Automated Profiling of Metabolomics Mixtures by NMR Spectral TOCSY Traces

Steven L Robinette[1,2], Fengli Zhang[1], Lei Brüschweiler-Li[1,3], and Rafael Brüschweiler[1,3]

[1]National High Magnetic Field Laboratory, Tallahassee, FL
[2]Department of Biochemistry and Molecular Biology, University of Florida, Gainesville, FL 32611
[3]Department of Chemistry and Biochemistry, Florida State University, Tallahassee, FL 32306

SUMMARY:

A procedure for automating the profiling of complex Metabolomics mixtures by Covariance Total Correlation Spectroscopy (TOCSY) NMR is presented. By selecting TOCSY traces from the spectrum of a complex mixture, the identification of compounds is achieved by the development of a query which ranks the compounds of a Metabolomics Database in order of similarity to the experimental spectrum. Development centers around the adoption and evaluation of several algorithms for matching experimental chemical shifts against a set of ^{1}H NMR spectra and the quantification of their similarity by assignment and heuristic methods. The effectiveness of this method is tested by creating a model Metabolomics mixture. All the components of the mixture are successfully identified by the query though some re-assembly of spin systems is necessary due to TOSCY spin system splitting. These results suggest that the query will be an effective tool for profiling complex Metabolomics mixtures in cases where the database includes spectral data for the target compound. Deployment to the web is in progress.

Introduction:

The emerging field of Metabolomics is based on the identification of chemical components in complex mixtures

- NMR and MS are currently the two techniques used to accomplish this goal and are often used in conjunction with each other.
- Mixture profiles, the identification of all components in a solution, are very difficult to obtain by either technique
- Recent developments in Covariance NMR and TOCSY tracing[5] allow the deconvolution of a mixture's spectrum by spin system without physical separation of compounds

Automated Profiling of Complex Mixtures would allow significant increases in Metabolomics throughput

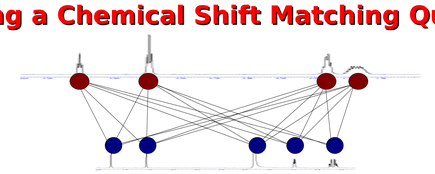
As such, the goal of this project is to develop a computational platform which would allow the comparison of both traditional single compound proton spectra and TOCSY traces with growing ^{1}H NMR databases

Databases:

Several databases have begun supporting Metabolomics sections with NMR spectral information on metabolites

- **Biological Magnetic Resonance Bank[1] (BMRB)**
  - -Offers comprehensive documentation including Bruker ser files of TOCSY, COSY, HSQC, and ^{1}H NMR of ~275 metabolites
  - -Small database, but information rich and easy to mine
- **Human Metabolome Database[2] (HMDB)**
  - -Offers ^{1}H NMR spectra for ~500 compounds
  - -Larger database, but spectra are harder to extract
- **Spectral Database for Organic Compounds (SDBS)**
  - -Offers ^{1}H NMR spectra for ~14,500 compounds
  - -Huge database, but specifically prohibits mining

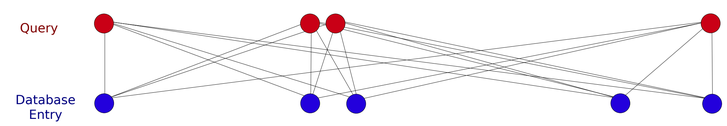
Developing a Chemical Shift Matching Query:



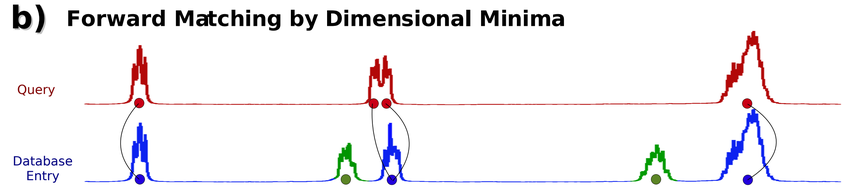
Scoring Problem: Determine matches and scoring by minimizing distance between shifts, but do so in such a way as to balance asymmetry, multiple matching, and distance. Ensure that the resultant score penalizes outliers for resolution, has enough tolerance for TOCSY isomers, but does not favor single shift references

Algorithms in Action:

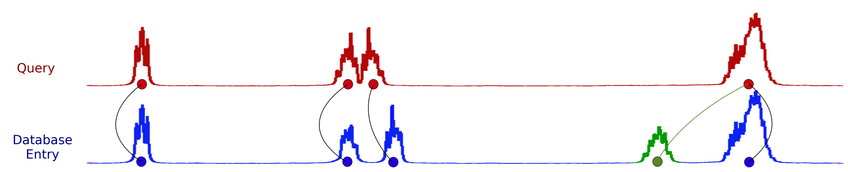
a) Complete Bipartite Graph[3]



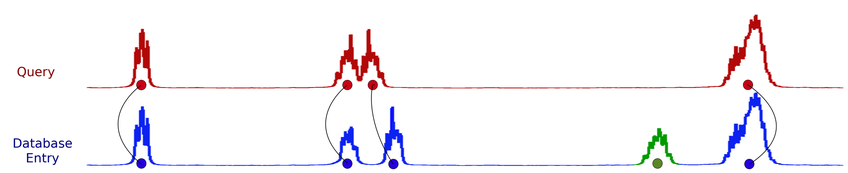
b) Forward Matching by Dimensional Minima



c) Reverse Matching by Dimensional Minima



d) Kuhn's Hungarian Algorithm[4]



The figure above is a visual representation of the necessity of a multiple algorithm solution. Sections b-d show algorithmic concepts used for automated matching in cases where each algorithm interprets the complete bipartite graph (shown in a) in a way which represents a correct spectral interpretation .

Algorithmic Action and Specialty

b) Forward Matching by the Dimensional Minima algorithm calculates the nearest reference analog to each experimental chemical shift by projecting the experimental peaklist onto the reference. This case shows that Forward Matching compensates for differences in fine structure.

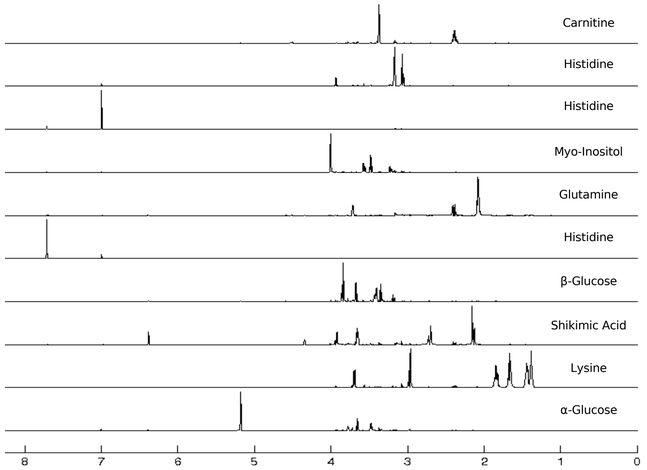
c) Reverse Matching by the Dimensional Minima algorithm calculates the nearest experimental analog to each reference shift. This is particularly useful in resolving out reference spectra with true outliers.

d) Assignment by Kuhn's Hungarian Algorithm is the canonical algorithm for bipartite matching. While it cannot deal with unbalanced graphs, it does not allow multiple matching and is very well studied.

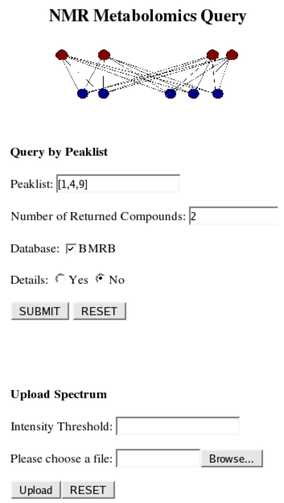
Due to the complexity of matching TOCSY traces against ^{1}H NMR spectra, no single algorithm was able to find the best match in all possible cases. As such, we adopted and made slight modifications to Kuhn's canonical Hungarian Algorithm for the assignment problem and developed a second "Dimensional Minima" algorithm to deal with NMR specific problems and to add increased resolution to the query. These three iterations test spectral fit by different criteria and the consensus score is used to determine the best database match.

Model Mixture:

The following spectra are Covariance TOCSY traces of the 10mM model mixture selected by the DemixC[5] protocol. Chemical shifts were selected at an intensity threshold of 10% and with a multiplet clustering of 0.03 ppm. Each trace was matched to a model compound, though several traces represented the same compound due to spin system splitting.



Source and Web Deployment:




```
function result=webquery
(peaklist,database,numreturns,details)

a=peaklist;
load(database)

for b=1:size(compounds,1)
    clear c
    c=peaks.(strcat('bmrb_',deblank
(compounds(b,:))));
    match1(b)=algorithm(a,c);
    match2(b)=doublealgorithm(a,c);
    match3(b)=hungry(a,c);
    Ny(b)=length(c);
end

[d,i]=sort(mean(k,1));
d=round(d(1:numreturns));

for b=1:numreturns
    matched1(b)=match1(i(b))/length(a);
    matched2(b)=match2(i(b))/(length(a)
    matched3(b)=match3(i(b))/length(a);
    blank(b,:)=blanks(2);
    Ny2(b)=Ny(i(b));
end

d=d';

if (length(details)==2)
    result=horzcat(j(:,:),num2str(d));
else
    matched=mean([])
    result=horzcat(j(:,:),num2str
end
```

References:

1. Seavey, B.R., Farr, E.A., Westler, W.M., Markley, J.L., "A Relational Database for Sequence-Specific Protein NMR Data," J. Biomolecular NMR 1, 217-236 (1991).
2. Wishart DS et al., HMDB: The Human Metabolome Database. Nucleic Acids Research 2007 Jan;35(Database issue):D521-6.
3. Papadimitriou, Chistos H. and Steiglitz, Kenneth. *Combinatorial Optimization Algorithms and Complexity*. Dover Publications, Inc., New York, 1982.
4. Kuhn, H.W., "The Hungarian Method for the Assignment Problem," *Naval Research Logistics Quarterly*, 2 (1955), 83-97.
5. F. Zhang and R. Brüschweiler, Robust deconvolution of complex mixtures by covariance TOCSY spectroscopy, Angewandte Chemie Int. Ed. Engl. 46, 2639-2642 (2007).

Acknowledgments:

We acknowledge funding from the National High Magnetic Field Laboratory's Research Experiences for Undergraduates (REU) Program and from the undergraduate Science for Life Program, an interdisciplinary program in the life sciences with support from the Howard Hughes Medical Institute