J.H.; Bienfait, B.; Wang, S.; Nicklaus, M.C. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 702-712.
3. Ihlenfeldt, W.-D.; Voigt, J.H.; Bienfait, B.; Oellien, F.; Nicklaus, M.C. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 46-57.
4. Chen, I.-J.; Neamati, N.; Nicklaus, M.C.; Orr, A.; Anderson, L.; Barchi, J.J., Jr.; Kelley, J.A.; Pommier, Y.; MacKerell, A.D., Jr. *Bioorg. Med. Chem.* **2000**, *8*, 2385-2398.