Article

Web Server Based Complex Mixture Analysis by NMR

Steven L. Robinette, Fengli Zhang, Lei Bru#schweiler-Li, and Rafael Bru#schweiler

Anal. Chem., **2008**, 80 (10), 3606-3611 • DOI: 10.1021/ac702530t • Publication Date (Web): 19 April 2008

Anal. Chem. **2008,** *80,* 3606–3611

Web Server Based Complex Mixture Analysis by NMR

Steven L. Robinette,[†,‡] Fengli Zhang,[†] Lei Brüschweiler-Li,[†,§] and Rafael Brüschweiler*[,†,§]

National High Magnetic Field Laboratory, University of Florida, Gainesville, Florida 32611, and Department of Chemistry and Biochemistry, Florida State University, Tallahassee, Florida 32310

Comprehensive metabolite identification and quantification of complex biological mixtures are central aspects of metabolomics. NMR shows excellent promise for these tasks. An automated fingerprinting strategy is presented, termed COLMAR query, which screens NMR chemical shift lists or raw 1D NMR cross sections taken from covariance total correlation spectroscopy (TOCSY) spectra or other multidimensional NMR spectra against an NMR spectral database. Cross peaks are selected using local clustering to avoid ambiguities between chemical shifts and scalar *J*-coupling effects. With the use of three different algorithms, the corresponding chemical shift list is then screened against chemical shift lists extracted from 1D spectra of a NMR database. The resulting query scores produced by forward assignment, reverse assignment, and bipartite weighted-matching algorithms are combined into a consensus score, which provides a robust means for identifying the correct compound. The approach is demonstrated for a metabolite model mixture that is screened against the metabolomics BioMagResDatabank (BMRB). This NMR-based compound identification approach has been implemented in a public Web server that allows the efficient analysis of a wide range of metabolite mixtures.

The study of the chemical composition of biological systems in response to a multitude of factors such as genetics, age, pathology, development, environment, and stress is the objective of metabolomics/metabonomics, which forms an essential part of systems biology.[1,2] The reliable and efficient identification of low molecular weight chemical components, known as fingerprinting, is a key aspect of metabolomics studies. Currently, the two main analytical techniques for this task are NMR spectroscopy[3] and mass spectrometry.[4,5]

Although the richness of a standard 1D proton NMR spectrum of a complex mixture of unknown composition typically precludes unambiguous interpretation, two-dimensional NMR spectra, such as correlation spectroscopy (COSY),[6] total correlation spectroscopy (TOCSY),[7] and heteronuclear $^{13}C-^{1}H$ correlation spectra and pseudo-2D spectra such as diffusion-ordered spectra (DOSY)[8] and statistical total correlation spectroscopy (STOCSY) spectra,[9] provide resonance separation along a second dimension that facilitates identification of individual chemical components.

The availability of public-domain metabolomics/metabonomics NMR databases, in particular the Biological Magnetic Resonance Data Bank (BMRB)[10] (http://www.bmrb.wisc.edu) and the Human Metabolome Database[11] (http://www.hmdb.ca) that contain NMR spectra and peak lists of an increasing number of compounds have the potential to greatly assist compound identification by NMR. An approach for database screening of 2D magnitude COSY spectra has been proposed, which directly compares 2D cross peaks.[12]

Covariance TOCSY spectroscopy[13–15] has recently been demonstrated to be suitable for complex metabolite mixture analysis.[16] TOCSY provides high sensitivity and high spectral resolution along both frequency dimensions displaying complete spin system connectivities for each compound.

The DemixC method[16] decomposes a complete 2D covariance TOCSY spectrum into a small number of nonredundant 1D cross sections or traces (Figure 1). Each of these traces represents a fingerprint of an individual mixture component with minimal overlap with other components. The number of identified traces depends on the mixture composition and the sensitivity of the subspectra of the individual components. The DemixC procedure

* Correspondence to be addressed to: Prof. Rafael Brüschweiler, Department of Chemistry and Biochemistry, National High Magnetic Field Laboratory, Florida State University, Tallahassee, FL 32306. Phone: 850-644-1768. Fax: 850-644-8281. E-mail: bruschweiler@magnet.fsu.edu.

† National High Magnet Field Laboratory.

‡ University of Florida.

§ Department of Chemistry and Biochemistry, Florida State University.

(1) Nicholson, J. K.; Wilson, I. D. *Nat. Rev. Drug Discovery* **2003**, *2*, 668–676.
(2) Fiehn, O. *Plant Mol. Biol.* **2002**, *48*, 155–171.
(3) Lenz, E. M.; Wilson, I. D. *J. Proteome Res.* **2007**, *6*, 443–458.
(4) Dettmer, K.; Aronov, P. A.; Hammock, B. D. *Mass Spectrom. Rev.* **2007**, *26*, 51–78.
(5) Pan, Z. Z.; Raftery, D. *Anal. Bioanal. Chem.* **2007**, *387*, 525–527.
(6) Aue, W. P.; Bartholdi, E.; Ernst, R. R. *J. Chem. Phys.* **1976**, *64*, 2229–2246.
(7) Braunschweiler, L.; Ernst, R. R. *J. Magn. Reson.* **1983**, *53*, 521–528.
(8) Johnson, C. S. *Prog. Nucl. Magn. Reson. Spectrosc.* **1999**, *34*, 203–256.
(9) Cloarec, O.; Dumas, M. E.; Craig, A.; Barton, R. H.; Trygg, J.; Hudson, J.; Blancher, C.; Gauguier, D.; Lindon, J. C.; Holmes, E.; Nicholson, J. *Anal. Chem.* **2005**, *77*, 1282–1289.
(10) Seavey, B. R.; Farr, E. A.; Westler, W. M.; Markley, J. L. *J. Biomol. NMR* **1991**, *1*, 217–236.
(11) Wishart, D. S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A. C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; Fung, C.; Nikolai, L.; Lewis, M.; Coutouly, M. A.; Forsythe, I.; Tang, P.; Shrivastava, S.; Jeroncic, K.; Stothard, P.; Amegbey, G.; Block, D.; Hau, D. D.; Wagner, J.; Miniaci, J.; Clements, M.; Gebremedhin, M.; Guo, N.; Zhang, Y.; Duggan, G. E.; MacInnis, G. D.; Weljie, A. M.; Dowlatabadi, R.; Bamforth, F.; Clive, D.; Greiner, R.; Li, L.; Marrie, T.; Sykes, B. D.; Vogel, H. J.; Querengesser, L. *Nucleic Acids Res.* **2007**, *35*, D521–D526.
(12) Xi, Y. X.; de Ropp, J. S.; Viant, M. R.; Woodruff, D. L.; Yu, P. *Metabolomics* **2006**, *2*, 221–233.
(13) Trbovic, N.; Smirnov, S.; Zhang, F.; Brüschweiler, R. *J. Magn. Reson.* **2004**, *171*, 277–283.
(14) Brüschweiler, R.; Zhang, F. *J. Chem. Phys.* **2004**, *120*, 5253–5260.
(15) Brüschweiler, R. *J. Chem. Phys.* **2004**, *121*, 409–414.
(16) Zhang, F.; Brüschweiler, R. *Angew. Chem., Int. Ed.* **2007**, *46*, 2639–2642.

10.1021/ac702530t CCC: $40.75
Published on Web 04/19/2008

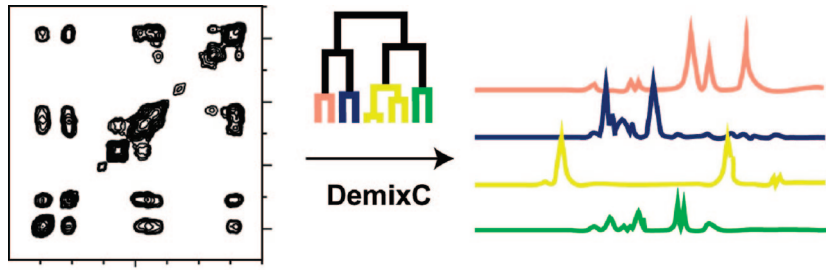

Figure 1. Schematic illustration of the DemixC method that decomposes a 2D covariance TOCSY spectrum into 1D traces of individual components that are subsequently subjected to peak picking and database searching using the COLMAR query Web portal at http://spinportal.magnet.fsu.edu.

does not require physical separation and was designed with the goal of enabling automated component identification using database screening. The DemixC method was successfully applied to a metabolomics study of the defensive secretion of a single insect.[17]

The approach presented here fully automatically screens 1D NMR subspectra, obtained by DemixC or other means, against an NMR database of small molecules and metabolites and returns a rank-ordered list of compounds that best match the query spectrum. The approach, which is termed COLMAR query for complex mixture analysis by NMR using database query, has been implemented on a publicly accessible Web server.

MATERIALS AND METHODS

NMR Sample. A model mixture was prepared by mixing the seven components histidine, lysine, glutamine, glucose, D-carnitine, myo-inositol, and shikimic acid (all purchased from Sigma-Aldrich) at 1 mM concentration each in 100% D_2O solution.

NMR Experiment. NMR measurements were done at 800 MHz with a total sample volume of 200 μL in a 3 mm NMR tube. The sample temperature was maintained at 298 K and residual water signal was suppressed by excitation sculpting.[18] A 2D TOCSY experiment[7] was collected using the MLEV-17 mixing sequence[19] with a 90 ms mixing time and 2048 t_2 and 1024 t_1 (complex) data points. States-TPPI was used for quadrature detection along the indirect dimension. The NMR data were processed by NMRPipe.[20] The time-domain data were apodized along the detection dimension (t_2) using a cosine function, zero-filled to 2048 data points, Fourier transformed, and polynomially baseline corrected to obtain the mixed time-frequency domain spectrum $\mathbf{F}(t_1,\omega_2)$.

Covariance and DemixC Processing. The covariance TOCSY spectrum (**C**) of the mixture was obtained from the mixed time-frequency domain spectrum (**F**) by the matrix multiplication and square root operation $\mathbf{C} = (\mathbf{F}^T\mathbf{F})^{1/2}$ using the covNMR module[13] in NMRPipe.[20] Because of its inherent symmetry, **C** displays the same high spectral resolution along both dimensions. Spectrum **C** is then subjected to the DemixC spectral deconvolution procedure sketched in Figure 1: for each trace (row) of matrix **C** the importance index was calculated as a measure of the cumulative overlap of this trace with all other traces of **C** by summing all elements of the corresponding row of $\mathbf{C}^2$.[16] On the basis of the importance index profile, a subset of traces of **C** was selected and clustered according to the agglomerative clustering algorithm.[21] For each cluster, a representative trace was identified as the trace with minimal importance index. In this way, the likelihood is maximized that the selected traces reflect individual components free of spurious contributions from other spin systems.

In order to identify the compound belonging to a DemixC trace, the trace is screened against all 1D spectra of the database by quantitative comparison of peak lists. For this purpose, chemical shift lists of resonances in the DemixC traces are generated using the same peak-picking procedure that is applied to the 1D database reference spectra.

Database Preprocessing. The complete BMRB metabolomics database was downloaded from http://www.bmrb.wisc.edu/metabolomics/, which at present contains over 250 different compounds. The ^{1}H 1D free-induction decays were locally reprocessed using NMRPipe with processing parameters, including parameters for phase correction and frequency calibration, taken from the BMRB. The final 1D spectra consist of 32K data points each.

Although BMRB data sets already include peak lists for each compound, independent peak picking of all 1D spectra was performed using the same peak-picking algorithm and criteria as those used for peak picking of covariance TOCSY traces. Peak picking was performed using both thresholding and domain clustering: the positions of all relative maxima with an intensity larger than 5% of the maximum spectral peak are stored in a vector and then clustered with a cluster resolution of 0.03 ppm. In this way the multiplet maxima that lie within 0.03 ppm of each other are combined into an average chemical shift.

Query Trace Processing. DemixC traces, which consist of 2048 real points each, are peak picked similar to the 1D spectra of the BMRB. Because of potential weak overlaps between traces, some of the smallest peaks in the DemixC traces may not belong to the same spin system and would bias database scoring if included in a list of trace chemical shifts. Because of the relatively high tolerance of the forward weighted matching and forward assignments algorithms for missing peaks, peaks smaller than 25% of the maximal peak are eliminated. A 0.03 ppm cutoff for domain clustering is then applied to determine average chemical shifts from multiplet components, which is fully analogous to peak picking of the BMRB spectra.

(17) Zhang, F.; Dossey, A. T.; Zachariah, C.; Edison, A. S.; Brüschweiler, R. *Anal. Chem.* **2007**, *79*, 7748–7752.
(18) Hwang, T. L.; Shaka, A. J. *J. Magn. Reson., Ser. A* **1995**, *112*, 275–279.
(19) Shaka, A. J.; Lee, C. J.; Pines, A. *J. Magn. Reson.* **1988**, *77*, 274–293.
(20) Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. *J. Biomol. NMR* **1995**, *6*, 277–93.
(21) Jain, A. K.; Murty, M. N.; Flynn, P. J. *ACM Computing Surveys* **1999**, *31*, 264–323.

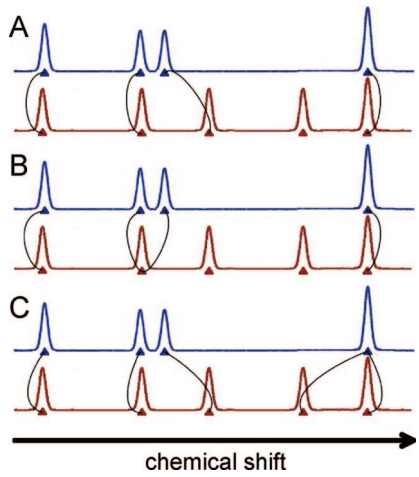

Figure 2. Visualization of different assignment algorithms of a query spectrum (blue) against a database spectrum (red). (A) forward weighted matching of the query trace against the database trace produces unambiguous resonance assignment of the query trace. (B) Forward assignment (each peak of the query trace is assigned to the closest peak of the database trace allowing multiple assignments). (C) Reverse assignment (each peak of the database trace is assigned to the closest peak in the query allowing multiple assignments).

Database Query and Scoring. The peak list of each query trace is scored against the peak lists of all database entries using the three algorithms depicted in Figure 2. All algorithms determine a score that is computed from the chemical shift difference matrix $\Delta^{(n)}$ with elements

$$\Delta_{jk}^{(n)} = |\delta_{Q,j} - \delta_{D,k}{}^{(n)}| \quad (1)$$

where $\delta_{Q,j}$ is the chemical shift of resonance j of the query trace and $\delta_{D,k}{}^{(n)}$ is resonance k of database spectrum n. In addition, in order to limit the effect of outliers, $\Delta_{j,k}{}^{(n)}$ is capped at 1 ppm, i.e. $\Delta_{j,k}^{(n)}$ ppm whenever $|\delta_{Q,j} - \delta_{D,k}^{(n)}| > 1$ ppm. If N is the number of peaks picked in the query trace and M is the number of peaks picked in the database trace, matrix $\Delta^{(n)}$ is a $N \times M$ matrix.

The *forward assignment algorithm* (Figure 2B) assigns to each chemical shift of the query trace the peak in the database peak list that is closest as measured by the chemical shift difference $\Delta_{j,k}^{(n)}$. This algorithm allows assignment of different query peaks to the same database peak by using the following penalty function

$$P_{\text{forw}}^{(n)} = \sum_j \min_k\{\Delta_{jk}^{(n)}\} \quad (2)$$

Thus, $P_{\text{forw}}^{(n)}$ is the sum of the minima of each row of $\Delta^{(n)}$. It follows that each resonance of the query spectrum is assigned to a resonance of the database spectrum but not necessarily vice versa.

The *reverse assignment algorithm* (Figure 2C) works identically to the forward algorithm except that the roles of the query trace and the database spectrum are exchanged. Therefore the penalty function is

$$P_{\text{rev}}^{(n)} = \sum_k \min_j\{\Delta_{jk}^{(n)}\} \quad (3)$$

Thus, $P_{\text{rev}}^{(n)}$ is the sum of the minima of each column of $\Delta_{(n)}$. It follows that each resonance of the database spectrum is assigned to a resonance of the database spectrum but not necessarily vice versa.

The *weighted matching algorithm*[22] (Figure 2A) produces in its standard form unambiguous NMR assignments.[23] If the query peak list has N entries and the peak list of the database spectrum has M entries, the algorithm matches the smaller of the two peak lists with the larger one so that each peak from the smaller list is uniquely assigned to a peak from the larger list, i.e., no two peaks from the smaller list are assigned to the same peak of the larger list. If $N < M$, there are $X = \binom{M}{N} N!$ possible assignments and the optimal assignment is the one whose penalty score

$$P_{\text{wm}}^{(n)} = \min_{c=1\ldots X} \sum_{j=1}^{N} \Delta_{j,f(c,j)}^{(n)} \quad (4)$$

is minimal. $f(c,j)$ denotes the number of the peak of database spectrum n that is assigned to peak j of the query spectrum in permutation c. The weighted matching problem of eq 4 is efficiently solved by the Hungarian algorithm,[22] which requires for $N = M$ only $O(N^3)$ arithmetic operations, i.e., it solves a factorial problem in polynomial time.

In contrast to the forward and reverse assignment algorithms, weighted matching yields unambiguous assignments. In the case of complete and fully resolved peaks both in the query trace and the database trace, weighted matching is the algorithm of choice. However, if $N > M$, the algorithm scores those database spectra highly whose peaks coincidentally fit a subset of query peaks well, which may be a database trace with $M = 1$ even if $N \gg 1$. To avoid this situation, for $N > M$, the weighted matching algorithm can be applied iteratively by eliminating in each iteration the query peaks that were assigned in the previous round. In this way, query peaks that lie far apart from any peak in the database spectrum are penalized and the accuracy of the algorithm is improved. While iterative weighted matching is most effective for identifying peaks that are well separated, noniterative weighted matching (and similarly reverse assignment) is capable of correctly identifying contaminated traces by ignoring outlier peaks in the query trace.

Web Server Implementation. The assignment and matching algorithms were implemented on our COLMAR query Web server at http://spinportal.magnet.fsu.edu/webquery/webquery.html, which runs on an Apache 2.0.52 Web Server using PHP 4.3.11 as a module. Matching and assignment operations are performed by Matlab 7.1 programs that operate in the background. For weighted matching, the Matlab program by A. Melin is used.[24] The COLMAR query is designed for compatibility with the Firefox Web browser, but it also runs under other Web browsers, such as Apple's Safari and Microsoft's Explorer.

RESULTS AND DISCUSSION

Application of the DemixC method to the 2D TOCSY spectrum of the model mixture yields a set consisting of eight 1D traces

(22) Papadimitriou, C. H.; Steiglitz, K. *Combinatorial Optmization, Algorithms and Complexity*; Dover: Mineola, NY, 1998.
(23) Hus, J. C.; Prompers, J. J.; Brüschweiler, R. *J. Magn. Reson.* **2002**, *157*, 119–123.
(24) Melin, A. http://www.mathworks.com/matlabcentral/fileexchange.

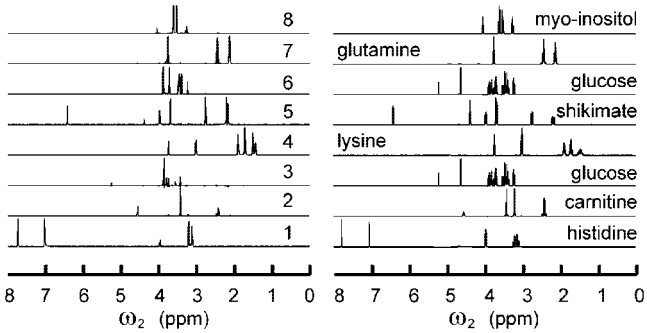

Figure 3. DemixC traces (left) of model mixture with seven compounds vs the best consensus database matches from the BMRB (right).

shown in Figure 3 (left panel) (these are the seven mixture components whereby glucose exists in the two isomeric forms α-D-glucose and β-D-glucose). Each of these traces was subjected to peak picking, and the resulting peak list was screened against the BMRB metabolomics database using the COLMAR query Web server. The results using the different assignment algorithms are summarized in Table 1. Penalty P together with the relative rank of the correct trace is given for all mixture components. The forward assignment algorithm and the iterative weighted matching identify the correct component (highest score) for all eight traces. The other algorithms do not identify the correct compound in all instances, but they are useful for other, less frequently encountered situations discussed below.

Figure 4 shows high scoring BMRB traces for lysine and for myo-inositol (blue traces). The blue and red triangles correspond to the peak lists obtained by peak picking of the DemixC traces and the BMRB traces and they serve as input for the assignment algorithms as implemented in the Web server. For example, peak picking of the DemixC query trace corresponding to myo-inositol yields the following peak list (in units of ppm): 4.05, 3.61, 3.52, and 3.26. When this peak list is entered on the COLMAR query server, one obtains myo-inositol as the best and correct hit (Figure 4F), which is followed by the next best scoring compounds D-glucuronate, D-glucose-6-phosphate, and choline with the associated 1D reference spectra provided in a summary file that can be downloaded for visual inspection.

The choline database spectrum in Figure 4H illustrates the balance between forward and reverse assignment for myo-inositol. Since M (= 3) < N (= 4), the misalignment of peaks is reflected more strongly in the forward assignment (choline ranks eighth) than in the reverse assignment (choline ranks second) with an intermediate consensus score that balances effects due to different numbers of peaks and chemical shift differences.

The scoring of dTMP (Figure 4D) illustrates the discriminatory power of reverse assignment for lysine. The rank of dTMP is 2 for forward assignment because the query peak list of lysine fits a subset of the dTMP database list (large M) well. However, the reverse discriminator penalizes the dTMP resonance outliers and lowers its rank to 9.

The accuracy of peak list matching can be affected by the fact that even in the case of the correct compound, the number of peaks and their positions may differ between the query trace and the 1D database spectrum. Reasons include (i) incomplete TOCSY transfer across the spin system, (ii) differences in peak patterns due to differences in B_0, (iii) contamination of DemixC trace by another spin system due to highly overlapping regions in the TOCSY spectrum, (iv) differences in temperature, pH, and calibration method, and (v) slow isomerism (on the chemical shift time scale) so that the DemixC traces reflect individual isomers whereas the 1D reference spectrum shows their superposition (see, e.g., α-D-glucose and β-D-glucose, traces 3 and 6 in Figure 3). It is thus useful to use different scoring functions so that these situations can be recognized and adequately addressed.

A major difference between the different algorithms concerns their scoring when the chemical shift difference matrix $\Delta^{(n)}$ is nonquadratic, i.e., when the number of query peaks differs from the number of database peaks ($N \neq M$). This situation is common when a query trace is screened against a spectral database. Even when the query trace is compared with the correct database entry, the $\Delta^{(n)}$ matrix may be nonquadratic because of incomplete magnetization transfer across the spin system during TOCSY mixing or because of slow isomerism, which reduces the number peaks observed in a DemixC trace. For this reason, the algorithms of eqs 1−3 have been modified and applied in various combinations to provide a P score that reflects the bidirectional fit of the TOCSY trace and the reference spectra. For example, the mixed 90%/10% forward/reverse algorithm uses the score

$$P_{9010} = 0.9P_{\text{forw}} + 0.1P_{\text{rev}} \tag{5}$$

This algorithm uses matching in both directions, and its score penalizes outliers in both the query trace and the reference spectrum. Forward assignment, on the other hand, assigns each resonance of the query spectrum to the closest resonance of the database spectrum but not necessarily vice versa. This makes forward assignment insensitive to missing peaks ($N < M$) (e.g., because of spectral overlaps or incomplete TOCSY transfers). On the other hand, reverse assignment is particularly useful as part of the $P_{90/10}$ score where it penalizes outlier peaks in the reference spectrum.

The differential performance of each of the algorithms for the model mixture is illustrated in Table 1. In some cases, such as myo-inositol and glutamine, the number of query peaks N is sufficiently small (4 and 3, respectively) to allow both the forward and reverse algorithms to find the correct reference spectrum. Reverse matching works well, primarily because none of the BMRB spectra with a small number of peaks M agrees better with a subset of the peaks of these two compounds. As metabolomic databases are increasing in size, this situation is likely to change. In this case, the balance between forward and reverse assignment provided by the 90/10 algorithm is expected to improve the robustness.

The different scores have been merged into a consensus score to provide a more robust ranking of the database compounds. The consensus score uses the average ranks of the three algorithms forward assignment, mixed 90%/10% forward/reverse assignment, and iterative matching. Table 1 illustrates how the consensus score can prevent false compound identification. Forward assignment algorithms generally rank those database traces higher that have a large number of peaks (large M), because the chances of a good match are increased. Conversely, reverse assignment algorithms tend to rank database traces higher with a small number of peaks (low M). Therefore, by combining

Table 1. Scoring Results of COLMAR Query Web Server Using BMRB Database

compound[a]	matching P_{wm}[b]	matching rank[c]	forward P_{forw}[b]	forward rank[c]	reverse P_{rev}[b]	reverse rank[c]	$P_{90/10}$[b]	90/10 rank[c]	iterative matching P_{wm}[b]	iterative rank[c]
myo-inositol	0.0072	1	0.0072	1	0.0072	1	0.0072	1	0.0072	1
β-glucose	0.1141	9	0.0980	1	1.9046	111	0.2787	2	0.1141	1
histidine	0.1275	8	0.1275	1	0.1275	8	0.1275	1	0.1275	1
α-glucose	0.1988	20	0.0940	1	0.8853	56	0.1731	1	0.1988	1
glutamine	0.0134	1	0.0134	1	0.0134	1	0.0134	1	0.0134	1
lysine	0.0555	3	0.1351	1	0.0555	3	0.1272	1	0.1351	1
carnitine	0.0235	2	0.0235	1	0.2275	10	0.0439	1	0.0235	1
shikimic acid	0.0349	1	0.0349	1	0.0580	3	0.0372	1	0.0349	1

[a] Compounds of model mixture. [b] Penalty P of BMRB spectrum of the same compound. [c] Scoring rank of BMRB spectrum of the same compound (1 = correct identification, 2 = second rank, etc.).

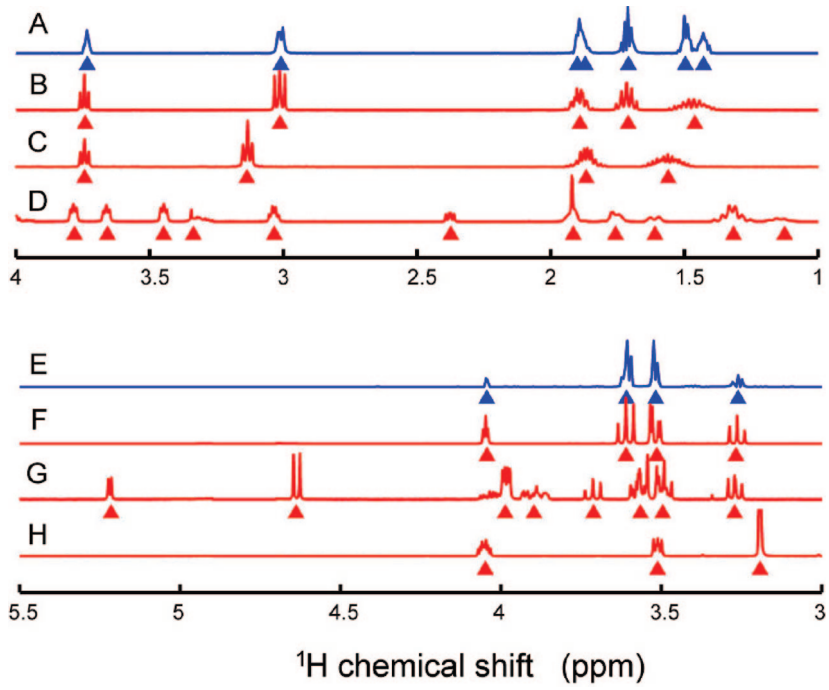

Figure 4. DemixC traces of A) lysine and (E) myo-inositol and their top rank-ordered scoring database spectra. For lysine, the top three database entries are (with decreasing rank order) (B) lysine, (C) D-citrulline, and (D) dTMP, and for myo-inositol they are (F) myo-inositol, D-glucuronate (not shown), (G) D-glucose-6-phosphate, and (H) choline. The triangles below each trace correspond to the chemical shift list obtained by peak picking.

forward assignment, iterative matching, and 90%/10% mixed assignment, consensus scoring balances the different effects. Because each of these algorithms includes a significant forward component, compound recognition generally works best if the number of false positive peaks in the query trace is small.

In the case of D-glucose, which exists as two slowly interconverting isomers, the 90/10 algorithm ranks the correct assignment as number 2. This is because each of the glucose isomers has only half of the chemical shifts found in the database spectrum, which shows the spectral superposition of both isomers. While the forward algorithm is insensitive to this property, the reverse algorithms penalize each unmatched shift in the database spectrum. A possible solution for the unambiguous identification of compounds that exhibit slow isomerism or incomplete magnetization transfer during TOCSY is the generation of a database consisting of 1D TOCSY traces of these compounds recorded at suitable mixing times.

It is anticipated that the ongoing expansion of NMR spectral databases will have different effects on the various query algorithms. As the number of reference spectra with similar shift values will increase, there will be more compounds with close scores in the database queries and the ability to predict the correct reference will become harder. In particular, reverse assignment will be more ambiguous as the chances increase that a compound closely matches a subset of peaks of the query peak list, which will have an adverse effect on the compound identification in contaminated traces. On the other hand, reverse matching is expected to remain a valuable component of the mixed forward/reverse assignment scores, eq 5.

CONCLUSION

We have introduced COLMAR query as a Web server based rapid screening method of NMR traces against an NMR spectral library. A chemical shift list of the query trace is first generated and then compared with the chemical shift lists of the compounds of the database. Because of the potentially imperfect nature of both the query chemical shift list as well as the automatically picked chemical shift list of the database spectra, three comple-

mentary scoring algorithms and their combinations are applied where each algorithm produces a rank-ordered list of the best scoring compounds. The three individual scores are then combined into a consensus score, and the spectra of the top scoring compounds superimposed on the chemical shifts of the query list can be downloaded for visual inspection.

The COLMAR query method has been demonstrated here using TOCSY traces identified by DemixC, but other types of 1D NMR traces including selective 1D TOCSY spectra,[25,26] NMR spin-wave traces,[27] DOSY traces,[8] or suitably chosen cross sections through other homonuclear and heteronuclear 2D NMR spectra can be analyzed as well (see COLMAR query Web site for examples). Together with the steadily growing NMR metabolomics databases, this Web server tool is expected to substantially facilitate and speed up the identification of metabolites of a wide range of biological mixtures.

(25) Kessler, H.; Oschkinat, H.; Griesinger, C.; Bermel, W. *J. Magn. Reson.* **1986**, *70*, 106–133.
(26) Sandusky, P.; Raftery, D. *Anal. Chem.* **2005**, *77*, 2455–2463.
(27) Madi, Z. L.; Brutscher, B.; Schulte-Herbruggen, T.; Brüschweiler, R.; Ernst, R. R. *Chem. Phys. Lett.* **1997**, *268*, 300–305.

ACKNOWLEDGMENT

We thank Art S. Edison for valuable discussion. S.L.R. is grateful for a REU NSF and a University of Florida HHMI fellowship during this work. The NMR experiments were conducted at the National High Magnetic Field Laboratory (NHMFL) supported by cooperative agreement DMR 0654118 between the NSF and the State of Florida. This work was supported by the National Institutes of Health (Grant R01 GM 066041).

Received for review December 13, 2007. Accepted March 5, 2008.

AC702530T