WHY DO WE NEED AN AUTOMATED NMR ANNOTATION TOOL
Metabolomic workflow

Control mice

Exposed mice

Urine, plasma, tissues...

Analytical analysis (NMR, MS...)

NMR and MS spectra

Metabolites Identification

Glucose
Taurine
Glycine
Glutamate
Lactate
Lysine
Leucine

Multivariate Statistical Analysis
Glycine

- Each group of protons gives one signal
- A singlet for Glycine
• 1 signal is observed for each type of proton
• 18 different signals for this compound
• Complex spectrum
1H NMR spectrum of biological matrices

- No separation of compounds before NMR analysis
- A metabolite can have many signals at different chemical shifts
- Many signals overlap

Identification of metabolites in mixtures is really complex
Metabolite identification (manually)

- Based on 1H chemical shift, coupling pattern and coupling constants
- Comparison with annotated spectra of similar matrices in literature
- Spiking: a reference compound is added in the sample
- Comparison with NMR spectra of reference compounds
  - database in-house
  - Chenomx (commercial)
  - HMDB (freeware)
- Multi-dimensional NMR
2D NMR experiments mostly used

- $^1$H-$^1$H COSY (Correlation Spectroscopy) experiment: correlations $^1$H-$^1$H via 3 bonds
- $^1$H-$^1$H TOCSY (TOtal Correlation Spectroscopy) experiment: correlations between all protons within a given spin system
- $^1$H-$^{13}$C HSQC (Heteronuclear Single Quantum Coherence) experiment: correlations $^1$H-$^{13}$C via 1 bond
- $^1$H-$^{13}$C HMBC (Heteronuclear Multiple Bond Coherence) experiment: correlations $^1$H-$^{13}$C via 2, 3 or 4 bonds
Example: $^1$H NMR spectrum of aqueous pup extract (PND2) of rat
2D $^1$H-$^1$H COSY NMR spectrum of aqueous pup extract (PND2)

- CH$_3$-CH(OH)COOH
- 4.11 ppm
- 1.33 ppm

- 1.33 ppm
- 4.11 ppm
- 3.78 ppm

- lactate
- alanine
- lysine
- glutamate
- glutathione
- glutamine
- leucine
- valine
- threonine
- choline
- taurine
- glucose
- citrate
- β-glucose
2D $^1$H-$^{13}$C HSQC NMR spectrum of aqueous pup extract (PND2) of rat
Comparison with NMR spectra of reference compounds:
Human Metabolome database

www.hmdb.ca/
Many responses (299) : the expert must check all the spectra to find the right metabolite.
Comparison with NMR spectra of reference compounds: Human Metabolome database
Comparison with NMR spectra of reference compounds: Human Metabolome database

<table>
<thead>
<tr>
<th>Name</th>
<th>Formula Weight</th>
<th>Structure</th>
<th>Library Matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-Lactic acid (HMDB001311)</td>
<td>0.0770</td>
<td><img src="structure1.png" alt="Structure" /></td>
<td>1:2</td>
</tr>
<tr>
<td></td>
<td>10226-41-7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-Octenolic acid (HMDB00282)</td>
<td>142.1900</td>
<td><img src="structure2.png" alt="Structure" /></td>
<td>1:7</td>
</tr>
<tr>
<td></td>
<td>1871-67-6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5α-Androstan-3β,17β-diol (HMDB00493)</td>
<td>292.4562</td>
<td><img src="structure3.png" alt="Structure" /></td>
<td>1:25</td>
</tr>
<tr>
<td></td>
<td>571-20-0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5α-Cholestanone (HMDB00871)</td>
<td>386.6530</td>
<td><img src="structure4.png" alt="Structure" /></td>
<td>1:29</td>
</tr>
<tr>
<td></td>
<td>566-85-1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

13 results
Comparison with NMR spectra of reference compounds:
Database in-house

- Glutamate
- Valine
- Alanine
- Taurine
- Creatine
- Lactate

Aqueous pup extract
Metabolite identification (manually)

- Time-consuming
- Complicated
- Some resonances are not identified
- An automated NMR annotation tool will be helpful:
  - Package R Batman: very complex and not user-friendly
  - BAYESIL (http://bayesil.ca/): web interface (limited to serum, plasma, and CSF biofluids)
  - Chenomx: commercial software


We developed a new tool for the annotation of NMR spectra: ASICS (Automatic Statistical Identification in Complex Spectra)
Objectives

- Automated identification of compounds in NMR spectra of complex mixtures
- Estimation of the proportion of the metabolites in the mixture
Modeling

Mixture Y spectrum

Library Z1, ..., Zn

Problems: noise and peaks warping
Noise Modeling

Several NMR spectra of glucose allow to model the noise
Warping problems

- Metabolite concentration is different in pure compound spectrum and in mixture spectrum: peaks do not have necessary the same shape.

- Experimental conditions are different between pure compound and mixture (pH, ionic strength, analytical variability...) : peaks can be shifted.
Warping modeling

Two NMR spectra of a same mixture obtained in different conditions
Warping modeling

Before modeling

After modeling

NMR spectra of choline pure or in the mixture

We used a warping function: $\Phi$
Warping

- Shift depend on pH: shift is not the same for all compounds
- Deformations are localized
- Peaks can be shifted, but also expanded or contracted
- Maximum deformation given by the expert
- NMR spectra are recorded at pH 7 to limit the shifts
Metabolite spectrum modeling

- \( f_1(t) \): NMR spectrum of pure compound
- NMR spectrum observed with noise

\[
Z_1(t) = f_1(\phi(t)) + \varepsilon_1(t)\sqrt{f_1(\phi(t))} + \varepsilon_2(t)
\]

- Library \( \{Z_1, \ldots, Z_{200}\} \)
A mixture $g$ is modeled as a positive function $g(t)$:

$$g(t) = \alpha_1 f_1(t) + \cdots + \alpha_{200} f_{200}(t)$$

Signal is noised and warped:

$$Y(t) = g\left(\phi(t)\right) + \varepsilon_1(t)\sqrt{g(\phi(t))} + \varepsilon_2(t)$$

$\alpha$ is the relative proportion of metabolite in the mixture spectrum: area under the curve of the metabolite divided by the total area of the spectrum.

- Determine non-zero proportions
- Estimate the proportions: quantification
Method

- A false detection is a metabolite identified but this metabolite is not present in the mixture.
- The detection threshold is the proportion from which we are sure that the metabolite is identified.
- Compromise between low probability of false detection and low threshold.
- In this method:
  - Control the probability to obtain one or several false detections.
  - Give the detection threshold of each metabolite.
  - Minimize the detection threshold.
Real data set

A mixture of 6 metabolites:

<table>
<thead>
<tr>
<th>Metabolites</th>
<th>Proportions</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-Glucose</td>
<td>5.8 %</td>
</tr>
<tr>
<td>Creatinine</td>
<td>21.0 %</td>
</tr>
<tr>
<td>L-Phenylalanine</td>
<td>3.3 %</td>
</tr>
<tr>
<td>L-Proline</td>
<td>6.7 %</td>
</tr>
<tr>
<td>Benzoic acid</td>
<td>8.4 %</td>
</tr>
<tr>
<td>Choline chloride</td>
<td>54.6 %</td>
</tr>
</tbody>
</table>

Library of 36 metabolites
Automatic cleaning

Black : mixture spectrum; Red : hydroxybenzoic spectrum

Elimination of 18 metabolites
## Identification

<table>
<thead>
<tr>
<th>Metabolites</th>
<th>Identification</th>
<th>Detection Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benzoic acid</td>
<td>Yes</td>
<td>1.6 %</td>
</tr>
<tr>
<td>Choline</td>
<td>Yes</td>
<td>2.3 %</td>
</tr>
<tr>
<td>Creatinine</td>
<td>Yes</td>
<td>1.1 %</td>
</tr>
<tr>
<td>Glucose</td>
<td>Yes</td>
<td>3.1 %</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>Yes</td>
<td>2.1 %</td>
</tr>
<tr>
<td>Proline</td>
<td>Yes</td>
<td>2.6 %</td>
</tr>
<tr>
<td>Acetic acid</td>
<td>No</td>
<td>0.9 %</td>
</tr>
<tr>
<td>Galactose</td>
<td>No</td>
<td>2.5 %</td>
</tr>
<tr>
<td>Mannitol</td>
<td>No</td>
<td>1.6 %</td>
</tr>
<tr>
<td>Mannose</td>
<td>No</td>
<td>2.4 %</td>
</tr>
<tr>
<td>Autres</td>
<td>No</td>
<td>&lt; 2.5 %</td>
</tr>
</tbody>
</table>
## Quantification

<table>
<thead>
<tr>
<th>Metabolites</th>
<th>Estimated proportions</th>
<th>Real proportions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benzoic acid</td>
<td>5.7 %</td>
<td>8.4 %</td>
</tr>
<tr>
<td>Choline</td>
<td>55.8 %</td>
<td>54.6 %</td>
</tr>
<tr>
<td>Creatinine</td>
<td>13.3 %</td>
<td>21 %</td>
</tr>
<tr>
<td>Glucose</td>
<td>5.5 %</td>
<td>5.8 %</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>3.9 %</td>
<td>3.3 %</td>
</tr>
<tr>
<td>Proline</td>
<td>5.7 %</td>
<td>5.9 %</td>
</tr>
</tbody>
</table>

Detection thresholds are too high.
## Improvement of detection thresholds

<table>
<thead>
<tr>
<th>Metabolites</th>
<th>Identification</th>
<th>New thresholds</th>
<th>Previous thresholds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetic acid</td>
<td>No</td>
<td>0.16 %</td>
<td>0.9 %</td>
</tr>
<tr>
<td>Galactose</td>
<td>No</td>
<td>0.19 %</td>
<td>2.5 %</td>
</tr>
<tr>
<td>Mannitol</td>
<td>No</td>
<td>0.24 %</td>
<td>1.6 %</td>
</tr>
<tr>
<td>Mannose</td>
<td>No</td>
<td>0.17 %</td>
<td>2.4 %</td>
</tr>
<tr>
<td>Others</td>
<td>No</td>
<td>&lt; 0.68%</td>
<td>&lt; 2.5 %</td>
</tr>
</tbody>
</table>
Synthesis Urine

• Mix in 500 ml of water:
  - 3.8 g of potassium chloride
  - 8.5 g of sodium chloride
  - 24.5 g of urea
  - 1.03 g of citric acid
  - 0.34 g of ascorbic acid
  - 1.18 g of potassium dihydrogenophosphate
  - 1.4 g of creatinine
  - 0.64 g of sodium hydroxide
  - 0.47 g of sodium bicarbonate
  - 0.28 ml of phosphoric acid
Synthesis Urine

- Addition of metabolites:
  - Hippuric acid : 8.2 mM
  - TMAO : 7 mM
  - acetic acid : 0.8 mM
  - Alanine : 1 mM
  - Betaine : 1 mM
  - Carnitine : 0.7 mM
  - Dimethylamine : 1 mM
  - Ethanolamine : 1.5 mM
  - Formate : 0.4 mM
  - Glucose : 1 mM
  - Glutamine : 2 mM
  - Glycine : 0.6 mM
  - Guanidinoacetate : 0.8 mM
  - Lactate : 0.6 mM
  - Lysine : 1 mM
  - Malonate : 1.7 mM
  - Trigonelline : 0.6 mM
  - Tyrosine : 0.3 mM

21 metabolites detectable in NMR
$^1$H NMR spectrum
## Annotation: comparison with other tools

<table>
<thead>
<tr>
<th></th>
<th>True Positive</th>
<th>False Positive</th>
<th>False Negative</th>
<th>True Negative</th>
<th>Accuracy (%)</th>
<th>Compounds in library</th>
<th>Computing Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASICS</td>
<td>17</td>
<td>10</td>
<td>4</td>
<td>145</td>
<td>92</td>
<td>176</td>
<td>2 mn 38s</td>
</tr>
<tr>
<td>MetaboHunter</td>
<td>4</td>
<td>51</td>
<td>17</td>
<td>795</td>
<td>92</td>
<td>867</td>
<td>&lt; 1 mn</td>
</tr>
<tr>
<td>Batman</td>
<td>21</td>
<td>125</td>
<td>0</td>
<td>1</td>
<td>18</td>
<td>147</td>
<td>74 hours</td>
</tr>
<tr>
<td>Bayesil</td>
<td>12</td>
<td>17</td>
<td>7</td>
<td>53</td>
<td>73</td>
<td>89</td>
<td>10 mn 48s</td>
</tr>
<tr>
<td>Chenomx</td>
<td>15</td>
<td>48</td>
<td>6</td>
<td>269</td>
<td>54</td>
<td>338</td>
<td>&lt; 1 mn</td>
</tr>
</tbody>
</table>

ASICS detects 17 metabolites among 21, but there are 10 false positive. The metabolites must be checked by the NMR expert.

The accuracy is acceptable compared to other tools.
The quantifications provided by ASICS or by Chenomx both fit quite well the order of magnitude of the real proportion of the different metabolites, but these quantifications are not accurate.
Conclusion

• The program is powerful for detecting the presence of compounds in mixture
• The program is not very powerful for estimating the true proportion, but we are working on this problem
• At the moment, pH of compounds in the library is 7, but we can import spectra at different pH
• At the moment, we have 175 compounds in the library but we will record more compounds (MetaboHub)
• More tests are needed (other mixtures)
• This tool can be used to help the expert to annotate the metabolites in a complex mixture
• Publication has been submitted
HOW TO DO WITH GALAXY?
FORM

- Name of the zip file: 1lume-Synthese.zip
- Exclusion zone
- Maximum shift
- Execute
MATRIX RESULT

Metabolites names

Relative concentrations
GRAPH RESULT
GRAPH RESULT
GRAPH RESULT
Example: Synthesis Urine

- Upload the $^1$H NMR spectrum of synthesis urine
- Region to exclude: 6.5-4.5 ppm for urine
- Determine the metabolites present in this sample
- Display the spectra
- Compare the results with the list of metabolites present in the mixture
Example: Synthesis Urine

List of metabolites present in the sample:

- Creatinine (24.8 mM; 0.378)
- Citric acid (10.7 mM; 0.163)
- Hippuric acid (8.2 mM; 0.125)
- TMAO (7 mM; 0.107)
- Glutamine (2 mM; 0.030)
- Malonate (1.7 mM; 0.026)
- Ethanolamine (1.5 mM; 0.023)
- Alanine (1 mM; 0.015)
- Betaine (1 mM; 0.015)
- Dimethylamine (1 mM; 0.015)
- D-Glucose (1 mM; 0.015)
- Lysine (1 mM; 0.015)
- Acetic acid (0.8 mM; 0.012)
- Guanidinoacetate (0.8 mM; 0.012)
- Carnitine (0.7 mM; 0.011)
- Glycine (0.6 mM; 0.009)
- Lactate (0.6 mM; 0.009)
- Trigonelline (0.6 mM; 0.009)
- Formate (0.4 mM; 0.006)
- Tyrosine (0.3 mM; 0.004)