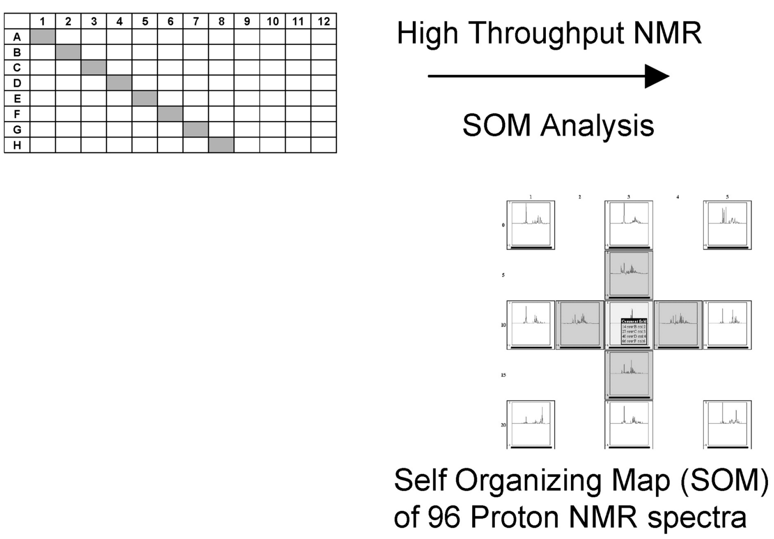
Article

Automated Analysis of Proton NMR Spectra from Combinatorial Rapid Parallel Synthesis Using Self-Organizing Maps

Sandeep Kalelkar, Ernst R. Dow, John Grimes, Matt Clapham, and Hong Hu

J. Comb. Chem., **2002**, 4 (6), 622-629• DOI: 10.1021/cc020031l • Publication Date (Web): 17 September 2002

Downloaded from http://pubs.acs.org on March 20, 2009



Journal of Combinatorial Chemistry is published by the American Chemical Society.
1155 Sixteenth Street N.W., Washington, DC 20036

Automated Analysis of Proton NMR Spectra from Combinatorial Rapid Parallel Synthesis Using Self-Organizing Maps

Sandeep Kalelkar,*,[†] Ernst R. Dow,[‡] John Grimes,[†] Matt Clapham,[†] and Hong Hu[†]

*Eli Lilly and Company, Sphinx Laboratories, P.O. Box 13951,
Research Triangle Park, North Carolina 27709, and Information Technology,
Lilly Research Laboratories, Indianapolis, Indiana 46285*

Received April 29, 2002

It is now quite routine to acquire proton NMR spectra of compounds in 96-well plates prepared in a rapid parallel synthesis fashion using a flow-NMR automation setup. However, the analysis of 96 NMR spectra obtained in this manner is often laborious and painstakingly slow. We have developed a new, automated method for rapidly analyzing 96 NMR spectra of compounds synthesized in an 8 × 12 matrix using self-organizing maps (SOM). This unsupervised neural network is capable of clustering together NMR spectra containing a common pattern of −R groups and identifying outliers from within such clusters. Analysis of these outlier spectra can quickly help indicate the presence of undesired products, impurities, starting materials, and other unexpected errors in a 96-well plate synthesis by focusing the chemists' attention on the aberrant NMR spectra. Thus, SOM can be a valuable tool in performing efficient quality control on combinatorial libraries.

Introduction

The evolution of combinatorial chemistry as a drug discovery enabling technology over the past few years has resulted in a tremendous increase in the need for high-throughput analytical methods to keep pace with the number of samples required to be analyzed. NMR spectroscopy is now capable of performing high-throughput spectral data acquisition on compounds in a combinatorial library. By use of a flow-NMR setup such as VAST, it is possible to routinely acquire NMR spectra on hundreds of samples a day.[1] Automated flow-NMR is particularly relevant in the analysis of combinatorial plates because of the direct sampling off a plate and subsequent near-complete recovery of the sample into the original well. This flow-NMR paradigm proves to be extremely convenient in conjunction with other analytical techniques in the quality control (QC) and analysis of combinatorial libraries.[2]

A major impediment to the use of high-throughput flow-NMR in the analysis of combinatorial plates is the conundrum of data analysis. A typical high-throughput NMR run of a rapid parallel synthesis (RPS) plate generates up to 96 separate proton NMR spectra of compounds synthesized from an 8 × 12 array of reagents. NMR spectra in the synthetic chemistry environment have traditionally been interpreted manually by considering the various parameters present in the NMR spectrum including chemical shift, intensities, and coupling constants. This detailed and manual process of spectral analysis is clearly far too laborious when one is looking to ensure the quality control of 96 compounds at once, and possibly many multiples of 96, depending on the number of plates being synthesized within the compound library. Hence, methods that display all 96 spectra at once provide a "bird's eye" view of the combinatorial plate and can be effective aids in this analysis.

The glued pseudo-2D maps described previously in this journal[1] are one such example wherein the spectra can be glued back-to-back in any specified order according to row or column, resulting in an interferogram that can then be Fourier transformed to provide a pseudo-2D spectral representation. Visually, such a display can point toward systematic problems with compounds across a row or column because of the presence or absence of intrinsic NMR patterns. However, the analysis of the resulting pseudo-2D map is not automated, and it is up to the chemist to interpret the glued map and to derive meaningful conclusions. Thus, the granularity of the information derived is only as good as the time and effort the chemist is willing to invest in analyzing the map.

Commercial software packages are available that perform automated spectral prediction and matching of the experimental spectra as a means of quality assessment.[3] The reliability of such methods depends on the somewhat questionable accuracy of proton spectral prediction. It has been suggested that one could employ ^{13}C information obtained from rapid inverse-detected {^1H, ^{13}C} correlation experiments on each compound in conjunction with the proton spectra to improve the reliability of QC using ^1H and ^{13}C prediction matching.[4] This would add a minimum of about 5 min to the NMR experiment time per sample over and above a standard eight-scan proton spectrum. On our system, that translates into a plate run time of almost 24 h at a minimum, up from about 4 h per 96-well plate for a proton spectrum alone. While this approach may be theoreti-

[†] Sphinx Laboratories.
[‡] Information Technology.

10.1021/cc020031l CCC: $22.00
Published on Web 09/17/2002

cally feasible, the extended time and effort for NMR analysis may in fact reverse the efficiencies earned through rapid parallel synthesis by performing lengthy analyses of each compound without necessarily adding value to the overall QC process. The evaluation as to whether such analyses may be warranted in any particular situation must take into consideration the final destination of samples from these combinatorial chemistry libraries. In our case, that destination is typically a “first-pass” high-throughput biological screen.

Efficiency in array combinatorial synthesis is gained by using reagents in multiple wells, with common reagents being added across rows and down columns of a 96-well plate. The problem of analyzing NMR spectra from a plate of 96 compounds synthesized using RPS in an 8 × 12 array of reagents is essentially one of observing relationships in patterns. One expects a “fingerprint” in the resulting NMR spectra arising from −R groups added across a row and other −R′ groups added down a column. Presence or absence of these fingerprint patterns can be used as an aid to profile compound quality in combinatorial plates. Such efforts have been described in the context of other analytical techniques such as MS and TLC and form the basis of the glued psuedo-2D maps described above.[5] We were interested in developing an automated method that would be able to discern these patterns from the 96 experimental spectra and then display that data in an intuitive form that would be useful to a chemist performing the QC.

We demonstrate here the development of a self-organizing map (SOM)[6] to perform this analysis of 96 NMR spectra in a reliable, robust, and automated fashion. The SOM algorithm is a nonlinear generalization of principal component analysis,[7] which can cluster together data that contain common patterns. SOM has previously been used by one of the authors to cluster genes with similar expression patterns over time in a DNA microarray experiment.[8] Elsewhere, SOM has demonstrated the ability to cluster proton NMR spectra of blood plasma samples according to clinically relevant lipid classifications.[9] In this instance, we are looking for the SOM to discern common features in the 96 NMR spectra and to identify those compounds (outliers) that do not fit the expected patterns in the row or column to which they physically belong. This focuses the chemists' attention primarily on the outlier NMR spectra while ensuring that all other spectra fit some “parent” pattern. This could save an enormous amount of time and labor in performing QC of plates from RPS using proton NMR.

Self-organizing maps were originally developed to understand how certain topographic features were formed in the brain. As such, when the clustering is done, nearby clusters are more similar to each other than clusters that are further away on the topographic map. This map provides the scientist with a two-dimensional graphical representation of the data, making it easier to understand the overall character of the data represented. Hence, SOM is a very useful tool to obtain a “holistic” view of the data, which can be used to help choose spectra that may require further investigation, instead of performing detailed, laborious analysis of each spectrum.

In this paper, we describe several experimental conditions that reflect typical problems encountered in the analysis of NMR spectra of compounds resulting from RPS. The SOM is used to help identify similarities and differences in the NMR spectra from a 96-well plate. We demonstrate that SOM_NMR is able to rapidly identify outliers among the NMR spectra, thus lowering the barrier for performing such studies, leading in turn to improved quality control of compounds in combinatorial chemistry libraries.

Experimental Section

Two plates were used for this study, one each from different combinatorial libraries previously synthesized. Plates were first analyzed by LC−MS to ensure the expected *m*/*z* match and an average sample purity of >90% using UV, ELSD, and CLND detectors. LC−MS data were acquired using an Agilent LC interfaced with a Waters ZQ MS operating under MassLynx software control. All plates were dissolved in 0.5 mL of $CDCl_3$ at an average well concentration of 5 mg/mL for the purposes of high-throughput NMR. NMR spectra were acquired on a 500 MHz Varian Unity Inova spectrometer equipped with a flow-NMR automation accessory (VAST), which involves a Gilson 215 liquid-handler injecting samples sequentially into a 60 μL NMR flow-probe. Proton NMR spectra were acquired on each of the 96 compounds in a plate, using an identical set of standard eight-scan proton parameters and 16K data points at 25 °C.

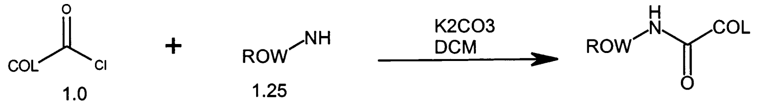
SOM_NMR Experiments. We describe below three different types of experiments that we performed on the SOM to test its strengths and limitations.

(1) The first involves the direct analysis of a 96-well plate (plate A) synthesized using an 8 × 12 reagent matrix. This test plate A depicted in Table 1 comprises 96 amides synthesized from a reaction between a series of 12 acid chlorides and 8 amines. This test plate was synthesized specifically for the purposes of NMR testing and passed our customary QC criteria using LC−MS and NMR.

(2) The second involves the digital production of NMR data for a hypothetical, hybrid plate C of 96 NMR spectra by combining spectra of distinct compounds from two different plates A (above) and B. Plate B contains 96 spiroheterocyclic compounds synthesized by reacting 12 acrylates and 8 cyclic secondary α-amino acids using a procedure described previously.[10] Spectra in plate A are substituted postacquisition by all spectra from column 9 of plate B, diagonally across wells A1 and B2 through H8 of plate A to produce the test NMR data set for plate C. The digital production of this hybrid test plate C is shown schematically in Table 2. The individual NMR spectra of the spiroheterocycles from plate B are quite dissimilar to those of the amides in plate A, but the compounds are all dissolved in the same NMR solvent, viz., $CDCl_3$. The peculiar pattern of the substitution to digitally “prepare” plate C was specifically chosen to avoid possible bias in the clustering arising from the row/column emphasis in the SOM algorithm.

(3) The third involves the doping of wells A1 through H8 diagonally across plate A of amides above with an “impurity” mixture comprising a 1:1:1 ratio of ethyl acetate (EtOAc), 2-propanol (^{i}Pr), and tetrahydrofuran (THF). In addition,

Table 1. Layout of 12 Acid Chlorides and 8 Amines Added down Columns and Across Rows, Respectively, during the Synthesis of Plate A of Amides Used in This Study



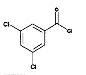
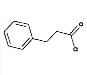
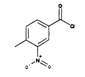
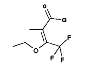
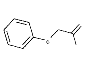
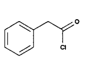
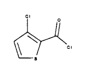
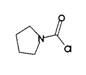
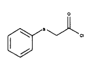
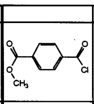
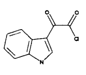
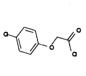
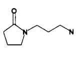
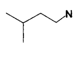
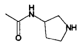
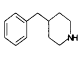
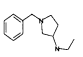
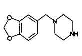
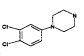
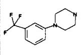
	COL-COCl												
		1	2	3	4	5	6	7	8	9	10	11	12
ROW-NH													
	A												
	B												
	C												
	D												
	E												
	F												
	G												
	H												

Table 2. Schematic Representation of Test Plate C Described in Experiment 2[a]

	1	2	3	4	5	6	7	8	9	10	11	12
A									▨			
B									▨			
C									▨			
D									▨			
E									▨			
F									▨			
G									▨			
H									▨			

Plate B

	1	2	3	4	5	6	7	8	9	10	11	12
A	▨											
B		▨										
C			▨									
D				▨								
E					▨							
F						▨						
G							▨					
H								▨				

Test Plate C

[a] Eight spectra from column 9 on plate B of spiro heterocycles were substituted diagonally across plate A of amides to produce hybrid test plate C, as depicted by the shaded wells.

wells H1 and H12 were also doped similarly to break any possible degeneracy arising from a single doped spectrum per row or column in our experiments. This doping mixture was added in a 1:1 ratio to the original compounds (amides) in these wells. This resulted in a new data set of 96 NMR spectra of which 10 were selectively "doped" as described. Table 3 shows a schematic representation of the doped test plate used in experiment 3.

These test experiments were designed to gauge the ability of the SOM to identify similarities and differences among the spectra and to discriminate among NMR spectra of markedly different compounds without any additional external input from the user. Experiments 2 and 3 above were designed to serve as surrogates for the kinds of problems that may typically arise during the automated processes used in RPS.

NMR Data Handling. The resulting 96 NMR spectra were glued according to procedures in the VNMR software described previously.[1] The spectra were then exported in 96 (X,Y-columnar) 2-column ASCII files. Each file had 16 384 data points, providing a digital resolution of 0.46 Hz for each spectrum. These files were then low-pass-filtered using a moving average of 40 points and subsampled with every 20th

Table 3. Schematic Representation of Test Plate A of Amides Doped 1:1 with an Equimolar "Impurity" Mixture Described in Experiment 3[a]

	1	2	3	4	5	6	7	8	9	10	11	12
A	▨											
B		▨										
C			▨									
D				▨								
E					▨							
F						▨						
G							▨					
H	▨							▨				▨

[a] The 10 shaded wells in this plate contain the "impurity" mixture in addition to the parent amide.

point. The filtering and subsampling were performed so that the peaks did not have to match exactly to be considered the same by the SOM algorithm. This produces a data matrix of 96 rows by 820 columns. This matrix file is submitted to the SOM program running on a LINUX server (1 GHz, 256 MB of RAM) on our intranet and is accessible through a web interface.

SOM Experiment. The SOM was programmed in-house using C, Java, and Perl. The NMR spectra were normalized using a unit hypersphere so that the comparisons were made on the basis of shape, i.e., relative peak heights within the spectrum, rather than on the magnitude of the peaks across all the spectra. This makes the SOM impervious to variations in peak intensity among spectra. The SOM was run for 100 epochs using a time-varying learning rate[11] and a time-varying Gaussian neighborhood function.[12] The time-varying neighborhood function starts out very wide, setting the overall landscape, and then becomes more focused, resulting in refinement of the clusters. For any given cluster, the SOM converges to the average of the spectra in the cluster and this “average” spectrum is graphically displayed (e.g., see Figure 1a). Empty clusters, containing no spectra, are represented by the average of the neighboring clusters, preserving the topology of the map. To simplify the visual appearance of the SOM, empty clusters are blanked out in the final display.

When the SOM is run, it is necessary to specify a priori the dimensions of the map that are to be produced. Choosing too many will simply order the spectra in two dimensions and may obviate the need for clustering, while choosing too few will introduce far too much variance within the clusters and make interpretation of the SOM potentially meaningless. For this work, SOMs were generated using matrixes of varying sizes such as 4 × 4 and 5 × 5, and after a number of runs, it seemed that the 5 × 5 matrices best represented our NMR data. This choice is also appropriate given that the number of reagents that gives rise to the source patterns in the spectra is 8 + 12, i.e., 20. The 25-cluster matrix thus provides the optimal size for RPS data and is not frequently altered. The default choice of matrix size for our SOM analysis is 5 × 5, while matrixes of other sizes can be chosen as options from a drop-down list if desired. However, if the RPS plate is designed to synthesize just 24 compounds in a 4 × 6 array, one could use a smaller matrix size such as 4 × 4 for the SOM. It is also possible to increase the dimensionality of SOM to 3D or higher and potentially gain some additional modeling power. However, the intuitive 2D map is lost as is the ability to print the results on paper.

The SOM updates the weight matrix in a global fashion instead of a local fashion (as in hierarchical clustering methods), making it more impervious to noise in the data. However, the SOM can be sensitive to the initial conditions of the weights, which are initialized with random numbers, if the clusters are not well defined. To test the validity of this method, SOMs were run many times with different random number seeds. In all cases, the results were either identical or qualitatively similar, with outlier spectra still being highlighted.

Row and Column Emphasis of the SOM. For compounds synthesized in a combinatorial RPS array, one encounters the presence of two intrinsic patterns in the NMR spectra, one along the rows of the plate and the other along the columns, arising from the rowwise and columnwise addition of reagents. When the SOM was initially run using the NMR data as specified so far, it would find the most prevalent patterns, be they from an entire row, column, or some partial grouping of row and column, depending on the number of clusters available. Such “global” pattern analysis may be quite revealing in some circumstances. However, this scrambling of all NMR spectra from all rows and columns into several clusters by the SOM algorithm could also lead to unnecessary conflict between the two intrinsic patterns and may cause confusion during the interpretation of the resulting SOM. Hence, we chose to force the clustering to emphasize the rows of the plate preferentially at the expense of the columns and vice versa. This would help identify outlier spectra within a row or column more directly, since there would now be only one intrinsic parent pattern (row or column) from which SOM must detect a departure. To emphasize such row or column clustering, additional 20 columns of data were generated (for a total of 840 columns of data), representing the 8 rows and 12 columns found in the combinatorial plate. For example, if row emphasis is desired, the maximum *Y* value for any of the spectra from that row is placed into an additional data column for all spectra from that row in the plate. The SOM treats these additional columns of data like it would any of the NMR spectra, i.e., the more closely the data columns match, the more likely they are to cluster together. A multiplier may be applied to adjust the weighting factor of the extra data columns in a SOM run, which adjusts the level of row or column emphasis desired. In our experience, it is advisable to run both row- and column-emphasized SOM using multiplier values of 1, 1.5, and 2 to determine the degree of emphasis that provides the most graphically useful clustering for any given “plate” of RPS spectra.

Results and Discussion

Figure 1a shows the row-emphasized SOM for plate A of amides. The layout shows a 5 × 5 matrix, resulting in a total of 25 clusters. Each cluster is identified by a number, which is the sum of its row and column coordinates on the 2D SOM. A floating tool-tip is displayed over each cluster listing the NMR spectra belonging to that cluster. Clicking on any one of these clusters leads to another web page that displays all the NMR spectra that were classified by the SOM as belonging to that cluster. Clusters that are blanked out in the display contain no spectra. Singleton clusters are those that contain exactly one NMR spectrum and are always colored in the darker shade of blue on the SOM. Clusters that are colored in lighter blue denote clusters with the highest variance.

It is instructive to note that most of the NMR spectra of the 96 compounds cluster neatly according to the rows they physically belong to on the plate. This is in fact expected, first, because of the presence of “fingerprint” patterns in the NMR spectra of these compounds arising from the specific

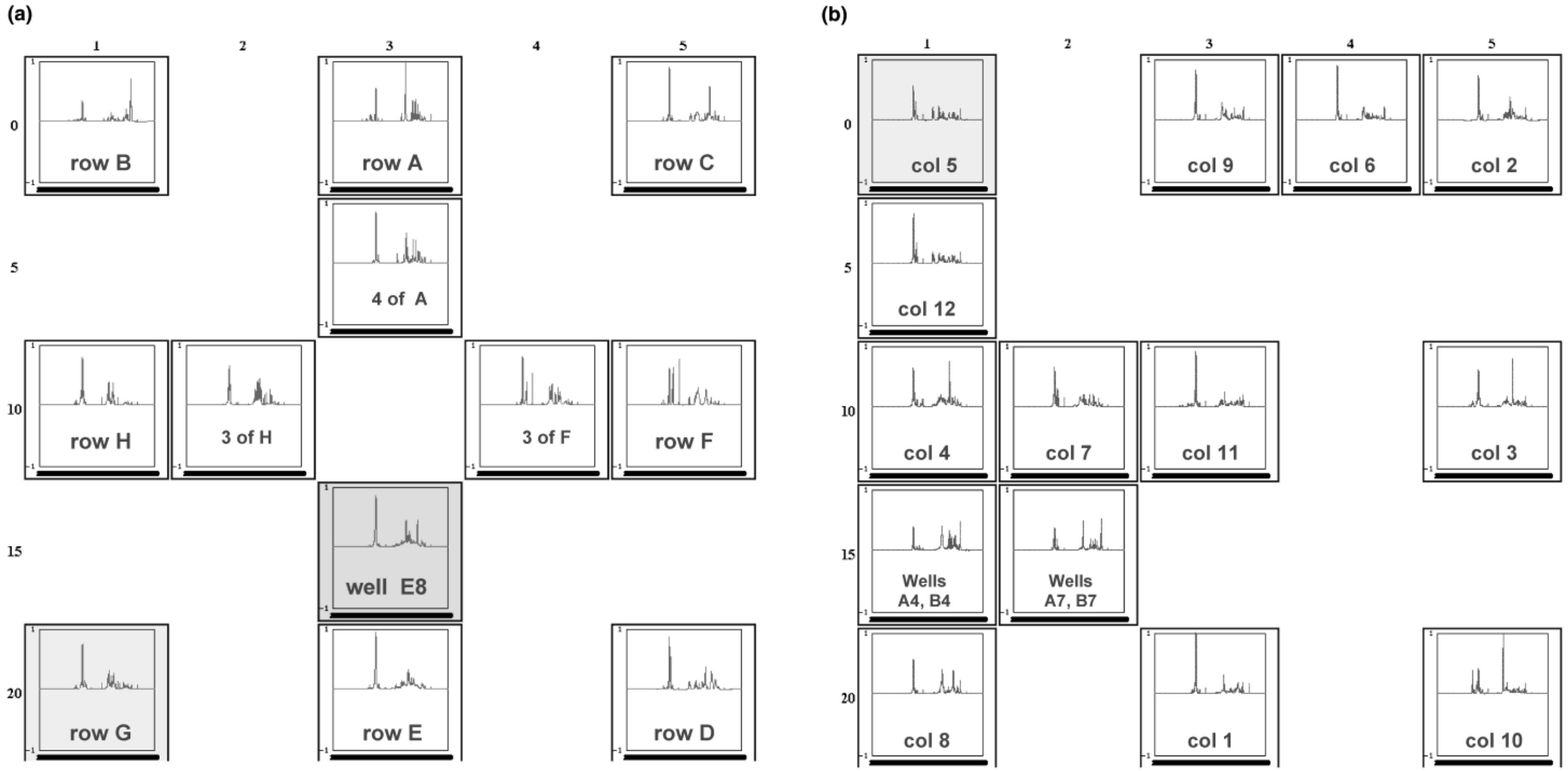


Figure 1. (a) Row-emphasized SOM from experiment 1 for plate A of amides. Clusters are numbered according to the sum of their row and column coordinates on the SOM. Blanked clusters contain no spectra. Note the positioning of spectrum E8 in the singleton (cluster 18) adjacent to the cluster containing the remaining spectra from row E (cluster 23). E8 is the only compound in row E synthesized using a nonaromatic acid chloride, and the SOM detects that difference in the NMR shift pattern from among the similarities it shares with other spectra from row E. (b) Column-emphasized SOM from experiment 1 for plate A of amides. Note the adjacent positioning on the map of clusters 1 and 6 containing the spectra of compounds from plate columns 5 and 12, respectively, which are synthesized using structurally similar acid chlorides.

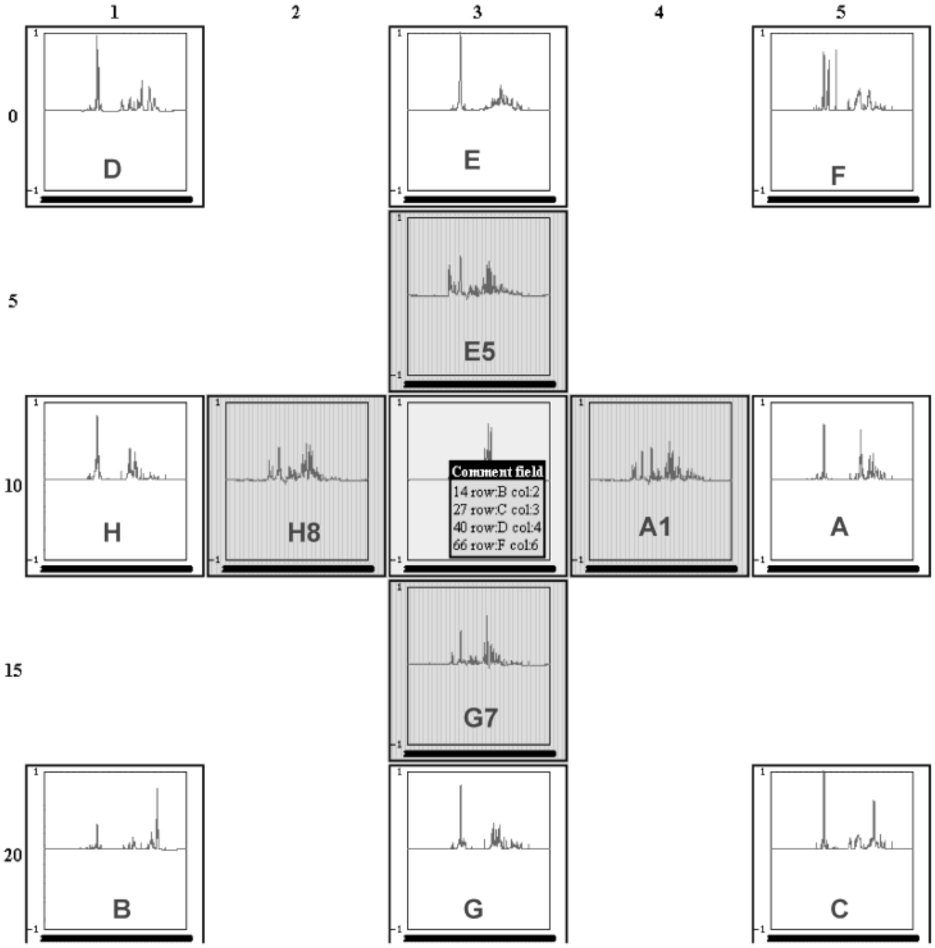


Figure 2. Row-emphasized SOM from experiment 2 for test plate C of amides substituted diagonally by eight spectra of spiro heterocycles from plate B. Note the positioning of all eight substituted spectra from wells A1 through H8 at the center of the SOM but yet adjacent to their parent wells.

−R groups added across each row and, second, because of our procedure of row emphasis described above. Also noteworthy is the fact that the majority of these spectral clusters have no neighbors, i.e., spectra in clusters immediately adjacent in any direction on the SOM. Neighboring clusters, when they do exist, typically contain spectra of compounds from the "parent" next-neighbor row, since those spectra closely resemble a "parent" pattern of NMR peaks. This can be observed in Figure 1a where clusters 8, 12, 14, and 18 contain respectively 4, 3, 3, and 1 spectra from their next-neighbor parent rows A, H, F, and E. This is testimony to the ability of the SOM to detect similar patterns in the NMR spectra without any external training.

Using the SOM in Figure 1a during a typical combinatorial plate analysis, a chemist would probe further into the outlier spectra, since the SOM appears to detect in them some departure from a "parent" pattern, which may signal a "flag" in terms of QC. One such outlier of note in this SOM is the NMR spectrum from well E8, which appears as a singleton in cluster 18. It is positioned on the SOM directly above cluster 23 containing all other NMR spectra from row E, as expected. The SOM is able to detect a common feature in the NMR spectra from row E and cluster them together, while at the same time being able to detect something different in the spectrum from well E8. As one can observe in Table 1, column 8 carries the only acid chloride of the 12 that contains a cyclic aliphatic group (−pyrrolidine), thus contributing to a unique pattern of NMR chemical shifts for the resulting amide in well E8. This increased our confidence in the ability of the SOM to discern the presence or absence of patterns in the NMR.

Figure 1b shows the column-emphasized SOM of experiment 1 on plate A. As in the previous row-emphasized case, the spectra from each column of the plate mostly cluster together except for those indicated in clusters 16 and 17 (wells A4, B4 and A7, B7). In this instance, a chemist performing QC could choose to investigate only these two clusters (four spectra) more closely and perhaps rapidly browse some others. The SOM indicates merely that it detects something different in these two clusters, which does not necessarily imply a QC problem with these wells. It simply indicates a departure from an expected "parent" pattern of NMR chemical shifts, prodding the chemist to analyze these some more. While this column-emphasized map appears visually different from the row-emphasized one in Figure 1a (12 vs 8 parent patterns), the principles guiding its interpretation and use are very similar.

Figure 1b shows all eight compounds from column 5 of the plate clustered together in cluster 1, while cluster 6 contains all compounds from column 12. The SOM is clearly able to discern that the amides resulting from the acid chlorides added down columns 5 and 12 have similar NMR

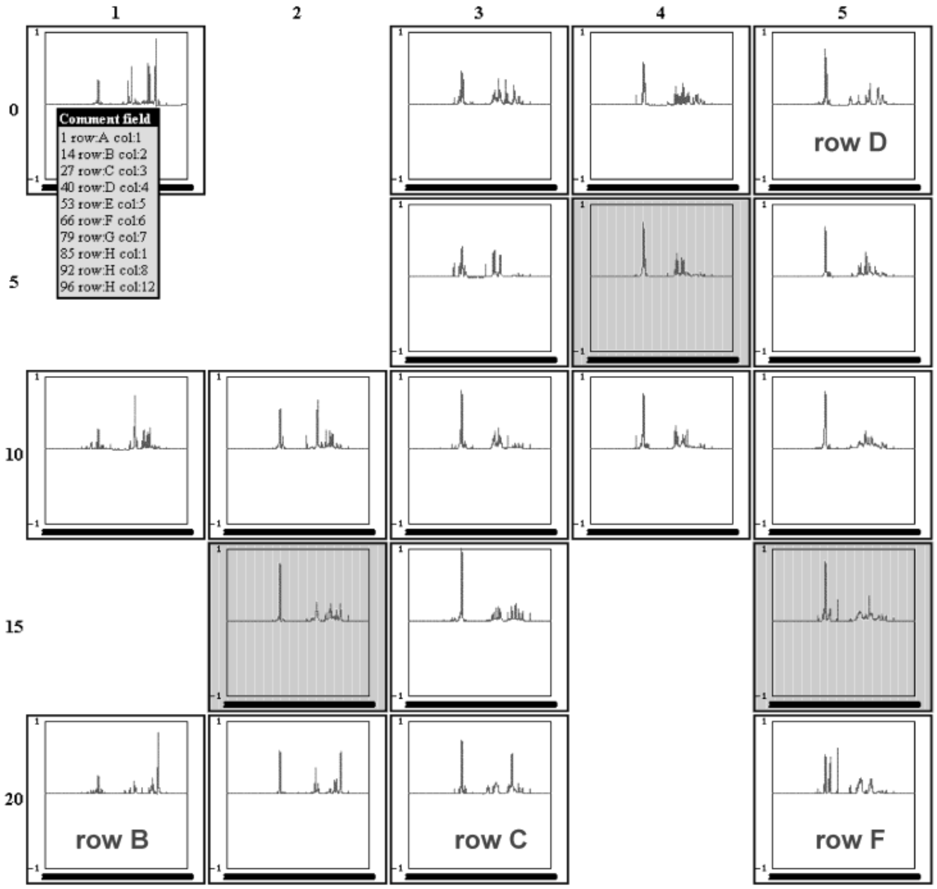


Figure 3. SOM from experiment 3 for test plate A of amides doped diagonally by an "impurity" mixture as shown in Table 3. Notice the clustering together of all 10 doped spectra from wells A1 through H8, H1, and H12 in cluster 1.

spectral features because of their structural similarity (phenoxyacetyl chloride and its *p*-chloro derivative, respectively). Also notice that clusters 4 and 5 contain all eight spectra from columns 6 and 2, respectively, which are also synthesized from structurally similar acid chlorides (phenylacetyl chloride and 3-phenylpropanoyl chloride). This illustrates that the SOM recognizes similarities among the spectra of these compounds and clusters them in proximity to each other on the map. One thus obtains a visual image of the 96 NMR spectra in just two (row- and column-emphasized) self-organizing maps, which can then be readily explored further, if necessary. At this point, we decided to further explore the ability of the SOM to cluster NMR spectra by testing it with experiments 2 and 3 described earlier.

Figure 2 shows the SOM for the data set described in experiment 2 above. This map graphically depicts the power of the SOM to cluster similar spectra together. Four of the eight "substitute" spectra are clustered together at the center in cluster 13, with the remaining four occupying adjacent clusters. This is due to the spectral similarities among these eight substituted spectra juxtaposed against the "artificial" row-emphasis criteria that have been enforced. Hence, the SOM is capable of discriminating between the NMR spectra of unrelated compound classes, the amides, and the spiro heterocycles. This ability of the SOM could be very useful in identifying those wells in a plate in which the intended product was not synthesized for any reason.

Figure 3 depicts the SOM for experiment 3 above where a mixture of solvents intended to serve as an "impurity surrogate" is added to the wells in a 1:1 ratio to the parent amide compounds in plate A across the diagonal A1 through H8 and in wells H1 and H12. Cluster 1 on this map clearly shows that the SOM was able to cluster all these 10 spectra together, since they contain a common "impurity" even though each of these 10 spectra do bear a similarity with spectra from their parent row/column. Of particular note is the fact that cluster 1 is the only cluster on this SOM that has no neighbors, implying that its spectra bear little resemblance to other spectra on the plate relative to each other. This experiment demonstrates that SOM can be a powerful tool in identifying "problem wells" on a plate.

Experiment 3 was conducted at decreasing ratios of the impurity mixture to the parent amides such as 5:1 down to 1:1, and the SOM performed similarly in each of those cases. However, as the ratio was decreased further below 1:1, the SOM was unable to consistently cluster these 10 spectra containing impurities as outliers. For example, at an impurity ratio of 0.8:1, with no row or column emphasis, there were five singleton clusters, two of which were impurity wells. With a column emphasis of 1, one cluster had four of nine members from the impurity set. The patterns of the R− groups in the NMR spectra simply become stronger than the patterns of the impurities at the lower concentrations. However, even this may be enough information to signify to the chemist that something is amiss in those wells. Experiments are in progress to extend this SOM methodology to situations in which one may have no more than just a few wells with trace levels of impurities.

Conclusions

The results discussed above indicate that SOM is capable of (a) clustering NMR spectra from 96-well combinatorial plates according to the similarities present in the spectra across rows and down columns of a plate, (b) positioning these clusters on a graphical 2D representation such that proximity on the map denotes structural similarity, and (c) identifying “outlier” spectra from among rows or columns that do not fit a “parent” pattern characteristic of that row or column on the plate.

These features of SOM are extremely helpful to a chemist to quickly identify “problem wells” in a plate. The laborious process of analyzing 96 NMR spectra sequentially and manually can thus be reduced to analyzing only the outlier spectra from two self-organizing maps (row- and column-emphasized) and to ensuring that all other spectra cluster neatly along their parent rows and columns. SOM requires no additional input beyond the NMR spectra from the combinatorial plate, making it far more efficient than other “automated” methods of analyzing such NMR data, which often require the creation of extensive databases of reagent spectra or prior iterative training of artificial neural networks. Each SOM_NMR run (row- and column-emphasized) on a 96-well combinatorial plate takes less than 60 s run time on our LINUX server. SOM_NMR is thus a very efficient tool to obtain a QC profile of a combinatorial plate from the proton NMR spectra.

It is clear from the results shown in Figures 2 and 3 that if a similar “chemical problem” occurs down a column or across a row during plate synthesis, the SOM might cluster all those spectra together. In the event of an excessive number of dissimilar deviations from the expected row/column patterns across a plate, we recommend expanding the 5 × 5 layout of the SOM map to accommodate the several outlier spectra, resulting in a more graphically intuitive SOM. It is important to use SOM_NMR as a guide and also to browse the consistent clusters for lurking signs of “problem” wells, which is best accomplished in conjunction with other analytical data. SOM may identify certain spectra as outliers within a given set for any number of reasons, and hence, it should be used as an aid to QC by delving deeper into the outlier spectra to identify the problems. It is important not to overinterpret the SOM result.

Most significantly, the use of SOM to analyze such large sets of NMR data may lower the barrier to actually acquiring such NMR data on large combinatorial libraries, thus ensuring improved quality in the compounds submitted for high-throughput screening. Clearly, the use of SOM in the context of combinatorial chemistry and high-throughput screening need not be limited to the analysis of NMR spectra but should be explored with other types of complex analytical and screening data as well.

Acknowledgment. The authors acknowledge comments and research support from Linda Dow, Philip F. Hughes, John Lampe, and Robert C. Anderson.

Supporting Information Available. Additional data are available. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Keifer, P. A.; Smallcombe, S. H.; Williams, E. H.; Salomon, K. E.; Mendez, G.; Belletire, J. L.; Moore, C. D. Direct-Injection NMR (DI-NMR): A Flow NMR Technique for the Analysis of Combinatorial Chemistry Libraries. *J. Comb. Chem.* **2000**, *2*, 151−171.

(2) Yurek, D. A.; Branch, D. L.; Kuo, M.-S. Development of a System To Evaluate Compound Identity, Purity, and Concentration in a Single Experiment and Its Application in Quality Assessment of Combinatorial Libraries and Screening Hits. *J. Comb. Chem.* **2002**, *4*, 138−148.

(3) ACD/CNMR Predictor, v5.0, ACD/HNMR Predictor v5.0, ACD/CombiNMR Predictor v5.0., Advanced Chemistry Development, Inc., Toronto (www.acdlabs.com).

(4) Spitzer, T.; Sefler, A. M.; Rutkowske, R. An improved DEPT-HMQC sequence for high-throughput NMR analysis. *Magn. Reson. Chem.* **2001**, *39*, 539−543.

(5) Gorlach, E.; Richmond, R.; Lewis, I. High-Throughput Flow Injection Analysis Mass Spectrometry with Networked Delivery of Color-Rendered Results. 2. Three-Dimensional Spectral Mapping of 96-Well Combinatorial Chemistry Racks. *Anal. Chem.* **1998**, *70*, 3227−3234.

(6) Kohonen, T. Self-organized formation of topologically correct feature maps. *Biol. Cybernet.* **1982**, *43*, 59−69. Kohonen, T. The self-organizing map. *Proc. IEEE* **1990**, *78*, 1464−1480.

(7) Ritter, H. Self-organizing feature maps: Kohonen maps. In *The Handbook of Brain Theory and Neural Networks*; Arbib, M. A., Ed.; MIT Press: Cambridge, MA, 1995; pp 846−851.

(8) Baker, T. K.; Carfagna, M. A.; Gao, H.; Dow, E. R.; Li, Q.; Searfoss, G. H.; Ryan, T. P. Temporal Gene Expression Analysis of Monolayer Cultured Rat Hepatocytes. *Chem. Res. Toxicol.* **2001**, *14*, 1218−1231.

(9) Kaartinen, J.; Hiltunen, Y.; Kovanen, P. T.; Ala-Korpela, M. Application of self-organizing maps for the detection and classification of human blood plasma lipoprotein lipid profiles on the basis of ^{1}H NMR spectroscopy data. *NMR Biomed*. **1998**, *11*, 168−176.

(10) Coulter, T.; Grigg, R.; Malone, J. F.; Sridharan, V. Chiral Induction in Cycloaddition Reactions of Azomethine Ylides Derived from Secondary α-Amino Acids by the Decarboxylative Route. *Tetrahedron Lett*. **1991**, *32*, 5417−5420. Mendoza, J. Unpublished results.

(11) Ritter, H.; Martinez, T.; Schulten, K. *Neural Computation and Self-Organizing Maps: An Introduction*; Addison-Wesley: Reading, MA, 1992. Kohonen, T. Exploration of very large databases by self-organizing maps. *Int. Conf. Neural Networks* **1997**, *1*, 1−6.

(12) Lo, Z. P.; Yu, Y.; Bavarian, B. Analysis of the convergence properties of topology preserving neural networks. *IEEE Trans. Neural Networks* **1993**, *4*, 207−220.

CC020031L